# 基于随机森林算法的贷款违约预测模型研究

摘要：如何在发放贷款前有效的评价和识别借款人潜在的违约风险，计算借款人的违约概率，是现代金融机构信用风险管理的基础和重要环节。本文主要研究借助非平衡数据分类的思想对银行等金融机构的历史贷款数据进行统计分析，并使用随机森林算法建立贷款违约预测模型。实验结果表型，随机森林算法在预测性能上超过了决策树和逻辑回归分类算法。此外通过使用随机森林算法对特征进行重要性排序，可以得到对最终是否违约影响较大的特征，从而能够更有效的进行金融领域的借贷风险判断。

关键词：随机森林，贷款违约预测，数据挖掘

**Research on loan default prediction model based on random forest algorithm**

**Abstract:** How to evaluate and identify the potential defaultrisk of the borrower before issuing the loan and calculate the defaultprobability of the borrower is the basis and important link of the credit riskmanagement of modern financial institutions. This paper mainly studies thestatistical analysis of historical loan data of banks and other financialinstitutions using the idea of non-equilibrium data classification, and uses arandom forest algorithm to establish a loan default prediction model. Theexperimental results show that the random forest algorithm surpasses thedecision tree and the logistic regression classification algorithm in theprediction performance. In addition, by using the random forest algorithm torank the importance of features, it is possible to obtain features that have agreater impact on the eventual default, so that it can more effectivelydetermine the risk of lending in the financial sector.

**Keywords:**Random forest, loan default forecast, data mining

**引言**

随着世界经济的蓬勃发展和中国改革开放的逐渐深入，无论是企业的发展还是从人们消费观念的转变，贷款已经成为企业和个人解决经济问题的一种重要方式。随着银行各种贷款业务的推出和人们日益膨胀的需求，不良贷款也就是贷款违约的概率也随之激增。为了避免贷款违约，银行等金融机构在发放贷款时会对借款人的信用风险进行评估或打分，预测贷款违约的概率并根据结果做出是否发放贷款的判断。如何在发放贷款前有效的评价和识别借款人潜在的违约风险，是金融机构信用风险管理的基础和重要环节，用一套科学的模型和系统来判定贷款违约的风险性可以将风险最小化和利润最大化。

本文主要研究如何借助非平衡数据分类的思想对银行等金融机构的历史贷款数据进行分析，并基于随机森林分类模型预测贷款违约的可能性。本文第一节主要介绍非平衡数据分类和随机森林算法；第二节主要进行数据预处理及数据分析。第三节主要构建预测贷款违约的随机森林分类模型，并得到该模型的评估结果AUC值，通过将随机森林算法与决策树和逻辑回归算法模型比较，得出随机森林算法更优的结论。最后通过对每个特征的重要性进行评估，得出哪些特征对最终是否违约的结果影响较大。第四节总结全文。

**1  随机森林算法**

**1.1  非平衡数据分类**

非平衡的数据即数据中某一类（多数类）的数据远远超过另一类（少数类）的数据，在网络入侵检测、金融欺诈交易检测、文本分类等很多领域普遍存在，并且在很多时候，我们只对少数类的分类情况更感兴趣。处理非平衡数据的分类问题可以通过正负样本的惩罚权重解决，思想是在算法实现过程中，对于分类中不同样本数量的类别分别赋予不同的权重，一般是小样本量类别权重高，大样本量类别权重低，然后进行计算和建模。

**1.2  随机森林简介**

随机森林（Random Forest）算法，是用随机的方式建立一个森林，是一种基于决策树的组合学习算法。随机森林的基本思想是在构造单个树的过程中，随机选取一些变量或特征参与树节点划分，重复多次并保证建立的这些树之间的独立性。在得到随机森林之后，当有一个新的输入样本进入的时候，森林中的每一棵决策树都会对该样本进行判断，得到该样本属于哪一个类的结果，最后看整个森林中属于哪一类的得票最高，就预测该样本为哪一类，其过程如图1所示。

图1 随机森林示意图

**1.3 随机森林算法原理及特点**

随机森林（Random Forest）算法，包含分类和回归问题，其算法步骤如下：

随机森林有以下特点：由上述算法过程可以看出，随机森林的随机性主要体现在两个方面：数据空间的随机性由Bagging（Bootstrap Aggregating）实现，特征空间的随机性由随机子样（Random Subspace）方式实现。对于分类问题，随机森林中的每一个决策树都对新的样本进行分类预测，然后以某种方式将这些树的决策结果集中起来，给出样本的最后分类结果。

1、数据中行（数据记录）和列（变量）两个随机性的引入，使得随机森林不容易陷入过拟合。

2、随机森林具有很好的抗噪声能力。

3、当数据集中存在大量缺失值（missing values），随机森林能对缺失值进行有效的估计和处理。//新用户

4、对数据集的适应能力强：既能处理离散型数据，也能处理连续型数据，数据集无需规范化。

5、可以能到变量重要性排序，方便对变量进行解释。随机森林中有两种计算变量重要性的方法：一种是基于OOB（Out of Bag）的平均下降准确率的方法。即在生长决策树的过程中，首先利用OOB样本进行测试并记录下来分错的样本，然后将Bootstrap样本中的某列变量的值顺序随机打乱，重新用决策树对其进行预测，再次记录下来分错的样本数。两次预测错误的数目除以OOB样本总数就是这个决策树的错误率变化情况，将随机森林中所有树的错误率变化情况汇总平均就得到平均下降准确率。另外一种是基于分裂时的GINI下降量方法，随机森林在生长决策树是按照GINI不纯度的下降情况在进行节点分裂的，把森林中的所有选择某变量为分裂变量的节点汇总就得到了GINI下降量。

**1.4 非平衡数据分类的随机森林方法**

随机森林算法默认每个类的权重为1，也就是假定所有类的错分代价是等同的。在scikit-learn中，随机森林算法提供class\_weight参数，其值可以为list或dict值，手动指明不同类别的权重。若该参数为"balanced"，那么随机森林算法使用y值自动调整权重，各类权重与输入数据中的类别频率成反比。

计算公式为：  。

"balanced\_subsample"与"balanced"模式类似，计算时使用有放回式的取样中的样本数，而不是总样本数。因此，我们可以通过这种方法解决非平衡数据分类问题。

**2  数据预处理与数据分析**

**2.1 数据集**

本文使用的贷款违约数据集来自Kaggle数据科学竞赛平台2011年的一个竞赛项目“Give Me Some Cerdit”，该数据集共包含25000个样本，其中150000个样本作为训练集，100000个样本用作测试集。

该训练集共有150000条借款人的历史数据，其中违约样本10026条，占样本总量的6.684%，贷款违约率为6.684%，未违约样本139974条，占样本总量的93.316%。可以看出该数据集是一个典型的高度不平衡的数据。数据集中包括了借款人的年龄、收入、家庭等情况和贷款情况，共11个变量，其中SeriousDlqin2yrs是label标签，另外10个变量为预测特征。下表列出了变量名称及数据类型情况：

表1 数据集变量情况

|  |  |  |
| --- | --- | --- |
| 变量名 | 变量描述 | 类型 |
| SeriousDlqin2yrs | 是否违约 | Y/N |
| RevolvingUtilizationOfUnsecuredLines | 信用卡和个人信用贷款的总额（不含房贷、类似车贷的分期付款等）除以授信额度之和 | 百分比 |
| age | 借款人年龄 | 整数 |
| NumberOfTime30-59DaysPastDueNotWorse | 过去两年中借款人逾期30-59天的次数 | 整数 |
| DebtRatio | 每月债务偿还数、赡养费、生活成本等除以每月总收入 | 百分比 |
| MonthlyIncome | 月收入 | 实数 |
| NumberOfOpenCreditLinesAndLoans | 开放贷款（Open loans，分期付款如车贷和抵押）和信用贷款（Lines of credit,如信用卡）的数量 | 整数 |
| NumberOfTimes90DaysLate | 过去两年中借款人逾期90天及以上的次数 | 整数 |
| NumberRealEstateLoansOrLines | 抵押和房地产贷款含房屋抵押式信用贷款的次数 | 整数 |
| NumberOfTime60-89DaysPastDueNotWorse | 过去两年中借款人逾期60-89天的次数 | 整数 |
| NumberOfDependents | 不包括本人在内家庭中需要抚养的人（配偶及子女等等）数 | 整数 |

**2.2 数据分析**

本文实验所使用的实验环境为Anaconda3+Python3.6。首先对数据进行初步分析，本实验主要分析违约率在每个自变量上的分布，及生成如表2所示的频率分布表（小数均为四舍五入）。

表2 变量age的频率分布表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| age | 人数 | 占比 | 违约人数 | 该区间违约占比 |
| 低于25 | 3028 | 2.02% | 338 | 11.16% |
| 26-35 | 18458 | 12.3% | 2053 | 11.12% |
| 36-45 | 29819 | 19.9% | 2628 | 8.8% |
| 46-55 | 36690 | 24.5% | 2786 | 7.6% |
| 56-65 | 33406 | 22.3% | 1531 | 4.6% |
| 高于65 | 28599 | 19.1% | 690 | 2.4% |

由表2可以看出来小于25岁的人群和26-35岁的人群，违约率都超过10%。随着年龄增加，违约率下降。

表3 变量NumberRealEstateLoansOrLines的频率分布表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NumberRealEstateLoansOrLines | 人数 | 占比 | 违约人数 | 该区间违约占比 |
| 低于5 | 149207 | 99.47% | 9884 | 6.6% |
| 6-10 | 699 | 0.47% | 121 | 17.3% |
| 11-15 | 70 | 0.05% | 16 | 22.8% |
| 16-20 | 14 | 0.009% | 3 | 21.4% |
| 高于20 | 10 | 0.007% | 2 | 20% |

由表2可以得知99.47%借款人的不动产和抵押贷款数目小于5笔，但是5笔以上的借贷人群违约率明显增加，其中超过10笔的借款人违约率均在20%以上。

表4 变量NumberOfTime30-59DaysPastDueNotWorse的频率分布表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NumberOfTime30-59Days  PastDueNotWorse | 人数 | 占比 | 违约人数 | 该区间违约占比 |
| 0 | 126018 | 84% | 5041 | 4% |
| 1 | 16032 | 10.7% | 2409 | 15% |
| 2 | 4598 | 3.1% | 1219 | 26.5% |
| 3 | 1754 | 1.2% | 618 | 35.2% |
| 4 | 747 | 0.5% | 318 | 42.6% |
| 5 | 342 | 0.23% | 154 | 45% |
| 6 | 140 | 0.09% | 74 | 52.9% |
| 7及以上 | 104 | 0.07% | 50 | 48.07% |

由表4可以得知没有发生过30-59天逾期的借款人违约率仅4%左右，但是随着逾期次数的增加，违约率明显升高。另外两个变量，借款人发生60-89天逾期次数和借款人发生90及以上的逾期次数频率分布表也呈与表4同样的趋势。因此可以得到结论，借款人发生逾期次数越多，违约率越高。

本实验所使用数据集有10个变量，我们对每个变量都统计分析并得到如上所示的频率分布表，除了变量NumberOfOpenCreditLinesAndLoans（开放贷款和信用贷款的数量）与违约率无明显相关性，其他变量均与借款人最终是否违约相关。

**2.3 数据预处理**

首先对数据进行初步探索，可以发现MonthlyIncome和NumberOfDependents这两个变量中存在缺失值，个数分别为29731和3924。

异常值情况：age变量中的最小值为0，属于异常值。

在NumberOfTime30-59DaysPastDueNotWorse、NumberOfTime30-59DaysPastDueNotWorse、NumberOfTimes90DaysLate三个逾期天数变量中，都存在少数96，98的值，可能是异常值也可能是某行为代码。

数据预处理：在使用Python中的pandas库读取数据时，将函数pd.read\_csv()中的na\_values参数设置为我们自己定义的list，将age变量中的0和三个逾期变量中的96，98处理为NaN 值，然后使用sklearn.preprocessing.Imputer库将数据集中所有的NaN替换为对应列的平均值

**3  建模及实验结果**

**3.1 随机森林模型**

本实验使用Python中的sklearn.ensemble.RandomForestClassifier建立随机森林模型。

部分参数设置为：

n\_estimators：决策树的个数设置为100，

oob\_score：是否使用袋外数据，设置为True,

min\_samples\_split：根据属性划分节点时，每个划分最少的样本数,设置为2，

min\_samples\_leaf：叶子结点最少的样本数，设置为50，

n\_jobs：并行数，设置为-1电脑CPU有多少core，就启动多少个job

class\_weight：设置为'balanced\_subsample'，使用y值自动调整权重，各类权重与输入数据中的类别频率成反比，

bootstrap：是否使用bootstrap样本抽样，设置为True。

**3.2 模型评估**

本实验使用的模型评估指标为AUC(Area under the ROC curve)值。AUC被定义为ROC(Receiver Operating Characteristic)曲线下的面积，显然这个面积的数值不会大于1。ROC曲线的横轴为False Positive Rate（假阳率，FPR），纵轴为True Positive Rate（真阳率，TPR），又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

我们将随机森林模型与逻辑回归分类模型和决策树分类模型进行了比较，结果如下表。

表5 随机森林和其他算法的比较

|  |  |
| --- | --- |
| 算法 | AUC值 |
| 随机森林 | 0.86 |
| 决策树 | 0.80 |
| 逻辑回归 | 0.80 |

从表中可以看出，随机森林算法的AUC值高于决策树和逻辑回归算法，因此随机森林的算法预测性能也优于其他两个算法。

**3.3 特征重要性度量**

本实验使用sklearn.ensemble.RandomForestClassifier的feature\_importances\_方法，得到每个特征的重要性程度如下表所示。

表6变量重要性

|  |  |
| --- | --- |
| 变量 | feature\_inportance |
| RevolvingUtilizationOfUnsecuredLines | 0.3411 |
| NumberOfTime30-59DaysPastDueNotWorse | 0.1694 |
| NumberOfTimes90DaysLate | 0.1594 |
| NumberOfTime60-89DaysPastDueNotWorse | 0.0727 |
| age | 0.0677 |
| DebtRatio | 0.0625 |
| MonthlyIncome | 0.0488 |
| NumberOfOpenCreditLinesAndLoans | 0.0442 |
| NumberRealEstateLoansOrLines | 0.0223 |
| NumberOfDependents | 0.0117 |

由上表可以看出，借贷人的贷款总额占授信总额的比率、过去两年逾期30-59天的次数和过去两年逾期超过90天的次数这三个特征的重要性程度在前三，对最终是否违约影响较大，因此在处理贷款申请时，可以特别关注借贷人的这些特征。

**4  小结**

本文主要研究了金融领域中常见的贷款违约问题，并使用了非平衡数据分类的随机森林方法建立了预测贷款违约模型，随机森林的基本思想是在构造单个树的过程中，随机选取一些变量或特征参与树节点划分，重复多次并保证建立的这些树之间的独立性，针对非平衡的数据，通过参数调整使得随机森林方法能够根据y值自动调整权重，从而有效解决非平衡数据的分类问题。

通过实验表明，随机森林算法比决策树和逻辑回归模型的分类性能更好，对金融领域的贷款违约预测问题有重要的参考意义。另外，通过对各特征的重要性进行度量，在本实验中可以得到借贷人的年龄、负债率和不动产和抵押贷款数目这三个特征对最终是否违约影响较大，该度量特征重要性的方法也对其他数据挖掘中的特征选择问题有较重要的参考意义。

References:

[1] 信用风险模型比较分析[J]. 梁世栋,郭仌,李勇,方兆本.  中国管理科学. 2002(01)

[2]商业银行贷款违约模型研究综述[J]. 刘铁牛.  湖南商学院学报. 2009(03)

[3] 基于KMV模型的上市公司信用风险预测[J]. 夏红芳,马俊海.  预测. 2008(06)

[4] 基于并行随机森林的在线贷款逾期预测研究[D]. 巫晓杰.华南理工大学 2016

[5]基于非平衡数据的集成学习分类及其应用[D]. 周宾宾.华南理工大学 2014

[6] 基于随机森林的不平衡数据分类方法研究[D]. 肖坚.哈尔滨工业大学 2013

[7]基于集成学习的非平衡数据集分类问题的研究[D]. 郭文利.西安电子科技大学 2013

[8] An experimental comparison of ensemble ofclassifiers for bankruptcy prediction and credit scoring[J] .LorisNanni,Alessandra Lumini.  Expert SystemsWith Applications . 2008 (2)