Coursework

2024/2025 Semester 1

Introduction to Data Science:

Report for Credit Risk Management Programme

Lanshun Yuan

Word Count:1492

Introduction

Credit cards have become essential for consumers, driving a global surge in usage and intense competition among issuers. To capture market share, companies often issue cards aggressively, sometimes with minimal risk assessment, increasing the likelihood of defaults from customers unable to repay on time. This project aims to address this issue by identifying high-risk customers prone to default. The objective is to reduce default rates through targeted risk management, protecting the financial health of both issuers and customers. The key focus will be predicting whether a customer defaults on their credit card payments.

Key predictive variables will include given credit, payment delay, bill statement and other financial indicators. These variables provide critical insights into a borrower's financial stability, spending habits, and credit risk profile. Additional variables such as gender, education, marriage etc will also be incorporated to enrich the analysis. Understanding how these factors interplay and contribute to default behavior is crucial for constructing a robust predictive model. This project aims to balance growth and risk using data-driven insights, helping credit card companies achieve sustainable profitability while promoting responsible credit use.

Data Exploration

This dataset profiles credit card users and their financial behavioral characteristics. According to the fig1, the users are primarily young (average age 35), with a balanced gender distribution and medium education levels. Financially, the average credit limit is 167,130.73, but most users have lower limits, as indicated by a right-skewed distribution. Users generally show good repayment behavior (PAY_1 to PAY_6), with negative mean values and mostly 0s, suggesting timely payments. Bill amounts (BILL_AMT1 to BILL_AMT6) increase monthly, while payment amounts (PAY_AMT1 to PAY_AMT6) remain stable, averaging 4,700-5,800. Notably, all financial variables are right-skewed, and the data quality is excellent with no missing values, providing a reliable foundation for subsequent in-depth analysis and modeling.

Name	Mean	Median	Min.	Max.	Missing
LIMIT_BAL	167130.7	140000	10000	1000000	0 (0 %)
SEX	1.60319	2	1	2	0 (0 %)
EDUCATION	1.85	2	0	6	0 (0 %)
MARRIAGE	1.55	2	0	3	0 (0 %)
AGE	35.47	34	21	75	0 (0 %)
PAY_1	-0.02	0	-2	8	0 (0 %)
PAY_2	-0.13	0	-2	7	0 (0 %)
PAY_3	-0.16	0	-2	8	0 (0 %)
PAY_4	-0.22	0	-2	8	0 (0 %)
PAY_5	-0.26	0	-2	8	0 (0 %)
PAY_6	-0.29	0	-2	8	0 (0 %)
BILL_AMT1	51005.64	22352.5	-165580	964511	0 (0 %)
BILL_AMT2	48962.83	21120.5	-67526	983931	0 (0 %)
BILL_AMT3	46803.85	20081	-157264	1664089	0 (0 %)
BILL_AMT4	43083.84	19052	-170000	891586	0 (0 %)
BILL_AMT5	40133.65	18077.5	-81334	987171	0 (0 %)
BILL_AMT6	38692.57	17072.5	-339603	961664	0 (0 %)
PAY_AMT1	5626.81	2100	0	873552	0 (0 %)
PAY_AMT2	5828.09	2007	0	1684259	0 (0 %)
PAY_AMT3	5226.95	1800	0	896040	0 (0 %)
PAY_AMT4	4761.34	1500	0	621000	0 (0 %)
PAY_AMT5	4752.97	1500	0	388071	0 (0 %)
PAY_AMT6	5224.23	1500	0	528666	0 (0 %)

Fig.1

Model Building

Data Preprocessing:

Based on our descriptive analysis, we firstly screen for obvious outliers such as values in education not between 1 and 4 or marriage not between 1 and 3. Secondly, isolation forest method is suitable for high-dimensional sparse and complex data, so we use it for advanced outlier filtering.

Logistic Regression:

Logistic regression is a statistical model that relies on the independence of predictors to estimate their effects on the outcome. Therefore, we followed a structured process using stepwise regression to avoid multicollinearity between variables (Senaviratna & Cooray, 2019). First, we converted all independent variables to numerical types to ensure compatibility with the model and included them in the initial analysis. Next, we applied backward elimination, progressively removing variables that were not statistically significant. We assessed each variable's contribution using a chi-square test and removed those with p-values exceeding the significance threshold ($\alpha = 0.05$) if their removal did not significantly affect the model's AUC or CA values. Through this repeated process, we narrowed down the dataset to 14 key variables. To improve the model's stability and prevent overfitting, we then applied LASSO (L1 regularization) with a regularization strength of $C=1$. To ensure fairness in variable comparison, we standardized all independent variables, transforming them to have a mean of 0 and a standard deviation of 1. Finally, this process resulted in a model that retained only significant variables, effectively reducing multicollinearity and improving overall performance.

Random Forest:

Random forest model is suitable for dealing with nonlinear relationships and binary variables, it also prefer complex variables (Archer and Kimes, 2008). We create a new variable Default to represent whether he has defaulted (Defaulted=1), and use the ratio of Bill statement to Credit limit to generate Credit Utilisation. We also created payout ratio, based on the ratio of Payment to bill statement, and set the value of the number greater than 30% to 1, and then created a new variable, payout ratio type. Finally, we generate Bill Volatility and Payment Volatility to observe internal change. Besides, we keep the original variables as random forests perform better when the number of variables is large. Additionally, we also tuned the hyperparameters to optimize the random forest model: the number of trees was set to 100 to reduce error and improve performance, the number of features per split was set to the square root of the number of variables to reduce the risk of overfitting, the tree depth was set to 10 to balance model complexity and overfitting, and the minimum number of samples per leaf node was set to 10 to enhance the model's prediction capability.

Naive Bayes:

The Naïve Bayes model assumes strong independence between input data (Yang, 2018). We used the Chi-square test to filter correlated variables and created new features to optimize model performance. Based on Random Forest variables, we developed several features: Default Count by converting PAY1-6 values to binary indicators (1 for default, 0 for no default) and summing them, reflecting credit risk (higher counts indicate lower credit); Activity Frequency by converting non-zero values in BILL_AMT1-6 and PAY_AMT1-6 to binary indicators (1 for activity, 0 for no activity) and summing them, reflecting the frequency of credit card usage and repayment; and Credit Balance Type by categorizing the difference between credit limit and bill amounts (0 for reaching limit, 1 for not reaching) to represent the user’s credit usage. We also retained original variables as Rish (2001) suggests Naïve Bayes performs well with functional features. To further optimize the model, we applied Principal Component Analysis (PCA) for dimensionality reduction, aiming to better adapt its assumption. Although PCA increased the model’s accuracy (from 75.4% to 78.0%), it also resulted in a higher rate of false positives in the confusion matrix. Given the nature of credit card risk prediction, we concluded that applying PCA was not preferable for this task.

Model Evaluation

In the model evaluation section, we selected AUC, F1, and false positive rate (FPR) as performance evaluation metrics. AUC measures the model's ability to distinguish between positive and negative samples, while F1 combines precision and recall to directly assess model performance. Both metrics align

with the goals of credit risk prediction. Moreover, FPR is another important metric, as in real-world credit card risk prediction, banks are particularly concerned about false positives, which represent cases where defaults are misclassified as non-defaults.

When comparing AUC and F1 across four models (including the baseline), we found that Random Forest performed the best, which is 78.5% and 80% respectively. However, when examining the confusion matrix to evaluate FPR (fig.2), the results differed significantly. Random Forest and Logistic Regression both showed high sensitivity (95% and above) but had FPRs below 50%, indicating poor performance in minimizing false positives. This suggests that the high AUC and F1 scores of Random Forest were driven by a large number of true positives, rather than a low FPR. In contrast, Naive Bayes demonstrated a sensitivity of 80% and a FPR of 59%, excelling in both identifying true positives and avoiding false positives.

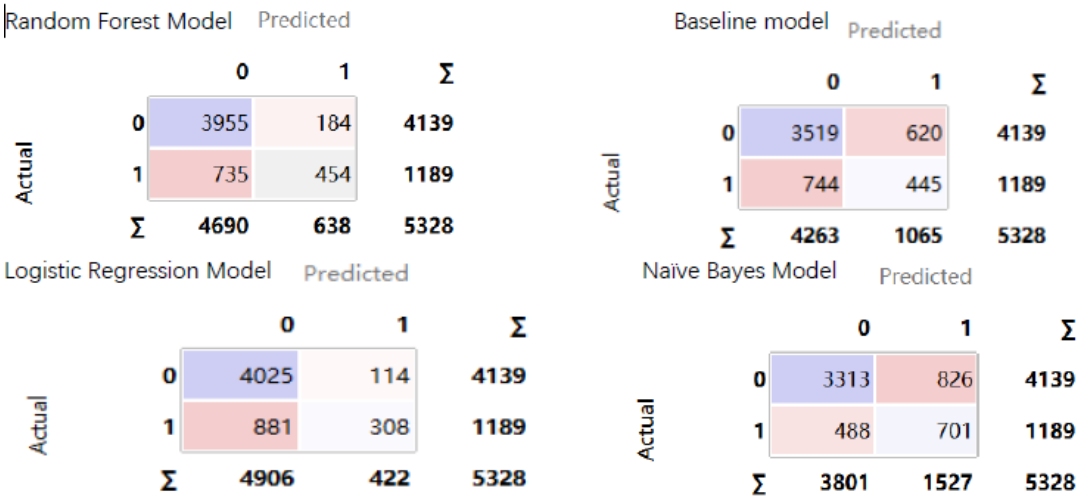

Fig.2

Conclusion & Recommendation

Based on the above analysis, we choose Naive Bayes as the winning classifier. Although Random Forest surpasses Naive Bayes in terms of AUC and F1, we find that the specificity of the other two models is much lower than that of Naive Bayes. In the analysis of default risk, false positives signify people with no default risk in reality is predicted to have default risk, which have minimal impact. In comparison, true negatives indicate that the bank has identified people with default risk as non-default risk, which will affect the bank's lending behavior. Lending to people with default risk will increase the probability of default and is likely to affect the bank's capital reserves. Although Naive Bayes' performance is slightly weaker in other aspects, its specificity performance is far more better, which is essential for the analysis of default risk on borrowers.

Then, to use the *customer predictive workflow* file we provide, customers

should first preprocess the data based on the existing dataset. This involves creating new variables in Excel based on the sample Excel file provided. Once the data has been updated, customers can open the workflow and upload the updated Excel file through the *input file* widget, with all variables converted to numeric format. After uploading, Orange3 will automatically process the data and make predictions, and customers can view the results by clicking on the *result data* widget.

While the credit limit, repayment history, and repayment amount do have an impact on predicting potential customers, these indicators alone do not provide a comprehensive or intuitive understanding of an individual's creditworthiness. We recommend that the company adopt more advanced algorithms for credit assessment of each customer and develop a scoring system, enabling more informed decision-making and personalized customer management.

At last, there are also some limitations in our dataset, such as the incomplete variety of variables, the insufficient number of defaulting customers, and the insufficient time span of data collection, which will affect the accuracy of the final results. In the future, banks should include information like income, years of work, mortgage and car loans, and borrowers' credit history at other financial institutions to judge default risk more comprehensively.

Reference

Senaviratna, N.A.M.R. and A Cooray, T.M.J. (2019) *'Diagnosing multicollinearity of logistic regression model'.* Asian Journal of Probability and Statistics. 5(2), pp.1-9.

Archer, K.J and Kimes, R.V. (2008) *'Empirical characterization of random forest variable importance measures'*. Computational Statistics & Data Analysis*.* 52(4), pp.2249-2260.

Rish, I. (2001). 'An empirical study of the naive Bayes classifier'. IJCAI 2001 workshop on empirical methods in artificial intelligence. pp. 41-46.

Yang, F.J. (2018) *'An Implementation of Naive Bayes Classifier'.* 2018 International Conference on Computational Science and Computational Intelligence. pp. 301-306.