Bank Customer Churn Based on Different Models, Oversampling, and Encoding Methods

Zekun An[1, a, †], Zhangdi Song[2, b, †], Xueying Wang[3, *, †]

[1]Business Administration Zhuhai College of Jilin University, Zhuhai, China

[2]Master of Data Science Monash University, Melbourne, Australia

[3]School of Psychology Jiangxi Normal University, Nanchang, China

*Corresponding author: honorwang@jxnu.edu.cn, [a]Kunken12864900@gmail.com, [b]songzhangdi2019@163.com

[†]These authors contributed equally

Abstract. Customer churn prediction (CCP) is one of the cornerstones of Customer Relationship Management (CRM), in which one seeks to forecast whether or not a customer will quit the organization. Currently, plenty of algorithmic focuses on CCP. To fill the gap in the current study, this paper builds different models to predict bank user churn based on data from Kaggle. Specifically, we investigate the difference between models with and without oversampling, as well as discuss the difference between models under different coding methods. According to the results, 'smote' does not necessarily improve the performance accuracy, one hot encoding is more effective than target encoding. Finally, after all aspects of comparison, the logistic regression model is more reliable in the future analysis of customer churn of commercial banks. These results offer a guideline for future bank customer churn prediction.

Keywords: Bank Churn Prediction; Machine Learning; Logistic Regression; Gradient boosting; Neural Network.

1. Introduction

Definitely, the financial industry is evolving at a rapid and increasing rate, in response to discernible variations in customs' preferences and requirements, which are fueled by state-of-art techniques and the wide capacity of various products and supports. Because of the threats posed by not only competitors in the same industry, but also new and inventive enterprises (e.g., Apple, Google), the banking business has become extremely competitive. Therefore, maintaining a competitive advantage for the sake of staying in the financial path of the customers is considered as the highest priorities (at least one of) in strategic planning related to predicting the probability of bank customer churn, they tested two models based on decision tree classifier and artificial neural network. customer attraction and, more importantly, retention in the retail banking sector. Although different types of CRM strategies have existed for decades, it has been regarded as the center of attention for many researchers and practitioners after the business world shifted its marketing focus from a product-centric strategy to a customer-centric strategy. As a result, the relationship between the customer and the company has evolved in such an approach that many novel marketing strategies have been created [1]. One of the cornerstones of CRM is churn prediction [2], where one tries to predict no matter whether a customer will leave the company or not. Furthermore, the churn rate is defined as an annual ratio when a customer stop subscribing to a service or terminates a business relationship. It is known that the cost of acquiring new consumers is 5 to 25 times higher than retaining existing customers [3]. Reducing customer churn and retaining existing customers is one of the most cost-effective marketing approaches to maximize shareholder value [4-7]. In such a competitive environment, companies need to focus on retaining existing customers by effectively meeting their needs [8]; instead, they will face the situation of reduction of customers quantity, offering competitors the opportunity to attract them.

Currently, there is a lot of algorithmic research on CCP, which focuses on exploring and improving algorithms for building customer churn prediction models. A study was carried out to examine the efficacy of two Spark packages in predicting the likelihood of banking clients churning (ML and MLlib)

[9]. To be more specific, the Kaggle dataset is applied to test the accuracy of the model when dealing with consumer transactional data. The results revealed that the ML Apache Spark package is more precise and accurate than the MLlib package. A dataset was used to forecast how customers would react to a bank's offer [10]. The results of the case study showed that the best model for predicting the target response variable was the random forest classifier, which had the strongest predictive power with an accuracy of 87% and an AUC value of 92.7%. Another study was carried out to predict the probability of bank customer churn, they tested two models based on decision tree classifier and artificial neural network [11]. The results revealed that the Neural Network model had the highest accuracy of 86.5%, while it is only about 79.8% for decision tree model.

However, previous studies ignored the treatments effects of models on performances, though they have compared the performance of different models. To fill the gap in the current study, this paper builds different models to predict bank user churn, investigates the difference between models using oversampling and those without oversampling, and investigates the difference between models under different coding methods. Specifically, models with different coding methods (one hot coding and target coding) and a data balancing (SMOTE) approach will be implemented with the building of a bank customer churn model. The second part will describe the data, present the model and its evaluation methods. The results will be presented and discussed in the third part and conclusions will be drawn in the fourth part.

2. Data & Model description

2.1 Data Description

With the financial derivatives and alternative investment developed more and more mature, customers quit the credit card market, thus the bank will face the risk of prepayment. If one can predict the numbers of churns precisely, we can base on the existing situation to lower the operational risk. From the data source (which came from the URL: https://leaps.analyttica.com/home), we got nearly 18 features and could use them to obtain well-performance models. Overall, 10,000 customers are recorded in the datasets including their sociodemographic and financial information, (e.g., age, salary, credit card limit, etc.).

2.2 Logistic Regression

The association between a dependent binary variable and one or more independent variables are modeled using logistic regression. This produces the standard coefficients and errors for the significance levels of metrics utilized to predict a logit transformation of the odds of an event occurring [12]. Based on independent factors, the classifier models the logarithm of the probability of an outcome. Finally, for each observation, the model forecasts a chance of occurrence with values in the range (0, 1). The model can be expressed mathematically by the following equation [13]:

$$logit(y) = c_0 + c_1x_1 + c_2x_2 + \cdots + c_kx_k \tag{1}$$

Where x_1~x_k are independent variables and y is the probability of the event, while logit(y) = ln[y/(1-y)]. Therefore, the equation above can be written as:

$$y = \frac{1}{1+e^{-(c_0+c_1x_1+c_2x_2+\cdots+c_kx_k)}} \tag{2}$$

2.3 Gradient Boosting

The gradient boosting model is a well-known approach [14]. The core idea of the gradient boosting model is to integrate numerous weak learners to improve the accuracy and robustness of the final model. Generating a single leaf and regression trees is the first step in the gradient boosting model. A

regression tree is not a classifier, but a decision tree to evaluate a continuous real-valued function. The regression tree is built in terms of an iterative process that divides the data into nodes or branches, which are further divided into smaller groups. Initially, all observations are grouped together. Then, the data is divided into two parts, with each conceivable partition applied to each available predictor.

Another tree is trained by the gradient boosting model based on the preceding tree's error, which continues in this way until the required number or fit cannot be improved any further. To avoid the problem of overfitting, the model scales the contribution from the new tree using a learning rate.

2.4 Neural network

Our daily activities are dominated by nerves, which involve nerve cells. The nerve cell (so-called neuron or neurogen), is one of the basic units of the nervous system, which is feasible to sense changes and convert the information to other nerve cells, and issue instructions to respond.

An artificial neural network (ANN) is a model proposed in consideration of the working basis of neural networks in biology. ANN can simulate the neural system of the human brain to realize the processing mechanism of complex information. As a complex network with many simple components connected, ANN has a high degree of nonlinearity. It is one of the important research directions of formal machine learning [15-17].

2.5 Evaluation Methods

- Accuracy_Score: classification accuracy score. The function returns a score, which is either the proportion of the correct results or the correction numbers.
- AUC (area under the curve) is defined as the area enclosed by the coordinate axis under the ROC curve.
- Kolmogorov Smirnov test (KS test for short) is a nonparametric hypothesis test in statistics, which is used to test whether a single sample obeys a certain distribution or whether two samples obey the same distribution. Generally speaking, if the KS statistic is more than 0.6, it indicates that the classification ability of the model is relatively strong [18].

3. Result & Discussion

During the experiment, we first performed a visual analysis of the dataset, the results indicate that the data is imbalance, and oversampling of the data can be considered. To show the correlation between variables, some new relevant variables were added to the original variables during feature engineering, and two different feature encoding methods were used for the categorical features. Finally, three different models were constructed based on the selected features. Whether the data has been sampled or not is treated as a variable .Combined with the different encoding methods, we compared the performance of the three models in different cases. The optimal model was obtained by adjusting the relevant parameters.

3.1 EDA

According to the type of data, the variables in this dataset can be divided into numerical variables and categorical variables, based on the meaning of each column, they can also be divided into customer personal information and bank card usage. Among them, "Attrition_Flag" can be used as label to determine whether a customer has churned or not.

Figures 1 and 2 show the distribution of numerical variable and categorical variable. From these figures, the dataset basically includes all categories of customers, but the number of customers in each category is not evenly distributed. The distribution of customers on personal information category variables is similar, while the difference between existing customers and churned customers is mainly shown in the use of cards (not shown here). Figure 3 illustrates the distribution of target customers, and it can be found that the number of churned customers is much less than that of existing customers,

which will cause an imbalance in the sample. In order to address the issue, SMOTE algorithm is chosen to oversample the data.

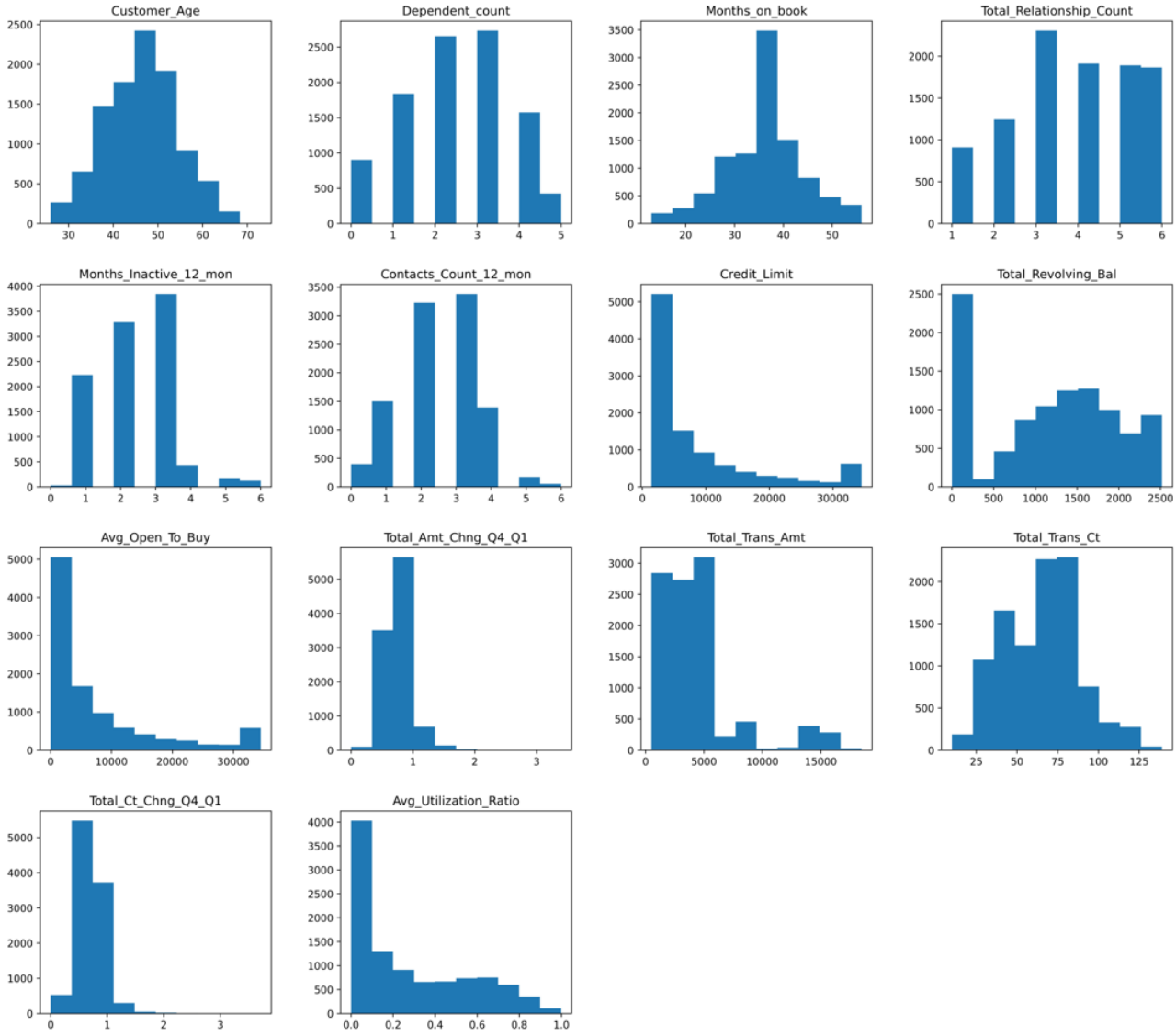


Figure 1. The variable dirstribution (numerical).

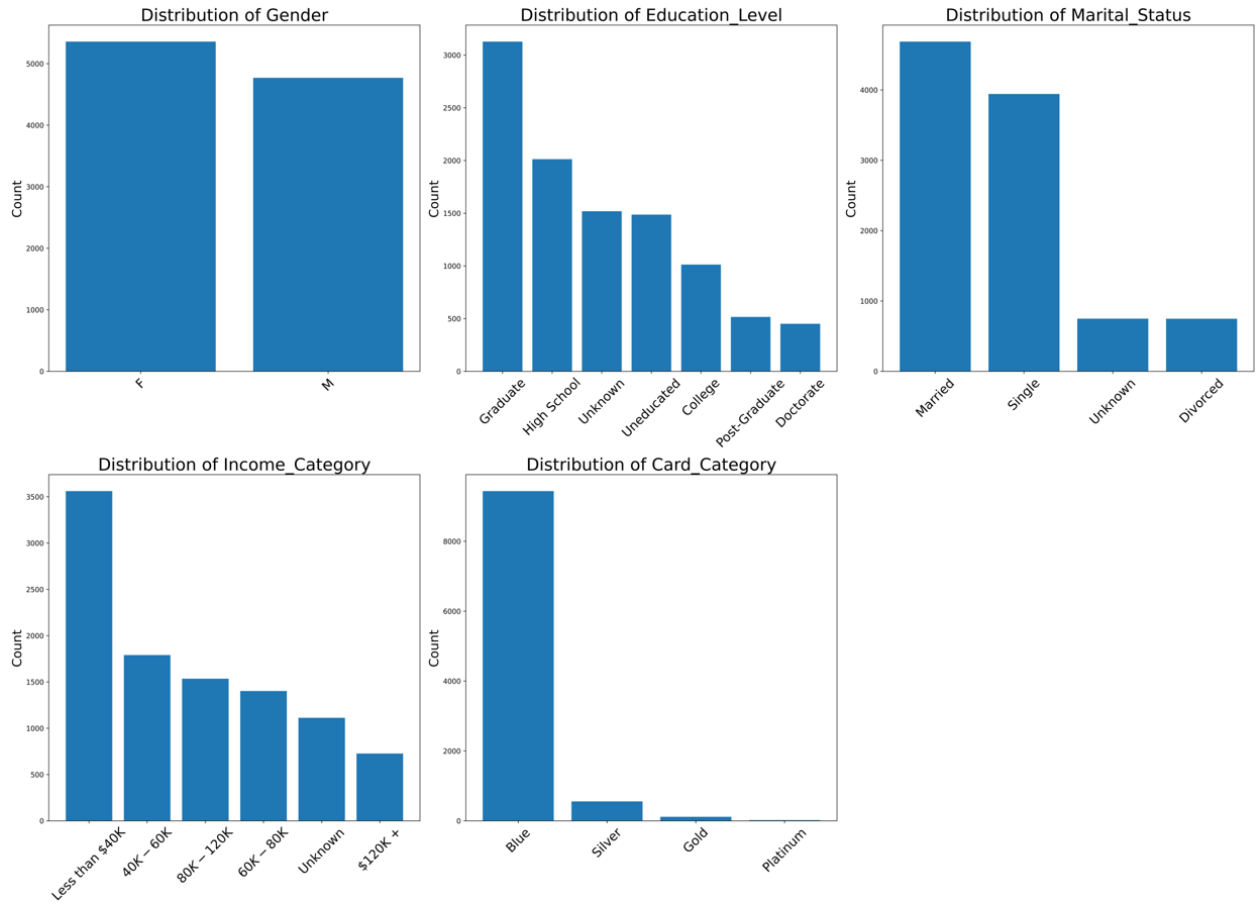


Figure 2. The variable dirstribution (Categorical).

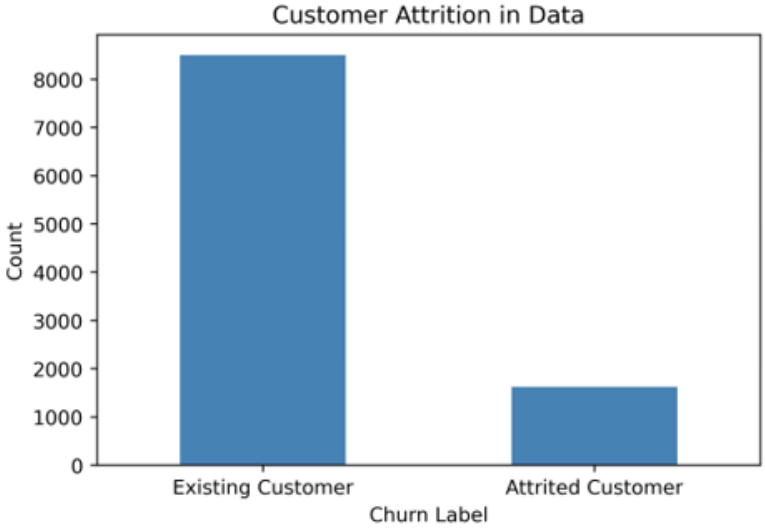


Figure 3. The dirstribution of target customs.

3.2 Feature Engineering & Feature Selection

In this part, we performed feature engineering to further investigate the correlation. To narrow the range of values and facilitate calculations, log transformations on numeric variables is performed. In addition, for categorical variables, the data were grouped according to the elements in each category, and statistical analysis variables (average, the number of values, max value, min value, the number of unique values, standard deviation, variance, skew and median) were added for each category in each group. After adding statistical variables, polynomial transformations were performed on all numeric columns, categorical variables were transformed by Target Encoding and One-Hot Encoding

respectively. Finally, 792 variables are obtained. It is also necessary to prevent overfitting while ensuring the accuracy of the model prediction. For this purpose, the features with higher correlation were filtered by Kolmogorov-Smirnov and Churn Detection Rate Univariate Filter, and then find the best number of features with Cross-Validated Recursive Feature Elimination.

Figures 4 and 5 exhibit the Spearman correlation matrix for each encoding methods. Apparently, One-Hot Encoding method get more variables and the correlation between variables is slightly higher than that of Target Encoding. With Cross-Validated Recursive Feature Elimination, for the Target Encoding dataset, we filtered out 9 features, while for the One-Hot Encoding dataset, it can get the highest accuracy with 17 features (as shown in Figure 6). Table I lists the best variables filtered for both datasets, comparison reveals that all the variables of Target Encoding are included in the features of One-Hot Encoding.

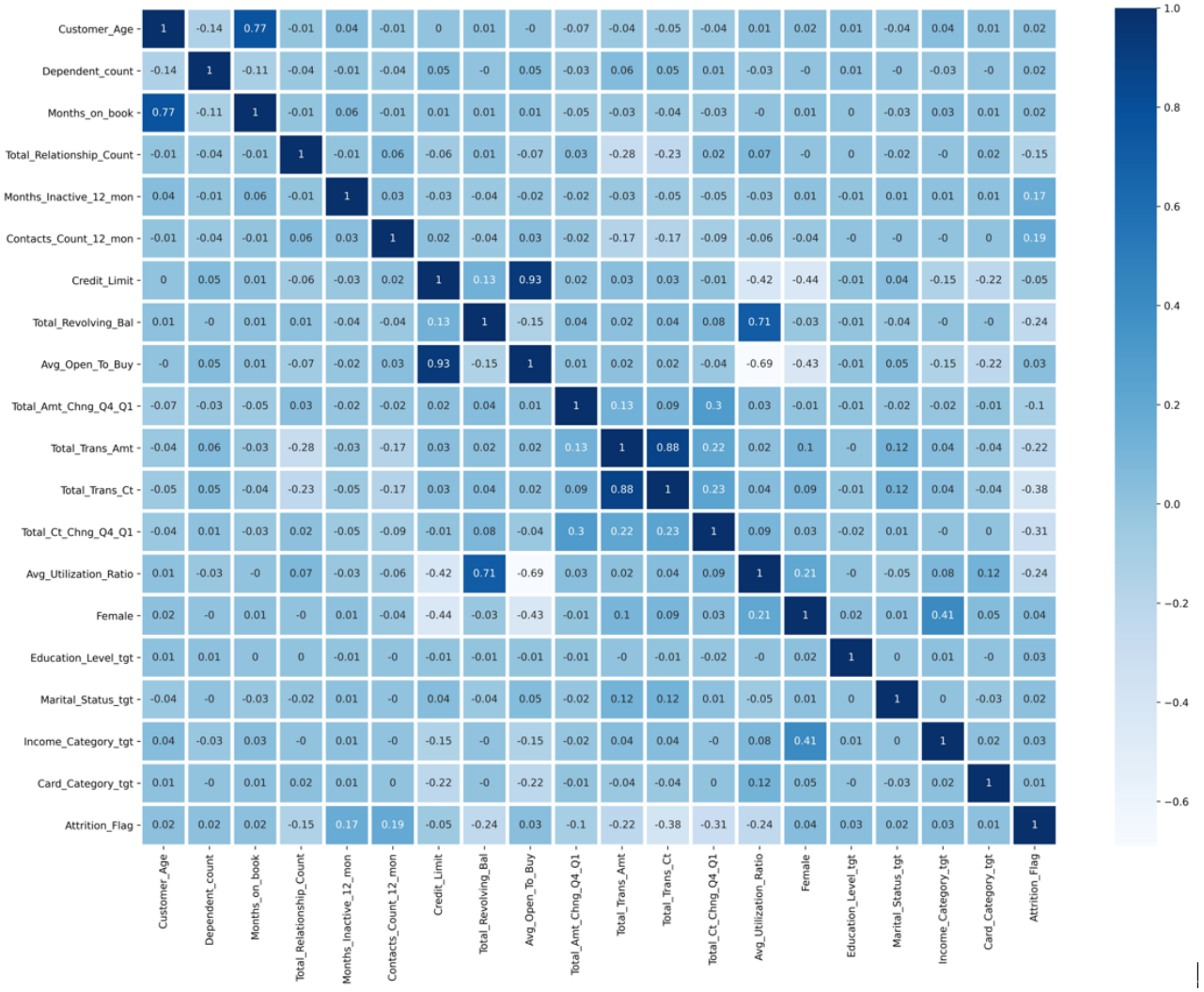


Figure 4. The Spearman correlation matrix of Target Encoding.

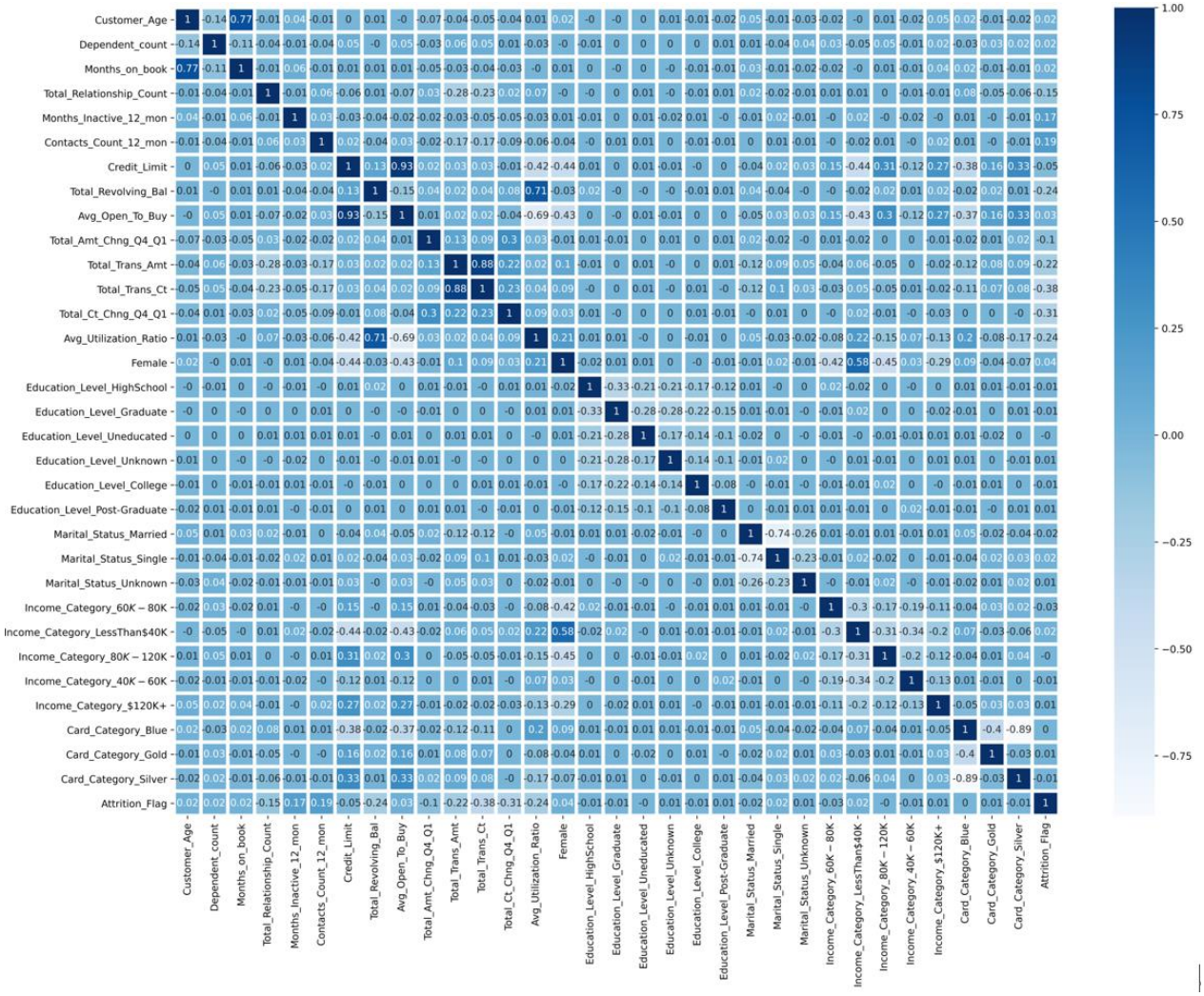


Figure 5. The Spearman correlation matrix of One-Hot Encoding.

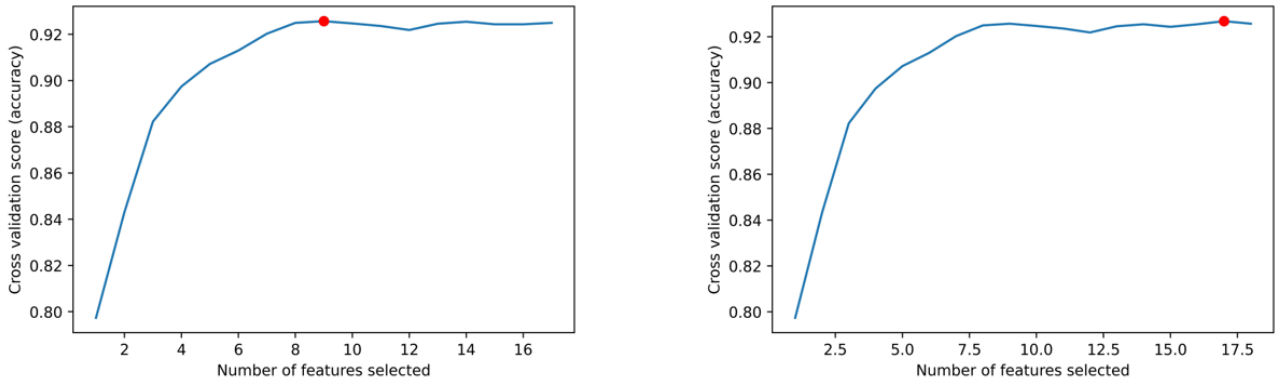


Figure 6. The number of features with the highest accuracy (the upper panel for the target encoding while the lower panel for the ONE-HOT encoding).

Table 1. Feature name for each encoding method.

Encoding Method	Feature name
Target Encoding & One-Hot Encoding	Total_Trans_Amt
	Total_Trans_Ct
	Total_Revolving_Bal
	Total_Ct_Chng_Q4_Q1
	Total_Relationship_Count
	Total_Amt_Chng_Q4_Q1
	Avg_Utilization_Ratio
	Customer_Age
	Credit_Limit
One-Hot Encoding	Avg_Open_To_Buy
	Contacts_Count_12_mon
	Months_on_book
	Months_Inactive_12_mon
	Dependent_count
	Marital_Status_Single
	Income_Category_LessThan$40K
	Education_Level_Graduate

3.3 Model performance

In this subsection, three models (Logistic Regression, Gradient Boosting and Neural Network) are constructed based on the filtered features. Besides, the performances of the models are compared on the original dataset and the oversampled dataset. In addition, the effect of imbalanced data on the models (as shown in Figure 7) are also analyzed. Moreover, we also compare the differences in model performance under different encoding methods (as shown in Figure 8). Finally, the optimal model for this dataset was selected by adjusting the corresponding parameters, and the accuracy of the optimal model was about 0.968. The hold-out test proved that the model did not have the risk of overfitting.

Figure 7 compares the performance of the three models on different dataset. Obviously, the performance of Logistic Regression under both encoding methods is almost independent of the imbalanced data, and the performance of Neural Network on the original dataset is better than that of sampled data. While for the Gradient Regression model, processing the data using the SMOTE algorithm with the One-Hot Encoding method reduces the accuracy of the model. On this basis, the data processing based on SMOTE does not necessarily improve the model accuracy for unbalanced data.

Figure 8 compares the performance of the model under different encoding methods. According to the results, the performance of Logistic Regression is more stable and does not change significantly. While the performance of Gradient Boosting on the original dataset is affected by the encoding method, the accuracy of One-Hot encoding is higher. The performance of Neural Network on both datasets was affected by the encoding method as well, and the difference was most obvious on the sampled data, similar with Gradient boosting, One-hot encoding can improve the model performance.

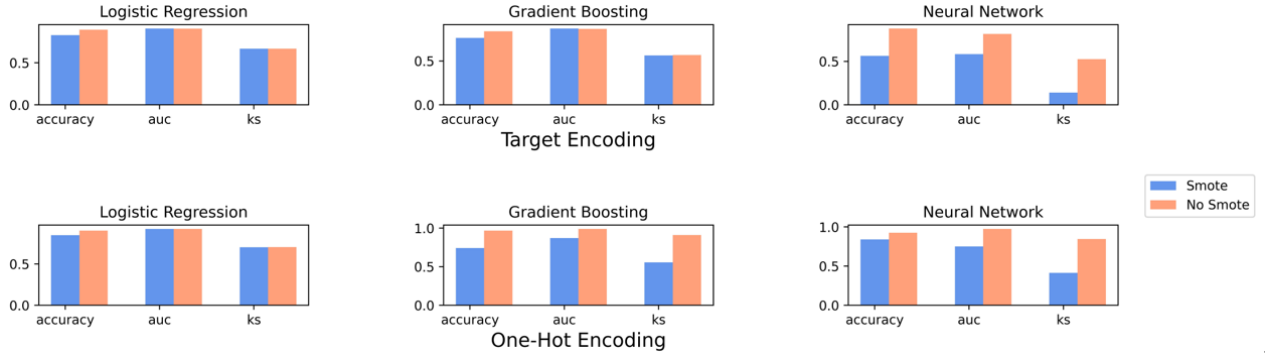


Figure 7. The model performance based on SMOTE.

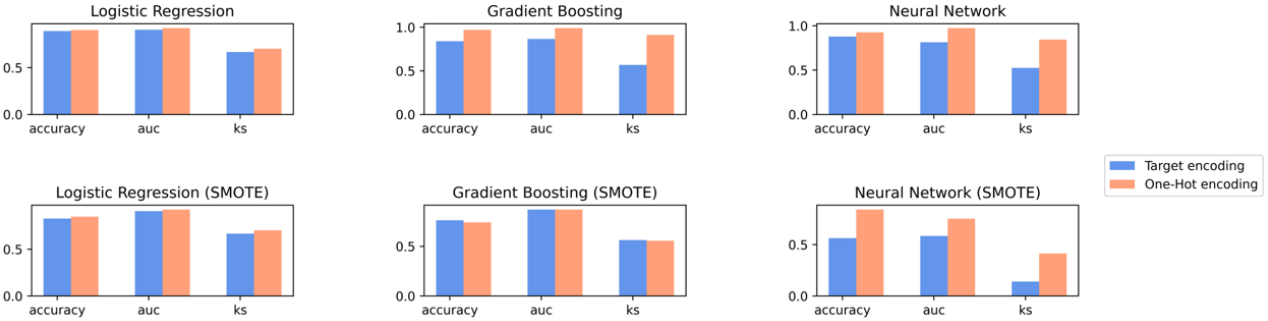


Figure 8. The model performance based on different encoding methods.

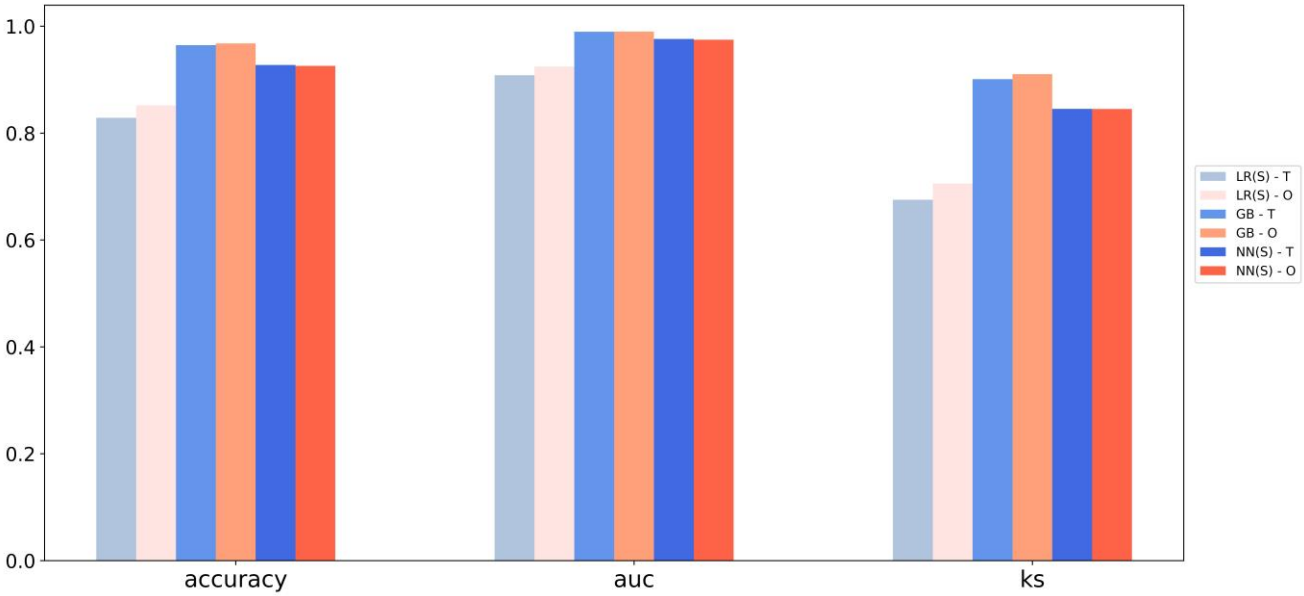


Figure 9. The model performance with the best parameters.

We adjusted the corresponding parameters of the models, selected the optimal parameters of the corresponding models (as shown in Table II), and tested the models for the risk of overfitting by hold-out set. The results showed that the performance of the three models with two encoding methods was not overfitting. Figure 9 compares the performance of the six models, and it can be found that Gradient boosting performs significantly better than the other two models. After tuning the parameters, the difference in the performance of the models with different encoding methods is not large. From Table II, Gradient boosting has the best performance on the original dataset. It is worth noting that the output results of the models are not consistent each time, and for this dataset, sometimes Gradient boosting on the oversampled dataset can also be optimal.

Table 2. Adjusting parameters.

Model	Parameters	Best Parameter	
		Target Encoding	**One-Hot Encoding**
Logistic Regression	c: [0.001, 0.01, 0.1, 1, 10, 100, 1000] Smote: [True, False]	C :1000 Smote = True	C :100 Smote = True
Gradient Boosting	max_depth=3, n_estimators: [1000, 1200], learning_rate: [0.005, 0.01] Smote: [True, False]	learning_rate: 0.01, max_depth: 3, n_estimators: 1200 Smote = False	learning_rate: 0.01, max_depth: 3, n_estimators: 1200 Smote = False
Neural Network	hidden_layer_sizes: [(5,), (10,), (20,), (10,5), (8,3)], activation: ['relu', 'logistic'] Smote: [True, False]	activation: relu, hidden_layer_sizes: (10, 5)} Smote = True	activation: relu, hidden_layer_sizes: (10, 5)} Smote = True

4. Conclusion

In summary, the impact of bank customers churn is investigated based on machine learning. Through the process of building machine learning model, SMOTE does not necessarily improve the performance accuracy. Since the data source only involves a relatively short time, there will be more errors in the application of the model for a long time. In the model established in this problem, One-Hot encoding is more effective than target encoding. Finally, logistic regression performances well after all aspects of comparison. These results summarize the factors in the loss of commercial bank customers and extract the key factors for analysis, i.e., strengthen the customer trust time and expand the income, avoid customer churn and reduce operational risk. In the future development of the financial industry, it is necessary to better understand the needs of customers and analyze the core factors in the hearts of customers to judge the development trend of banks in the future. Moreover, we have judged the effectiveness of the three models. Specifically, the logistics expression model has the best effect, which proves that the logistics expression model is more reliable in the future analysis of customer churn of commercial banks. Overall, these results offer a guideline for machine learning model construction for bank customers churn.

References

[1] Ngai, E. W. “Customer Relationship Management Research (1992-2002).” *Marketing* Intelligence *& Planning*, vol. 23, no. 6, Oct. 2005, pp. 582–605, https://doi.org/10.1108/02634500510624147.

[2] Coussement, Kristof, et al. “Preventing Customers from Running Away! Exploring Generalized Additive Models for Customer Churn Prediction.” *The Sustainable Global Marketplace*, Oct. 2014, pp. 238–38, https://doi.org/10.1007/978-3-319-10873-5_134.

[3] Gallo, Amy. “The Value of Keeping the Right Customers.” *Harvard Business Review*, 5 Nov. 2014, hbr.org/2014/10/the-value-of-keeping-the-right-customers.

[4] Ekinci, Yeliz, et al. “A Customer Lifetime Value Model for the Banking Industry: A Guide to Marketing Actions.” *European Journal of Marketing*, vol. 48, no. 3/4, Apr. 2014, pp. 761–84, https://doi.org/10.1108/ejm - 12 - 2011 - 0714.

[5] Ngai, E. W. T., et al. "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification." *Expert Systems with Applications*, vol. 36, no. 2, Mar. 2009, pp. 2592–602, https://doi.org/10.1016/j.eswa.2008.02.021.

[6] Ryals, Lynette J., and Simon Knox. "Measuring Risk-Adjusted Customer Lifetime Value and Its Impact on Relationship Marketing Strategies and Shareholder Value." *European Journal of Marketing*, vol. 39, no. 5/6, May 2005, pp. 456–72, https://doi.org/10.1108/03090560510590665.

[7] Van den Poel, Dirk, and Bart Larivière. "Customer Attrition Analysis for Financial Services Using Proportional Hazard Models." *European Journal of Operational Research*, vol. 157, no. 1, Aug. 2004, pp. 196–217, https://doi.org/10.1016/s0377-2217 (03) 00069 - 9.

[8] Lin, Chiun-Sin, et al. "Combined Rough Set Theory and Flow Network Graph to Predict Customer Churn in Credit Card Accounts." *Expert Systems with Applications*, vol. 38, no. 1, Jan. 2011, pp. 8–15, https://doi.org/10.1016/j.eswa.2010.05.039.

[9] Sayed, Hend, et al. "Predicting Potential Banking Customer Churn Using Apache Spark ML and MLlib Packages: A Comparative Study." *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018, https://doi.org/10.14569/ijacsa.2018.091196.

[10] Asare-Frempong, Justice, and Manoj Jayabalan. "Predicting Customer Response to Bank Direct Telemarketing Campaign." *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*, Sept. 2017, https://doi.org/10.1109/ice2t.2017.8215961.

[11] Kumar, G. Ravi, et al. "Client Churn Prediction of Banking and Fund Industry Utilizing Machine Learning Techniques." *International Journal of Computer Sciences and Engineering*, vol. 7, no. 6, June 2019, pp. 842–46, https://doi.org/10.26438/ijcse/v7i6.842846.

[12] Sperandei, Sandro. "Understanding Logistic Regression Analysis." *Biochemia Medica*, vol. 24, no. 1, 2014, pp. 12–18, https://doi.org/10.11613/bm.2014.003.

[13] Yanchang Zhao. *R and Data Mining: Examples and Case Studies*. Academic Press, An Imprint of Elsevier, 2013, pp. 1189 – 232.

[14] Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals* of *Statistics*, vol. 29, no. 5, Oct. 2001, pp. 1189–232, https://doi.org/10.1214/aos/1013203451.

[15] Pan, Indranil, et al. "Data-Centric Engineering: Integrating Simulation, Machine Learning and Statistics. Challenges and Opportunities." *Chemical Engineering Science*, vol. 249, Feb. 2022, p. 117271, https://doi.org/10.1016/j.ces.2021.117271.

[16] Wen, Fei, et al. "A Survey on Nonconvex Regularization-Based Sparse and Low-Rank Recovery in Signal Processing, Statistics, and Machine Learning." *IEEE Access*, vol. 6, 2018, pp. 69883 – 906, https://doi.org/10.1109/access.2018.2880454.

[17] Barrow, Devon, et al. "Automatic Robust Estimation for Exponential Smoothing: Perspectives from Statistics and Machine Learning." *Expert Systems with Applications*, vol. 160, Dec. 2020, p. 113637, https://doi.org/10.1016/j.eswa.2020.113637.

[18] Željko Ivezić, et al. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press, 2020.