用户信贷违约风险

预测方案设计

队员：黄凤翔、汤荣杰、高梦真

指导老师： 曹琼 黄贤英

目录

**[摘要](#_Toc31595_WPSOffice_Level1)** **[3](#_Toc31595_WPSOffice_Level1)**

**[一、数据来源及说明](#_Toc27279_WPSOffice_Level1)** **[4](#_Toc27279_WPSOffice_Level1)**

**[二、问题背景](#_Toc15349_WPSOffice_Level1)** **[4](#_Toc15349_WPSOffice_Level1)**

**[三、数据分析及模型处理](#_Toc11345_WPSOffice_Level1)** **[4](#_Toc11345_WPSOffice_Level1)**

[1.任务目标](#_Toc27279_WPSOffice_Level2) [4](#_Toc27279_WPSOffice_Level2)

[2.数据集的分析及预处理](#_Toc15349_WPSOffice_Level2) [5](#_Toc15349_WPSOffice_Level2)

[3.特征工程](#_Toc11345_WPSOffice_Level2) [9](#_Toc11345_WPSOffice_Level2)

[4.模型训练](#_Toc1924_WPSOffice_Level2) [15](#_Toc1924_WPSOffice_Level2)

**[四、解决方案](#_Toc1924_WPSOffice_Level1)** **[17](#_Toc1924_WPSOffice_Level1)**

[1.方案主要工作](#_Toc9030_WPSOffice_Level2) [17](#_Toc9030_WPSOffice_Level2)

[2.方案优点](#_Toc29020_WPSOffice_Level2) [18](#_Toc29020_WPSOffice_Level2)

[3.方案使用建议](#_Toc26292_WPSOffice_Level2) [18](#_Toc26292_WPSOffice_Level2)

**摘要**

如今信贷方式的贷款越来越多，在普及的同时，便也带来了许许多多的问题。而信贷中最重要的则是要判断用户是否会违约。因而有了对用户信贷违约的预测，以用于降低用户违约的风险。

借助于Lending Club 2018年第一季度的信贷数据，我们通过一些数据处理和分析的基本方法来对用户信贷是否会违约进行预测。从而给投资人一些帮助，以决定是否可放款。

方案中首先是对数据进行了预处理，除去部分没有用处的信息，接着是对数据进行分析，包括对不同种类型特征的不同处理，缺失值的填充等，然后则是进行特征工程的探索，比如对非数值特征的转换，新特征的寻找，特征的选择等。最后，将处理好的数据放入模型。同时在多个模型训练下，比较每个的预测效果，判断模型的好坏。

当数据和模型都处理好后，我们根据过程中遇到的一些问题进行分析，针对信贷过程中可能会遇到的一些问题，总结了一些建议和方案，希望可以在信贷方面上做出一些贡献。

**一、数据来源及说明**

本数据来源于财务平台Lending Club（P2P平台，公司位于美国旧金山）2018年第一季度的部分贷款的完整贷款数据。包括当前贷款状况(正在执行、迟交、全额偿还等)，以及最新的付款信息；包括贷款用户公开可见的身份信息等。正在执行的贷款数据包含上一个季度发放的贷款数据相关信息。数据除隐藏了用户的私有信息外，基本未做处理，因而更能体现数据分析的应用。本贷款数据有107864条用户贷款记录，特征属性包括贷款人贷款金额，贷款利率，贷款人填报年收入等共145个属性。

**二、问题背景**

如今，Lending Club的商业模式已转变为证券模式。在此模式下，第一步便是平台对借款人的信用等级进行评级评分，进行评分后才能供其他投资者根据评价标准投资。然而在贷款的发放审核中的主要依据就是借款人的个人征用评级，所以对于这一步的不断完善与维护是及其重要的。

从网上的数据来看，美国在个人征信方面做得相对于国内要好得多。在美国，个人伴随的各种信用以及各种指标是比较可靠的，在贷款方面便可以以此为较大的依据。而在国内，个人征信这方面的工作不够完善，所以对贷款方面的违约预测的工作还应该加大，加强用户的信用评估，同时也可以从这个过程中发现问题，针对问题找到解决方案来完善征信体系，评价指标等。

综上所述，本方案针对Lending Club某季度的用户信贷数据，通过数据的处理、分析等，进行用户信贷违约风险的预测，即对新增贷款申请人预测是否会违约，从而决定是否可以放款。同时也给出相应的预测模型和方案，从而能通过方案去解决相应的一些问题。

**三、数据分析及模型处理**

**1.任务目标**

用户信贷风险预测的任务是通过公司给出的用户贷款行为（如贷款数据的各种特征和贷款状态是否违约）来训练模型，通过模型对新增的贷款申请人进行预测，判断他（她）是否具有还款能力，因而判断该用户是否会发生贷款后的违约行为。

由此可知，该任务的本质是一个有监督的学习，因为贷款数据除具有各种特征属性列以外，还有目标列，即用户是否会违约的一个标记特征。此外，该目标列在结果中只有两种取值，即违约和不违约；故可判定贷款申请人是否违约是一个二分类问题，因此可以选择合适的模型进行最后数据的训练。

**2.数据集的分析及预处理**

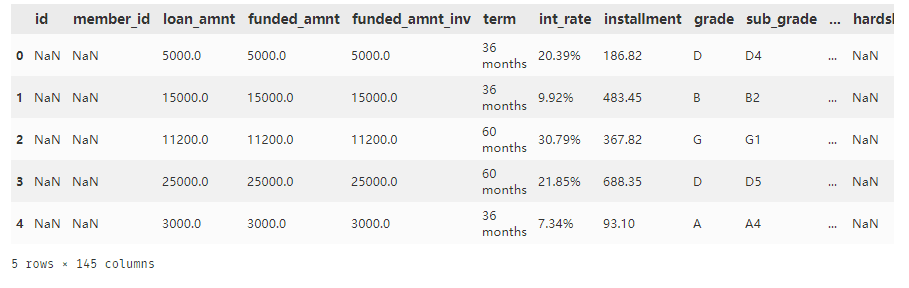
**（1）数据概况**

图1 贷款数据集头部数据预览

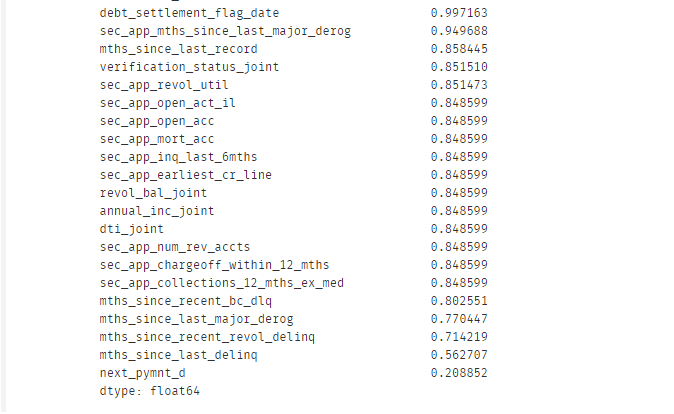


图2 数据集部分特征缺失比例

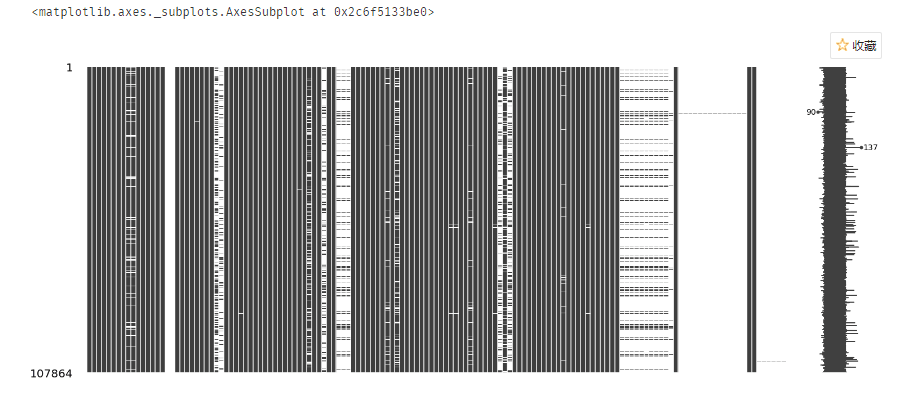


图3 数据整体缺失情况

【分析】：从图1可知，数据集特征包括数值特征和非数值特征。图3展示了数据所有特征列中的缺失情况（图片由missingno库评估而得，在missingno库所得图中，某一列特征若有缺失值，则会用一条白线表示，因而可根据白线的程度判断缺失的情况）。图2中则列出了部分特征的缺失值比例，从图2和图3中可以看到，有较多的特征缺失较为严重。

【解决方法】：从数据所表现的信息情况来看，数据中严重缺失的数据特征列对分析的意义是不大的，而缺失较少的特征，通过处理则会有益。在这里，有两种处理方式：一是直接删除，而是填充处理。

结合特征的意义，删除了缺失值比例超过0.4的数据特征，选择这个比例的理由如下：缺失比列超过0.4的特征除一个缺失比例为0.5（此特征对模型影响并不大）的以外，都是0.7以上的，这些特征在经过分析后发现都是对模型预测意义不大的，是可以直接删除的。对于缺失较少的则分为非数值特征和数值特征处理。

从方案角度想，严重缺失的特征信息可能也是不怎么强调用户填写的，因为它本身对用户画像的描写意义就不大。

**（2）非数值/类别特征的处理**

首先查看数据集中的类别特征数据及缺失情况，如下图：

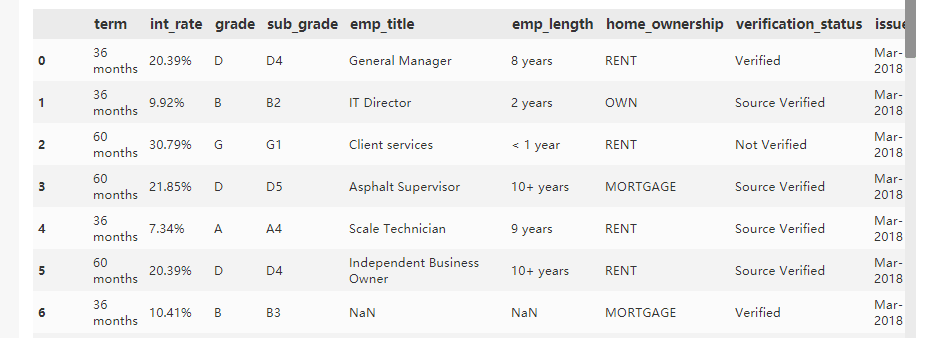


图4 类别特征头部数据

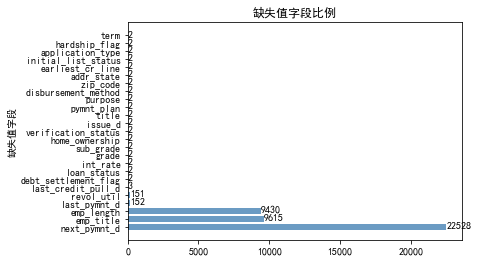


图5 类别特征缺失情况

【分析】：图3中，仔细查找后有两个比较重要的特征（int\_rate, revol\_util）实际上是数值特征，而因为含有“%”后缀而被pandas认为是字符串，因而被当成了类别特征，故需要将其转换为数值特征。

图4中是部分非数值特征的少数缺失情况，仍有几个特征的缺失值比较多，但可能会对模型的训练造成一定的影响。由于已经是经过第一步删除处理过后的数据，不能再进行删除，所以考虑填充。

对于这些缺失值，实际中可能因为是涉及到用户部分不愿意透露的信息，所以用户未详细填写，但大部分用户是不介意的。所以从方案角度来说，这一部分缺失值是有意义的。

【解决方法】：对于图3中的问题，通过一些基本操作将该非数值特征转换为float类型的数值特征。对于图4中，由于是非数值特征，而且大部分的特征缺失值比较少，故暂时先用unknown填充，表示未知。处理前后的缺失值比较如下图：

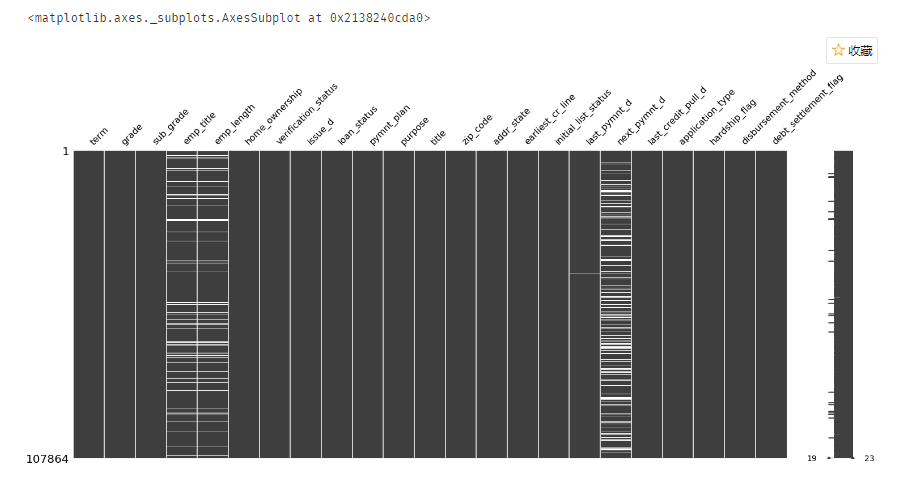


图6 非数值特征填充缺失情况

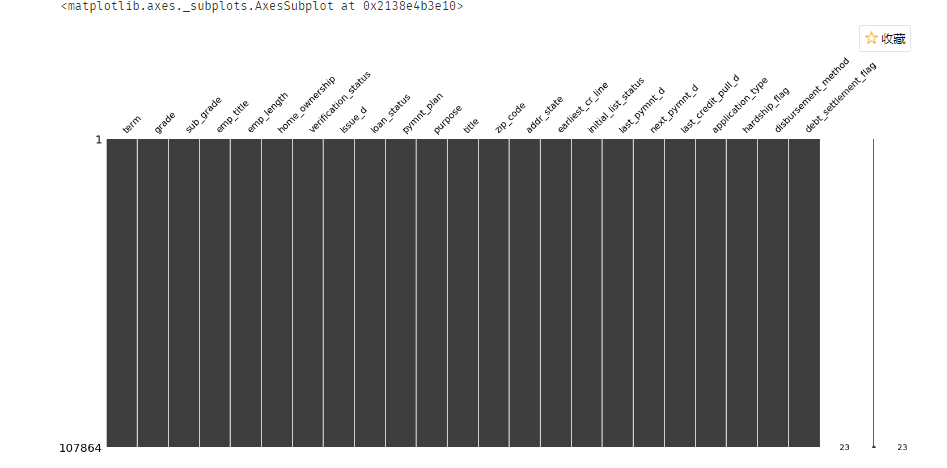


图7 非数值特征填充后情况

**（3）数值特征的处理**

同样，首先也是查看数值特征数据以及缺失情况，如下图所示：

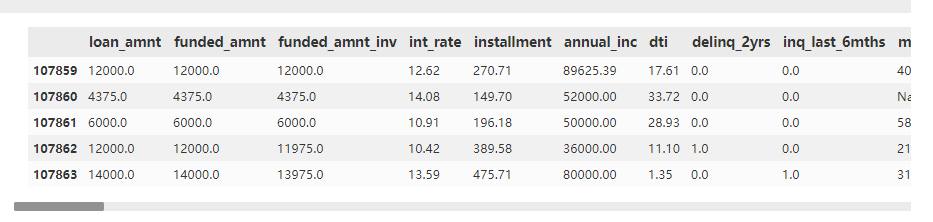


图8 数值特征尾部数据

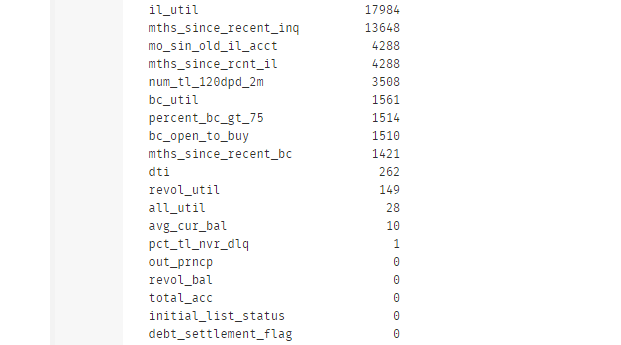


图9 数值特征缺失情况

【解决方法】：对于剩下的数值特征的缺失不是很严重，同时填充的处理方法也比较多，比如用平均值，众数，最大数等填充，或者也可以预测填充。在我们所用的数据环境中，我们选择了众数填充的方式。不过大多数数据使用众数填充的效果是比较好的。

**（4）数据再清洗**

再次对数据集的各个特征查看，特别是非数值特征，因为有的模型是不能对非数值特征进行处理的，而且有的非数值特征因为取值过多，可能还需要考虑其它方法，这些则放到后面的特征抽象中讨论；而有的非数值特征几乎每条数据都不一样，这种特征大部分都是对模型没有用处的，因此，数据再清洗则是考虑到这些特征，根据特征属性的含义，人为的判断出对模型没有意义的特征并把它删除掉。

通过斟酌后，像“title”，“disbursement\_method”等这样的特征，从字面上便可知对模型的作用是不大的，故删除掉。

**3.特征工程**

从解决方案的角度出发，特征工程的作用是更加详细的描述用户的画像，使得根据用户填写的信息特征更加能用于判断用户的信贷诚信行为。对于这一部分则分特征衍生、特征抽象、特征标准化和特征选择4个方面阐述。

**（1）特征衍生**

**<1> "installment\_feat"：用户每月还款支出占月收入的比。**

数据集中，"installment"代表贷款每月分期应还的金额（月负债），"annual\_inc"代表申请人的年收入，我们将"annual\_inc"除以12得到贷款申请人的平均月收入金额，然后再把"installment"（月负债）与（"annual\_inc"/12）（月收入）相除生成新的特征"installment\_feat"，新特征"installment\_feat"代表客户每月还款支出占月收入的比，"installment\_feat"的值越大，意味着贷款人的还款压力越大，违约的可能性则越大。

**<2> "emergency\_repayment"：用户是否能完成紧急情况还款**

数据集中，"avg\_cur\_bal"代表用户当前所有账户的平均余额，"num\_actv\_bc\_tl"代表用户当前活跃的银行卡账户数，"loan\_amnt"代表用户贷款数目，"avg\_cur\_bal"乘上"num\_actv\_bc\_tl"便可在一定程度上理解成用户所有账户的总余额，再减去"loan\_amnt"，则可以根据结果的正负判断在紧急情况下用户是否可以马上还款。用户是否能马上还款对于放款者的判断在一定程度上也是有影响的。

**<3> "finish\_in\_time"：用户是否能在指定时间内完成还款**

数据集中，"term"代表用户所填的还款月份。我们用"loan\_amnt"除以"annual\_inc"得到用户还款的年数，再减去"term"/12，根据得到结果的正负则可理解为用户是否可在所填的时间内完成还款。对于结果为正的数据，在一般情况下，当然有理由认为该用户具有偿还贷款的能力。

**<4> "reputation"：用户的信誉**

数据集中，"num\_accts\_ever\_120\_pd"代表用户当前逾期120天的账户数，"num\_bc\_tl"代表用户银行卡账户总数；用"num\_accts\_ever\_120\_pd"除以银行卡账户总数"num\_bc\_tl"得到用户逾期的银行账户比例。不过对于这一个特征，可能不是用于大部分，因为大多数的人多的账户都是比较少的，但是可以通过设置一个阈值，筛选出信誉特别差或者特征好的，另外也可以对结果进行分级，即在这方面反映用户一定程度的信誉。

**（2）特征抽象**

**<1> 对label特征的抽象**

Lending Club的数据中，贷款状态loan\_status这一特征就相当于目标列，但是数据中的这一特征有多种情况，不过对这些状态都可以分为是否违约两个类别。所以有必要对这一目标特征进行抽象，以将问题转化为为二分类问题。

loan\_status的不同状态中，我们将Default，In Grace Period，Late (31-120 days)，Late (16-30 days)，Charged Off映射为1，即已违约；将Current，Issued，Fully Paid映射为0，即未违约。这样，我们便也把这一列类别特征转化为了数值特征，同时也转化为了二分类问题。

处理过后通过可视化查看该特征的情况，如图9，很显然以及转化完毕。同时可以看出0,1两个类别之间的数据大小差异，其中违约的数据相比未违约的是比较少的，这也符合现实情况。

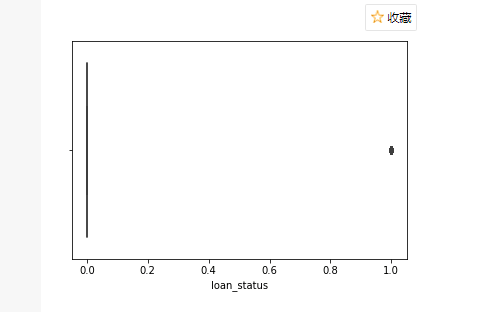


图10 贷款状态情况

**<2> 对就业年限emp\_length的抽象**

由于就业年限这一特征在很大程度上反应一个人的诚信程度，所以也需要对这一特征进行处理，除此之外，这是一个有序特征，跟抽象后的数字大小是有联系的，所以有必要进行一些大小上的区分。

这一特征的取值是在0-10之间的，单位是年，所以除开将10+year映射为10，将<1 year和unknown映射为0之外，其余映射为对应的年限数字。比如8year映射为8，这样处理之后，便将这一特征转化为了有序的数值特征，因此可以直接放入模型之中进行训练。

**<3> 信用证指定贷款等级grade的抽象**

grade是由Lending Club指定的一个贷款等级，体现用户贷款信用的程度。以给放款方一些决策帮助。可以知道的是，这一指标也是放款方着重参考的一个点，所以它的重要性也是挺高的。

在数据中，grade有A，B，C，D，E，F，G共7个等级，因此可以直接对应一个数值，即1-7，这样也将其转换成为了一个有序特征。

**<4> 对剩下的几个类别特征：home\_ownership, verification\_status, application\_type, purpose, term的处理。**

考虑到在最后需要使用到的模型Logistic Regression，RandomForest等不支持字符串类型的非数值特征，所以需要对剩下的几个类别特征进行处理。当然有的模型比如LightGBM可以直接放进去，但处理效果不一定能和人为预期一样。

为了较好的区分每个特征的不同类别取值，可以使用one-hot编码，对剩下的几个特征进行one-hot处理后，将离散特征的取值扩展到了欧式空间，分类器计算通过距离计算相似度时也更加方便了，即解决了分类器不易处理属性数据的问题，同时，也起到了扩充特征的作用。

**（3）特征标准化**

由于数据的问题，可能有的数据在处理之后仍然可能有一些异常值，或者分布与实际分布可能存在差异，或者可能收集到的数据有一些小的量纲问题等，所以需要对特征进行标准化处理，采用的是sklearn模块preprocessing的子模块StandardScaler。

**（4）特征选择**

**<1> 初步筛选**

在前面的所有处理完毕之后，特征还有100个左右，很明显的是，有很多特征是没有必要的，而且是冗余的，接下来要做的则是初步选择出一些相关性较高的特征，再从剩下的特征中进行筛选，同时也能在后面模型处理部分降低模型训练的难度和时间消耗。

在这里，以Recursive Feature Elimination ( RFE ：递归式特征消除)方法进行特征选择为例，还有其它的特征选择方法也适用。该方法根据学习器计算出的特征重要性，递归的删除重要性最低的特征，直到达到所需要的特征个数为止。经过特征重要性的一些查看以及特征训练的难易和时间消耗程度，我们最后选择筛选出30个特征。

通过筛选过后，画出剩下特征的皮尔森相关性图谱，通过观察再进行下一步的筛选。

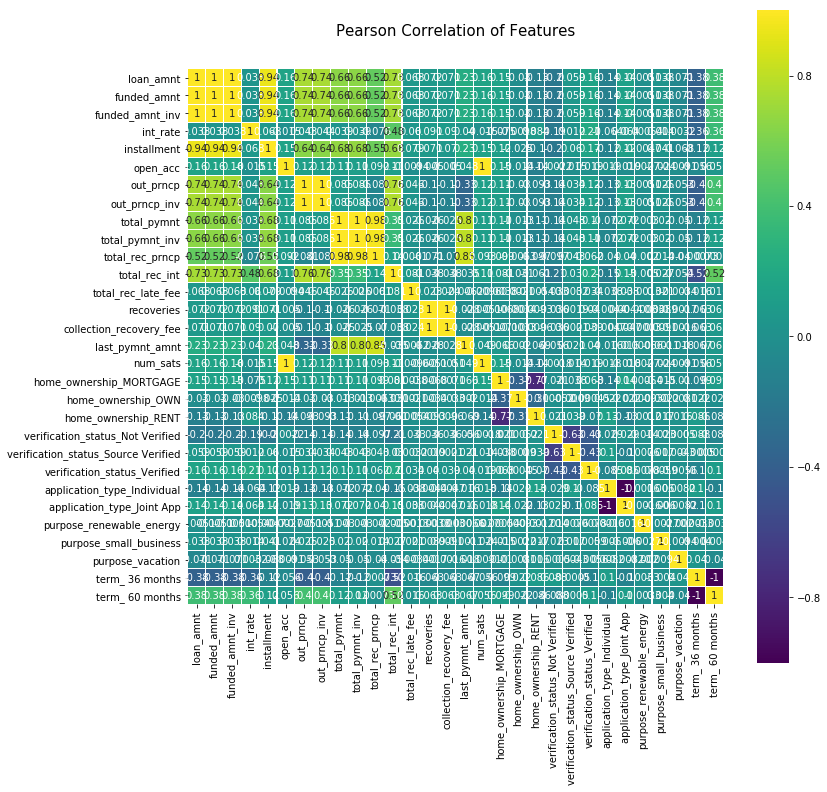


图11 初次筛选后特征相关性图

图中显示了所有特征两两之间的相关性，这就很容易观察到特征之间的关系了。从图中便可以看出哪些特征是多余的，哪些太低而没有作用的等。

**<2> 再次筛选**

通过观察图10发现，很多特征之间的相关性都是较小的，只有少数是关系比较大的，而且有很多特征是多余的。

所以在这里删除了15个重要性低的特征，留下15个更为重要的特征。再次查看其皮尔森相关性图谱如下：

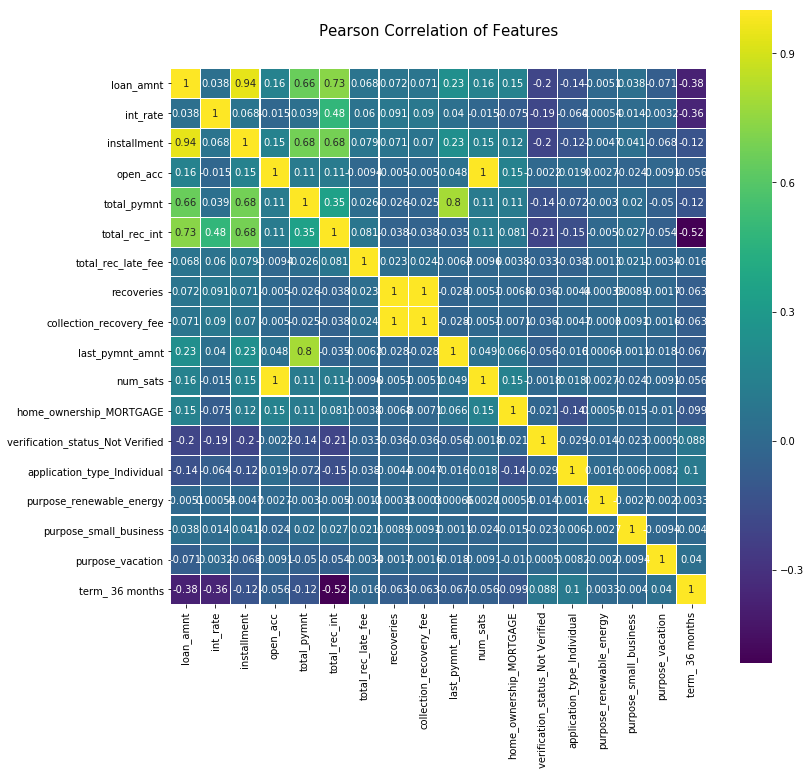


图12 再次筛选后特性相关性图

**<3> 特征重要性评判以及排序**

选择好了特征之后，还可以通过学习器学习之后，输出特征重要性进行相关的评判，用以判断所做的特征选择是否正确。另外则是可以挖掘哪些特征变量的重要性很好，可以降低学习的难度，在优化模型上起到一定的效果。

留下的特征的重要性排序如下图12，特征的重要性柱状图及步阶图如图13：

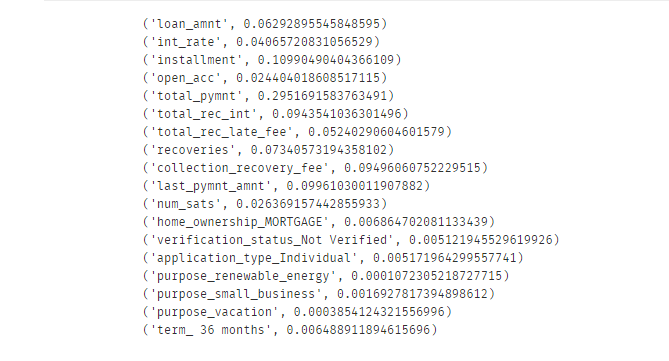


图13 特征重要性排序输出

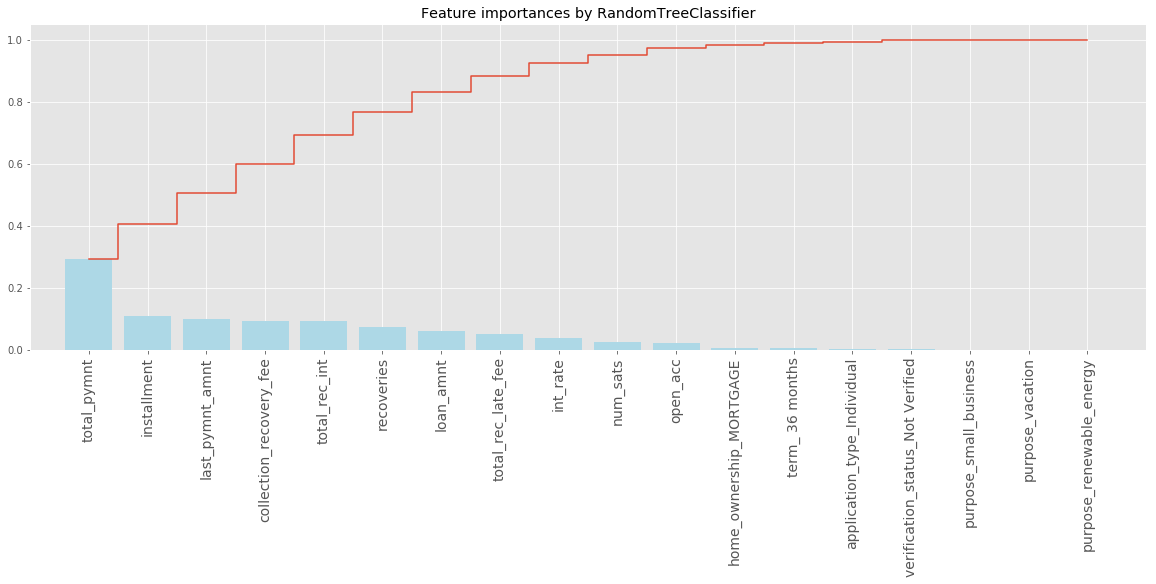


图14 特征重要性柱状图和步阶图

**4.模型训练**

**（1）样本数量均衡处理**

在前面对目标列loans\_status输出的时候我们发现，违约数据和未违约数据之间的差距是很大的，而且这种差距会对模型训练造成很大的影响。所以为了使训练效果更好，采用过采样的方法来处理样本不均衡的问题。

算法则是使用SMOTE（Synthetic Minority Oversampling Technique）算法进行处理，算法原理是采样最邻近算法，计算出每个少数类样本的K个近邻，从K个近邻中随机挑选N个样本进行随机线性插值，构造新的少数样本，同时将新样本与原数据合成，产生新的训练集。

**（2）模型构建与训练**

模型的训练采用交叉验证法划分数据集，将数据划分为3部分：训练集（training set）、验证集（validation set）和测试集（test set）。让模型在训练集进行学习，在验证集上进行参数调优，最后使用测试集数据评估模型的性能。

模型调优采用网格搜索调优参数（grid search），通过构建参数候选集合，然后网格搜索穷举各种参数组合，根据设定的评分机制找到最好的一组设置。

在模型上课根据需要选择，这里使用了五个常见的处理分类问题的模型，即：LogisticRegression，决策树，随机森林（包括以gini系数和信息增益为分类依据两种），XGBoost，LightGBM。它们各自都有自己的优点，我们则通过效果选择好的模型，或者做融合。

**（4）评判指标**

评判指标使用的是一般常用的AUC评估方法。相比于ROC曲线，它定义为其曲线下面的面积，更加能清晰的反应模型的好坏。所以，AUC值更大的模型，效果则越好。

通过模型训练后，将每个模型测试得到的AUC值输出查看。其结果列在下表中：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模型 | Logistic  Regression | 决策树 | 随机森林（Gini系数） | 随机森林  （信息熵） | XGBoost | LightGBM |
| AUC值 | 0.930629 | 0.957753 | 0.976464 | 0.977710 | 0.981896 | 0.987091 |

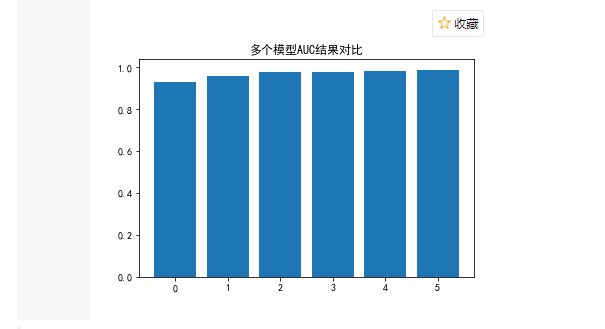


图15 多个模型的AUC结果值比较

【分析】：从表中以及以上的几个模型的预测结果对比来看，都是90%以上，不过我们不应该觉得很好，因为像贷款这样的工作，在100个人中若有10个人违约，那样的结果将是难以想象的。

从结果来看，LogisticRegression的预测结果是最差的，而且几乎是投资者不愿意接受的。当然，也可以看到XGBoost和LightGBM的效果达到了98%，甚至快达到99%，这样的效果还可以，的确，这两者的模型是经过了各种优化了的，对数据的处理比较到位。另外，在决策树和随机森林的两种模型上，其效果也还差强人意。

当然，对这些模型的结果我们可以做融合。对于分类问题，可以使用投票的机制进行。不过，在这里经过我们的尝试，发现效果是和LightGBM是差不多的。通过更多实例来看，组合的效果还是最好的，因为它能减小一些异常的影响。

**四、解决方案**

**1.方案主要工作**

通过以上所做的工作，我们可以思考这些工作的每个方面有哪些可以加强的地方，或者从这些方面，提高获取的数据质量，从而利于模型的预估，使得方案更加完整。我们总结了方案的主要工作如下：

1. 涉及信贷方面的公司或者银行等得到用户的历史贷款记录数据，并处理成表格的形式作为训练集；同时将需要预测的用户的信息汇总成同样维度的表格，作为测试集。
2. 将得到的训练集和测试集，通过我们方案中的数据分析方法进行数据处理和清洗，并做好特征工程方面的工作，得到最后的数据。
3. 将处理后的训练集放入我们设计的模型框架中进行训练，得到预测模型，然后则将测试集数据放入模型中进行预测，得到预测的AUC值。

通过我们的方案得到结果后，公司或者银行可以将得到的结果作为一个参考，再结合用户的一些其他行为进行判断是否可以对用户进行放款。

**2.方案优点**

（1）从得到杂乱的数据到输出预测结果，我们的方案有一个完整的数据分析和处理框架，能适用于大部分维数多，数据杂乱的数据。

（2）在性能上，我们的预测模型比较稳定，不会随数据的杂乱变化而差异很大，这也给了我们方案有力的支撑。

（3）在我们数据分析的基础上，可以根据需要添加一些特殊的处理，因此，我们方案的泛化能力也较强。

（4）除开一些处理操作外，我们的模型训练速度快，具有实时性和准确性。

**3.方案使用建议**

此外在以下几个方面，我们分别提出了一些可以提高方案在实际环境中的应用程度的建议。

\*数据质量方面：

（1）政府可完善当前的征信体系，或者建立更好的征信系统，加强用户信用的记录，且使得用户信用更接近于真实情况。

（2）信用贷款管理方可以根据信息的重要性，强调用户在贷款时对这些信息的填写。

\*数据处理方面：

（1）做好数据处理的基本工作，要详细处理，比如缺失问题，根据实际情况删除。

（2）因为同类但不同样的数据可能存在差异，每一次的数据可以尝试使用与当前不同的处理方法，而不是每次都使用相同的处理操作。

（3）建立多套完整的数据处理方法框架，同时附加一些可根据实际情况添加的方法，根据数据情况使用不同的处理。

\*特征工程方面：

（1）挑选出特征重要度高的信息列，让用户优先填写；此外，如果认为有的缺失较为严重的信息可能比较重要，可以多次让用户填写，直到其质量比较稳定。

（2）加强特征衍生的工作，挖掘更有用的信息，使得对用户画像的描述更加清晰，这样才能更加了解用户。

（3）根据业务的逻辑选择或者构造特征，也可以采用数据挖掘中的一些合适的算法。

\*模型方面：

（1）如果当个模型效果好，可以根据效果选择适合的模型，不过尽量还是多个模型的结合，考虑各个模型的优点，取长补短。

（2）根据数据的分布或者类型，设计适用的评估方法，使其能在信贷方面广泛应用，这里则适用的是一般的AUC方法来评估。

（3）通过模型的作用，选择出信贷违约的大体人群，对这些人群进行跟踪。

总的来说，要保证信贷预测更加准确，则需要做好在每一方面的工作，从分析中我们可以发现数据的质量是很重要的，如果说数据质量都不能保证，那么分析和预测自然也没了保障。当然，数据分析与处理，特征工程和模型预测也同样重要，只有加强了这些方面的工作，我们才能更好的挖掘出更多的有用的信息。