基于网格搜索和随机森林的汽车信贷违约预测研究

陈 滢

(上海立达学院 科研与发展规划处, 上海 201608)

摘要:基于某金融机构的汽车信贷违约数据构建随机森林风险预测模型,用主成分分析法对数据进行降维,利用上采样的方法解决样本不平衡的问题,同时通过综合五折交叉验证法和网格搜索对随机森林模型调参。此外,还与其他机器学习算法的预测结果进行比较。研究表明,相对于其他两种预测模型,随机森林的性能都是最优的,性能较佳。同时,采用随机森林计算特征重要性时发现,个人抵押资产的价值对汽车信贷违约有显著的影响。

关键词:信用指标体系;随机森林;上采样;网格搜索

中图分类号:F832.479 **文献标志码**:A **文章编号**:1671-1807(2023)09-0116-06

近年来,随着消费信贷在汽车金融市场的广泛应用,汽车信贷业务得以迅速发展。汽车信用贷款作为消费金融的一个组成部分,虽然其市场渗透率较低,但是在内外部环境的推动下,未来仍有较大的发展空间。据华夏时报估计,到 2025 年,国内的汽车金融市场规模将达 2.32 万亿元。但是,由于汽车信贷消费方式还处于初步发展阶段,个人信用体系不健全,风险评估机制不完善,而且多方面的压力也造成了个人不能按期归还借款,一旦发生此类问题,金融机构将无法追踪借款人去向等相关信息。此外,由于借款主体的多样化,这也会导致汽车信用贷款违约现象层出不穷,在消费金融市场的推动下,这种信用违约风险更是造成了多方面的损失。对金融机构来说,复杂多变的信用贷款违约风险在一定程度上给它们带来了经济危机,由此阻碍了其信贷业务的发展,无法在激烈的市场竞争中获取利润。对国家来说,金融机构可以传输国家经济政策,是调节经济生产和消费的有效工具,如果金融机构不能发挥其应有的作用,国家经济的发展将受到一定的阻力。因此,对汽车信用贷款违约风险进行有效的研究,不仅能够进一步剖析个人信用体系的构建依据,还可以帮助金融机构减少损失,规避类似的风险。

1 文献综述

1.1 关于汽车信贷违约原因的研究

蔡姗姗[1]利用 SWOT 方法从贷前、贷中、贷后 3 个角度分析汽车信贷风险中的问题,认为汽车金融市场融资渠道单一、行业缺乏专业人才、经济政策的调整和恶性的市场竞争是导致信贷违约风险的重要原因,同时指出风险管理控制问题是汽车金融的核心。周博和黄奕涵[2]从内部人力资源配置的角度指出经销商经理的年龄越大,汽车消费信贷业务量将增加,反之将减少。李雯晶等[3]认为社会传染会影响用户消费信贷行为,并且在社会传染的中介效应下,其中年龄、用户金融知识和教育水平都对用户信贷行为产生了一定的影响。

1.2 对汽车信贷违约预测模型的研究

王方方[4]以某汽车金融公司的信贷数据为例,结合外部各种影响因素,构建了逻辑回归预测模型。周宏杰[5]基于汽车消费信贷金融机构的角度,从用户贷款信息、历史信息和个人信息 3 个方面构建了 stacking 预测模型,并结合改进版的特征选择方法和 SMOTE-Tomek 采样方法,发现该模型具有良好的分类能力和稳健性。陈倩等[6]通过分析个人信贷违约的因素,清洗了信贷数据后,构建了一个融合随机森林算法和逻辑回归算法的模型,发现综合使用机器学习算法的模型预测效果更好。

综上所述,学者们针对汽车信贷违约预测模型的研究相对较少,并且预测模型呈现出单一化的特点,多是采用 Logistic 算法构建信贷违约风险评估模型。而逻辑回归模型在遇到样本非常不平衡的时候性能不是很好,且对于高维度数据无法进行有效的处理。因此,本文从贷款者特征和贷款信息两

收稿日期:2022-12-09

基金项目:2022 年上海立达学院校级项目(AKY-2022-01-91)。

作者简介:陈滢(1994—),女,江苏镇江人,上海立达学院科研与发展规划处,助教,企业管理硕士,研究方向为商务智能。

个角度构建信用指标评估体系,先对数据进行初步的处理和清洗,然后采用主成分分析法对数据进行降维处理以此提高模型运行效率,接着利用上采样的方法解决信贷数据不平衡的问题,再构建随机森林汽车信贷违约预测模型处理高维度数据和计算各个特征重要性,并且用网格搜索法对模型参数进行寻优,使得随机森林模型性能最优,最后将该模型结果与 Logistic 模型、k 最近邻模型进行对比。

2　汽车信用指标体系的构建

实验的资料来源于 kaggle 机器学习数据库关于汽车信贷违约的数据。样本总数为 233 154 个,其中正例(违约)有 50 611 个,反例(没有违约)有 182 543 个。

在所有原始数据中,每个样本包括标签变量(loan_default),一共有 41 个特征,由于 uniqueid、branch_id、supplier_id、manufacturer_id、current_pincode_id、state_id、employee_code_id 这 7 个特征明显与汽车信贷违约结果无关,因此不作为输入变量。特征 date.of.birth 是顾客的出生日期,由于不清楚顾客的出生年份,因此在记录时选择了顾客的出生月份,二分类特征 employment.type 存在 7 661 个空白值,在对结果影响不大的情况下作删除处理,其属性'salaried'被编码为 0,属性'self employed'被编码为 1,特征 disbursaldate 也是日期形式,为了方便特征处理,同时在对结果影响不大的情况下只记录其年份。而多分类特征 perform_cns.score.description 有 20 个类别值,先将 a-very low risk、b-very low risk、c-very low risk、d-very low risk、e-low risk、f-low risk、g-low risk 记录为 low risk,h-medium risk、i-medium risk 记录为 medium risk,将 j-high risk、k-high risk、l-very high risk、m-very high risk 记录为 high risk,将 no bureau history a-vailable、not scored: no activity seen on the customer (inactive)、not scored: not enough info a-vailable on the customer、not scored: only a guarantor、not scored: sufficient history not available、not scored: more than 50 active accounts found、not scored: no updates available in last 36 months 记录为 not scored,然后对这 4 个类别(low risk、medium risk、high risk、not scored)进行编码,并且也利用 get_dummies()函数对四分类特征 perform_cns.score.description 进行编码。此外,特征 average.acct.age 和特征 credit.history.length 的属性值是日期的字符串形式,在记录时将年份数提取出来并转化成月份数,1 年计算成 24 个月,比如'4yrs 8 mon'就记录为 54,特征 loan_default 是违约结果,其值 0 代表没有违约,1 代表违约,其余特征的属性值按资料给定的方式记录。实验中原数据样本一共有 7 661 个缺省值,经综合考虑均作删除处理。此外,经过数据的初步整理之后,构建的指标体系包含 3 个方面内容:贷款者信息、贷款信息和信用局数据及历史记录,一共有 33 个信用评估指标。

3　实证分析

3.1　实验环境

实验的电脑处理器为 Intel(R) Core(TM) i5_3317U CPU@1.70 GHz。使用软件 Python3.6 来编写相关代码,进行数据分析与建模。

3.2　数据降维

主成分分析法是学者较为青睐的数据降维处理方法,它不仅可以消除变量间的相关性,还可以将多特征数据降低维度,提高模型运行效率和性能。在本次实验中,一共有 33 个输入变量,经过 get_dummies()编码和初步处理后,变成 36 个输入特征,维数较多。因此,实验使用主成分分析法来对数据进行降维,将累计方差贡献率设置为 0.95,经过 Python 软件运行后得到了 19 个因子,减少了 17 个输入变量,数量减少了将近原先输入的一半,在没有降低模型性能的情况下,主成分分析极大地降低了数据集维度。

3.3　平衡数据类别

实验中,数据样本总数(经过数据预处理后)有 225 493 个,其中正例(违约)有 48 967 个,反例(没有违约)有 176 526 个,类别非常不平衡。解决数据不平衡问题的最常用方法主要是上采样和下采样。其中上采样就是对少数类进行处理,最后使得正反类别样本比例为 1:1。其中上采样还包括 smote 和 adasyn 算法,smote 算法就是随机选取一个少数类样本,然后随机找一与其相近的样本,在这两个样本间再选取一个点作为新样本,而 adasyn 算法是在被错误分类的样本附近新生成一个少数类样本。下采样就是对多数类样本进行处理,由于在处理类别不平衡的众多方法中,如果正反例数据类别非常不平衡进行下采样分析将会丢失很多重要信息。因此,经综合考虑,研究使用上采样的方法来解决数据不平衡,并采用二分类模型评估性能指标来具体说明使用上采样前后对模型预测结果的影响。上采样前后的预测模型 ROC 曲线(接受者操作特性曲线)对比结果如图 1 所示。

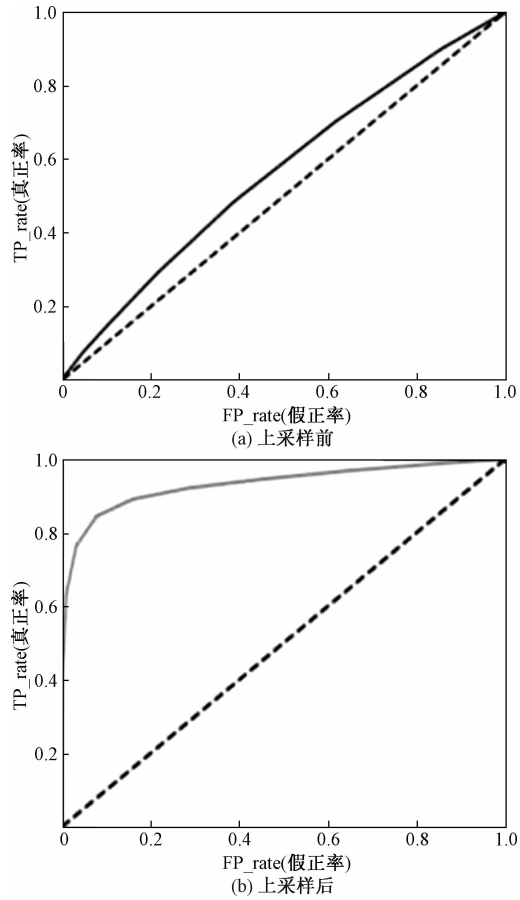

图 1　上采样前后的 ROC 曲线对比

从图 1 的对比结果中可以看出，图 1(a)没有经过上采样的随机森林风险预测模型 AUC(ROC 曲线下的面积)为 0.56，ROC 曲线接近中间对角线，而图 1(b)经过上采样的模型 AUC 为 0.93，且 ROC 曲线比较接近上方横线，这些结果说明随机森林模型经过上采样的处理之后，其性能得到了极大的改善。

3.4　分类性能评估指标

为验证建立的随机森林模型是否能够成功预测汽车信贷违约情况，需要利用一些指标来对模型预测结果进行评估，而不同的预测问题采用的性能评估指标是不一样的，本次实验解决的是二分类预测问题，相应的性能评价指标有 precision(准确率)、recall(召回率)、f1-score、ROC 曲线图和 AUC，其中准确率、召回率的计算公式为

$$\text{precision} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

式中：TP 为实际上是正例，预测是正例的样本个数；TN 为实际上是负例，预测是负例的样本个数；FP 为实际上是正例，预测是负例的样本个数；FN 为实际上是负例，预测是正例的样本个数；precision(准确率)表示在所有样本中被正确分类的比例，该比例越高，说明分类器越好；recall(召回率)表示在所有正样本中被正确分类为正样本的比例。f1-score(精准率)综合了 precision 和 recall 这两个指标，当训练样本极不平衡时，结果越好，说明分类器效果越好。而 ROC 曲线是以假正率(FP_rate)为横轴，真正率(TP_rate)为纵轴的曲线，ROC 曲线下面的面积 AUC 越大，分类器的性能越好。

3.5　构建随机森林模型

分析了数据特点后根据汽车信用评价指标构建随机森林模型，先在原始样本中采用 bootstrap 方法进行有放回随机抽取样本，构建 225 493 个分类决策树。然后根据主成分分析降维后得到的 19 个成分用作每个决策树的分裂节点处的分裂特征，并利用 Gini 计算节点不纯性来衡量特征重要性，且计算每个特征所包含的信息，然后在 19 个特征中选择最优的特征进行分裂。最后，将所有决策树组合在一起形成随机森林，并对结果进行投票，预测结果为大多数决策树输出的类别。其建模过程如图 2 所示。

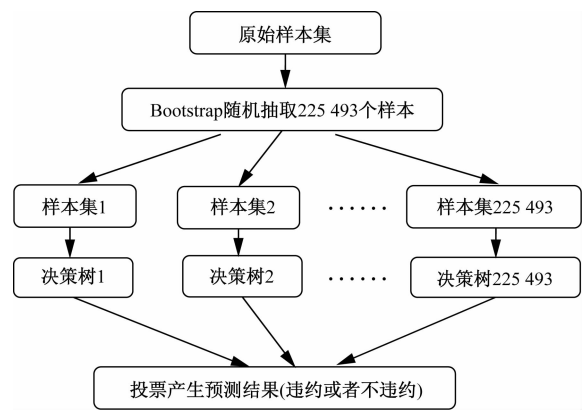

图 2　随机森林建模过程

3.6　调参和结果分析

网格搜索(Grid Search)是一种常见的模型参数寻优方法[7]。其基本原理是通过穷举法将所给定范围内的参数代入模型中，并用交叉验证法找出使得模型性能最好的参数。本次实验中，随机森林(RandomForest)模型有两部分内容，一个是框架参数，其中包括 n_estimators、oob_score 和 criterion，另一个是决策树参数，其中包括 max_features、max_depth、min_samples_split、min_samples_leaf、min_weight_fraction_leaf、max_leaf_nodes、min_impuri-

ty_split。如果用网格搜索法对这所有的参数进行寻优,将会耗费大量的时间,本次实验根据数据的特点,综合考虑对 n_estimators 和 max_depth 进行寻优,先用五折交叉验证法计算出 n_estimators 五次袋外误差率,结果如图 3 所示。

从图 3 中可以看出袋外误差率在 100 附近是趋于稳定且是最低的,于是继续细分调参范围,找到最佳参数组合提升模型性能,发现 n_estimators 为 120,max_depth 为 40,说明此时的模型效果最好。将最优参数值代入随机森林模型中,分析其预测性能,具体情况见表 1。

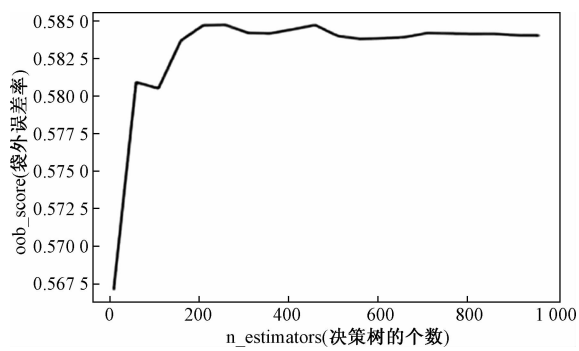

图 3　袋外误差率曲线

表 1　随机森林性能指标结果

类别	precision(准确率)	recall(召回率)	f1-score(精准率)
反例(class 0)	0.90	0.85	0.88
正例(class 1)	0.86	0.91	0.88

表 1 列出了分别正例(class 1)和反例(class 0)的性能评估指标结果,然后平均正例和反例的数据,经计算可知,随机森林预测模型的准确率为 0.88,召回率为 0.88,精准率为 0.88。这 3 个性能指标结果均已超过 0.8,说明模型分类器效果较好。此外,本次实验还制作了 ROC 曲线图来进一步了解随机森林分类器的性能,具体情况如图 4 所示。

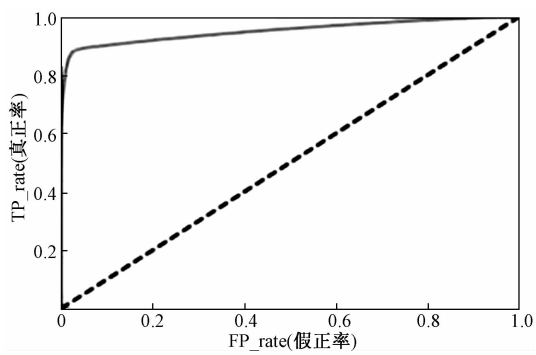

图 4　随机森林的 ROC 曲线

从图 4 中可以看出,ROC 曲线逼近坐标轴上方横线,AUC 为 0.95,这说明经过上采样、主成分分析和网格搜索调参法使得随机森林的分类性能在一定程度上得到了改善。为了进一步分析商业银行个人信用贷款违约原因,本文运用随机森林算法计算特征重要性,由于个人信用评估指标较多,此处仅列举出重要性大于 0.05 的变量,其具体情况见表 2。

从表 2 中可以看出,在个人信用所有的评估指标中,资产成本(asset_cost)对汽车信贷违约结果有最重要的影响,其次是支付金额(disbursed_amount),列举出的 7 个指标中,发现主要是贷款者特征和贷款信息影响信用贷款的结果,说明金融机构在对个人进行信用评估时,不仅需要仔细核实贷款相关的信息活动,还需要重视贷款者自身的特质,如果贷款金额明显超出了个人支付能力,可不予以借款。

表 2　变量重要性

序号	变量	重要性
1	asset_cost(资产成本)	0.162 250
2	disbursed_amount(支付金额)	0.156 601
3	disbursal date(支付日期)	0.104 479
4	ltv(资产贷款价值比)	0.090 101
5	date. of. birth(出生日期)	0.063 264
6	credit. history. length(自首次贷款以来的时间)	0.061 903
7	average. acct. age(平均贷款期限)	0.057 801

4　其他模型预测结果对比

4.1　逻辑回归模型(Logistic)

Logistic 回归是比较经典的二分类预测模型,因此,实验通过 Python3.6 软件中 LogisticRegression()包进行建模,并用准确率、召回率和 f1-score 这 3 个性能指标对其进行评估,具体结果见表 3。

从表 3 中可以看出,无论正例(class 1)还是反例(class 0),最后的结果都不是特别理想,尽管进行了上采样处理,使得正例和反例预测结果相差不大,但是 f1_score 都未超过 0.7,并且 Logistic 分类器的 AUC 为 0.62,这都说明该模型分类性能不佳。

表 3　Logistic 性能评估指标结果

类别	precision(准确率)	recall(召回率)	f1-score(精准率)
反例(class 0)	0.60	0.55	0.57
正例(class 1)	0.58	0.63	0.61

4.2 k 最近邻模型(KNN)

k 最近邻分类算法是数据挖掘中比较简单的机器学习方法。在本次实验中,利用 Python3.6 软件中 KNeighborsClassifier()包进行建模,并用准确率、召回率和精准率这 3 个性能指标对其进行评估,具体结果见表 4。

表 4 KNN 性能指标评估结果

类别	precision (准确率)	recall (召回率)	f1-score (精准率)
反例(class 0)	0.69	0.57	0.63
正例(class 1)	0.64	0.74	0.69

从表 4 中可以发现,正例和反例的 precision 和 f1-score 都达到了 0.6 以上,其 AUC 为 0.72,与逻辑回归模型预测结果相比,其预测性能要好一点,但是与随机森林预测结果的差距还是较大的。

综上所述,由各个性能指标值可以明显看出,随机森林预测模型的准确率、召回率、f1_score 和 AUC 都比逻辑回归模型和最近邻分类器的效果好。这说明,相比较于大多数学者运用逻辑回归模型预测汽车信贷违约的状况,随机森林的优势更加明显,并且可以被应用于汽车信贷违约预测的领域。

5 结论、建议与展望

5.1 研究结论

从贷款者和信用评估机构的角度出发,确定了 33 个信用指标,采用主成分分析法进行降维,选取了 19 个因子,构建汽车信用评估体系。然后利用上采样的方法解决数据不平衡的问题,并构建了随机森林违约预测模型,同时用网格搜索法对模型参数进行寻优。最后将预测结果与经典的 Logistic 模型、KNN 模型进行对比,发现随机森林分类器的准确率达到了 0.88,AUC 为 0.95,说明该模型预测模型性能更佳,具备更好的拟合效果。此外,通过随机森林对各个特征进行重要性分析,发现贷款支付的金额、支付的日期、资产的成本和贷款与资产的价值比对汽车信贷违约有着较大的影响,并且抵押资产价值的影响最大。这说明金融机构在构建信用评估体系时,如果遇到抵押贷款的情况,应该着重考虑抵押资产的价值。

5.2 对策建议

随着汽车金融市场的渗透率逐渐增加,汽车信贷业务也在快速发展,对其违约风险的研究有以下建议:

1)进一步完善信用指标体系。良好的信用指标体系有利于更好地评估个人信用,还可以建立分类性能更好的风险预测模型,具体可以采用德尔斐专家法、层次分析法和回归分析等方法,找到最具有代表性的个人信用指标,然后确定每个指标的权重,最后一定要动态管理评估体系,使得建立的信用指标评级系统更具风险抵抗性。

2)加强风险管控。由于汽车信贷违约涉及到个人道德方面的问题,极具主观性,不可控力较大,虽然各大金融机构致力于开发信贷风险规避的最佳方法,但是一直以来都无法彻底解决信贷违约的现象。因此,金融机构应该以管控和规避风险为主,尽力降低风险损失,基于机器学习集成方法的思想,综合利用各具优势的分类器,开发通用性较强的风险管控方法。

5.3 研究展望

本次研究可为金融公司在汽车信用评估指标体系的构建提供参考,进一步完善汽车信用指标构建机制,同时为其他行业遇到信贷违约问题时提供借鉴,增强自身规避风险的能力,减少经济损失。研究还存在一些不足之处。一方面,本次研究选取的是某金融机构提供的信贷违约数据,典型性和代表性有待提升;另一方面,研究只构建了随机森林模型、经典的逻辑回归模型和 k 最近邻分类器模型,没有将更多的机器学习模型预测结果进行对比,当遇到性能更好的分类模型时,随机森林可能不具有说服力。未来可针对这些不足展开更多、更丰富的研究。

参考文献

[1] 蔡姗姗.B汽车金融公司信贷风险管理优化研究[D].西安:西北大学,2019.

[2] 周博,黄奕涵.汽车企业人力资源配置对于汽车消费信贷业务影响研究[J].市场周刊,2021,34(3):172-176.

[3] 李雯晶,蒋青云,刘婷.社会传染对消费信贷行为的影响:来自汽车消费信贷网络数据的证据[J].管理现代化,2021,41(4):10-16.

[4] 王方方.C汽车金融公司信贷审批决策模型研究[D].天津:天津大学,2020.

[5] 周宏杰.汽车消费信贷违约风险评估研究[D].武汉:中南财经政法大学,2021.

[6] 陈倩,贺兴时,杨新社.基于 RF 的 Elastic Net-Logistic 个人信用违约风险评估[J].西安工程大学学报,2021,35(3):116-122.

[7] 刘佳星.基于网格搜索超参数优化的支持向量回归[J].科学技术创新,2022(13):71-74.

Research on Automobile Credit Default Prediction Based on Grid Search and Random Forest

CHEN Ying

(Department of Scientific Research and Development Planning, Shanghai Lida University, Shanghai 201608, China)

Abstract: Based on the automobile credit default data of a financial institution, a random forest risk prediction model is constructed. The principal component analysis method is used to reduce the dimensions of the data, and the method of up sampling was used to solve the problem of sample imbalance. The random forest model parameters were adjusted by integrating the 50 fold cross validation method and grid search. In addition, the prediction results are compared with those of other machine learning algorithms. The research shows that, compared with the other two prediction models, the performance of random forest is optimal and better. At the same time, when using stochastic forests to calculate the importance of characteristics, the value of personal mortgage assets has a significant impact on automobile credit default.

Keywords: credit indicator system; random forest; upsampling; grid search