Yellowbrick Documentation

发布 *v0.5*

The scikit-yb developers

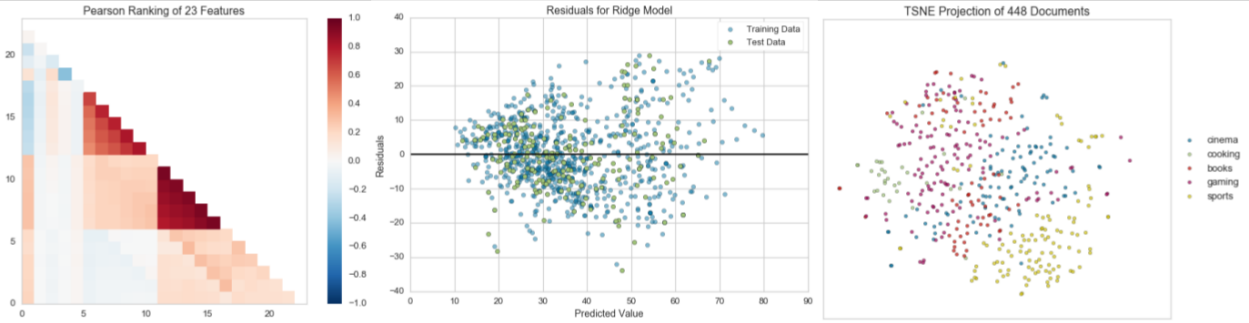
2019 年 04 月 07 日

Contents

欢迎来到 Yellowbrick!

我们最近正在将文档翻译成中文中，请稍后再来。

并且，我们总是在寻求帮助。如果你愿意帮助我们翻译，请向 yellowbrick-docs-zh 提交一个 pull request。如果你对支持 Yellowbrick 感兴趣，请向 codebase 提交一个 pull request。



Yellowbrick 是由一套被称为”Visualizers” 组成的可视化诊断工具组成的套餐，其由 Scikit-Learn API 延伸而来，对模型选择过程其指导作用。总之，Yellowbrick 结合了 Scikit-Learn 和 Matplotlib 并且最好得传承了 Scikit-Learn 文档，对 你的模型进行可视化！想要更多地了解 Yellowbrick，请访问*About* 。

如果你第一次接触 Yellowbrick，请查看快速开始 或者直接跳到模型选择教程 。Yellowbrick 是一个丰富的库，并且定期加入多个 Visualizers。想要对了解特定 Visualizers 的更多细节并且扩展对其使用，请前往*Visualizers and API* 。如果你想对 Yellowbrick 作出贡献，请查看 contributing guide 。如果你已经报名参加用户测试，请前往*User Testing Instructions* （谢谢!）。

CHAPTER 1

Visualizers

Visualizers 也是 estimators（从数据中习得的对象），其主要任务是产生可对模型选择过程有更深入了解的视图。从 Scikit-Learn 来看，当可视化数据空间或者封装一个模型 estimator 时，其和转换器（transformers）相似，就像"ModelCV"（比如 RidgeCV, LassoCV ）的工作原理一样。Yellowbrick 的主要目标是创建一个和 Scikit-Learn 类似的有意义的 API。其中最受欢迎的 visualizers 包括：

1.1 特征可视化

- *Rank Features*: 对单个或者两两对应的特征进行排序以检测其相关性
- *Parallel Coordinates*: 对实例进行水平视图
- *Radial Visualization*: 在一个圆形视图中将实例分隔开
- *PCA Projection*: 通过主成分将实例投射
- *Feature Importances*: 基于它们在模型中的表现对特征进行排序
- *Scatter and Joint Plots*: 用选择的特征对其进行可视化

1.2 分类可视化

- *Class Balance*: 看类的分布怎样影响模型
- *Classification Report*: 用视图的方式呈现精确率，召回率和 F1 值
- *ROC/AUC Curves*: 特征曲线和 ROC 曲线子下的面积

- *Confusion Matrices*: 对分类决定进行视图描述

1.3 回归可视化

- *Prediction Error Plot*: 沿着目标区域对模型进行细分
- *Residuals Plot*: 显示训练数据和测试数据中残差的差异
- *Alpha Selection*: 显示不同 alpha 值选择对正则化的影响

1.4 聚类可视化

- *K-Elbow Plot*: 用肘部法则或者其他指标选择 k 值
- *Silhouette Plot*: 通过对轮廓系数值进行视图来选择 k 值

1.5 文本可视化

- *Term Frequency*: 对词项在语料库中的分布频率进行可视化
- *t-SNE Corpus Visualization*: 用随机邻域嵌入来投射文档

... 以及更多！Visualizers 随时在增加中，请务必查看示例（甚至是 develop branch 上的），并且随时欢迎你对 Visualizers 贡献自己的想法。

CHAPTER 2

获取帮助

Yellowbrick 是一个传承自 Matplotlib 和 Scikit-Learn 的热情包容的项目。和这些项目一样，我们遵守 Python Software Foundation Code of Conduct 。如果需要帮助、或者想要对项目进行贡献、或者发现有漏洞需要报告，请不要犹豫，随时和我们联系。

寻求帮助最主要的方法是在我们的 Google Groups Listserv 上发帖。这是社区会员可以加入以及互相回应的一个邮件列表/论坛；在这里你应该能得到最快的回应。希望你也能考虑加入这个组，这样你也可以回答问题！你也可以在 Stack Overflow and tag them with "yellowbrick". Or you can add issues on GitHub. You can also tweet or direct message us on Twitter @DistrictDataLab 上问问题。

CHAPTER 3

开源

Yellowbrick license 使用开源 Apache 2.0 许可证。Yellowbrickx 拥有一个非常活跃的开发者社区；请考虑加入他们并且 贡献 ！

Yellowbrick 在 GitHub 上托管。issues 和 pull requests 都记录在上面。

CHAPTER 4

目录

这个版本库的 Yellowbrick 文档的完整清单如下：

4.1 快速开始

如果你对 Yellowbrick 还不熟悉，这个教程可以帮助你很快上手将可视化运用到你机器学习的流程中去。不过在我们开始用 Yellowbrick 之前，有几个开发环境相关的问题需要注意。

Yellowbrick 主要依赖于两个包: Scikit-Learn 和 Matplotlib. 如果你没有安装这两个包也没关系，当你安装 Yellowbrick 的时候，它会帮你将它们一起装上。需要注意的是，要想 Yellowbrick 达到最佳效果，最好是结合 Scikit-Learn 0.18 和 Matplotlib 2.0 及以上版本使用。因为上述两个包都通过 C 语言编译，在某些系统上（比如 Windows）安装时可能会有一些困难。如果你安装有困难，可以使用 Anaconda 等其他版本 Python。

4.1.1 安装

Yellowbrick 虽然和 Python 2.7 及以后版本也兼容，但是如果你想更好得利用其全部功能，建议其与和 Python 3.5 及以后版本一起使用。安装 Yellowbrick 最简单的方法是从 PyPI_ 用 pip_（Python 包安装的首选安装程序）安装。

```
$ pip install yellowbrick
```

需要注意的是 Yellowbrick 是一个在建的项目，目前常规发布新的版本，并且每一个新版本都将会有新的可视化功能更新。为了将 Yellowbrick 升级到最新版本，你可以用如下 pip 命令.

```
$ pip install -u yellowbrick
```

你也可以用 -u 标记对 Scikit-Learn，matplotlib 或者其他和 Yellowbrick 兼容的第三方包进行升级.

如果你使用的是 Windows 或者 Anaconda，你也可以充分利用 conda 的功能安装 Anaconda Yellowbrick package：

```
conda install -c districtdatalabs yellowbrick
```

一旦安装好，不管你是在 Python 内部还是在 Jupyter notebooks 上运行 Yellowbrick 应该就没问题了。需要注意的是，因为 Yellowbrick 用的是 matplotlib，其并不能在虚拟环境中运行。如果你一定要用的话，可能需要费一些周折。

4.1.2 使用 Yellowbrick

为了更好得配合 Scikit-Learn 一起使用，我们特意对 Yellowbrick API 进行了一些特殊设计。当然其最主要的接口就是“Visualizer”——一个可以运用数据产生图片的对象。visualizer 是一系列 Scikit-Learn 的 Estimator 对象并且和画图的方法有很多接口。和用 Scikit-Learn 建模相同的流程相同，用 visualizer 也需要先将其载入，对其初始化，调用其 “fit()“方法，然后调用其 “poof()“方法——然后就是见证奇迹的那一刻了。

比如，有很多 visualizer 可以作为转换器来使用，用于在模型拟合之前对特征进行分析。下面这个例子展示的就是如何用平行坐标的方法对高纬数据进行作图。

```
from yellowbrick.features import ParallelCoordinates

visualizer = ParallelCoordinates()
visualizer.fit_transform(X, y)
visualizer.poof()
```

正如你所看到的一样，这个工作流程和用 Scikit-Learn 的转换器是一样的，并且 visualizer 的目的就是要和 Scikit-Learn 的应用程序相整合。和 Scikit-Learn 模型中的超参数一样，visualizer 的参数也可以在其实例化同时就就传递进去，而这个参数可以决定画图的方式。

> poof() 方法用来完成最终的绘画（加标题，轴标签等等），然后根据你自己的要求对其进行渲染。如果你用的是 Jupyter notebook 的话，立刻就应该看到图。如果你运行的是 Python 脚本的话，图片将会在一个图形界面窗口以交互图片形式显示。当然，你也可以将图片传递到一个文件路径，将其保存在本地磁盘：

```
visualizer.poof(outpath="pcoords.png")
```

文件扩展名决定其不同的渲染方式。除了.png，.pdf 扩展名也是常用的一种。

Visualizer 还可封装 Scikit-Learn 模型然后对其进行评估，超参数调节和算法选择。比如，可以用 heatmap 方式对分类结果进行可视化，用来显示其精确度，召回率，F1 值，并且对分类器中的所以类都支持。将 estimator 封装在 visualizer 的方法如下：

```
from yellowbrick.classifier import ClassificationReport
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
visualizer = ClassificationReport(model)

visualizer.fit(X_train, y_train)
visualizer.score(X_test, y_test)
visualizer.poof()
```

只需要在分类模型产生之后加入两行代码就可将模型评估可视化。首先实例化一个名为 ClassificationReport 的 visualizer，并将分类 estimator 封装进去，然后调用其 poof() 方法。这样，Visualizers 既 * 增强 * 了机器学习的工作流程又不对其进行干扰。

基于分类的 API 是注定要直接和 Scikit-Learn 进行整合的。然而有时候你只需要一个快速视图的时候怎么办呢。Yellowbrick 有一些功能可以支持快速视图。比如这两个诊断视图可以用如下方法实现：

```
from sklearn.linear_model import LogisticRegression

from yellowbrick.features import parallel_coordinates
from yellowbrick.classifier import classification_report

# Displays parallel coordinates
g = parallel_coordinates(X, y)

# Displays classification report
g = classification_report(LogisticRegression(), X, y)
```

这些快速视图的方法可能会有些减弱你对整个机器学习工作流程的控制，但是可以帮你很快得根据你的要求对模型进行诊断而且在数据探索过程中非常有效。

4.1.3 逐步解说

这里用一个回归分析作为简单的例子来展示怎样在机器学习流程中使用 visualizers。用上传到 UCI 机器学习数据库 的 共享单车数据集 ，我们可以用季节、天气、或者是否假日等信息对某个小时内被租自行车的数量进行预测。

在你下载并且将数据集解压缩到你目前的工作目录之后，我们可以将数据用如下方法载入：

```
import pandas as pd

data = pd.read_csv('bikeshare.csv')
```

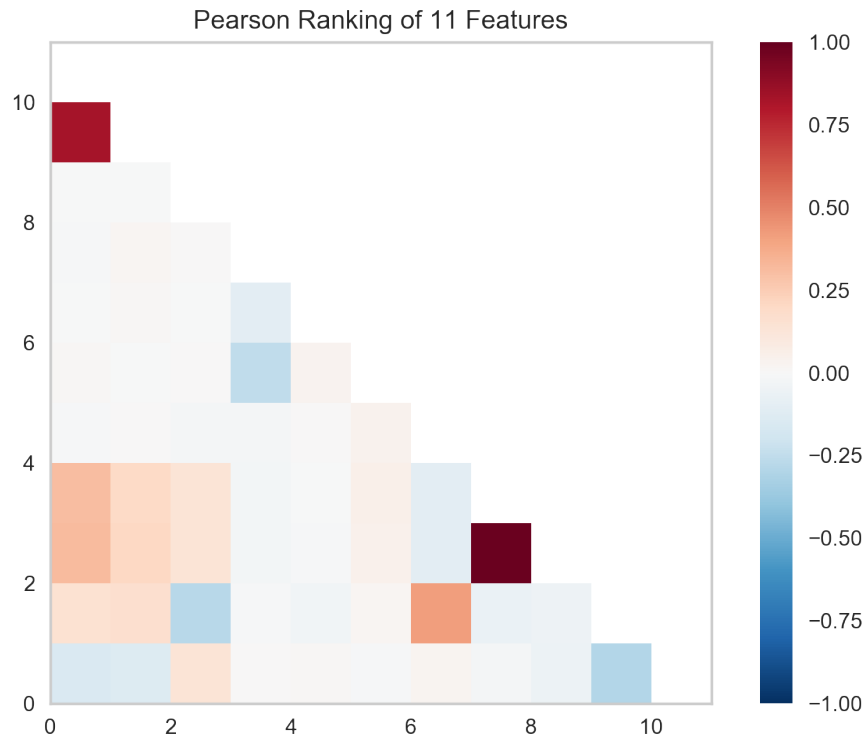
(下页继续)

(续上页)

```
X = data[[
    "season", "month", "hour", "holiday", "weekday", "workingday",
    "weather", "temp", "feelslike", "humidity", "windspeed"
]]
y = data["riders"]
```

机器学习的流程是创作 * 模型选择三重奏 * 的艺术，将特征、算法和超参数柔和在一起独特地组成一个模型并将其运用到特定的数据集上。作为特征选择的一部分，我们需要将和其他有线性关系的一部分特征识别出来。因为这部分特征有可能将协方差引入到模型中并且破坏 OLS（将我们引入到移除特征或者使用正则化的道路上）。**我们可以用 Rank2D_** visualizer 将所有特征两两之间的 Pearson 相关系数计算出来，具体操作如下：

```
from yellowbrick.features import Rank2D

visualizer = Rank2D(algorithm="pearson")
visualizer.fit_transform(X)
visualizer.poof()
```

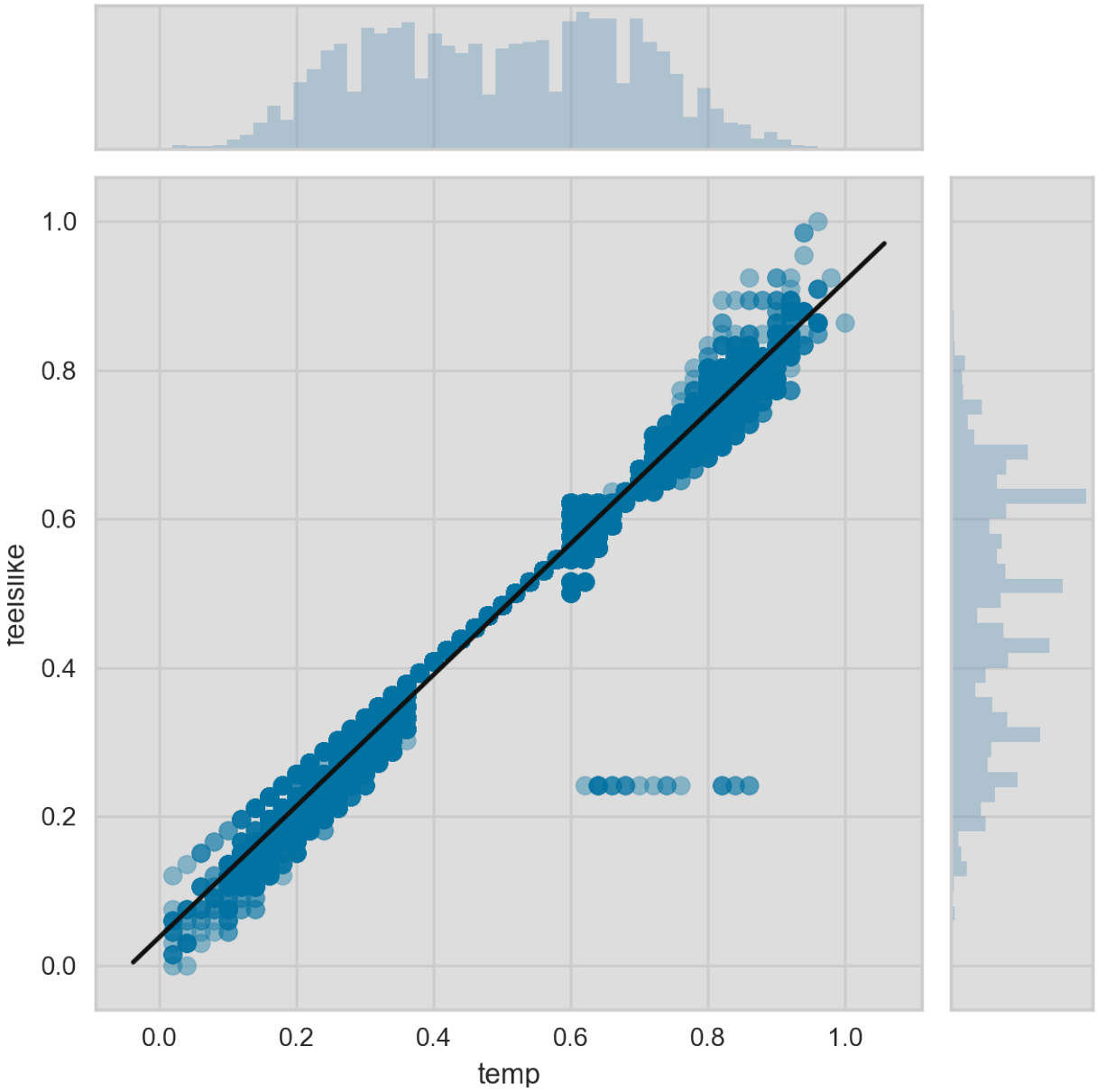


上图表示的是两两特征之间 Pearson 相关系数，其中坐标中的每一个小格代表 x 和 y 轴上两个相交特征的相关系数，其颜色的深浅和相关系数的值大小相关。当 Pearson 系数为 1.0 时，表示两个特征之间有强烈的

正的线性相关，而-1.0 则表示两个特征之间有强烈的负的线性相关（而 0 则表示没有任何相关）。因此我们需要找到深红色或者深蓝色的方块对其进行深入研究。

在这个图上我们可以看到特征 7（temperature）和特征 9（feelslike）有很强的相关性，并且特征 0（season）和特征 1（month）也有很强的相关性。这好像有些道理；我们感觉的温度依赖于实际温度以及其他空气质量相关的因子，并且每年中的季节是通过月份来描述的。为了对其进行更深入的分析，我们还可以用 JointPlotVisualizer 来考察这些相关性。

```
from yellowbrick.features import JointPlotVisualizer

visualizer = JointPlotVisualizer(feature='temp', target='feelslike')
visualizer.fit(X['temp'], X['feelslike'])
visualizer.poof()
```

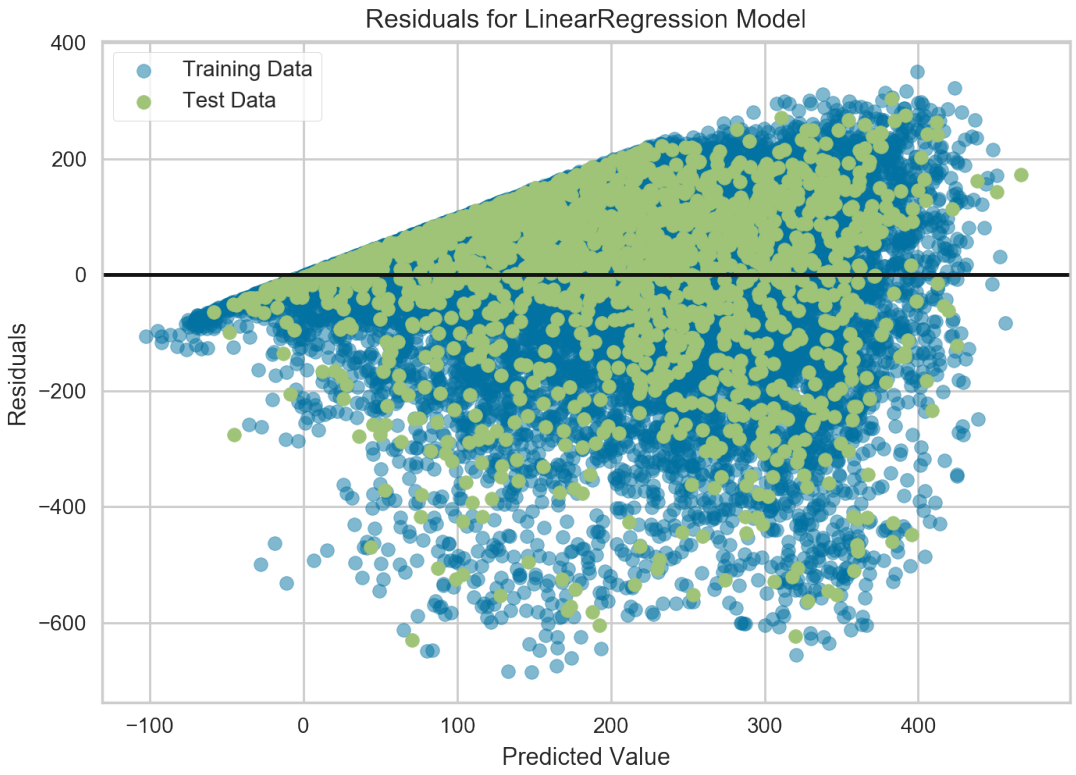


上面用 visualizer 做了一个散点图，其中 y 轴是感觉的温度，x 轴是实际温度，然后再将一个用简单线性回归训练的最佳模型的回归线添加上去。另外，还可将各个变量的分布情况用直方图的形式分别在 x 轴（temp）上方和 y 轴（feelslike）右侧显示。`JointPlotVisualizer` 让我们能快速浏览有强相关性的特征，以及各个特征的范围和分布情况。需要注意的是图中的各个轴都已经标准化到 0 到 1 之间的值，这是机器学习中一中非常常用的减少一个特征对另一个影响的技术。

这个图非常有趣；首先在 feelslike 等于 0.25 处好像有一些异常值。为了增加最终模型的质量，也许需要我们对其进行人工移除，因为这些也许是数据输入造成错误。其次，我们可以看到更多的极端温度可以对感知温度造成夸大的效应；温度越低，人们就越容易感觉越冷，温度越高，人们就感觉天气越暖和。适中的温度则让人感觉起来和实际温度不相上下。这给我们一个直觉好像 feelslike 是一个比 temp 更好的特征，并且如果其对我们的回归分析造成问题的话，我们应该移除 temp 变量而保留 feelslike。

到这，我们就可以训练我们模型了；我们来训练一个线性回归模型，并且绘制其残差。

```
from yellowbrick.regressor import ResidualsPlot
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

# Create training and test sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.1
)

visualizer = ResidualsPlot(LinearRegression())
visualizer.fit(X_train, y_train)
visualizer.score(X_test, y_test)
visualizer.poof()
```



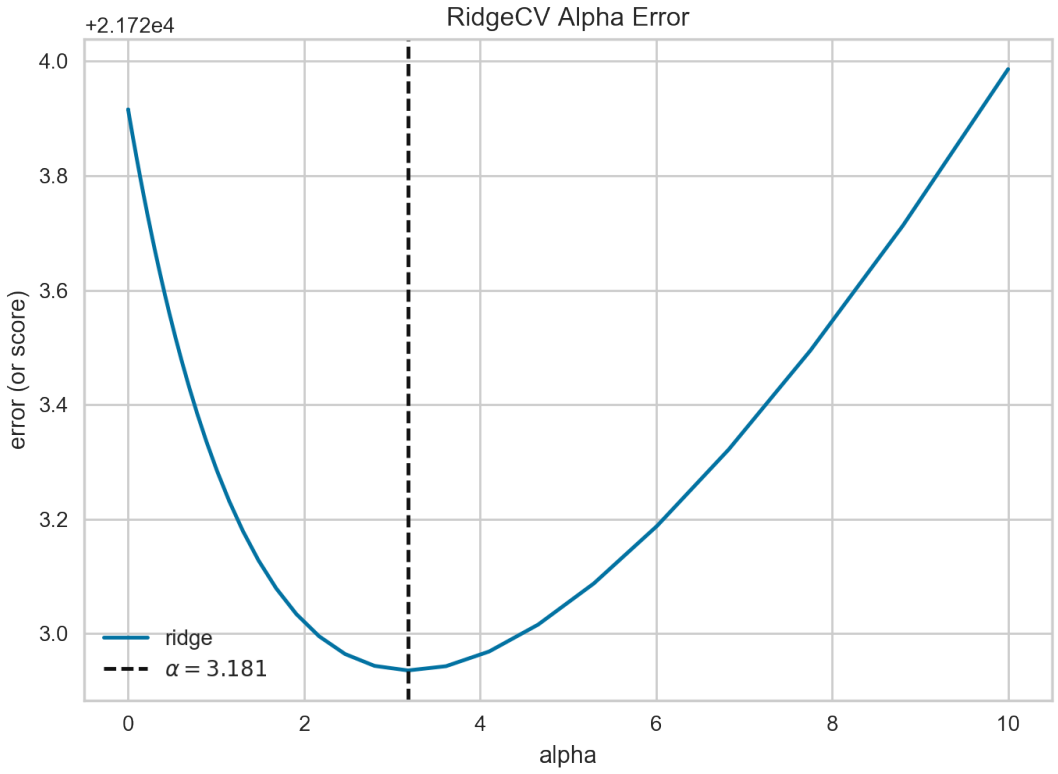
残差图不但可以显示与预测数据对应的误差，并且可以让我们观察到模型中的异方差性；比如，方差最大的区域。残差的形状可以让我们很好得意识模型中的组成部分对 OLS（普通最小二乘法）的影响最大区域。在这种情况下，我们可以看到预测的值越小（骑车人数越小），误差就越小，而预测的骑车人数越大，误差就越大。这就意味着我们的模型在某些目标区域有更多的噪音或者那两个变量是共线性的，也就是说在他们关系中的噪音发生变化时就产生了误差。

残差图还向我们展示了模型的误差是怎么产生的：那根加粗的水平线表示的是 `residuals = 0`，也就是没有误差；线上方或者下方的点则表示误差值的大小。比如大部分残差是负值，并且其值是由 `actual - expected` 算得，也就是说大部分时间预测值比实际值要大，比如和实际相比我们的模型总是预测有更多的骑手。还有，在残差图的右上角还有一个非常有趣的分界线，显示模型空间中有一种非常有趣的效应；也许在这个模型中有一些特征权重比较大。

最后，残差图中的训练样本和测试样本还可以用不同的颜色标注。这可以帮助我们更好的发现在训练和测试样本生成时产生的误差。如果测试样本误差和训练样本误差不相符的话，那我们的样本不是过拟合就是欠拟合。否则就是产生两个样本前随机整理数据集时产生了误差。

因为这个模型的决定系数为 0.328，让我们看看能不能用 * 正则化 * 训练一个更好的模型，并同时探索另一个 visualizer。

```
import numpy as np

from sklearn.linear_model import RidgeCV
from yellowbrick.regressor import AlphaSelection

alphas = np.logspace(-10, 1, 200)
visualizer = AlphaSelection(RidgeCV(alphas=alphas))
visualizer.fit(X, y)
visualizer.poof()
```



在探索模型家族的过程中，第一个要考虑的是模型是怎样变得更 * 复杂 * 的。当模型的复杂度增加，由于方差增加形成的误差也相应增加，因为模型会变得过拟合并且不能泛化到未知数据上。然而，模型越简单由于偏差造成的误差就会越大；模型欠拟合，因此有更多的未中靶预测。大部分机器学习的目的就是要产生一个 * 复杂度适中 * 的模型，在偏差和方差之间找到一个中间点。

对一个线性模型来说，复杂度来自于特征本身以及根据模型赋予它们的值。因此对线性模型期望用 * 最少的特征 * 达到最好的阐释结果。*正则化 * 是实现如上目标的其中一种技术，即引入一个 alpha 参数来对其相互之间系数的权重进行标准化并且惩罚其复杂度。Alpha 和复杂度之间是一个负相关。alpha 值越大，复杂度越小，反之亦然。

因此现在的问题就变成怎样选取 alpha 值了。其中的一项技术是用交叉验证的方法训练一系列模型，然后选择使误差值最小的 alpha。`AlphaSelection` 就是实现以上技术的一个 visualizer，其以图表形式呈现正则化的效果。正如上图所示，误差随着 alpha 值的增加而减小直到我们需要的值（目前情况下为 3.181），然后误差开始增加。这让我们可以实现偏差/方差平衡的目标，并且可以对不同的正则化方法之间的关系进行探索（比如 Ridge 对阵 Lasso）。

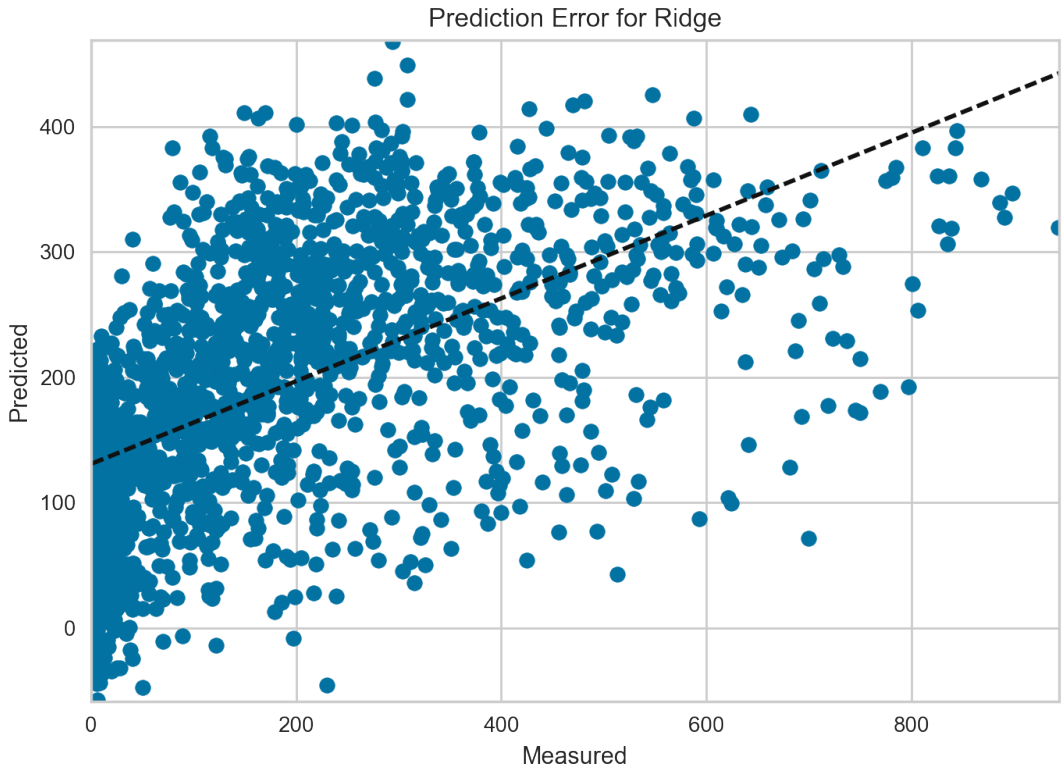
我们现在可以训练我们最终的模型并且用 `PredictionError` 对其进行可视化了：

```
from sklearn.linear_model import Ridge
from yellowbrick.regressor import PredictionError

visualizer = PredictionError(Ridge(alpha=3.181))
```

（下页继续）

(续上页)

```
visualizer.fit(X_train, y_train)
visualizer.score(X_test, y_test)
visualizer.poof()
```



用预测误差 visualizer 将实际（测量）值对期望（预测）值进行可视化。黑色的 45 度虚线表示误差为 0 的点。和残差图一样，这让我们可以看到误差在何处出现，值为多大。

在这个图上，我们可以看到大部分的点集中在小于 200 骑手的位置。我们也许想要尝试用正交匹配追踪算法（OMP）或者样条（spline）来训练一个将更多区域性考虑进来的回归模型。我们还可以看到残差图中奇怪的拓扑结构好像已被 Ridge 回归纠正，而且在我们的模型中大值和小值之间有了更多的平衡。Ridge 正则化可能纠正了两个特征之间的协方差问题。当我们用其他模型的形式将我们的数据分析推进的同时，我们可以继续 visualizers 来快速比较并且可视化我们的结果。

希望这个流程让你对怎样将 Visualizers 通过 Scikit-Learn 整合到机器学习中去有一个概念，并且给你启发让你将其运用到你的工作中！如果想要了解更多的有关怎样开始使用 Yellowbrick 的信息，请查看模型选择教程 。然后你就在*Visualizers and API* 上快速查看更多的特定 visualizers 了。

翻译：Juan L. Kehoe

4.2 模型选择教程

在本教程中，我们将查看各种 Scikit-Learn 模型的分数，并使用 Yellowbrick 的可视化诊断工具对其进行比较，以便为我们的数据选择最佳模型。

4.2.1 模型选择三元组

关于机器学习的讨论常常集中在模型选择上。无论是逻辑回归、随机森林、贝叶斯方法，还是人工神经网络，机器学习实践者通常都能很快地展示他们的偏好。这主要是因为历史原因。尽管现代的第三方机器学习库使得各类模型的部署显得微不足道，但传统上，即使是其中一种算法的应用和调优也需要多年的研究。因此，与其他模型相比，机器学习实践者往往对特定的 (并且更可能是熟悉的) 模型有强烈的偏好。

然而，模型选择比简单地选择“正确”或“错误”算法更加微妙。实践中的工作流程包括:

1. 选择和/或设计最小和最具预测性的特性集
2. 从模型家族中选择一组算法，并且
3. 优化算法超参数以优化性能。

模型选择三元组是由 Kumar 等人，在 2015 年的 SIGMOD 论文中首次提出。在他们的论文中，谈论到下一代为预测建模而构建的数据库系统的开发。作者很中肯地表示，由于机器学习在实践中具有高度实验性，因此迫切需要这样的系统。“模型选择，”他们解释道，“是迭代的和探索性的，因为 (模型选择三元组) 的空间通常是无限的，而且通常不可能让分析师事先知道哪个 (组合) 将产生令人满意的准确性和/或洞察力。”

最近，许多工作流程已经通过网格搜索方法、标准化 API 和基于 GUI 的应用程序实现了自动化。然而，在实践中，人类的直觉和指导可以比穷举搜索更有效地专注于模型质量。通过可视化模型选择过程，数据科学家可以转向最终的、可解释的模型，并避免陷阱和陷阱。

Yellowbrick 库是一个针对机器学习的可视化诊断平台，它允许数据科学家控制模型选择过程。Yellowbrick 用一个新的核心对象扩展了 Scikit-Learn 的 API: Visualizer。Visualizers 允许可视化模型作为 Scikit-Learn 管道过程的一部分进行匹配和转换，从而在高维数据的转换过程中提供可视化诊断。

4.2.2 关于数据

本教程使用来自 UCI Machine Learning Repository 的修改过的蘑菇数据集版本。我们的目标是基于蘑菇的特定，去预测蘑菇是有毒的还是可食用的。

这些数据包括与伞菌目 (Agaricus) 和环柄菇属 (Lepiota) 科中 23 种烤蘑菇对应的假设样本描述。每一种都被确定为绝对可食用，绝对有毒，或未知的可食用性和不推荐（后一类与有毒物种相结合)。

我们的文件“agaricus-lepiota.txt”，包含 3 个名义上有价值的属性信息和 8124 个蘑菇实例的目标值 (4208 个可食用，3916 个有毒)。

让我们用 Pandas 加载数据。

```
import os
import pandas as pd

names = [
    'class',
    'cap-shape',
    'cap-surface',
    'cap-color'
]

mushrooms = os.path.join('data','agaricus-lepiota.txt')
dataset   = pd.read_csv(mushrooms)
dataset.columns = names
dataset.head()
```

.	class	cap-shape	cap-surface	cap-color
0	edible	bell	smooth	white
1	poisonous	convex	scaly	white
2	edible	convex	smooth	gray
3	edible	convex	scaly	yellow
4	edible	bell	smooth	white

```
features = ['cap-shape', 'cap-surface', 'cap-color']
target   = ['class']

X = dataset[features]
y = dataset[target]
```

4.2.3 特征提取

我们的数据，包括目标参数，都是分类型数据。为了使用机器学习，我们需要将这些值转化为数值型数据。为了从数据集中提取这一点，我们必须使用 Scikit-Learn 的转换器（transformers）将输入数据集转换为适合模型的数据集。幸运的是，Sckit-Learn 提供了一个转换器，用于将分类标签转换为整数: sklearn.preprocessing.LabelEncoder。不幸的是，它一次只能转换一个向量，所以我们必须对它进行调整，以便将它应用于多个列。

```
from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.preprocessing import LabelEncoder, OneHotEncoder


```

(下页继续)

(续上页)

```
class EncodeCategorical(BaseEstimator, TransformerMixin):
    """
    Encodes a specified list of columns or all columns if None.
    """

    def __init__(self, columns=None):
        self.columns  = [col for col in columns]
        self.encoders = None

    def fit(self, data, target=None):
        """
        Expects a data frame with named columns to encode.
        """
        # Encode all columns if columns is None
        if self.columns is None:
            self.columns = data.columns

        # Fit a label encoder for each column in the data frame
        self.encoders = {
            column: LabelEncoder().fit(data[column])
            for column in self.columns
        }
        return self

    def transform(self, data):
        """
        Uses the encoders to transform a data frame.
        """
        output = data.copy()
        for column, encoder in self.encoders.items():
            output[column] = encoder.transform(data[column])

        return output
```

4.2.4 建模与评估

评估分类器的常用指标

精确度 (Precision) 是正确的阳性结果的数量除以所有阳性结果的数量 (例如，我们预测的可食用蘑菇实际上有多少?)

召回率 (Recall) 是正确的阳性结果的数量除以应该返回的阳性结果的数量 (例如，我们准确预测了多少有毒蘑菇是有毒的?)

F1 分数 (F1 score) 是测试准确度的一种衡量标准。它同时考虑测试的精确度和召回率来计算分数。F1 得分可以解释为精度和召回率的加权平均值，其中 F1 得分在 1 处达到最佳值，在 0 处达到最差值。

```
precision = true positives / (true positives + false positives)

recall = true positives / (false negatives + true positives)

F1 score = 2 * ((precision * recall) / (precision + recall))
```

现在我们准备好作出一些预测了!

让我们构建一种评估多个估算器 (multiple estimators) 的方法——首先使用传统的数值分数（我们稍后将与 Yellowbrick 库中的一些可视化诊断进行比较）。

```
from sklearn.metrics import f1_score
from sklearn.pipeline import Pipeline


def model_selection(X, y, estimator):
    """
    Test various estimators.
    """
    y = LabelEncoder().fit_transform(y.values.ravel())
    model = Pipeline([
         ('label_encoding', EncodeCategorical(X.keys())),
         ('one_hot_encoder', OneHotEncoder()),
         ('estimator', estimator)
    ])

    # Instantiate the classification model and visualizer
    model.fit(X, y)

    expected  = y
    predicted = model.predict(X)
```

(下页继续)

(续上页)

```
    # Compute and return the F1 score (the harmonic mean of precision and recall)
    return (f1_score(expected, predicted))
```

```
# Try them all!
from sklearn.svm import LinearSVC, NuSVC, SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegressionCV, LogisticRegression, SGDClassifier
from sklearn.ensemble import BaggingClassifier, ExtraTreesClassifier,␣
↪RandomForestClassifier
```

```
model_selection(X, y, LinearSVC())
```

```
0.65846308387744845
```

```
model_selection(X, y, NuSVC())
```

```
0.63838842388991346
```

```
model_selection(X, y, SVC())
```

```
0.66251459711950167
```

```
model_selection(X, y, SGDClassifier())
```

```
0.69944182052382997
```

```
model_selection(X, y, KNeighborsClassifier())
```

```
0.65802139037433149
```

```
model_selection(X, y, LogisticRegressionCV())
```

```
0.65846308387744845
```

```
model_selection(X, y, LogisticRegression())
```

```
0.65812609897010799
```

```
model_selection(X, y, BaggingClassifier())
```

```
0.687643484132343
```

```
model_selection(X, y, ExtraTreesClassifier())
```

```
0.68713648045448383
```

```
model_selection(X, y, RandomForestClassifier())
```

```
0.69317131158367451
```

初步模型评估

根据上面 F1 分数的结果，哪个模型表现最好?

4.2.5 可视化模型评估

现在，让我们重构模型评估函数，使用 Yellowbrick 的 ClassificationReport 类，这是一个模型可视化工具，可以显示精确度、召回率和 F1 分数。这个可视化的模型分析工具集成了数值分数以及彩色编码的热力图，以支持简单的解释和检测，特别是对于我们用例而言非常相关 (性命攸关!) 的第一类错误 (Type I error) 和第二类错误 (Type II error) 的细微差别。

第一类错误 (或 **" 假阳性 (false positive)"**) 是检测一种不存在的效应 (例如，当蘑菇实际上是可以食用的时候，它是有毒的)。

第二类错误 (或 **"假阴性" "false negative"**) 是未能检测到存在的效应 (例如，当蘑菇实际上有毒时，却认为它是可以食用的)。

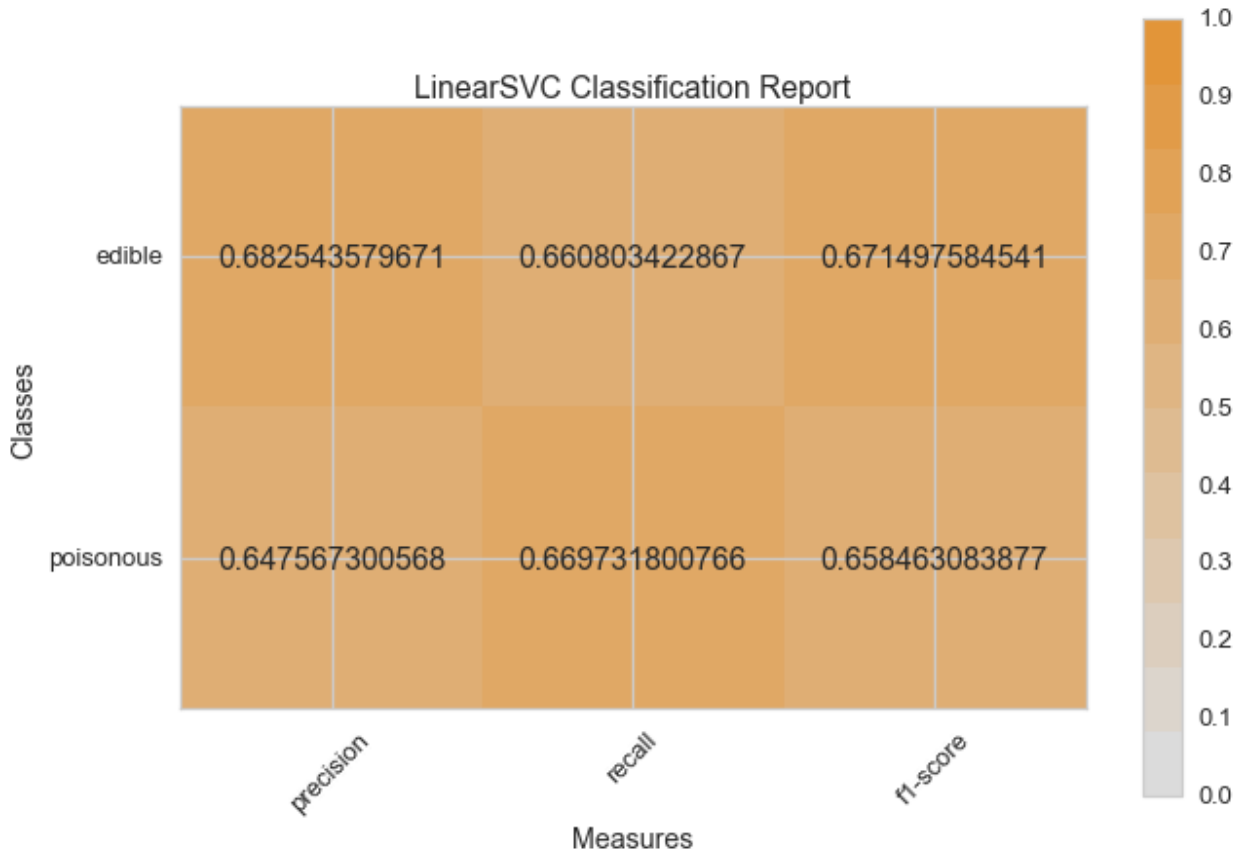
```
from sklearn.pipeline import Pipeline
from yellowbrick.classifier import ClassificationReport


def visual_model_selection(X, y, estimator):
    """
    Test various estimators.
    """
    y = LabelEncoder().fit_transform(y.values.ravel())
```

(下页继续)

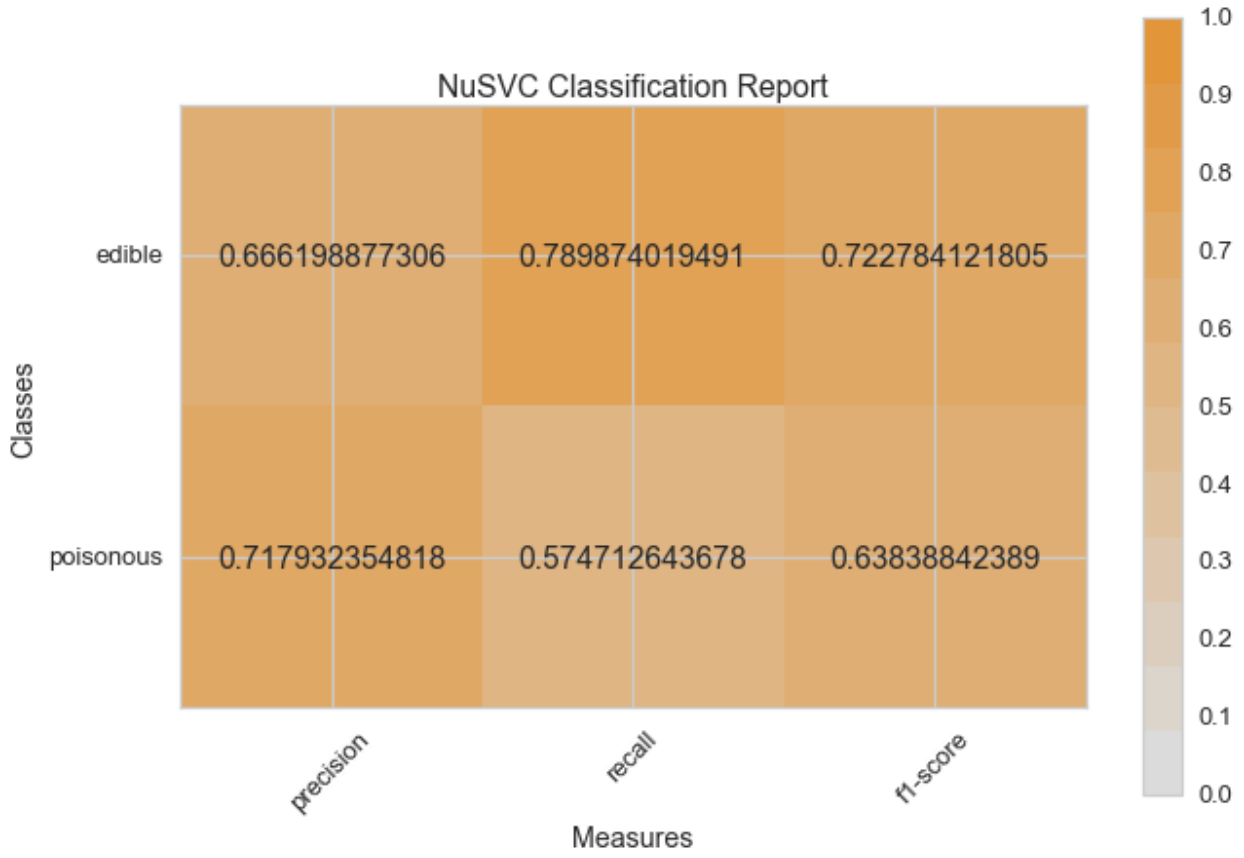
(续上页)

```
    model = Pipeline([
        ('label_encoding', EncodeCategorical(X.keys())),
        ('one_hot_encoder', OneHotEncoder()),
        ('estimator', estimator)
    ])

    # Instantiate the classification model and visualizer
    visualizer = ClassificationReport(model, classes=['edible', 'poisonous'])
    visualizer.fit(X, y)
    visualizer.score(X, y)
    visualizer.poof()
```

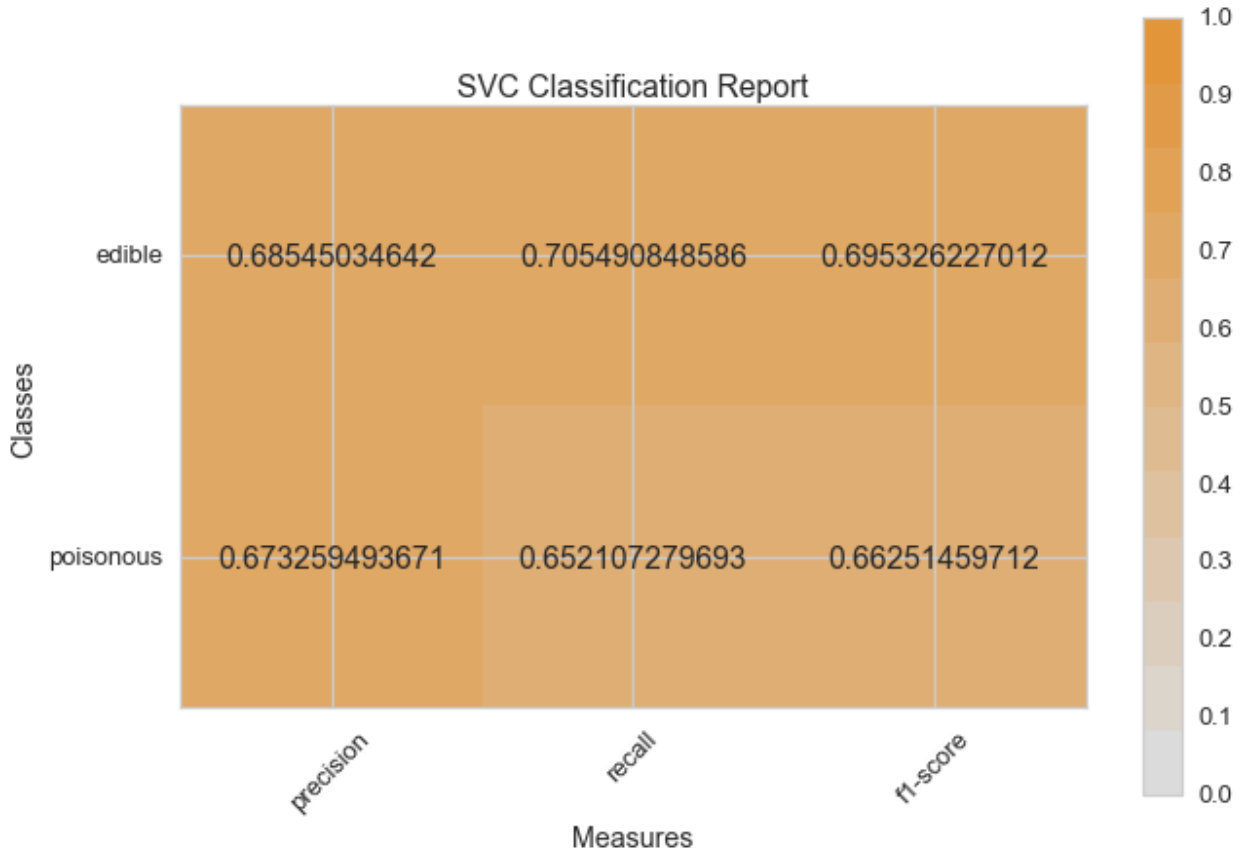
```
visual_model_selection(X, y, LinearSVC())
```



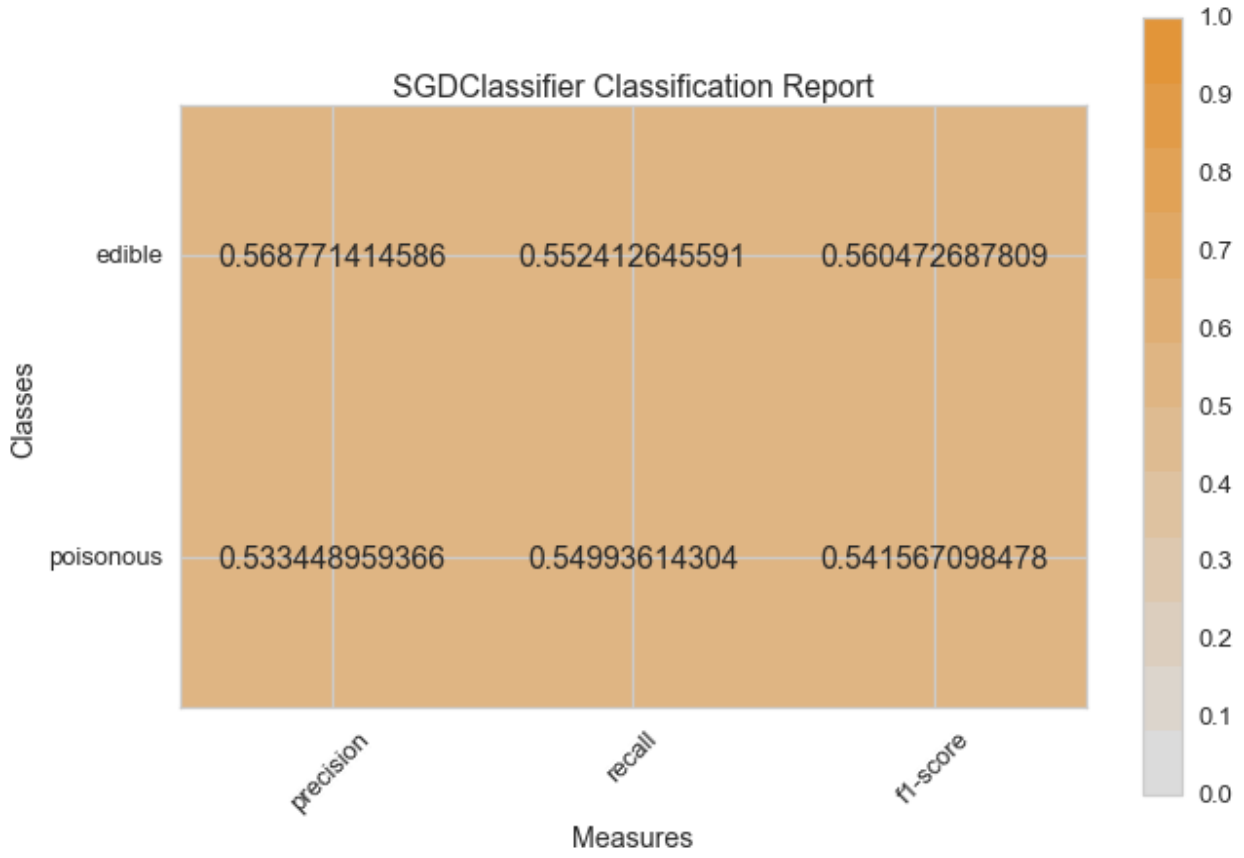

```
visual_model_selection(X, y, NuSVC())
```



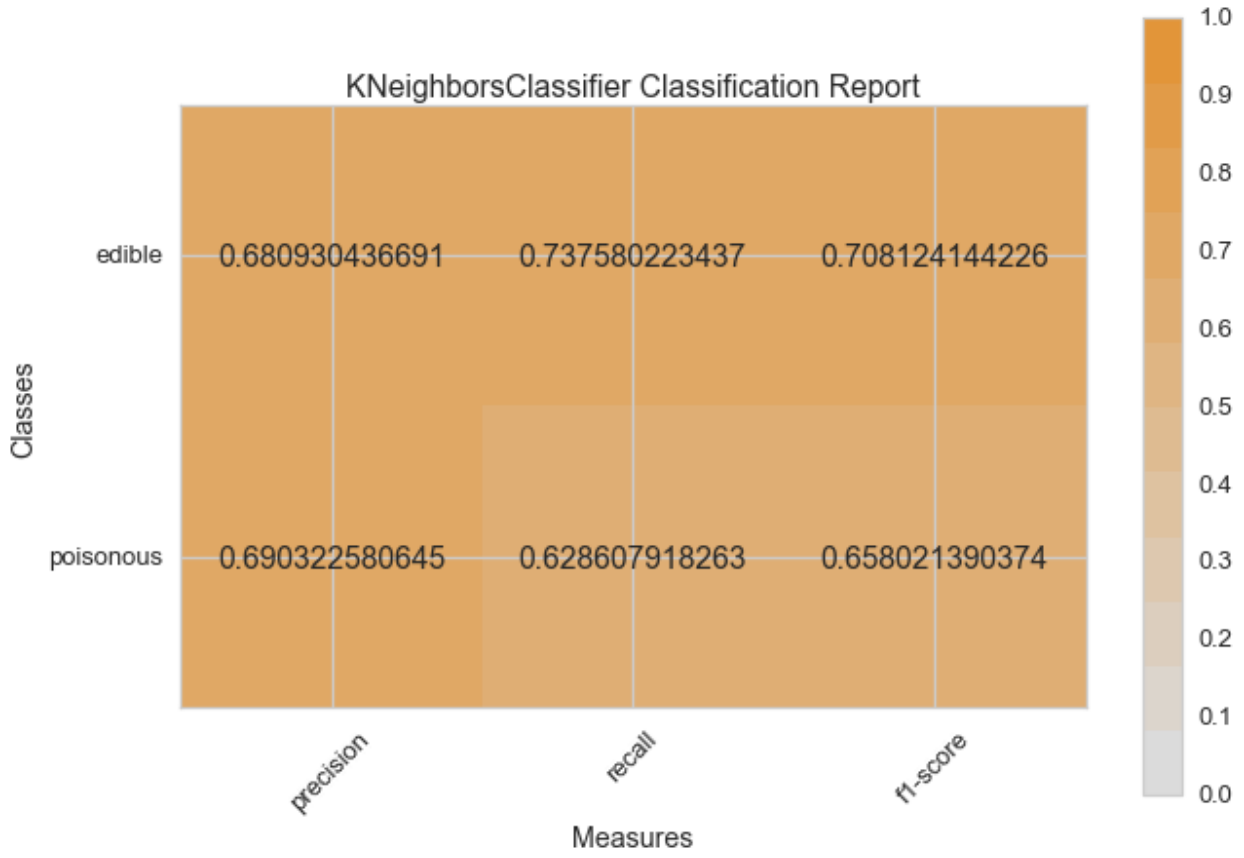

```
visual_model_selection(X, y, SVC())
```



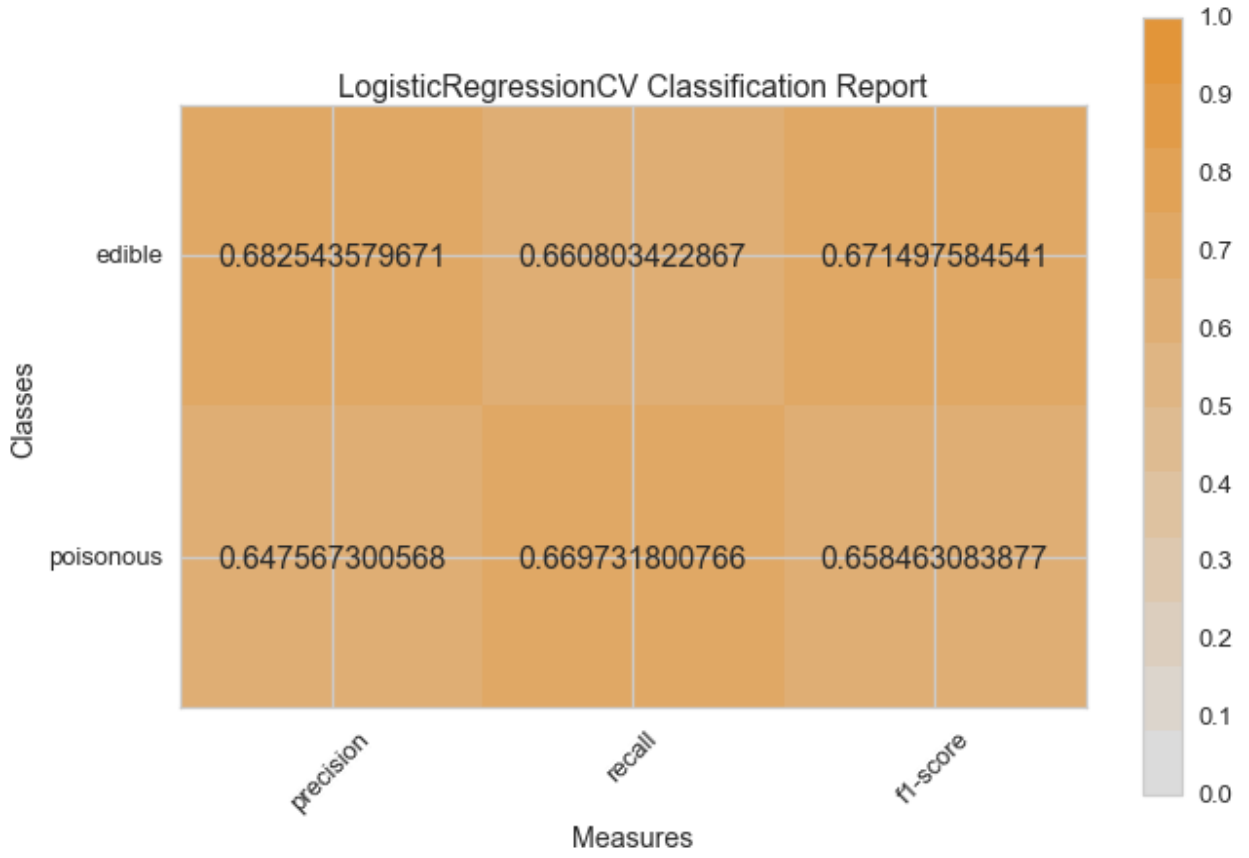

```
visual_model_selection(X, y, SGDClassifier())
```



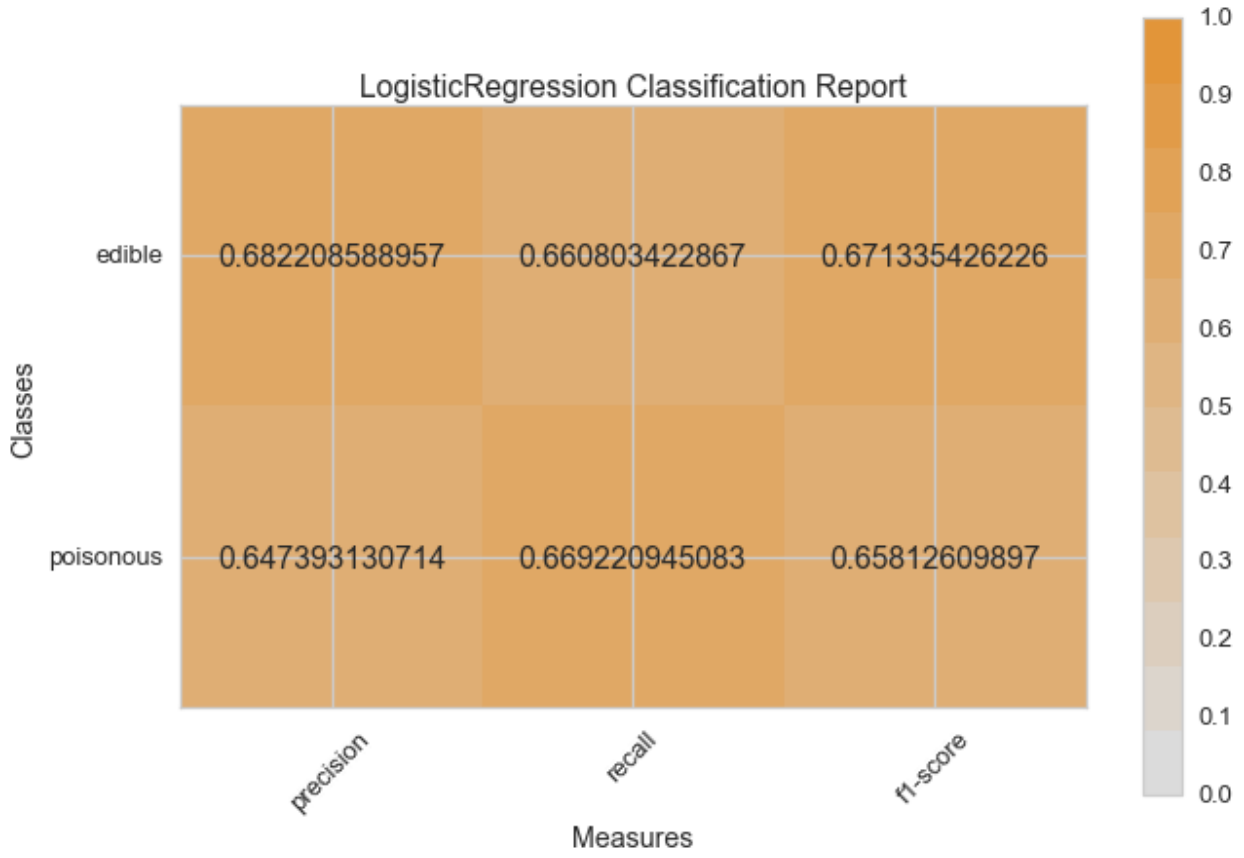

```
visual_model_selection(X, y, KNeighborsClassifier())
```



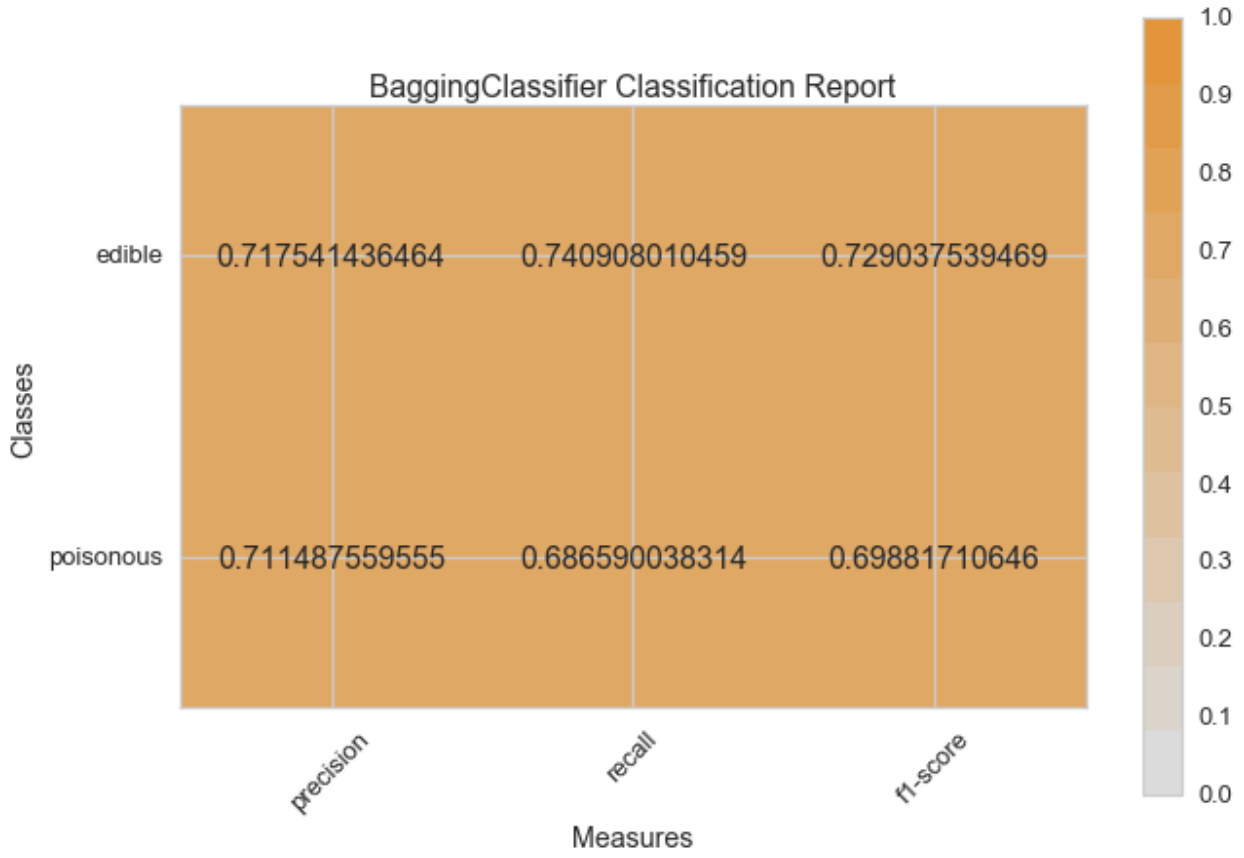

```
visual_model_selection(X, y, LogisticRegressionCV())
```



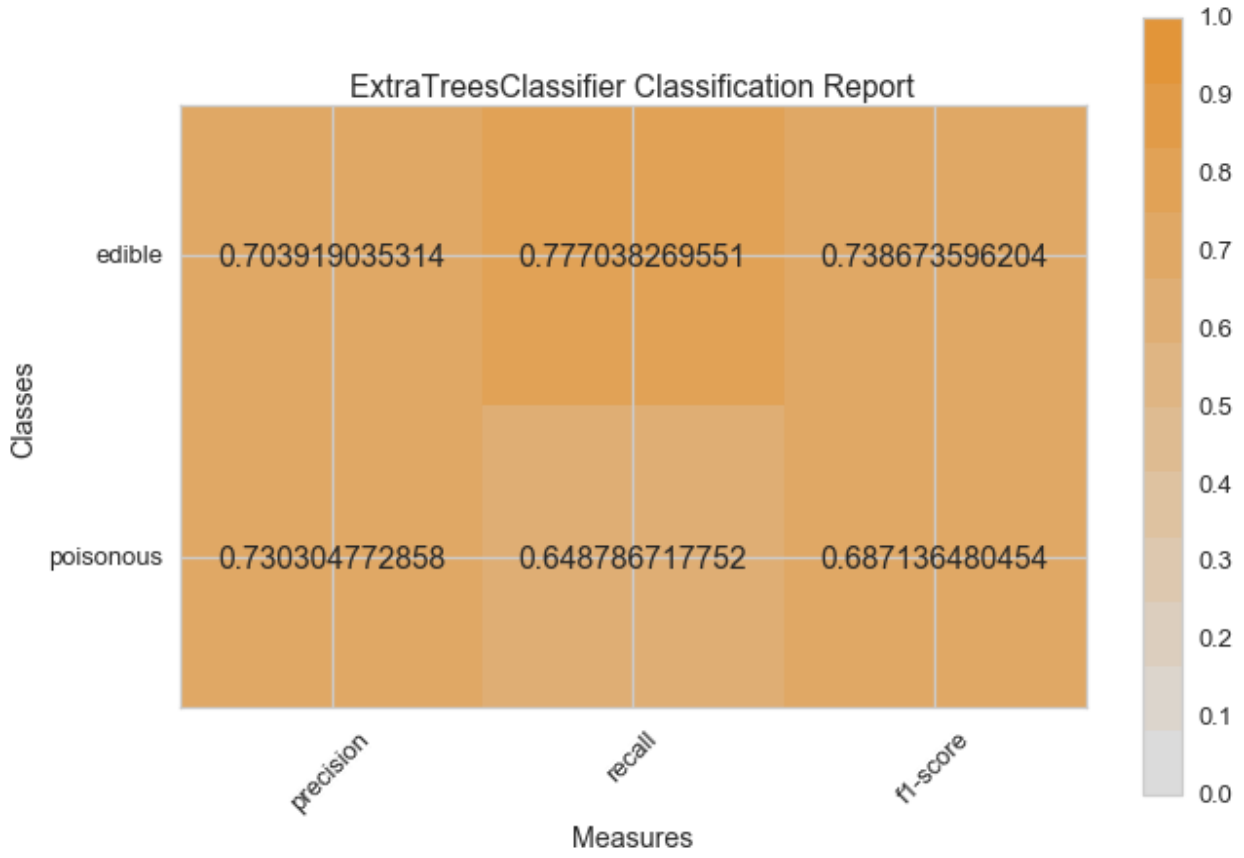

```
visual_model_selection(X, y, LogisticRegression())
```



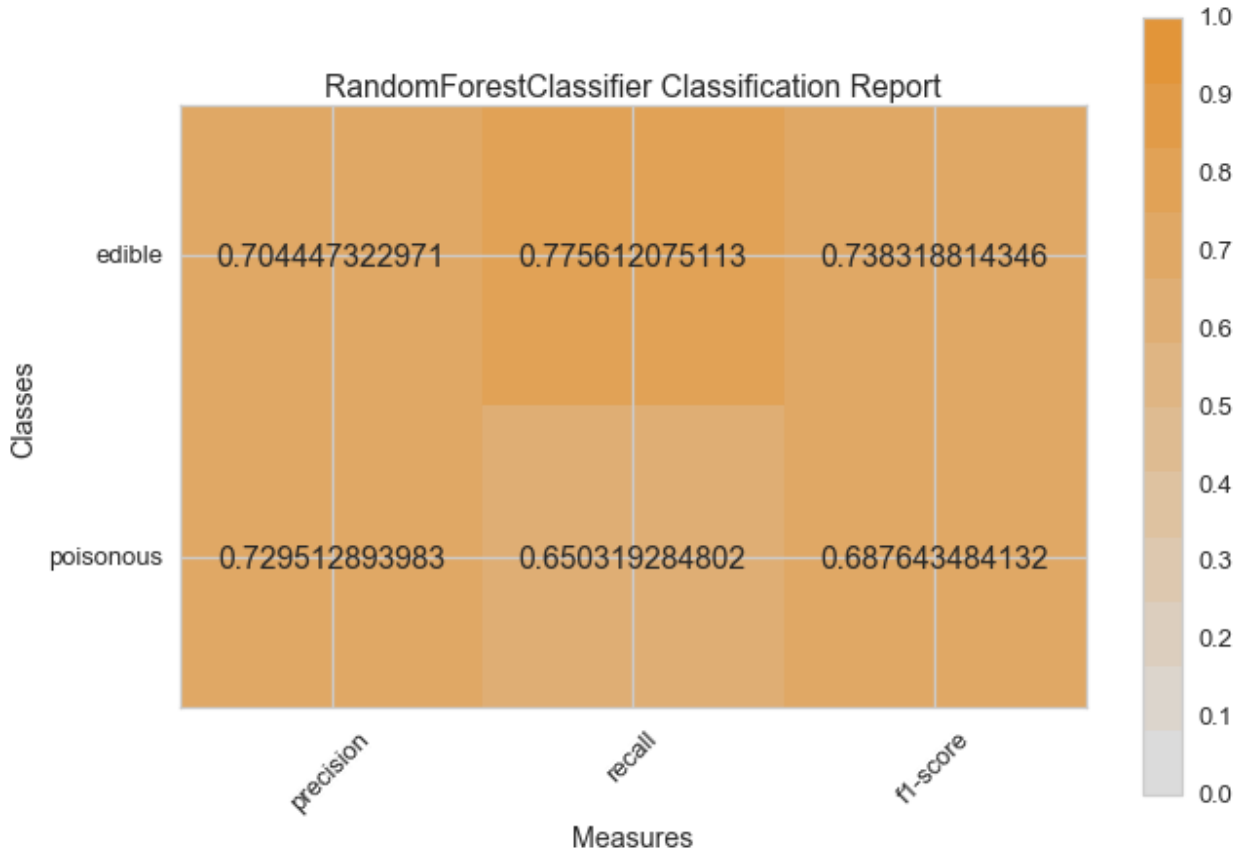

```
visual_model_selection(X, y, BaggingClassifier())
```




```
visual_model_selection(X, y, ExtraTreesClassifier())
```




```
visual_model_selection(X, y, RandomForestClassifier())
```



4.2.6 检验

1. 现在, 哪种模型看起来最好? 为什么?
2. 哪一个模型最有可能救你的命?
3. 可视化模型评估与数值模型评价，体验起来有何不同?

4.3 Visualizers and API

Welcome the API documentation for Yellowbrick! This section contains a complete listing of all currently available, production-ready visualizers along with code examples of how to use them. Use the links below to navigate to the reference for each visualization.

4.3.1 Example Datasets

Yellowbrick hosts several datasets wrangled from the UCI Machine Learning Repository to present the examples in this section. If you haven't downloaded the data, you can do so by running:

```
$ python -m yellowbrick.download
```

This should create a folder called `data` in your current working directory with all of the datasets. You can load a specified dataset with `pandas.read_csv` as follows:

```
import pandas as pd

data = pd.read_csv('data/concrete/concrete.csv')
```

The following code snippet can be found at the top of the `examples/examples.ipynb` notebok in Yellowbrick. Please reference this code when trying to load a specific data set:

```
from yellowbrick.download import download_all

## The path to the test data sets
FIXTURES  = os.path.join(os.getcwd(), "data")

## Dataset loading mechanisms
datasets = {
    "bikeshare": os.path.join(FIXTURES, "bikeshare", "bikeshare.csv"),
    "concrete": os.path.join(FIXTURES, "concrete", "concrete.csv"),
    "credit": os.path.join(FIXTURES, "credit", "credit.csv"),
    "energy": os.path.join(FIXTURES, "energy", "energy.csv"),
    "game": os.path.join(FIXTURES, "game", "game.csv"),
    "mushroom": os.path.join(FIXTURES, "mushroom", "mushroom.csv"),
    "occupancy": os.path.join(FIXTURES, "occupancy", "occupancy.csv"),
}


def load_data(name, download=True):
    """
    Loads and wrangles the passed in dataset by name.
    If download is specified, this method will download any missing files.
    """

    # Get the path from the datasets
    path = datasets[name]

    # Check if the data exists, otherwise download or raise
    if not os.path.exists(path):
        if download:
```

(下页继续)

(续上页)

```
            download_all()
        else:
            raise ValueError((
                "'{}' dataset has not been downloaded, "
                "use the download.py module to fetch datasets"
            ).format(name))


    # Return the data frame
    return pd.read_csv(path)
```

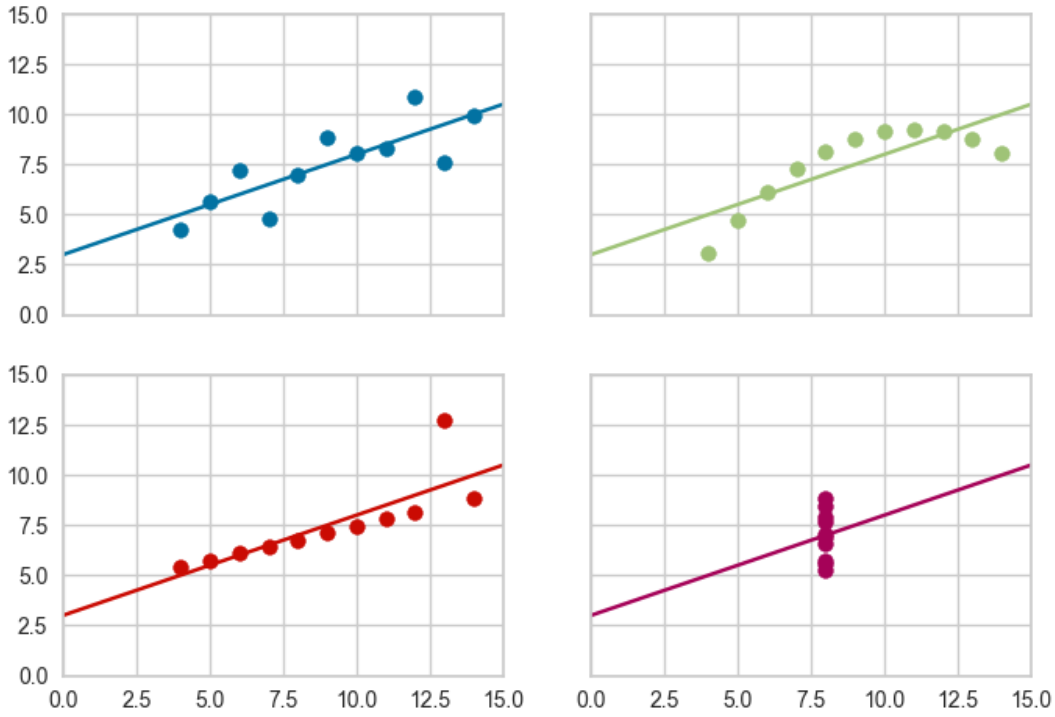
Note that most of the examples currently use one or more of the listed datasets for their examples (unless specifically shown otherwise). Each dataset has a `README.md` with detailed information about the data source, attributes, and target. Here is a complete listing of all datasets in Yellowbrick and their associated analytical tasks:

- **bikeshare**: suitable for regression
- **concrete**: suitable for regression
- **credit**: suitable for classification/clustering
- **energy**: suitable for regression
- **game**: suitable for classification
- **hobbies**: suitable for text analysis
- **mushroom**: suitable for classification/clustering
- **occupancy**: suitable for classification

4.3.2 Anscombe's Quartet

Yellowbrick has learned Anscombe's lesson - which is why we believe that visual diagnostics are vital to machine learning.

```
import yellowbrick as yb
import matplotlib.pyplot as plt

g = yb.anscombe()
plt.show()
```



API Reference

Plots Anscombe's Quartet as an illustration of the importance of visualization.

`yellowbrick.anscombe.anscombe()`

Creates 2x2 grid plot of the 4 anscombe datasets for illustration.

4.3.3 Feature Analysis Visualizers

Feature analysis visualizers are designed to visualize instances in data space in order to detect features or targets that might impact downstream fitting. Because ML operates on high-dimensional data sets (usually at least 35), the visualizers focus on aggregation, optimization, and other techniques to give overviews of the data. It is our intent that the steering process will allow the data scientist to zoom and filter and explore the relationships between their instances and between dimensions.

At the moment we have five feature analysis visualizers implemented:

- *Rank Features*: rank single and pairs of features to detect covariance
- *RadViz Visualizer*: plot data points along axes ordered around a circle to detect separability
- *Parallel Coordinates*: plot instances as lines along vertical axes to detect classes or clusters

- *PCA Projection*: project higher dimensions into a visual space using PCA
- *Feature Importances*: rank features by relative importance in a model
- *Direct Data Visualization*: plot instances by selecting subsets of features

Feature analysis visualizers implement the `Transformer` API from Scikit-Learn, meaning they can be used as intermediate transform steps in a `Pipeline` (particularly a `VisualPipeline`). They are instantiated in the same way, and then fit and transform are called on them, which draws the instances correctly. Finally `poof` or `show` is called which displays the image.

```
# Feature Analysis Imports
# NOTE that all these are available for import directly from the `yellowbrick.features`␣
↪module
from yellowbrick.features.rankd import Rank1D, Rank2D
from yellowbrick.features.radviz import RadViz
from yellowbrick.features.pcoords import ParallelCoordinates
from yellowbrick.features.jointplot import JointPlotVisualizer
from yellowbrick.features.pca import PCADecomposition
from yellowbrick.features.importances import FeatureImportances
from yellowbrick.features.scatter import ScatterVisualizer
```

RadViz Visualizer

RadViz is a multivariate data visualization algorithm that plots each feature dimension uniformly around the circumference of a circle then plots points on the interior of the circle such that the point normalizes its values on the axes from the center to each arc. This mechanism allows as many dimensions as will easily fit on a circle, greatly expanding the dimensionality of the visualization.

Data scientists use this method to detect separability between classes. E.g. is there an opportunity to learn from the feature set or is there just too much noise?

If your data contains rows with missing values (numpy.nan), those missing values will not be plotted. In other words, you may not get the entire picture of your data. RadViz will raise a DataWarning to inform you of the percent missing.

If you do receive this warning, you may want to look at imputation strategies. A good starting place is scikit-learn Imputer.

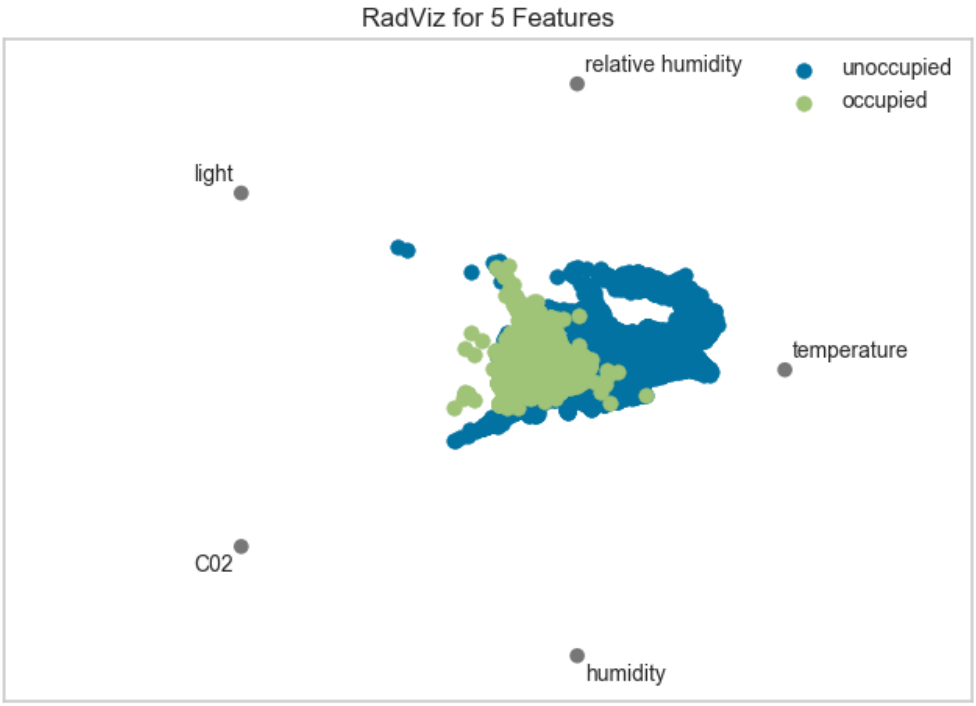
```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']
```

(下页继续)

(续上页)

```
# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.occupancy.as_matrix()
```

```
# Import the visualizer
from yellowbrick.features import RadViz

# Instantiate the visualizer
visualizer = RadViz(classes=classes, features=features)

visualizer.fit(X, y)       # Fit the data to the visualizer
visualizer.transform(X)    # Transform the data
visualizer.poof()          # Draw/show/poof the data
```



For regression, the RadViz visualizer should use a color sequence to display the target information, as opposed to discrete colors.

API Reference

Implements radviz for feature analysis.

class yellowbrick.features.radviz.RadialVisualizer(*ax=None*, *features=None*, *classes=None*, *color=None*, *colormap=None*, ***kwargs*)

基类：yellowbrick.features.base.DataVisualizer

RadViz is a multivariate data visualization algorithm that plots each axis uniformely around the circumference of a circle then plots points on the interior of the circle such that the point normalizes its values on the axes from the center to each arc.

Parameters

ax [matplotlib Axes, default: None] The axis to plot the figure on. If None is passed in the current axes will be used (or generated if required).

features [list, default: None] a list of feature names to use If a DataFrame is passed to fit and features is None, feature names are selected as the columns of the DataFrame.

classes [list, default: None] a list of class names for the legend If classes is None and a y value is passed to fit then the classes are selected from the target vector.

color [list or tuple, default: None] optional list or tuple of colors to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

colormap [string or cmap, default: None] optional string or matplotlib cmap to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

Examples

```
>>> visualizer = RadViz()
>>> visualizer.fit(X, y)
>>> visualizer.transform(X)
>>> visualizer.poof()
```

draw(*X*, *y*, ***kwargs*)

Called from the fit method, this method creates the radviz canvas and draws each instance as a class or target colored point, whose location is determined by the feature data set.

finalize(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

static normalize(*X*)

MinMax normalization to fit a matrix in the space [0,1] by column.

yellowbrick.features.radviz.RadViz

yellowbrick.features.radviz.RadialVisualizer 的别名

Rank Features

Rank1D and Rank2D evaluate single features or pairs of features using a variety of metrics that score the features on the scale [-1, 1] or [0, 1] allowing them to be ranked. A similar concept to SPLOMs, the scores are visualized on a lower-left triangle heatmap so that patterns between pairs of features can be easily discerned for downstream analysis.

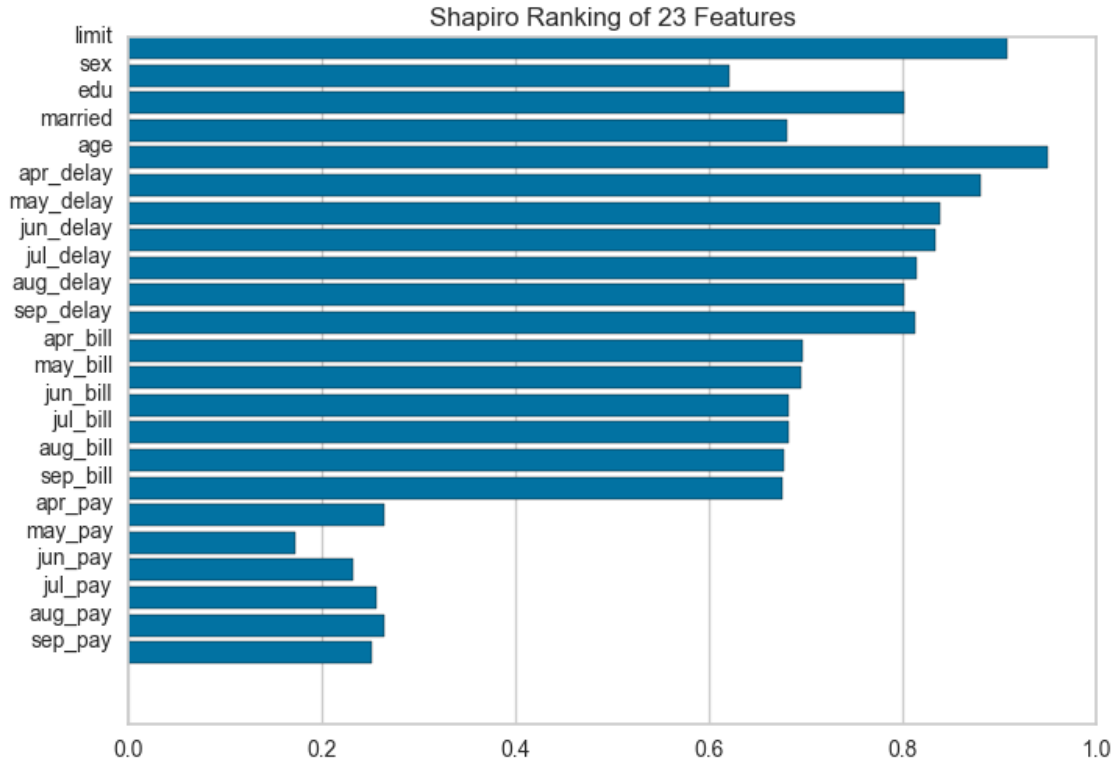
In this example, we'll use the credit default data set from the UCI Machine Learning repository to rank features. The code below creates our instance matrix and target vector.

```
# Load the dataset
data = load_data('credit')

# Specify the features of interest
features = [
        'limit', 'sex', 'edu', 'married', 'age', 'apr_delay', 'may_delay',
        'jun_delay', 'jul_delay', 'aug_delay', 'sep_delay', 'apr_bill', 'may_bill',
        'jun_bill', 'jul_bill', 'aug_bill', 'sep_bill', 'apr_pay', 'may_pay', 'jun_pay',
        'jul_pay', 'aug_pay', 'sep_pay',
    ]

# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.default.as_matrix()
```

Rank 1D

A one dimensional ranking of features utilizes a ranking algorithm that takes into account only a single feature at a time (e.g. histogram analysis). By default we utilize the Shapiro-Wilk algorithm to assess the normality of the distribution of instances with respect to the feature. A barplot is then drawn showing the relative ranks of each feature.

```
# Instantiate the 1D visualizer with the Sharpiro ranking algorithm
visualizer = Rank1D(features=features, algorithm='shapiro')

visualizer.fit(X, y)                # Fit the data to the visualizer
visualizer.transform(X)             # Transform the data
visualizer.poof()                   # Draw/show/poof the data
```



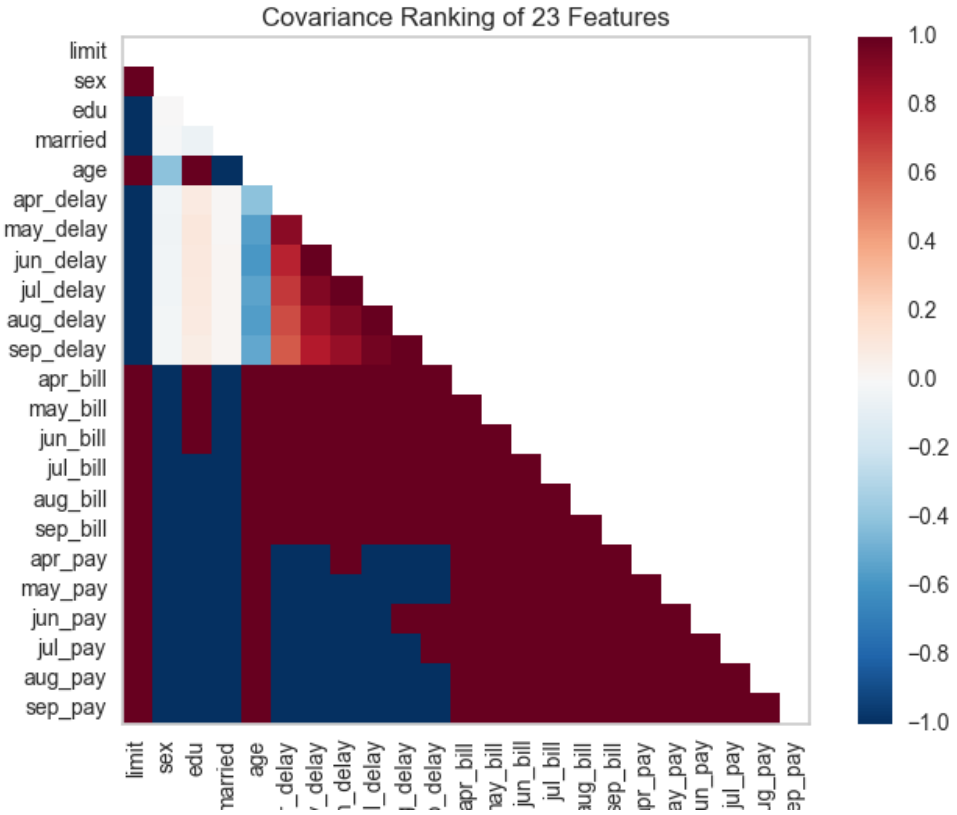
Rank 2D

A two dimensional ranking of features utilizes a ranking algorithm that takes into account pairs of features at a time (e.g. joint plot analysis). The pairs of features are then ranked by score and visualized using the lower left triangle of a feature co-occurence matrix.

The default ranking algorithm is covariance, which attempts to compute the mean value of the product

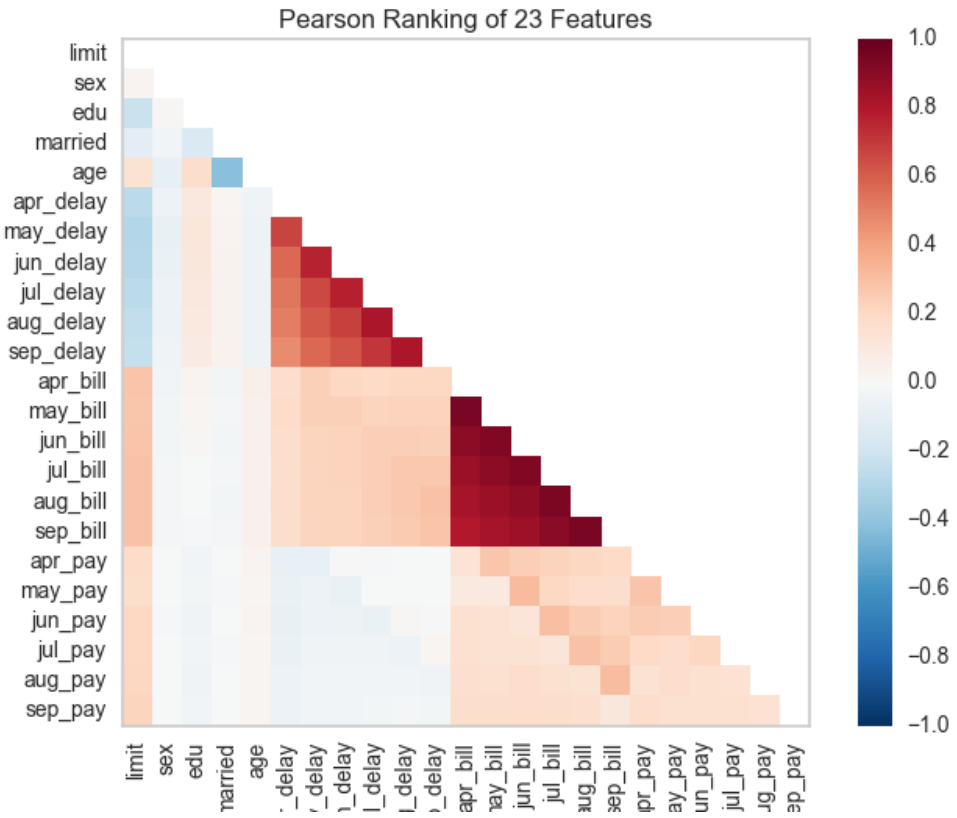
of deviations of variates from their respective means. Covariance loosely attempts to detect a colinear relationship between features.

```
# Instantiate the visualizer with the Covariance ranking algorithm
visualizer = Rank2D(features=features, algorithm='covariance')

visualizer.fit(X, y)                # Fit the data to the visualizer
visualizer.transform(X)             # Transform the data
visualizer.poof()                   # Draw/show/poof the data
```



Alternatively we can utilize a linear correlation algorithm such as a Pearson score to similarly detect colinear relationships. Compare the output from Pearson below to the covariance ranking above.

```
# Instantiate the visualizer with the Pearson ranking algorithm
visualizer = Rank2D(features=features, algorithm='pearson')

visualizer.fit(X, y)                # Fit the data to the visualizer
visualizer.transform(X)             # Transform the data
visualizer.poof()                   # Draw/show/poof the data
```



API Reference

Implements 1D (histograms) and 2D (joint plot) feature rankings.

`class yellowbrick.features.rankd.Rank1D`(*ax=None*, *algorithm='shapiro'*, *features=None*, *orient='h'*, *show_feature_names=True*, ***kwargs*)

基类: `yellowbrick.features.rankd.RankDBase`

Rank1D computes a score for each feature in the data set with a specific metric or algorithm (e.g. Shapiro-Wilk) then returns the features ranked as a bar plot.

Parameters

ax [matplotlib Axes, default: None] The axis to plot the figure on. If None is passed in the current axes will be used (or generated if required).

algorithm [one of {'shapiro', }, default: 'shapiro'] The ranking algorithm to use, default is 'Shapiro-Wilk.

features [list] A list of feature names to use. If a DataFrame is passed to fit and features is None, feature names are selected as the columns of the DataFrame.

orient ['h' or 'v'] Specifies a horizontal or vertical bar chart.

show_feature_names [boolean, default: True] If True, the feature names are used to label the x and y ticks in the plot.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Examples

```
>>> visualizer = Rank1D()
>>> visualizer.fit(X, y)
>>> visualizer.transform(X)
>>> visualizer.poof()
```

Attributes

ranks_ [ndarray] An array of rank scores with shape (n,), where n is the number of features. It is computed during *fit*.

draw(***kwargs*)

Draws the bar plot of the ranking array of features.

ranking_methods = {'shapiro': <function Rank1D.<lambda>>}

class yellowbrick.features.rankd.**Rank2D**(*ax=None*, *algorithm='pearson'*, *features=None*, *colormap='RdBu_r'*, *show_feature_names=True*, ***kwargs*)

基类: yellowbrick.features.rankd.RankDBase

Rank2D performs pairwise comparisons of each feature in the data set with a specific metric or algorithm (e.g. Pearson correlation) then returns them ranked as a lower left triangle diagram.

Parameters

ax [matplotlib Axes, default: None] The axis to plot the figure on. If None is passed in the current axes will be used (or generated if required).

algorithm [one of {'pearson', 'covariance'}, default: 'pearson'] The ranking algorithm to use, default is Pearson correlation.

features [list] A list of feature names to use. If a DataFrame is passed to fit and features is None, feature names are selected as the columns of the DataFrame.

colormap [string or cmap, default: 'RdBu_r'] optional string or matplotlib cmap to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

show_feature_names [boolean, default: True] If True, the feature names are used to label the axis ticks in the plot.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

Examples

```
>>> visualizer = Rank2D()
>>> visualizer.fit(X, y)
>>> visualizer.transform(X)
>>> visualizer.poof()
```

Attributes

ranks_ [ndarray] An array of rank scores with shape (n,n), where n is the number of features. It is computed during *fit*.

draw(***kwargs*)

Draws the heatmap of the ranking matrix of variables.

ranking_methods = {'covariance': <function Rank2D.<lambda>>, 'pearson': <function Rank2D.<lambda>>

Parallel Coordinates

Parallel coordinates displays each feature as a vertical axis spaced evenly along the horizontal, and each instance as a line drawn between each individual axis. This allows many dimensions; in fact given infinite horizontal space (e.g. a scrollbar), an infinite number of dimensions can be displayed!

Data scientists use this method to detect clusters of instances that have similar classes, and to note features that have high variance or different distributions.

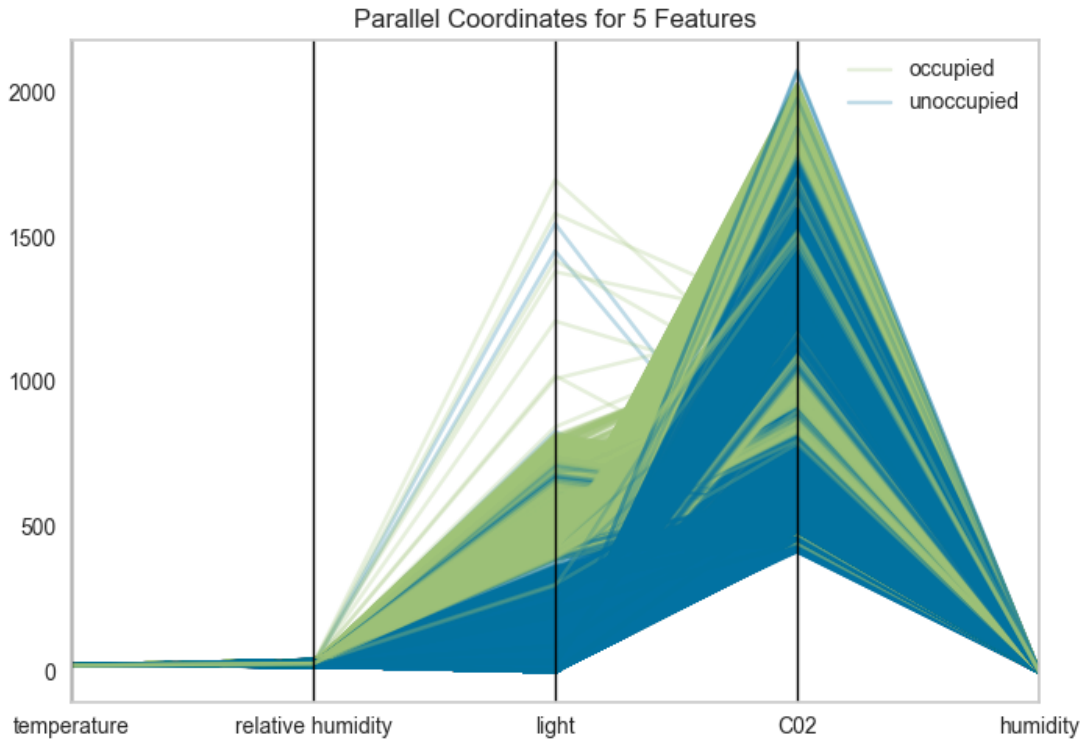
```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']

# Extract the numpy arrays from the data frame
```

(下页继续)

(续上页)

```
X = data[features].as_matrix()
y = data.occupancy.as_matrix()
```

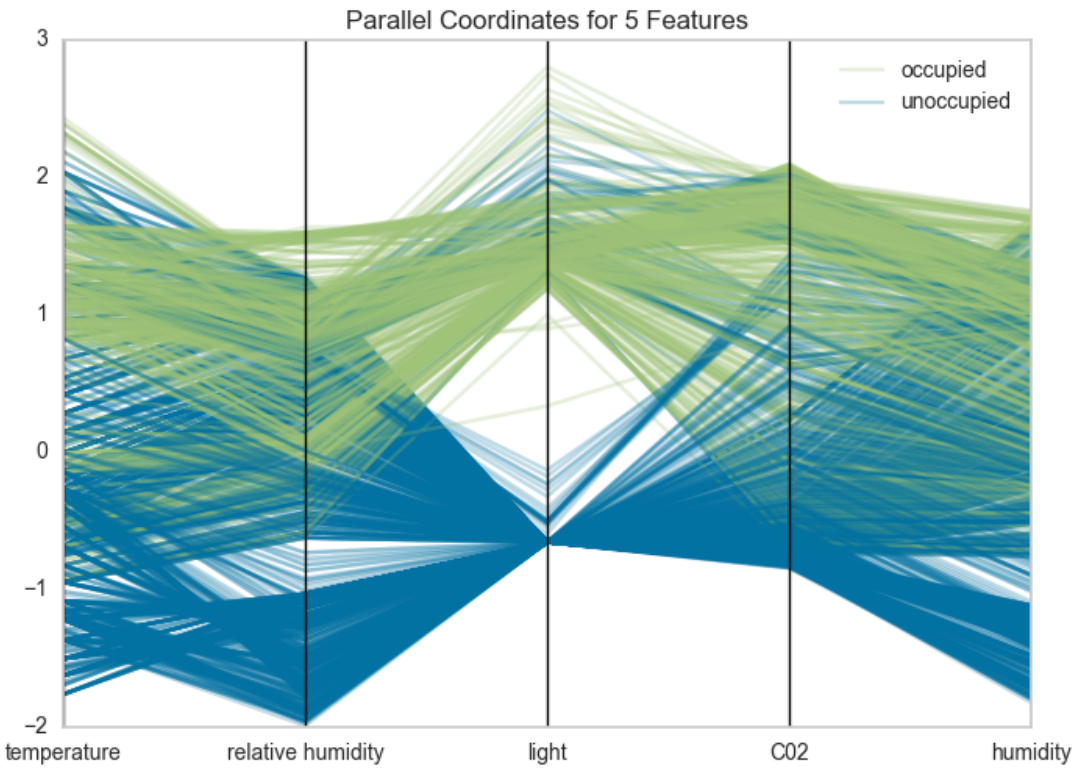
```
# Instantiate the visualizer
visualizer = ParallelCoordinates(classes=classes, features=features)

visualizer.fit(X, y)       # Fit the data to the visualizer
visualizer.transform(X)    # Transform the data
visualizer.poof()          # Draw/show/poof the data
```



Parallel coordinates can take a long time to draw since each instance is represented by a line for each feature. Worse, this time is not well spent since a lot of overlap in the visualization makes the parallel coordinates less understandable. To fix this, pass the `sample` keyword argument to the visualizer with a percentage to randomly sample from the dataset.

Additionally the domain of each feature may make the visualization hard to interpret. In the above visualization, the domain of the `light` feature is from in `[0, 1600]`, far larger than the range of temperature in `[50, 96]`. A normalization methodology can be applied to change the range of features to `[0,1]`. Try using `minmax`, `minabs`, `standard`, `l1`, or `l2` normalization to change perspectives in the parallel coordinates:

```
# Instantiate the visualizer
visualizer = ParallelCoordinates(
    classes=classes, features=features,
    normalize='standard', sample=0.1,
)

visualizer.fit(X, y)       # Fit the data to the visualizer
visualizer.transform(X)    # Transform the data
visualizer.poof()          # Draw/show/poof the data
```



API Reference

Implementations of parallel coordinates for multi-dimensional feature analysis. There are a variety of parallel coordinates from Andrews Curves to coordinates that optimize column order.

class yellowbrick.features.pcoords.**ParallelCoordinates**(*ax=None*, *features=None*, *classes=None*, *normalize=None*, *sample=1.0*, *color=None*, *colormap=None*, *vlines=True*, *vlines_kwds=None*, ***kwargs*)

基类：`yellowbrick.features.base.DataVisualizer`

Parallel coordinates displays each feature as a vertical axis spaced evenly along the horizontal, and each instance as a line drawn between each individual axis.

Parameters

ax [matplotlib Axes, default: None] The axis to plot the figure on. If None is passed in the current axes will be used (or generated if required).

features [list, default: None] a list of feature names to use If a DataFrame is passed to fit and features is None, feature names are selected as the columns of the DataFrame.

classes [list, default: None] a list of class names for the legend If classes is None and a y value is passed to fit then the classes are selected from the target vector.

normalize [string or None, default: None] specifies which normalization method to use, if any Current supported options are 'minmax', 'maxabs', 'standard', 'l1', and 'l2'.

sample [float or int, default: 1.0] specifies how many examples to display from the data If int, specifies the maximum number of samples to display. If float, specifies a fraction between 0 and 1 to display.

color [list or tuple, default: None] optional list or tuple of colors to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

colormap [string or cmap, default: None] optional string or matplotlib cmap to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

vlines [boolean, default: True] flag to determine vertical line display

vlines_kwds [dict, default: None] options to style or display the vertical lines, default: None

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

Examples

```
>>> visualizer = ParallelCoordinates()
>>> visualizer.fit(X, y)
```

(下页继续)

(续上页)

```
>>> visualizer.transform(X)
>>> visualizer.poof()
```

draw(*X*, *y*, ***kwargs*)

Called from the fit method, this method creates the parallel coordinates canvas and draws each instance and vertical lines on it.

finalize(****kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

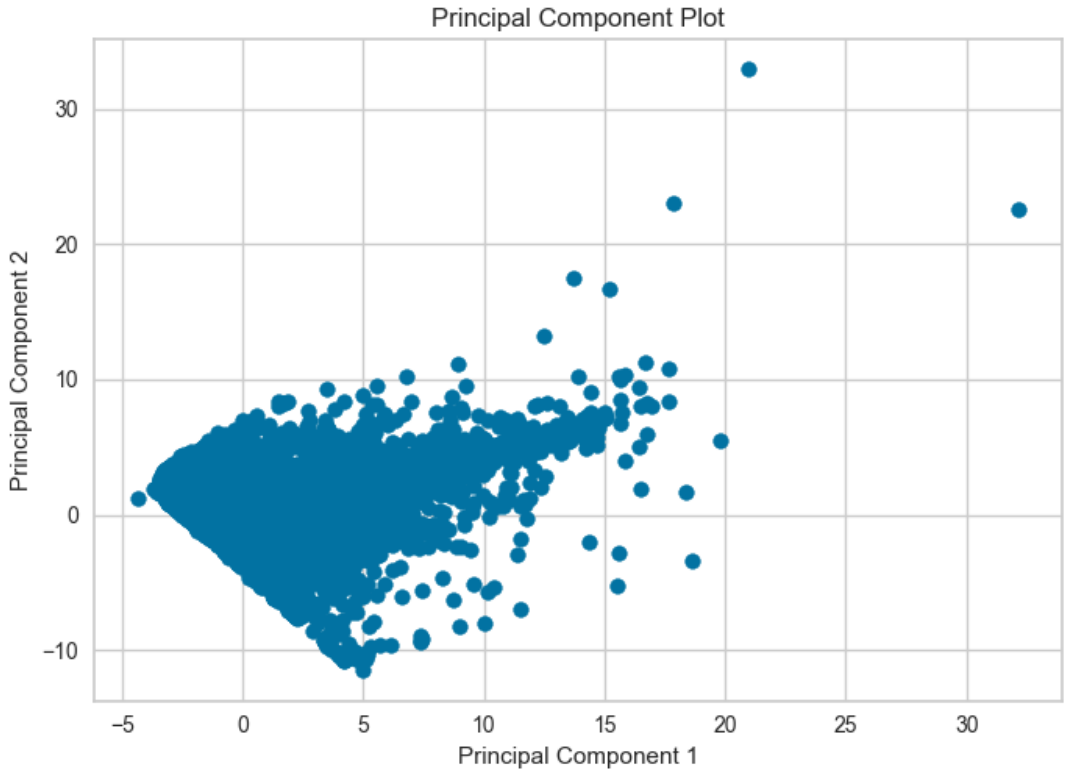
```
normalizers = {'l1': Normalizer(copy=True, norm='l1'), 'l2': Normalizer(copy=True, norm='l2'), 'ma
```

PCA Projection

The PCA Decomposition visualizer utilizes principle component analysis to decompose high dimensional data into two or three dimensions so that each instance can be plotted in a scatter plot. The use of PCA means that the projected dataset can be analyzed along axes of principle variation and can be interpreted to determine if spherical distance metrics can be utilized.

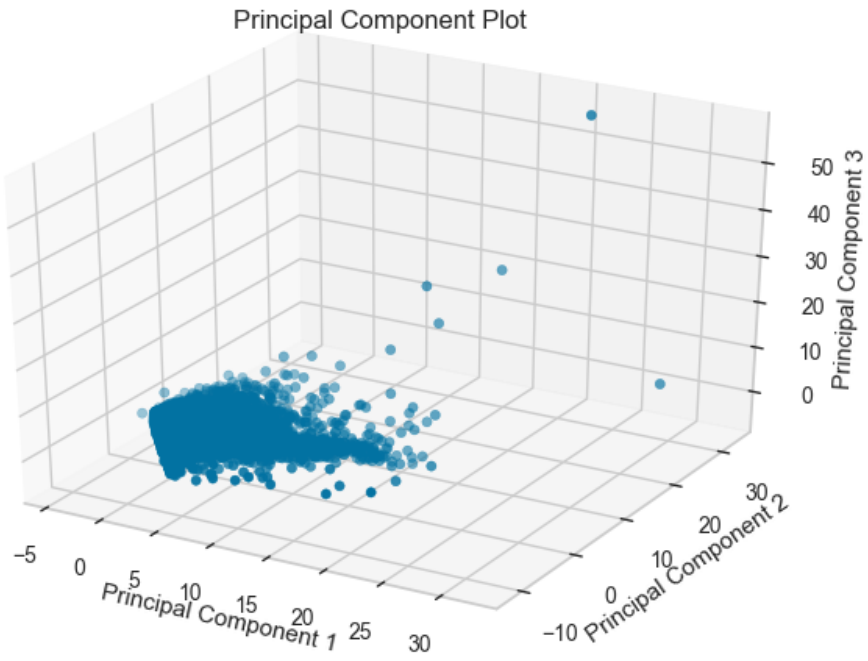
```
# Load the classification data set
data = load_data('credit')

# Specify the features of interest
features = [
    'limit', 'sex', 'edu', 'married', 'age', 'apr_delay', 'may_delay',
    'jun_delay', 'jul_delay', 'aug_delay', 'sep_delay', 'apr_bill', 'may_bill',
    'jun_bill', 'jul_bill', 'aug_bill', 'sep_bill', 'apr_pay', 'may_pay', 'jun_pay',
    'jul_pay', 'aug_pay', 'sep_pay',
]

# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.default.as_matrix()
```

```
visualizer = PCADecomposition(scale=True, center=False, col=y)
visualizer.fit_transform(X,y)
visualizer.poof()
```



The PCA projection can also be plotted in three dimensions to attempt to visualize more princple components and get a better sense of the distribution in high dimensions.

```
visualizer = PCADecomposition(scale=True, center=False, col=y, proj_dim=3)
visualizer.fit_transform(X,y)
visualizer.poof()
```



API Reference

Decomposition based feature visualization with PCA.

class yellowbrick.features.pca.**PCADecomposition**(*ax=None*, *scale=True*, *color=None*, *proj_dim=2*, *colormap='RdBu'*, ***kwargs*)

基类: yellowbrick.features.base.FeatureVisualizer

Produce a two or three dimensional principal component plot of the data array `X` projected onto it's largest sequential principal components. It is common practice to scale the data array `X` before applying a PC decomposition. Variable scaling can be controlled using the `scale` argument.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features.

y [ndarray or Series of length n] An array or series of target or class values.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes. will be used (or generated if required).

scale [bool, default: True] Boolean that indicates if user wants to scale data.

proj_dim [int, default: 2] Dimension of the PCA visualizer.

color [list or tuple of colors, default: None] Specify the colors for each individual class.

colormap [string or cmap, default: None] Optional string or matplotlib cmap to colorize lines. Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Examples

```
>>> from sklearn import datasets
>>> iris = datasets.load_iris()
>>> X = iris.data
>>> y = iris.target
>>> params = {'scale': True, 'center': False, 'col': y}
>>> visualizer = PCADecomposition(**params)
>>> visualizer.fit(X)
>>> visualizer.transform(X)
>>> visualizer.poof()
```

`draw`(***kwargs*)

The fitting or transformation process usually calls draw (not the user). This function is implemented for developers to hook into the matplotlib interface and to create an internal representation of the data the visualizer was trained on in the form of a figure or axes.

Parameters

kwargs: dict generic keyword arguments.

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps.

Parameters

kwargs: dict generic keyword arguments.

Notes

The user calls poof and poof calls finalize. Developers should implement visualizer-specific finalization methods like setting titles or axes labels, etc.

`fit`(*X*, *y=None*, ***kwargs*)

Fits a visualizer to data and is the primary entry point for producing a visualization. Visualizers are Scikit-Learn Estimator objects, which learn from data in order to produce a visual analysis

or diagnostic. They can do this either by fitting features related data or by fitting an underlying model (or models) and visualizing their results.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

kwargs: dict Keyword arguments passed to the drawing functionality or to the Scikit-Learn API. See visualizer specific details for how to use the kwargs to modify the visualization or fitting process.

Returns

self [visualizer] The fit method must always return self to support pipelines.

`transform`(*X*, *y=None*, ***kwargs*)

Primarily a pass-through to ensure that the feature visualizer will work in a pipeline setting. This method can also call drawing methods in order to ensure that the visualization is constructed.

This method must return a numpy array with the same shape as X.

Feature Importances

The feature engineering process involves selecting the *minimum* required features to produce a valid model because the more features a model contains, the more complex it is (and the more sparse the data), therefore the more sensitive the model is to errors due to variance. A common approach to eliminating features is to describe their relative importance to a model, then eliminate weak features or combinations of features and re-evalute to see if the model fairs better during cross-validation.

Many model forms describe the underlying impact of features relative to each other. In Scikit-Learn, Decision Tree models and ensembles of trees such as Random Forest, Gradient Boosting, and Ada Boost provide a `feature_importances_` attribute when fitted. The Yellowbrick `FeatureImportances` visualizer utilizes this attribute to rank and plot relative importances. Let's start with an example; first load a classification dataset as follows:

```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest
features = [
    "temperature", "relative humidity", "light", "CO2", "humidity"
]

# Extract the instances and target
```

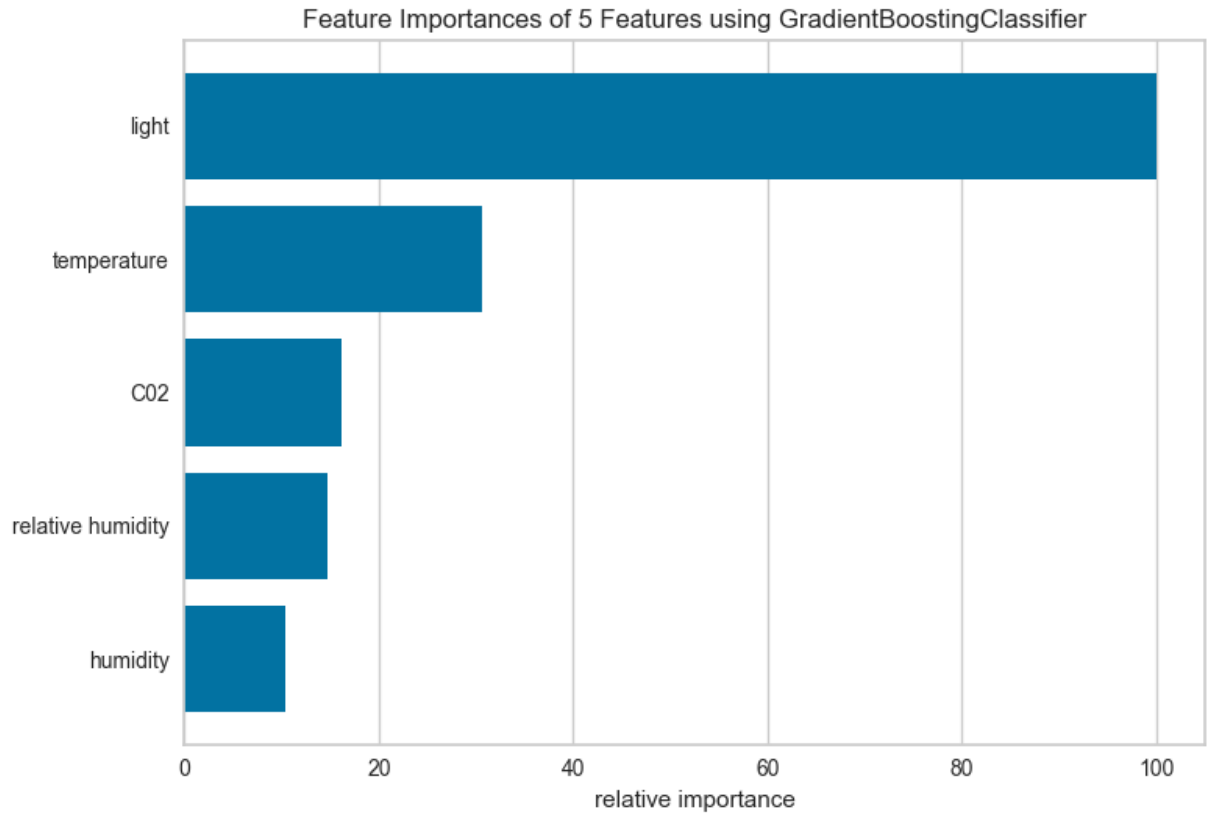
(下页继续)

(续上页)

```
X = data[features]
y = data.occupancy
```

Once the dataset has been loaded, we can create a new figure (this is optional, if an `Axes` isn't specified, Yellowbrick will use the current figure or create one). We can then fit a `FeatureImportances` visualizer with a `GradientBoostingClassifier` to visualize the ranked features:

```
from sklearn.ensemble import GradientBoostingClassifier
from yellowbrick.features import FeatureImportances

# Create a new matplotlib figure
fig = plt.figure()
ax = fig.add_subplot()

viz = FeatureImportances(GradientBoostingClassifier(), ax=ax)
viz.fit(X, y)
viz.poof()
```



The above figure shows the features ranked according to the explained variance each feature contributes to the model. In this case the features are plotted against their *relative importance*, that is the percent importance of

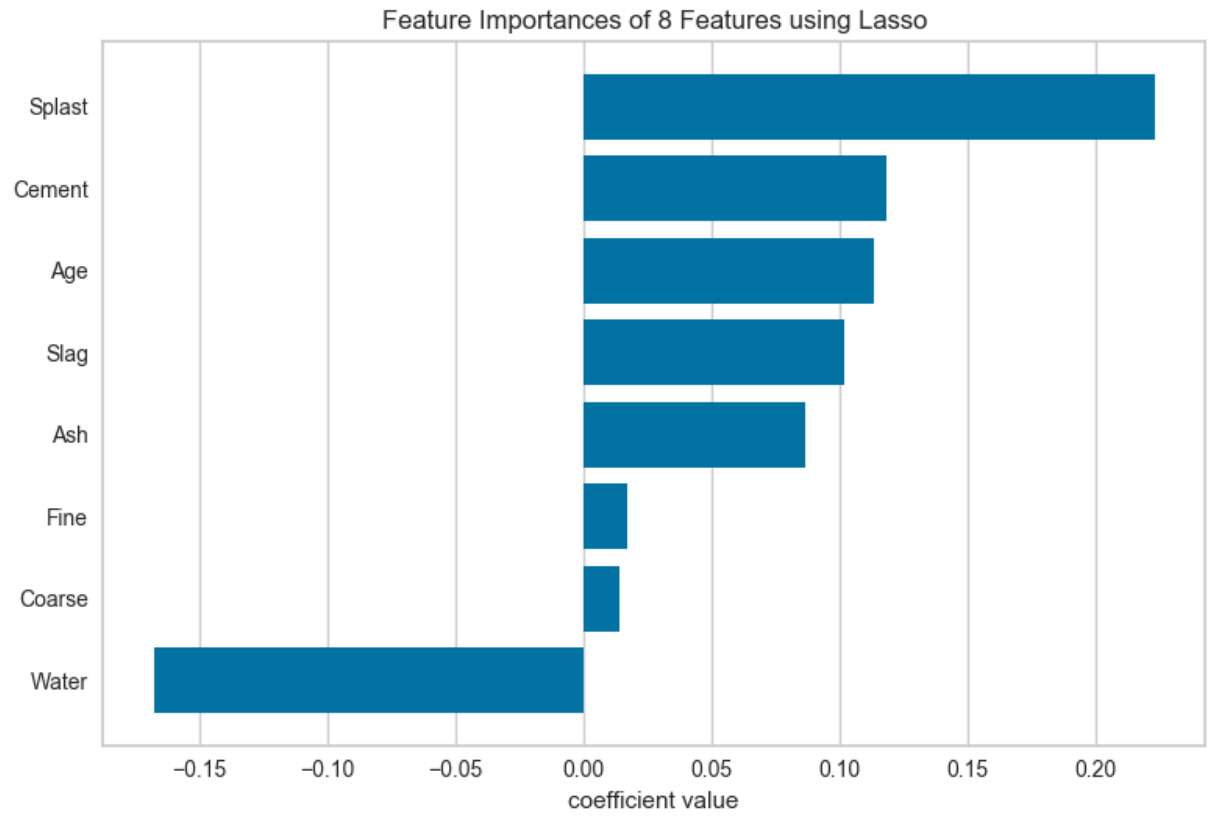
the most important feature. The visualizer also contains `features_` and `feature_importances_` attributes to get the ranked numeric values.

For models that do not support a `feature_importances_` attribute, the `FeatureImportances` visualizer will also draw a bar plot for the `coef_` attribute that many linear models provide. First we start by loading a regression dataset:

```
# Load a regression data set
data = load_data("concrete")

# Specify the features of interest
features = [
    'cement','slag','ash','water','splast','coarse','fine','age'
]

# Extract the instances and target
X = concrete[feats]
y = concrete.strength
```

When using a model with a `coef_` attribute, it is better to set `relative=False` to draw the true magnitude of the coefficient (which may be negative). We can also specify our own set of labels if the dataset does not have column names or to print better titles. In the example below we title case our features for better readability:

```
# Create a new figure
fig = plt.figure()
ax = fig.add_subplot()

# Title case the feature for better display and create the visualizer
labels = list(map(lambda s: s.title(), features))
viz = FeatureImportances(Lasso(), ax=ax, labels=labels, relative=False)

# Fit and show the feature importances
viz.fit(X, y)
viz.poof()
```



注解: The interpretation of the importance of coeficients depends on the model; see the discussion below for more details.

Discussion

Generalized linear models compute a predicted independent variable via the linear combination of an array of coefficients with an array of dependent variables. GLMs are fit by modifying the coefficients so as to minimize error and regularization techniques specify how the model modifies coefficients in relation to each other. As a result, an opportunity presents itself: larger coefficients are necessarily "more informative" because they contribute a greater weight to the final prediction in most cases.

Additionally we may say that instance features may also be more or less "informative" depending on the product of the instance feature value with the feature coefficient. This creates two possibilities:

1. We can compare models based on ranking of coefficients, such that a higher coefficient is "more informative".
2. We can compare instances based on ranking of feature/coefficient products such that a higher product is "more informative".

In both cases, because the coefficient may be negative (indicating a strong negative correlation) we must rank features by the absolute values of their coefficients. Visualizing a model or multiple models by most informative feature is usually done via bar chart where the y-axis is the feature names and the x-axis is numeric value of the coefficient such that the x-axis has both a positive and negative quadrant. The bigger the size of the bar, the more informative that feature is.

This method may also be used for instances; but generally there are very many instances relative to the number models being compared. Instead a heatmap grid is a better choice to inspect the influence of features on individual instances. Here the grid is constructed such that the x-axis represents individual features, and the y-axis represents individual instances. The color of each cell (an instance, feature pair) represents the magnitude of the product of the instance value with the feature's coefficient for a single model. Visual inspection of this diagnostic may reveal a set of instances for which one feature is more predictive than another; or other types of regions of information in the model itself.

API Reference

Implementation of a feature importances visualizer. This visualizer sits in kind of a weird place since it is technically a model scoring visualizer, but is generally used for feature engineering.

`class yellowbrick.features.importances.FeatureImportances`(*model*, *ax=None*, *labels=None*, *relative=True*, *absolute=False*, *xlabel=None*, ***kwargs*)

基类：`yellowbrick.base.ModelVisualizer`

Displays the most informative features in a model by showing a bar chart of features ranked by their importances. Although primarily a feature engineering mechanism, this visualizer requires a model that has either a `coef_` or `feature_importances_` parameter after fit.

Parameters

model [Estimator] A Scikit-Learn estimator that learns feature importances. Must support either `coef_` or `feature_importances_` parameters.

ax [matplotlib Axes, default: None] The axis to plot the figure on. If None is passed in the current axes will be used (or generated if required).

labels [list, default: None] A list of feature names to use. If a DataFrame is passed to fit and features is None, feature names are selected as the column names.

relative [bool, default: True] If true, the features are described by their relative importance as a percentage of the strongest feature component; otherwise the raw numeric description of the feature importance is shown.

absolute [bool, default: False] Make all coeficients absolute to more easily compare negative coeficients with positive ones.

xlabel [str, default: None] The label for the X-axis. If None is automatically determined by the underlying model and options provided.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Examples

```
>>> from sklearn.ensemble import GradientBoostingClassifier
>>> visualizer = FeatureImportances(GradientBoostingClassifier())
>>> visualizer.fit(X, y)
>>> visualizer.poof()
```

Attributes

features_ [np.array] The feature labels ranked according to their importance

feature_importances_ [np.array] The numeric value of the feature importance computed by the model

`draw`(***kwargs*)

Draws the feature importances as a bar chart; called from fit.

`finalize`(***kwargs*)

Finalize the drawing setting labels and title.

`fit`(*X*, *y=None*, ***kwargs*)

Fits the estimator to discover the feature importances described by the data, then draws those importances as a bar plot.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

kwargs [dict] Keyword arguments passed to the fit method of the estimator.

Returns

self [visualizer] The fit method must always return self to support pipelines.

Direct Data Visualization

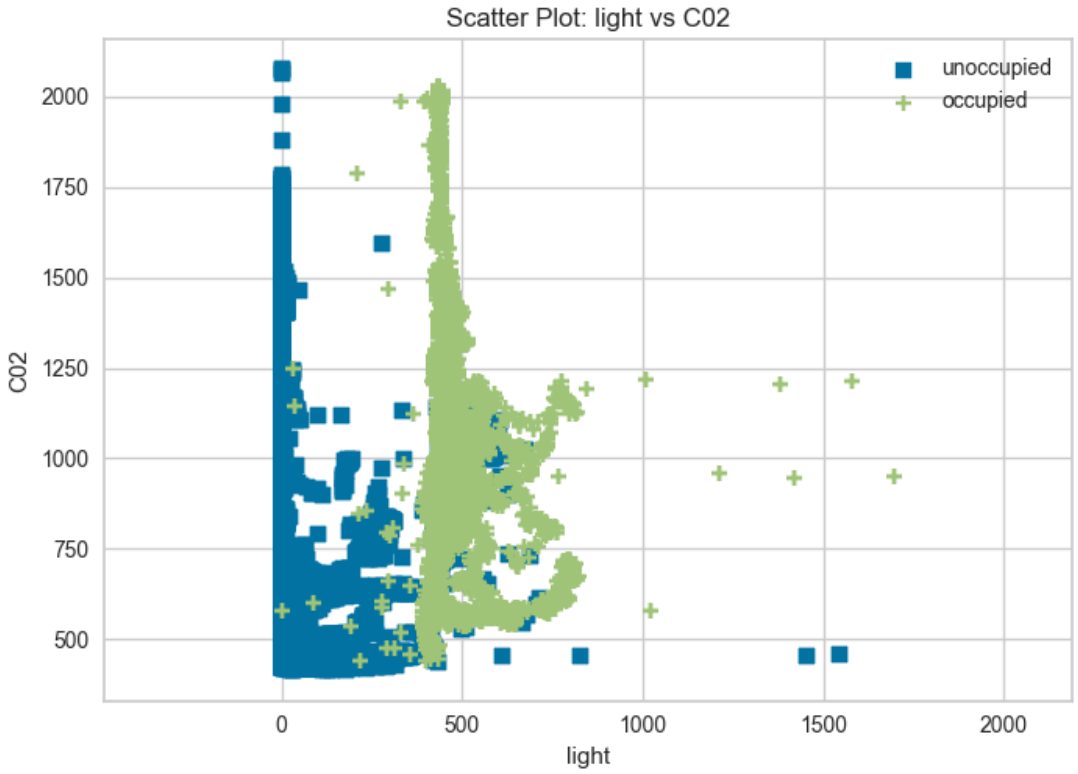
Sometimes for feature analysis you simply need a scatter plot to determine the distribution of data. Machine learning operates on high dimensional data, so the number of dimensions has to be filtered. As a result these visualizations are typically used as the base for larger visualizers; however you can also use them to quickly plot data during ML analysis.

Scatter Visualization

A scatter visualizer simply plots two features against each other and colors the points according to the target. This can be useful in assessing the relationship of pairs of features to an individual target.

```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']

# Extract the numpy arrays from the data frame
X = data[features]
y = data.occupancy
```

```
from yellowbrick.features import ScatterVisualizer

visualizer = ScatterVisualizer(x='light', y='CO2', classes=classes)

visualizer.fit(X, y)
visualizer.transform(X)
visualizer.poof()
```

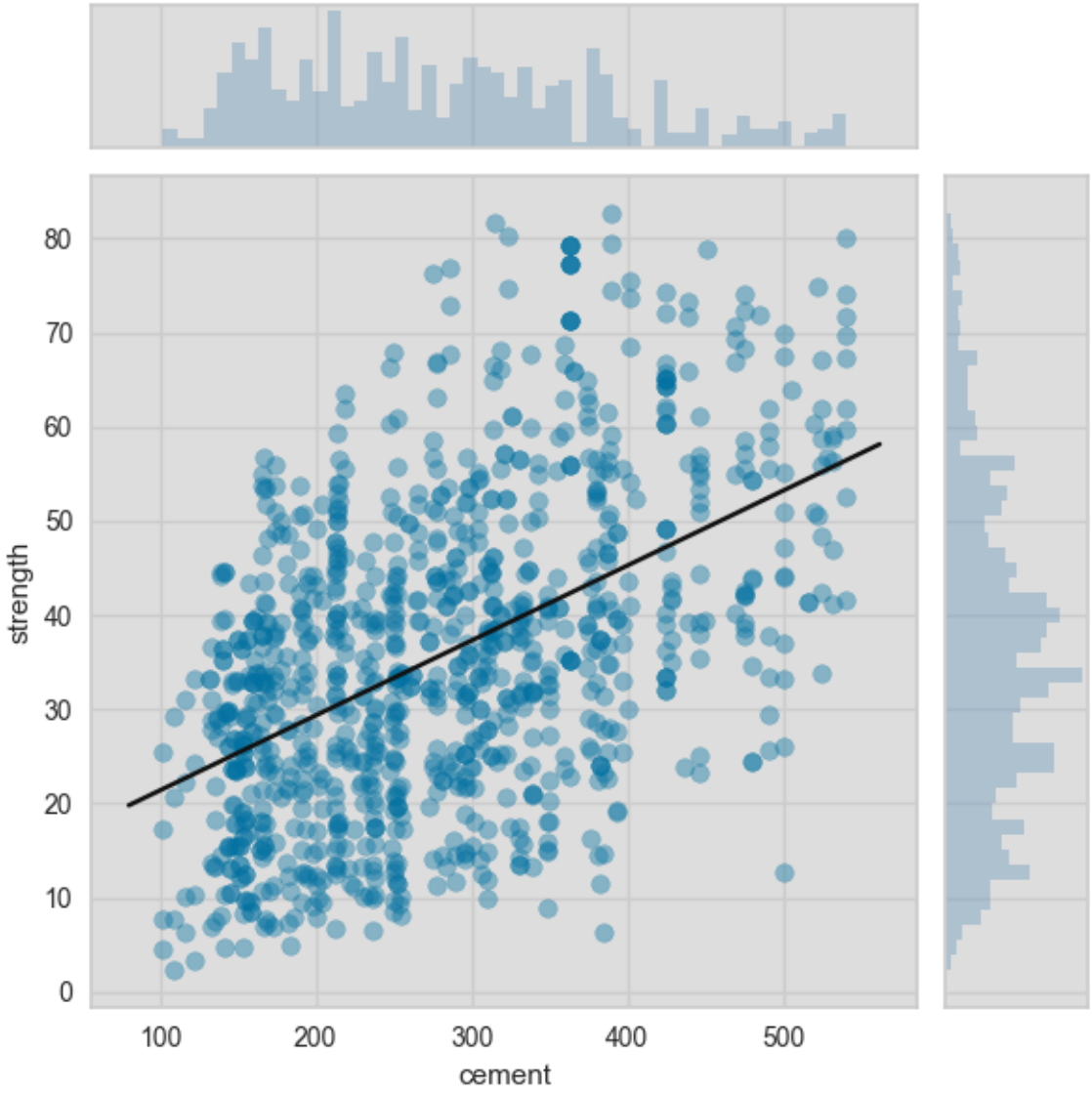


Joint Plot Visualization

A joint plot visualizer plots a feature against the target and shows the distribution of each via a histogram on each axis.

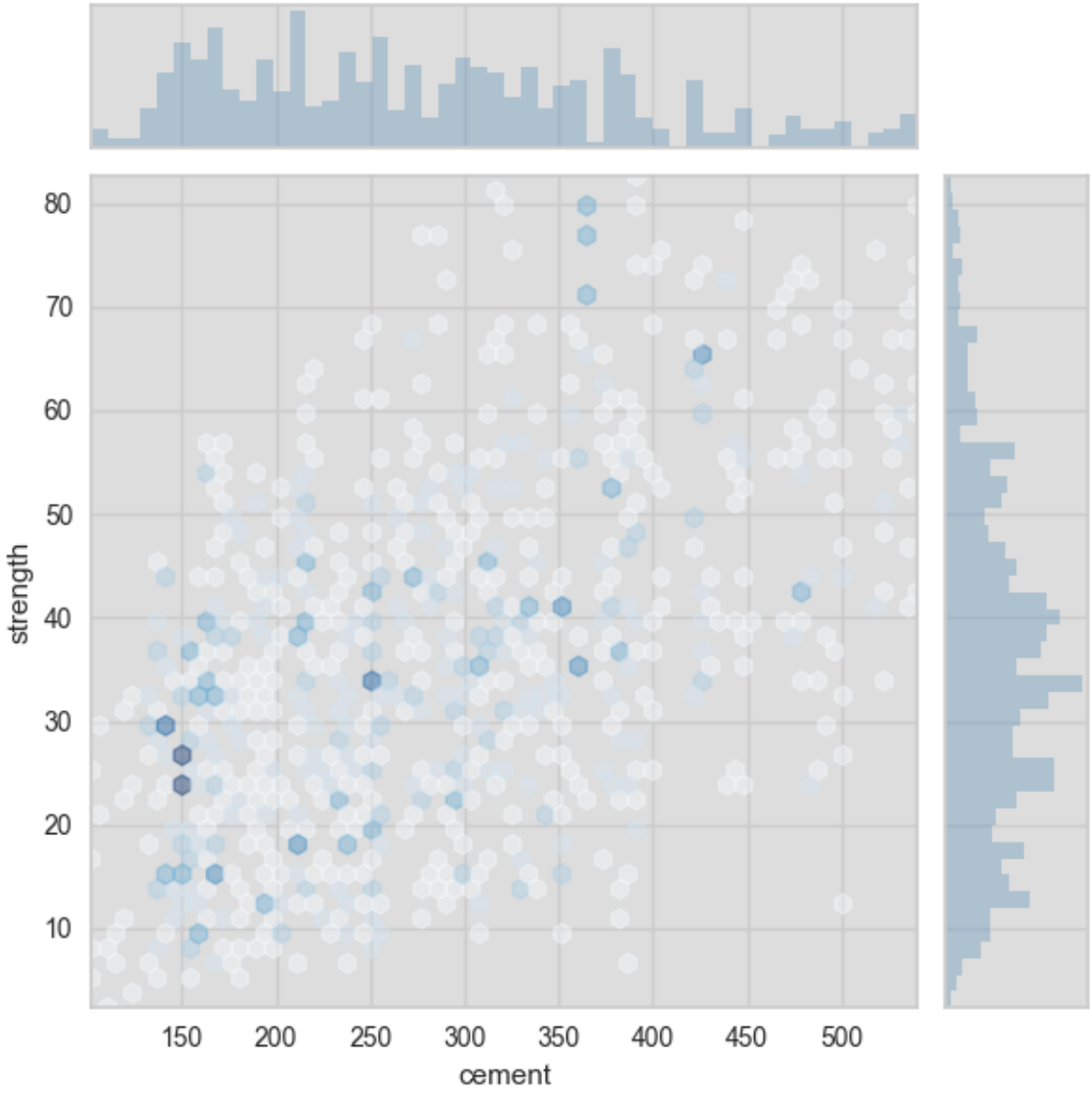
```
# Load the data
df = load_data('concrete')
feature = 'cement'
target = 'strength'

# Get the X and y data from the DataFrame
X = df[feature]
y = df[target]
```

```
visualizer = JointPlotVisualizer(feature=feature, target=target)

visualizer.fit(X, y)
visualizer.poof()
```



The joint plot visualizer can also be plotted with hexbins in the case of many, many points.

```
visualizer = JointPlotVisualizer(
    feature=feature, target=target, joint_plot='hex'
)

visualizer.fit(X, y)
visualizer.poof()
```



API Reference

Implements a 2D scatter plot for feature analysis.

`class yellowbrick.features.scatter.ScatterVisualizer`(*ax=None*, *x=None*, *y=None*, *features=None*, *classes=None*, *color=None*, *colormap=None*, *markers=None*, ***kwargs*)

基类：`yellowbrick.features.base.DataVisualizer`

ScatterVisualizer is a bivariate feature data visualization algorithm that plots using the Cartesian coordinates of each point.

Parameters

ax [a matplotlib plot, default: None]

The axis to plot the figure on.

x [string, default: None] The feature name that corresponds to a column name or index postion in the matrix that will be plotted against the x-axis

y [string, default: None] The feature name that corresponds to a column name or index postion in the matrix that will be plotted against the y-axis

features [a list of two feature names to use, default: None] List of two features that correspond to the columns in the array. The order of the two features correspond to X and Y axes on the graph. More than two feature names or columns will raise an error. If a DataFrame is passed to fit and features is None, feature names are selected that are the columns of the DataFrame.

classes [a list of class names for the legend, default: None] If classes is None and a y value is passed to fit then the classes are selected from the target vector.

color [optional list or tuple of colors to colorize points, default: None] Use either color to colorize the points on a per class basis or colormap to color them on a continuous scale.

colormap [optional string or matplotlib cmap to colorize points, default: None] Use either color to colorize the points on a per class basis or colormap to color them on a continuous scale.

markers [iterable of strings, default: ,+o*vhd] Matplotlib style markers for points on the scatter plot points

kwargs : keyword arguments passed to the super class.

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

`draw`(*X*, *y*, ***kwargs*)

Called from the fit method, this method creates a scatter plot that draws each instance as a class or target colored point, whose location is determined by the feature data set.

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`fit`(*X*, *y=None*, ***kwargs*)

The fit method is the primary drawing input for the parallel coords visualization since it has both the X and y data required for the viz and the transform method does not.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with 2 features

y [ndarray or Series of length n] An array or series of target or class values

kwargs [dict] Pass generic arguments to the drawing method

Returns

self [instance] Returns the instance of the transformer/visualizer

`class yellowbrick.features.jointplot.JointPlotVisualizer`(*ax=None*, *feature=None*, *target=None*, *joint_plot='scatter'*, *joint_args=None*, *xy_plot='hist'*, *xy_args=None*, *size=600*, *ratio=5*, *space=0.2*, ***kwargs*)

基类：`yellowbrick.features.base.FeatureVisualizer`

JointPlotVisualizer allows for a simultaneous visualization of the relationship between two variables and the distrbution of each individual variable. The relationship is plotted along the joint axis and univariate distributions are plotted on top of the x axis and to the right of the y axis.

Parameters

ax: matplotlib Axes, default: None This is inherited from FeatureVisualizer but is defined within JointPlotVisualizer since there are three axes objects.

feature: string, default: None The name of the X variable If a DataFrame is passed to fit and feature is None, feature is selected as the column of the DataFrame. There must be only one column in the DataFrame.

target: string, default: None The name of the Y variable If target is None and a y value is passed to fit then the target is selected from the target vector.

joint_plot: one of {'scatter', 'hex'}, default: 'scatter' The type of plot to render in the joint axis Currently, the choices are scatter and hex. Use scatter for small datasets and hex for large datasets

joint_args: dict, default: None Keyword arguments used for customizing the joint plot:

Property	Description
alpha	transparency
facecolor	background color of the joint axis
aspect	aspect ratio
fit	used if scatter is selected for joint_plot to draw a best fit line - values can be True or False. Uses `Yellowbrick.bestfit`
estimator	used if scatter is selected for joint_plot to determine the type of best fit line to use. Refer to Yellowbrick.bestfit for types of estimators that can be used.
x_bins	used if hex is selected to set the number of bins for the x value
y_bins	used if hex is selected to set the number of bins for the y value
cmap	string or matplotlib cmap to colorize lines Use either color to colorize the lines on a per class basis or colormap to color them on a continuous scale.

xy_plot: one of {'hist'}, default: 'hist' The type of plot to render along the x and y axes Currently, the choice is hist

xy_args: dict, default: None Keyword arguments used for customizing the x and y plots:

Property	Description
alpha	transparency
facecolor_x	background color of the x axis
facecolor_y	background color of the y axis
bins	used to set up the number of bins for the hist plot
histcolor_x	used to set the color for the histogram on the x axis
histcolor_y	used to set the color for the histogram on the y axis

size: float, default: 600 Size of each side of the figure in pixels

ratio: float, default: 5 Ratio of joint axis size to the x and y axes height

space: float, default: 0.2 Space between the joint axis and the x and y axes

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

Examples

```
>>> visualizer = JointPlotVisualizer()
>>> visualizer.fit(X,y)
>>> visualizer.poof()
```

`draw`(*X*, *y*, ***kwargs*)

Sets up the layout for the joint plot draw calls `draw_joint` and `draw_xy` to render the visualizations.

`draw_joint`(*X*, *y*, ***kwargs*)

Draws the visualization for the joint axis.

`draw_xy`(*X*, *y*, ***kwargs*)

Draws the visualization for the x and y axes

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`fit`(*X*, *y*, ***kwargs*)

Sets up the X and y variables for the jointplot and checks to ensure that X and y are of the correct data type

Fit calls draw

Parameters

X [ndarray or DataFrame of shape n x 1] A matrix of n instances with 1 feature

y [ndarray or Series of length n] An array or series of the target value

kwargs: dict keyword arguments passed to Scikit-Learn API.

`poof`(***kwargs*)

Creates the labels for the feature and target variables

4.3.4 Regression Visualizers

Regression models attempt to predict a target in a continuous space. Regressor score visualizers display the instances in model space to better understand how the model is making predictions. We currently have implemented three regressor evaluations:

- *Residuals Plot*: plot the difference between the expected and actual values
- *Prediction Error Plot*: plot the expected vs. actual values in model space
- *Alpha Selection*: visual tuning of regularization hyperparameters

Estimator score visualizers *wrap* Scikit-Learn estimators and expose the Estimator API such that they have `fit()`, `predict()`, and `score()` methods that call the appropriate estimator methods under the hood. Score visualizers can wrap an estimator and be passed in as the final step in a `Pipeline` or `VisualPipeline`.

```
# Regression Evaluation Imports

from sklearn.linear_model import Ridge, Lasso
from sklearn.model_selection import train_test_split

from yellowbrick.regressor import PredictionError, ResidualsPlot
from yellowbrick.regressor.alphas import AlphaSelection
```

Residuals Plot

A residuals plot shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

```
# Load the data
df = load_data('concrete')
feature_names = ['cement', 'slag', 'ash', 'water', 'splast', 'coarse', 'fine', 'age']
target_name = 'strength'

# Get the X and y data from the DataFrame
X = df[feature_names].as_matrix()
y = df[target_name].as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Instantiate the linear model and visualizer
ridge = Ridge()
visualizer = ResidualsPlot(ridge)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
visualizer.score(X_test, y_test)  # Evaluate the model on the test data
g = visualizer.poof()             # Draw/show/poof the data
```

API Reference

Regressor visualizers that score residuals: prediction vs. actual data.

class yellowbrick.regressor.residuals.**ResidualsPlot**(*model*, *ax=None*, ***kwargs*)

基类：yellowbrick.regressor.base.RegressionScoreVisualizer

A residual plot shows the residuals on the vertical axis and the independent variable on the horizontal axis.

If the points are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Parameters

model [a Scikit-Learn regressor] Should be an instance of a regressor, otherwise a will raise a YellowbrickTypeError exception on instantiation.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes will be used (or generated if required).

train_color [color, default: 'b'] Residuals for training data are ploted with this color but also given an opacity of 0.5 to ensure that the test data residuals are more visible. Can be any matplotlib color.

test_color [color, default: 'g'] Residuals for test data are plotted with this color. In order to create generalizable models, reserved test data residuals are of the most analytical interest, so these points are highlighted by hvaing full opacity. Can be any matplotlib color.

line_color [color, default: dark grey] Defines the color of the zero error line, can be any matplotlib color.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

ResidualsPlot is a ScoreVisualizer, meaning that it wraps a model and its primary entry point is the *score()* method.

Examples

```
>>> from yellowbrick.regressor import ResidualsPlot
>>> from sklearn.linear_model import Ridge
>>> model = ResidualsPlot(Ridge())
>>> model.fit(X_train, y_train)
>>> model.score(X_test, y_test)
>>> model.poof()
```

`draw`(*y_pred*, *residuals*, *train=False*, ***kwargs*)

Parameters

y_pred [ndarray or Series of length n] An array or series of predicted target values

residuals [ndarray or Series of length n] An array or series of the difference between the predicted and the target values

train [boolean] If False, *draw* assumes that the residual points being plotted are from the test data; if True, *draw* assumes the residuals are the train data.

Returns

——

ax [the axis with the plotted figure]

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`fit`(*X*, *y=None*, ***kwargs*)

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target values

kwargs: keyword arguments passed to Scikit-Learn API.

`score`(*X*, *y=None*, *train=False*, ***kwargs*)

Generates predicted target values using the Scikit-Learn estimator.

Parameters

X [array-like] X (also X_test) are the dependent variables of test set to predict

y [array-like] y (also y_test) is the independent actual variables to score against

train [boolean] If False, *score* assumes that the residual points being plotted are from the test data; if True, *score* assumes the residuals are the train data.

Returns

——

ax [the axis with the plotted figure]

Prediction Error Plot

A prediction error plot shows the actual targets from the dataset against the predicted values generated by our model. This allows us to see how much variance is in the model. Data scientists can diagnose regression models using this plot by comparing against the 45 degree line, where the prediction exactly matches the model.

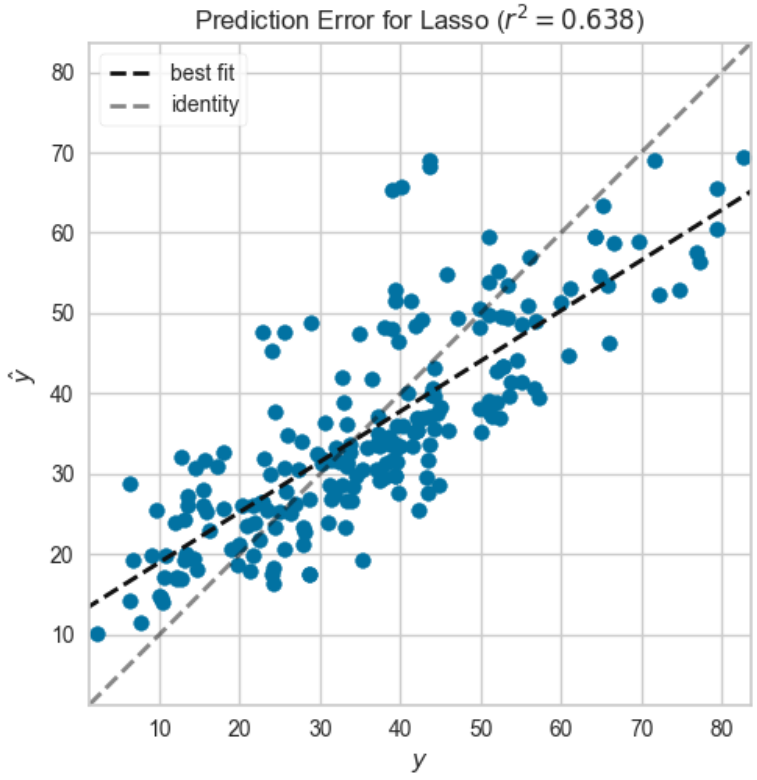
```
# Load the data
df = load_data('concrete')
feature_names = ['cement', 'slag', 'ash', 'water', 'splast', 'coarse', 'fine', 'age']
```

(下页继续)

(续上页)

```
target_name = 'strength'

# Get the X and y data from the DataFrame
X = df[feature_names].as_matrix()
y = df[target_name].as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Instantiate the linear model and visualizer
lasso = Lasso()
visualizer = PredictionError(lasso)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
visualizer.score(X_test, y_test)  # Evaluate the model on the test data
g = visualizer.poof()             # Draw/show/poof the data
```



API Reference

Regressor visualizers that score residuals: prediction vs. actual data.

`class yellowbrick.regressor.residuals.PredictionError`(*model*, *ax=None*, *shared_limits=True*, *bestfit=True*, *identity=True*, ***kwargs*)

基类: `yellowbrick.regressor.base.RegressionScoreVisualizer`

The prediction error visualizer plots the actual targets from the dataset against the predicted values generated by our model(s). This visualizer is used to dectect noise or heteroscedasticity along a range of the target domain.

Parameters

model [a Scikit-Learn regressor] Should be an instance of a regressor, otherwise a will raise a YellowbrickTypeError exception on instantiation.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes will be used (or generated if required).

shared_limits [bool, default: True] If shared_limits is True, the range of the X and Y axis limits will be identical, creating a square graphic with a true 45 degree line. In this form, it is easier to diagnose under- or over- prediction, though the figure will become more sparse. To localize points, set shared_limits to False, but note that this will distort the figure and should be accounted for during analysis.

besfit [bool, default: True] Draw a linear best fit line to estimate the correlation between the predicted and measured value of the target variable. The color of the bestfit line is determined by the `line_color` argument.

identity: bool, default: True Draw the 45 degree identity line, y=x in order to better show the relationship or pattern of the residuals. E.g. to estimate if the model is over- or under- estimating the given values. The color of the identity line is a muted version of the `line_color` argument.

point_color [color] Defines the color of the error points; can be any matplotlib color.

line_color [color] Defines the color of the best fit line; can be any matplotlib color.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

PredictionError is a ScoreVisualizer, meaning that it wraps a model and its primary entry point is the *score()* method.

Examples

```
>>> from yellowbrick.regressor import PredictionError
>>> from sklearn.linear_model import Lasso
>>> model = PredictionError(Lasso())
>>> model.fit(X_train, y_train)
>>> model.score(X_test, y_test)
>>> model.poof()
```

`draw`(*y*, *y_pred*)

Parameters

y [ndarray or Series of length n] An array or series of target or class values

y_pred [ndarray or Series of length n] An array or series of predicted target values

Returns

——

ax [the axis with the plotted figure]

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`score`(*X*, *y=None*, ***kwargs*)

The score function is the hook for visual interaction. Pass in test data and the visualizer will create predictions on the data and evaluate them with respect to the test values. The evaluation will then be passed to draw() and the result of the estimator score will be returned.

Parameters

X [array-like] X (also X_test) are the dependent variables of test set to predict

y [array-like] y (also y_test) is the independent actual variables to score against

Returns

score [float]

Alpha Selection

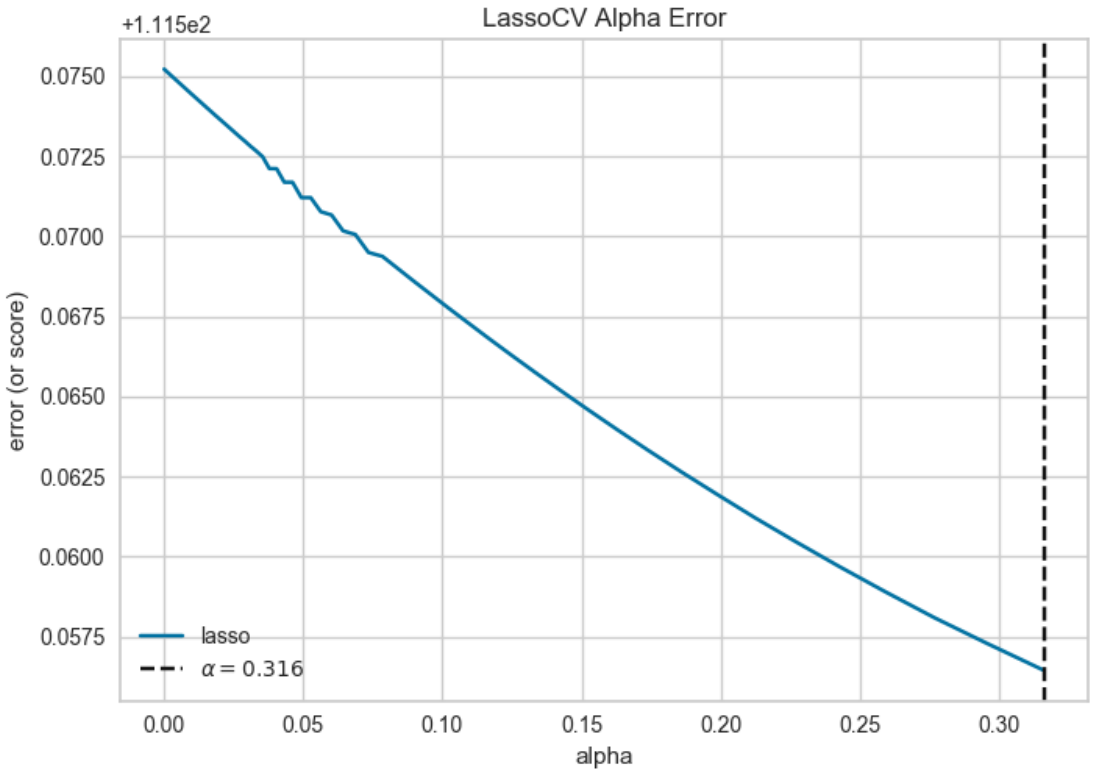
Regularization is designed to penalize model complexity, therefore the higher the alpha, the less complex the model, decreasing the error due to variance (overfit). Alphas that are too high on the other hand increase

the error due to bias (underfit). It is important, therefore to choose an optimal alpha such that the error is minimized in both directions.

The AlphaSelection Visualizer demonstrates how different values of alpha influence model selection during the regularization of linear models. Generally speaking, alpha increases the affect of regularization, e.g. if alpha is zero there is no regularization and the higher the alpha, the more the regularization parameter influences the final model.

```
# Load the data
df = load_data('concrete')
feature_names = ['cement', 'slag', 'ash', 'water', 'splast', 'coarse', 'fine', 'age']
target_name = 'strength'

# Get the X and y data from the DataFrame
X = df[feature_names].as_matrix()
y = df[target_name].as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Create a list of alphas to cross-validate against
alphas = np.logspace(-12, -0.5, 400)

# Instantiate the linear model and visualizer
model = LassoCV(alphas=alphas)
visualizer = AlphaSelection(model)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
g = visualizer.poof()             # Draw/show/poof the data
```



API Reference

Implements alpha selection visualizers for regularization

class yellowbrick.regressor.alphas.**AlphaSelection**(*model*, *ax=None*, ***kwargs*)

基类: yellowbrick.regressor.base.RegressionScoreVisualizer

The Alpha Selection Visualizer demonstrates how different values of alpha influence model selection during the regularization of linear models. Generally speaking, alpha increases the affect of regularization, e.g. if alpha is zero there is no regularization and the higher the alpha, the more the regularization parameter influences the final model.

Regularization is designed to penalize model complexity, therefore the higher the alpha, the less complex the model, decreasing the error due to variance (overfit). Alphas that are too high on the other hand increase the error due to bias (underfit). It is important, therefore to choose an optimal Alpha such that the error is minimized in both directions.

To do this, typically you would you use one of the "RegressionCV" models in Scikit-Learn. E.g. instead of using the `Ridge` (L2) regularizer, you can use `RidgeCV` and pass a list of alphas, which will be selected based on the cross-validation score of each alpha. This visualizer wraps a "RegressionCV" model and visualizes the alpha/error curve. Use this visualization to detect if the model is responding to regularization, e.g. as you increase or decrease alpha, the model responds and error is decreased.

If the visualization shows a jagged or random plot, then potentially the model is not sensitive to that type of regularization and another is required (e.g. L1 or `Lasso` regularization).

Parameters

model [a Scikit-Learn regressor] Should be an instance of a regressor, and specifically one whose name ends with "CV" otherwise a will raise a YellowbrickTypeError exception on instantiation. To use non-CV regressors see: `ManualAlphaSelection`.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes will be used (or generated if required).

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

This class expects an estimator whose name ends with "CV". If you wish to use some other estimator, please see the `ManualAlphaSelection` Visualizer for manually iterating through all alphas and selecting the best one.

This Visualizer hoooks into the Scikit-Learn API during `fit()`. In order to pass a fitted model to the Visualizer, call the `draw()` method directly after instantiating the visualizer with the fitted model.

Note, each "RegressorCV" module has many different methods for storing alphas and error. This visualizer attempts to get them all and is known to work for RidgeCV, LassoCV, LassoLarsCV, and ElasticNetCV. If your favorite regularization method doesn't work, please submit a bug report.

For RidgeCV, make sure `store_cv_values=True`.

Examples

```
>>> from yellowbrick.regressor import AlphaSelection
>>> from sklearn.linear_model import LassoCV
>>> model = AlphaSelection(LassoCV())
>>> model.fit(X, y)
>>> model.poof()
```

`draw()`

Draws the alpha plot based on the values on the estimator.

`finalize()`

Prepare the figure for rendering by setting the title as well as the X and Y axis labels and adding the legend.

`fit`(*X*, *y*, ***kwargs*)

A simple pass-through method; calls fit on the estimator and then draws the alpha-error plot.

class yellowbrick.regressor.alphas.**ManualAlphaSelection**(*model*, *ax=None*, *alphas=None*, *cv=None*, *scoring=None*, ***kwargs*)

基类：*yellowbrick.regressor.alphas.AlphaSelection*

The `AlphaSelection` visualizer requires a "RegressorCV", that is a specialized class that performs cross-validated alpha-selection on behalf of the model. If the regressor you wish to use doesn't have an associated "CV" estimator, or for some reason you would like to specify more control over the alpha selection process, then you can use this manual alpha selection visualizer, which is essentially a wrapper for `cross_val_score`, fitting a model for each alpha specified.

Parameters

model [a Scikit-Learn regressor] Should be an instance of a regressor, and specifically one whose name doesn't end with "CV". The regressor must support a call to `set_params(alpha=alpha)` and be fit multiple times. If the regressor name ends with "CV" a `YellowbrickValueError` is raised.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes will be used (or generated if required).

alphas [ndarray or Series, default: np.logspace(-10, 2, 200)] An array of alphas to fit each model with

cv [int, cross-validation generator or an iterable, optional] Determines the cross-validation splitting strategy. Possible inputs for cv are:

- None, to use the default 3-fold cross validation,
- integer, to specify the number of folds in a *(Stratified)KFold*,
- An object to be used as a cross-validation generator.
- An iterable yielding train, test splits.

This argument is passed to the `sklearn.model_selection.cross_val_score` method to produce the cross validated score for each alpha.

scoring [string, callable or None, optional, default: None] A string (see model evaluation documentation) or a scorer callable object / function with signature `scorer(estimator, X, y)`.

This argument is passed to the `sklearn.model_selection.cross_val_score` method to produce the cross validated score for each alpha.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

This class does not take advantage of estimator-specific searching and is therefore less optimal and more time consuming than the regular "RegressorCV" estimators.

Examples

```
>>> from yellowbrick.regressor import ManualAlphaSelection
>>> from sklearn.linear_model import Ridge
>>> model = ManualAlphaSelection(
...     Ridge(), cv=12, scoring='neg_mean_squared_error'
... )
...
>>> model.fit(X, y)
>>> model.poof()
```

`draw()`

Draws the alphas values against their associated error in a similar fashion to the AlphaSelection visualizer.

`fit(`*X*, *y*, ***args*`)`

The fit method is the primary entry point for the manual alpha selection visualizer. It sets the alpha param for each alpha in the alphas list on the wrapped estimator, then scores the model using the passed in X and y data set. Those scores are then aggregated and drawn using matplotlib.

4.3.5 Classification Visualizers

Classification models attempt to predict a target in a discrete space, that is assign an instance of dependent variables one or more categories. Classification score visualizers display the differences between classes as well as a number of classifier-specific visual evaluations. We currently have implemented four classifier evaluations:

- *Classification Report*: Presents the classification report of the classifier as a heatmap
- *Confusion Matrix*: Presents the confusion matrix of the classifier as a heatmap
- *ROCAUC*: Presents the graph of receiver operating characteristics along with area under the curve
- *Class Balance*: Displays the difference between the class balances and support
- *Threshold*: Shows the bounds of precision, recall and queue rate after a number of trials.

Estimator score visualizers wrap Scikit-Learn estimators and expose the Estimator API such that they have fit(), predict(), and score() methods that call the appropriate estimator methods under the hood. Score visualizers can wrap an estimator and be passed in as the final step in a Pipeline or VisualPipeline.

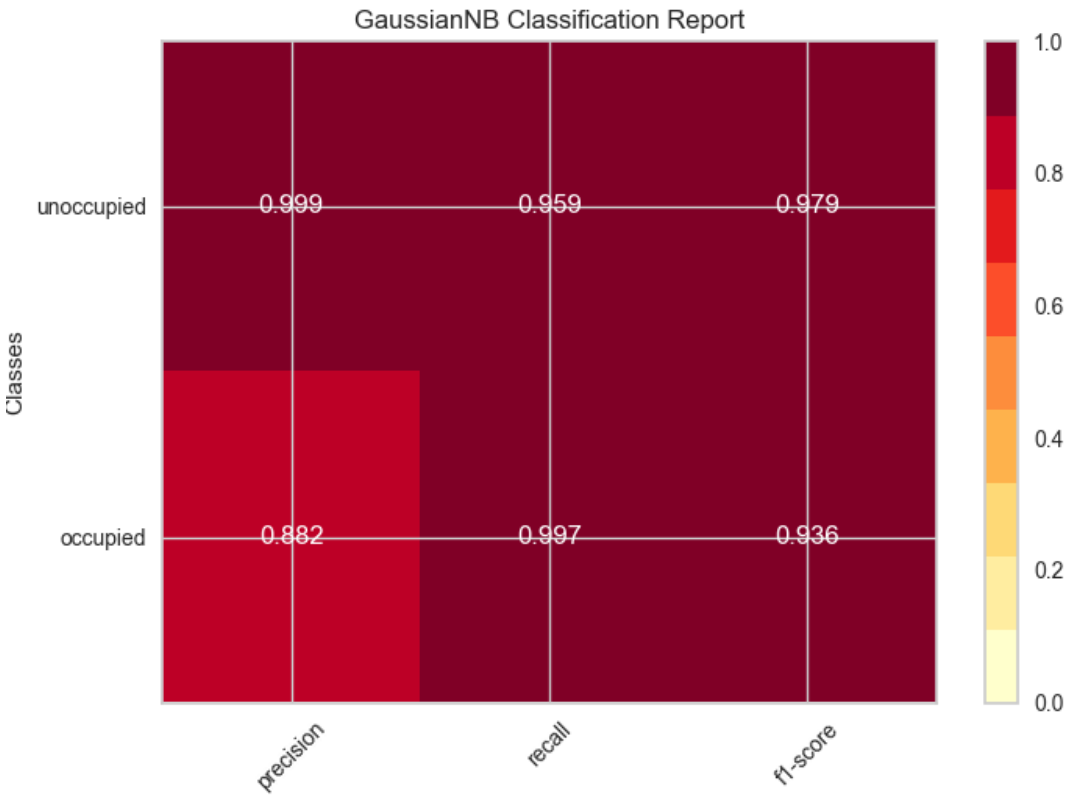
```
# Classifier Evaluation Imports

from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

from yellowbrick.classifier import ClassificationReport, ROCAUC, ClassBalance,␣
↪ThresholdViz
```

Classification Report

The classification report visualizer displays the precision, recall, and F1 scores for the model. In order to support easier interpretation and problem detection, the report integrates numerical scores with a color-coded heatmap.

```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']

# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.occupancy.as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Instantiate the classification model and visualizer
bayes = GaussianNB()
visualizer = ClassificationReport(bayes, classes=classes)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
visualizer.score(X_test, y_test)  # Evaluate the model on the test data
g = visualizer.poof()             # Draw/show/poof the data
```



API Reference

Visual classification report for classifier scoring.

class yellowbrick.classifier.classification_report.**ClassificationReport**(*model*, *ax=None*, *classes=None*, ***kwargs*)

基类：yellowbrick.classifier.base.ClassificationScoreVisualizer

Classification report that shows the precision, recall, and F1 scores for the model. Integrates numerical scores as well as a color-coded heatmap.

Parameters

ax [The axis to plot the figure on.]

model [the Scikit-Learn estimator] Should be an instance of a classifier, else the ___init___ will return an error.

classes [a list of class names for the legend] If classes is None and a y value is passed to fit then the classes are selected from the target vector.

colormap [optional string or matplotlib cmap to colorize lines] Use sequential heatmap.

kwargs [keyword arguments passed to the super class.]

Examples

```
>>> from yellowbrick.classifier import ClassificationReport
>>> from sklearn.linear_model import LogisticRegression
>>> viz = ClassificationReport(LogisticRegression())
>>> viz.fit(X_train, y_train)
>>> viz.score(X_test, y_test)
>>> viz.poof()
```

`draw`(*y*, *y_pred*)

Renders the classification report across each axis.

Parameters

y [ndarray or Series of length n] An array or series of target or class values

y_pred [ndarray or Series of length n] An array or series of predicted target values

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`score`(*X*, *y=None*, ***kwargs*)

Generates the Scikit-Learn classification_report

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

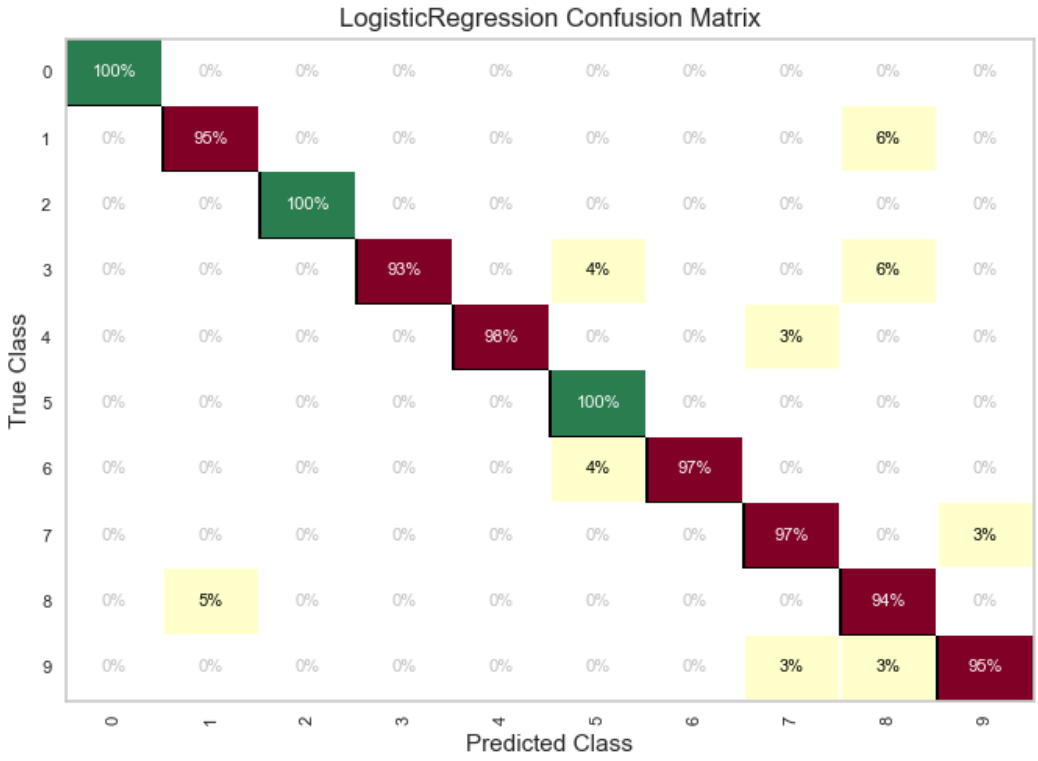
Confusion Matrix

The `ConfusionMatrix` visualizer is a ScoreVisualizer that takes a fitted Scikit-Learn classifier and a set of test X and y values and returns a report showing how each of the test values predicted classes compare to their actual classes. Data scientists use confusion matrices to understand which classes are most easily confused. These provide similar information as what is available in a ClassificationReport, but rather than top-level scores they provide deeper insight into the classification of individual data points.

Below are a few examples of using the ConfusionMatrix visualizer; more information can be found by looking at the Scikit-Learn documentation on confusion matrices.

```
#First do our imports
import yellowbrick

from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

from yellowbrick.classifier import ConfusionMatrix
```

```
# We'll use the handwritten digits data set from scikit-learn.
# Each feature of this dataset is an 8x8 pixel image of a handwritten number.
# Digits.data converts these 64 pixels into a single array of features
digits = load_digits()
X = digits.data
y = digits.target

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size =0.2, random_state=11)

model = LogisticRegression()

#The ConfusionMatrix visualizer taxes a model
cm = ConfusionMatrix(model, classes=[0,1,2,3,4,5,6,7,8,9])

#Fit fits the passed model. This is unnecessary if you pass the visualizer a pre-fitted␣
↪model
cm.fit(X_train, y_train)

#To create the ConfusionMatrix, we need some test data. Score runs predict() on the data
#and then creates the confusion_matrix from scikit learn.
cm.score(X_test, y_test)

#How did we do?
cm.poof()
```



API Reference

Visual confusion matrix for classifier scoring.

class yellowbrick.classifier.confusion_matrix.**ConfusionMatrix**(*model*, *ax=None*, *classes=None*, *label_encoder=None*, ***kwargs*)

基类：yellowbrick.classifier.base.ClassificationScoreVisualizer

Creates a heatmap visualization of the sklearn.metrics.confusion_matrix(). A confusion matrix shows each combination of the true and predicted classes for a test data set.

The default color map uses a yellow/orange/red color scale. The user can choose between displaying values as the percent of true (cell value divided by sum of row) or as direct counts. If percent of true mode is selected, 100% accurate predictions are highlighted in green.

Requires a classification model

Parameters

model [the Scikit-Learn estimator] Should be an instance of a classifier or ___init___ will return an error.

ax [the matplotlib axis to plot the figure on (if None, a new axis will be created)]

classes [list, default: None] a list of class names to use in the confusion_matrix. This is passed to the 'labels' parameter of sklearn.metrics.confusion_matrix(), and follows the behaviour indicated by that function. It may be used to reorder or select a subset of labels. If None, values that appear at least once in y_true or y_pred are used in sorted order.

label_encoder [dict or LabelEncoder, default: None] When specifying the `classes` argument, the input to `fit()` and `score()` must match the expected labels. If the `X` and `y` datasets have been encoded prior to training and the labels must be preserved for the visualization, use this argument to provide a mapping from the encoded class to the correct label. Because typically a Scikit-Learn `LabelEncoder` is used to perform this operation, you may provide it directly to the class to utilize its fitted encoding.

Examples

```
>>> from yellowbrick.classifier import ConfusionMatrix
>>> from sklearn.linear_model import LogisticRegression
>>> viz = ConfusionMatrix(LogisticRegression())
>>> viz.fit(X_train, y_train)
>>> viz.score(X_test, y_test)
>>> viz.poof()
```

`draw`(*percent=True*)

Renders the classification report Should only be called internally, as it uses values calculated in Score and score calls this method.

Parameters

percent: Boolean Whether the heatmap should represent "% of True" or raw counts

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps.

Parameters

kwargs: dict generic keyword arguments.

Notes

The user calls poof and poof calls finalize. Developers should implement visualizer-specific finalization methods like setting titles or axes labels, etc.

score(*X*, *y*, *sample_weight=None*, *percent=True*)

Generates the Scikit-Learn confusion_matrix and applies this to the appropriate axis

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

sample_weight: optional, passed to the confusion_matrix

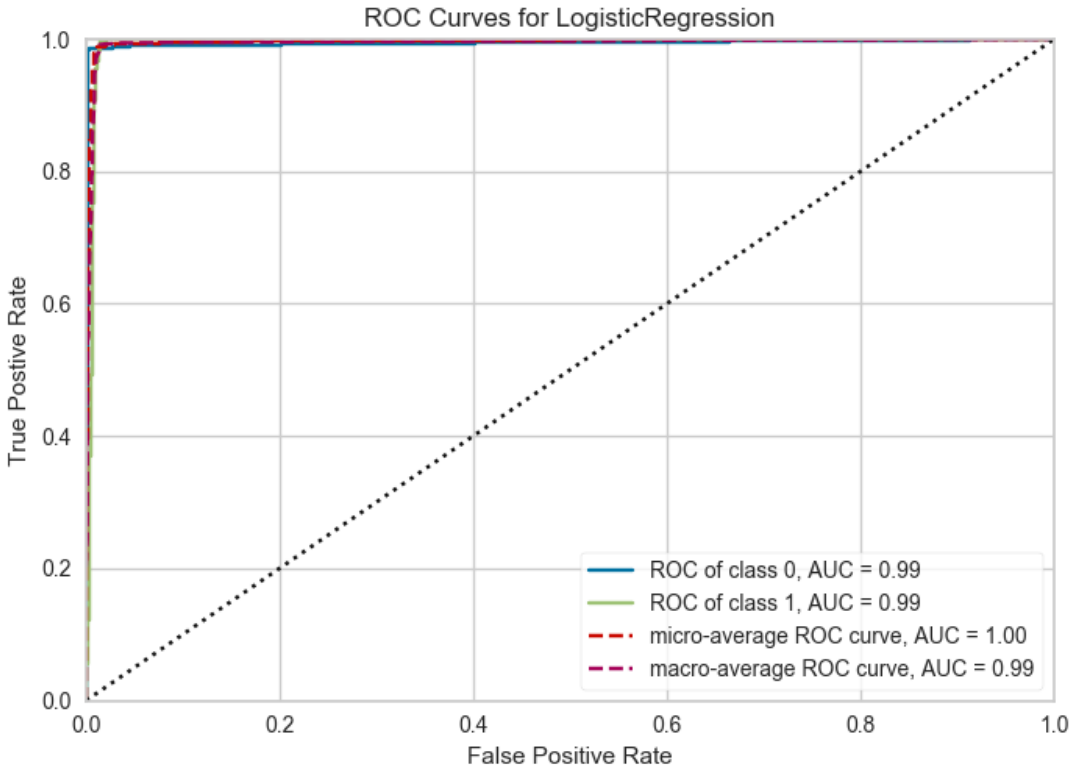
percent: optional, Boolean. Determines whether or not the confusion_matrix should be displayed as raw numbers or as a percent of the true predictions. Note, if using a subset of classes in ___init___, percent should be set to False or inaccurate percents will be displayed.

ROCAUC

A ROCAUC (Receiver Operating Characteristic/Area Under the Curve) plot allows the user to visualize the tradeoff between the classifier's sensitivity and specificity.

```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']

# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.occupancy.as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Instantiate the classification model and visualizer
logistic = LogisticRegression()
visualizer = ROCAUC(logistic)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
visualizer.score(X_test, y_test)  # Evaluate the model on the test data
g = visualizer.poof()             # Draw/show/poof the data
```



API Reference

Implements visual ROC/AUC curves for classification evaluation.

class yellowbrick.classifier.rocauc.**ROCAUC**(*model*, *ax=None*, *classes=None*, *micro=True*, *macro=True*, *per_class=True*, ***kwargs*)

基类: yellowbrick.classifier.base.ClassificationScoreVisualizer

Receiver Operating Characteristic (ROC) curves are a measure of a classifier's predictive quality that compares and visualizes the tradeoff between the models' sensitivity and specificity. The ROC curve displays the true positive rate on the Y axis and the false positive rate on the X axis on both a global average and per-class basis. The ideal point is therefore the top-left corner of the plot: false positives are zero and true positives are one.

This leads to another metric, area under the curve (AUC), a computation of the relationship between false positives and true positives. The higher the AUC, the better the model generally is. However, it is also important to inspect the "steepness" of the curve, as this describes the maximization of the true positive rate while minimizing the false positive rate. Generalizing "steepness" usually leads to discussions about convexity, which we do not get into here.

Parameters

ax [the axis to plot the figure on.]

model [the Scikit-Learn estimator] Should be an instance of a classifier, else the ___init___ will return an error.

classes [list] A list of class names for the legend. If classes is None and a y value is passed to fit then the classes are selected from the target vector. Note that the curves must be computed based on what is in the target vector passed to the `score()` method. Class names are used for labeling only and must be in the correct order to prevent confusion.

micro [bool, default = True] Plot the micro-averages ROC curve, computed from the sum of all true positives and false positives across all classes.

macro [bool, default = True] Plot the macro-averages ROC curve, which simply takes the average of curves across all classes.

per_class [bool, default = True] Plot the ROC curves for each individual class. Primarily this is set to false if only the macro or micro average curves are required.

kwargs [keyword arguments passed to the super class.] Currently passing in hard-coded colors for the Receiver Operating Characteristic curve and the diagonal. These will be refactored to a default Yellowbrick style.

Notes

ROC curves are typically used in binary classification, and in fact the Scikit-Learn `roc_curve` metric is only able to perform metrics for binary classifiers. As a result it is necessary to binarize the output or to use one-vs-rest or one-vs-all strategies of classification. The visualizer does its best to handle multiple situations, but exceptions can arise from unexpected models or outputs.

Another important point is the relationship of class labels specified on initialization to those drawn on the curves. The classes are not used to constrain ordering or filter curves; the ROC computation happens on the unique values specified in the target vector to the `score` method. To ensure the best quality visualization, do not use a LabelEncoder for this and do not pass in class labels.

参见:

http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

Examples

```
>>> from sklearn.datasets import load_breast_cancer
>>> from yellowbrick.classifier import ROCAUC
>>> from sklearn.linear_model import LogisticRegression
>>> from sklearn.model_selection import train_test_split
>>> data = load_breast_cancer()
```

(下页继续)

(续上页)

```
>>> X = data['data']
>>> y = data['target']
>>> X_train, X_test, y_train, y_test = train_test_split(X, y)
>>> viz = ROCAUC(LogisticRegression())
>>> viz.fit(X_train, y_train)
>>> viz.score(X_test, y_test)
>>> viz.poof()
```

draw()

Renders ROC-AUC plot. Called internally by score, possibly more than once

Returns

ax [the axis with the plotted figure]

finalize(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

score(*X*, *y=None*, ***kwargs*)

Generates the predicted target values using the Scikit-Learn estimator.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

Returns

score [float] The micro-average area under the curve of all classes.

Class Balance

Oftentimes classifiers perform badly because of a class imbalance. A class balance chart can help prepare the user for such a case by showing the support for each class in the fitted classification model.

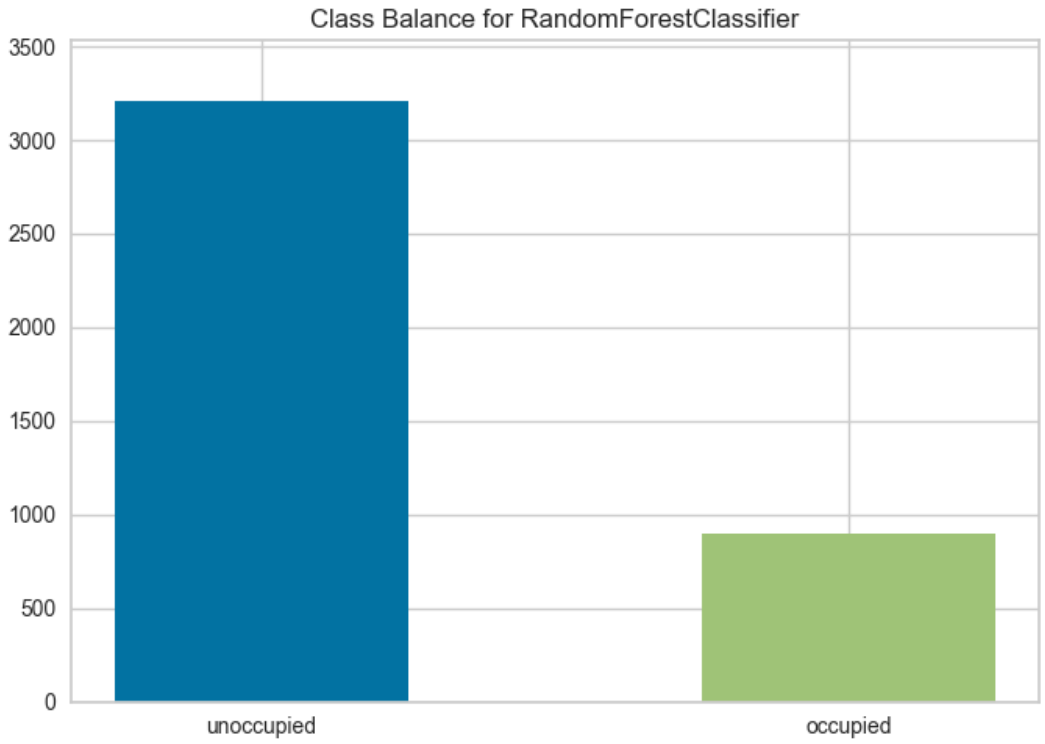
```
# Load the classification data set
data = load_data('occupancy')

# Specify the features of interest and the classes of the target
features = ["temperature", "relative humidity", "light", "CO2", "humidity"]
classes = ['unoccupied', 'occupied']
```

(下页继续)

(续上页)

```
# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.occupancy.as_matrix()

# Create the train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```
# Instantiate the classification model and visualizer
forest = RandomForestClassifier()
visualizer = ClassBalance(forest, classes=classes)

visualizer.fit(X_train, y_train)  # Fit the training data to the visualizer
visualizer.score(X_test, y_test)  # Evaluate the model on the test data
g = visualizer.poof()             # Draw/show/poof the data
```



API Reference

Class balance visualizer for showing per-class support.

`class yellowbrick.classifier.class_balance.ClassBalance`(*model*, *ax=None*, *classes=None*, ***kwargs*)

基类：`yellowbrick.classifier.base.ClassificationScoreVisualizer`

Class balance chart that shows the support for each class in the fitted classification model displayed as a bar plot. It is initialized with a fitted model and generates a class balance chart on draw.

Parameters

ax: axes the axis to plot the figure on.

model: estimator Scikit-Learn estimator object. Should be an instance of a classifier, else `__init__()` will raise an exception.

classes: list A list of class names for the legend. If classes is None and a y value is passed to fit then the classes are selected from the target vector.

kwargs: dict Keyword arguments passed to the super class. Here, used to colorize the bars in the histogram.

Notes

These parameters can be influenced later on in the visualization process, but can and should be set as early as possible.

`draw()`

Renders the class balance chart across the axis.

Returns

ax [the axis with the plotted figure]

`finalize`(***kwargs*)

Finalize executes any subclass-specific axes finalization steps. The user calls poof and poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`score`(*X*, *y=None*, ***kwargs*)

Generates the Scikit-Learn precision_recall_fscore_support

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features

y [ndarray or Series of length n] An array or series of target or class values

Returns

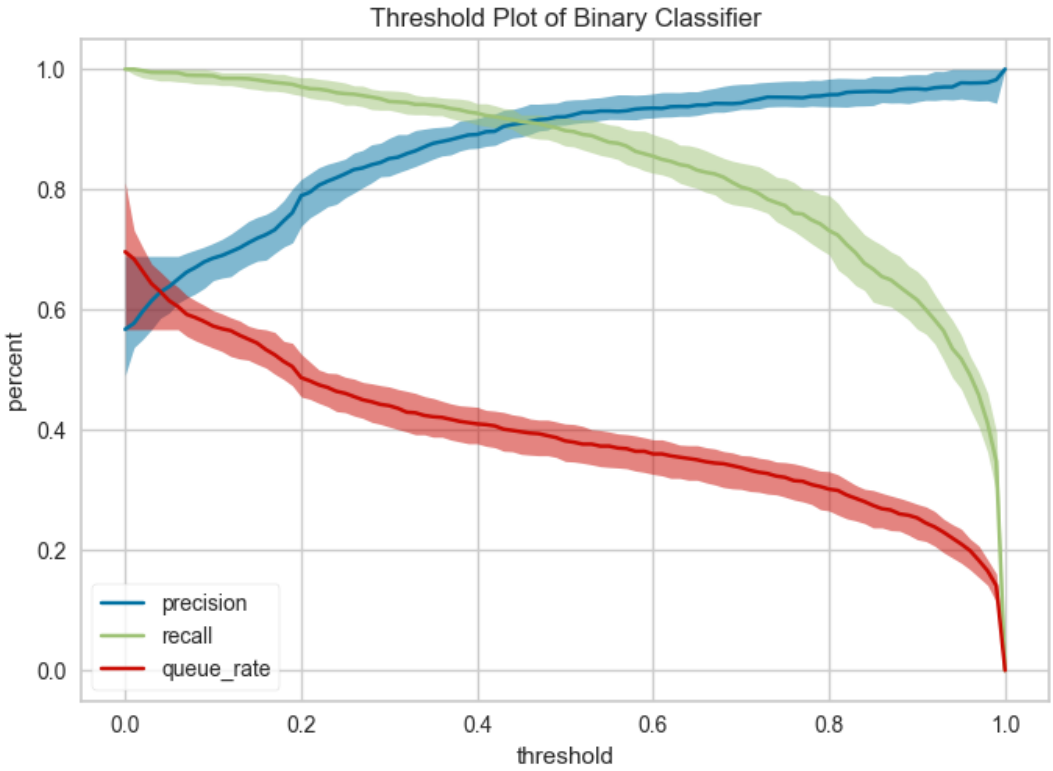
ax [the axis with the plotted figure]

Threshold

The Threshold visualizer shows the bounds of precision, recall and queue rate for different thresholds for binary targets after a given number of trials.

```
# Load the data set
data = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/
↪spambase.data', header=None)
data.rename(columns={57:'is_spam'}, inplace=True)

features = [col for col in data.columns if col != 'is_spam']

# Extract the numpy arrays from the data frame
X = data[features].as_matrix()
y = data.is_spam.as_matrix()
```

```
# Instantiate the classification model and visualizer
logistic = LogisticRegression()
visualizer = ThreshViz(logistic)

visualizer.fit(X, y)  # Fit the training data to the visualizer
g = visualizer.poof() # Draw/show/poof the data
```



API Reference

`yellowbrick.classifier.threshold.ThreshViz`
`yellowbrick.classifier.threshold.ThresholdVisualizer` 的别名

4.3.6 Clustering Visualizers

Clustering models are unsupervised methods that attempt to detect patterns in unlabeled data. There are two primary classes of clustering algorithm: *agglomerative* clustering links similar data points together, whereas *centroidal* clustering attempts to find centers or partitions in the data. Yellowbrick provides the *yellowbrick.cluster* module to visualize and evaluate clustering behavior. Currently we provide two visualizers to evaluate *centroidal* mechanisms, particularly K-Means clustering, that help us to discover an optimal K parameter in the clustering metric:

- *Elbow Method*: visualize the clusters according to some scoring function, look for an "elbow" in the curve.
- *Silhouette Visualizer*: visualize the silhouette scores of each cluster in a single model.

Because it is very difficult to *score* a clustering model, Yellowbrick visualizers wrap Scikit-Learn "clusterer" estimators via their *fit()* method. Once the clustering model is trained, then the visualizer can call *poof()* to

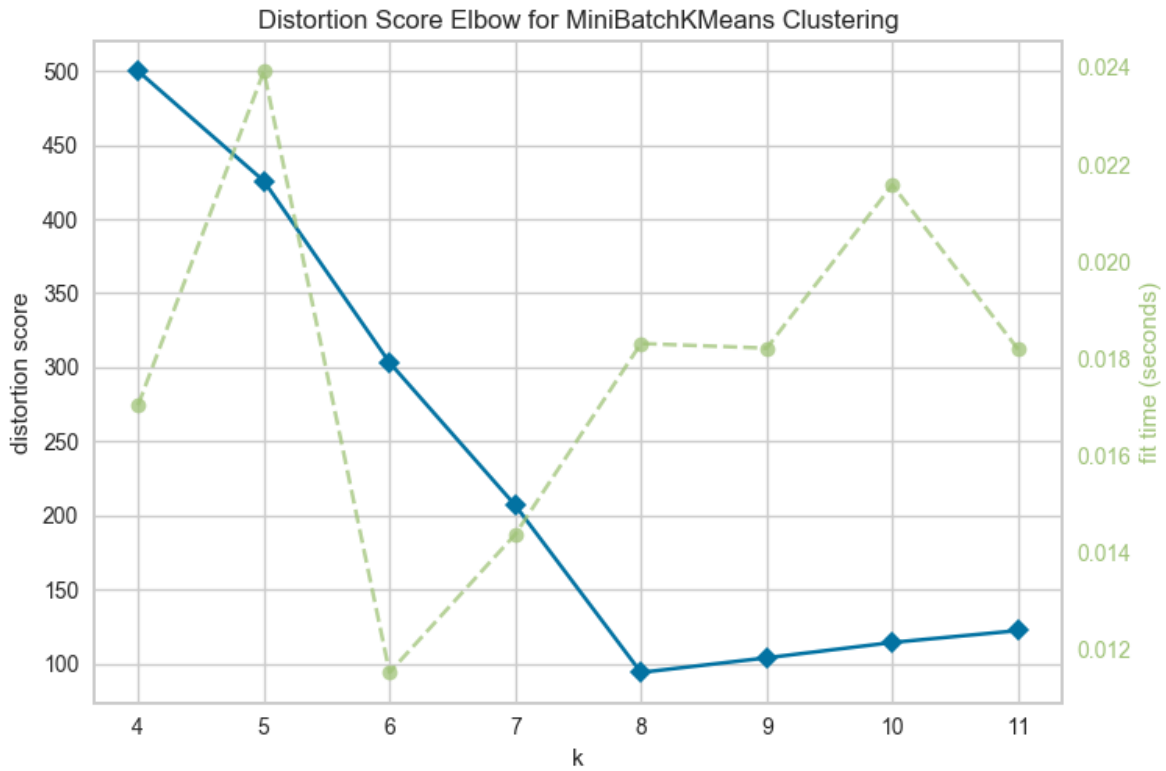
display the clustering evaluation metric.

Elbow Method

The elbow method for K selection visualizes multiple clustering models with different values for K. Model selection is based on whether or not there is an "elbow" in the curve; e.g. if the curve looks like an arm, if there is a clear change in angle from one part of the curve to another.

```
# Make 8 blobs dataset
X, y = make_blobs(centers=8)
```

```
# Instantiate the clustering model and visualizer
visualizer = KElbowVisualizer(MiniBatchKMeans(), k=(4,12))

visualizer.fit(X) # Fit the training data to the visualizer
visualizer.poof() # Draw/show/poof the data
```



API Reference

Implements the elbow method for determining the optimal number of clusters. https://bl.ocks.org/rpgove/0060ff3b656618e9136b

`class yellowbrick.cluster.elbow.KElbowVisualizer`(*model*, *ax=None*, *k=10*, *metric='distortion'*, *timings=True*, ***kwargs*)

基类: `yellowbrick.cluster.base.ClusteringScoreVisualizer`

The K-Elbow Visualizer implements the "elbow" method of selecting the optimal number of clusters for K-means clustering. K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters. Because the user must specify in advance what k to choose, the algorithm is somewhat naive – it assigns all members to k clusters even if that is not the right k for the dataset.

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the `distortion_score` is computed, the sum of square distances from each point to its assigned center. Other metrics can also be used such as the `silhouette_score`, the mean silhouette coefficient for all samples or the `calinski_harabaz_score`, which computes the ratio of dispersion between and within clusters.

When these overall metrics for each model are plotted, it is possible to visually determine the best value for K. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

Parameters

model [a Scikit-Learn clusterer] Should be an instance of a clusterer, specifically `KMeans` or `MiniBatchKMeans`. If it is not a clusterer, an exception is raised.

ax [matplotlib Axes, default: None] The axes to plot the figure on. If None is passed in the current axes will be used (or generated if required).

k [integer or tuple] The range of k to compute silhouette scores for. If a single integer is specified, then will compute the range (2,k) otherwise the specified range in the tuple is used.

metric [string, default: `"distortion"`] Select the scoring metric to evaluate the clusters. The default is the mean distortion, defined by the sum of squared distances between each observation and its closest centroid. Other metrics include:

- **distortion**: mean sum of squared distances to centers
- **silhouette**: mean ratio of intra-cluster and nearest-cluster distance
- **calinski_harabaz**: ratio of within to between cluster dispersion

timings [bool, default: True] Display the fitting time per k to evaluate the amount of time required to train the clustering model.

kwargs [dict] Keyword arguments that are passed to the base class and may influence the visualization as defined in other Visualizers.

Notes

If you get a visualizer that doesn't have an elbow or inflection point, then this method may not be working. The elbow method does not work well if the data is not very clustered; in this case you might see a smooth curve and the value of k is unclear. Other scoring methods such as BIC or SSE also can be used to explore if clustering is a correct choice.

For a discussion on the Elbow method, read more at Robert Gove's Block.

Examples

```
>>> from yellowbrick.cluster import KElbowVisualizer
>>> from sklearn.cluster import KMeans
>>> model = KElbowVisualizer(KMeans(), k=10)
>>> model.fit(X)
>>> model.poof()
```

`draw()`

Draw the elbow curve for the specified scores and values of K.

`finalize()`

Prepare the figure for rendering by setting the title as well as the X and Y axis labels and adding the legend.

`fit(`*X*, *y=None*, ***kwargs*`)`

Fits n KMeans models where n is the length of `self.k_values_`, storing the silhoutte scores in the `self.k_scores_` attribute. This method finishes up by calling draw to create the plot.

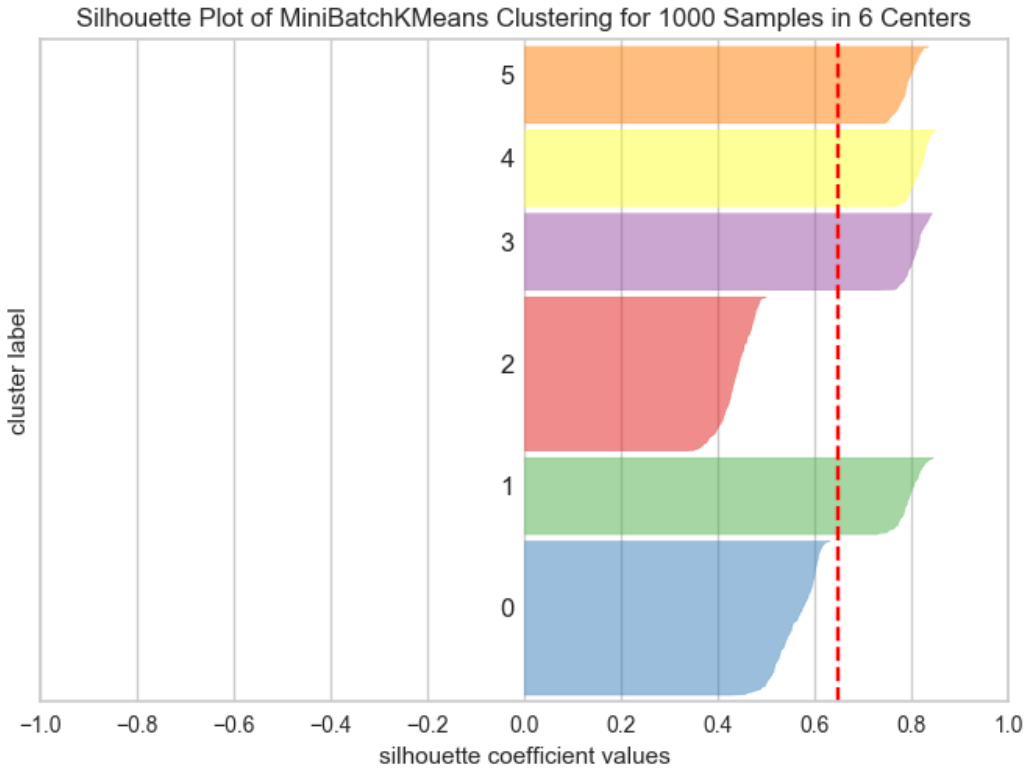
Silhouette Visualizer

The Silhouette Coefficient is used when the ground-truth about the dataset is unknown and computes the density of clusters computed by the model. The score is computed by averaging the silhouette coefficient for each sample, computed as the difference between the average intra-cluster distance and the mean nearest-cluster distance for each sample, normalized by the maximum value. This produces a score between 1 and -1, where 1 is highly dense clusters and -1 is completely incorrect clustering.

The Silhouette Visualizer displays the silhouette coefficient for each sample on a per-cluster basis, visualizing which clusters are dense and which are not. This is particularly useful for determining cluster imbalance, or for selecting a value for K by comparing multiple visualizers.

```
# Make 8 blobs dataset
X, y = make_blobs(centers=8)
```

```
# Instantiate the clustering model and visualizer
model = MiniBatchKMeans(6)
visualizer = SilhouetteVisualizer(model)

visualizer.fit(X) # Fit the training data to the visualizer
visualizer.poof() # Draw/show/poof the data
```



API Reference

Implements visualizers that use the silhouette metric for cluster evaluation.

class yellowbrick.cluster.silhouette.**SilhouetteVisualizer**(*model*, *ax=None*, ***kwargs*)

基类：yellowbrick.cluster.base.ClusteringScoreVisualizer

TODO: Document this class!

draw(*labels*)

Draw the silhouettes for each sample and the average score.

Parameters

labels [array-like] An array with the cluster label for each silhouette sample, usually computed with `predict()`. Labels are not stored on the visualizer so that the figure can be redrawn with new data.

`finalize()`

Prepare the figure for rendering by setting the title and adjusting the limits on the axes, adding labels and a legend.

`fit`(*X*, *y=None*, ***kwargs*)

Fits the model and generates the the silhouette visualization.

TODO: decide to use this method or the score method to draw. NOTE: Probably this would be better in score, but the standard score is a little different and I'm not sure how it's used.

4.3.7 Text Modeling Visualizers

Yellowbrick provides the *yellowbrick.text* module for text-specific visualizers. The *TextVisualizer* class specifically deals with datasets that are corpora and not simple numeric arrays or DataFrames, providing utilities for analyzing word distribution, showing document similarity, or simply wrapping some of the other standard visualizers with text-specific display properties.

We currently have two text-specific visualizations implemented:

- *Token Frequency Distribution*: plot the frequency of tokens in a corpus
- *t-SNE Corpus Visualization*: plot similar documents closer together to discover clusters

Note that the examples in this section require a corpus of text data, see *loading a text corpus* for more information.

```
from yellowbrick.text import FreqDistVisualizer
from yellowbrick.text import TSNEVisualizer

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
```

Loading a Text Corpus

As in the previous sections, Yellowbrick has provided a sample dataset to run the following cells. In particular, we are going to use a text corpus wrangled from the Baleen RSS Corpus to present the following examples. If you haven't already downloaded the data, you can do so by running:

```
$ python -m yellowbrick.download
```

Note that this will create a directory called `data` in your current working directory that contains subdirectories with the provided datasets.

注解: If you've already followed the instructions from *downloading example datasets*, you don't have to repeat these steps here. Simply check to ensure there is a directory called `hobbies` in your data directory.

The following code snippet creates a utility that will load the corpus from disk into a Scikit-Learn Bunch object. This method creates a corpus that is exactly the same as the one found in the "working with text data" example on the Scikit-Learn website, hopefully making the examples easier to use.

```
import os
from sklearn.datasets.base import Bunch

def load_corpus(path):
    """
    Loads and wrangles the passed in text corpus by path.
    """

    # Check if the data exists, otherwise download or raise
    if not os.path.exists(path):
        raise ValueError((
            "'{}' dataset has not been downloaded, "
            "use the yellowbrick.download module to fetch datasets"
        ).format(path))

    # Read the directories in the directory as the categories.
    categories = [
        cat for cat in os.listdir(path)
        if os.path.isdir(os.path.join(path, cat))
    ]

    files  = [] # holds the file names relative to the root
    data   = [] # holds the text read from the file
    target = [] # holds the string of the category

    # Load the data from the files in the corpus
    for cat in categories:
        for name in os.listdir(os.path.join(path, cat)):
            files.append(os.path.join(path, cat, name))
            target.append(cat)
```

(下页继续)

(续上页)

```
            with open(os.path.join(path, cat, name), 'r') as f:
                data.append(f.read())


    # Return the data bunch for use similar to the newsgroups example
    return Bunch(
        categories=categories,
        files=files,
        data=data,
        target=target,
    )
```

This is a fairly long ibt of code, so let's walk through it step by step. The data in the corpus directory is stored as follows:

```
data/hobbies
  README.md
  books
|     56d62a53c1808113ffb87f1f.txt
|     5745a9c7c180810be6efd70b.txt
  cinema
|     56d629b5c1808113ffb87d8f.txt
|     57408e5fc180810be6e574c8.txt
  cooking
|     56d62b25c1808113ffb8813b.txt
|     573f0728c180810be6e2575c.txt
  gaming
|     56d62654c1808113ffb87938.txt
|     574585d7c180810be6ef7ffc.txt
  sports
      56d62adec1808113ffb88054.txt
      56d70f17c180810560aec345.txt
```

Each of the documents in the corpus is stored in a text file labeled with its hash signature in a directory that specifies its label or category. Therefore the first step after checking to make sure the specified path exists is to list all the directories in the `hobbies` directory – this gives us each of our categories, which we will store later in the bunch.

The second step is to create placeholders for holding filenames, text data, and labels. We can then loop through the list of categories, list the files in each category directory, add those files to the files list, add the category name to the target list, then open and read the file to add it to data.

To load the corpus into memory, we can simply use the following snippet:

```
corpus = load_corpus("data/hobbies")
```

We'll use this snippet in all of the text examples in this section!

Token Frequency Distribution

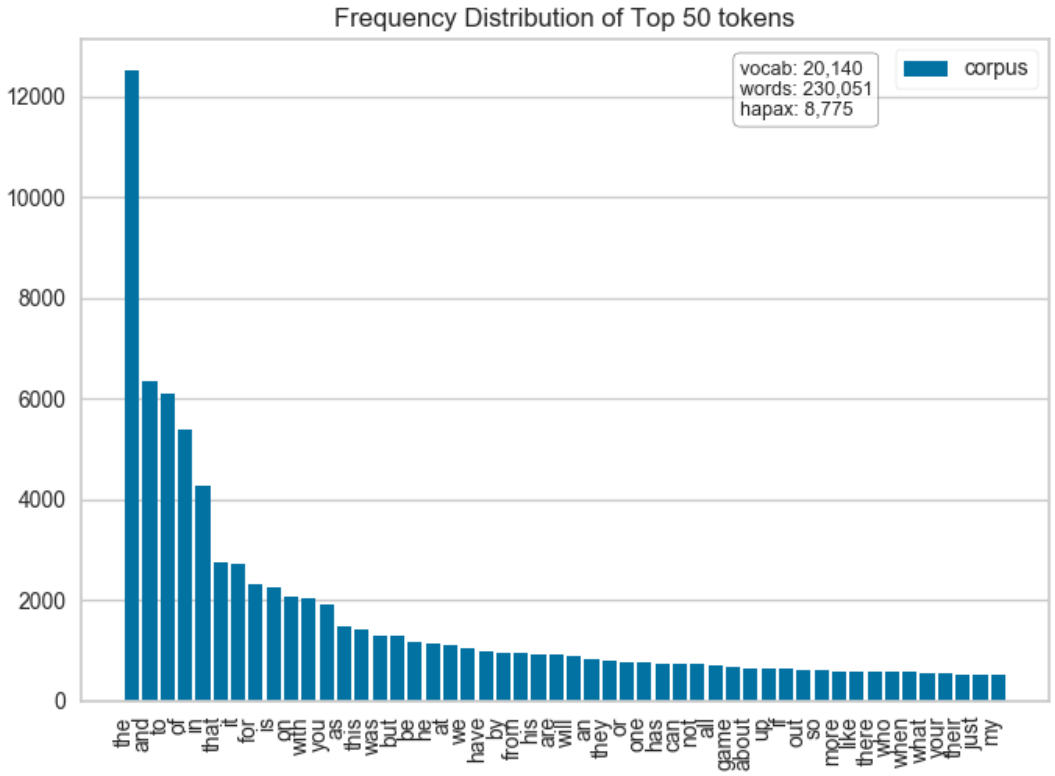
A method for visualizing the frequency of tokens within and across corpora is frequency distribution. A frequency distribution tells us the frequency of each vocabulary item in the text. In general, it could count any kind of observable event. It is a distribution because it tells us how the total number of word tokens in the text are distributed across the vocabulary items.

```
from yellowbrick.text.freqdist import FreqDistVisualizer
from sklearn.feature_extraction.text import CountVectorizer
```

Note that the `FreqDistVisualizer` does not perform any normalization or vectorization, and it expects text that has already be count vectorized.

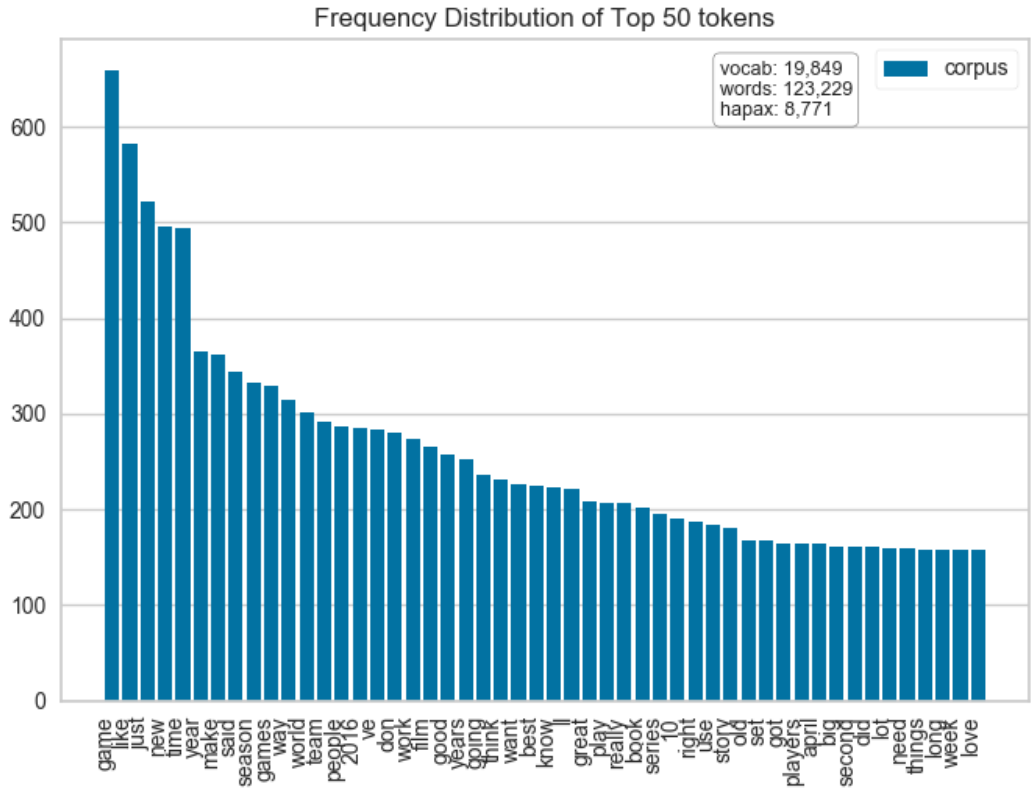
We first instantiate a `FreqDistVisualizer` object, and then call `fit()` on that object with the count vectorized documents and the features (i.e. the words from the corpus), which computes the frequency distribution. The visualizer then plots a bar chart of the top 50 most frequent terms in the corpus, with the terms listed along the x-axis and frequency counts depicted at y-axis values. As with other Yellowbrick visualizers, when the user invokes `poof()`, the finalized visualization is shown.

```
vectorizer = CountVectorizer()
docs       = vectorizer.fit_transform(corpus.data)
features   = vectorizer.get_feature_names()

visualizer = FreqDistVisualizer(features=features)
visualizer.fit(docs)
visualizer.poof()
```



It is interesting to compare the results of the `FreqDistVisualizer` before and after stopwords have been removed from the corpus:

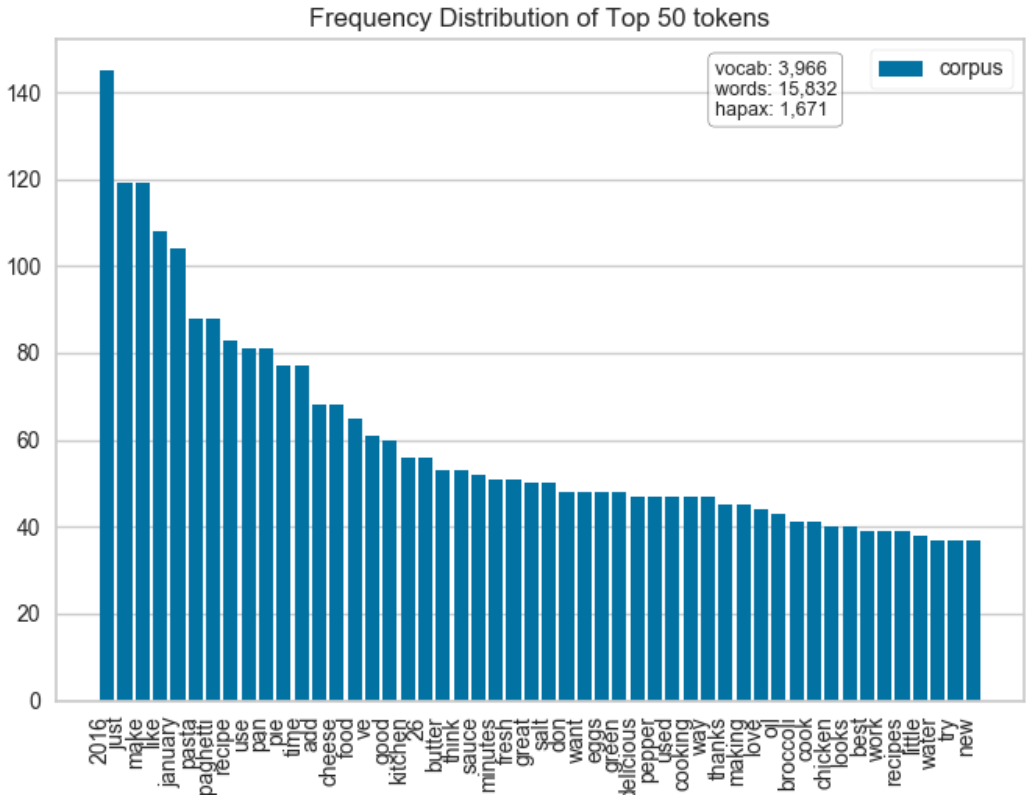
```
vectorizer = CountVectorizer(stop_words='english')
docs       = vectorizer.fit_transform(corpus.data)
features   = vectorizer.get_feature_names()

visualizer = FreqDistVisualizer(features=features)
visualizer.fit(docs)
visualizer.poof()
```



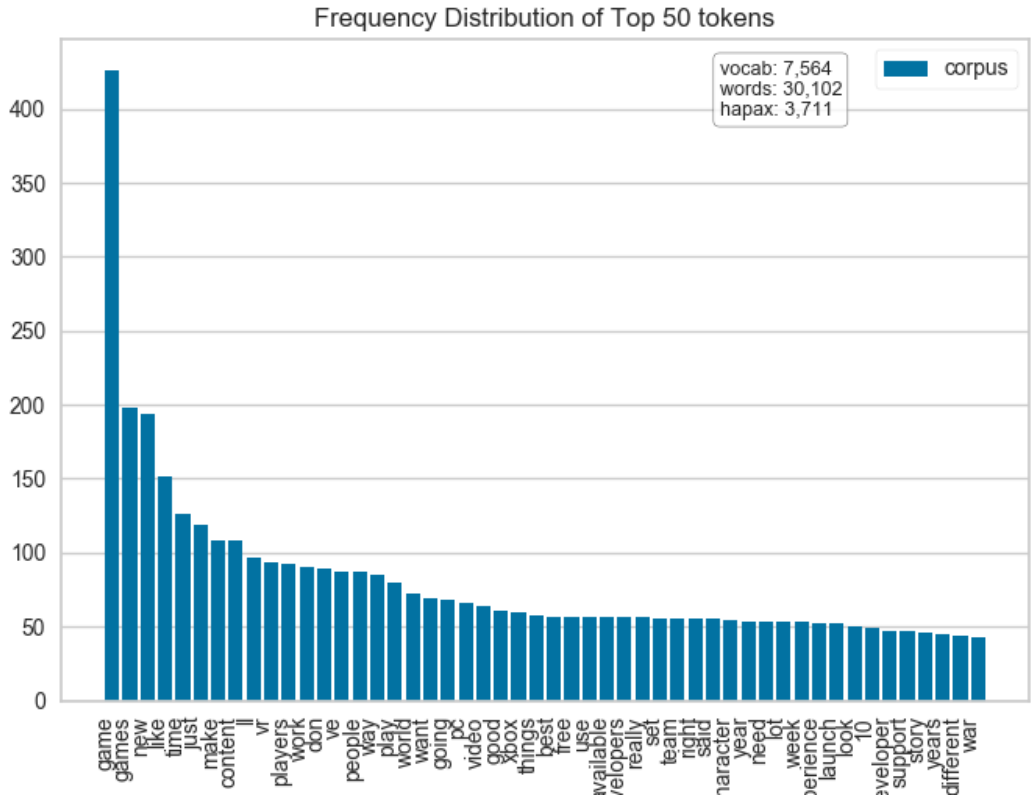
It is also interesting to explore the differences in tokens across a corpus. The hobbies corpus that comes with Yellowbrick has already been categorized (try `corpus['categories']`), so let's visually compare the differences in the frequency distributions for two of the categories: *"cooking"* and *"gaming"*.

```
from collections import defaultdict

hobbies = defaultdict(list)
for text, label in zip(corpus.data, corpus.label):
    hobbies[label].append(text)
```

```
vectorizer = CountVectorizer(stop_words='english')
docs       = vectorizer.fit_transform(text for text in hobbies['cooking'])
features   = vectorizer.get_feature_names()

visualizer = FreqDistVisualizer(features=features)
visualizer.fit(docs)
visualizer.poof()
```




```
vectorizer = CountVectorizer(stop_words='english')
docs       = vectorizer.fit_transform(text for text in hobbies['gaming'])
features   = vectorizer.get_feature_names()

visualizer = FreqDistVisualizer(features=features)
visualizer.fit(docs)
visualizer.poof()
```



API Reference

Implementations of frequency distributions for text visualization

`class yellowbrick.text.freqdist.FrequencyVisualizer`(*features*, *ax=None*, *n=50*, *orient='h'*, *color=None*, ***kwargs*)

基类: `yellowbrick.text.base.TextVisualizer`

A frequency distribution tells us the frequency of each vocabulary item in the text. In general, it could count any kind of observable event. It is a distribution because it tells us how the total number of word tokens in the text are distributed across the vocabulary items.

Parameters

features [list, default: None] The list of feature names from the vectorizer, ordered by index. E.g. a lexicon that specifies the unique vocabulary of the corpus. This can be typically fetched using the `get_feature_names()` method of the transformer in Scikit-Learn.

ax [matplotlib axes, default: None] The axes to plot the figure on.

n: integer, default: 50 Top N tokens to be plotted.

orient ['h' or 'v', default: 'h'] Specifies a horizontal or vertical bar chart.

color [list or tuple of colors] Specify color for bars

kwargs [dict] Pass any additional keyword arguments to the super class.

These parameters can be influenced later on in the visualization

process, but can and should be set as early as possible.

`count`(*X*)

Called from the fit method, this method gets all the words from the corpus and their corresponding frequency counts.

Parameters

X [ndarray or masked ndarray] Pass in the matrix of vectorized documents, can be masked in order to sum the word frequencies for only a subset of documents.

Returns

counts [array] A vector containing the counts of all words in X (columns)

`draw`(***kwargs*)

Called from the fit method, this method creates the canvas and draws the distribution plot on it.

Parameters

kwargs: generic keyword arguments.

`finalize`(***kwargs*)

The finalize method executes any subclass-specific axes finalization steps. The user calls poof & poof calls finalize.

Parameters

kwargs: generic keyword arguments.

`fit`(*X*, *y=None*)

The fit method is the primary drawing input for the frequency distribution visualization. It requires vectorized lists of documents and a list of features, which are the actual words from the original corpus (needed to label the x-axis ticks).

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features representing the corpus of frequency vectorized documents.

y [ndarray or DataFrame of shape n] Labels for the documents for conditional frequency distribution.

.. note:: Text documents must be vectorized before ‘‘fit()‘‘.

t-SNE Corpus Visualization

One very popular method for visualizing document similarity is to use t-distributed stochastic neighbor embedding, t-SNE. Scikit-Learn implements this decomposition method as the `sklearn.manifold.TSNE` transformer. By decomposing high-dimensional document vectors into 2 dimensions using probability distributions from both the original dimensionality and the decomposed dimensionality, t-SNE is able to effectively cluster similar documents. By decomposing to 2 or 3 dimensions, the documents can be visualized with a scatter plot.

Unfortunately, `TSNE` is very expensive, so typically a simpler decomposition method such as SVD or PCA is applied ahead of time. The `TSNEVisualizer` creates an inner transformer pipeline that applies such a decomposition first (SVD with 50 components by default), then performs the t-SNE embedding. The visualizer then plots the scatter plot, coloring by cluster or by class, or neither if a structural analysis is required.

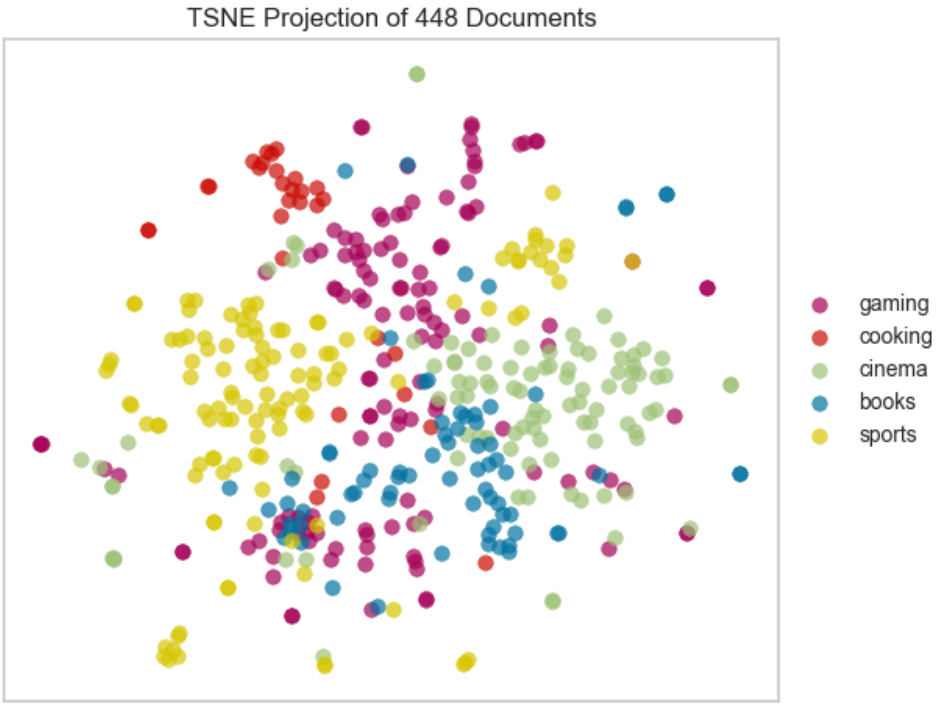
```
from yellowbrick.text import TSNEVisualizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

After importing the required tools, we can *load the corpus* and vectorize the text using TF-IDF.

```
# Load the data and create document vectors
corpus = load_corpus('hobbies')
tfidf  = TfidfVectorizer()

docs   = tfidf.fit_transform(corpus.data)
labels = corpus.target
```

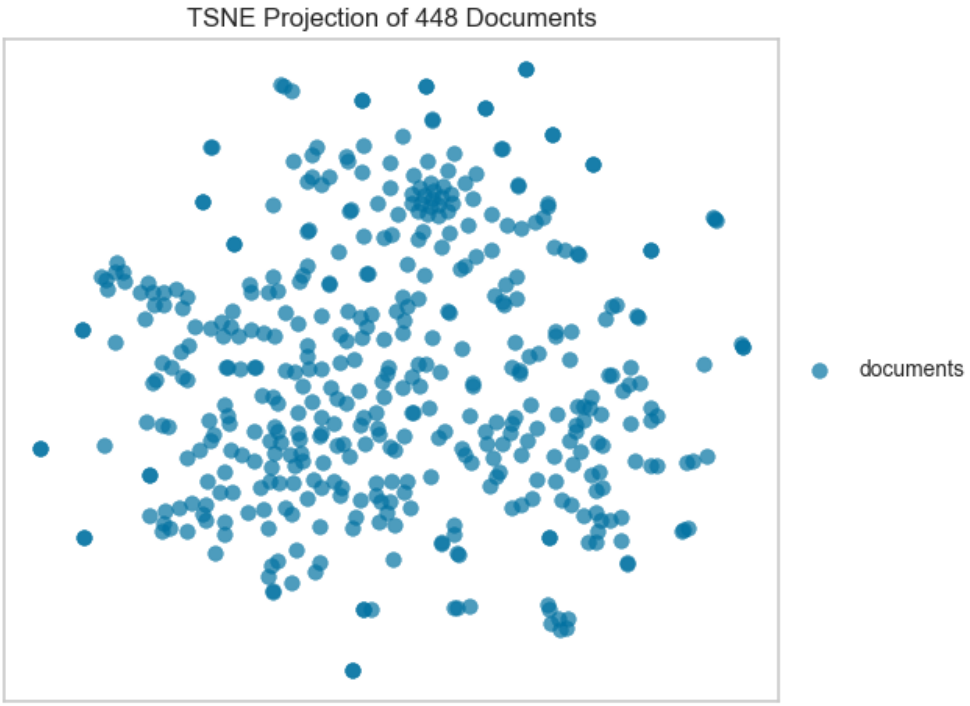
Now that the corpus is vectorized we can visualize it, showing the distribution of classes.

```
# Create the visualizer and draw the vectors
tsne = TSNEVisualizer()
tsne.fit(docs, labels)
tsne.poof()
```



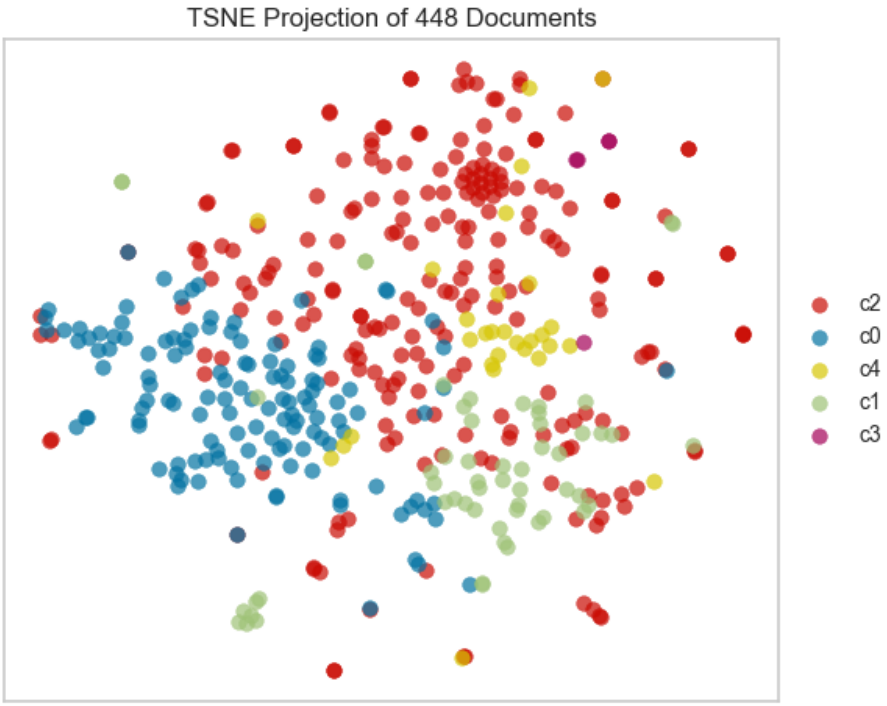
If we omit the target during fit, we can visualize the whole dataset to see if any meaningful patterns are observed.

```
# Don't color points with their classes
tsne = TSNEVisualizer(labels=["documents"])
tsne.fit(docs)
tsne.poof()
```



This means we don't have to use class labels at all, instead we can use cluster membership from K-Means to label each document, looking for clusters of related text by their contents:

```
# Apply clustering instead of class names.
from sklearn.cluster import KMeans

clusters = KMeans(n_clusters=5)
clusters.fit(docs)

tsne = TSNEVisualizer()
tsne.fit(docs, ["c{}".format(c) for c in clusters.labels_])
tsne.poof()
```



API Reference

Implements TSNE visualizations of documents in 2D space.

`class yellowbrick.text.tsne.TSNEVisualizer`(*ax=None*, *decompose='svd'*, *decompose_by=50*, *labels=None*, *classes=None*, *colors=None*, *colormap=None*, *random_state=None*, ***kwargs*)

基类：`yellowbrick.text.base.TextVisualizer`

Display a projection of a vectorized corpus in two dimensions using TSNE, a nonlinear dimensionality reduction method that is particularly well suited to embedding in two or three dimensions for visualization as a scatter plot. TSNE is widely used in text analysis to show clusters or groups of documents or utterances and their relative proximities.

TSNE will return a scatter plot of the vectorized corpus, such that each point represents a document or utterance. The distance between two points in the visual space is embedded using the probability distribution of pairwise similarities in the higher dimensionality; thus TSNE shows clusters of similar documents and the relationships between groups of documents as a scatter plot.

TSNE can be used with either clustering or classification; by specifying the `classes` argument, points will be colored based on their similar traits. For example, by passing `cluster.labels_` as `y` in `fit()`, all points in the same cluster will be grouped together. This extends the neighbor embedding with

more information about similarity, and can allow better interpretation of both clusters and classes.

For more, see https://lvdmaaten.github.io/tsne/

Parameters

ax [matplotlib axes] The axes to plot the figure on.

decompose [string or None, default: `'svd'`] A preliminary decomposition is often used prior to TSNE to make the projection faster. Specify `"svd"` for sparse data or `"pca"` for dense data. If None, the original data set will be used.

decompose_by [int, default: 50] Specify the number of components for preliminary decomposition, by default this is 50; the more components, the slower TSNE will be.

labels [list of strings] The names of the classes in the target, used to create a legend. Labels must match names of classes in sorted order.

colors [list or tuple of colors] Specify the colors for each individual class

colormap [string or matplotlib cmap] Sequential colormap for continuous target

random_state [int, RandomState instance or None, optional, default: None] If int, random_state is the seed used by the random number generator; If RandomState instance, random_state is the random number generator; If None, the random number generator is the RandomState instance used by np.random. The random state is applied to the preliminary decomposition as well as tSNE.

kwargs [dict] Pass any additional keyword arguments to the TSNE transformer.

`NULL_CLASS = None`

`draw`(*points*, *target=None*, ***kwargs*)

Called from the fit method, this method draws the TSNE scatter plot, from a set of decomposed points in 2 dimensions. This method also accepts a third dimension, target, which is used to specify the colors of each of the points. If the target is not specified, then the points are plotted as a single cloud to show similar documents.

`finalize`(***kwargs*)

Finalize the drawing by adding a title and legend, and removing the axes objects that do not convey information about TNSE.

`fit`(*X*, *y=None*, ***kwargs*)

The fit method is the primary drawing input for the TSNE projection since the visualization requires both X and an optional y value. The fit method expects an array of numeric vectors, so text documents must be vectorized before passing them to this method.

Parameters

X [ndarray or DataFrame of shape n x m] A matrix of n instances with m features representing the corpus of vectorized documents to visualize with tsne.

y [ndarray or Series of length n] An optional array or series of target or class values for instances. If this is specified, then the points will be colored according to their class. Often cluster labels are passed in to color the documents in cluster space, so this method is used both for classification and clustering methods.

kwargs [dict] Pass generic arguments to the drawing method

Returns

self [instance] Returns the instance of the transformer/visualizer

`make_transformer`(*decompose='svd'*, *decompose_by=50*, *tsne_kwargs={}*)

Creates an internal transformer pipeline to project the data set into 2D space using TSNE, applying an pre-decomposition technique ahead of embedding if necessary. This method will reset the transformer on the class, and can be used to explore different decompositions.

Parameters

decompose [string or None, default: `'svd'`] A preliminary decomposition is often used prior to TSNE to make the projection faster. Specify `"svd"` for sparse data or `"pca"` for dense data. If decompose is None, the original data set will be used.

decompose_by [int, default: 50] Specify the number of components for preliminary decomposition, by default this is 50; the more components, the slower TSNE will be.

Returns

transformer [Pipeline] Pipelined transformer for TSNE projections

4.3.8 Colors and Style

Yellowbrick believes that visual diagnostics are more effective if visualizations are appealing. As a result, we have borrowed familiar styles from Seaborn and use the new Matplotlib 2.0 styles. We hope that these out of the box styles will make your visualizations publication ready, though of course you can customize your own look and feel by directly modifying the visualization with matplotlib.

Yellowbrick prioritizes color in its visualizations for most visualizers. There are two types of color sets that can be provided to a visualizer: a palette and a sequence. Palettes are discrete color values usually of a fixed length and are typically used for classification or clustering by showing each class, cluster or topic. Sequences are continuous color values that do not have a fixed length but rather a range and are typically used for regression or clustering, showing all possible values in the target or distances between items in clusters.

In order to make the distinction easy, most matplotlib colors (both palettes and sequences) can be referred to by name. A complete listing can be imported as follows:

```
import matplotlib.pyplot as plt
from yellowbrick.style.palettes import PALETTES, SEQUENCES, color_palette
```

Palettes and sequences can be passed to visualizers as follows:

```
visualizer = Visualizer(color="bold")
```

Refer to the API listing of each visualizer for specifications about how each color argument is handled. In the next two sections we will show every possible color palette and sequence currently available in Yellowbrick.

Color Palettes

Color palettes are discrete color lists that have a fixed length. The most common palettes are ordered as "blue", "green", "red", "maroon", "yellow", "cyan", and an optional "key". This allows you to specify these named colors or by the first character, e.g. 'bgrmyck' for matplotlib visualizations.

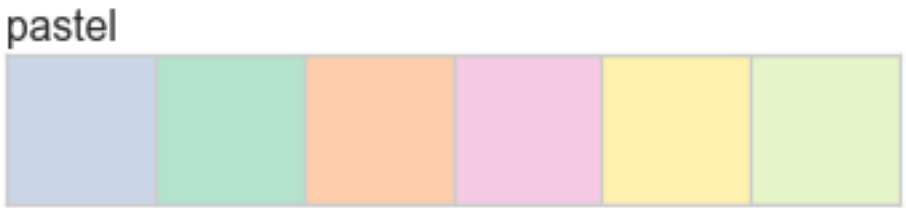
To change the global color palette, use the *set_palette* function as follows:

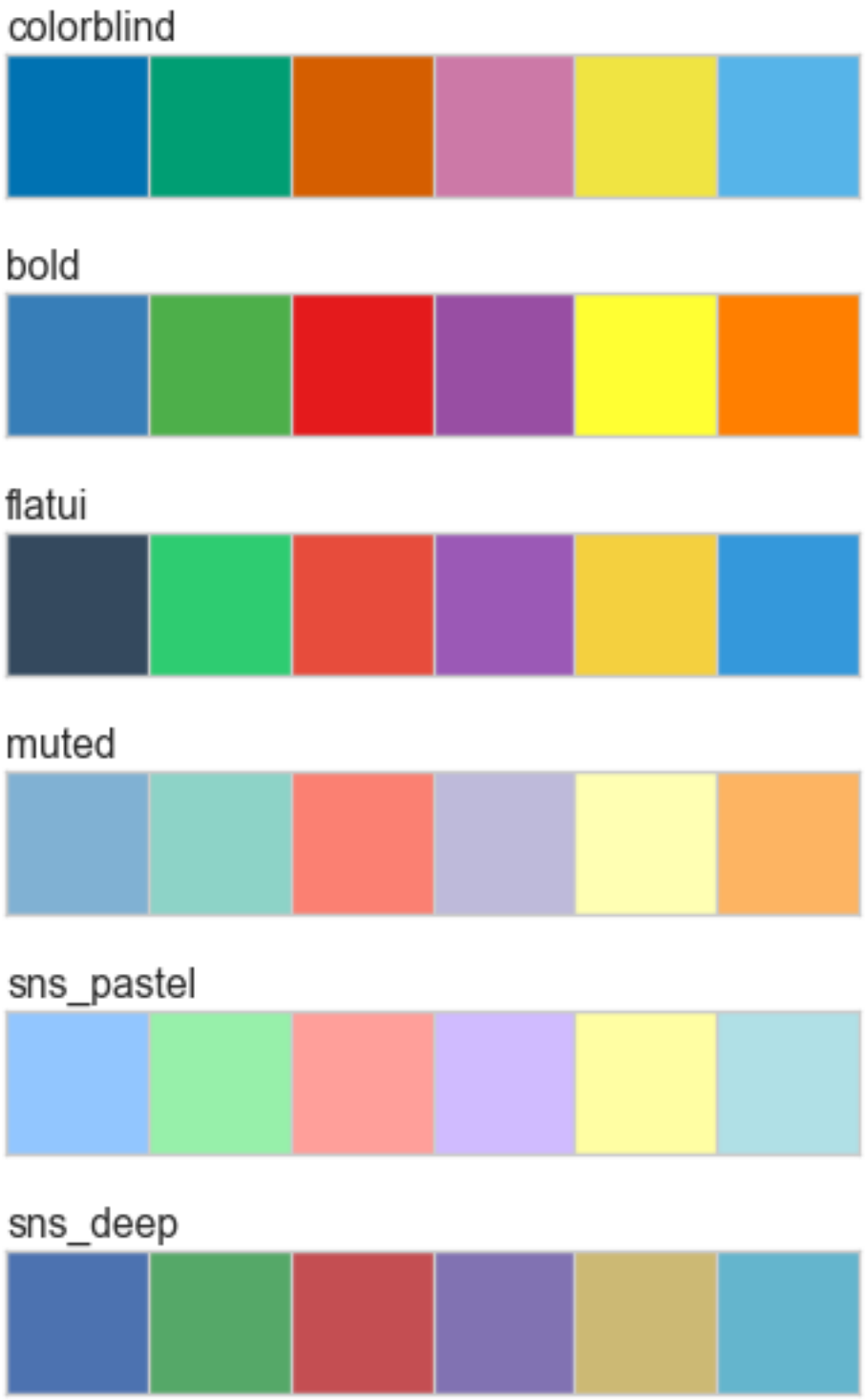
```
from yellowbrick.style import set_palette
set_palette('flatui')
```

Color palettes are most often used for classifiers to show the relationship between discrete class labels. They can also be used for clustering algorithms to show membership in discrete clusters.

A complete listing of the Yellowbrick color palettes can be visualized as follows:

```
# ['blue', 'green', 'red', 'maroon', 'yellow', 'cyan']
for palette in PALETTES.keys():
    color_palette(palette).plot()
    plt.title(palette, loc='left')
```




colorblind
bold
flatui
muted
sns_pastel
sns_deep

accent

sns_dark

dark

paired

sns_muted

sns_colorblind

set1

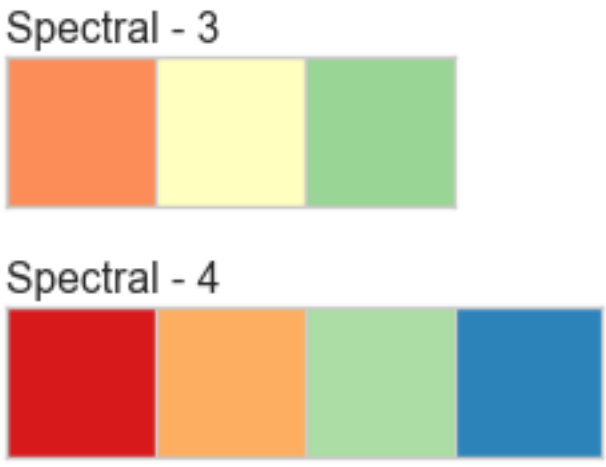
Color Sequences

Color sequences are continuous representations of color and are usually defined as a fixed number of steps between a minimum and maximal value. Sequences must be created with a total number of bins (or length) before plotting to ensure that values are assigned correctly. In the listing below, each sequence is shown with varying lengths to describe the range of colors in detail.

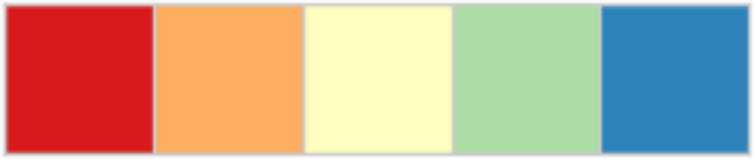
Color sequences are most often used in regressions to show the distribution in the range of target values. They can also be used in clustering and distribution analysis to show distance or histogram data.

Below is a complete listing of all the sequence names available in Yellowbrick:

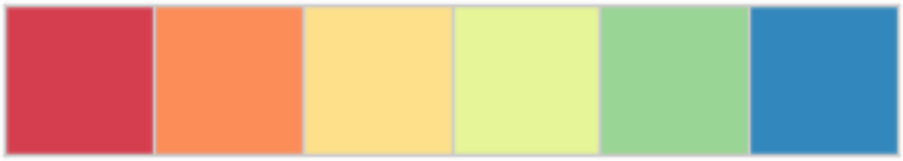
```
for name, maps in SEQUENCES.items():
    for num, palette in maps.items():
        color_palette(palette).plot()
        plt.title("{} - {}".format(name, num), loc='left')
```



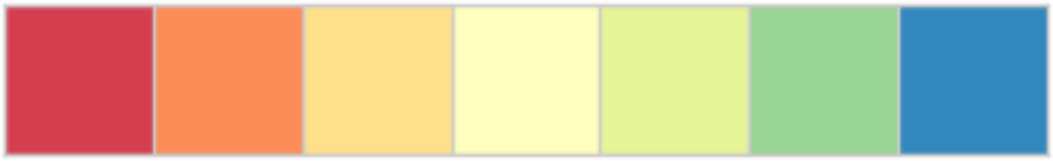
Spectral - 5



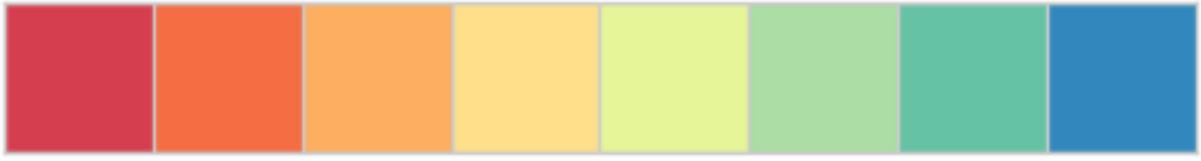
Spectral - 6



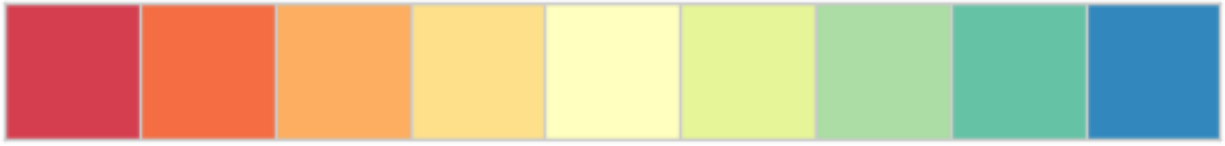
Spectral - 7



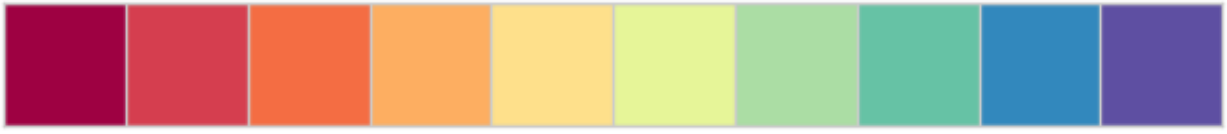
Spectral - 8



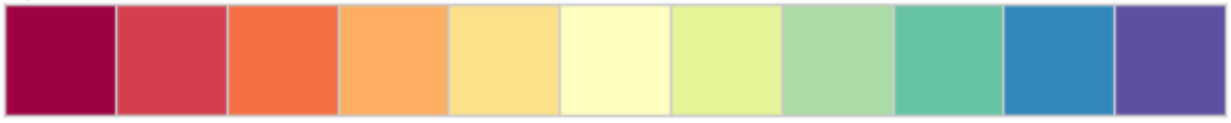
Spectral - 9



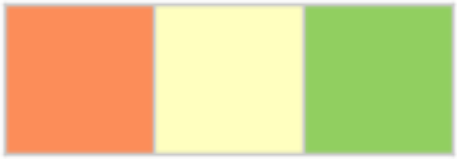
Spectral - 10



Spectral - 11

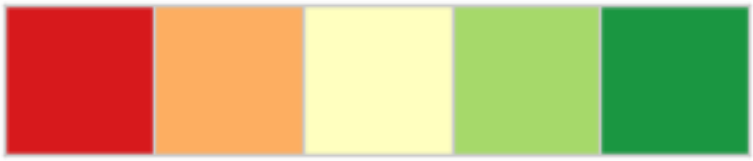


RdYlGn - 3

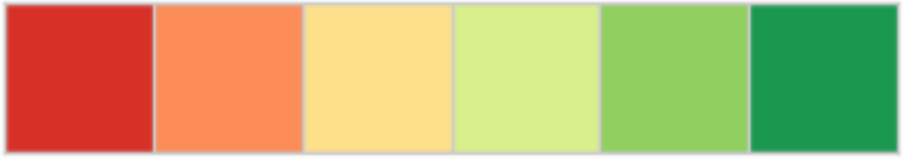


RdYlGn - 4

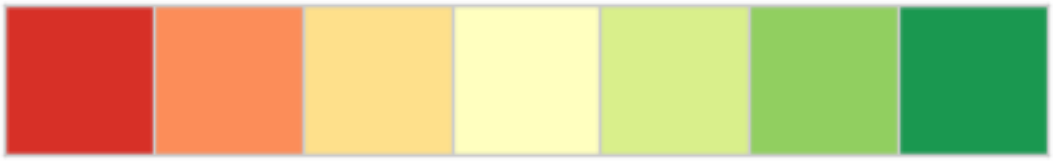
RdYlGn - 5



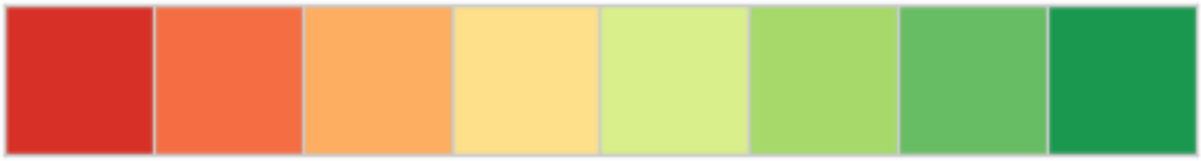
RdYlGn - 6



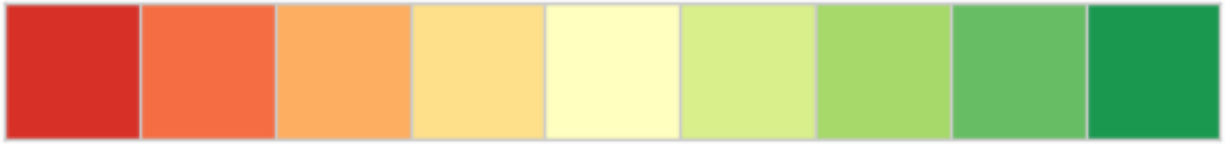
RdYlGn - 7



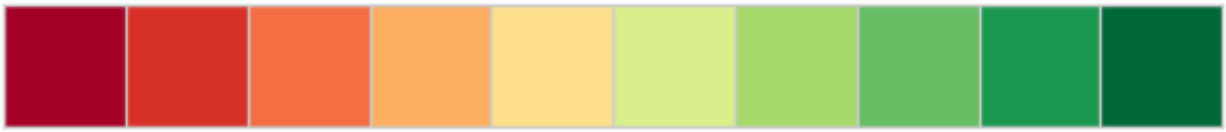
RdYlGn - 8



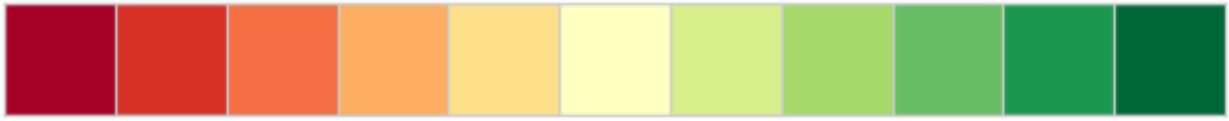
RdYlGn - 9



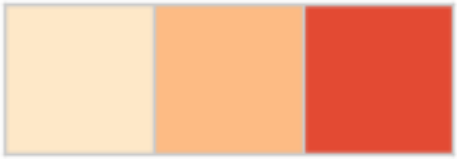
RdYlGn - 10



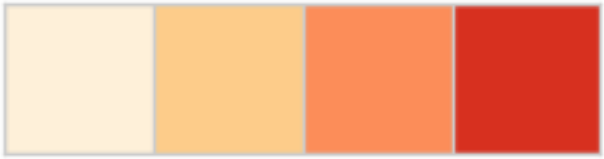
RdYlGn - 11



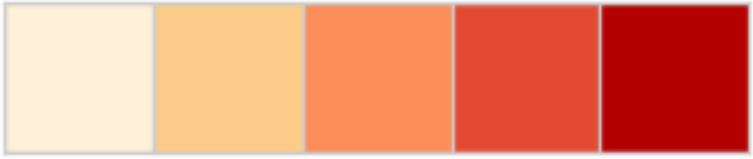
OrRd - 3



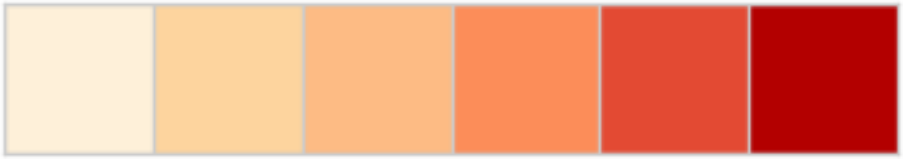
OrRd - 4



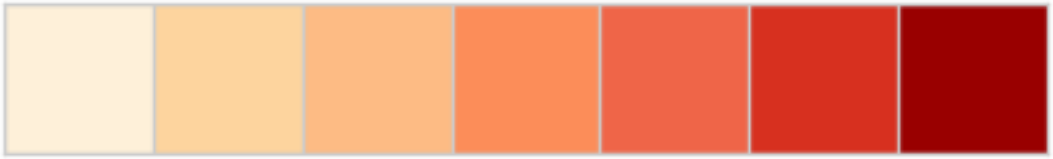
OrRd - 5



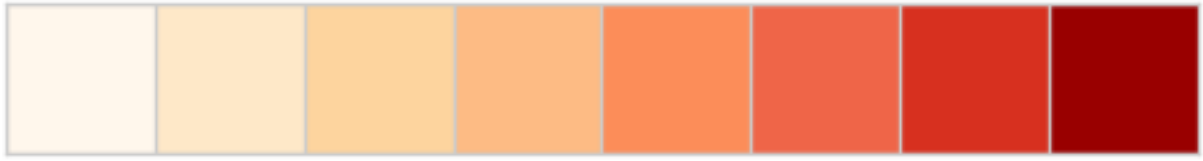
OrRd - 6



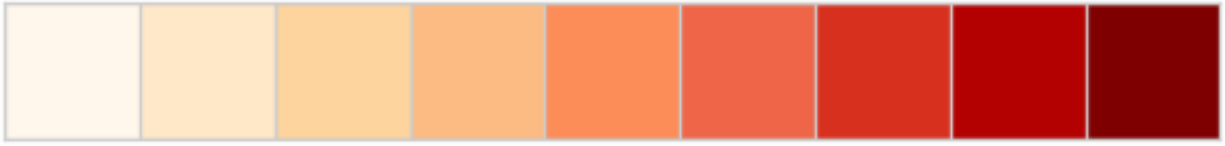
OrRd - 7



OrRd - 8



OrRd - 9



PuBu - 3

PuBu - 4

PuBu - 5

PuBu - 6

PuBu - 7

PuBu - 8

PuBu - 9

BuPu - 3

BuPu - 4

BuPu - 5

BuPu - 6

BuPu - 7

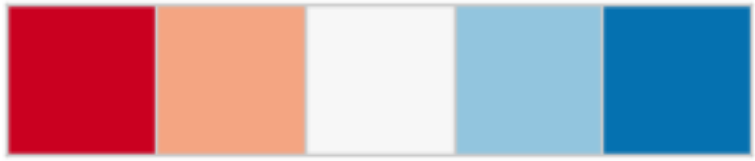
BuPu - 8

BuPu - 9

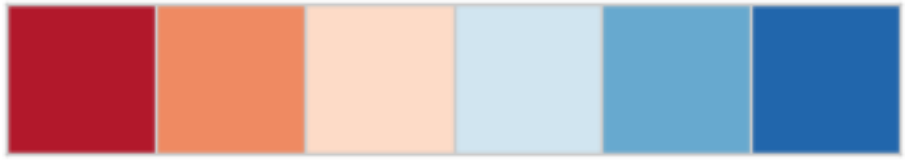
RdBu - 3

RdBu - 4

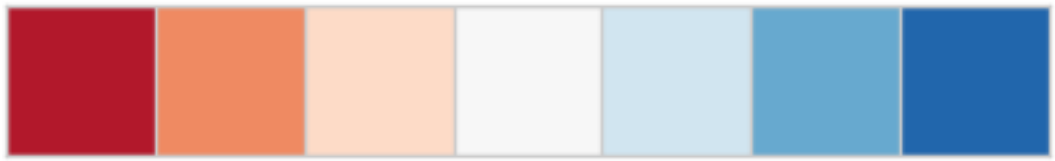
RdBu - 5



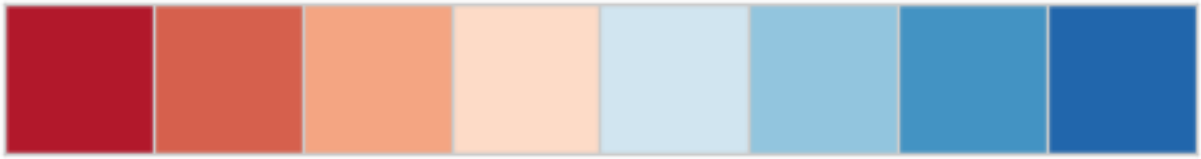
RdBu - 6



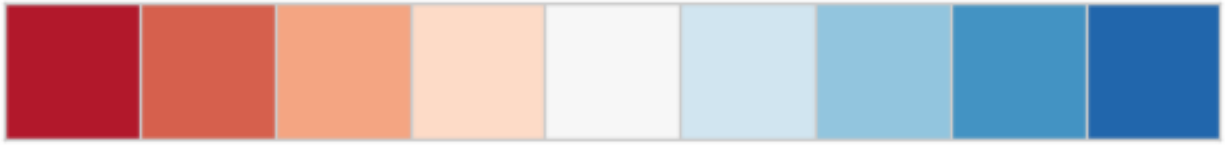
RdBu - 7



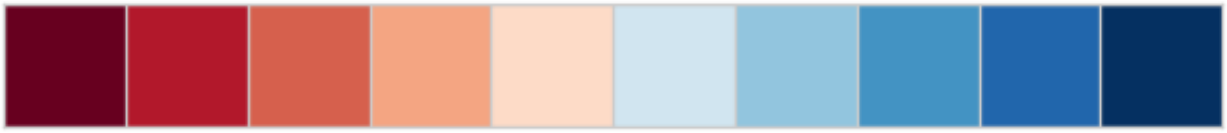
RdBu - 8



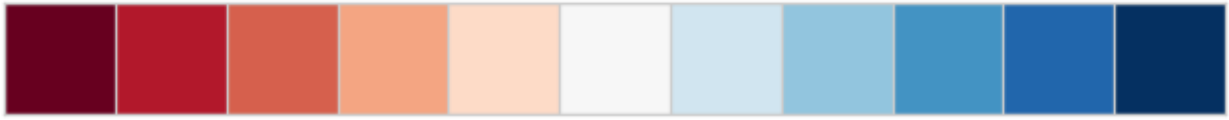
RdBu - 9



RdBu - 10



RdBu - 11



Oranges - 3

Oranges - 4

Oranges - 5

Oranges - 6

Oranges - 7

Oranges - 8

Oranges - 9

BuGn - 3

BuGn - 4

BuGn - 5

BuGn - 6

BuGn - 7

BuGn - 8

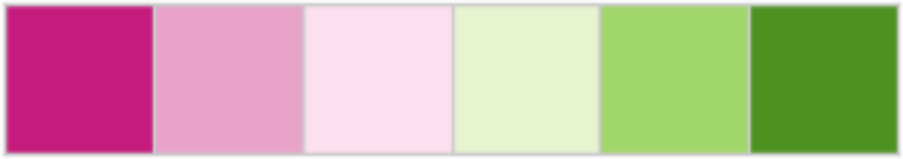
BuGn - 9

PiYG - 3

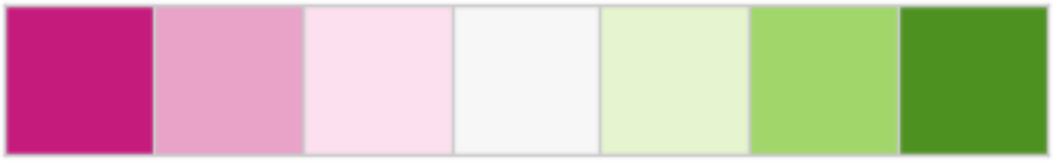
PiYG - 4

PiYG - 5

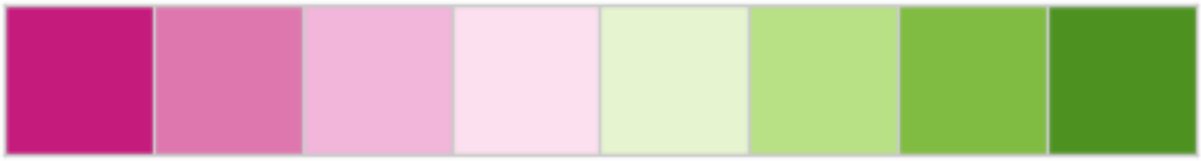
PiYG - 6



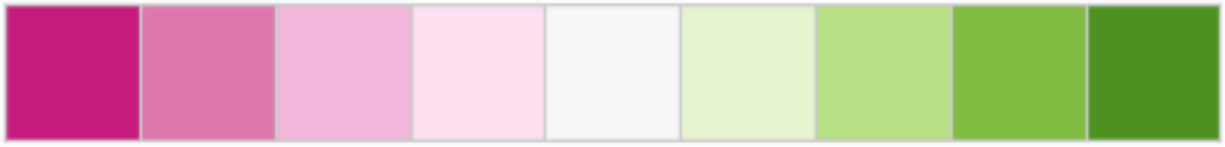
PiYG - 7



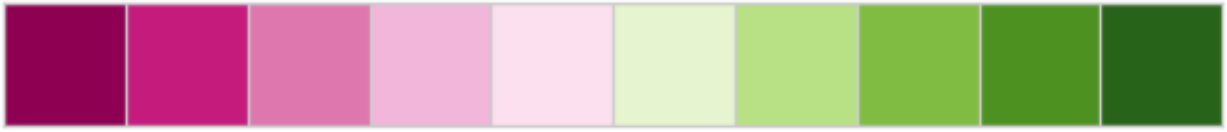
PiYG - 8



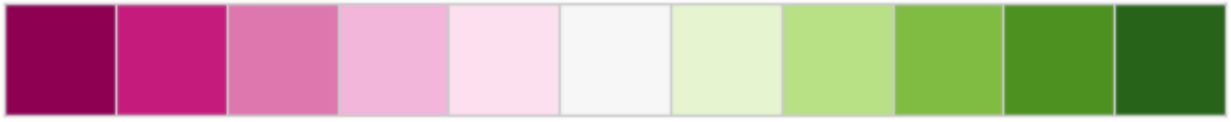
PiYG - 9



PiYG - 10

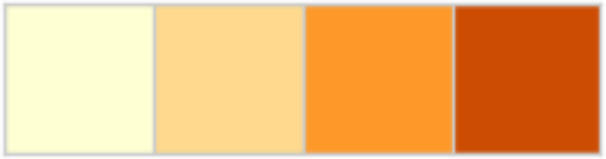


PiYG - 11



YlOrBr - 3

YlOrBr - 4



YlOrBr - 5

YlOrBr - 6

YlOrBr - 7

YlOrBr - 8

YlOrBr - 9

YlGn - 3

YlGn - 4

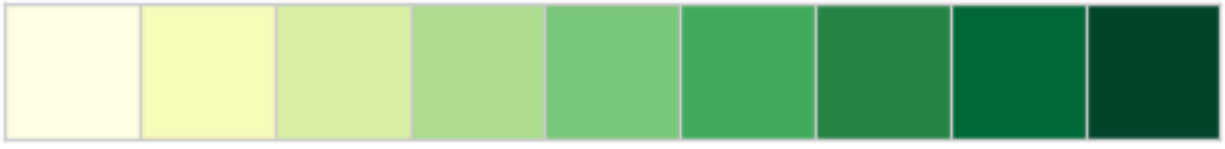
YlGn - 5

YlGn - 6

YlGn - 7

YlGn - 8

YlGn - 9



RdPu - 3

RdPu - 4

RdPu - 5

RdPu - 6

RdPu - 7

RdPu - 8

RdPu - 9

Greens - 3

Greens - 4

Greens - 5

Greens - 6

Greens - 7

Greens - 8

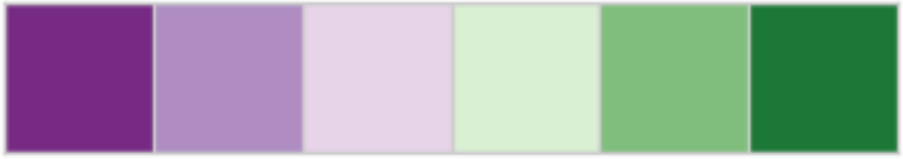
Greens - 9

PRGn - 3

PRGn - 4

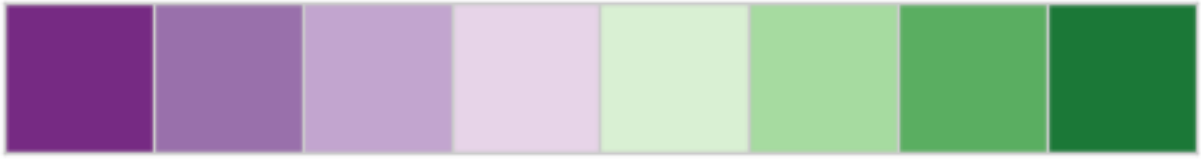
PRGn - 5

PRGn - 6

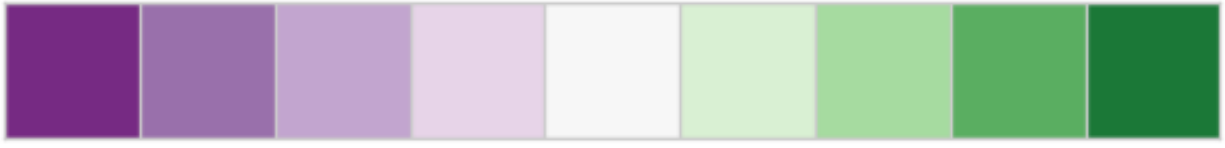


PRGn - 7

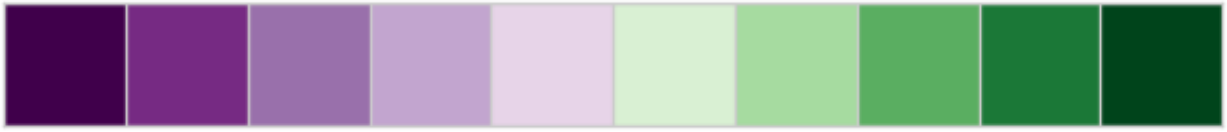
PRGn - 8



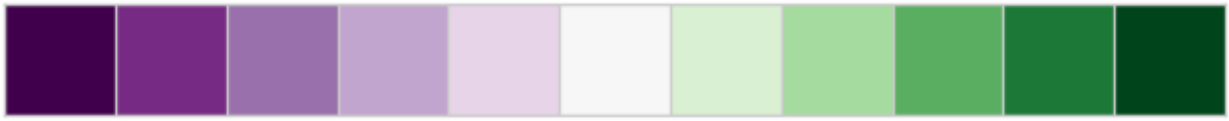
PRGn - 9



PRGn - 10



PRGn - 11



YlGnBu - 3

YlGnBu - 4

YlGnBu - 5

YlGnBu - 6

YlGnBu - 7

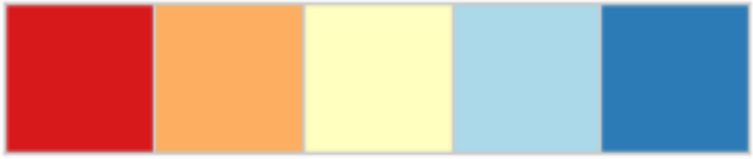
YlGnBu - 8

YlGnBu - 9

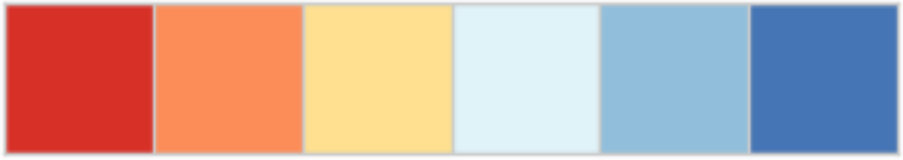
RdYlBu - 3

RdYlBu - 4

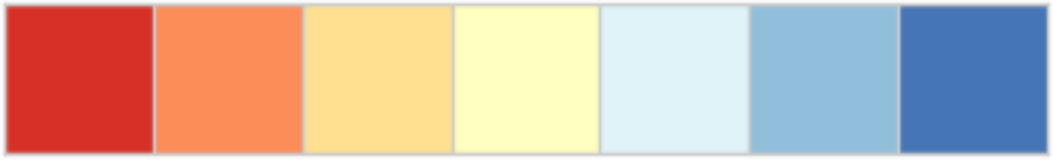
RdYlBu - 5



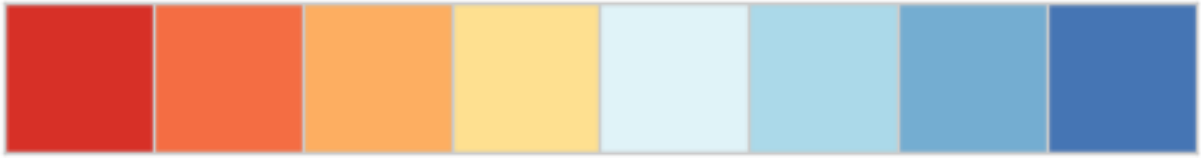
RdYlBu - 6



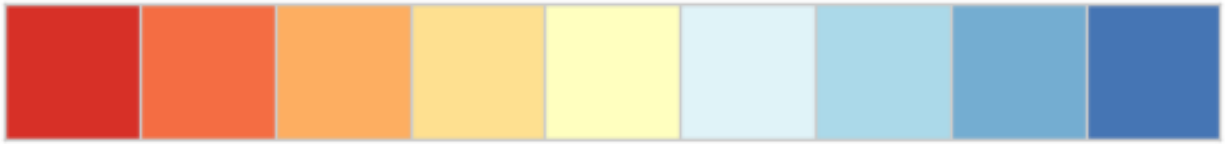
RdYlBu - 7



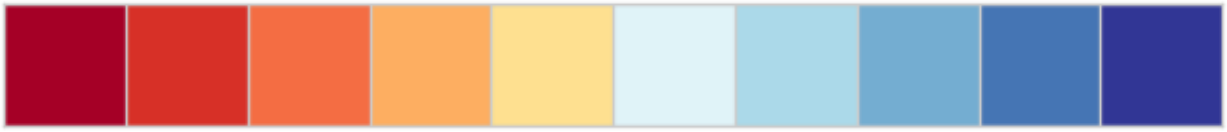
RdYlBu - 8



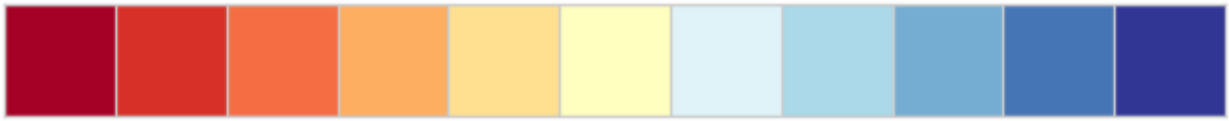
RdYlBu - 9



RdYlBu - 10



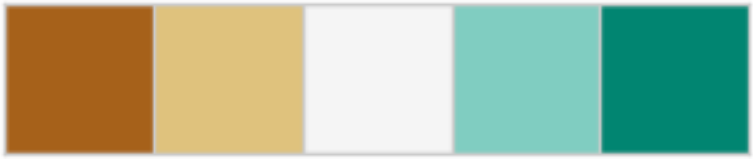
RdYlBu - 11



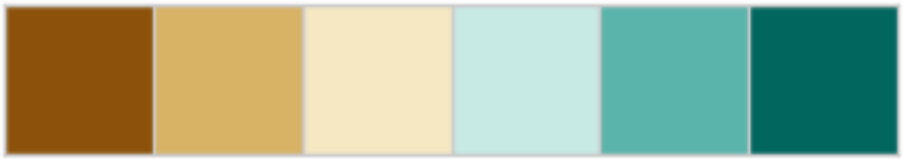
BrBG - 3

BrBG - 4

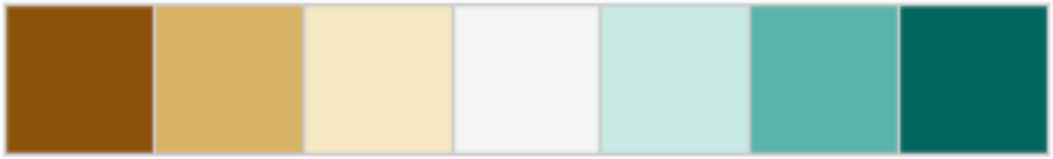
BrBG - 5



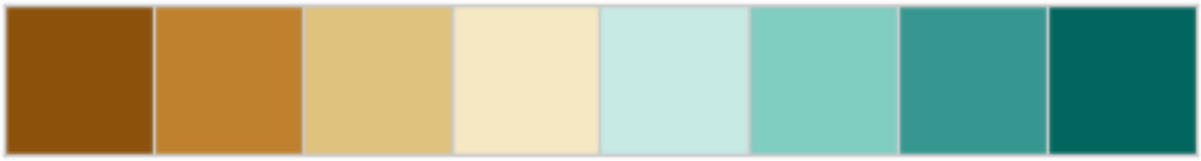
BrBG - 6



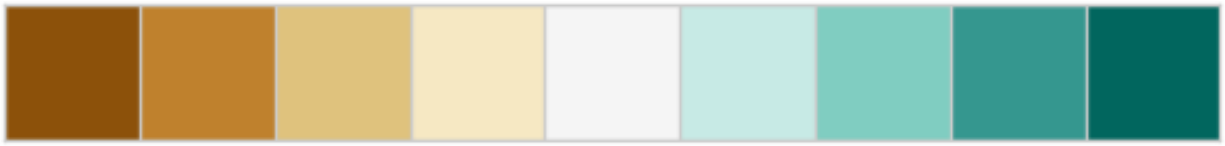
BrBG - 7



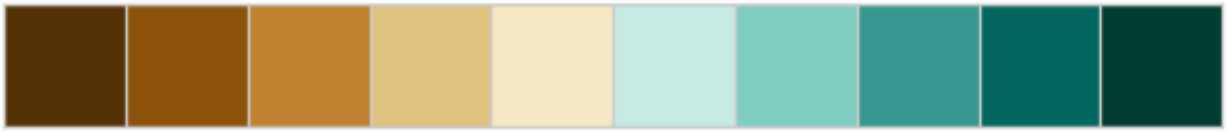
BrBG - 8



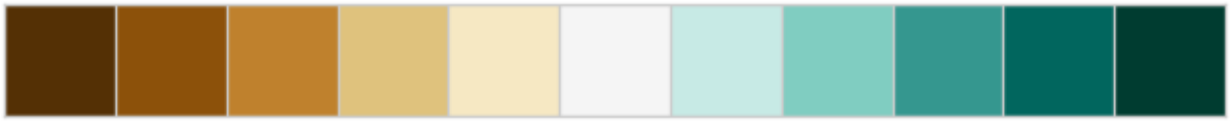
BrBG - 9



BrBG - 10



BrBG - 11



Purples - 3

Purples - 4

Purples - 5

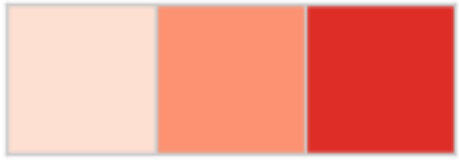
Purples - 6

Purples - 7

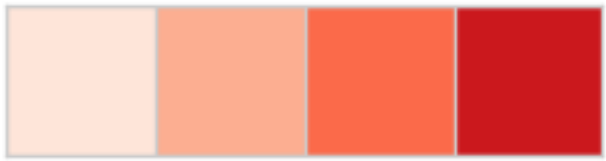
Purples - 8

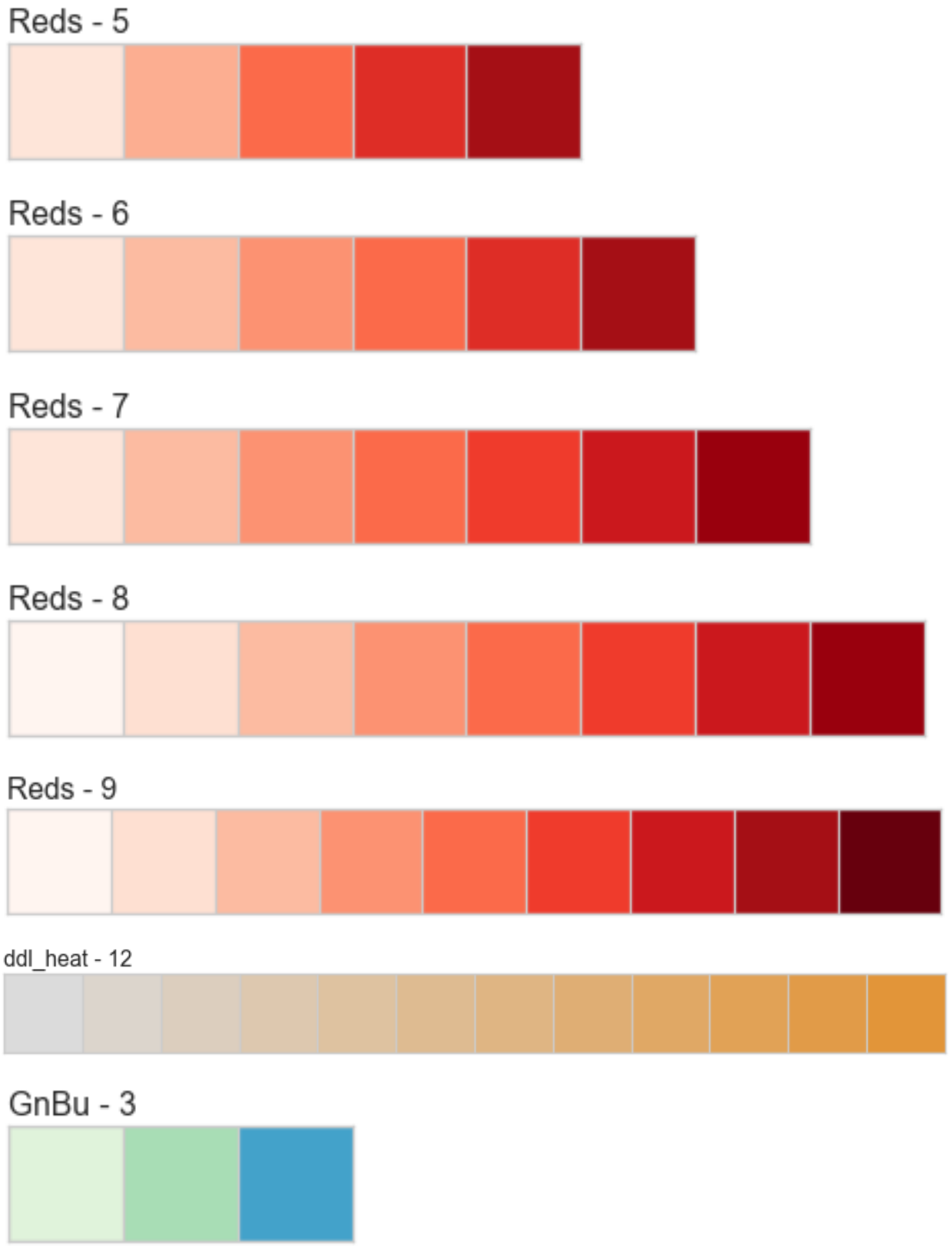
Purples - 9

Reds - 3



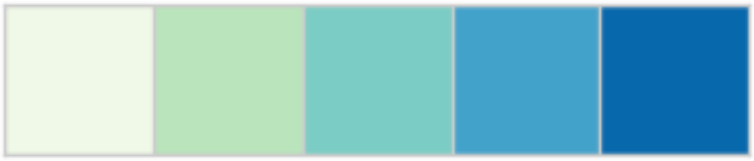
Reds - 4




Reds - 5
Reds - 6
Reds - 7
Reds - 8
Reds - 9
ddl_heat - 12
GnBu - 3

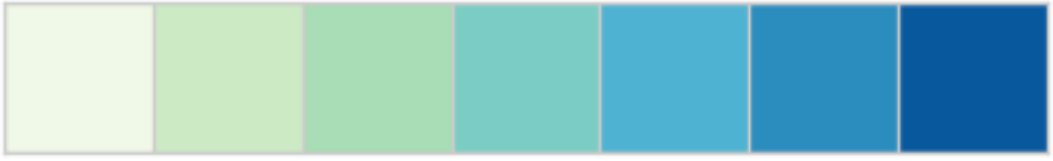
GnBu - 4

GnBu - 5



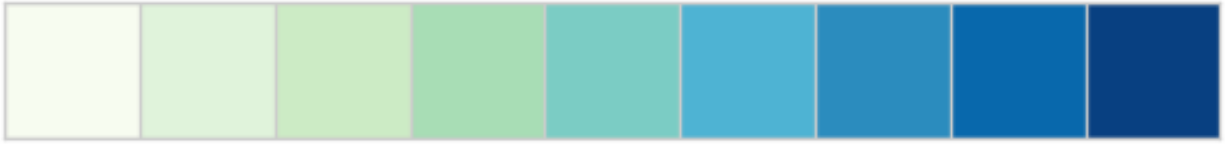
GnBu - 6

GnBu - 7



GnBu - 8

GnBu - 9



Greys - 3

Greys - 4

Greys - 5

Greys - 6

Greys - 7

Greys - 8

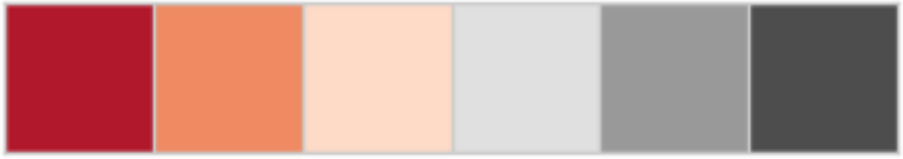
Greys - 9

RdGy - 3

RdGy - 4

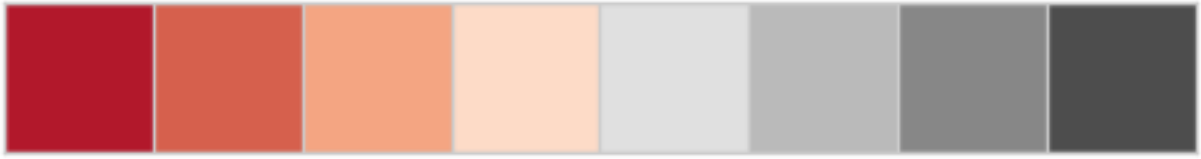
RdGy - 5

RdGy - 6



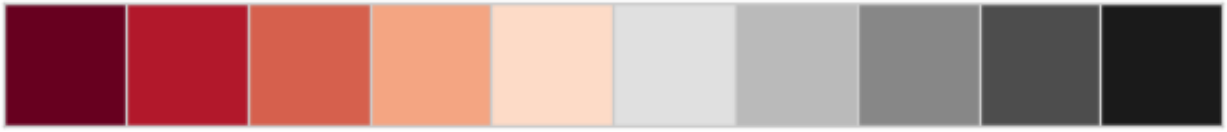
RdGy - 7

RdGy - 8

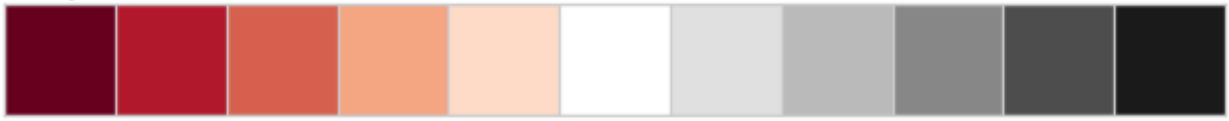


RdGy - 9

RdGy - 10



RdGy - 11

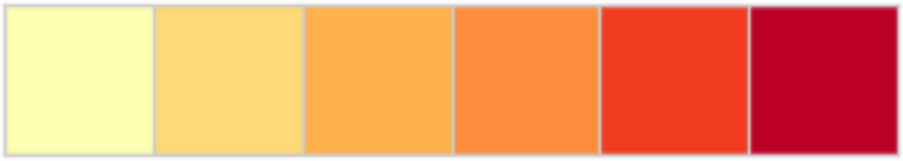


YlOrRd - 3

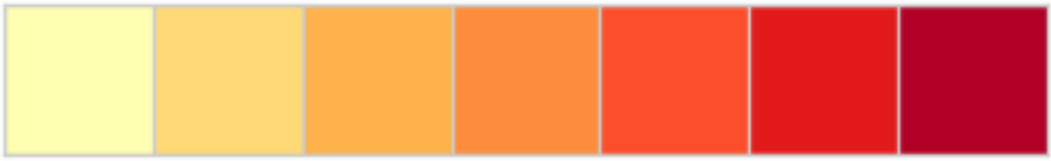
YlOrRd - 4

YlOrRd - 5

YlOrRd - 6



YlOrRd - 7



YlOrRd - 8

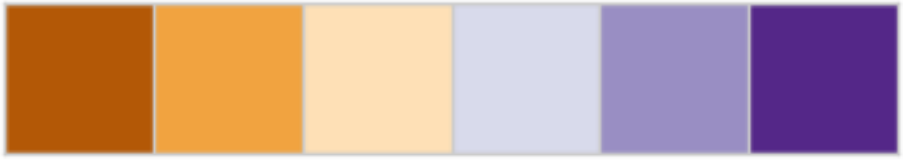
YlOrRd - 9

PuOr - 3

PuOr - 4

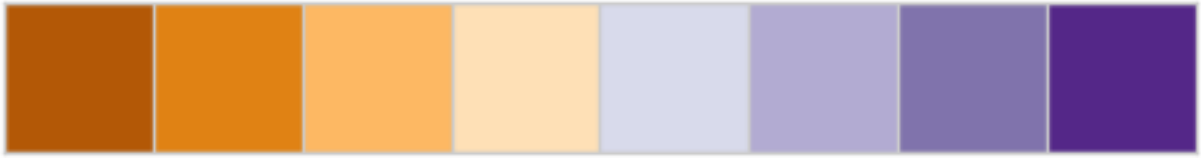
PuOr - 5

PuOr - 6



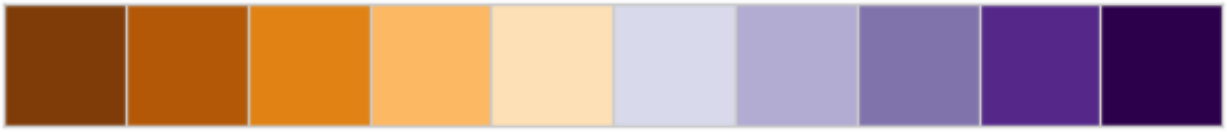
PuOr - 7

PuOr - 8



PuOr - 9

PuOr - 10



PuOr - 11

PuRd - 3

PuRd - 4

PuRd - 5

PuRd - 6

PuRd - 7

PuRd - 8

PuRd - 9

Blues - 3

Blues - 4

Blues - 5

Blues - 6

Blues - 7

Blues - 8

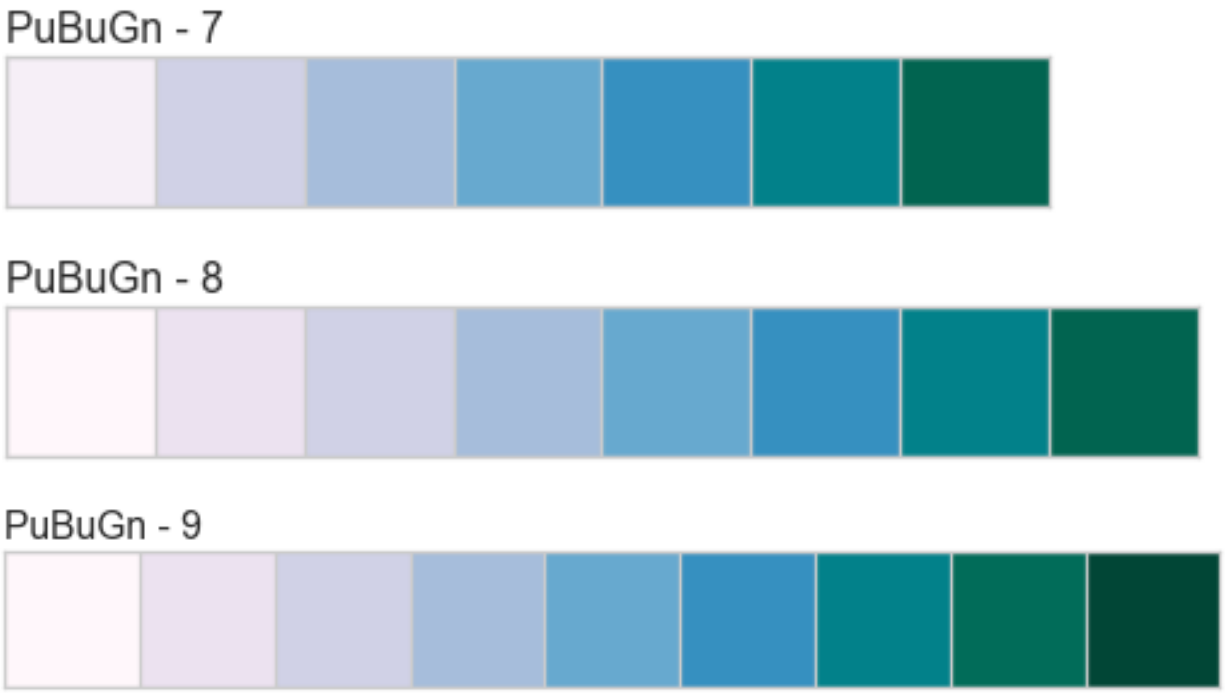
Blues - 9

PuBuGn - 3

PuBuGn - 4

PuBuGn - 5

PuBuGn - 6



API Reference

yellowbrick.style.colors module

Colors and color helpers brought in from an alternate library. See https://bl.ocks.org/mbostock/5577023

class yellowbrick.style.colors.**ColorMap**(*colors='flatui'*, *shuffle=False*)

基类: object

A helper for mapping categorical values to colors on demand.

colors

yellowbrick.style.colors.**get_color_cycle**()

Returns the current color cycle from matplotlib.

yellowbrick.style.colors.**resolve_colors**(*n_colors=None*, *colormap=None*, *colors=None*)

Generates a list of colors based on common color arguments, for example the name of a colormap or palette or another iterable of colors. The list is then truncated (or multiplied) to the specific number of requested colors.

Parameters

n_colors [int, default: None] Specify the length of the list of returned colors, which will either truncate or multiple the colors available. If None the length of the colors will not be modified.

colormap [str, default: None] The name of the matplotlib color map with which to generate colors.

colors [iterable, default: None] A collection of colors to use specifically with the plot.

Returns

colors [list] A list of colors that can be used in matplotlib plots.

Notes

This function was originally based on a similar function in the pandas plotting library that has been removed in the new version of the library.

yellowbrick.style.palettes module

Implements the variety of colors that yellowbrick allows access to by name. This code was originally based on Seaborn's rcmody.py but has since been cleaned up to be Yellowbrick-specific and to dereference tools we don't use. Note that these functions alter the matplotlib rc dictionary on the fly.

`yellowbrick.style.palettes.color_palette`(*palette=None*, *n_colors=None*)

Return a color palette object with color definition and handling.

Calling this function with `palette=None` will return the current matplotlib color cycle.

This function can also be used in a `with` statement to temporarily set the color cycle for a plot or set of plots.

Parameters

palette [None or str or sequence] Name of a palette or `None` to return the current palette. If a sequence the input colors are used but possibly cycled.

Available palette names from `yellowbrick.colors.palettes` are:

- `accent`
- `dark`
- `paired`
- `pastel`
- `bold`
- `muted`
- `colorblind`
- `sns_colorblind`
- `sns_deep`
- `sns_muted`
- `sns_pastel`
- `sns_bright`

- sns_dark
- flatui
- neural_paint

n_colors [None or int] Number of colors in the palette. If None, the default will depend on how palette is specified. Named palettes default to 6 colors which allow the use of the names "bgrmyck", though others do have more or less colors; therefore reducing the size of the list can only be done by specifying this parameter. Asking for more colors than exist in the palette will cause it to cycle.

Returns

list(tuple) Returns a ColorPalette object, which behaves like a list, but can be used as a context manager and possesses functions to convert colors.

.. seealso::

set_palette() Set the default color cycle for all plots.

set_color_codes() Reassign color codes like "b", "g", etc. to colors from one of the yellowbrick palettes.

colors.resolve_colors() Resolve a color map or listed sequence of colors.

yellowbrick.style.palettes.**set_color_codes**(*palette='accent'*)

Change how matplotlib color shorthands are interpreted.

Calling this will change how shorthand codes like "b" or "g" are interpreted by matplotlib in subsequent plots.

Parameters

palette [str] Named yellowbrick palette to use as the source of colors.

参见:

set_palette Color codes can also be set through the function that sets the matplotlib color cycle.

yellowbrick.style.rcmod module

Modifies the matplotlib rcParams in order to make yellowbrick more appealing. This has been modified from Seaborn's rcmod.py: github.com/mwaskom/seaborn in order to alter the matplotlib rc dictionary on the fly.

NOTE: matplotlib 2.0 styles mean we can simply convert this to a stylesheet!

yellowbrick.style.rcmod.**set_aesthetic**(*palette='yellowbrick'*, *font='sans-serif'*, *font_scale=1*, *color_codes=True*, *rc=None*)

Set aesthetic parameters in one step.

Each set of parameters can be set directly or temporarily, see the referenced functions below for more information.

Parameters

palette [string or sequence] Color palette, see `color_palette()`

font [string] Font family, see matplotlib font manager.

font_scale [float, optional] Separate scaling factor to independently scale the size of the font elements.

color_codes [bool] If `True` and `palette` is a yellowbrick palette, remap the shorthand color codes (e.g. "b", "g", "r", etc.) to the colors from this palette.

rc [dict or None] Dictionary of rc parameter mappings to override the above.

`yellowbrick.style.rcmod.set_style(`*style=None*, *rc=None*`)`

Set the aesthetic style of the plots.

This affects things like the color of the axes, whether a grid is enabled by default, and other aesthetic elements.

Parameters

style [dict, None, or one of {darkgrid, whitegrid, dark, white, ticks}] A dictionary of parameters or the name of a preconfigured set.

rc [dict, optional] Parameter mappings to override the values in the preset seaborn style dictionaries. This only updates parameters that are considered part of the style definition.

`yellowbrick.style.rcmod.set_palette(`*palette*, *n_colors=None*, *color_codes=False*`)`

Set the matplotlib color cycle using a seaborn palette.

Parameters

palette [yellowbrick color palette | seaborn color palette (with `sns_` prepended)] Palette definition. Should be something that `color_palette()` can process.

n_colors [int] Number of colors in the cycle. The default number of colors will depend on the format of `palette`, see the `color_palette()` documentation for more information.

color_codes [bool] If `True` and `palette` is a seaborn palette, remap the shorthand color codes (e.g. "b", "g", "r", etc.) to the colors from this palette.

`yellowbrick.style.rcmod.reset_defaults()`

Restore all RC params to default settings.

`yellowbrick.style.rcmod.reset_orig()`

Restore all RC params to original settings (respects custom rc).

注解: Many examples utilize data from the UCI Machine Learning repository, in order to run the examples,

make sure you follow the instructions in *Example Datasets* to download and load required data.

A guide to finding the visualizer you're looking for: generally speaking, visualizers can be data visualizers which visualize instances relative to the model space; score visualizers which visualize model performance; model selection visualizers which compare multiple model forms against each other; and application specific-visualizers. This can be a bit confusing, so we've grouped visualizers according to the type of analysis they are well suited for.

Feature analysis visualizers are where you'll find the primary implementation of data visualizers. Regression, classification, and clustering analysis visualizers can be found in their respective libraries. Finally visualizers for text analysis are also available in Yellowbrick! Other utilities like styles, best fit lines, and anscombe's visualization can also be found in the links above.

4.4 User Testing Instructions

We are looking for people to help us Alpha test the Yellowbrick project! Helping is simple: simply create a notebook that applies the concepts in this *Getting Started* guide to a small-to-medium size dataset of your choice. Run through the examples with the dataset, and try to change options and customize as much as possible. After you've exercised the code with your examples, respond to our alpha testing survey!

4.4.1 Step One: Questionaire

Please open the quesionaire, in order to familiarize yourself with the feedback that we are looking to receive. We are very interested in identifying any bugs in Yellowbrick. Please include al cells in your jupyter notebook that produce errors so that we may reproduce the problem.

4.4.2 Step Two: Dataset

Select a multivariate dataset of your own; the more (e.g. different) datasets that we can run through Yellowbrick, the more likely we'll discover edge cases and exceptions! Note that your dataset must be well-suited to modeling with Scikit-Learn. In particular we recommend you choose a dataset whose target is suited to the following supervised learning tasks:

- Regression (target is a continuous variable)
- Classification (target is a discrete variable)

There are datasets that are well suited to both types of analysis; either way you can use the testing methodology from this notebook for either type of task (or both). In order to find a dataset, we recommend you try the following places:

- UCI Machine Learning Repository
- MLData.org

- Awesome Public Datasets

You're more than welcome to choose a dataset of your own, but we do ask that you make at least the notebook containing your testing results publicly available for us to review. If the data is also public (or you're willing to share it with the primary contributors) that will help us figure out bugs and required features much more easily!

4.4.3 Step Three: Notebook

Create a notebook in a GitHub repository. We suggest the following:

1. Fork the Yellowbrick repository
2. Under the `examples` directory, create a directory named with your GitHub username
3. Create a notebook named `testing`, i.e. examples/USERNAME/testing.ipynb

Alternatively, you could just send us a notebook via Gist or your own repository. However, if you fork Yellowbrick, you can initiate a pull request to have your example added to our gallery!

4.4.4 Step Four: Model with Yellowbrick and Scikit-Learn

Add the following to the notebook:

- A title in markdown
- A description of the dataset and where it was obtained
- A section that loads the data into a Pandas dataframe or NumPy matrix

Then conduct the following modeling activities:

- Feature analysis using Scikit-Learn and Yellowbrick
- Estimator fitting using Scikit-Learn and Yellowbrick

You can follow along with our `examples` directory (check out examples.ipynb) or even create your own custom visualizers! The goal is that you create an end-to-end model from data loading to estimator(s) with visualizers along the way.

IMPORTANT: please make sure you record all errors that you get and any tracebacks you receive for step three!

4.4.5 Step Five: Feedback

Finally, submit feedback via the Google Form we have created:

https://goo.gl/forms/naoPUMFa1xNcafY83

This form is allowing us to aggregate multiple submissions and bugs so that we can coordinate the creation and management of issues. If you are the first to report a bug or feature request, we will make sure you're notified (we'll tag you using your Github username) about the created issue!

4.4.6 Step Six: Thanks!

Thank you for helping us make Yellowbrick better! We'd love to see pull requests for features you think would be extend the library. We'll also be doing a user study that we would love for you to participate in. Stay tuned for more great things from Yellowbrick!

4.5 Contributing

Yellowbrick is an open source project that is supported by a community who will gratefully and humbly accept any contributions you might make to the project. Large or small, any contribution makes a big difference; and if you've never contributed to an open source project before, we hope you will start with Yellowbrick!

Principally, Yellowbrick development is about the addition and creation of *visualizers* — objects that learn from data and create a visual representation of the data or model. Visualizers integrate with scikit-learn estimators, transformers, and pipelines for specific purposes and as a result, can be simple to build and deploy. The most common contribution is a new visualizer for a specific model or model family. We'll discuss in detail how to build visualizers later.

Beyond creating visualizers, there are many ways to contribute:

- Submit a bug report or feature request on GitHub Issues.
- Contribute an Jupyter notebook to our examples gallery.
- Assist us with user testing.
- Add to the documentation or help with our website, scikit-yb.org
- Write unit or integration tests for our project.
- Answer questions on our issues, mailing list, Stack Overflow, and elsewhere.
- Translate our documentation into another language.
- Write a blog post, tweet, or share our project with others.
- Teach someone how to use Yellowbrick.

As you can see, there are lots of ways to get involved and we would be very happy for you to join us! The only thing we ask is that you abide by the principles of openness, respect, and consideration of others as described in the Python Software Foundation Code of Conduct.

4.5.1 Getting Started on GitHub

Yellowbrick is hosted on GitHub at https://github.com/DistrictDataLabs/yellowbrick.

The typical workflow for a contributor to the codebase is as follows:

1. **Discover** a bug or a feature by using Yellowbrick.
2. **Discuss** with the core contributors by adding an issue.
3. **Assign** yourself the task by pulling a card from our Waffle Kanban
4. **Fork** the repository into your own GitHub account.
5. Create a **Pull Request** first thing to connect with us about your task.
6. **Code** the feature, write the tests and documentation, add your contribution.
7. **Review** the code with core contributors who will guide you to a high quality submission.
8. **Merge** your contribution into the Yellowbrick codebase.

注解: Please create a pull request as soon as possible, even before you've started coding. This will allow the core contributors to give you advice about where to add your code or utilities and discuss other style choices and implementation details as you go. Don't wait!

We believe that *contribution is collaboration* and therefore emphasize *communication* throughout the open source process. We rely heavily on GitHub's social coding tools to allow us to do this.

Forking the Repository

The first step is to fork the repository into your own account. This will create a copy of the codebase that you can edit and write to. Do so by clicking the **"fork"** button in the upper right corner of the Yellowbrick GitHub page.

Once forked, use the following steps to get your development environment set up on your computer:

1. Clone the repository.

   After clicking the fork button, you should be redirected to the GitHub page of the repository in your user account. You can then clone a copy of the code to your local machine.:

   ```
   $ git clone https://github.com/[YOURUSERNAME]/yellowbrick
   $ cd yellowbrick
   ```

2. Create a virtual environment.

Yellowbrick developers typically use virtualenv (and virtualenvwrapper), pyenv or conda envs in order to manage their Python version and dependencies. Using the virtual environment tool of your choice, create one for Yellowbrick. Here's how with virtualenv:

```
$ virtualenv venv
```

3. Install dependencies.

   Yellowbrick's dependencies are in the `requirements.txt` document at the root of the repository. Open this file and uncomment the dependencies that are for development only. Then install the dependencies with `pip`:

   ```
   $ pip install -r requirements.txt
   ```

   Note that there may be other dependencies required for development and testing; you can simply install them with `pip`. For example to install the additional dependencies for building the documentation or to run the test suite, use the `requirements.txt` files in those directories:

   ```
   $ pip install -r tests/requirements.txt
   $ pip install -r docs/requirements.txt
   ```

4. Switch to the develop branch.

   The Yellowbrick repository has a `develop` branch that is the primary working branch for contributions. It is probably already the branch you're on, but you can make sure and switch to it as follows:

   ```
   $ git fetch
   $ git checkout develop
   ```

At this point you're ready to get started writing code. If you're going to take on a specific task, we'd strongly encourage you to check out the issue on Waffle and create a pull request *before you start coding* to better foster communication with other contributors. More on this in the next section.

Pull Requests

A pull request (PR) is a GitHub tool for initiating an exchange of code and creating a communication channel for Yellowbrick maintainers to discuss your contribution. In essenence, you are requesting that the maintainers merge code from your forked repository into the develop branch of the primary Yellowbrick repository. Once completed, your code will be part of Yellowbrick!

When starting a Yellowbrick contribution, *open the pull request as soon as possible*. We use your PR issue page to discuss your intentions and to give guidance and direction. Every time you push a commit into your forked repository, the commit is automatically included with your pull request, therefore we can review as

you code. The earlier you open a PR, the more easily we can incorporate your updates, we'd hate for you to do a ton of work only to discover someone else already did it or that you went in the wrong direction and need to refactor.

注解: For a great example of a pull request for a new feature visualizer, check out this one by Carlo Morales.

When you open a pull request, ensure it is from your forked repository to the develop branch of github.com/districtdatalabs/yellowbrick; we will not merge a PR into the master branch. Title your Pull Request so that it is easy to understand what you're working on at a glance. Also be sure to include a reference to the issue that you're working on so that correct references are set up.

After you open a PR, you should get a message from one of the maintainers. Use that time to discuss your idea and where best to implement your work. Feel free to go back and forth as you are developing with questions in the comment thread of the PR. Once you are ready, please ensure that you explicitly ping the maintainer to do a code review. Before code review, your PR should contain the following:

1. Your code contribution
2. Tests for your contribution
3. Documentation for your contribution
4. A PR comment describing the changes you made and how to use them
5. A PR comment that includes an image/example of your visualizer

At this point your code will be formally reviewed by one of the contributors. We use GitHub's code review tool, starting a new code review and adding comments to specific lines of code as well as general global comments. Please respond to the comments promptly, and don't be afraid to ask for help implementing any requested changes! You may have to go back and forth a couple of times to complete the code review.

When the following is true:

1. Code is reviewed by at least one maintainer
2. Continuous Integration tests have passed
3. Code coverage and quality have not decreased
4. Code is up to date with the yellowbrick develop branch

Then we will "Squash and Merge" your contribution, combining all of your commits into a single commit and merging it into the develop branch of Yellowbrick. Congratulations! Once your contribution has been merged into master, you will be officially listed as a contributor.

4.5.2 Developing Visualizers

In this section, we'll discuss the basics of developing visualizers. This of course is a big topic, but hopefully these simple tips and tricks will help make sense. First thing though, check out this presentation that

we put together on yellowbrick development, it discusses the expected user workflow, our integration with scikit-learn, our plans and roadmap, etc:

One thing that is necessary is a good understanding of scikit-learn and Matplotlib. Because our API is intended to integrate with scikit-learn, a good start is to review "APIs of scikit-learn objects" and "rolling your own estimator". In terms of matplotlib, use Yellowbrick's guide *Effective Matplotlib*. Additional resources include Nicolas P. Rougier's Matplotlib tutorial and Chris Moffitt's Effectively Using Matplotlib.

Visualizer API

There are two basic types of Visualizers:

- **Feature Visualizers** are high dimensional data visualizations that are essentially transformers.
- **Score Visualizers** wrap a scikit-learn regressor, classifier, or clusterer and visualize the behavior or performance of the model on test data.

These two basic types of visualizers map well to the two basic objects in scikit-learn:

- **Transformers** take input data and return a new data set.
- **Estimators** are fit to training data and can make predictions.

The scikit-learn API is object oriented, and estimators and transformers are initialized with parameters by instantiating their class. Hyperparameters can also be set using the `set_attrs()` method and retrieved with the corresponding `get_attrs()` method. All scikit-learn estimators have a `fit(X, y=None)` method that accepts a two dimensional data array, `X`, and optionally a vector `y` of target values. The `fit()` method trains the estimator, making it ready to transform data or make predictions. Transformers have an associated `transform(X)` method that returns a new dataset, `Xprime` and models have a `predict(X)` method that returns a vector of predictions, `yhat`. Models also have a `score(X, y)` method that evaluate the performance of the model.

Visualizers interact with scikit-learn objects by intersecting with them at the methods defined above. Specifically, visualizers perform actions related to `fit()`, `transform()`, `predict()`, and `score()` then call a `draw()` method which initializes the underlying figure associated with the visualizer. The user calls the visualizer's `poof()` method, which in turn calls a `finalize()` method on the visualizer to draw legends, titles, etc. and then `poof()` renders the figure. The Visualizer API is therefore:

- `draw()`: add visual elements to the underlying axes object
- `finalize()`: prepare the figure for rendering, adding final touches such as legends, titles, axis labels, etc.
- `poof()`: render the figure for the user (or saves it to disk).

Creating a visualizer means defining a class that extends `Visualizer` or one of its subclasses, then implementing several of the methods described above. A barebones implementation is as follows:

```
import matplotlib.pyplot as plot

from yellowbrick.base import Visualizer

class MyVisualizer(Visualizer):

    def __init__(self, ax=None, **kwargs):
        super(MyVisualizer, self).__init__(ax, **kwargs)

    def fit(self, X, y=None):
        self.draw(X)
        return self

    def draw(self, X):
        if self.ax is None:
            self.ax = self.gca()

        self.ax.plot(X)

    def finalize(self):
        self.set_title("My Visualizer")
```

This simple visualizer simply draws a line graph for some input dataset X, intersecting with the scikit-learn API at the `fit()` method. A user would use this visualizer in the typical style:

```
visualizer = MyVisualizer()
visualizer.fit(X)
visualizer.poof()
```

Score visualizers work on the same principle but accept an additional required `model` argument. Score visualizers wrap the model (which can be either instantiated or uninstantiated) and then pass through all attributes and methods through to the underlying model, drawing where necessary.

Testing

The test package mirrors the yellowbrick package in structure and also contains several helper methods and base functionality. To add a test to your visualizer, find the corresponding file to add the test case, or create a new test file in the same place you added your code.

Visual tests are notoriously difficult to create — how do you test a visualization or figure? Moreover, testing scikit-learn models with real data can consume a lot of memory. Therefore the primary test you should create is simply to test your visualizer from end to end and make sure that no exceptions occur. To assist with

this, we have two primary helpers, `VisualTestCase` and `DatasetMixin`. Create your unittest as follows:

```
import pytest
from tests.base import VisualTestCase
from tests.dataset import DatasetMixin

class MyVisualizerTests(VisualTestCase, DatasetMixin):

    def test_my_visualizer(self):
        """
        Test MyVisualizer on a real dataset
        """
        # Load the data from the fixture
        dataset = self.load_data('occupancy')

        # Get the data
        X = dataset[[
            "temperature", "relative_humidity", "light", "CO2", "humidity"
        ]]
        y = dataset['occupancy'].astype(int)

        try:
            visualizer = MyVisualizer()
            visualizer.fit(X)
            visualizer.poof()
        except Exception as e:
            pytest.fail("my visualizer didn't work")
```

Tests can be run as follows:

```
$ make test
```

The Makefile uses the pytest runner and testing suite as well as the coverage library, so make sure you have those dependencies installed! The `DatasetMixin` also requires requests.py to fetch data from our Amazon S3 account.

Image Comparison Tests

Writing an image based comparison test is only a little more difficult than the simple testcase presented above. We have adapted matplotlib's image comparison test utility into an easy to use assert method : `self.assert_images_similar(visualizer)`

The main consideration is that you must specify the "baseline" , or expected, image in the `tests/`

baseline_images/ folder structure.

For example, create your unittest located in tests/test_regressor/test_myvisualizer.py as follows:

```
from tests.base import VisualTestCase
...
    def test_my_visualizer_output(self):
        ...
        visualizer = MyVisualizer()
        visualizer.fit(X)
        visualizer.poof()
        self.assert_images_similar(visualizer)
```

The first time this test is run, there will be no baseline image to compare against, so the test will fail. Copy the output images (in this case tests/actual_images/test_regressor/test_myvisualizer/test_my_visualizer_output.png) to the correct subdirectory of baseline_images tree in the source directory (in this case tests/baseline_images/test_regressor/test_myvisualizer/test_my_visualizer_output.png). Put this new file under source code revision control (with git add). When rerunning the tests, they should now pass.

We also have a helper script, tests/images.py to clean up and manage baseline images automatically. It is run using the python -m command to execute a module as main, and it takes as an argument the path to your *test file*. To copy the figures as above:

```
$ python -m tests.images tests/test_regressor/test_myvisualizer.py
```

This will move all related test images from actual_images to baseline_images on your behalf (note you'll have had to run the tests at least once to generate the images). You can also clean up images from both actual and baseline as follows:

```
$ python -m tests.images -C tests/test_regressor/test_myvisualizer.py
```

This is useful particularly if you're stuck trying to get an image comparison to work. For more information on the images helper script, use python -m tests.images --help.

Documentation

The initial documentation for your visualizer will be a well structured docstring. Yellowbrick uses Sphinx to build documentation, therefore docstrings should be written in reStructuredText in numpydoc format (similar to scikit-learn). The primary location of your docstring should be right under the class definition, here is an example:

```
class MyVisualizer(Visualizer):
    """
```

(下页继续)

(续上页)

```
    This initial section should describe the visualizer and what
    it's about, including how to use it. Take as many paragraphs
    as needed to get as much detail as possible.

    In the next section describe the parameters to __init__.

    Parameters
    ----------

    model : a scikit-learn regressor
        Should be an instance of a regressor, and specifically one whose name
        ends with "CV" otherwise a will raise a YellowbrickTypeError exception
        on instantiation. To use non-CV regressors see:
        ``ManualAlphaSelection``.

    ax : matplotlib Axes, default: None
        The axes to plot the figure on. If None is passed in the current axes
        will be used (or generated if required).

    kwargs : dict
        Keyword arguments that are passed to the base class and may influence
        the visualization as defined in other Visualizers.

    Examples
    --------

    >>> model = MyVisualizer()
    >>> model.fit(X)
    >>> model.poof()

    Notes
    -----

    In the notes section specify any gotchas or other info.
    """
```

When your visualizer is added to the API section of the documentation, this docstring will be rendered in HTML to show the various options and functionality of your visualizer!

To add the visualizer to the documentation it needs to be added to the `docs/api` folder in the correct subdirectory. For example if your visualizer is a model score visualizer related to regression it would go in

the `docs/api/regressor` subdirectory. If you have a question where your documentation should be located, please ask the maintainers via your pull request, we'd be happy to help!

There are two primary files that need to be created:

1. **mymodule.rst**: the reStructuredText document
2. **mymodule.py**: a python file that generates images for the rst document

There are quite a few examples in the documentation on which you can base your files of similar types. The primary format for the API section is as follows:

```
.. -*- mode: rst -*-

My Visualizer
=============

Intro to my visualizer

.. code:: python

    # Example to run MyVisualizer
    visualizer = MyVisualizer(LinearRegression())

    visualizer.fit(X, y)
    g = visualizer.poof()


.. image:: images/my_visualizer.png

Discussion about my visualizer


API Reference
-------------

.. automodule:: yellowbrick.regressor.mymodule
    :members: MyVisualizer
    :undoc-members:
    :show-inheritance:
```

This is a pretty good structure for a documentation page; a brief introduction followed by a code example with a visualization included (using the `mymodule.py` to generate the images into the local directory's `images` subdirectory). The primary section is wrapped up with a discussion about how to interpret the visualizer

and use it in practice. Finally the `API Reference` section will use `automodule` to include the documentation from your docstring.

At this point there are several places where you can list your visualizer, but to ensure it is included in the documentation it *must be listed in the TOC of the local index*. Find the `index.rst` file in your subdirectory and add your rst file (without the `.rst` extension) to the `..toctree::` directive. This will ensure the documentation is included when it is built.

Speaking of, you can build your documentation by changing into the `docs` directory and running `make html`, the documentation will be built and rendered in the `_build/html` directory. You can view it by opening `_build/html/index.html` then navigating to your documentation in the browser.

There are several other places that you can list your visualizer including:

- `docs/index.rst` for a high level overview of our visualizers
- `DESCRIPTION.rst` for inclusion on PyPI
- `README.md` for inclusion on GitHub

Please ask for the maintainer's advice about how to include your visualizer in these pages.

4.5.3 Advanced Development

In this section we discuss more advanced contributing guidelines including setting up branches for development as well as the release cycle. This section is intended for maintainers and core contributors of the Yellowbrick project. If you would like to be a maintainer please contact one of the current maintainers of the project.

Branching Convention

The Yellowbrick repository is set up in a typical production/release/development cycle as described in "A Successful Git Branching Model." The primary working branch is the `develop` branch. This should be the branch that you are working on and from, since this has all the latest code. The `master` branch contains the latest stable version and release, which is pushed to PyPI. No one but core contributors will generally push to master.

注解: All pull requests should be into the `yellowbrick/develop` branch from your forked repository.

You can work directly in your fork and create a pull request from your fork's develop branch into ours. We also recommend setting up an `upstream` remote so that you can easily pull the latest development changes from the main Yellowbrick repository (see configuring a remote for a fork). You can do that as follows:

```
$ git remote add upstream https://github.com/DistrictDataLabs/yellowbrick.git
$ git remote -v
origin    https://github.com/YOUR_USERNAME/YOUR_FORK.git (fetch)
origin    https://github.com/YOUR_USERNAME/YOUR_FORK.git (push)
upstream  https://github.com/DistrictDataLabs/yellowbrick.git (fetch)
upstream  https://github.com/DistrictDataLabs/yellowbrick.git (push)
```

When you're ready, request a code review for your pull request. Then, when reviewed and approved, you can merge your fork into our main branch. Make sure to use the "Squash and Merge" option in order to create a Git history that is understandable.

注解: When merging a pull request, use the "squash and merge" option.

Core contributors have write access to the repository. In order to reduce the number of merges (and merge conflicts) we recommend that you utilize a feature branch off of develop to do intermediate work in:

```
$ git checkout -b feature-myfeature develop
```

Once you are done working (and everything is tested) merge your feature into develop.:

```
$ git checkout develop
$ git merge --no-ff feature-myfeature
$ git branch -d feature-myfeature
$ git push origin develop
```

Head back to Waffle and checkout another issue!

Releases

When ready to create a new release we branch off of develop as follows:

```
$ git checkout -b release-x.x
```

This creates a release branch for version x.x. At this point do the version bump by modifying `version.py` and the test version in `tests/__init__.py`. Make sure all tests pass for the release and that the documentation is up to date. There may be style changes or deployment options that have to be done at this phase in the release branch. At this phase you'll also modify the `changelog` with the features and changes in the release.

Once the release is ready for prime-time, merge into master:

```
$ git checkout master
$ git merge --no-ff --no-edit release-x.x
```

Tag the release in GitHub:

```
$ git tag -a vx.x
$ git push origin vx.x
```

You'll have to go to the release page to edit the release with similar information as added to the changelog. Once done, push the release to PyPI:

```
$ make build
$ make deploy
```

Check that the PyPI page is updated with the correct version and that `pip install -U yellowbrick` updates the version and works correctly. Also check the documentation on PyHosted, ReadTheDocs, and on our website to make sure that it was correctly updated. Finally merge the release into develop and clean up:

```
$ git checkout develop
$ git merge --no-ff --no-edit release-x.x
$ git branch -d release-x.x
```

Hotfixes and minor releases also follow a similar pattern; the goal is to effectively get new code to users as soon as possible!

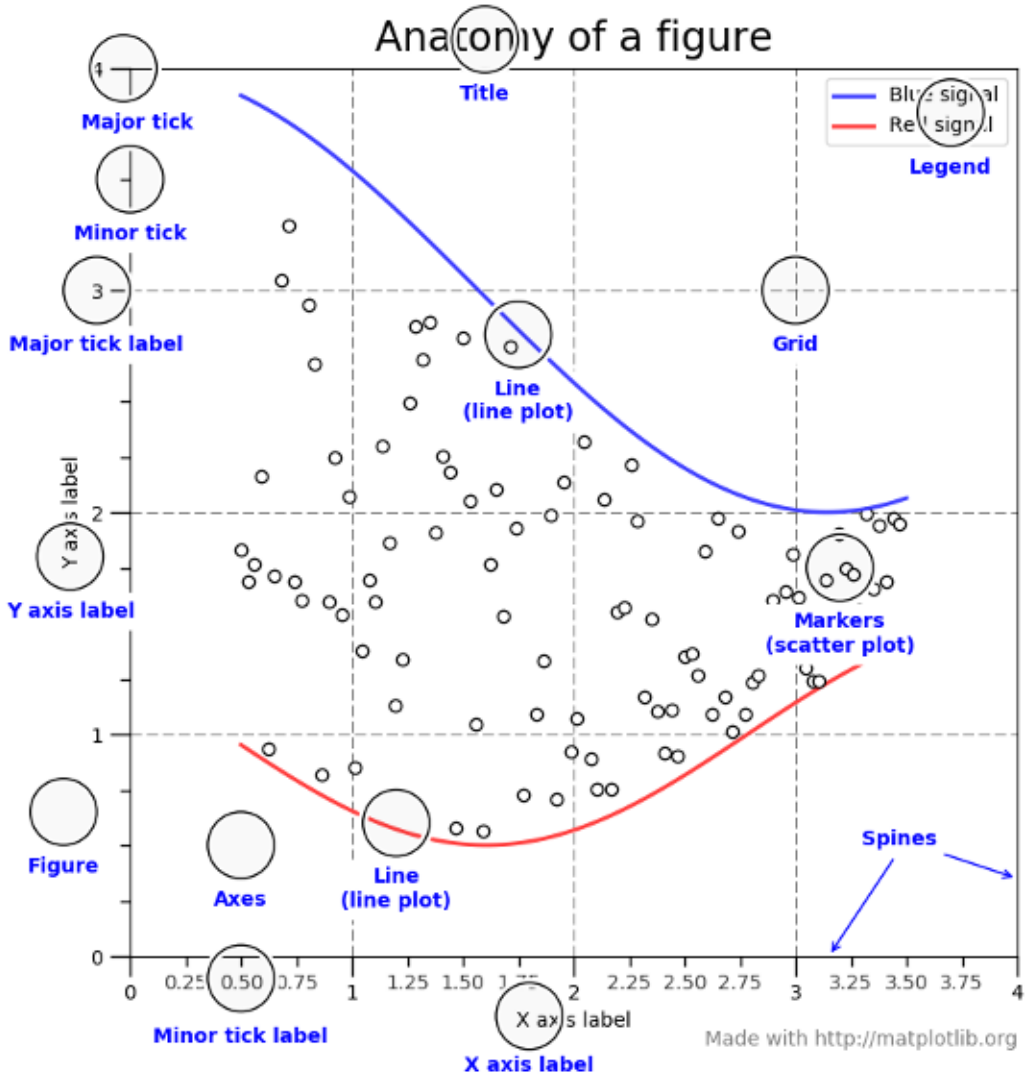
4.6 Effective Matplotlib

Yellowbrick generates visualizations by wrapping matplotlib, the most prominent Python scientific visualization library. Because of this, Yellowbrick is able to generate publication-ready images for a variety of GUI backends, image formats, and Jupyter notebooks. Yellowbrick strives to provide well-styled visual diagnostic tools and complete information. However, to customize figures or roll your own visualizers, a strong background in using matplotlib is required.

With permission, we have included part of Chris Moffitt's Effectively Using Matplotlib as a crash course into Matplotlib terminology and usage. For a complete example, please visit his excellent post on creating a visual sales analysis! Additionally we recommend Nicolas P. Rougier's Matplotlib tutorial for an in-depth dive.

4.6.1 Figures and Axes

This graphic from the matplotlib faq is gold. Keep it handy to understand the different terminology of a plot.



Most of the terms are straightforward but the main thing to remember is that the `Figure` is the final image that may contain 1 or more axes. The `Axes` represent an individual plot. Once you understand what these are and how to access them through the object oriented API, the rest of the process starts to fall into place.

The other benefit of this knowledge is that you have a starting point when you see things on the web. If you take the time to understand this point, the rest of the matplotlib API will start to make sense.

Matplotlib keeps a global reference to the global figure and axes objects which can be modified by the `pyplot` API. To access this import matplotlib as follows:

```
import matplotlib.pyplot as plt

axes = plt.gca()
```

The `plt.gca()` function gets the current axes so that you can draw on it directly. You can also directly create a figure and axes as follows:

```
fig = plt.figure()
ax = fig.add_subplot(111)
```

Yellowbrick will use `plt.gca()` by default to draw on. You can access the `Axes` object on a visualizer via its `ax` property:

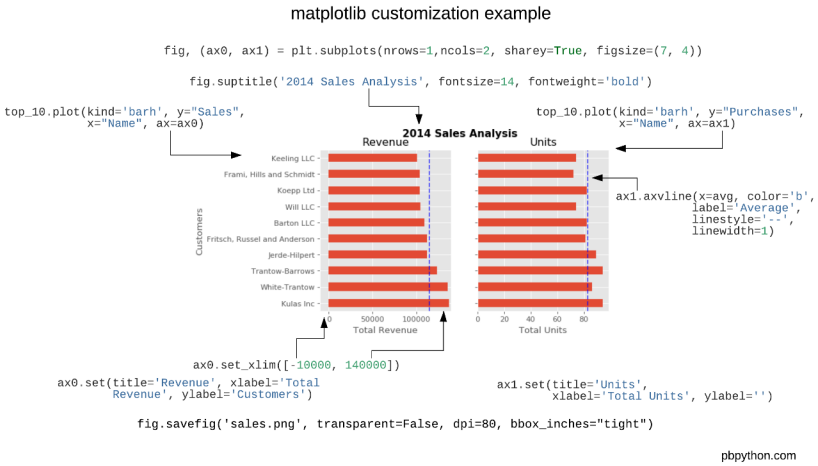
```
from sklearn.linear_model import LinearRegression
from yellowbrick.regressor import PredictionError

# Fit the visualizer
model = PredictionError(LinearRegression() )
model.fit(X_train, y_train)
model.score(X_test, y_test)

# Call finalize to draw the final yellowbrick-specific elements
model.finalize()

# Get access to the axes object and modify labels
model.ax.set_xlabel("measured concrete strength")
model.ax.set_ylabel("predicted concrete strength")
plt.savefig("peplot.pdf")
```

You can also pass an external `Axes` object directly to the visualizer:

```
model = PredictionError(LinearRegression(), ax=ax)
```

Therefore you have complete control of the style and customization of a Yellowbrick visualizer.

4.6.2 Creating a Custom Plot



The first step with any visualization is to plot the data. Often the simplest way to do this is using the standard pandas plotting function (given a `DataFrame` called `top_10`):

```
top_10.plot(kind='barh', y="Sales", x="Name")
```

The reason I recommend using pandas plotting first is that it is a quick and easy way to prototype your visualization. Since most people are probably already doing some level of data manipulation/analysis in pandas as a first step, go ahead and use the basic plots to get started.

Assuming you are comfortable with the gist of this plot, the next step is to customize it. Some of the customizations (like adding titles and labels) are very simple to use with the pandas plot function. However, you will probably find yourself needing to move outside of that functionality at some point. That's why it is recommended to create your own `Axes` first and pass it to the plotting function in Pandas:

```
fig, ax = plt.subplots()
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
```

The resulting plot looks exactly the same as the original but we added an additional call to `plt.subplots()` and passed the `ax` to the plotting function. Why should you do this? Remember when I said it is critical to get access to the axes and figures in matplotlib? That' s what we have accomplished here. Any future customization will be done via the `ax` or `fig` objects.

We have the benefit of a quick plot from pandas but access to all the power from matplotlib now. An example should show what we can do now. Also, by using this naming convention, it is fairly straightforward to adapt others' solutions to your unique needs.

Suppose we want to tweak the x limits and change some axis labels? Now that we have the axes in the `ax` variable, we have a lot of control:

```
fig, ax = plt.subplots()
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
ax.set_xlim([-10000, 140000])
ax.set_xlabel('Total Revenue')
ax.set_ylabel('Customer');
```

Here' s another shortcut we can use to change the title and both labels:

```
fig, ax = plt.subplots()
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
ax.set_xlim([-10000, 140000])
ax.set(title='2014 Revenue', xlabel='Total Revenue', ylabel='Customer')
```

To further demonstrate this approach, we can also adjust the size of this image. By using the `plt.subplots()` function, we can define the `figsize` in inches. We can also remove the legend using `ax.legend().set_visible(False)`:

```
fig, ax = plt.subplots(figsize=(5, 6))
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
ax.set_xlim([-10000, 140000])
ax.set(title='2014 Revenue', xlabel='Total Revenue')
ax.legend().set_visible(False)
```

There are plenty of things you probably want to do to clean up this plot. One of the biggest eye sores is the formatting of the Total Revenue numbers. Matplotlib can help us with this through the use of the `FuncFormatter` . This versatile function can apply a user defined function to a value and return a nicely formatted string to place on the axis.

Here is a currency formatting function to gracefully handle US dollars in the several hundred thousand dollar range:

```
def currency(x, pos):
    """
    The two args are the value and tick position
    """
    if x >= 1000000:
        return '${:1.1f}M'.format(x*1e-6)
    return '${:1.0f}K'.format(x*1e-3)
```

Now that we have a formatter function, we need to define it and apply it to the x axis. Here is the full code:

```
fig, ax = plt.subplots()
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
ax.set_xlim([-10000, 140000])
ax.set(title='2014 Revenue', xlabel='Total Revenue', ylabel='Customer')
formatter = FuncFormatter(currency)
ax.xaxis.set_major_formatter(formatter)
ax.legend().set_visible(False)
```

That' s much nicer and shows a good example of the flexibility to define your own solution to the problem.

The final customization feature I will go through is the ability to add annotations to the plot. In order to draw a vertical line, you can use `ax.axvline()` and to add custom text, you can use `ax.text()`.

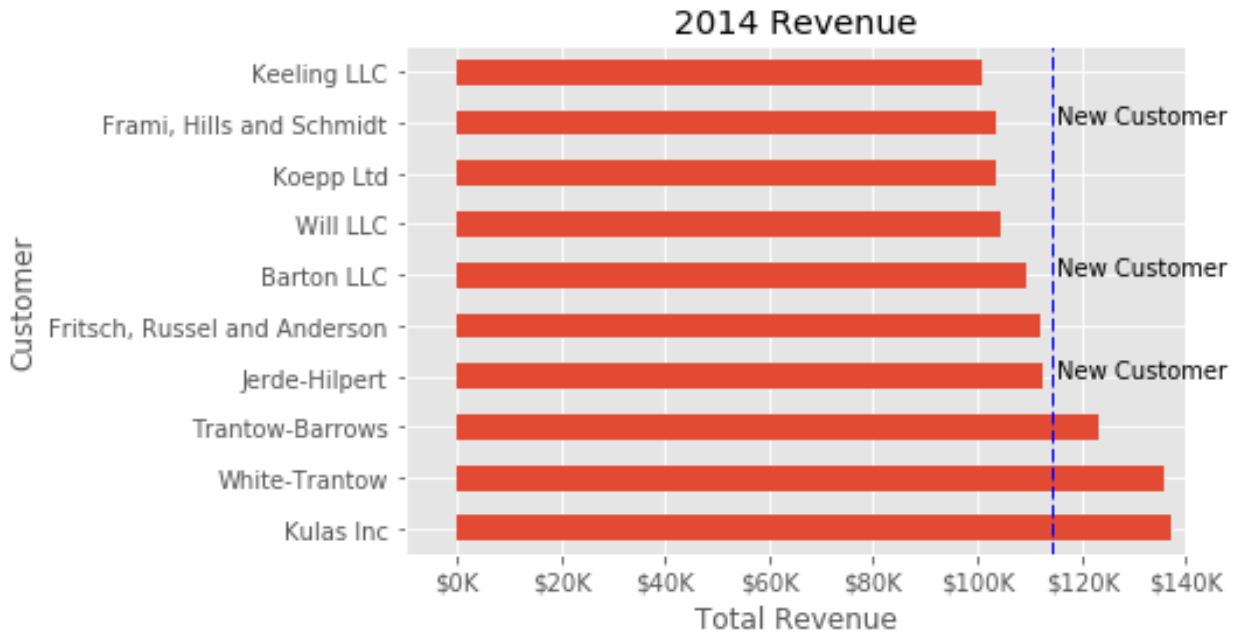
For this example, we' ll draw a line showing an average and include labels showing three new customers. Here is the full code with comments to pull it all together.

```
# Create the figure and the axes
fig, ax = plt.subplots()
```

(下页继续)

(续上页)

```
# Plot the data and get the average
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax)
avg = top_10['Sales'].mean()

# Set limits and labels
ax.set_xlim([-10000, 140000])
ax.set(title='2014 Revenue', xlabel='Total Revenue', ylabel='Customer')

# Add a line for the average
ax.axvline(x=avg, color='b', label='Average', linestyle='--', linewidth=1)

# Annotate the new customers
for cust in [3, 5, 8]:
ax.text(115000, cust, "New Customer")

# Format the currency
formatter = FuncFormatter(currency)
ax.xaxis.set_major_formatter(formatter)

# Hide the legend
ax.legend().set_visible(False)
```



While this may not be the most exciting plot it does show how much power you have when following this approach.

Up until now, all the changes we have made have been with the individual plot. Fortunately, we also have the ability to add multiple plots on a figure as well as save the entire figure using various options.

If we decided that we wanted to put two plots on the same figure, we should have a basic understanding of how to do it. First, create the figure, then the axes, then plot it all together. We can accomplish this using `plt.subplots()`:

```
fig, (ax0, ax1) = plt.subplots(nrows=1, ncols=2, sharey=True, figsize=(7, 4))
```

In this example, I' m using `nrows` and `ncols` to specify the size because this is very clear to the new user. In sample code you will frequently just see variables like 1,2. I think using the named parameters is a little easier to interpret later on when you' re looking at your code.

I am also using `sharey=True` so that the y-axis will share the same labels.

This example is also kind of nifty because the various axes get unpacked to `ax0` and `ax1`. Now that we have these axes, you can plot them like the examples above but put one plot on `ax0` and the other on `ax1`.

```
# Get the figure and the axes
fig, (ax0, ax1) = plt.subplots(nrows=1,ncols=2, sharey=True, figsize=(7, 4))
top_10.plot(kind='barh', y="Sales", x="Name", ax=ax0)
ax0.set_xlim([-10000, 140000])
ax0.set(title='Revenue', xlabel='Total Revenue', ylabel='Customers')

# Plot the average as a vertical line
avg = top_10['Sales'].mean()
ax0.axvline(x=avg, color='b', label='Average', linestyle='--', linewidth=1)

# Repeat for the unit plot
top_10.plot(kind='barh', y="Purchases", x="Name", ax=ax1)
avg = top_10['Purchases'].mean()
ax1.set(title='Units', xlabel='Total Units', ylabel='')
ax1.axvline(x=avg, color='b', label='Average', linestyle='--', linewidth=1)

# Title the figure
fig.suptitle('2014 Sales Analysis', fontsize=14, fontweight='bold');

# Hide the legends
ax1.legend().set_visible(False)
ax0.legend().set_visible(False)
```

When writing code in a Jupyter notebook you can take advantage of the `%matplotlib inline` or `%matplotlib notebook` directives to render figures inline. More often, however, you probably want to save your images to disk. Matplotlib supports many different formats for saving files. You can use `fig.`

canvas.get_supported_filetypes() to see what your system supports:

```
fig.canvas.get_supported_filetypes()
```

```
{'eps': 'Encapsulated Postscript',
 'jpeg': 'Joint Photographic Experts Group',
 'jpg': 'Joint Photographic Experts Group',
 'pdf': 'Portable Document Format',
 'pgf': 'PGF code for LaTeX',
 'png': 'Portable Network Graphics',
 'ps': 'Postscript',
 'raw': 'Raw RGBA bitmap',
 'rgba': 'Raw RGBA bitmap',
 'svg': 'Scalable Vector Graphics',
 'svgz': 'Scalable Vector Graphics',
 'tif': 'Tagged Image File Format',
 'tiff': 'Tagged Image File Format'}
```

Since we have the fig object, we can save the figure using multiple options:

```
fig.savefig('sales.png', transparent=False, dpi=80, bbox_inches="tight")
```

This version saves the plot as a png with opaque background. I have also specified the dpi and bbox_inches=”tight” in order to minimize excess white space.

4.7 About

Image by QuatroCinco, used with permission, Flickr Creative Commons.

Yellowbrick is an open source, pure Python project that extends the Scikit-Learn API with visual analysis and diagnostic tools. The Yellowbrick API also wraps Matplotlib to create publication-ready figures and interactive data explorations while still allowing developers fine-grain control of figures. For users, Yellowbrick can help evaluate the performance, stability, and predictive value of machine learning models, and assist in diagnosing problems throughout the machine learning workflow.

Recently, much of this workflow has been automated through grid search methods, standardized APIs, and GUI-based applications. In practice, however, human intuition and guidance can more effectively hone in on quality models than exhaustive search. By visualizing the model selection process, data scientists can steer towards final, explainable models and avoid pitfalls and traps.

The Yellowbrick library is a diagnostic visualization platform for machine learning that allows data scientists to steer the model selection process. Yellowbrick extends the Scikit-Learn API with a new core object: the Visualizer. Visualizers allow visual models to be fit and transformed as part of the Scikit-Learn Pipeline process, providing visual diagnostics throughout the transformation of high dimensional data.

4.7.1 Model Selection

Discussions of machine learning are frequently characterized by a singular focus on model selection. Be it logistic regression, random forests, Bayesian methods, or artificial neural networks, machine learning practitioners are often quick to express their preference. The reason for this is mostly historical. Though modern third-party machine learning libraries have made the deployment of multiple models appear nearly trivial, traditionally the application and tuning of even one of these algorithms required many years of study. As a result, machine learning practitioners tended to have strong preferences for particular (and likely more familiar) models over others.

However, model selection is a bit more nuanced than simply picking the "right" or "wrong" algorithm. In practice, the workflow includes:

1. selecting and/or engineering the smallest and most predictive feature set
2. choosing a set of algorithms from a model family, and
3. tuning the algorithm hyperparameters to optimize performance.

The **model selection triple** was first described in a 2015 SIGMOD paper by Kumar et al. In their paper, which concerns the development of next-generation database systems built to anticipate predictive modeling, the authors cogently express that such systems are badly needed due to the highly experimental nature of machine learning in practice. "Model selection," they explain, "is iterative and exploratory because the space of [model selection triples] is usually infinite, and it is generally impossible for analysts to know a priori which [combination] will yield satisfactory accuracy and/or insights."

Recently, much of this workflow has been automated through grid search methods, standardized APIs, and GUI-based applications. In practice, however, human intuition and guidance can more effectively hone in on quality models than exhaustive search. By visualizing the model selection process, data scientists can steer towards final, explainable models and avoid pitfalls and traps.

The Yellowbrick library is a diagnostic visualization platform for machine learning that allows data scientists to steer the model selection process. Yellowbrick extends the Scikit-Learn API with a new core object: the Visualizer. Visualizers allow visual models to be fit and transformed as part of the Scikit-Learn Pipeline process, providing visual diagnostics throughout the transformation of high dimensional data.

4.7.2 Name Origin

The Yellowbrick package gets its name from the fictional element in the 1900 children's novel **The Wonderful Wizard of Oz** by American author L. Frank Baum. In the book, the yellow brick road is the path that the protagonist, Dorothy Gale, must travel in order to reach her destination in the Emerald City.

From Wikipedia: "The road is first introduced in the third chapter of The Wonderful Wizard of Oz. The road begins in the heart of the eastern quadrant called Munchkin Country in the Land of Oz. It functions as a guideline that leads all who follow it, to the road's ultimate destination—the imperial capital of Oz called Emerald City that is located in the exact center of the entire continent. In the

book, the novel's main protagonist, Dorothy, is forced to search for the road before she can begin her quest to seek the Wizard. This is because the cyclone from Kansas did not release her farmhouse closely near it as it did in the various film adaptations. After the council with the native Munchkins and their dear friend the Good Witch of the North, Dorothy begins looking for it and sees many pathways and roads nearby, (all of which lead in various directions). Thankfully it doesn't take her too long to spot the one paved with bright yellow bricks."

4.7.3 Team

Yellowbrick is is developed by data scientists who believe in open source and the project enjoys contributions from Python developers all over the world. The project was started by @rebeccabilbro and @bbengfort as an attempt to better explain machine learning concepts to their students; they quickly realized, however, that the potential for visual steering could have a large impact on practical data science and developed it into a high-level Python library.

Yellowbrick is incubated by District Data Labs, an organization that is dedicated to collaboration and open source development. As part of District Data Labs, Yellowbrick was first introduced to the Python Community at PyCon 2016 in both talks and during the development sprints. The project was then carried on through DDL Research Labs (semester-long sprints where members of the DDL community contribute to various data related projects).

4.7.4 License

Yellowbrick is an open source project and its license is an implementation of the FOSS Apache 2.0 license by the Apache Software Foundation. In plain English this means that you can use Yellowbrick for commercial purposes, modify and distribute the source code, and even sublicense it. We want you to use Yellowbrick, profit from it, and contribute back if you do cool things with it.

There are, however, a couple of requirements that we ask from you. First, when you copy or distribute Yellowbrick source code, please include our copyright and license found in the LICENSE.txt at the root of our software repository. In addition, if we create a file called "NOTICE" in our project you must also include that in your source distribution. The "NOTICE" file will include attribution and thanks to those who have worked so hard on the project! Finally you can't hold District Data Labs or any Yellowbrick contributor liable for your use of our software, nor use any of our names, trademarks, or logos.

We think that's a pretty fair deal, and we're big believers in open source. If you make any changes to our software, use it commercially or academically, or have any other interest, we'd love to hear about it.

4.7.5 Presentations

Yellowbrick has enjoyed the spotlight at a few conferences and in several presentations. We hope that these videos, talks, and slides will help you understand Yellowbrick a bit better.

Videos:

- Visual Diagnostics for More Informed Machine Learning: Within and Beyond Scikit-Learn (PyCon 2016)
- Visual Diagnostics for More Informed Machine Learning (PyData Carolinas 2016)
- Yellowbrick: Steering Machine Learning with Visual Transformers (PyData London 2017)

Slides:

- Visualizing the Model Selection Process
- Visualizing Model Selection with Scikit-Yellowbrick
- Visual Pipelines for Text Analysis (Data Intelligence 2017)

4.8 Changelog

4.8.1 Version 0.5

- Tag: v0.5
- Deployed: Wednesday, August 9, 2017
- Contributors: Benjamin Bengfort, Rebecca Bilbro, Nathan Danielsen, Carlo Morales, Jim Stearns, Phillip Schafer, Jason Keung

Changes:

- Added VisualTestCase.
- New PCADecomposition Visualizer, which decomposes high dimensional data into two or three dimensions so that each instance can be plotted in a scatter plot.
- New and improved ROCAUC Visualizer, which now supports multiclass classification.
- Prototype Decision Boundary Visualizer, which is a bivariate data visualization algorithm that plots the decision boundaries of each class.
- Added Rank1D Visualizer, which is a one dimensional ranking of features that utilizes the Shapiro-Wilks ranking that takes into account only a single feature at a time (e.g. histogram analysis).
- Improved Prediction Error Plot with identity line, shared limits, and r squared.
- Updated FreqDist Visualizer to make word features a hyperparameter.
- Added normalization and scaling to Parallel Coordinates.
- Added Learning Curve Visualizer, which displays a learning curve based on the number of samples versus the training and cross validation scores to show how a model learns and improves with experience.

- Added data downloader module to the yellowbrick library.
- Complete overhaul of the yellowbrick documentation; categories of methods are located in separate pages to make it easier to read and contribute to the documentation.
- Added a new color palette inspired by ANN-generated colors

Bug Fixes:

- Repairs to PCA, RadViz, FreqDist unit tests
- Repair to matplotlib version check in JointPlot Visualizer

4.8.2 Hotfix 0.4.2

Update to the deployment docs and package on both Anaconda and PyPI.

- Tag: v0.4.2
- Deployed: Monday, May 22, 2017
- Contributors: Benjamin Bengfort, Jason Keung

4.8.3 Version 0.4.1

This release is an intermediate version bump in anticipation of the PyCon 2017 sprints.

The primary goals of this version were to (1) update the Yellowbrick dependencies (2) enhance the Yellowbrick documentation to help orient new users and contributors, and (3) make several small additions and upgrades (e.g. pulling the Yellowbrick utils into a standalone module).

We have updated the Scikit-Learn and SciPy dependencies from version 0.17.1 or later to 0.18 or later. This primarily entails moving from `from sklearn.cross_validation import train_test_split` to `from sklearn.model_selection import train_test_split`.

The updates to the documentation include new Quickstart and Installation guides as well as updates to the Contributors documentation, which is modeled on the Scikit-Learn contributing documentation.

This version also included upgrades to the KMeans visualizer, which now supports not only `silhouette_score` but also `distortion_score` and `calinski_harabaz_score`. The `distortion_score` computes the mean distortion of all samples as the sum of the squared distances between each observation and its closest centroid. This is the metric that K-Means attempts to minimize as it is fitting the model. The `calinski_harabaz_score` is defined as ratio between the within-cluster dispersion and the between-cluster dispersion.

Finally, this release includes a prototype of the `VisualPipeline`, which extends Scikit-Learn's `Pipeline` class, allowing multiple Visualizers to be chained or sequenced together.

- Tag: v0.4.1

- Deployed: Monday, May 22, 2017
- Contributors: Benjamin Bengfort, Rebecca Bilbro, Nathan Danielsen

Changes:

- Score and model visualizers now wrap estimators as proxies so that all methods on the estimator can be directly accessed from the visualizer
- Updated Scikit-learn dependency from >=0.17.1 to >=0.18
- Replaced `sklearn.cross_validation` with `model_selection`
- Updated SciPy dependency from >=0.17.1 to >=0.18
- ScoreVisualizer now subclasses ModelVisualizer; towards allowing both fitted and unfitted models passed to Visualizers
- Added CI tests for Python 3.6 compatibility
- Added new quickstart guide and install instructions
- Updates to the contributors documentation
- Added `distortion_score` and `calinski_harabaz_score` computations and visualizations to KMeans visualizer.
- Replaced the `self.ax` property on all of the individual `draw` methods with a new property on the `Visualizer` class that ensures all visualizers automatically have axes.
- Refactored the utils module into a package
- Continuing to update the docstrings to conform to Sphinx
- Added a prototype visual pipeline class that extends the Scikit-learn pipeline class to ensure that visualizers get called correctly.

Bug Fixes:

- Fixed title bug in Rank2D FeatureVisualizer

4.8.4 Version 0.4

This release is the culmination of the Spring 2017 DDL Research Labs that focused on developing Yellowbrick as a community effort guided by a sprint/agile workflow. We added several more visualizers, did a lot of user testing and bug fixes, updated the documentation, and generally discovered how best to make Yellowbrick a friendly project to contribute to.

Notable in this release is the inclusion of two new feature visualizers that use few, simple dimensions to visualize features against the target. The `JointPlotVisualizer` graphs a scatter plot of two dimensions in the data set and plots a best fit line across it. The `ScatterVisualizer` also uses two features, but also colors the graph by the target variable, adding a third dimension to the visualization.

This release also adds support for clustering visualizations, namely the elbow method for selecting K, `KElbowVisualizer` and a visualization of cluster size and density using the `SilhouetteVisualizer`. The release also adds support for regularization analysis using the `AlphaSelection` visualizer. Both the text and classification modules were also improved with the inclusion of the `PosTagVisualizer` and the `ConfusionMatrix` visualizer respectively.

This release also added an Anaconda repository and distribution so that users can `conda install` yellowbrick. Even more notable, we got yellowbrick stickers! We've also updated the documentation to make it more friendly and a bit more visual; fixing the API rendering errors. All-in-all, this was a big release with a lot of contributions and we thank everyone that participated in the lab!

- Tag: v0.4
- Deployed: Thursday, May 4, 2017
- Contributors: Benjamin Bengfort, Rebecca Bilbro, Nathan Danielsen, Matt Andersen, Prema Roman, Neal Humphrey, Jason Keung, Bala Venkatesan, Paul Witt, Morgan Mendis, Tuuli Morril

Changes:

- Part of speech tags visualizer – `PosTagVisualizer`.
- Alpha selection visualizer for regularized regression – `AlphaSelection`
- Confusion Matrix Visualizer – `ConfusionMatrix`
- Elbow method for selecting K vis – `KElbowVisualizer`
- Silhouette score cluster visualization – `SilhouetteVisualizer`
- Joint plot visualizer with best fit – `JointPlotVisualizer`
- Scatter visualization of features – `ScatterVisualizer`
- Added three more example datasets: mushroom, game, and bike share
- Contributor's documentation and style guide
- Maintainers listing and contacts
- Light/Dark background color selection utility
- Structured array detection utility
- Updated classification report to use colormesh
- Added anacondas packaging and distribution
- Refactoring of the regression, cluster, and classification modules
- Image based testing methodology
- Docstrings updated to a uniform style and rendering
- Submission of several more user studies

4.8.5 Version 0.3.3

Intermediate sprint to demonstrate prototype implementations of text visualizers for NLP models. Primary contributions were the `FreqDistVisualizer` and the `TSNEVisualizer`.

The `TSNEVisualizer` displays a projection of a vectorized corpus in two dimensions using TSNE, a nonlinear dimensionality reduction method that is particularly well suited to embedding in two or three dimensions for visualization as a scatter plot. TSNE is widely used in text analysis to show clusters or groups of documents or utterances and their relative proximities.

The `FreqDistVisualizer` implements frequency distribution plot that tells us the frequency of each vocabulary item in the text. In general, it could count any kind of observable event. It is a distribution because it tells us how the total number of word tokens in the text are distributed across the vocabulary items.

- Tag: v0.3.3
- Deployed: Wednesday, February 22, 2017
- Contributors: Rebecca Bilbro, Benjamin Bengfort

Changes:

- TSNEVisualizer for 2D projections of vectorized documents
- FreqDistVisualizer for token frequency of text in a corpus
- Added the user testing evaluation to the documentation
- Created scikit-yb.org and host documentation there with RFD
- Created a sample corpus and text examples notebook
- Created a base class for text, `TextVisualizer`
- Model selection tutorial using Mushroom Dataset
- Created a text examples notebook but have not added to documentation.

4.8.6 Version 0.3.2

Hardened the Yellowbrick API to elevate the idea of a Visualizer to a first principle. This included reconciling shifts in the development of the preliminary versions to the new API, formalizing Visualizer methods like *draw()* and *finalize()*, and adding utilities that revolve around Scikit-Learn. To that end we also performed administrative tasks like refreshing the documentation and preparing the repository for more and varied open source contributions.

- Tag: v0.3.2
- Deployed: Friday, January 20, 2017
- Contributors: Benjamin Bengfort, Rebecca Bilbro

Changes:

- Converted Mkdocs documentation to Sphinx documentation
- Updated docstrings for all Visualizers and functions
- Created a DataVisualizer base class for dataset visualization
- Single call functions for simple visualizer interaction
- Added yellowbrick specific color sequences and palettes and env handling
- More robust examples with downloader from DDL host
- Better axes handling in visualizer, matplotlib/sklearn integration
- Added a finalize method to complete drawing before render
- Improved testing on real data sets from examples
- Bugfix: score visualizer renders in notebook but not in Python scripts.
- Bugfix: tests updated to support new API

4.8.7 Hotfix 0.3.1

Hotfix to solve pip install issues with Yellowbrick.

- Tag: v0.3.1
- Deployed: Monday, October 10, 2016
- Contributors: Benjamin Bengfort

  Changes:

  - Modified packaging and wheel for Python 2.7 and 3.5 compatibility
  - Modified deployment to PyPI and pip install ability
  - Fixed Travis-CI tests with the backend failures.

4.8.8 Version 0.3

This release marks a major change from the previous MVP releases as Yellowbrick moves towards direct integration with Scikit-Learn for visual diagnostics and steering of machine learning and could therefore be considered the first alpha release of the library. To that end we have created a Visualizer model which extends sklearn.base.BaseEstimator and can be used directly in the ML Pipeline. There are a number of visualizers that can be used throughout the model selection process, including for feature analysis, model selection, and hyperparameter tuning.

In this release specifically we focused on visualizers in the data space for feature analysis and visualizers in the model space for scoring and evaluating models. Future releases will extend these base classes and add more functionality.

- Tag: v0.3
- Deployed: Sunday, October 9, 2016
- Contributors: Benjamin Bengfort, Rebecca Bilbro, Marius van Niekerk

  Enhancements:

  - Created an API for visualization with machine learning: Visualizers that are BaseEstimators.
  - Created a class hierarchy for Visualizers throughout the ML process particularly feature analysis and model evaluation
  - Visualizer interface is draw method which can be called multiple times on data or model spaces and a poof method to finalize the figure and display or save to disk.
  - ScoreVisualizers wrap Scikit-Learn estimators and implement fit and predict (pass-throughs to the estimator) and also score which calls draw in order to visually score the estimator. If the estimator isn't appropriate for the scoring method an exception is raised.
  - ROCAUC is a ScoreVisualizer that plots the receiver operating characteristic curve and displays the area under the curve score.
  - ClassificationReport is a ScoreVisualizer that renders the confusion matrix of a classifier as a heatmap.
  - PredictionError is a ScoreVisualizer that plots the actual vs. predicted values and the 45 degree accuracy line for regressors.
  - ResidualPlot is a ScoreVisualizer that plots the residuals (y - yhat) across the actual values (y) with the zero accuracy line for both train and test sets.
  - ClassBalance is a ScoreVisualizer that displays the support for each class as a bar plot.
  - FeatureVisualizers are Scikit-Learn Transformers that implement fit and transform and operate on the data space, calling draw to display instances.
  - ParallelCoordinates plots instances with class across each feature dimension as line segments across a horizontal space.
  - RadViz plots instances with class in a circular space where each feature dimension is an arc around the circumference and points are plotted relative to the weight of the feature.
  - Rank2D plots pairwise scores of features as a heatmap in the space [-1, 1] to show relative importance of features. Currently implemented ranking functions are Pearson correlation and covariance.

  - Coordinated and added palettes in the bgrmyck space and implemented a version of the Seaborn set_palette and set_color_codes functions as well as the ColorPalette object and other matplotlib.rc modifications.
  - Inherited Seaborn's notebook context and whitegrid axes style but make them the default, don't allow user to modify (if they'd like to, they'll have to import Seaborn). This gives Yellowbrick a consistent look and feel without giving too much work to the user and prepares us for Matplotlib 2.0.
  - Jupyter Notebook with Examples of all Visualizers and usage.

Bug Fixes:

  - Fixed Travis-CI test failures with matplotlib.use('Agg').
  - Fixed broken link to Quickstart on README
  - Refactor of the original API to the Scikit-Learn Visualizer API

4.8.9 Version 0.2

Intermediate steps towards a complete API for visualization. Preparatory stages for Scikit-Learn visual pipelines.

- Tag: v0.2
- Deployed: Sunday, September 4, 2016
- Contributors: Benjamin Bengfort, Rebecca Bilbro, Patrick O'Melveny, Ellen Lowy, Laura Lorenz

  Changes:

  - Continued attempts to fix the Travis-CI Scipy install failure (broken tests)
  - Utility function: get the name of the model
  - Specified a class based API and the basic interface (render, draw, fit, predict, score)
  - Added more documentation, converted to Sphinx, autodoc, docstrings for viz methods, and a quickstart
  - How to contribute documentation, repo images etc.
  - Prediction error plot for regressors (mvp)
  - Residuals plot for regressors (mvp)
  - Basic style settings a la seaborn
  - ROC/AUC plot for classifiers (mvp)
  - Best fit functions for "select best", linear, quadratic
  - Several Jupyter notebooks for examples and demonstrations

4.8.10 Version 0.1

Created the yellowbrick library MVP with two primary operations: a classification report heat map and a ROC/AUC curve model analysis for classifiers. This is the base package deployment for continuing yellowbrick development.

- Tag: v0.1
- Deployed: Wednesday, May 18, 2016
- Contributors: Benjamin Bengfort, Rebecca Bilbro

  Changes:

  - Created the Anscombe quartet visualization example
  - Added DDL specific color maps and a stub for more style handling
  - Created crplot which visualizes the confusion matrix of a classifier
  - Created rocplot_compare which compares two classifiers using ROC/AUC metrics
  - Stub tests/stub documentation

CHAPTER 5

索引和表格

- genindex
- modindex

翻译：Juan L. Kehoe

Python 模块索引

y

索引