

**机器学习与模式识别研究报告**

题目：家庭信用风险评估

学院：计算机科学与软件工程学院

专业：计算机科学与技术

姓名：xuzhongyou

学号：51184506050

# 一、研究背景与现状

## 背景

随着生活水平的提高、保守消费观念的改变、亦或者出于投资方面的考虑，人们对于贷款业务的需求越来越多。p2p网络信贷作为微型金融领域的一项创新，通过电子商务公司提供在线交易平台向其他人提供小额信贷的金融模式，随着行业认知度的提高，网络信贷公司的数量和成交金额迅猛增长。与此同时，信贷风险是运营过程中的主要风险，风险管理对行业发展起着至关重要的作用。

## 现状

传统金融机构对个人信贷风险的评估集中于对债务人的风险评级，对债务人的风险评级就是对借款人进行资信评估，得到个人信用得分，划分信用等级，以此为借贷风险大小的判断依据。建设银行的个人资信评级体系符合我国个人信贷基本情况，是一种适用性广泛，科学性较强的个人信贷风险评估方法。这种评估从业人员要求专业高，对于小额借贷评估和贷款人数增加面临着挑战。

从信贷技术方面，p2p信贷机构从业人员专业技能水平偏低。此外，参与借贷交易的投资人文化层次差别大，对信贷风险相关知识的理解程度深浅不一。在选定p2p网络信贷风险评估时应避免使用过于复杂，难以掌握的方法。

在信贷模式以及评估对象特征的差异方面，对于分期还款方式，非常重视债务人还款期间的现金状况。例如德国的IPC技术的信贷风险评估坚持以‘现金流’为核心，分析真实的财务情况，客观的判断债务人生意状况或者企业价值。比较中国建设银行个人资信评估对象侧重与稳定单位的个人，德国IPC技术评估的对象大多是个体工商户。IPC公司信贷技术的核心，是评估客户偿还贷款的能力。主要包括三个部分：一是考察借款人偿还贷款的能力，二是衡量借款人偿还贷款的意愿，三是银行内部操作风险的控制，FICO评分则是为美国民众设计的，FICO 评分系统得出的信用分数范围在300- 850分之间。分数越高, 说明客户的信用风险越小。具体可参考【13】【14】。

## 方法

目前在传统评估方法基础上，已有采用机器学习并取得不错效果的图模型方法有：基于树增强朴素贝叶斯模型，基于马尔科夫毯模型，以及基于特征选择的马尔科夫毯模型等。然而熟练掌握这些模型并兼顾复杂度，效率，准确率，可解释的情况下，往往取得效果未能尽如人意。本课题旨在用工业界和竞赛中常用方法，训练得到较高准确率的简单的线性分类器和解释性更强的树分类器以及集成策略来构建个人信贷风险评估模型。

# 二、研究内容

## （一）、数据集来源

Kaggle上今年Home Credit提供的一个竞赛数据集，致力于向无银行账户的人群提供信贷，预测客户是否偿还。该数据集有七种数据来源，主要分为训练集和测试集。

## （二）、数据预处理

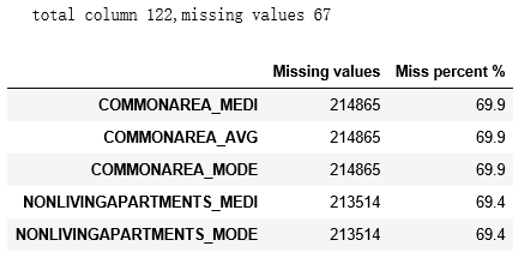
### 1、统计训练数据



其中按期还款记录282686，逾期记录24825。这里存在一个典型样本不平衡问题。在用复杂机器学习模型时，可采用对类进行加权来解决。

### 2、缺失值，异常值处理

统计训练中的缺失值：

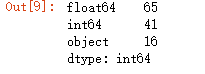


这里只展示了前五个缺失值的column,发现缺失率接近70%，在特征工程部分可以参考这些column的缺失情况。

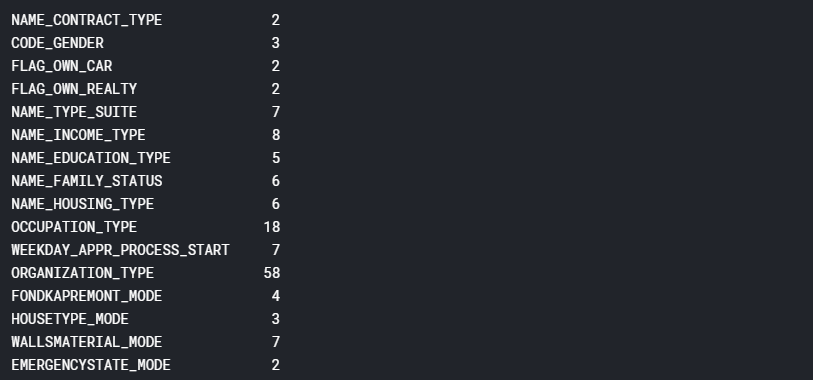
对于少数column值缺失采用插补(平均，众数等)，对于一些较多缺失的column考虑删除。需要去检查特征的异常值。比如：DAY\_BIRTH为负。

### 3、categorical to numeric

查看column的数据类型



这里有16个column为object类型。继续查看这16个column对应的种类：



可以看到大多数column的类别还是少的，在10个以内。

column type为categorical需要转为numeric。编码categorical类型变量有两种：1.Label encoding。2.one-hot encoding。Label encoding 给类别任意顺序且类别值是随机，但是因为是任意的会带来一些模型权重上的问题。因此一般类别两个时可选用Label encoding,类别更多时选用one-hot encoding。然而选用one-hot的问题是导致feature变多，造成维度灾难，这会一定程度上使得建模困难。一般做法除了人工选取有用的特征，会选择降维手段减少feature：如基于距离矩阵的等度量映射（Isometric Mapping）,主成成分分析（Principal Component Analysis）,核化线性降维。

### 4、关于对齐数据部分

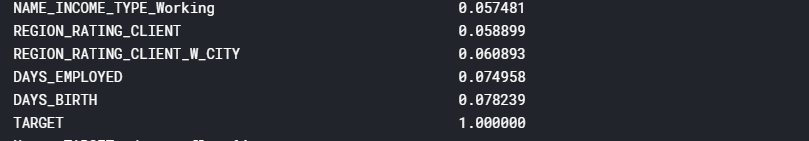
训练数据属性和测试数据的属性位置应当一致(除训练数据的label)，大部分数据训练集和测试集column的位置是一致的。需要注意的是对于监督学习任务，训练集多个label属性。

### 5、异常数据的分析与处理

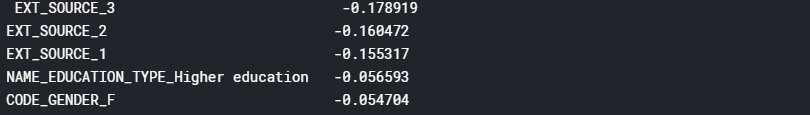
如数据集中的‘DAYS\_BITRH’和‘DAYS\_EMPLOYED’column值为负数，原因是从开户那天开始计算，出生和工作日期相对于开户是更早的时间，为了方便建模会将负值转为正值。其次，在分析过程中发现性。‘DAYS\_EMPLOYED’值为‘365243’的异常值，那么是否异常值的存在更有可能导致逾期还款呢？统计结果显示：在非异常数据中，有8.66%的逾期还款。在异常数据中，有5.40%的逾期还款，发现异常数据反而有更低的逾期还款率。

## （三）、特征分析

在特征提取过程：特征的相关性分析，尝试和理解特征和目标之间的相关。可通过Pearson系数计算每一个column和target的关系来观察。分别看下前五的正相关：



前五的负相关：



可以观察并初步分析出：随着年龄的增长，借贷者更大可能按时还款。进一步我们可以看下整个年龄的分布：



这张图仍不是很直观：进一步挖掘年龄对违约的影响：可以将年龄划分为间隔为5的一个区域，求出每个区域违约还款的概率（因为违约的label为1）。



可以看出与刚才皮尔逊系数分析的结果相同，随着年龄增长，更大的可能性按时还款。

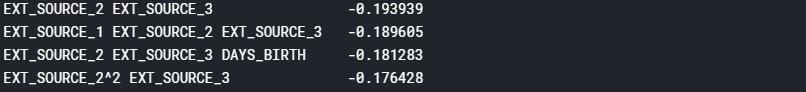
对于无关的特征，可以直接去除，一定程度上降维（谨慎使用）。除此之外可以通过简单的线性模型去测试，先提取部分有效特征也是常用的方法。特征工程中要去掉一些无用的特征如数据条目的id号。数据降维的方法有很多大致分为线性和非线性降维，其中非线性降维又分为基于核函数和基于特征值的方法。常用的降维方法PCA。

## （四）、特征工程

好的机器学习效果建立在好的特征上，特征工程是最为关键的一步。特征工程包括选取原数据集中有用的特征即feature selection ，还包括构造出来的新特征即feature construction。这里的特征工程是建立在对target影响较大的几个属性上所做，如‘EXT\_SOURCE\_1’,‘EXT\_SOURCE\_2’,‘EX\_SOURCE\_3’,‘DAYS\_BIRTH’。

feature construction：

方法一：Polynomial Feature,其方法思路其实很简单。假设有feature（a,b）,那么经过2项多项式得到的feature为(a,b,ab,a^2,b^2)。该方法的好处是：单个变量本身可能对目标没有太大的影响，但是变量组合而成的交互项，捕捉变量之间的交互可能会突出显示出与目标之间的关系。当然是否采用新的feature取决于实际实验中的效果。在本项目中，我们选用了‘EXT\_SOURCE\_1’,‘EXT\_SOURCE\_2’,‘EX\_SOURCE\_3’,‘DAYS\_BIRTH’来构造new feature。为了对比效果，利用Polynomial Feature做feature construction和不添加新的feature看对模型效果产生的影响。但是计算新的feature和label之间的皮尔逊系数发现：新构造的feature的相关性比原始feature的相关性更强一些。除此之外，一般情况下都是选取几个new feature，它的选取也是多次尝试实验的结果。其中EXT\_SOURCE\_表示的是来自外部数据的标准化得分，可以理解为信用评级。



上表格为多项式得到部分feature与目标之间的皮尔逊系数

方法二：Domain Knowledge Feature。相比于方法一，需要专业的知识来构造特征。这种先验知识下构造往往能选出对模型有很大的提升的feature,并且缩短feature construction的时间。本项目中，根据金融的一些知识，构造新feature:Credit\_income\_percent,Loan\_income\_percent,Credit\_term,Days\_employed\_percent，分别表示信用卡占收入的比率，借款金额占收入的比率，支付期限，工作时间占年龄的比率（天为单位）。

## （五）、模型选择

用简单的线性模型做baseline，在此基础上比较不同的模型，可使用交叉验证的方式选出最优模型，亦或者使用集成学习策略确定模型，再通过trick来提高分类的准确率，这里模型好坏的标准以提高Kaggle上的准确率为标准。可尝试的模型有：Logistic Regression,RandomForest,SVM。复杂模型相比于简单模型不一定就能取得更好的效果，取决于应用场景以及数据清洗和特征工程，尤其是后两者在真实应用场景下尤为重要。

# 研究方法：

## （一）、Logistic Regression:

一个分类器，将线性回归的结果通过Sigmoid映射到[0,1]的域值空间里。线性回归指的是构造一条直线拟合历史数据用这条直线对新的数据进行预测。总的来说就是用线性回归模型的预测结果去逼近真实标记的对数几率。

Logistic Regression分类器的优点是：直接对分类可能性建模，无需事先假设数据分布，避免了假设分布不准确所带来的问题；它不是仅预测出‘类别’，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用。十分适用于本项目，给出按时还贷的概率。

采用sklearn包里面的Logistic regression,它的目标函数为：



采用的L2正则。

使用Logistic Regression 作为baseline分类器，feature是feature engineering之前的，做了简单的数据异常，数据缺失值处理。Baseline的提交到kaggle上的结果为：



## （二）、Random Forest:

Random Forest是Bagging的一个扩展变体，其基学习器是一个决策树，在决策树的训练过程中引入了随机属性选择。其基本思想是：对每个决策树的每个节点，从该节点的属性集合中随机选择包含K个属性的子集，再从子集中选择一个最优属性划分，参数K控制了随机性的引入程度，推荐是K=log2d。

Random Forest思想简单，容易实现，计算开销小，却展现出强大的性能。主要原因是：样本的扰动来自于Bagging策略，属性的扰动来自于(Random Forest 中Decision Tree 随机选择包含K个属性的子集)。

Random的结合策略主要有：简单平均法，加权平均法，投票法，加权投票法以及学习法。

关于决策树，Bagging策略，以及Random Forest具体细节原理参考【1】【2】。



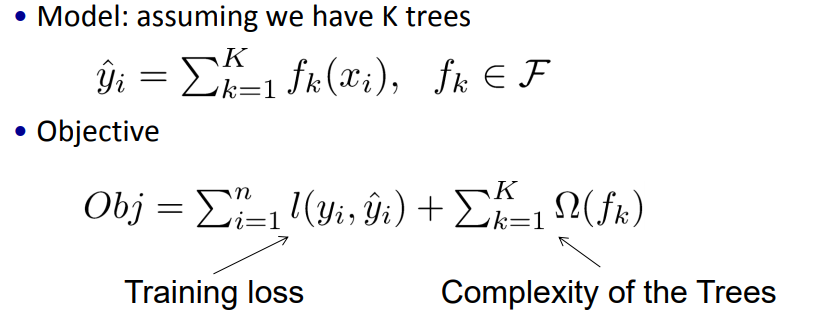
## （三）、SVM:

在维度空间中构造超平面，将训练数据可分。为了缩小分类的泛化误差，分类边界距离最近的训练数据点越远越好。在原始维度中线性不可分的情况，通过核函数映射到高维空间再找超平面。

## （四）、GBDT，xgboost，lightgbm:

Gradient Boosting Decision Tree,Boosting是一族可将弱学习器提升为强学习器的算法：先在初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，使得先前基学习器做错的训练样本在后续收到更多关注，然后基于调整后的样本分布来训练下一个基学习器，如此反复，直至基学习器数目达到事先指定的值T，最终将这T个基学习器进行加权结合。不同Boosting的策略区别在于采用何种方式？Adaboost采用增加上一轮学习错误样本的权重的策略，Gradient Boosting 中则是将负梯度作为上一轮基学习器犯错的衡量指标，在下一轮学习中通过拟合负梯度来纠正上一轮犯的错误。GBDT 与传统的Boosting区别较大，它是一种基于前向策略的加法模型，它的每一次计算都是为了减少上一模型的残差，为了消除残差，可以在残差减少的梯度方向上建立模型。GBDT 的每一基学习器是一个回归树，累加每个树的结果作为模型的输出。因此GBDT在做分类任务时选择 deviance loss 和 exponential loss。

xgboost是在GBDT上改进而来，传统GBDT在优化时只用到了一阶导数信息，xgboost则是对代价函数进行了泰勒展开，只要函数可一阶二阶求导。其次，xgboost在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数，每个叶子节点上输出的L2模的平方和。正则项降低了模型的variance,使学习出来的模型更加简单，防止过拟合。寻找分割点时，采用的是传统的枚举每个特征的所有可能分割点的贪心算法。



xgboost 的代价函数

xgboost在计算开销上非常大，预排序算法在选择好分裂特征计算分裂收益时需要遍历所有样本特征值。

Lightgbm是微软的gradient boosting 框架，相比xgboost有更快的训练效率，更低的内存使用，同样支持并行，可处理大规模数据。采用的是基于Histogram的决策树算法，直方图做差加速，带深度限制的leaf-wise叶子生长策略。该项目中关于boosting策略的方法采用的是lightgbm。

## （五）、神经网络:

训练数据的原始维度为122，采用两层全连接神经网络层数分别为100,64层。选择层数较低，因为原始维度不高，为了防止过拟合。通过神经网络训练来进行特征提取，最后加一个二分类器。通过调参训练，训练结果在0.70左右，并且仍然会出现过拟合现象。通过增加drop out等技巧，最终结果接近0.71。

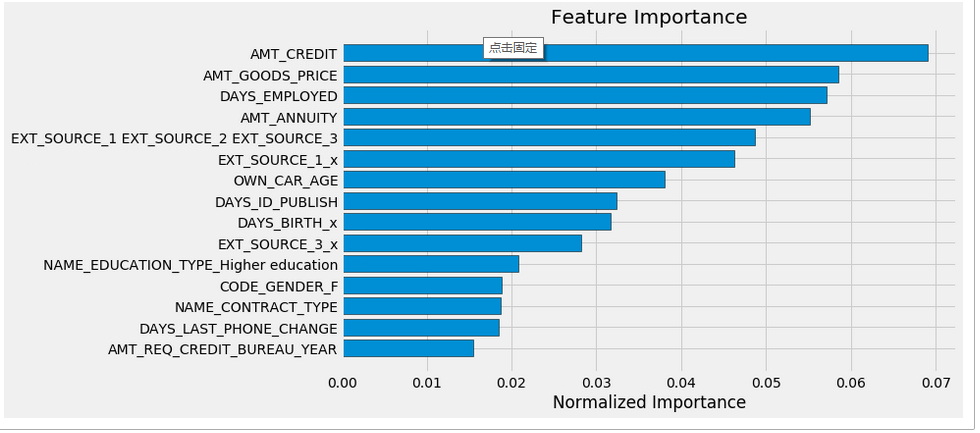
# 实验结果

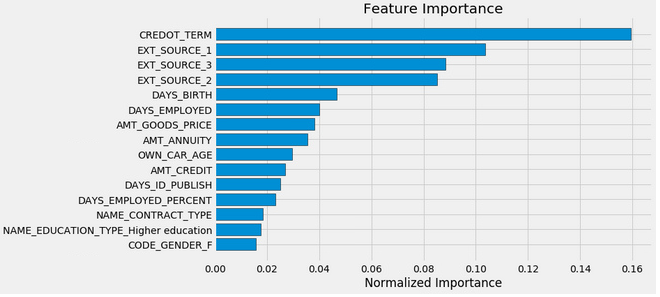
|  |  |  |  |
| --- | --- | --- | --- |
| Method | Polynomial Feature/yes | No feature engineer | Domain Knowledge Feature/yes |
| Logistic Regression | 0.71090 | 0.67041 | 0.67123 |
| Random  Forest | 0.60462 | 0.67863 | 0.67952 |
| GBM | 0.73237 | 0.73227 | 0.75097 |
| Neural Network |  | 0.71786 |  |

对比试验结果基于算法和特征

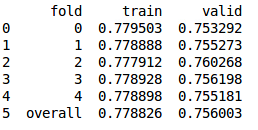


基于domain feature的lightgbm





同一模型下不同特征工程学习特征的重要性



Lightgbm模型下交叉验证的结果

# 总结

由不同方法不同的特征对比结果可知：（1）不同模型的复杂程度在原始特征上取得的分数不同，复杂模型在维度较多的数据集中往往能取得不错的得分。（2）采用集成学习比单学习器总体上取得的效果要好(实验中RF在Poly feature后得到较低的分数原因可能是特征工程生成的feature不理想)（3）交叉验证和正则化能够有效的降低过拟合带来的影响，在BGM中尤其明显。（4）不同模型，不同特征工程对最终学习到的特征权重不同。（5）当模型一定，做特征工程往往取得收益是巨大的，domain feature的效果明显优于poly feature，domain feature 需要一定的专业背景知识。（6）简单的统计观察，了解属性，初步分析属性对目标的影响，以及可视化对把握整个数据集和后续的建模有很大的帮助。（7）根据kaggle上的提交记录，我们的模型仍然有提升的空间，比较不同之处，在特征工程上仍旧可以提取更多有用的特征。（8）该项目还有多个其他的数据文件考虑到专业知识上不足，并没有选择全部文件，想取得更高的分数，这些文件可以适当利用。

在本项目中按时还款和逾期还款分别为282686,24825条，存在类别不平衡问题，但是测试提交的结果是逾期还款的概率，并不是真正意义上的二分类，在实际应用中可通过控制阈值来缓解该问题。

关于svm算法的实验结果，由于数据量太大，svm的训练量过大，训练时间过长（30w条数据服务器跑了一天仍然没有结果，换成1w条数据可能因为数据量的问题，最终得分只有0.55093，远低于其他算法），没有作为主要的机器学习算法去实践。关于使用神经网络自动提取特征，对调参技巧有一定的要求，对于任务简单，分类效果不一定优于有专业指导下特征提取加简单分类器。

# 六、参考文献

【1】机器学习 周志华 清华大学出版社

【2】统计学习方法 李航 清华大学出版社

【3】Home Credit Default Risk https://www.kaggle.com/c/home-credit-default-risk/data

【4】https://scikit-learn.org/stable/modules/classes.html

【5】https://www.cnblogs.com/infaraway/p/7890558.html

【6】<https://zhuanlan.zhihu.com/p/24498293>

【7】<https://feisky.xyz/machine-learning/basic/feature-engineering.html>

【8】<https://blog.csdn.net/keepreder/article/details/47273297>

【9】<https://zhuanlan.zhihu.com/p/24825503>

【10】我国p2p网络信贷风险评估研究 广西大学

【11】我国p2p网络信贷借款人信用评估研究 江西财经大学

【12】基于贝叶斯网络理论的个人信用评价模型研究 安徽财经大学

【13】<https://www.cnblogs.com/nxld/p/6364341.html>

【14】https://www.jianshu.com/p/06121460ae8c