The Bank Conducts Credit Evaluation on Credit Holders

Ye Zeng*

Beijing Institute of Technology (Zhuhai), Zhuhai, China

*Corresponding author: stu2102928@cgt.bitzh.edu.cn

Abstract. The study started with data preprocessing, identifying and removing 88 potential outliers in contact duration based on the 3σ rule. Descriptive stats showed a customer base primarily aged 30-50 with a low median deposit balance (440), brief (4-minute) calls, and infrequent interactions. Graphical analysis via bar and pie charts illuminated key demographics: managerial and blue-collar workers dominated, most customers were married, had a low default rate, and approximately 15.3% had active loans. Most customers didn't use deposit products (89.4%). Scatter plots revealed significant correlations among continuous variables, with 'duration' having a moderate positive link to the dependent variable. Multivariate categorical variables underwent variance tests and multiple comparisons, revealing differences in deposit subscriptions by occupation and education levels. Binary categorical variables were assessed using T-tests. Logistic regression models were trained on five randomly divided subsets, leveraging SMOTE due to class imbalance. The model performed well (accuracy 0.8935, F1 score 0.7107), yet lower recall suggested scope for improvement. The study furnishes insights into customer behavior and proposes avenues for refining credit risk assessment in banking.

Keywords: Credit risk; T-test; model performance; bank risk.

1. Introduction

In the context of commercial banking operations, where credit risk constitutes approximately 60% of the aggregate risk portfolio among the eight principal business risks, enhancing credit risk management is paramount [1]. This is especially critical considering the increasing complexity of both the banking industry's external environment and borrowers' backgrounds, imposing heightened challenges to commercial banks' risk control capacities [2]. To fortify these abilities, credit risk assessment assumes central importance throughout the lending cycle-before, during, and after disbursing loans, as it significantly impacts the operational soundness of these institutions [3].

The quest for refining credit risk assessment methods in light of diverse and intricate borrower profiles drives the need to explore strategies for effectively managing high-dimensional credit data. Consequently, the aim of this study is to devise a multi-dimensional approach to assess borrowers' credit risk, thereby assisting commercial banks to mitigate credit risk and bolster profitability [4]. In this pursuit, literature reveals several innovative methodologies. Xu highlighted that integrating multi-dimensional borrower data like tax records and personal credit checks into logistic regression analyses via loan datasets could convert the correlations between indicators and default probabilities into tangible credit scores [5]. By constructing bank credit scores using scores assigned to various attribute variables, lenders can implement more direct and efficient risk controls.

Ke employed random forests to build a robust personal credit assessment model, adept at extracting significant features and handling outliers, thereby enhancing the model's classification precision. Further, Ke constructed a personal credit evaluation index system utilizing a machine learning confidence network, pinpointing 15 critical items across nine attribute categories, inclusive of crucial aspects such as credit history, citizenship, occupation, and housing status. Ke combined deep networks and random forests to formulate a composite personal credit evaluation model [6]. Fang addressed the complexities inherent in personal credit data-large feature dimensions, intricate correlations, and class imbalance-by leveraging the particle swarm optimization algorithm. This approach integrates a variation mechanism and nonlinear transformation method, allowing for quick and efficient selection of impactful features, thus effectively reducing data dimensionality [7].

Yu contended that incomplete or inaccessible raw data in bank credit risk assessments hinder model effectiveness. Hence, he proposed a multi-view imputation method grounded in BP neural networks, exploiting the interdependence between indicators to fill missing values and utilize the powerful non-linear predictive capabilities of BP neural networks [8]. Lastly, Run conducted extensive research on the performance of multiple credit evaluation models and concluded that unsupervised learning techniques, specifically self-organizing feature map neural networks and K-means clustering models, offer practical advantages in addressing credit evaluation challenges due to their aptitude for fitting and scoring potential borrowers [9].

In sum, each researcher contributes to the collective effort of refining credit risk management practices within commercial banks by exploring innovative ways to optimize feature selection, handle missing data, and harness the power of advanced analytical models for improved credit risk assessment and profitability.

2. Methods

2.1. Data Source

The dataset used in this paper comes from Kaggle about bank lenders. First, the missing value and the duplicate value of the data are tested, no duplicate value and the missing value, no special treatment is done. The 3σ principle is a common principle in statistics, also known as the "3x standard deviation principle". It is based on the assumption of a normal distribution and is used to determine whether an observation has deviated from the mean value. Based on the 3σ principle, standard deviations can be used to estimate the dispersion of data and identify potential outliers or outliers. When an observation exceeds a range of plus or minus 3 standard deviations from the mean, this paper can consider it an outlier or outlier. According to the distribution of contact duration with customers, 88 data are distributed outside 3sigma according to the principle of 3 sigma, which is highly likely to be outliers, and will be directly deleted considering the relatively small proportion [10].

2.2. Method Introduction

The methodology entailed a rigorous statistical examination of the dataset to discern meaningful patterns and relationships pertinent to credit risk assessment. Continuous variables were analyzed, revealing moderate to weak inter-correlations, with 'duration' displaying a notable positive correlation with the dependent variable, suggesting an increased credit risk with longer borrowing periods. Categorical variables were subjected to variance testing, uncovering substantial differences across groups based on occupation, marital status, education level, and deposit product subscription behavior, emphasizing their relevance in discriminating credit risk profiles.

Binary variables ('default', 'housing', and 'loan') were assessed using T-tests, which confirmed significant group disparities at the 0.05 level, reinforcing their importance in credit risk modeling. To address the imbalance in the dependent variable, Synthetic Minority Over-sampling Technique (SMOTE) was employed for oversampling, ensuring a more balanced representation in the dataset. The resulting model exhibited strong overall performance, achieving an accuracy of 0.8935, recall rate of 0.6548, and F1 score of 0.7107 on the test set. While these metrics attest to the model's effectiveness, the lower recall rate indicates a need for further refinement in accurately identifying true positive cases, i.e., borrowers with heightened credit risk.

3. Results and Discussion

3.1. Descriptive Analysis

Descriptive statistics are conducted to show the results of relevant numerical variables. The mean age is 41 years old, and the standard deviation is 10. According to the results of quantile, the data are

concentrated in the 30-50 years old range, the average age of customers is relatively high, and the median deposit balance (440) is relatively low. day indicates that the customer is usually contacted in the middle of each month, duration is 4 minutes on average, and the communication time with them is relatively short. The campaign is concentrated in 1-3 times, and the number of contacts is not much. In addition, the average time since customers were last contacted was 39.77 days ago, which indicates that the frequency of interaction between customers and banks is relatively low. Customers were also contacted relatively few times on average (0.54 times), further indicating a low frequency of interaction.

The professional composition of the surveyed cohort, as deduced from the bar chart (Figure 1) generated through the employment of the matplotlib's pyplot functionality, predominantly consists of managerial roles and blue-collar occupations. These individuals exhibit characteristics indicative of either elevated earning capacities or a discernible inclination towards deposit placements, thereby positioning them as prospective clients with a likelihood of subscribing to deposit-based financial products. The graphical representation serves as compelling evidence to substantiate the assertion that these occupational segments embody a segment of the market with untapped potential for deposit product acquisition.

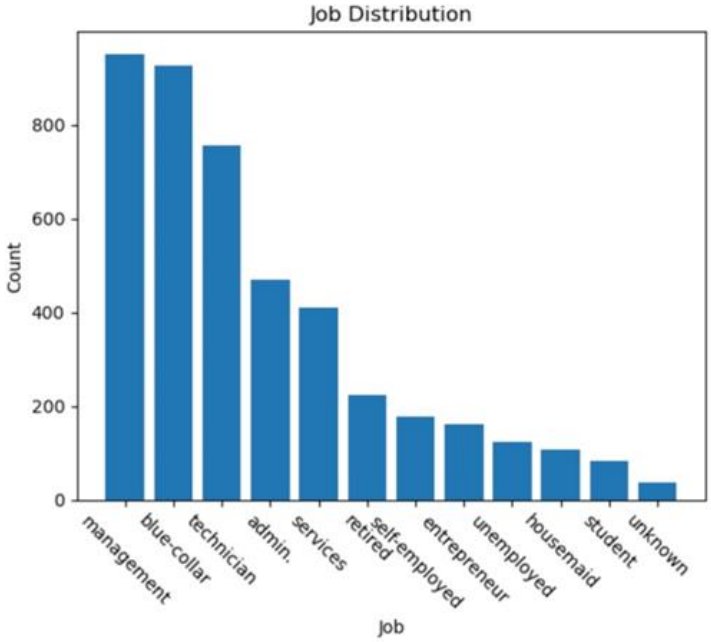

Fig. 1 Job Distribution plot

Then the pie chart is used as a graphical representation tool to illustrate the distribution of various categorical variables. According to Figure 2, 62.1% of the study subjects were married. The prevalence of loan defaults was very low, affecting only 1.7% of participants, while 15.3% of this group had outstanding loans.

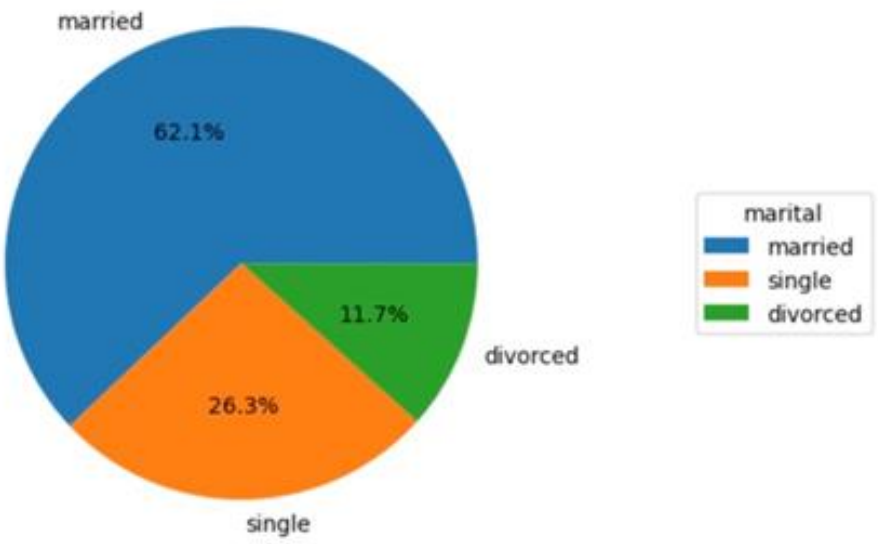

Fig. 2 Married distribution plot

Utilizing the pie chart as a visual representation tool, categorical variables within the surveyed population were subjected to comprehensive examination. The graphical display explicated that the educational attainment levels among the participants were not notably elevated, with over half (50%+) of the individuals possessing merely primary education qualifications. Furthermore, 62.1% of the subjects were reported to be married, indicative of a majority demographic status. Notably, a marginal fraction (1.7%) of the sample exhibited a history of loan defaults, while a somewhat larger proportion (15.3%) held existing loan obligations.

Regarding residential property ownership, the majority of individuals had been reached through cellular communication channels. However, the outcome of the last recorded contact was far from encouraging-a substantial 81.9% of these communications lacked definitive information regarding the recipients' disposition, leaving the intent and receptiveness of the majority populace indeterminate. Conversely, among the minority whose stance was clearly discernible, a disappointing 10.9% manifested a negative response.

Moreover, a strikingly high figure of 89.4% of the clientele had refrained from subscribing to relevant deposit products. Despite this, a closer scrutiny of the demographic attributes, particularly their ages and occupational profiles, suggested the presence of untapped potential savings capacity and latent demand. This discrepancy implies a significant skewness within the dependent variable distribution, highlighting a substantial imbalance that calls for further investigation and targeted marketing strategies.

In the given dataset, the continuous variables under investigation consist of 'age', 'day', 'previous', 'duration', 'campaign', 'pdays', and 'balance'. The scatter plot analysis unveils a discernible pattern of inter-variable correlations. To illustrate this phenomenon, consider Figure 3 which specifically depicts the scatter plot of 'duration' against the dependent variable (y). Herein, the Pearson correlation coefficient between 'duration' and y registers at 0.3794, accompanied by a corresponding p-value of 8.910, thereby indicating the statistical significance of the correlation.

At the established significance level of $\alpha = 0.05$, the results of the correlation analysis highlight that the correlation coefficients for 'age', 'previous', 'duration', 'campaign', and 'pdays' are indeed statistically significant. Among these, the associations between 'age', 'previous', and 'pdays' exhibit weak correlation tendencies. Conversely, 'duration' displays a moderate positive linear correlation ($r = 0.3794$) with the dependent variable (y). Furthermore, 'campaign' demonstrates a weak but negative linear correlation with y, adding another dimension to the intricate web of relationships within the dataset.

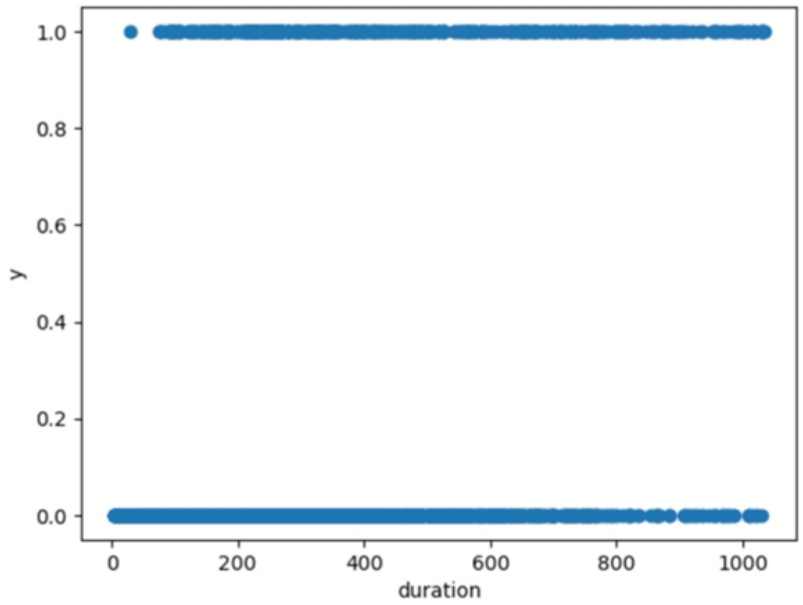

Fig. 3 Scatter Plot-day vs

3.2. Model Results

Regarding the multivariate categorical variables, could observe a positive linear correlation, while an activity exhibits a weak negative linear association. Variables such as 'work', 'marriage', 'education', 'contact', 'month', and 'outcome' have been subjected to variance testing post the execution of a preliminary test for homogeneity of variances. Subsequent multiple comparison analyses were conducted following the variance tests.

For the multi-class categorical variables, the 'education', 'contact', 'month', and 'poutcome' datasets met the criteria for the test. Within the 'job' category, significant discrepancies were found between retirees and other occupational groups including administrative staff, managers, blue-collar workers, and entrepreneurs, particularly regarding subscription to deposit products. In the context of marital status ('marriage'), disparities were detected among the married, single, and divorced groups. Educational attainment ('education') also revealed differences between those with higher education and those with primary or secondary education levels. However, no significant distinction was observed between communication methods ('contact') -cellular phones versus landlines. Variability across different months ('month') did not always show uniformity, but some differences emerged. As for 'outcome', divergences existed across all categories, thus suggesting its potential inclusion in the independent variable scope based on the outcome of the difference test.

Regarding binary categorical variables ('default', 'housing', and 'loan'), the T-test was employed. As shown in Table 1, According to the T-test results, the value of default variable is greater than 0.05, suggesting that there is no difference between groups. The data of housing and loan variables are significant at the significance level of 0.05 and can be included in the scope of independent variables.

Table 1. T-test results

Numble	Statistic	P value
LeveneResult	0.0002	0.995
TtestResult	0.005	0.995
LeveneResult	54.663	<0.0001
TtestResult	-7.393	<0.0001
LeveneResult	24.202	<0.0001
TtestResult	-4.919	<0.0001

Use grid search cross validation to select the best valid values that will be included in the independent variables, and use a simple logistic regression model to find the coefficient values of

each independent variable. The coefficient values are the weights of the corresponding features of the logistic regression model, and they represent the degree of influence of each feature on the predicted target. The positive or negative of the coefficient determines the direction of the relationship between the feature and the target, and the absolute value of the coefficient indicates the importance of the feature to the target.

Figure 4 shows the print coefficient values. Taking "age" as an example, the age coefficient is negative, indicating that the older you are, the less likely you are to buy a product. As for "occupation", the coefficient of different occupation is different. Positive value indicates that the occupation tends to buy products, negative value indicates that the occupation tends not to buy products. For example, a higher coefficient for retirees and students indicates that these groups are more likely to buy products.

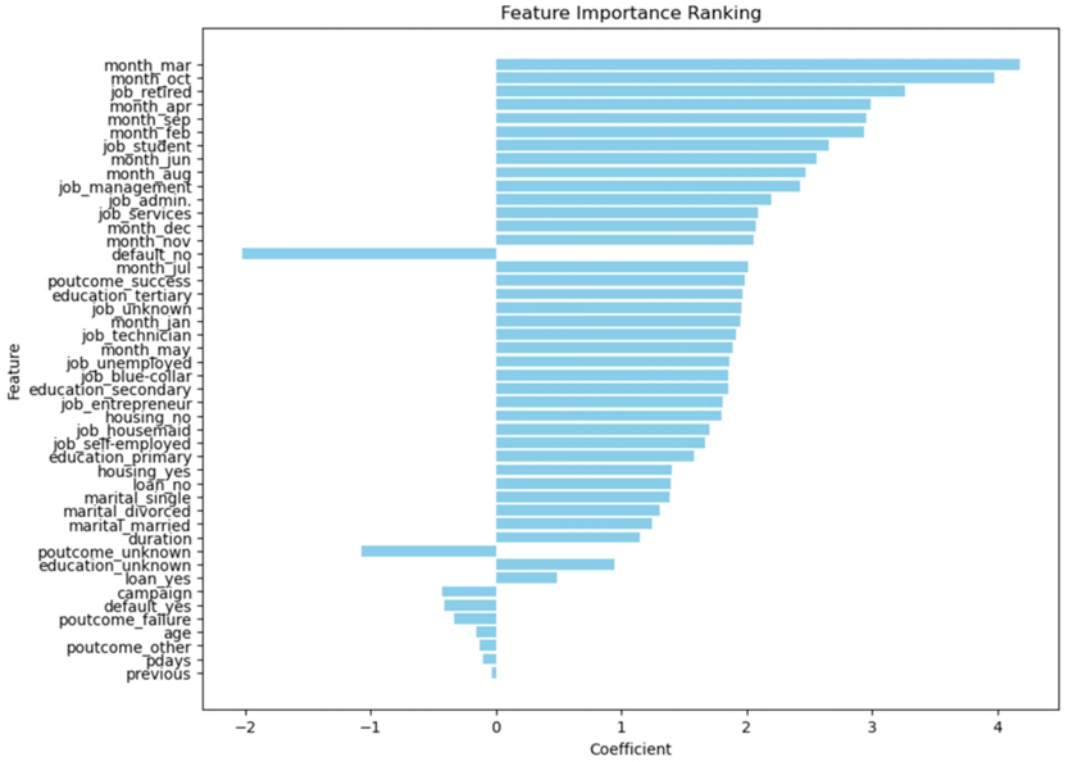

Fig. 4 Importance of different variables

4. Conclusion

This study underscores the vital role of refining credit risk assessment methods in commercial banking operations, given the dominance of credit risk and its impact on banks' operational stability and profitability. Researchers explored multifaceted approaches to enhance credit risk evaluation, including the integration of multidimensional borrower data in logistic regression models, application of random forests for precise credit assessment, utilization of particle swarm optimization to handle complex personal credit data, BP neural networks for imputing missing values, and unsupervised learning techniques such as self-organizing feature maps and K-means clustering.

The empirical analysis revealed insightful findings regarding the customer base. A substantial portion of the surveyed population comprised high-income earners or those with a propensity to save, particularly managers and blue-collar workers. The dataset showed a low educational level among respondents, with most being married and having minimal loan defaults. However, there was a significant imbalance in deposit product subscriptions, as evidenced by the large percentage of customers not subscribed to any deposit product, despite demonstrating potential eligibility based on demographic and occupational characteristics.

Statistical analysis of continuous variables demonstrated moderate to weak correlations, with 'duration' showing a noteworthy positive correlation with the dependent variable. Categorical variables underwent variance testing, revealing significant differences among groups related to occupation, marital status, education level, and deposit product subscription behavior.

Applying T-tests to binary variables ('default', 'housing', and 'loan') indicated group differences under the significance level of 0.05. Due to the dependent variable's imbalance, SMOTE was used for oversampling. The resultant model achieved a commendable accuracy of 0.8935 on the test set, with a recall rate of 0.6548 and an F1 score of 0.7107, signifying effective overall performance albeit suggesting a need for improvement in identifying true positive cases.

In conclusion, the adoption of innovative strategies drawn from various methodologies has proven beneficial in refining credit risk assessment in commercial banks. The study's outcomes provide valuable insights for future enhancements, emphasizing the importance of tailored risk management strategies that account for customer demographics, interaction patterns, and behavioral characteristics to optimize credit decisions and boost profitability.

References

[1] Zhou Shanghua. Research on Risk management of innovative SME loans in A Bank. Southwest university of science and technology, 2024.

[2] Zhou Xiaoping. H bank credit risk management research. Southwest university of finance and economics, 2022.

[3] Yi Mengyun. Research on risk assessment of Y Bank's small and micro credit based on Analytic Hierarchy Process. The three gorges university, 2023.

[4] Permission, Ding Pan, Yan Lei, et al. Evaluation of the effectiveness of innovative direct monetary policy tools: Evidence from quasi-natural experiments of local corporate banks. Southern Finance, 2021, 10: 10-21.

[5] Shen Lvzhu, Guo Wenlong. Empirical Analysis of Credit Risk Assessment and Measurement of Chinese commercial Banks: Based on Multiple Linear Discriminant Model and Logistic Regression Model. Proceedings of International Conference on Engineering and Business Management, 2011.

[6] Feng Yuxue. Multi-view filling method of bank credit risk assessment data based on BP network. Public Standardization, 2021, 8: 86-89.

[7] Su C. Research on credit risk assessment of commercial banks based on Logistic regression model. Chinese Urban Economy, 2011, 12: 72.

[8] Lan N V D T. Research on Credit risk Assessment of Listed commercial banks in China and Vietnam based on neural network model. Hunan University, 2014.

[9] Zhu Jinhua. Research on Combination model in credit risk assessment of commercial banks. Computer Simulation, 2011, 28(9): 361-364.

[10] Niu Xuecheng. Empirical Analysis of credit risk default evaluation model of Chinese commercial banks. Wuhan Finance, 2008, 7: 39-42.