目录

[Machine learning modeling framework 1](#_Toc81876)

[一． 数据观察 2](#_Toc81877)

[1.1 Data\_Explore模块 2](#_Toc81878)

[1.2 Data\_Distribution模块 3](#_Toc81879)

[1.3 Data\_Sampling模块 4](#_Toc81880)

[二． 数据处理 5](#_Toc81881)

[2.1 Missing\_Data\_Impute模块 5](#_Toc81882)

[2.2 Cut\_Merge模块 5](#_Toc81883)

[2.2.1 Chi2\_merge模块 6](#_Toc81884)

[2.2.2 Dummy模块 6](#_Toc81885)

[三． 特征筛选/降维 7](#_Toc81886)

[3.1.1 Corr\_Vif模块 7](#_Toc81887)

[3.1.2 PSI/Calc\_psi模块 8](#_Toc81888)

[3.2.1 Woe\_Iv模块 9](#_Toc81889)

[3.2.2 PDO\_Score模块 9](#_Toc81890)

[PDO\_Score [2]： 9](#_Toc81891)

[PDO\_Score\_Convert： 10](#_Toc81892)

[3.3.1 Feature\_Select\_From\_Ensemble模块 10](#_Toc81893)

[3.3.2 Feature\_Select\_From\_LR模块 10](#_Toc81894)

[四． 模型表现 10](#_Toc81895)

[4 Model\_Performance模块 10](#_Toc81896)

[五． 应用实例 12](#_Toc81897)

[1. 泰坦尼克号生存率预测建模 12](#_Toc81898)

Machine learning modeling framework

本工具箱全部基于Python3构建，代码中除画图调用scikitplot源包外，内容上无整段照搬或直接调用第三方包；所以执行中难免仍可能遇到代码bug、效率问题甚至逻辑错误，欢迎使用者多提意见和建议（Version1.0于20Dec2018）。

本工具箱的构建旨在为使用者简化建模流程中的冗长编程，其中每个模块独立，可供建模者依照建模流程拼接使用。当前版本基于4个阶段（数据观察、数据处理、建模和模型表现）进行展开，贯穿机器学习模型和评分卡模型两个条线。由于建模多可调用Sklearn或者Xgboost包实现，所以工具包本期较多笔墨在数据观察和数据处理两个部分，后期将会补全模型表现部分的其他图表。

通用建模流程如下（其中模块标目带x的未开发完成或使用位置不限）：



1. 数据观察

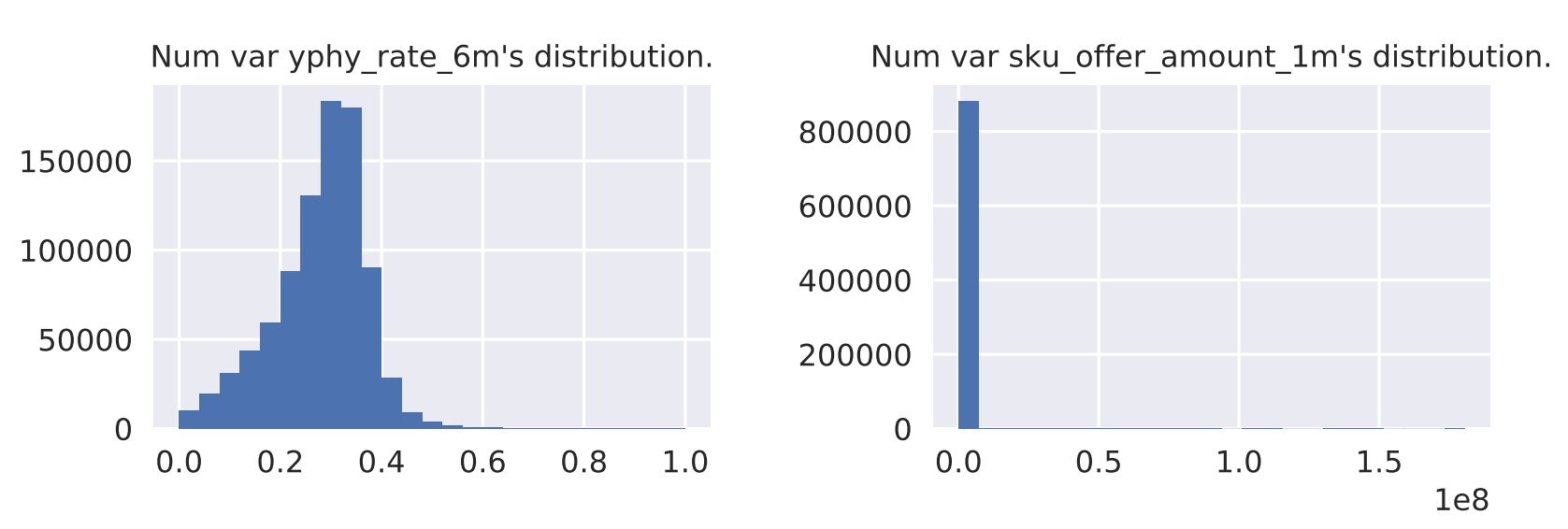
1.1 Data\_Explore模块

模块功能介绍：对输入数据集各字段数据质量进行固定模式的探查，包括：字段类型、最大最小值（字符型则为值宽度）、0值数/率、唯一值数/率、缺失值数/率、众数值/率、计算值数（按Mean±N\*Std划定）。

* 入参：
  + Datain（Dataframe/must）：需要统计描述的数据集
  + Ext\_n（Int/3）：按Mean±N\*Std方式划定极端值时的N
  + Plt\_out（Url/’’）：如定义，则将各个变量的频数分布图输出至路径
  + Plt\_show（Boolean/Fasle）：是否在log中输出图，少量特征适用，默认False
  + Char\_cate\_threshold（Int/10）：在绘制分类型变量的占比饼图数，设定类数上限，超过上限则跳过
  + Sug\_drop\_na\_rate（float/0.6）：标记出缺失率高于该阈值的特征
  + Sug\_drop\_ct\_rate（float/0.2）：标记出唯一值率高于该阈值的分类特征
  + sps（Int /4）：频数分布图按照4宫格或9宫格输出。
  + Calc\_obs\_na（Boolean/False）：计算观测缺失率（行缺失），并绘制观测缺失量占比图
  + Ecd（Str/’utf-8’）：encoding
* 返回：
  + Variable\_pre\_exp （Dataframe）：数据质量表
  + obs\_miss\_rate（Series）：行缺失率
* 结果举例：
  + Variable\_pre\_exp：



* + 数值型频数分布图、字符型饼图、相关性热力图：



1.2 Data\_Distribution模块

模块功能介绍：此为数据观察的子模块，用途为对连续性序列值按照指定规则切分后，各个箱内样本量情况。

该模块统计表格已可用，出图仍需调整。

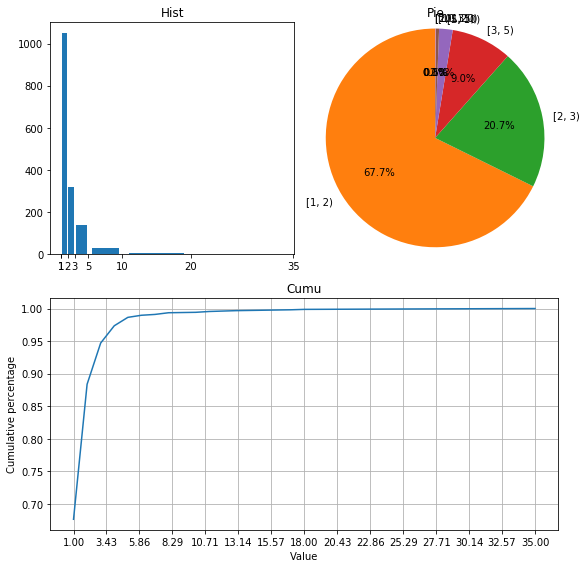
* 入参：
  + Serin（Series/must）：需要统计分布的连续性变量
  + bins（list/must）：手动给出分箱规则
  + close（str/’left’）：每个分箱边界开闭方式
* 返回：

例如调用函数为：Data\_Distribution (serin = datain\_agged['fmerchantid'], bins = [1, 2, 3, 5, 10, 20])

Table：返回表格中id为划分组数interval为区间，N为区间包含样本量，rate为占比，cumN为累计占比数，cunRate为累计占比率。



Figure：图片尚有优化空间。左上为每个区间的样本量，横轴按照区间距离标注；右上为每个区间的样本率；下图为按照最大最小值等宽划分15份后的样本量累计图。



1.3 Data\_Sampling模块

模块功能介绍：该模块对数据集进行抽样。从抽样意图可分为欠采样和过采样，过采样为随机简单复制。

从抽样方式可分为：

1. 单纯抽样：全部数据按照一定比例或指定数量进行抽样，如sample\_rule = 100; sample\_rule = 0.3；

2. 分层抽样：按照某个变量进行排序，按照等频分箱，每箱内抽取均匀样本量，如stratify\_by = ['age', 20] + sample\_rule = 100 or 0.3；

3. 非均衡抽样：

* 设定某个分组变量，对其包含值个数取最小，按此最小值对所有分组 1:1 抽样，如sample\_rule = 'align' + group\_by = 'survived'
* 指定规则分组抽样，按照分组变量的每个值设定规则，如sample\_rule = {0:100, 1:'max'} + group\_by = 'survived'
* 非均衡 + 分层抽样，如sample\_rule = {0:100, 1:0.3} + group\_by = 'survived' + stratify\_by = ['age', 20]
* 入参：
  + datain（Dataframe/must）：需要抽样的数据集
  + sample\_rule（int/float/dict/’align’）：
    1. int：按照定量抽取，例如：1000条
    2. float：按照比例抽取，例如：0.5，抽50%样本
    3. dict：按字典给出的规则抽取（需配合分组变量），例如：{0:'max',1:4000}，分组变量=0的情况全部保留，=1的情况抽4000条
    4. ‘align’：按照分组变量中每个组样本量最小值，进行所有组的样本量定义。例如一个数据集逾期1000例，非逾期10000例；则抽样时将全部逾期保留，非逾期抽1000例
  + stratify\_by（list/[]）：[需要分层抽样的特征名，箱数]
  + group\_by（str/’ ’）：分组抽样的特征名
* 返回：
  + sample\_out（Dataframe）：抽样好的数据集

1. 数据处理

2.1 Missing\_Data\_Impute模块

模块功能介绍：为缺失值填补模块，目前已开发的填补方法为：均数填补、中位数填补、众数填补（为字符型变量默认填补）、指定值填补。

另有文章介绍可用随机森林填补或kNN填补，随机森林填补经测试对模型效果提升不显著，已废弃；kNN模块待开发。

* 核心函数：pd.fillna
* 入参：
  + datain（Dataframe/must）：需要填补的数据集
  + Imp\_config（Str、float、dict/ 'imp\_mean'）：该参数用于指示对输入数据集的哪些变量采用何种填补方式。
    1. 如传入字符型，并限定于以下列表['imp\_mean'/'imp\_median'/'imp\_mode'/'imp\_knn']，则对所有有缺失变量进行填补，其中数值型变量按照如上填补方法进行填补、字符型变量按照众数填补。如imp\_config = 'imp\_median'
    2. 如传入数值型，则对所有有缺失变量进行填补，其中数值型变量按照传入值进行填补、字符型变量按照众数填补。如imp\_config = -999
    3. 如传入字典，则按照字典指示对每个变量单独填补，字典键值对为{需填补的值 : 需填补的变量}。

例如：Age和Sex填均值、Income填中位数、Debt填-1，Class填'Unknow'，则imp\_config = {'imp\_mean':['Age','Sex'], 'imp\_median':'Income', -1:'Debt', 'Unknow':'Class'}

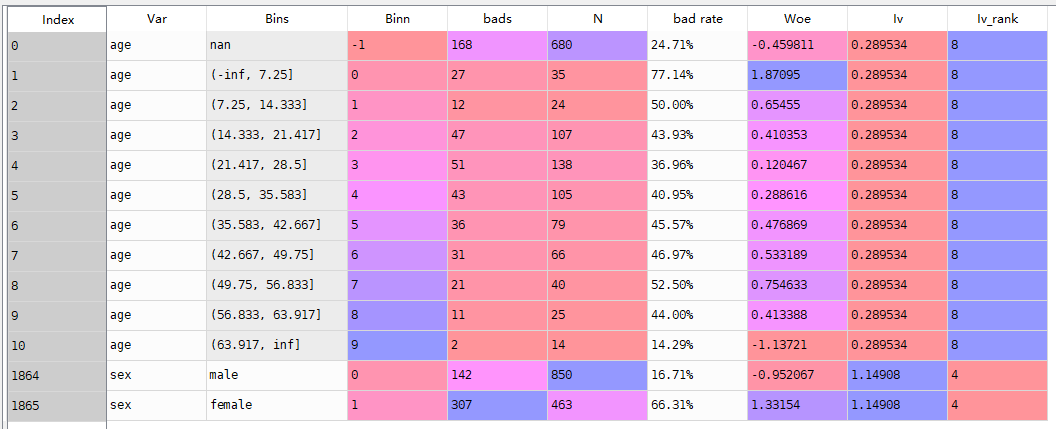
* + define\_key\_parameter（List/['imp\_mean', 'imp\_median', 'imp\_mode', 'imp\_knn']）：一般无需该动，如果你需要填补的字符型变量值正好例如'imp\_mean'，则可以改动。
* 返回：
  + imp\_dataout：填补后的数据集

2.2 Cut\_Merge模块

模块功能介绍：该模块功能为对输入数据集中特征进行简单分箱操作（等频、等宽分箱），其中数值型按照指定分箱方式进行分箱、字符型按照每个值给出分箱。已从逻辑上避免空箱现象，待完备点：① 需要识别每个箱中是否有坏样本；② 需增设手动分箱入参。

模块默认设置自动计算WoE，计算逻辑和公式请参照3.2.1模块。

* 核心函数：等频：quantile
* 入参：
  + datain（Dataframe、Series/must）：需进行分箱的数据
  + Cut\_way（Str/’ew’）：对数值型变量分箱方式（ef：等频、ew：等宽）
  + Cut\_y（Str/’’）：Y变量。用途：① 在对数据分箱时，可屏蔽误处理；② 在末尾输出分箱统计表时，可计算每个箱内坏样本量/率
  + Cut\_n（Int/3）：预分箱数。注意：若分箱后，组内出现空箱，等频和等宽过程采用不同处理方式。例如Age变量有50%缺失，均按照均数填补。① 等频分箱核心函数为：quantile，分箱规则中会出现大量的闭合箱（如：[3.14, 3.14]），等频分箱将闭合箱剔除；② 等宽分箱按照最大最小值等分切割，如数据分布不匀则会出现空箱，等宽分箱将目标分箱数减1后重新分箱。
  + Calc\_woe(Boolean/True)：连带计算WOE-IV，数据集直接以WoE转化后输出，WOE和IV计算逻辑详见2.3.1模块。
* 返回：
  + cut\_df（Dataframe/Series）：切割好后的数据集
  + fa\_df（Dataframe）：分箱情况统计集
* 结果举例：
  + fa\_df：



2.2.1 Chi2\_merge模块

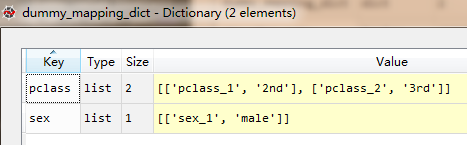
模块功能介绍：该模块为卡方分箱。因为分箱阈值设定存在问题，程序修改测试完毕后完成此模块介绍。

2.2.2 Dummy模块

模块功能介绍：该模块用于将数据集中字符型变量值哑变量化

* 入参：
  + datain（Dataframe、Series/must）：用于哑变量处理的数据集（可直接输入原数据集，结果会merge到原数据集中）
  + Dummy\_var（Str、List/[]）：指明哪些字段需要进行哑变量处理。如果不定义该参数，则认定将输入数据集中所有字符型变量哑变量化。
  + Drop\_one（Boolean/True）：是否将dummy后的第0个哑变量剔除，以防止模型共线性。
  + Drop\_orig（Boolean/True）：是否将原值剔除。
* 返回：
  + dummy\_mapping\_df：处理后的数据集
  + dummy\_mapping\_dict：特征哑变量化映射字典
* 结果举例：
  + dummy\_mapping\_df, dummy\_mapping\_dict = Dummy\_char(datain\_test[['pclass','survived','age','sex']])
  + 输入集、输出集和映射字典：



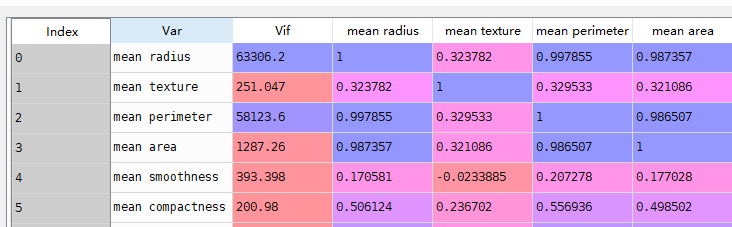
1. 特征筛选/降维

3.1.1 Corr\_Vif模块

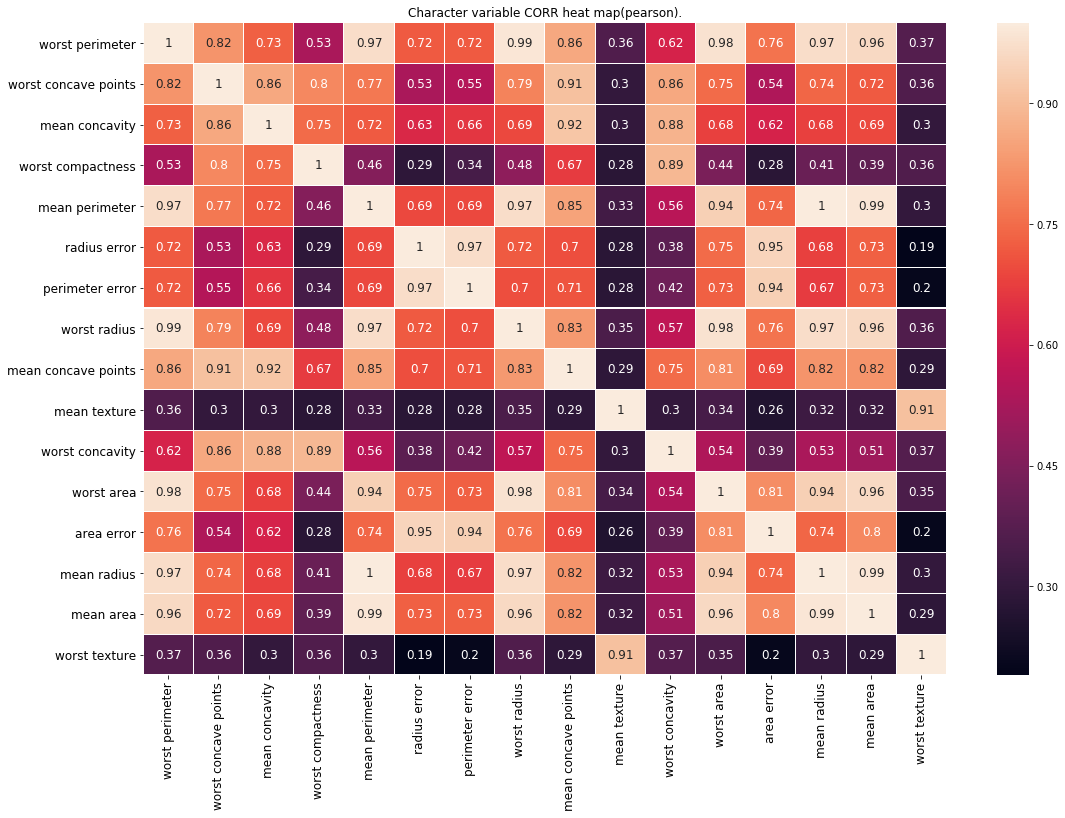
模块功能介绍：分别计算数据数据集各个特征之间的相关性和方差膨胀因子。

* 核心函数：pandas.corr、statsmodels.variance\_inflation\_factor
* 入参：
  + datain（DataFrame/must）：需要统计的数据集。
  + corr\_method（str/’pearson’）：特征相关性计算方式，可选：'pearson', 'kendall', 'spearman'
  + corr\_plt\_filter（int、float/20）：如果输入为float，例如0.8，则按此阈值筛选相关性矩阵，留下绝对值高于0.8的所有情况；如果输入为int，例如20，则将相关性绝对值从大到小排序，保留前20个（或21个）特征
  + font\_size（int/12）：相关性矩阵图中所有字体大小
  + figsize（tumple/（14，10））：相关性图大小
  + rounds（int/2）：相关性矩阵热力图中相关性值保留位数
* 返回：

Corr Table：



Corr Heatmap：



3.1.2 PSI/Calc\_psi模块

模块功能介绍：PSI模块用于计算两个序列值的PSI值，该方法来自@Huyao；Calc\_psi模块调用PSI方法，可输入数据集也可输入序列。

* 入参：
  + bench（DataFrame、Series、array/must）：需要计算的一侧序列值或数据集。
  + comp（DataFrame、Series、array/must）：需要计算的另一侧序列值或数据集。

Notice：如果输入类型均为数据集，则取两个数据集的共同特征依次计算PSI，最后返回一个list

* + Group（Int/10）：分箱组数
* 返回：
  + psi\_value（float/Series）：如比较两个Series，则直接返回PSI值；如比较两个DataFrame，则返回公共特征的PSI序列

3.2.1 Woe\_Iv模块

模块功能介绍：将输入已分箱完成的数据集计算WoE和IV值，并将两值附加至分箱情况统计集fa\_df。

此模块功能已整合至2.2 Cut\_Merge，在此留作备份。

* 计算公式[1]：

假设对于某个变量Var分箱后的第i箱，其WoE计算如下：

其中，是指：该分组中所有y=1的样本占全量数据y=1的比例；是指：该分组中所有y≠1的样本占全量数据y≠1的比例。

如果某个组没有坏样本，采用拉普拉斯平滑处理：上述算式分子分母均加1。

则某个变量的IV值定义为：

* 入参：
  + datain（Dataframe/must）：传入需已分箱好、需计算WoE-IV的数据集
  + fa\_df（Dataframe/‘’）：分箱过程传入的分箱情况统计集，传入则会将每个变量的IV值、每个箱的WoE值计算并merge入fa\_df
* 返回：
  + wi\_dataout（Dataframe）：WoE转化后的训练数据集
  + woe\_dict（dict）：为每个变量、每个箱的WoE权重值
  + fa\_df（Dataframe）：① 如果传入了fa\_df，则将WoE值、IV值、IV重要性排序merge如fa\_df；② 否则传出已排序的每个变量的IV值

3.2.2 PDO\_Score模块

模块功能介绍：本模块包含PDO\_Score和PDO\_Score\_Convert

PDO\_Score：用途为根据逻辑回归训练的模型，外加①PDO基础分、②PDO翻倍分、③odds，计算返回每个变量、每个分箱的Score + Basic\_Score。

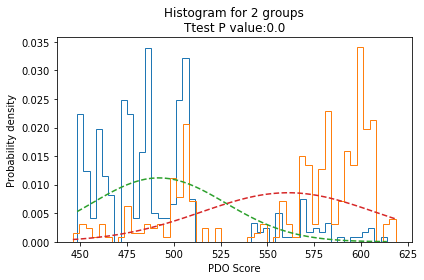
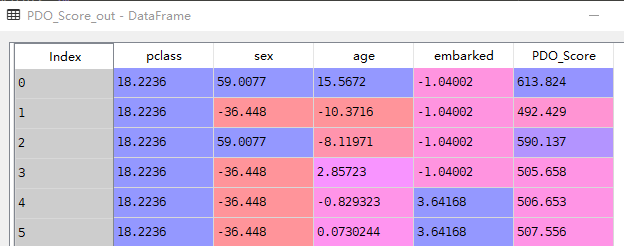
PDO\_Score\_Convert：根据前面函数计算的两个返回值，还有fa\_df，对原始输入的数据集直接进行打分，最终返回打分数据集。并绘制正负样本打分分布图。

PDO\_Score [2]：

* 入参：
  + X\_train（DataFrame/must）：训练集，在本版程序中为必须，用途为提取逻辑回归参数列表、另外还会协助计算odds。
  + y\_train（Series/must）：训练集对应的标签
  + Clf（model/must）：逻辑回归模型
  + fa\_df（DataFrame/must）：前面函数计算得到的规则集
  + pdo\_p0（int/600）：pdo基础分
  + pdo\_pdo（int/20）：pdo翻倍分
  + odds（int/0）：好坏比。如果不设定，则按照ks切点的百分位数区间内统计好坏占比。
* 输出
  + fa\_df（DataFrame）：附带score的规则集

PDO\_Score\_Convert：

* 入参：
  + PDO\_datain（DataFrame/must）：需要打分的数据集
  + fa\_df（DataFrame/must）：前面函数计算得到的规则集
  + y\_var（String/’ ‘）：如定义，则用于绘制打分后数据集的分布图
  + Basic\_score（Int/600）：基础分
  + num\_bins（Int/50）：绘制频数分布图时的分箱数
* 输出
  + PDO\_Score\_out（DataFrame）：打分集

3.3.1 Feature\_Select\_From\_Ensemble模块

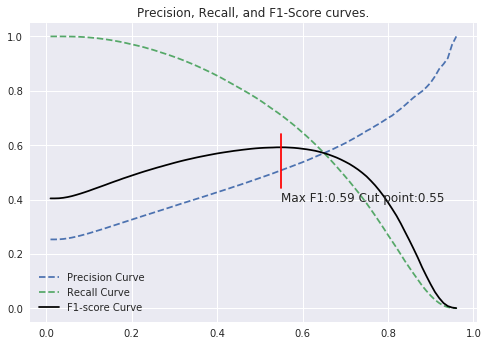
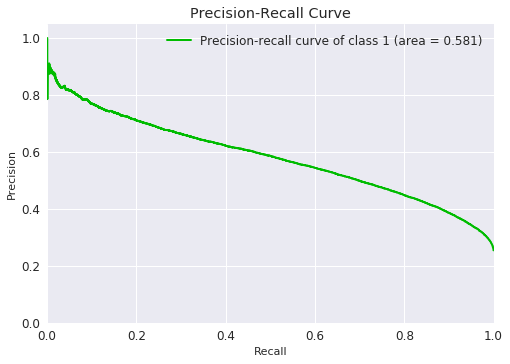
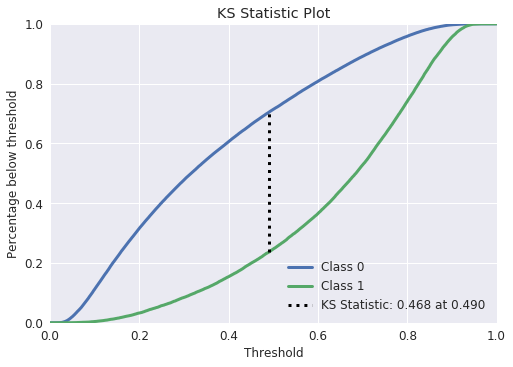
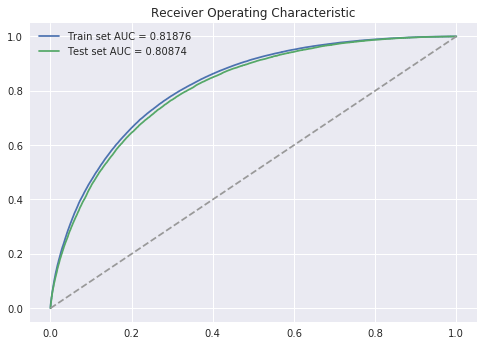
3.3.2 Feature\_Select\_From\_LR模块

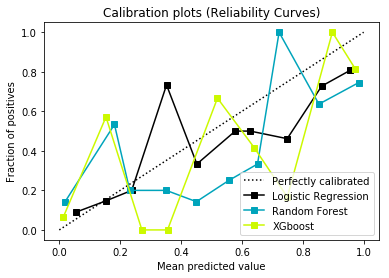
1. 模型表现

4 Model\_Performance模块

模块功能介绍：该模块为最终对模型进行评价的部分；主要包括：百分位切点准召情况、特征重要性表/图、ROC图、KS图、PR图、学习率图、calibration图。

* 核心函数：scikitplot包
* 入参：
  + Clf（model/must）：需要查看表现的主模型。
  + X\_train（DataFrame/must）：训练集
  + y\_train（Series/must）：训练集对应的标签
  + X\_test（DataFrame/[]）：测试集，非必须
  + y\_test（Series/[]）：测试集对应的标签，非必须
  + out\_path（Url/’’）：如定义，则将特征重要性和百分位切点准召情况输出
  + calc\_list（list/ ['pr','roc','ks']）：需要绘制哪些模型表现，完全指令如下：
    1. 'cut'：百分位切点准召
    2. 'pr'：PR图
    3. 'roc'：ROC图
    4. 'ks'：KS图
    5. 'fi'：特征重要性表/图
    6. 'lc'：学习率图（绘制过程很慢）
  + Calibration（dict/{}）：绘制Calibration图的模型列表例如：calibration = { 'Random Forest':clf\_rf, 'Logistic Regression':clf\_lr}，将给出两个模型的预测能力比较图
* 结果举例：



1. 应用实例

各个模块已经同时汇总到MLMF.py，使用时需要导入即可。如：

|  |
| --- |
| #import sys  #sys.path.append(r'D:\Desktop\Machine learning modeling framework')  #import MLMF as mf  #Variable\_pre\_exp, obs\_miss\_rate = mf.Data\_Explore(datain,calc\_obs\_na=True) |

以下为应用实例：

1. 泰坦尼克号生存率预测建模

项目介绍：泰坦尼克号生存预测问题是kaggle中的经典问题，也是机器学习建模入门实例。主要为利用乘客登记信息对最终乘客生存率进行预测。官方kaggle连接如：<https://www.kaggle.com/c/titanic>

泰坦尼克号包含特征较复杂，数据质量较差，但是特征量较少；适合用作本建模包开荒实例。

参考文章：

[1] https://blog.csdn.net/shenxiaoming77/article/details/78771698

[2] https://blog.csdn.net/sscc\_learning/article/details/78591210?utm\_source=blogxgwz0