**基于大数据技术的互联网金融风控预测用户贷款违约情况的探索性分析研究报告**

目录

[1摘要 2](#_Toc29039)

[1.1 研究背景 2](#_Toc16046)

[1.2 研究概括 4](#_Toc1224)

[2 关键词 4](#_Toc10279)

[3 正文部分 4](#_Toc8050)

[3.1 所用技术 4](#_Toc3597)

[3.2数据来源 5](#_Toc6837)

[3.3 数据分析与清洗： 7](#_Toc10625)

[3.4 特征工程 24](#_Toc25670)

[3.5 构建机器学习模型 30](#_Toc11820)

[3.6 基于模型的预测与评估 35](#_Toc4804)

**1摘要**

**1.1 研究背景**

在互联网金融迅猛发展的背景下，风险控制问题已成为行业关注的焦点。基于大数据的风控模型正成为互联网金融领域的研究热点。大数据风控的内涵及其与传统风控在风险识别方面的差异，是本文探讨的核心议题。大数据在商业模式中的应用主要体现在精准营销和大数据风控两个方面，其中精准营销的典型应用场景包括商品推荐和精准广告投放，而大数据风控则在互联网金融领域中扮演着重要角色。金融的本质在于风险管理，风控是金融业务的核心。在金融借贷业务中，如抵押贷款、消费贷款、P2P、供应链金融以及票据融资等，均需依赖数据风控来识别欺诈用户并评估用户信用等级。

在探讨互联网金融风控时，不得不回顾传统金融风控的实践。传统金融风控主要依赖于信用属性强大的金融数据，通常采用约20个维度的数据进行评分，以识别客户的还款能力和还款意愿。信用相关性高的数据维度大约有十个，包括年龄、职业、收入、学历、工作单位、借贷情况、房产、汽车、单位和还贷记录等。金融企业依据用户提交的数据进行评分，最终得出申请人的信用评分，并据此决定是否放贷及确定贷款额度。此外，与信用相关的数据还包括区域、产品、理财方式、行业、缴款方式、缴款记录、金额、时间和频率等。

互联网金融的大数据风控并非完全摒弃传统风控，而是对传统风控的数据维度进行了丰富。在互联网风控中，首先利用信用属性强的金融数据评估借款人的还款能力和还款意愿，随后利用信用属性较弱的行为数据进行补充。通常采用数据关联分析来判断借款人的信用状况，并借助数据模型揭示行为特征与信用风险之间的关系。互联网金融公司运用大数据进行风控时，会采用多维度数据来识别借款人风险。与信用相关的数据越多地被用于借款人风险评估，借款人的信用风险就越能被充分揭示，信用评分也越客观，更接近借款人实际风险水平。

**1.2 研究概括**

贷款业务作为借贷公司的核心业务，既带来了可观的营业收入，也带来了潜在的违约风险。为了在贷款收益与风险之间找到平衡点，借贷方必须对发放的贷款进行风险控制。因此，对每笔贷款进行风险评估成为了一项不可或缺的工作。然而，人工审核贷款发放不仅耗时费力，还会产生高昂的人力成本。为了应对日益增长的贷款审核需求，公司有必要构建一个更高效的审核系统以取代人工审核。

为了满足不断增长的贷款需求，借贷公司可以利用机器学习模型自动审核贷款申请人的信息。为了构建一个有效的分析模型，必须收集贷款申请人的各项特征，并分析这些特征与违约之间的关系。特征筛选是模型构建过程中的关键环节，不当的特征筛选可能导致模型结果过于宽松或过于严苛。当模型过于严苛时，借贷公司可能会错失潜在收益；而当模型过于宽松时，借贷公司将面临较高的风险。因此，特征筛选至关重要。通过描述性统计分析，我们可以了解每个特征变量与违约的相关性。在掌握了特征变量与违约结果之间的相互关系后，我们可以通过调整特征变量来提高其有效性（例如，将数值型数据中的少数变量转换为分类数据）。最终，通过对清洗后的数据进行建模，我们可以获得一个可靠的模型，以区分潜在的违约者并降低违约风险。

**2 关键词**

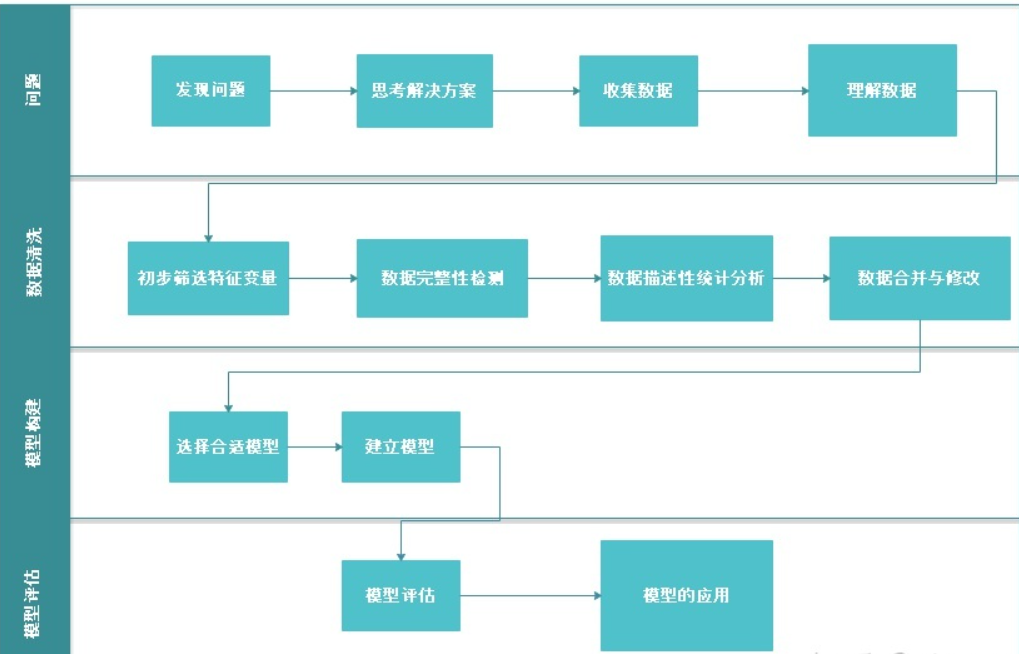
金融风控，数据分析，贷款违约，机器学习，人工智能

**3 正文部分**

**3.1 所用技术**

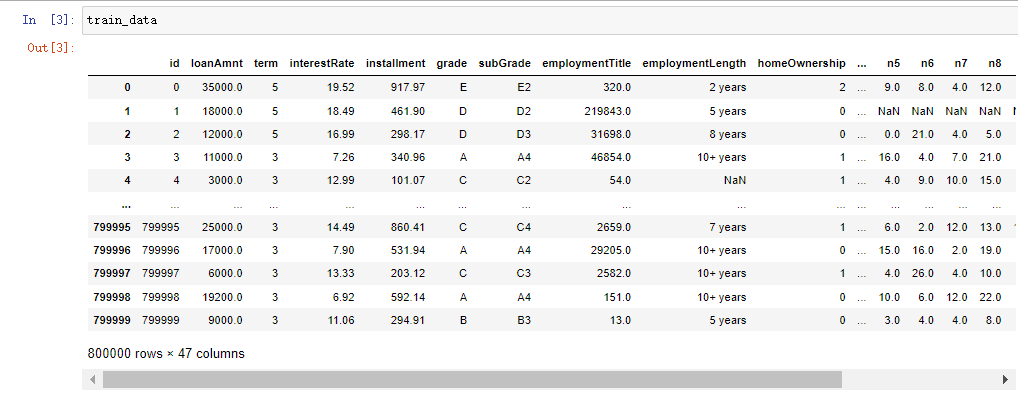
本研究基于Python语言进行分析，使用的开发环境为Jupyter Notebook。所涉及的库包括：数据分析库（如NumPy、Pandas）、可视化库（如Matplotlib、Seaborn）以及机器学习库（如Scikit-learn）。

分析步骤包括：一、数据采集；二、数据清洗与预处理；三、特征工程；四、建立机器学习模型；五、模型融合。

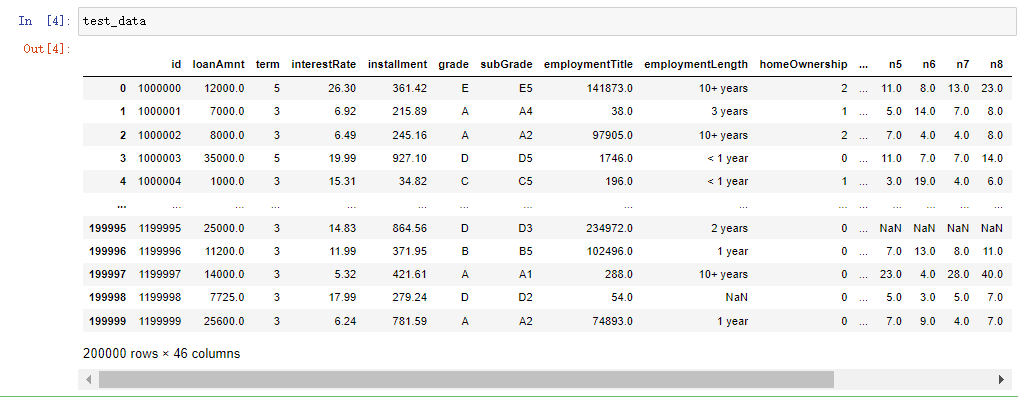


**3.2数据来源**

本研究数据源自某信贷平台的贷款记录数据库，数据总量逾120万条，涵盖47个变量字段，其中15个字段为匿名化处理。为便于模型验证，本研究将从中随机抽取80万条记录作为训练集，另外各20万条记录分别作为测试集A和测试集B。同时，将对employmentTitle、purpose、postCode和title等字段进行信息脱敏处理。



**训练集数据**



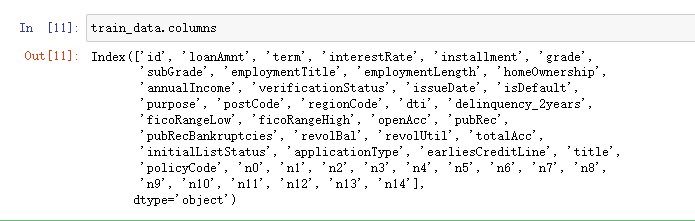
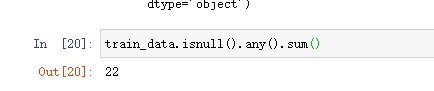
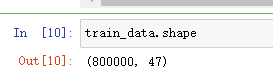
**测试集数据**

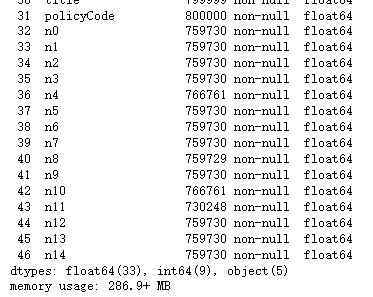
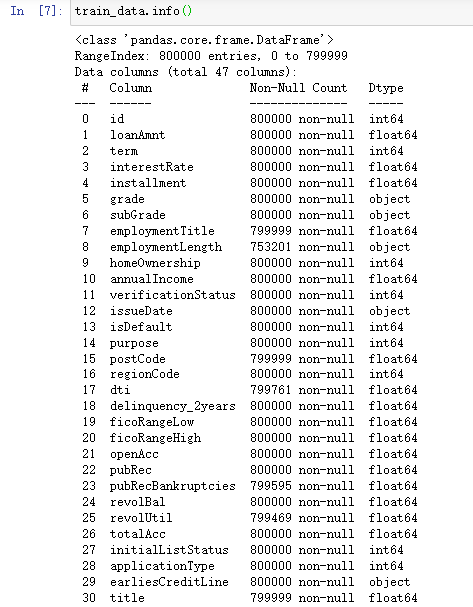


**各特征含义**

**3.3 数据分析与清洗：**

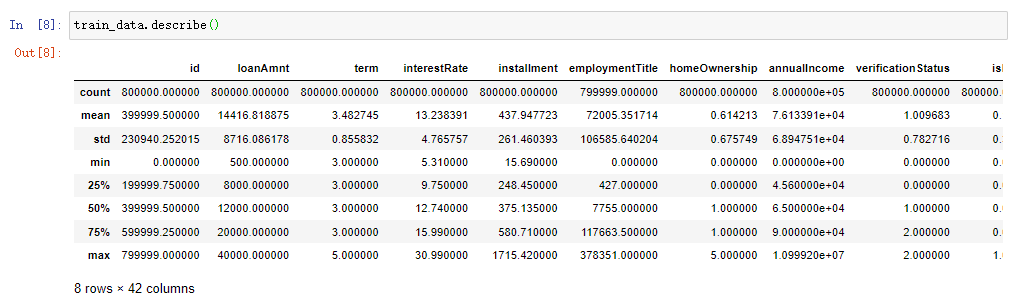
**3.3.1：对训练集数据进行量级及特征维度的审视，分析数据缺失情况以及数据类型。据图示，训练集数据量达800,000条，特征维度为47，其中22个特征存在缺失值。数据类型涵盖整型、浮点型及文本型。**





**分析数据缺失情况以及数据类型**

**3.3.2：对训练集中的各个特征变量进行数据量、平均值、标准偏差以及极值（包括前百分位数的最大值）的统计分析**

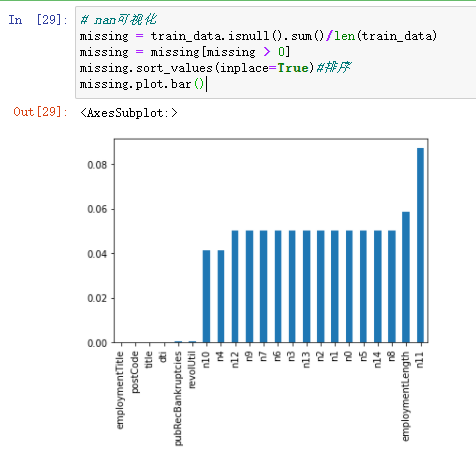


**特征变量进行统计分析**

**3.3.3：缺失数据的分析与填补**

根据3.3.3节的分析结果，训练集中共有22个特征列存在缺失值。

为探究缺失值特征之间的相关性，本研究对缺失特征进行了深入分析。首先，纵向审视数据集，识别出含有缺失值（"nan"）的列，并统计各列中"nan"的数量。此举旨在评估某一列中"nan"值的频率是否显著，若某一列中"nan"值过多，则该列对标签（label）的影响可能微乎其微，从而可考虑将其排除。反之，若缺失值较少，则通常采用填充策略。此外，通过横向比较，若数据集中存在大量样本在多数列上均缺失数据，且样本量充足，可考虑将这些样本删除。

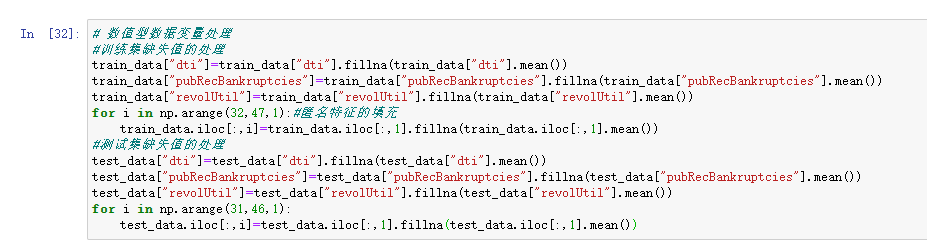


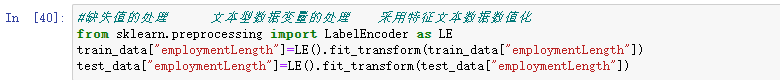
**Nan可视化**

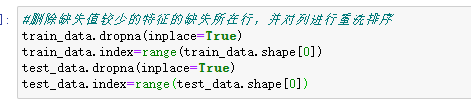
观察图表数据，可发现就业年限及若干匿名特征的缺失数据量较大，而其他部分特征的缺失值较少，可予以忽略（直接删除，但需注意数据重新排序）。

接下来，将对缺失值进行清洗与填充处理：

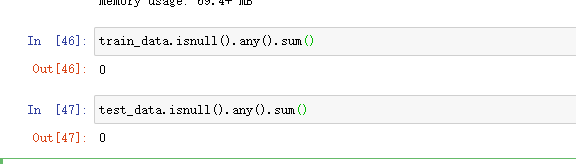
分析表明，在本数据集中，可采用填充特征平均值的方法来替代缺失数据（在数据分析实践中，对于数值型变量，通常采用均值或极值填充缺失值；对于文本型变量，则常利用sklearn.preprocessing模块中的LabelEncoder进行数值编码，以实现对数据及缺失值的直接处理）。







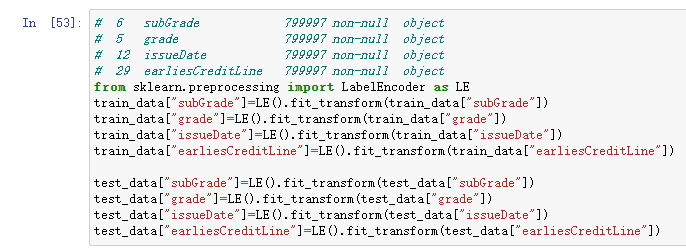
清洗结果：

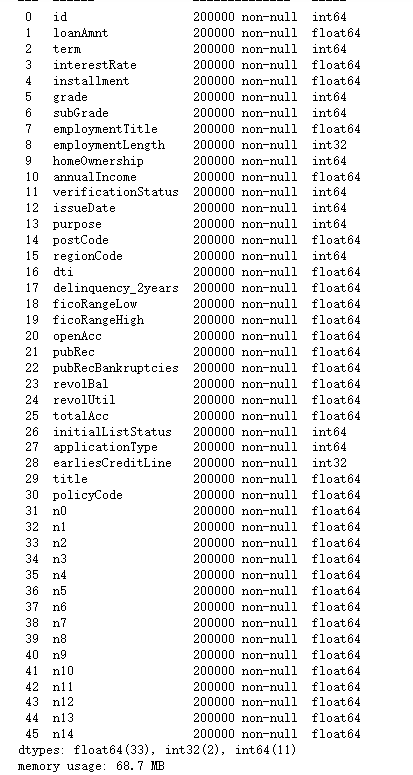
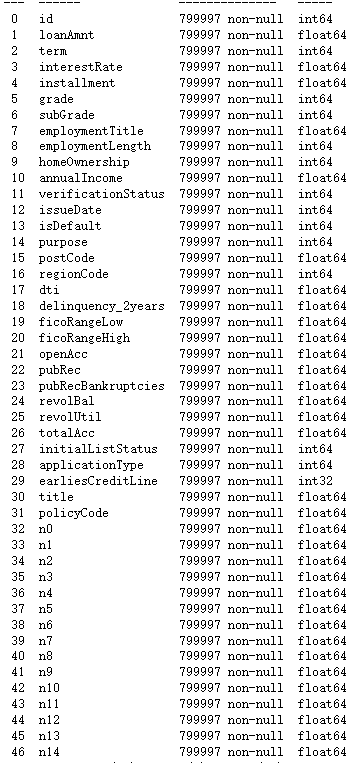


**缺失值进行清洗与填充处理**

**3.3.4：处理文本型数据**

在文本型数据处理过程中，通常采纳两种编码策略。其一为直接将文本型数据通过标签编码器（label encoder）进行数值化处理（标签编码器的作用在于将标签转换为数值形式。例如，若标签为一系列地名，由于地名无法直接输入至sklearn的分类模型中作为训练标签，因此需先将地名转换为数字形式）。其二为独热编码（one-hot编码），该方法通过将离散型特征转换为one-hot编码形式，使得特征间距离的计算更为合理。经过one-hot编码处理的离散特征，其每一维度均可视为连续特征，从而可以采用与连续型特征相同的归一化方法，对每一维度特征进行归一化处理。在本研究中，根据特征的具体属性，本人同时采用了标签编码器方法对数据进行编码处理（鉴于文本型特征值数据存在较多的不同值）。

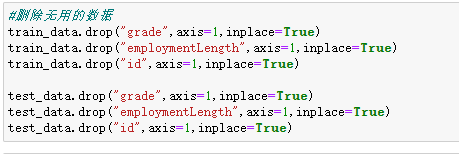




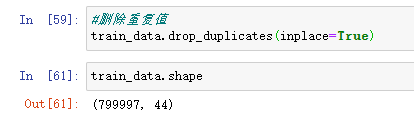
**处理文本型数据**

**3.3.5 删除冗余特征列**

**在特征选择过程中，通常会移除对模型构建影响较小的特征列。在此过程中，需依据数据特性进行差异化分析。例如，若特征列中的数据具有唯一性，则该特征列对模型的预测能力通常贡献有限。**



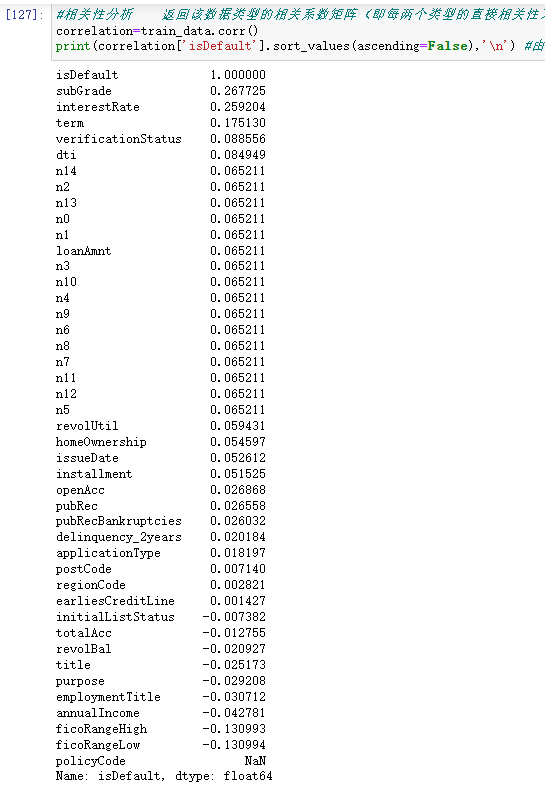
**3.3.6：移除数据集中的重复项**

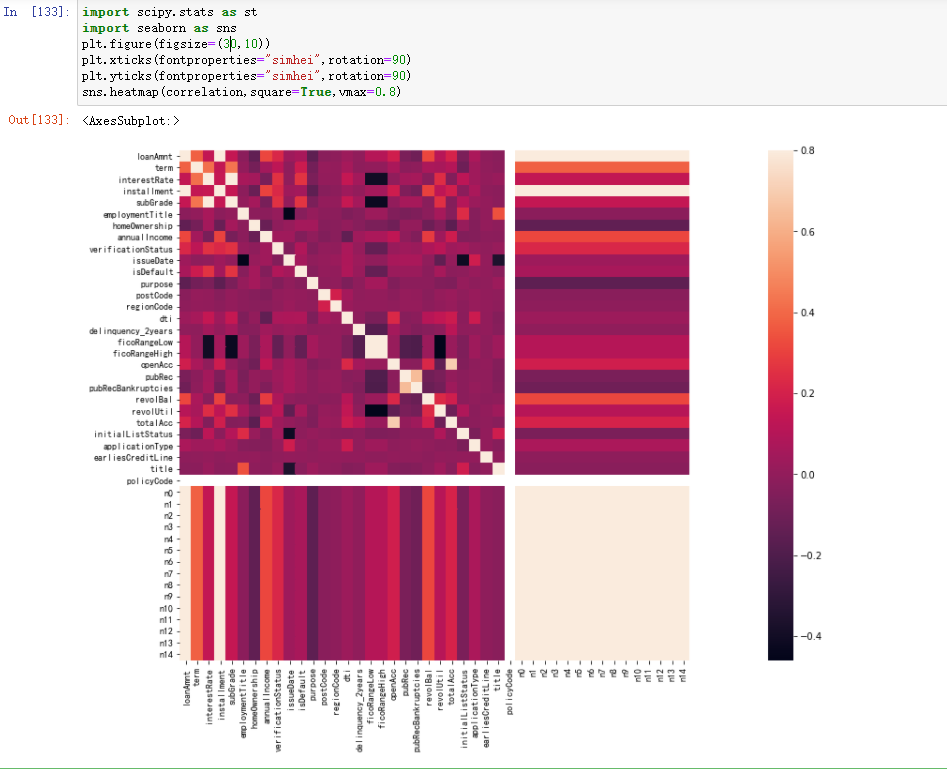


**3.3.7 特征间相关性分析**

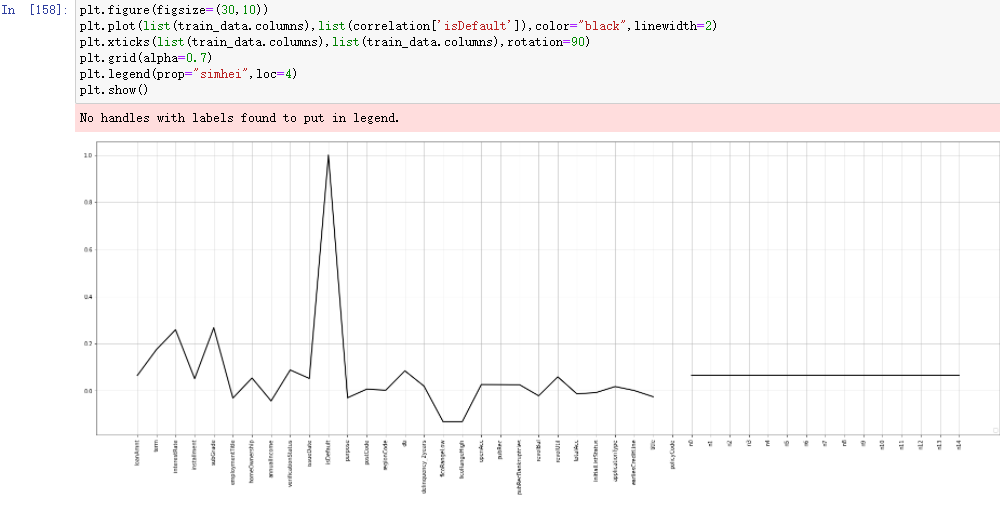
在进行数据相关性分析时，通常利用data.corr()函数来评估数据集中各变量间的相关性。该函数的返回值范围为[-1,1]，其中值接近-1表示变量间存在负相关关系，类似于反比例函数；值接近1则表示变量间存在正相关关系。data.corr()函数默认采用皮尔森相关系数，这是一种基础且直观的方法，用于揭示特征与响应变量之间的关系，主要衡量变量间的线性相关性。此外，该方法亦可应用于特征选择过程。

通过相关性筛选，可以识别出与目标变量（标签）相关且具有解释意义的特征。同时，依据相关性分析结果，可对与目标变量相关性较高的特征进行深入探讨。例如，数据分析揭示了subGrade（贷款等级子级）、interestRate（贷款利率）以及term（贷款期限，以年为单位）等特征与违约行为之间存在较高的相关性。



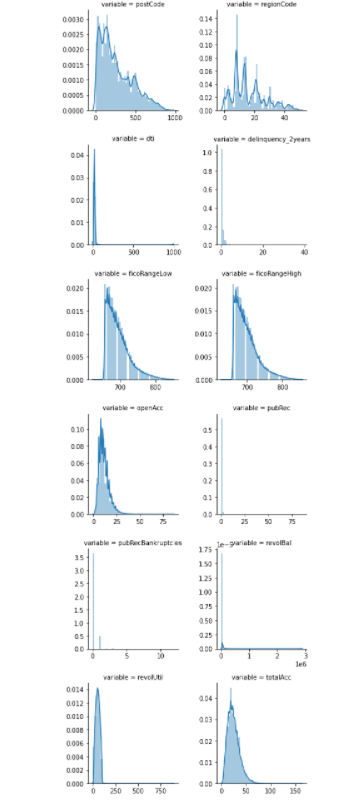
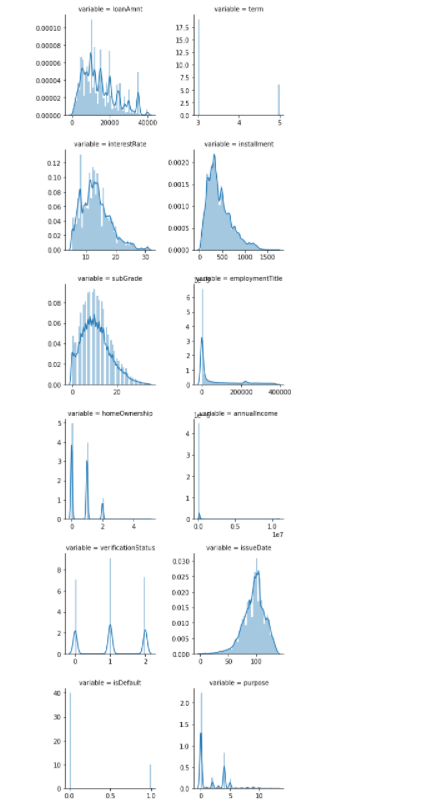


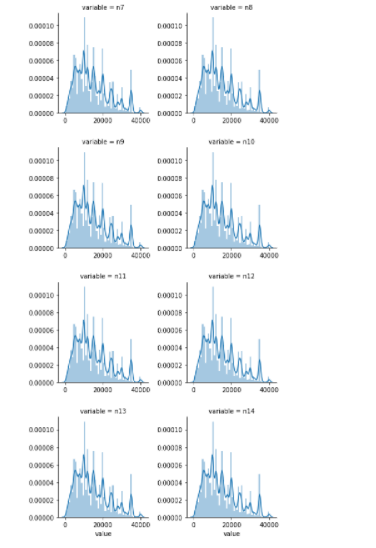
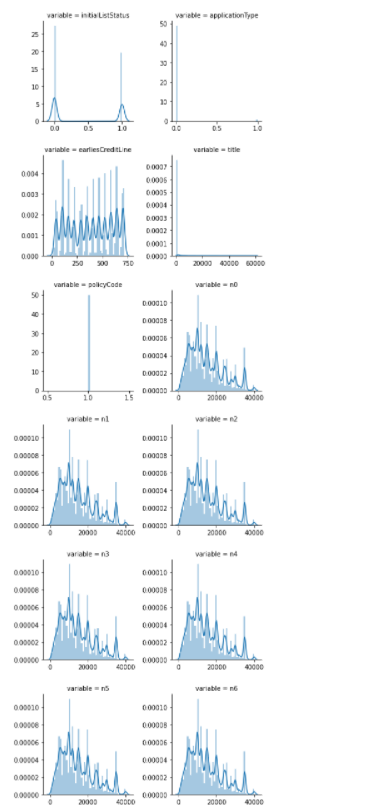
**各个特征之间的相关性**



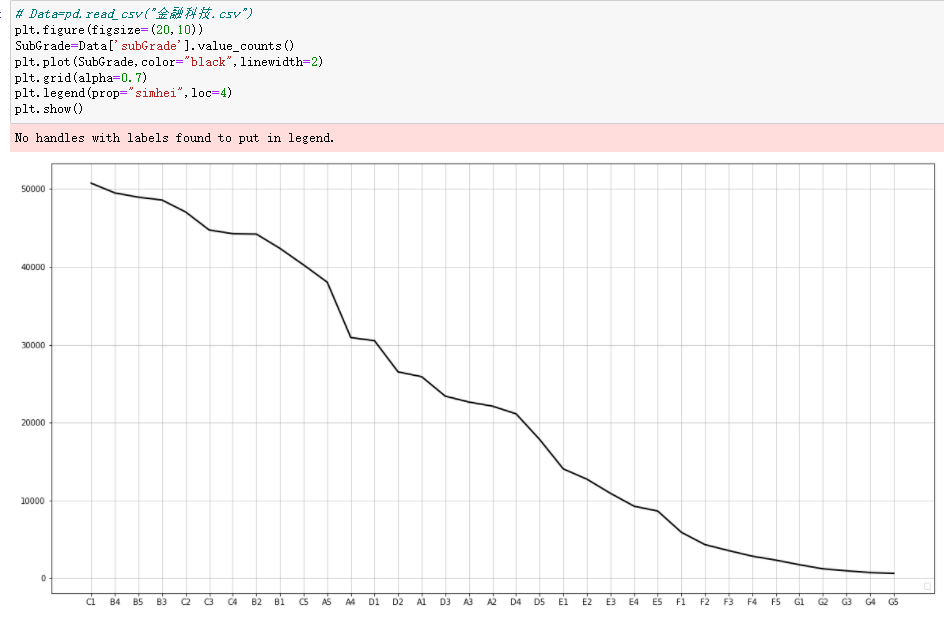
经相关性分析，subGrade贷款等级的子级、interestRate（贷款利率）以及term（贷款期限，以年为单位）与违约行为之间存在显著的相关性。因此，本研究通过

**3.3.8数字特征的分布可视化（特征的分布情况分析）**

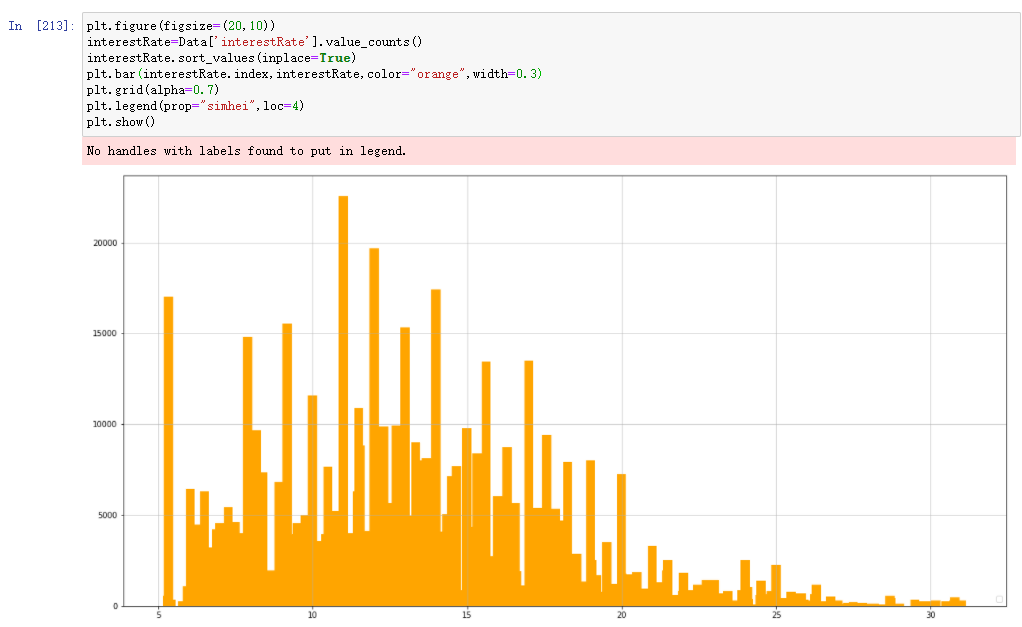




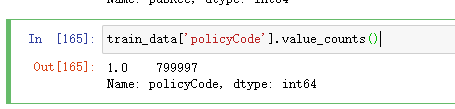
数值可视化的功能：通过观察数据的数值分布，实现对数据的深入分析，并识别特征趋势以指导相应的决策制定。

相关性分析揭示：SubGrade（信誉等级）对标签具有显著影响。因此，有必要对这一特征进行专项的数据可视化分析（见下图）。分析结果表明，信誉等级中人数最多的为C1，而最少的为G5。人数较多的等级包括C1、B4、B5、B3和C2，而人数较少的则为G1、G2、G3、G4和G5。基于此，若旨在降低违约率，策略上应考虑提升用户的信誉等级，进而制定针对性的决策。

同时，贷款利率（interestRate）作为一项具有显著影响的特征，其在单独特征可视化分析中显示出，贷款利率主要分布在8至15的区间内。其中，以10.99的贷款利率为最多，尽管存在少数个体的贷款利率较高，但这些高值仅占少数。基于此，可以针对贷款利率制定相应的策略，以期降低违约率。



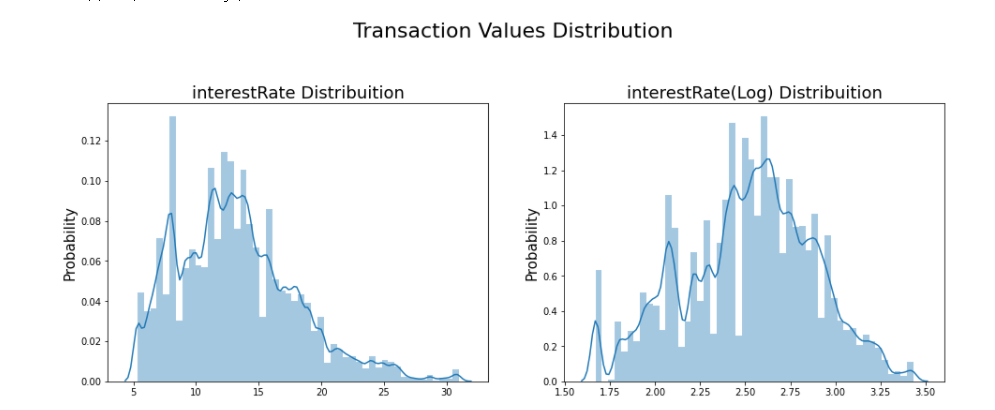
通过数值可视化筛选出一些对标签影响较小的数据（像唯一值）可直接删除。比如：特征policyCode唯一：可通过train\_data['policyCode'].value\_counts()验证



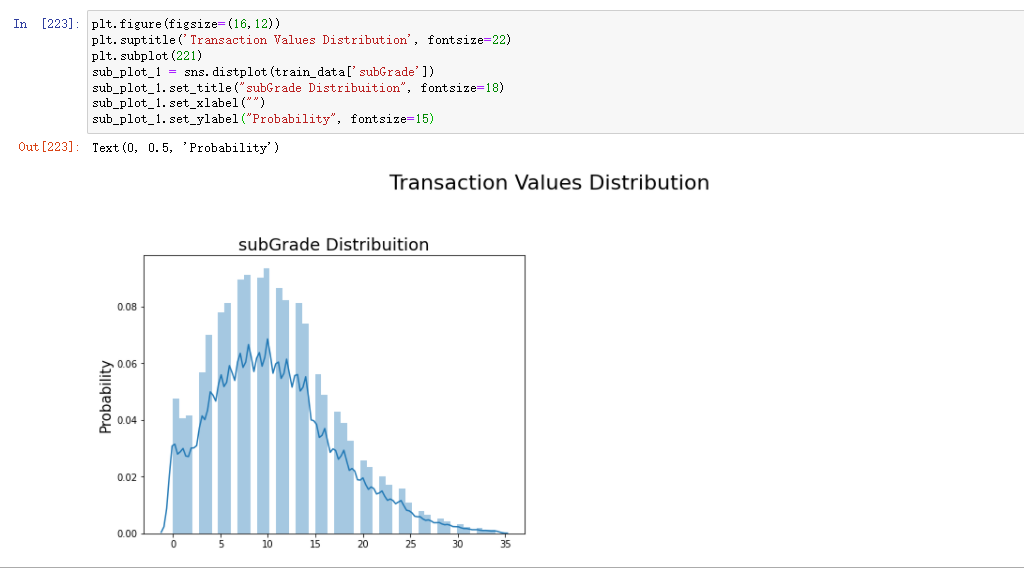
在进行数据分析时，可对特征变量进行正态性检验，以判断其是否遵循正态分布规律。对于那些不满足正态分布假设的变量，可考虑采用对数转换（log transformation）以检验其是否能更接近正态分布。在处理数据集时，若需进行标准化处理，必须明确先前进行正态化处理的原因。正态化的原因在于，在某些情况下，正态分布的数据能够加速模型的收敛过程，同时某些模型如高斯混合模型（GMM）和K最近邻（KNN）算法要求输入数据必须符合正态分布。确保数据分布不过度偏斜是重要的，因为极端的偏态分布可能会对模型的预测性能产生不利影响。

对贷款利率数据进行正态分布的观察，并采用对数转换（log transformation），鉴于贷款利率数据通常呈现正态分布特性。

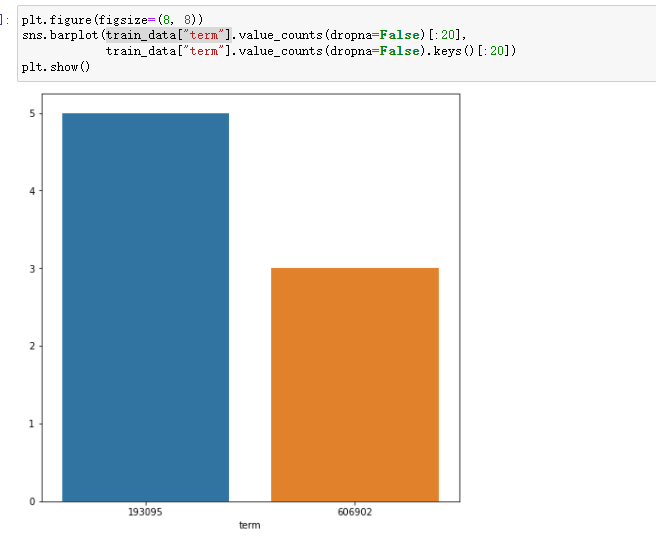




在对subGrade进行数值化处理后，即赋予其数值（0对应A1，1对应A2，3对应A3，依此类推至35对应G5），其分布情况亦呈现出与正态分布相吻合的特征。

**3.3.9单一变量分布可视化**

Term贷款年限分布可视化情况（由贷款年限可知主要分为五年和三年两种，在这两种中更多的人选择了5年）

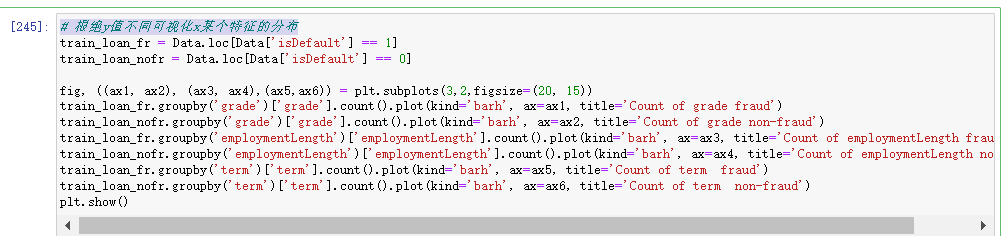


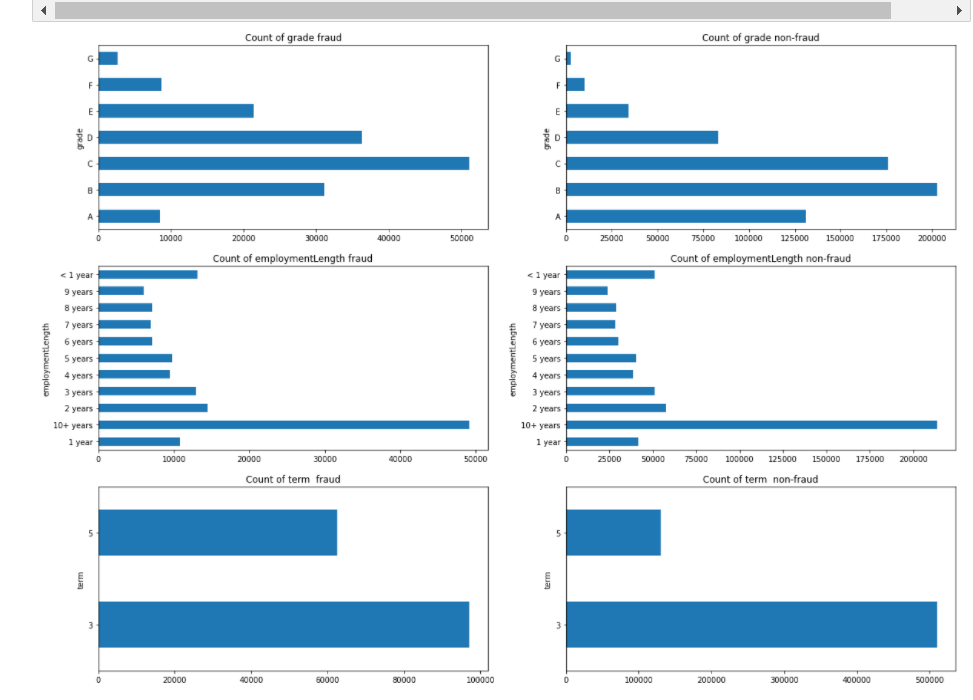
**3.3.10根据y值的不同观察特征的分布情况（根据是否违约来观察特征）**

观察图示数据，左侧分布代表违约用户，右侧分布代表非违约用户。分析图中grade信誉等级数据，违约用户中C级占比最高，而B级在非违约用户中占据较大比例。据此可推断，非违约用户的平均信誉等级高于违约用户，从而得出结论：违约用户群体中信誉等级较低。

在考察employmentLength（就业年限）这一指标时，发现违约用户与非违约用户在就业年限上差异不显著，难以作为有效区分依据。

针对term（贷款期限）这一变量进行分析，结果显示在贷款期限为5年的群体中，违约用户所占比例高于非违约用户。因此，从贷款期限角度出发，可制定相应策略以降低违约率。



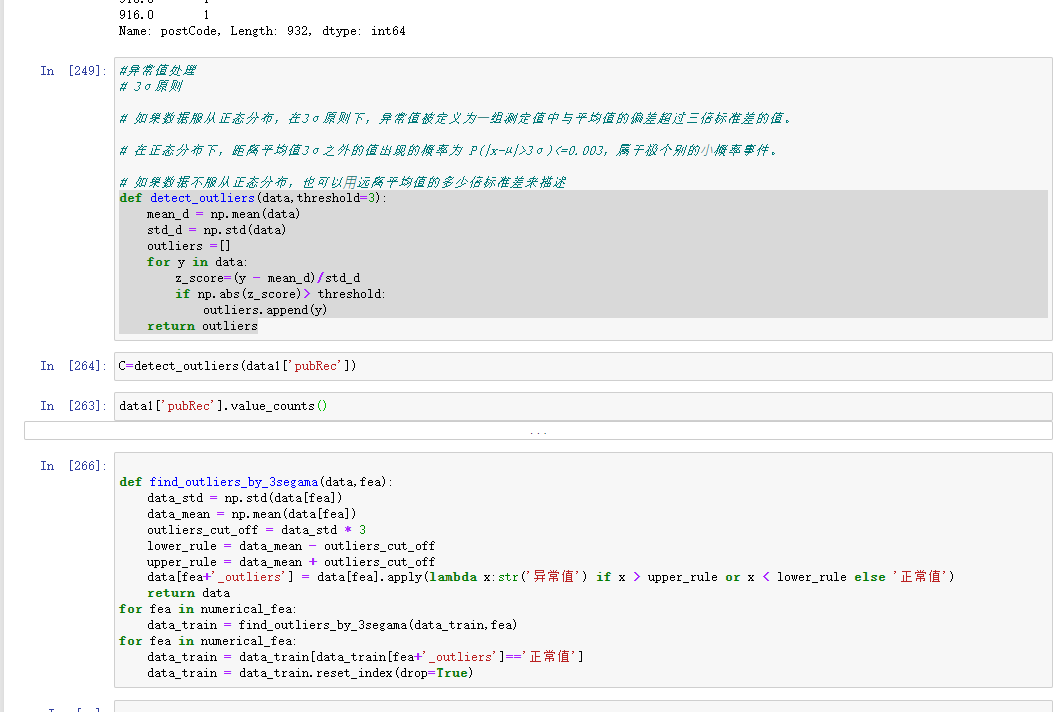


**3.3.11异常值的处理**

常用方法3σ原则（前提是数据服从正太分布）

如果数据服从正态分布，在3σ原则下，异常值被定义为⼀组测定值中与平均值的偏差超过三倍标准差的值。

在正态分布下，距离平均值3σ之外的值出现的概率为 P(|x-µ|>3σ)<=0.003，）（概率约99.7%）属于极个别的⼩概率事件。



如果数据不服从正态分布，也可以⽤远离平均值的多少倍标准差来描述



**箱线图（处理函数如图）**

**3.4 特征工程**

特征工程在机器学习乃至深度学习领域占据着至关重要的地位，其在实际应用中的时间投入通常也是最多的。

（在建模之前，特征的选择、新特征的构建以及无量纲化处理等特征处理步骤）对模型构建具有决定性影响。

特征工程的核心方法论可概括为：提出问题、建立假设、进行论证。在特征工程的不同阶段，提出不同的问题，建立相应的假设，并采用适宜的方法，构成了特征工程的全部内容。

无量纲化（亦称无量纲化或无量纲化处理）是指通过适当的变量替换，去除方程中涉及物理量的部分或全部单位，旨在简化实验或计算过程，是科学研究中一种关键的处理策略。

**3.4.1 特征的无量纲化处理**

（一）特征归一化的优势

1、数据量纲的差异性；数量级的显著差异

（经过标准化处理后，原始数据转化为无量纲化指标测评值，各指标值处于同一数量级别，便于进行综合测评分析。

若直接使用原始指标值进行分析，则数值较高的指标在综合分析中的作用会被过分强调，而数值水平较低的指标的作用则相对减弱。）

2、避免数值问题：过大的数值可能引发数值问题。

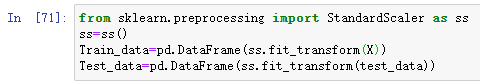
3、平衡各特征的贡献度

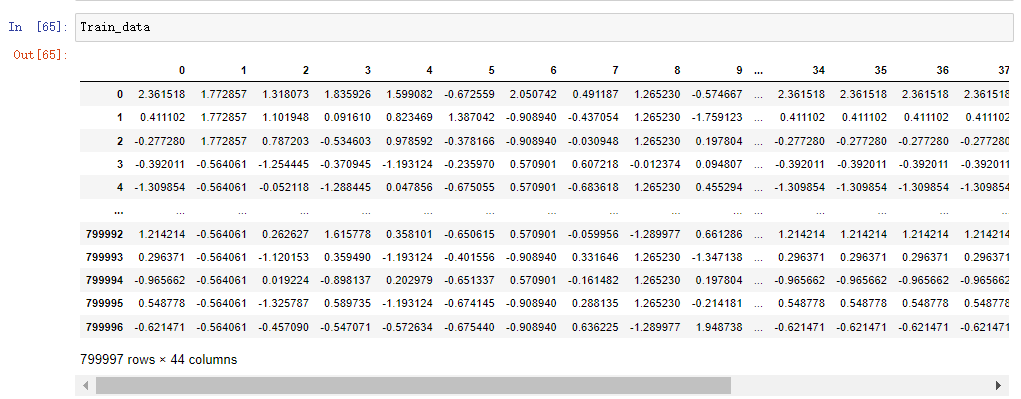
某些分类器需要计算样本间的距离（例如欧氏距离），如KNN算法。若某一特征的值域范围极大，则距离计算将主要依赖于该特征，这与实际情况可能不符（例如，在某些情况下，值域范围较小的特征可能更为重要）。

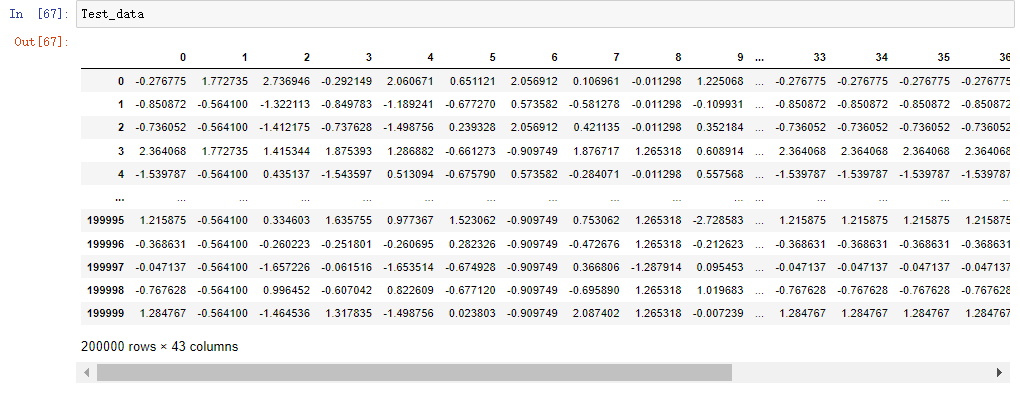
4、满足某些模型求解的需求：加速梯度下降法寻找最优解的过程）

注意：在进行特征的无量纲化处理时，需将标签单独提取出来。









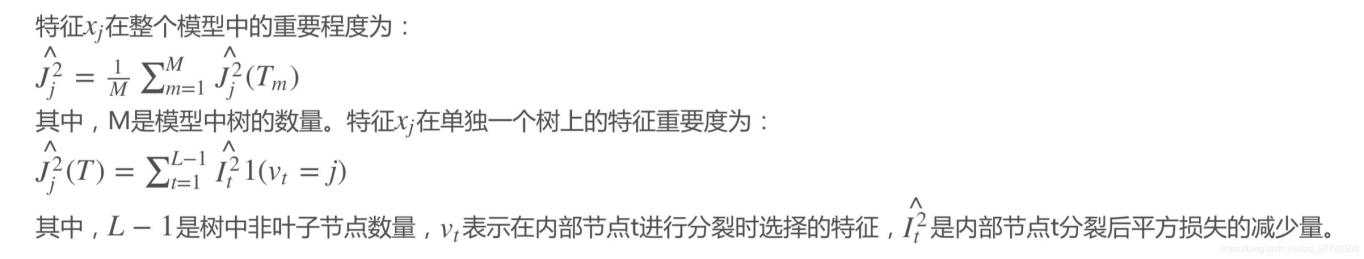
**3.4.2 特征选择及其对贷款违约预测的重要性分析**

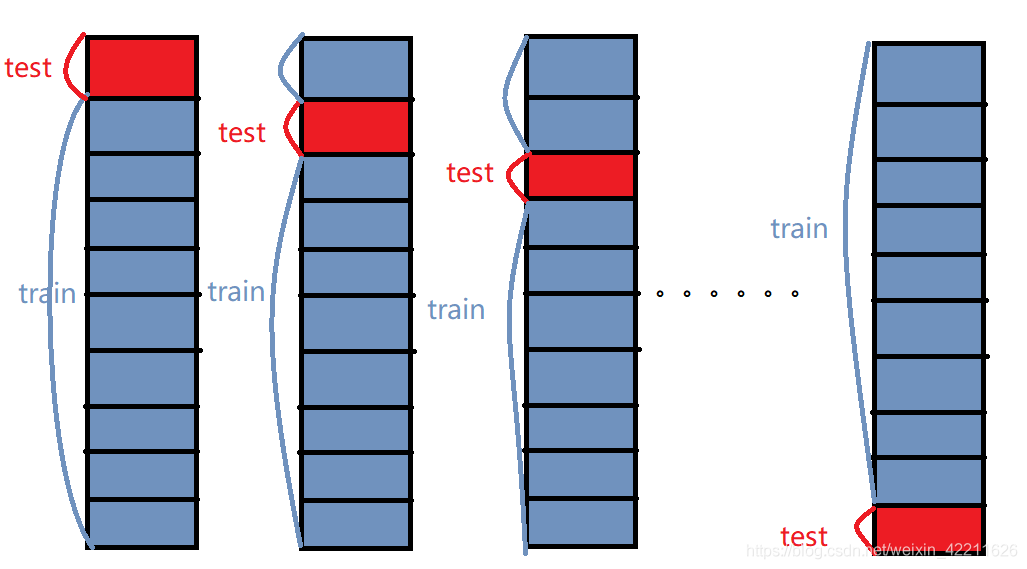
在特征提取过程中，通常选取与预测结果相关性较高的特征。为此，本研究采用机器学习库sklearn中的SelectFromModel方法进行特征选择，旨在通过减少特征数量以提升机器学习模型的运行效率。在大多数情况下，通过剔除少量特征即可实现模型运行速度的提升。

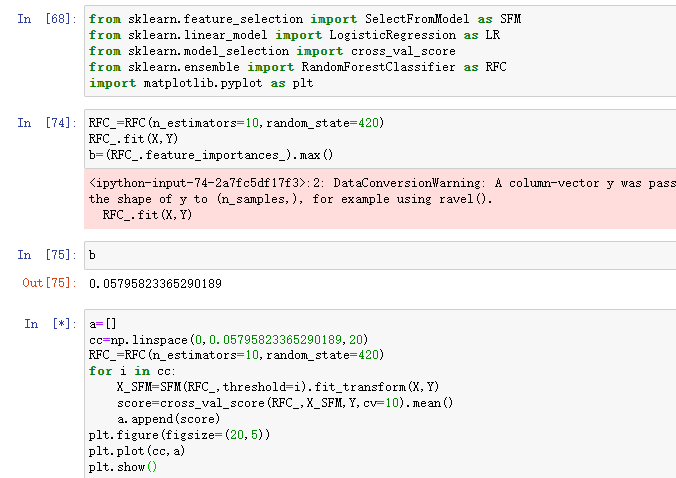
（1）SelectFromModel 是一种基础分类器，它依据特征的重要性权重进行特征选择。该方法适用于任何具有coef\_或feature\_importances\_属性的估计器。若估计器的coef\_或feature\_importances\_值低于设定的threshold参数阈值，则认为这些特征不重要或可被移除。）

（2）threshold 参数用于设定特征选择的阈值。该参数保留重要性更高或相等的特征，而剔除重要性较低的特征。当threshold设置为None时，若估计器的惩罚参数显式或隐式设置为l1（例如Lasso），则默认阈值为1e-5；否则，默认使用特征重要性的均值（mean）作为阈值。

（3）特征重要性评估：通过以下公式对特征的重要性进行评估。



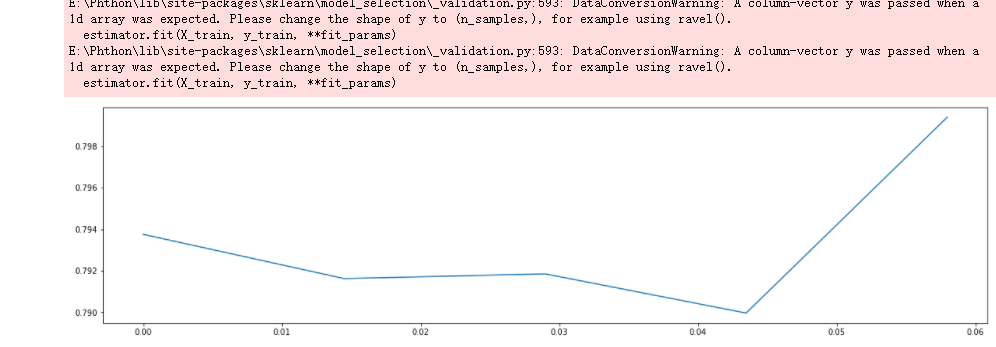
（4）交叉验证：本研究采用10折交叉验证方法，将数据集划分为10个子集，并进行一次交叉验证。实际上，该过程涉及十次模型训练和测试，每次将一个子集作为测试集，其余九个子集作为训练集。通过十次循环，最终对模型的十次测试结果求平均值。交叉验证的优势在于：1）它能够评估模型在新数据上的预测性能，从而在一定程度上减少过拟合现象；2）它能够从有限的数据集中尽可能多地提取有效信息。



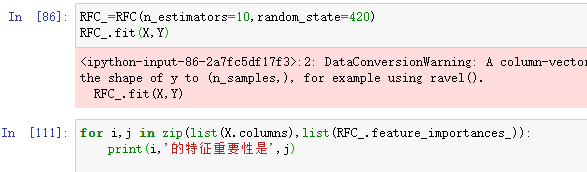
在本研究中，我们采纳随机森林算法作为分类器，其阈值设定为特征重要性评分的最大值，即X\_SFM=SFM(RFC\_,threshold=i).fit\_transform(X,Y)。

其中，b=(RFC\_.feature\_importances\_).max()。

通过调整阈值threshold的数值，我们能够控制训练集中特征的数量。为了对预测结果进行初步评估，我们采用了交叉验证方法，并利用随机森林模型辅助选择模型。此外，通过matplotlib绘制不同特征的预测分数图谱，我们旨在选取一个特征数量较少但不影响整体预测准确性的模型，从而提升模型的运行效率。



观察图示数据，可见阈值（threshold）与特征重要性（feature\_importances\_）在0.05至0.06区间内呈现出较高的数值。

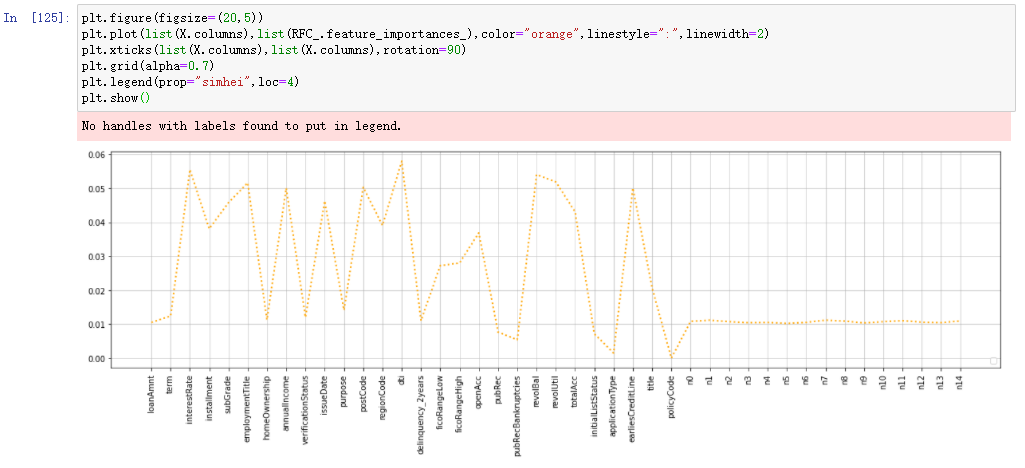


-



利用feature\_importances\_函数进行特征重要性分析，发现dti（债务收入比）、interestRate（贷款利率）、revolBal（信贷周转余额合计）、revolUtil（循环额度利用率，即借款人使用的相对于所有可用循环信贷的信贷金额）、postCode（公开可用的策略\_代码=1新产品不公开可用的策略\_代码=2）、earliesCreditLine（借款人最早报告的信用额度开立的月份）以及annualincome（年收入）等特征对贷款违约结果具有显著影响。其中，债务收入比对贷款违约的影响最为显著。

相对而言，policyCode（借款人在贷款申请中提供的邮政编码的前3位数字）、pubRec（贬损公共记录的数量）、pubRecBankruptcies（公开记录清除的数量）、initialListStatus（贷款的初始列表状态）以及applicationType（表明贷款是个人申请还是与两个共同借款人的联合申请）等特征对贷款违约的影响微乎其微。



**3.5 构建机器学习模型**

到目前为止，已处理的训练集数据包括以下三个：

经过数据预处理后的数据集。

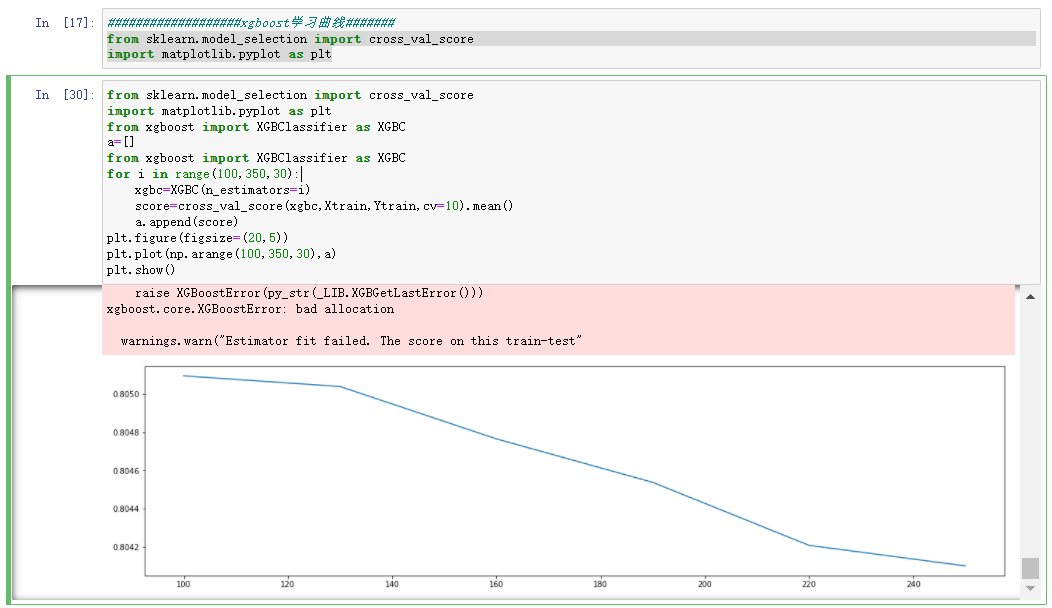
在数据预处理基础上实施特征工程后的数据集。

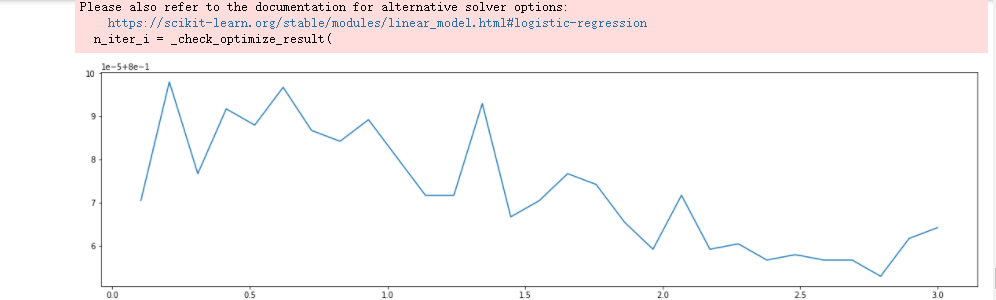
在数据预处理和异常值剔除后进行特征工程的数据集。

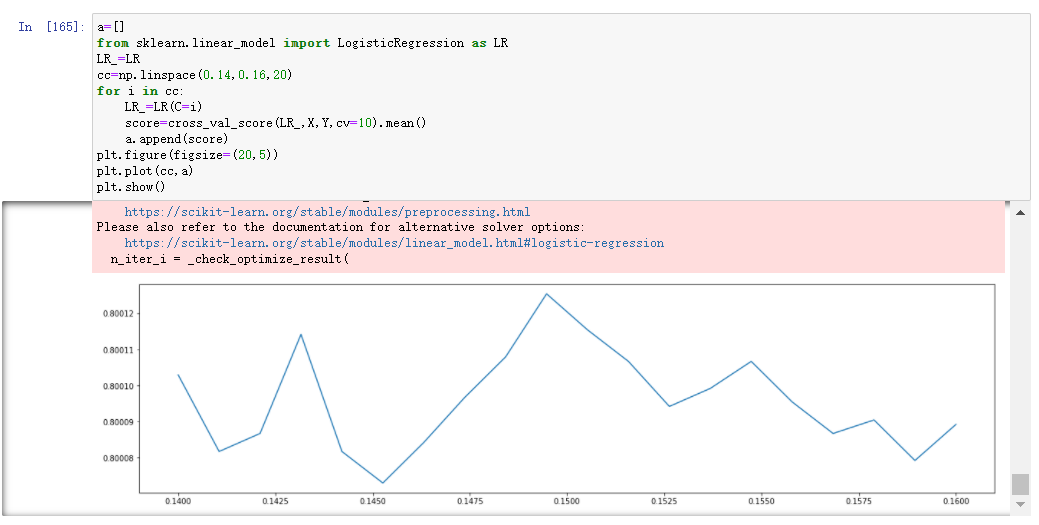
鉴于该数据集属于离散型数据类别，因此选择分类算法进行处理。

由于数据集规模较大，支持向量机和KNN算法并不适宜处理此类数据。故选取随机森林分类器、逻辑回归分类器和梯度提升树分类器作为初始模型。通过这些模型对数据进行评分（score），并返回预测的R2值。R2的计算公式为：R2 =（1-u/v），其中u为（y\_true - y\_pred）\*\* 2的总和，v为（y\_true - y\_true.mean()）\*\* 2的总和。

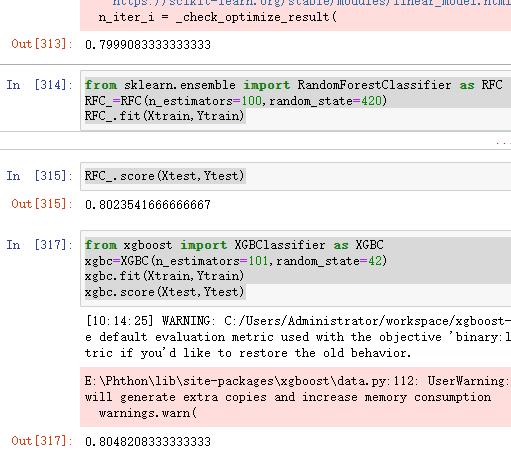
参数调整：采用交叉验证方法对模型参数进行调整，并绘制相应的学习曲线。图示为梯度提升树的学习曲线，主要调整参数为树的数量，观察到树的数量在100至120之间时模型拟合效果较佳；在逻辑回归中，初步调整参数包括C值、正则化方式和max\_iter等（正则化强度的倒数。默认值为1.0。C值越小，对损失函数的惩罚越重，正则化效果越强，对过拟合的惩罚也越严厉，参数Theta也会相应减小）。随机森林中最重要的调整参数为树的数量，此外还包括计算属性的方法（基尼系数或信息增益）、树的最大深度、叶子节点的最小和最大样本数，以及在分支时考虑的特征数量等。



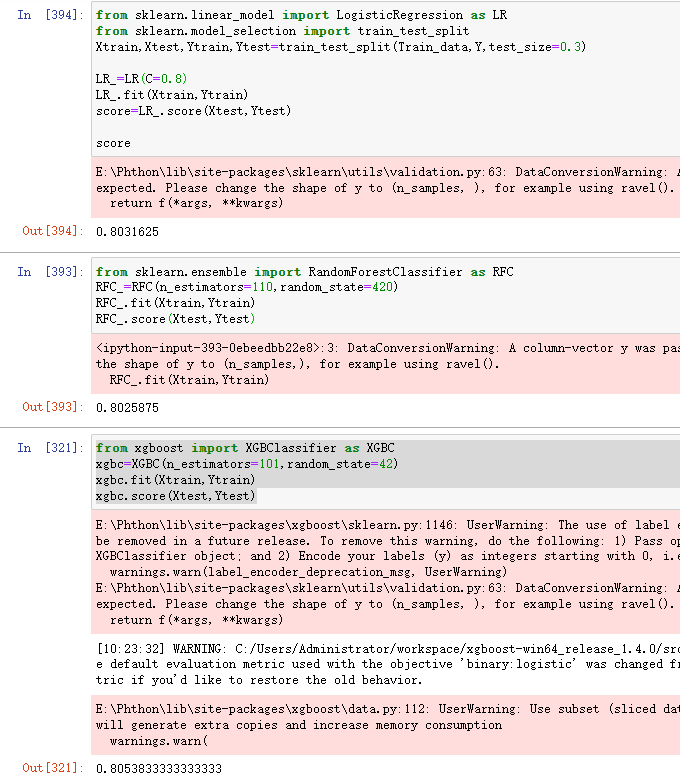




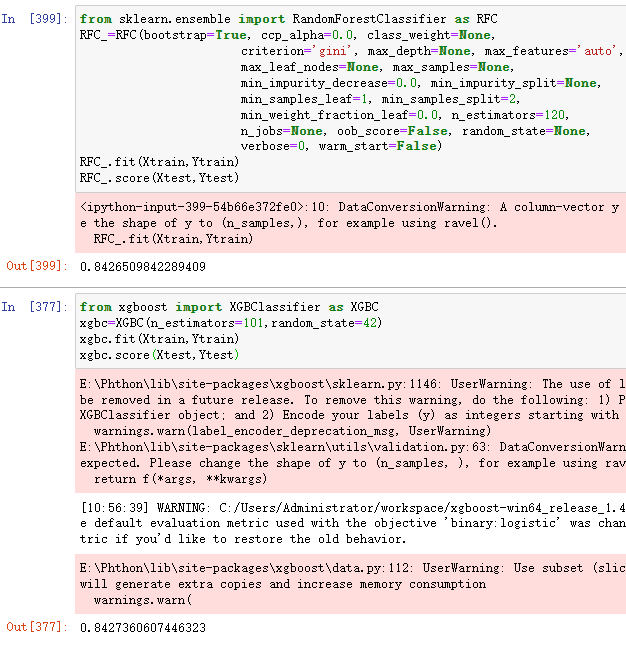
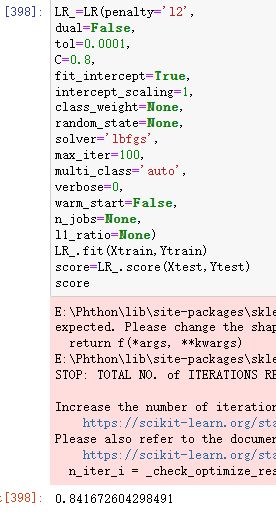
通过实施模型测试与参数调整，对三组数据进行建模分析后发现：在三组数据的测试过程中，第三组数据（经过数据清洗、特征工程以及异常值处理）的性能最优，显著优于第一组数据（仅进行简单清洗）和第二组数据（简单清洗与特征工程相结合）。进一步分析表明，第二组数据相较于第一组数据亦展现出更佳的性能（简单清洗数据的R^2平均值介于0.7900至0.8050之间，而加入特征工程后的平均值稳定在0.8至0.81之间，且运行速度较仅进行简单清洗的数据有显著提升）。此外，结合异常值处理与特征工程的数据处理结果显示，R^2平均值稳定在0.84至0.85之间。由此可见，异常值处理和特征工程对于构建更为精确的模型具有积极影响。



**简单清洗的数据进行的跑分**



**特征工程后的建立的模型**

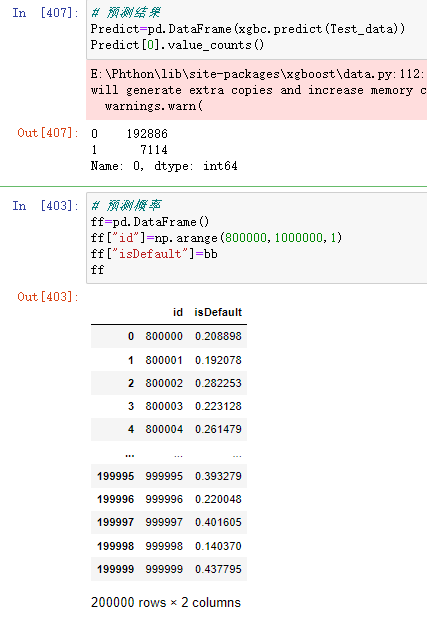
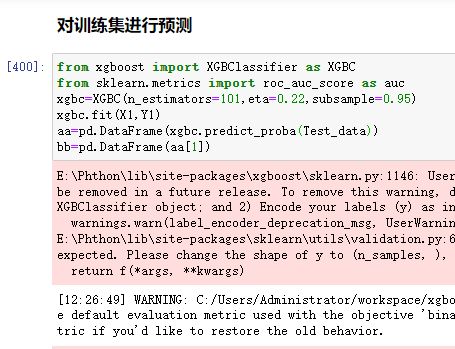


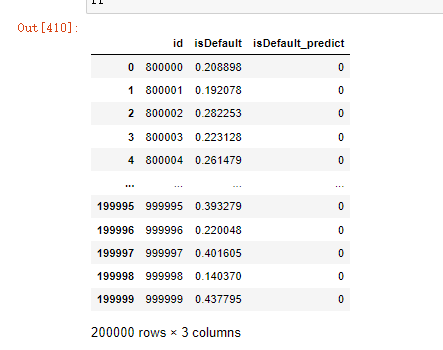
**对异常值进行处理和特征工程后的数据建立的模型**

经过综合考量，本研究最终选定第三组数据作为训练集以开展模型训练。通过反复的实验和参数调整，最终确立了模型结构（如图所示）。

**3.6 基于模型的预测与评估**

模型构建完成后，利用训练集对新的测试集进行预测，以获得预测数据。predict\_proba函数与predict函数不同，其输出为各类结果的概率值。而predict函数则直接给出预测结果（0代表非违约，1代表违约）。

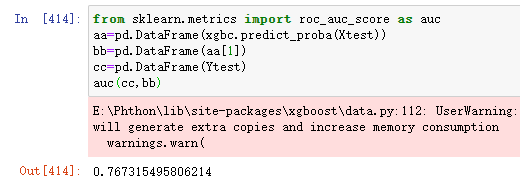




**模型的评估**

在本研究中，我们采用了ROC曲线下面积（Area Under Curve, AUC）作为评估指标。该指标定义为ROC曲线与坐标轴所围成的区域面积，其值域上限为1。鉴于ROC曲线通常位于等误判率直线（y=x）之上，AUC的取值范围介于0.5至1之间。AUC值越接近1，表明检测方法的鉴别能力越强；而当AUC值等于0.5时，表明该检测方法无鉴别能力，不具备实际应用价值。

在模型正负样本比例失衡的情况下，AUC评估模型展现出其独特的优势，能够有效地评估模型性能。



**模型的预测**