# Investigation on credit card customer cancellations: A case study of AMEX credit cards

Ruidi Gao[1, *]

[1]College of Letters and Science, University of California, Davis, The U.S.

*Corresponding author: rdgao@ucdavis.edu

**Abstract.** American Express (AMEX) is one of the most popular credit card services in the U.S. Lately, the company is facing loss of some valued customers for specific reasons. This paper investigates the customers' cancellation behavior based on XGBoost in terms of the AMEX data. Besides, the factors contributing to customers' cancellations are identified, e.g., customer age, total transaction counts, total transaction amount. According to the analysis, the proposed XGBoost model can reach 96.5% in accuracy. By implementation of such tool, it is feasible for the bank to predict their customer behavior and take measures proactively. These results shed light on guiding further exploration of credit card services.

**Keywords:** XGBoost, Credit Card Customer, Machine Learning, Data Mining.

## 1. Introduction

American Express (AMEX), is one of the largest and the most popular credit card companies in the U.S. [1, 2]. It is a leader in providing first-class service, ranking No. 1 in customer satisfaction among issuers nationwide, according to J.D. Power's 2021 American Credit Card Customer Satisfaction Study. What's more, American Express is accepted in 99% of places that accept credit cards in the US, with more participating merchants accepting contactless payments for secure purchases. ns in the United States. In addition, in 2017, AMEX was annotated as the 23rd most valuable brand in the world by Forbes.

However, in the 12 months in 2021, a large number of customers are stopping the credit card services. This is probably related to the simulated checks government issued with COVID-19 pandemic. Customers have more cash flow owing to the promulgated policies so that they would like to cancel their credit card services.

In this paper, it will predict who will be affected in the future. In this way, the bank can proactively provide customers with appropriate services in order to affect their decisions. This study will be presented in the following orders. Primarily, the first part will demonstrate an exploratory analysis, which clarifies the influencing variable that strongly affect the cancellation of credit card service customers. Subsequently, the second part will discuss the feature engineering techniques and other related methodologies. Afterwards. the third act consists of applying the state-of-art machine learning scenario to find the best resources for building the model. Subsequently, after the completion of all steps, a machine learning model: XGBoost model will be developed, capable of predicting, based on the data of a system. Eventually, a brief summary will be given in the last section.

## 2. Data & Method

### 2.1 Data

In general, there are about 10,000 customers information collected in the data set, including the social demographics and credit relevant information (e.g., credit card limit). According to the analysis of simple calculation, only about 16.07% of customers have canceled, leading to certain difficulty to train the statistical model.

### 2.2 Exploratory data analysis (EDA)

In this part, it is aimed to discover the main elements that affect the predicted target. The data visualization and statistical descriptive analysis will be provided, along with some satisfactory insights.

#### 2.2.1 Correlation analysis

Spearman, a non-parametric statistical test, can measure statistical dependence between two variables [3-6]. The correlation analysis can help us to verify the significant variables related to the rate of customers who leave the credit card services. The spearman correlation equation is shown below:

$$r_R = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)} \quad (1)$$

Where n is the number of data points for the two variables and di is the difference in reach of the n element. The Spearman coefficient, ρ, can be calculated through the equation and the result will be a value between -1 and 1. The interpretation of Spearman's correlation coefficient is ρ=1 express the high positive correlation, ρ=0 represents no classification association and ρ=-1 means the perfect negative association between the two intervals.

The correlation coeffiients heatmap is shown in Fig. 1. The variables that demonstrate a considerable negative association in relation to the dependent attribute. A negative value of correaltion coefficients indicates that the features or factors have impacts on the customer's permanence. As for positive correlation coefficients, one has the variables Contacts_Count_12_mon and Months_inactive_112_mon.

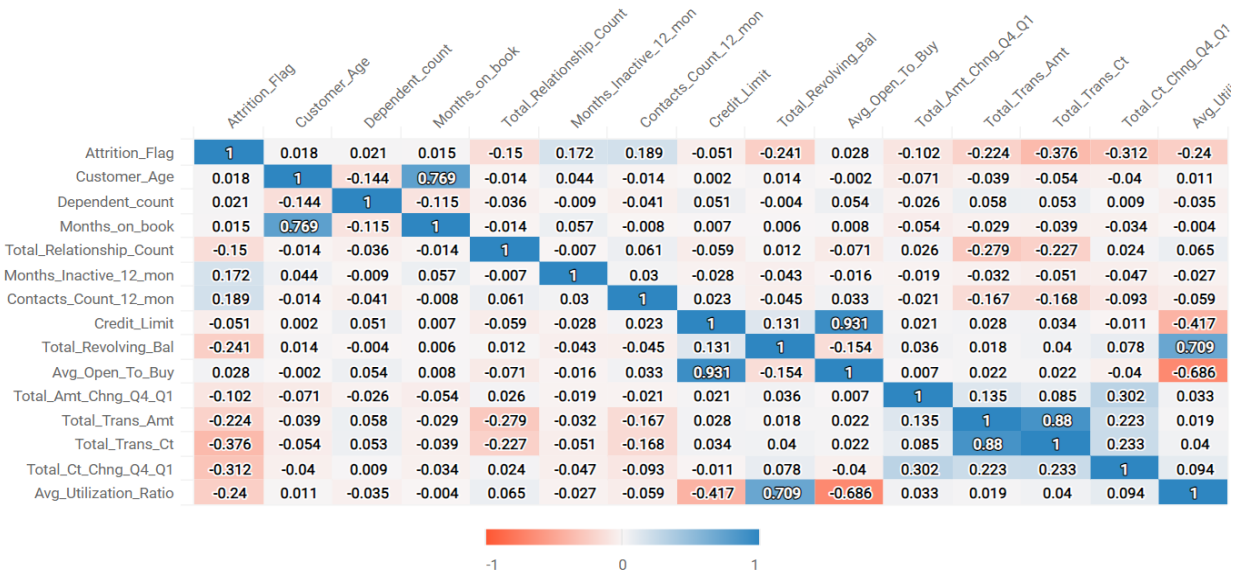


**Fig. 1** The Spearman's correlation matrix.

#### 2.2.2 Number of cancellations for each qualitative variable category

In general, the tree map graphs (seen from Fig. 2-Fig. 5) can help us to visualize hierarchical data using nested rectangles and better visualize the categorical variables in the dataset. Over a third of credit card cancellations (35.16%) are from people with an annual income less than 40k as shown in Fig. 2. It is interesting to find that majority of credit card cancellations (93.17%) are from customers who own the blue card and the lowest number of cancellations (0.19%) are from customers who have Platinum card type as shown in Fig. 3. Besides, 30.88% of credit card cancellations are from people who have graduated, and the lowest number of cancellations (4.45%) are from customers who have post Doctorate and 46,28% of credit card cancellations are from married people as shown in Figs. 4 and 5.

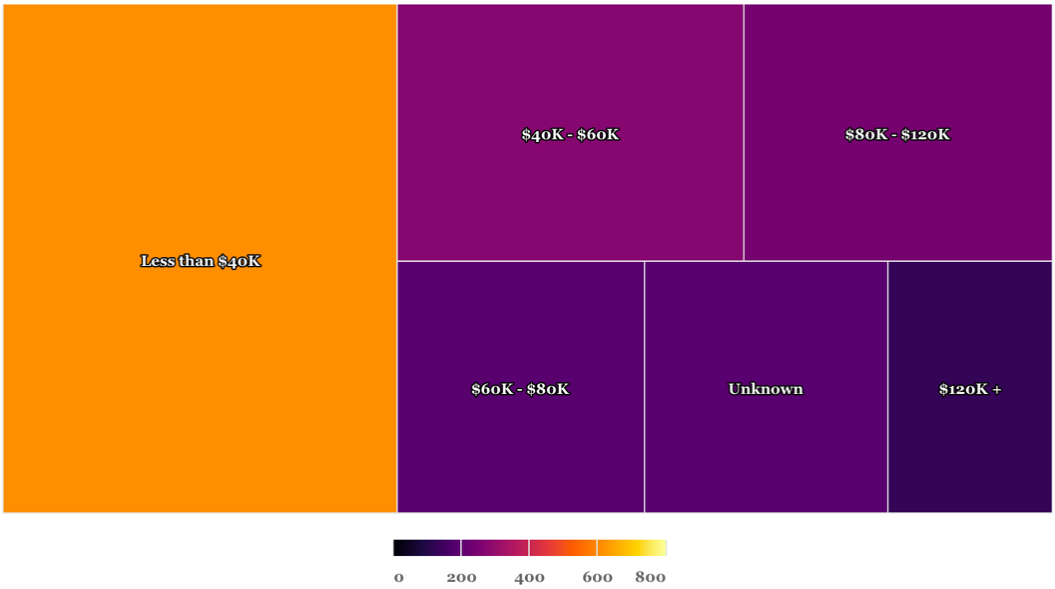


**Fig. 2** Different annual come levels.

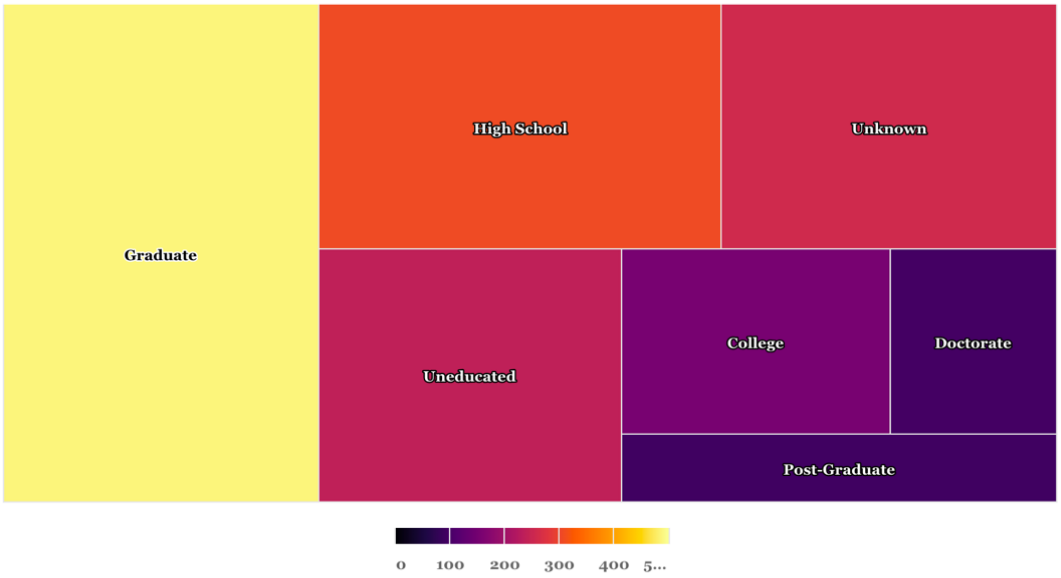


**Fig. 3** Different card types.

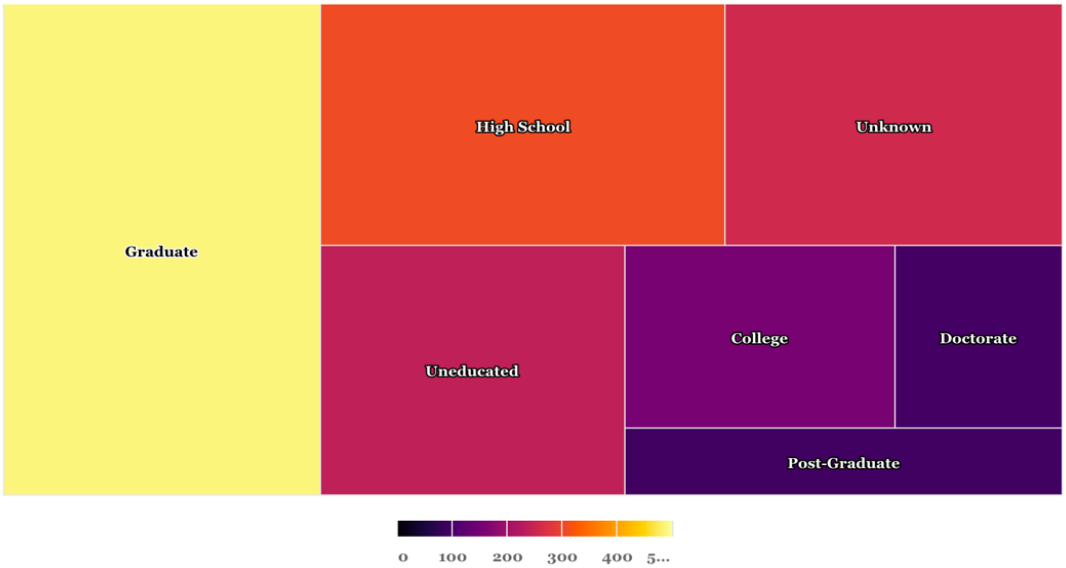


**Fig. 4** Number of cancellations by educational level

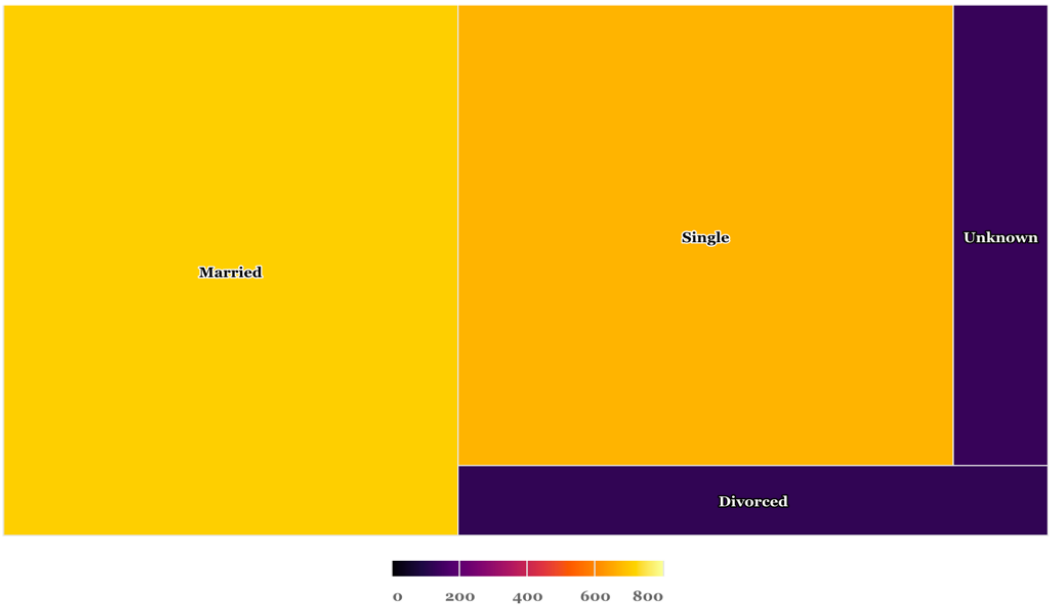


**Fig. 5** Distribution of number of cancellations by marital status.

### 2.2.3 Behavior of total transactions made by customers

As seen in Pearson's statistical test, where the variables 'Total_Trans_Ct' have a correlation coefficient of -0.376, 'Avg_Utilization_Ratio' with 0.24 and 'Total_Ct_Chng_Q4_Q1' with -0.312 as shown in Fig. 6, indicating that they all positively influence customers stay. This session whose name is 'Behavior of the total customer transaction' aims to understand the behavior of the quantitative variables listed above in relation to the target variable 'Attrition_Flag', which informs whether the customer has left the card service or not.

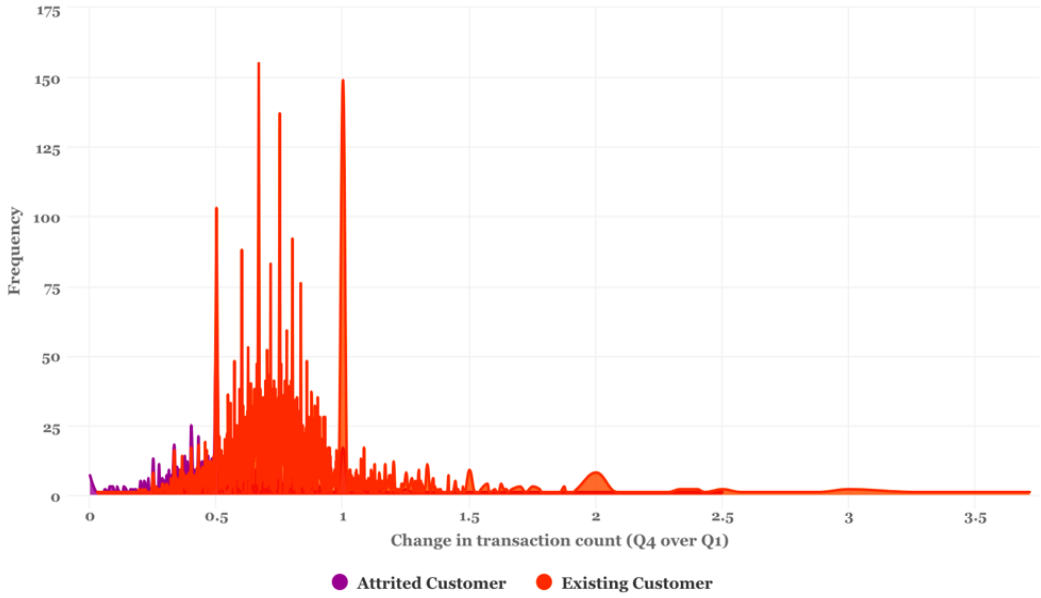


**Fig. 6** Changes in transaction count

### 2.2.4 Analysis of customer transactions

The Total_Trans_Ct variable shows an interesting behavior when it is observed its distribution in relation to customers who left or did not leave credit card services. To obtain more information about this variable and the way it relates to the other attributes of the data set, investigations will be carried out in this session accordingly.

Seen from Fig. 7, one can tell 50% of customers who left credit card services had a number of transactions in the last 12 months less than or equal to 43, remembering that the maximum number

of transactions in the last 12 months of customers who left the service is 72. While the median of the people who remained with the card services is 71 transactions. Besides, 75% of them had a number of transactions for the same lasting period equal to or less than 51. The third quartile of people who remained with card services is 82 transactions.

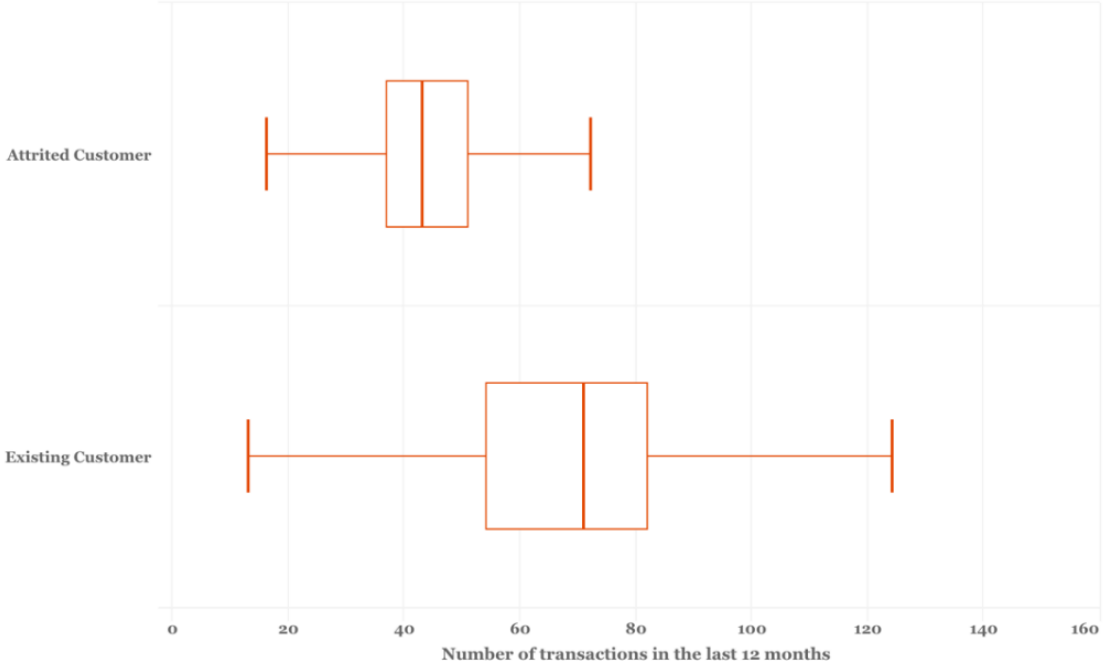


**Fig. 7** Total transactions in terms of percentages

### 2.3 Feature engineering

All the decisions made in the resource engineering process were the ones that were the best to prepare the data for the machine learning process. This study will use basic techniques including data banning to reduce the high number of unique values, one-hot coding for nominal qualitative variables.

#### 2.3.1 The group data into bins

The values of the datasets are separated to different bins to inhibit overfitting. To define the number of bins, the Sturges rule will be applied that allows us to create fixed amplitude classes from the following equation:

$$k = 1 + \frac{10}{3} log_{10} n \tag{2}$$

After converting the data into bins, the counting number is given in Fig. 8.

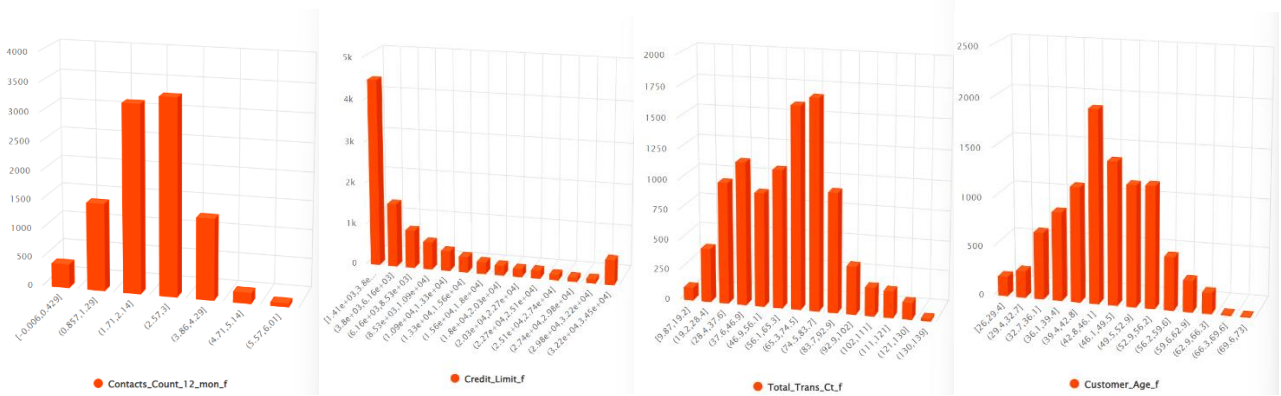


**Fig. 8** Categorical variable for contracts count, credit limit categories, total transaction count and customer age groups (from left to the right).

### 2.3.2 The Over-sampling

To correct the problem of unbalancing the classes of the data set, the SMOTE (Synthetic Minority Oversampling Technique) method will be utilized to enlarge the dataset preserving the feature of the class [7, 9-11]. The data after implementation of SMOTE is depicted in Fig. 9.

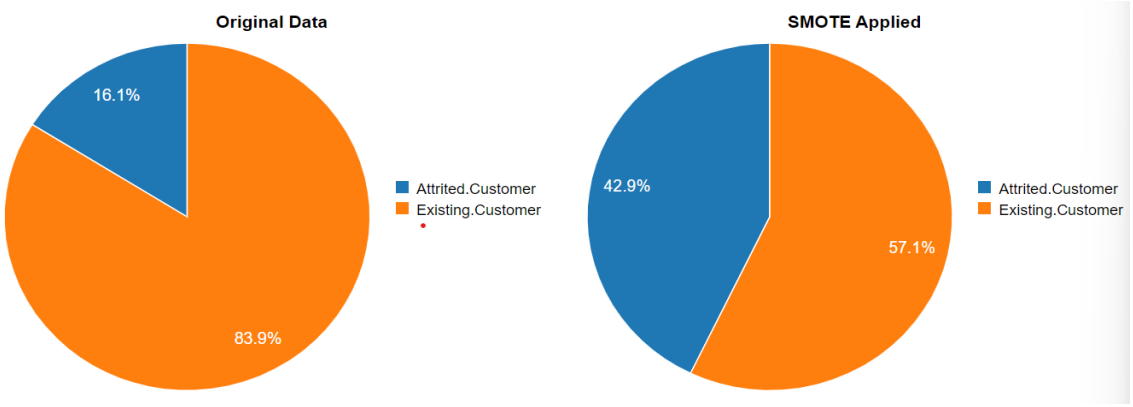


**Fig. 9** SMOTE applied data

## 3. Results & Discussion

Based on the feature selection, one obtained the list of all the significant variables with their importance. The variables are shown in Figure 10. The XGBoost is an algorithm based on gradient boosting decision trees. While in the model, a series of decision trees were employed. These trees learned from the prior tree in the last layer and then influences the following tree to increase the model performance. The advantage of XGBoost includes XGBoost is not affected by the multicollinearity as other machine learning models. XGBoost is first introduced by Chen in 2016. Chen and Guestrin made further improvements to traditional gradient boosting model [3].

One of the most important changes they made is the regularization of loss function. The regularized objective $L_k$ for the $k^{th}$ iteration is show in the equation below

$$L_k = \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}_k^{(i)}\right) + \sum_{j=1}^{k} \Omega(f_j) \tag{3}$$

Where n is the number of samples, $l$ is the original loss function and $\hat{y}_k^{(i)}$ is the predicted value of $i$ at $k$th iteration. Here, $\Omega$ is the regularization term as shown below.

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Here, T is the number of leaf nodes, $\Upsilon$ is a constant and $\lambda$ is the degree of regularization.

Since an additive learning approach is applied in XGBoost to combine the most popular tree model $f_k(x^i)$ into the classification model to predict. The equation could further be expressed as

$$L_k = \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}_{k-1}^{(i)} + f_k(x^i)\right) + \Omega(f_k) + \sum_{j=1}^{k-1} \Omega(f_j) \tag{5}$$

Moreover, second-order Taylor expansion to the objective function is also employed in XGBoost. Thus, the objective function is

$$L_k = \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}_{k-1}^{(i)} + g_i * f_k(x^i) + \frac{1}{2} h_i * f_k(x^i)\right) + \Omega(f_k) + C \tag{6}$$

Here, $g_i$ and $h_i$ are the first and second derivatives of the loss function, respectively, and C is constant in the equation. The confusion matrix is shown in Fig. 11. The model has accuracy around 96.5% in prediction. Thus, using this proposed model, one can predict the customers' cancellation ahead of time. In addition, the identified features can be utilized for the company to improve the costumers' satisfaction and reduce the rate of cancellations.

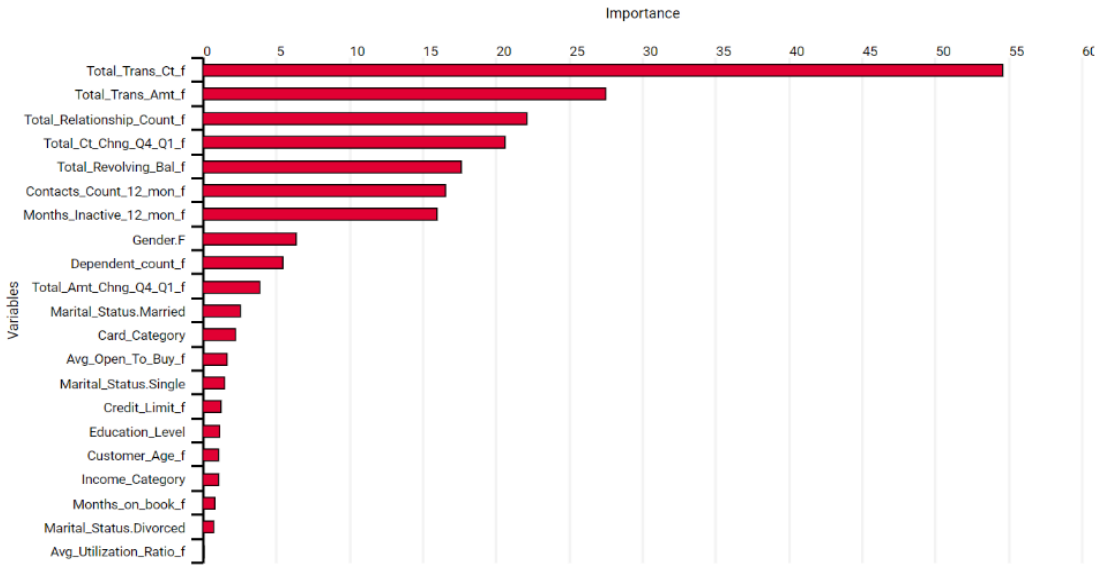


**Fig. 10** The importance of the Features.

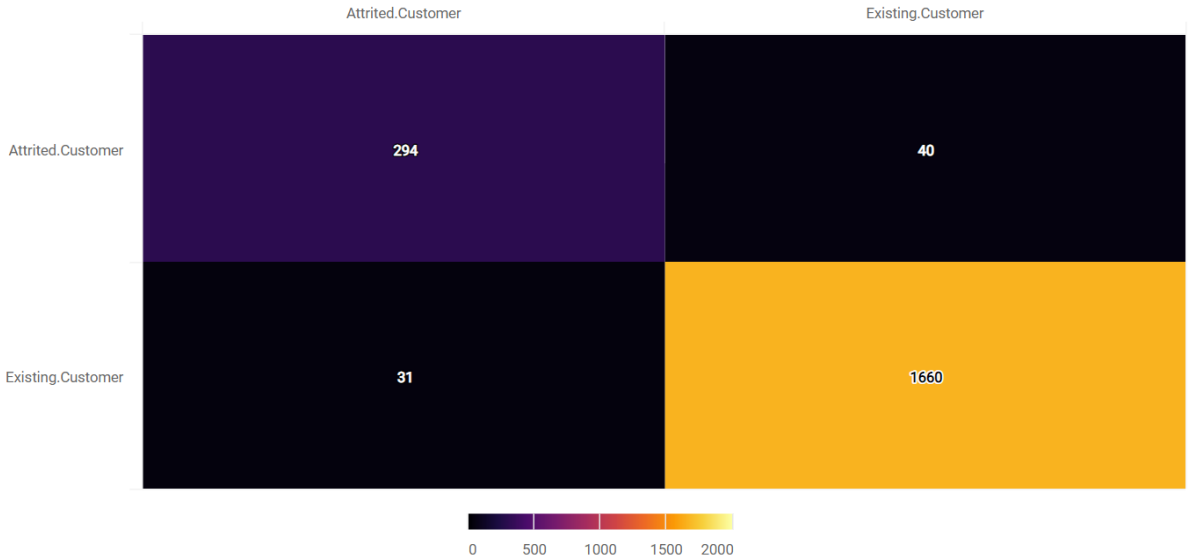


**Fig. 11** The confusion matrix.

Nevertheless, this paper has some drawbacks and shortcomings. Primarily, the paper found some significant factors related to customers' cancellations in AMEX, e.g., total transaction count and card type. In other words, customers who own premium cards are less likely to cancel their cards, which indicates that the company can improve their services to customers owning other types of cards. In addition, the study proposed a prediction model which can tell customers willing to cancel in advance. This prediction model can help the company the identify those customers and improve their services accordingly. Therefore, this model can save the rate of leaving customers for AMEX company. However, the paper used the past 12-month data to predict the customer behavior. This may cause some inaccurate prediction since the customer behavior changes through time. It will be better if the model can be real-time or near-real-time. In this way, the model can take the actual customer data to

predict in time and the service team could take strategies at the right time. Moreover, XGBoost is a robust model with high accuracy, whereas it could be better if some other state-of-art machine learning model results could be provided for comparison and external validation. In the future, more real-time data and models should be employed in customer behavior analysis and prediction to better help the service industry to improve their customers' satisfaction level.

## 4. Summary

In conclusion, this paper selects AMEX credit card service data to predict the customer cancellations using the latest 12-month data based on XGBoost. Specifically, card type, total transaction count and card usage fee are carried out. According to the analysis, the proposed XGBoost model can predict the existing customer to leave ahead with accuracy over 95%. Besides, the descriptive statistics found that customers who own premium cards are less likely to cancel their cards, which indicates that the company can improve their services to customers owning other types of cards. Nevertheless, it should be noted that the temporal and spatial factors should be considered. In other words, some near-real-time models should be used for further studies in customer behavior prediction. In the future, more real-time data and models should be employed in customer behavior analysis and prediction to better help the service industry to improve their customers' satisfaction level. Overall, these results offer a guideline for machine learning prediction model application in customer behavior analysis to improve their satisfaction levels.

## References

[1] D. W. Carlton, The anticompetitive effects of vertical most-favored-nation restraints and the Error of AMEX, Colum. Bus. L. Rev., p. 93, 2019.

[2] M. Dumiak, Advertising Campaigns: Amex Unrivaled in Advertising Spending: The three top spenders in 1999 were credit card companies, but it was also the year of the dot-com effect, Financ. Serv. Mark., vol. 2, no. 4, p. 8, 2000.

[3] D. G. Bonett and T. A. Wright, Sample size requirements for estimating Pearson, Kendall and Spearman correlations, Psychometrika, vol. 65, no. 1, pp. 23–28, 2000.

[4] Z. Wang et al., Temporospatial variations and Spearman correlation analysis of ozone concentrations to nitrogen dioxide, sulfur dioxide, particulate matters and carbon monoxide in ambient air, China, Atmos. Pollut. Res., vol. 10, no. 4, pp. 1203–1210, 2019.

[5] J. O. May and S. W. Looney, "Sample size charts for Spearman and Kendall coefficients," J. Biom. Biostat., vol. 11, no. 6, pp. 1–7, 2020.

[6] D. J. Bartholomew, Spearman and the origin and development of factor analysis, Br. J. Math. Stat. Psychol., vol. 48, no. 2, pp. 211–220, 1995.

[7] Y. Peng, C. Li, K. Wang, Z. Gao, and R. Yu, Examining imbalanced classification algorithms in predicting real-time traffic crash risk, Accid. Anal. Prev., vol. 144, no. March, 2020.

[8] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[9] J. Wang, M. Xu, H. Wang, and J. Zhang, Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding, in 2006 8th international Conference on Signal Processing, 2006, vol. 3.

[10] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "Smote for regression," in Portuguese conference on artificial intelligence, 2013, pp. 378–389.

[11] L. Camacho, G. Douzas, and F. Bacao, Geometric SMOTE for regression, Expert Syst. Appl., p. 116387, 2022.

[12] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.