# Predicting Credit Card Churn with Logistic Regression

*by Hieu Phi Nguyen, John Vishwas Nelson, Seif Abuhashish, Mariana Tollko, and Haodong Wu*

**Abstract** Because of the continuously changing features of customers, today's market encounters short-lived concepts and strategies. Churn prediction is becoming a key concern for banks looking to retain clients by meeting their demands while working within limited resources. Real-world data still has numerous critical but difficult challenges that might destroy the effectiveness of churn prediction. The issues encompass data distribution imbalance, outliers, noise, multicollinearity, and the curse of dimensionality. This study aims to address these issues by answering the following research questions: (i) which determinants influence customer decisions to stay or leave credit card services, (ii) which approach helps the most to improve the churn prediction model for current bank data. To address practical difficulties, we present Logistic Regression and show its application to churn prediction with the improvement of three data processing approaches, namely Synthetic Minority Oversampling Technique (SMOTE), Weight of Evidence (WoE), and Principal Component Analysis (PCA). It has been discovered that I demographic information, account information, and transaction behavior all have an influence on bank customers' churn decisions, and (ii) SMOTE methods enhance Logistic Regression the best. Other invaluable insights have been gleaned from the business perspectives.

## Executive Summary

## Introduction and Literature review

Customer relationship management is a collection of procedures and enabling systems that help a firm create long-term, lucrative connections with particular customers, as defined by Ling and Yen (2001). Furthermore, the fast expansion of the Internet and its related technologies has considerably boosted marketing options and altered the way firms and their consumers manage their interactions, see Ngai (2005). Customer relationship management is defined by Ngai et al. (2009) as assisting businesses in better discriminating and allocating resources to the most lucrative set of customers through the cycle of customer identification, customer acquisition, customer retention, and customer development. The primary issue here is customer retention, which is strongly weighted on customer satisfaction, which refers to a comparison of consumers' expectations with their sense of being satisfied, see Kracklauer et al. (2004). As a result, aspects of customer retention include loyalty programs, which comprise campaigns or supporting activities aimed at sustaining a long-term connection with customers, as shown in Ngai et al. (2009). Churn analysis, in particular, is a component of loyalty programs.

Customer churn, defined as the proclivity of consumers to discontinue doing business with a firm in a particular time period, has become a serious issue and is one of the primary issues that many organizations across the world are facing, see Laha. Many businesses are resorting to data mining strategies to retain existing consumers in order to compete in an increasingly competitive industry. According to Nie et al. (2011), increasing the retention rate by up to 5% may improve a bank's earnings by up to 85%. Furthermore, client retention is regarded as more essential than in the past. This survey aims to uncover common churned customer attributes in order to develop a customer churn prediction model. A variety of research utilizing various algorithms, such as logistic regression (Pen, 2014), (Jain et al., 2020), (Jain et al.), (Dijendra and Sisodia), sequential patterns (Thi, 2019), (Culbert et al.), (Stojanovski), genetic modeling (Faris et al., 2014), (Abbasimehr and Alizadeh, 2013), (Stripling et al., 2018), classification trees (Höppner et al., 2020), (Dorokhov et al., 2020), (Ahmad et al., 2019), neural networks (Khan et al., 2019), (Brandusoiu and Toderean, 2020), (Saghir et al.), and support vector machine (Rodan et al., 2014), (Xia and Jin, 2008), (Li and Xia), have been conducted to investigate customer turnover and illustrate the potential of data mining.

However, due to the unique nature of the churn prediction problem, existing churn analysis algorithms still have certain limitations. According to Xie et al. (2009), there are three key characteristics: (1) The data is typically unbalanced; that is, the number of turnover consumers accounts for just a relatively tiny proportion of the data (usually 2% of the total samples) (Zhao et al.); (2) Large learning applications will invariably generate some form of noise in the data. (Shah, 1996); and (3) Predicting turnover necessitates rating subscribers based on their probability to churn (Au et al., 2003). Furthermore, according to Manďák and Hančlová (2019), those in charge of churn management should consider the following three dimensions: *who* or which customers are likely to churn, *when* will the customers churn, and *why* will they decide to leave.

To solve this issue, several techniques have been offered. Although decision-tree-based algorithms

may be extended to calculate ranking, certain leaves in a decision tree may have identical class probabilities, making the technique susceptible to noise. The neural network method does not communicate the discovered patterns directly in a symbolic, easily comprehensible manner. Although genetic algorithms can generate accurate predictive models, they cannot assess the likelihood of their predictions. These issues make the aforementioned approaches inapplicable to the churn prediction problem (Au et al., 2003). Other approaches, such as the Bayesian multi-net classifier (Luo and Mu, 2004), support vector machine, sequential patterns, and survival analysis (Larivière and Van den Poel, 2004), have made reasonable attempts to forecast churn, but their error rates remain unacceptable. Last but not least, while the aforementioned techniques can identify which customers are likely to churn, they do not address the causes for customer turnover. Dahiya and Talwar (2015) emphasizes that machine learning models perform effectively provided sufficient effort is spent preparing useful features. As a result, possessing the proper characteristics is typically the most crucial factor. Banking firms now have access to a variety of data sources, which can be useful for analyzing client turnover, thanks to the continuing decline in the cost of data storage. As a result, it is important to devote time in feature engineering, because well-prepared features may also assist us in identifying the causes of churn.

In this work, we propose the logistic regression technique in response to the limitations of current methods. It belongs to a class of models known as generalized linear models (Mount and Zumel, 2019). The goal of generalized linear models for binary dependent variables is to estimate a regression equation that connects the predicted value of the dependent variable $\mathbf{y}$ to one or more predictor variables, denoted by $\mathbf{x}$ (Heeringa et al., 2017). With such features, logistic regression has the potential to provide explanations for why clients opt to discontinue using financial services. Although numerous implementations of logistic regression in a customer churn environment have been published to the best of our knowledge, our study adds to the existing literature not only by investigating determinants in predicting customer churn in the banking domain, but also by incorporating sampling techniques and other data processing techniques into logistic regression to achieve better performance.

In addition to the imbalance in distribution, real-life data is dealing with certain significant concerns in the banking industry. With so many statistical methodologies at the analyst's disposal, an ideal method for doing churn analysis is dependent on need and context. A research by SEbASTIAN and Wagh (2017) in the telecoms business was beneficial with very few parameters. However, when it comes to client bases for financial businesses, there are many more policies and characteristics that will play into the churn study, raising the possibility of the curse of dimensionality and multicollinearity in the data.

Furthermore, due to outliers, noise, and non-linear connections, the underlying distribution of many real-world datasets is uncertain or complicated. An outlier is a data point that differs from the rest of the data, as defined by Barnett and Lewis (1984). Whereas noise is defined as mislabeled instances (class noise) or mistakes in attribute values (attribute noise) (Salgado et al., 2016), an outlier is a wider notion that encompasses not only errors but also discordant data that may result from natural variance within the population or process. Outliers, as a result, frequently include fascinating and helpful information about the underlying data. As a result of removing outliers, banks may experience some information loss in their customer behavior.

The rest of this paper is organized as follows. The next part explains the methodological underpinnings of logistic regression as well as data processing techniques, followed by dataset preparation and exploratory analysis, and finally the findings and discussion. In the final part, some concluding thoughts and recommendations for future work are included.

## Methodologies

In this part, we explain the technique's methodological foundations as well as the assessment criteria we utilize to assess the method's performance.

### Logistic regression

The primary goal of regression analysis approaches is to study and evaluate the connections between a collection of features. A simplistic method would be to represent $\mathbf{y}$ as a linear function of $\mathbf{x}$, however linear regression does not capture the connection between $\mathbf{y}$ and $\mathbf{x}$ in the binary extent, and it may also yield predictions that are beyond the allowed range 0-1. A better alternative is a nonlinear function, which produces a regression model with linear coefficients and allows the predicted values to be transformed to the range 0-1. These functions are known as link functions in the vocabulary of generalized linear models (Heeringa et al., 2017). The logit and the probit are the two most popular link functions used to model binary survey data. A linear regression model may be used to model the

logit, or natural logarithm of the odds:

$$\log \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{1}$$

where $\pi(x)$ is the conditional probability that $y = 1$ given the covariate vector $\mathbf{x}$, the $\beta$s are estimated regression coefficients of the logit model. The left-hand side of the Equation (1) is called the *log-odds* or *logit* and can take values from the interval $(-\infty, \infty)$. The term inside the brackets is called the *odds* and can take on any value between 0 and $\infty$. Values close to 0 indicate very low and values close to $\infty$ indicate very high probability.

Following the estimate of the model coefficients, it is common practice to assess the importance of the explanatory variables (Hosmer Jr et al., 2013). *Wald test* can be used to test the statistical significance of the coefficients $\beta$ in the model. Wald test calculates a $Z$ statistic (2), which is for $i$-th variable computed as:

$$Z = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \tag{2}$$

where $\hat{SE}(\hat{\beta}_i)$ is an estimated standard error of the estimated regression coefficient $\hat{\beta}_i$. The $Z$ value is then squared to get a Wald statistic with a chi-square distribution. Maximum likelihood is the classic technique of estimate that leads to the least squares function under the linear regression model and serves as the foundation for estimating the parameters of a logistic regression model. Instead of a novel technique, this study is connected to knowledge discovery based on logistic regression. Other studies (Hosmer, 1989) go into further depth on fitting the logistic regression.

To determine the power of various variables, we construct a model with various variable combinations. Because real-world data typically contains hundreds of variables, including all of them in the algorithms would cause the model to be misled. Furthermore, if we create the model with all of the variables, the cost will be rather high when the model is utilized since it will be time-consuming to calculate all of the variables. As a result, a stepwise method is utilized to choose variables throughout the model-building process.

### Synthetic Minority Oversampling Technique

Standard machine learning techniques for classification problems presume that the proportion of various classes in the real dataset is about equal. However, in many practical applications, classes of the response variable are not provided in an equal proportion, particularly minority situations, which are under-represented. One potential consequence of such under-representation is that accuracy, a frequently used statistic to evaluate model performance, may mislead outcomes on the unbalanced data, and more suitable metrics, like as precision and recall, are generated from the confusion matrix. There are two techniques for this: (i) undersampling and (ii) oversampling. Oversampling techniques are often favored over undersampling approaches. The reason for this is because when we undersample data, we tend to eliminate occurrences that may contain valuable information. Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique that generates synthetic samples for the minority class, as described by Chawla et al. (2002). This approach aids in overcoming the overfitting issue caused by random oversampling. To summarize, SMOTE first chooses a minority instance at random and then determines the $k$-nearest neighbors of that instance. The SMOTE samples are linear combinations of two similar samples from the minority class ($\mathbf{x}$ and $\mathbf{x}^R$) and are defined as

$$\mathbf{s} = \mathbf{x} + u \cdot (\mathbf{x}^R - \mathbf{x}) \tag{3}$$

where $0 \leq u \leq 1$; $\mathbf{x}^R$ is randomly chosen among the $k$ minority class nearest neighbors of $\mathbf{x}$. In other words, a synthetic instance is produced at a random point on the line linking the instance and one random instance from the feature space's $k$ neighbors. Chawla et al. (2002) has further information on this method.

### Weight of Evidence

Weight of evidence (WoE) is a phrase that appears often in the published scientific and policy-making literature, most notably in the context of risk assessment (see Weed (2005)). Its definition, however, is ambiguous. Decisions are primarily based on evaluating the likelihood of a single event occurring. The difficulty of decision making ranges from basic to complicated, requiring more involved processing of input from many sources. The outcome of this probabilistic decision making is dependent on facts that may or may not be interdependent, as Chater and Oaksford (2008) demonstrates. For each choice,

we must assess the importance of the facts that impact it. This allows us to map the risk associated with a certain decision or information on a linear scale.

Good (1950) was the first to develop the notion of quantitatively weighing evidence. WoE of the $i$-th value of the feature $A$ is defined as follows

$$WoE_i^A = \log\left(\frac{\frac{N_i^A}{SN}}{\frac{P_i^A}{SP}}\right) = \log\frac{N_i^A}{SN} - \log\frac{P_i^A}{SP} \tag{4}$$

where $N_i^A$ is the number of data points that were labeled as negative, and $P_i^A$ is the number of data points that were labeled as positive for the $i$-th value of the feature $A$. $SN$ is the total number of negatively labeled data points, $PN$ is the total number of positively labeled data points in the set.

As can be seen from Equation 4, WoE is made up of two parts: a variable component and a constant component. These numbers are unrelated to the machine learning method that will be used in the data mining process. They are computed at the pre-processing stage. The variable component is generated using data points with a specific value of feature $A$, whereas the constant component is derived using the whole sample.

The WoE framework is based on the following relationship in the extent of binary outcome:

$$\log\frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \log\frac{P(y=1)}{P(y=0)} + \log\frac{f(\mathbf{x}|y=1)}{f(\mathbf{x}|y=0)} \tag{5}$$

where $f(\mathbf{x}|y=1)$ and $f(\mathbf{x}|y=0)$ denote the conditional probability density function (or a discrete probability distribution if $\mathbf{x}$ is categorical). The first term is called sample log-odds whereas the latter is WoE. The equation 4 in fact is discrete version of WoE. Equation 5 implies that WoE and the conditional log odds of $\mathbf{y}$ are perfectly correlated since the *intercept* is constant. Hence, the greater the value of WOE, the higher the chance of observing $y = 1$. Indeed, as WoE is positive the chance of observing $y = 1$ is above average (for the sample), and vice versa when WOE is negative. As WoE equals to zero, the odds are simply equal to the sample average.

Information Value (IV) is a term closely related to WoE that refers to the value of information or information. See Chen et al. for further information on IV, which is an essential metric for determining the degree of effect of independent factors on target variables. The formula is as follows:

$$\begin{aligned} IV^A &= \int \log\frac{f(\mathbf{x}|y=1)}{f(\mathbf{x}|y=0)}(f(\mathbf{x}|y=1) - f(\mathbf{x}|y=0))dx \\ &= \sum_{i=1}^{p}\left(\frac{N_i^A}{SN} - \frac{P_i^A}{SP}\right)\cdot WoE_i^A \end{aligned} \tag{6}$$

if the feature $A$ has $p$ categories.

Equation 6 demonstrates that the IV is basically a weighted total of all the individual WoE values, with the weights taking into account the absolute difference between the numerator and the denominator (WoE captures the relative difference). Generally, the variable with (i) $IV < 0.02$ has very little predictive power and will not add any meaningful predictive power to your model, (ii) $0.02 \leq IV < 0.1$ has a weak predictive power, (iii) $0.1 \leq IV < 0.3$ has a medium predictive power, (iv) $IV \geq 0.3$ has a strong predictive power, (v) $IV \geq 0.5$ is suspicious for over-predicting.

### Principal Component Analysis

Normally, realistic data contains both numerical and category characteristics.Multiple factor analysis (Escofier and Pages, 1994), (Abdi et al., 2013) is used with multi-table data in which the type of variables vary from one data table to the next but the variables must be of the same kind within a given data table. As a result, multivariate analysis of mixed data with observations characterized by a combination of numerical and categorical variables becomes necessary. Based on a Generalized Singular Value Decomposition (GSVD) of pre-processed data, Chavent et al. (2014) has developed principal component analysis (PCA) techniques for dealing with a mixture of numerical and categorical variables. As special instances, this method incorporates naturally standard principal component analysis and standard multiple correspondence analysis. The algorithms are described in detail in Chavent et al. (2017).

Multicollinearity of independent variables is common in real-world data. It has a detrimental impact on the performance of regression and classification models. PCA utilizes multicollinearity to integrate highly correlated data into a collection of uncorrelated variables. As a result, PCA may effectively remove feature multicollinearity. Furthermore, PCA employs a mathematical technique to

calculate a reduced number of new variables known as principle components (PCs), which are linear functions of the original dataset. As a result, PCA reduces the dimensionality of a big dataset while retaining as much statistical information as feasible.

### Evaluation criteria

Marketers will utilize these categorization models to anticipate future customer behaviour after developing a predictive model. It is critical to assess the performance of the classifiers. The receiver operating curve (ROC) is commonly employed as a criterion. The ROC is a graphical representation of the sensitivity (number of true positives vs total number of events) and 1−specificity (number of true negatives versus total number of non-events). The ROC may also be depicted by graphing the percentage of true positives vs the percentage of false positives, as shown by Coussement and Van den Poel (2008). Area under the ROC curve is the best summary number for this ROC curve (AUC). The AUC evaluates a classifier's behavior while ignoring class distribution, classification cutoff, and misclassification costs.

Precision and recall are more acceptable techniques for evaluating model performance in the setting of under-representation of minority situations. Precision is intuitively a measure of correctness (i.e., how many positively labeled examples are genuinely positive), whereas recall (or sensitivity) is a measure of completeness or accuracy of positive instances (i.e., how many instances of the positive class are labeled correctly/positively). Furthermore, the two types of mistakes are studied: Type I error, which occurs when a customer who did not churn is misclassified as a churner (False Positive), and Type II error, which occurs when a customer who churned is misclassified as an un-churner (False Negative). The loss produced by Type II mistake is typically thought to be 5–20 times more than the loss caused by Type I error, according to Lee et al. (2006). The bank's goal is to find all potential clients who want to cancel their Credit Card Services. Predicting that consumers would not terminate their Card Services but will instead attrit will result in a loss. As a result, the False Negative values must be decreased. In other words, the Recall must be increased to reduce the likelihood of False Negatives.

## Result and Discussion

### Data

The dataset in this study of this study is from Kaggle[1]. The dataset contains customers' statistical data with over 10000 observations including 19 explanatory features. 16% of the observations have the target variable *attrited* and 84% observations have the value *existing*, implying a severe imbalanced problem. All of the data is integrated at the level of the customer. There are no inaccurate data, e.g. customer age are ranged from 26 years old to 73 years old, which is legal in several countries. Also, we have 6 categorical variables.

We eliminate descriptors that clearly have nothing to do with the prediction, such as a driver's license number. We investigate three primary descriptor groups that include possible explanatory descriptors from our input. Personal demographics, account level, and consumer behavior are the three categories. They are labeled as follows. Age, gender, number of dependents, education, marital status, and income are all examples of personal demographics. Credit card type, years with bank, amount of services utilized by bank, inactivity period, and credit limit are all considered at the account level. Customer behavior includes revolving balance, available credit, credit usage percentage, transaction amount, and transaction frequency. There are no errors in the statistics, for example, client ages range from 26 to 73, which is allowed in various nations.

Interestingly, Figure 1 notes that the among categorical variables the percentage of Attrited Customers seems to be fairly equal across all categories of all the Variables. Despite having a large imbalance in the proportions across the categories, the attrition however is quite similar. There seems to be no significant categorical variable that shows a strong indicator for Attrition. A more formal test, Chi-squared test, can be employed to verify this assertion. In this test, $H_0$ refers to the two variables are independent and $H_1$ is the two variables relate to each other. Table 1 shows that with p-value less than 0.05, we reject the $H_0$ for Gender and Income_Category.

Table 2 and Figure 2 show several Normal-distributed variables with bell shape, skewness of zero and excess kurtosis of zero, namely Customer_Age, Dependent_count, Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon, Total_Revolving_Bal, Total_Trans_Ct and Avg_Utilization_Ratio. We employ a formal test for normality, Jarque-Bera test

[1]Retrieved from https://www.kaggle.com/sakshigoyal7/credit-card-customers


**Figure 1:** Categorical features

**Table 1:** Chi-squared test

| Variable | ChiSq | DF | PVal |
|---|---|---|---|
| Gender | 13.866 | 1 | 0.0002 |
| Education_Level | 12.511 | 6 | 0.051 |
| Marital_Status | 6.056 | 3 | 0.109 |
| Income_Category | 12.832 | 5 | 0.025 |
| Card_Category | 2.234 | 3 | 0.525 |

based on skewness and kurtosis that matches a normal distribution, that is, $H_0$ refers to the hypothesis that data are normally distributed, see Table 3. As we can see, only Contacts_Count_12_mon is failed to reject the $H_0$ and therefore, only Contacts_Count_12_mon follows the Normal distribution. Aslo, several variables show remarkably skewed to the right in their distributions, including Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy, Total_Trans_Amt, Total_Ct_Chng_Q4_Q1, and Avg_Utilization_Ratio. Furthermore, there are several variables having extreme values based on IQR rule, such as Customer_Age, Months_on_book, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, and Total_Ct_Chng_Q4_Q1. This explains partly about the right skewness in aforementioned variables, that is, there are extreme values in the right tail of their distributions.

**Table 2:** Continuous variable summary

| | mean | sd | median | min | max | skew | kurtosis |
|---|---|---|---|---|---|---|---|
| Customer_Age | 46.326 | 8.017 | 46 | 26 | 73 | -0.034 | -0.290 |
| Dependent_count | 2.346 | 1.299 | 2 | 0 | 5 | -0.021 | -0.684 |
| Months_on_book | 35.928 | 7.986 | 36 | 13 | 56 | -0.107 | 0.399 |
| Total_Relationship_Count | 3.813 | 1.554 | 4 | 1 | 6 | -0.162 | -1.007 |
| Months_Inactive_12_mon | 2.341 | 1.011 | 2 | 0 | 6 | 0.633 | 1.097 |
| Contacts_Count_12_mon | 2.455 | 1.106 | 2 | 0 | 6 | 0.011 | -0.0003 |
| Credit_Limit | 8,631.954 | 9,088.777 | 4,549 | 1,438.300 | 34,516 | 1.666 | 1.807 |
| Total_Revolving_Bal | 1,162.814 | 814.987 | 1,276 | 0 | 2,517 | -0.149 | -1.146 |
| Avg_Open_To_Buy | 7,469.140 | 9,090.685 | 3,474 | 3 | 34,516 | 1.661 | 1.796 |
| Total_Amt_Chng_Q4_Q1 | 0.760 | 0.219 | 0.736 | 0 | 3.397 | 1.732 | 9.985 |
| Total_Trans_Amt | 4,404.086 | 3,397.129 | 3,899 | 510 | 18,484 | 2.040 | 3.890 |
| Total_Trans_Ct | 64.859 | 23.473 | 67 | 10 | 139 | 0.154 | -0.368 |
| Total_Ct_Chng_Q4_Q1 | 0.712 | 0.238 | 0.702 | 0 | 3.714 | 2.063 | 15.677 |
| Avg_Utilization_Ratio | 0.275 | 0.276 | 0.176 | 0 | 0.999 | 0.718 | -0.796 |

Figure 3 shows numerous insights of our dataset. A correlation matrix is located above the diagonal. Several associations need to be highlighted between continuous variables: Customer_Age and Months_on_book (0.79 Pearson correlation), Total_Relationship_Count and Total_Trans_Amt (-0.35), Total_Relationship_Count and Total_Trans_Ct (-0.24), Total_Trans_Amt and Total_Trans_Ct

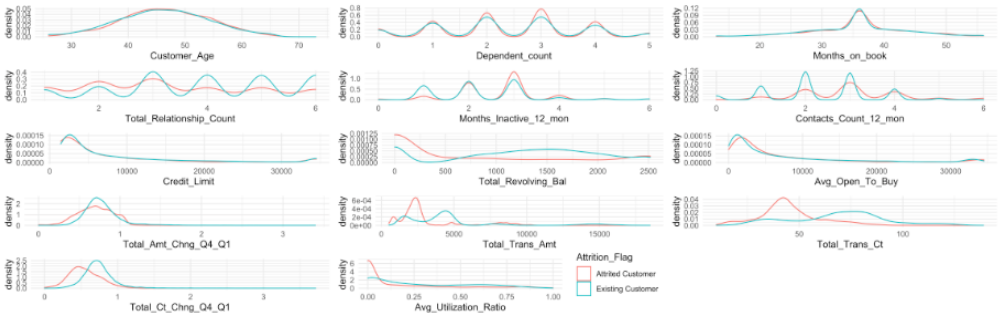


**Figure 2:** Numeric features

**Table 3:** Jarque-Bera test

| Variable | ChiSq | DF | PVal |
|---|---|---|---|
| Customer_Age | 37.165 | 2 | 0 |
| Dependent_count | 197.727 | 2 | 0 |
| Months_on_book | 86.442 | 2 | 0 |
| Total_Relationship_Count | 471.759 | 2 | 0 |
| Months_Inactive_12_mon | 1,184.374 | 2 | 0 |
| Contacts_Count_12_mon | 0.204 | 2 | 0.903 |
| Credit_Limit | 6,065.937 | 2 | 0 |
| Total_Revolving_Bal | 591.561 | 2 | 0 |
| Avg_Open_To_Buy | 6,021.924 | 2 | 0 |
| Total_Amt_Chng_Q4_Q1 | 47,156.490 | 2 | 0 |
| Total_Trans_Amt | 13,418.990 | 2 | 0 |
| Total_Trans_Ct | 96.858 | 2 | 0 |
| Total_Ct_Chng_Q4_Q1 | 110,944.700 | 2 | 0 |
| Avg_Utilization_Ratio | 1,136.684 | 2 | 0 |

(0.81), Credit_Limit and Avg_Utilization_Ratio (-0.48), Avg_Open_To_Buy and Avg_Utilization_Ratio (-0.54), Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1 (0.38). This seems logical when a customer who has tight relationship with a bank and possesses a small number of products of that bank usually transacts more usually and hence, increasing her transaction amount and the number of transactions. The correlation between Credit_Limit and Avg_Open_To_Buy is close to 1 (0.99). This means that these two features are strongly correlated. Usually we can remove one of them and keep the feature with the larger Pearson coefficient of the target value.

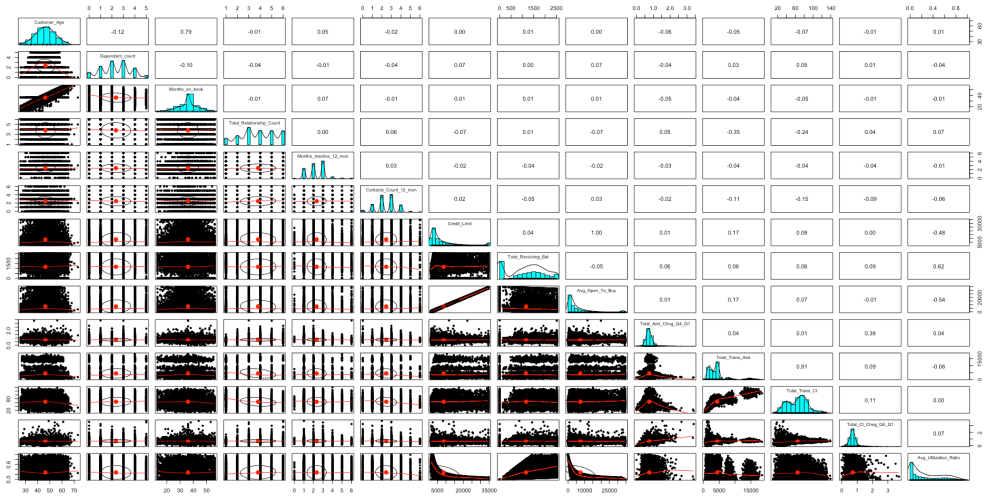

**Figure 3:** Numeric pairs of features

Figure 3 also shows much weak linear connection between continuous variables in dataset. Another highlight can be taken from the figure: age and number of dependent persons curve is an upside-down U that peaks around middle age, meaning that the sample's oldest and youngest customers have fewer dependent persons than those in their forties and fifties. Because this tendency is non-linear, the correlations alone could not have led to this conclusion. Similarly, there are other non-linear relationships despite small correlation, such as Total_Amt_Chng_Q4_Q1 and

Total_Trans_Amt, Total_Amt_Chng_Q4_Q1 and Total_Trans_Ct. We next to use the ANOVA test to consider the discrimination power of each or continuous variables in terms of Attrition_Flag. The null is that the means of the different groups are the same. Table 4 reports that the null is rejected at 0.05 significance level for Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Total_Revolving_Bal, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng_Q4_Q1, Avg_Utilization_Ratio. Due to trivial discrimination power of Avg_Open_To_Buy, we exclude it and at the same time solving the problem perfect correlation with Credit_Limit which significantly contributes to discriminating Attrition_Flag.

**Table 4:** ANOVA test

| Variable | F | PVal |
|---|---|---|
| Customer_Age | 3.356 | 0.067 |
| Dependent_count | 3.653 | 0.056 |
| Months_on_book | 1.897 | 0.168 |
| Total_Relationship_Count | 233.073 | 0 |
| Months_Inactive_12_mon | 240.910 | 0 |
| Contacts_Count_12_mon | 441.868 | 0 |
| Credit_Limit | 5.774 | 0.016 |
| Total_Revolving_Bal | 752.702 | 0 |
| Avg_Open_To_Buy | 0.001 | 0.977 |
| Total_Amt_Chng_Q4_Q1 | 176.962 | 0 |
| Total_Trans_Amt | 296.228 | 0 |
| Total_Trans_Ct | 1,620.122 | 0 |
| Total_Ct_Chng_Q4_Q1 | 930.078 | 0 |
| Avg_Utilization_Ratio | 332.877 | 0 |

From Table 5 and Figure 4, with Factor analysis of mixed data, we need 17 principal components to describe 71.5% of total variation of the dataset, according to the suggestion of Jolliffe (1972). With Kaiser (1958) approach, we need 14 principal components with eigenvalue larger than 1, making up around 62.2% total variation. Scree plot suggests 2 points of elbow, .i.e 7 principal components with 39.8% total variation and 28 principal components with 98.4% total variation. This study selects 17 components for further analyses.

Figure 5 consists of four panels. Panel (a) shows the principal component map where customers are colored by their income level. The first dimension (left hand side) highlights observations with low income level (<$60K). Panel (b) confirms this interpretation and suggests that customers with a low income status have usually female. The correlation circle in Panel (c) confirms that Credit_Limit is negatively correlated with Avg_Utilization_Ratio (recall Pearson correlation -0.48) and that these two variables discriminate the income levels on the first dimension.

Panel (d) plots the variables (categorical or numerical) using squared loadings as coordinates. For numerical variables, squared loadings are squared correlations and for categorical variables squared loadings are correlation ratios. In both cases, they measure the link between the variables and the principal components. One observes that the 3 numerical variables Credit_Limit and Avg_Open_To_Buy and Avg_Utilization_Ratio and the two categorical variables Gender and Income_Category are linked to the first component. On the contrary, the variables Total_Trans_Ct and Total_Trans_Amt are almost orthogonal to these variables and associated to the second component.

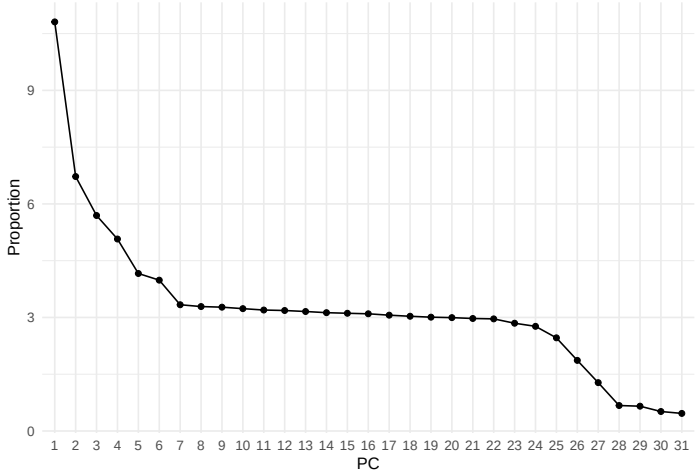


**Figure 4:** Scree plot

**Table 5:** PCA summary

| Eigenvalue | Proportion | Cumulative |
|---|---|---|
| 3.459 | 10.809 | 10.809 |
| 2.152 | 6.724 | 17.533 |
| 1.822 | 5.694 | 23.226 |
| 1.623 | 5.072 | 28.299 |
| 1.332 | 4.162 | 32.460 |
| 1.275 | 3.985 | 36.446 |
| 1.067 | 3.336 | 39.782 |
| 1.052 | 3.288 | 43.070 |
| 1.047 | 3.272 | 46.341 |
| 1.035 | 3.233 | 49.574 |
| 1.023 | 3.196 | 52.770 |
| 1.019 | 3.184 | 55.954 |
| 1.010 | 3.157 | 59.111 |
| 1.000 | 3.126 | 62.238 |
| 0.995 | 3.110 | 65.348 |
| 0.991 | 3.097 | 68.445 |
| 0.979 | 3.060 | 71.505 |
| 0.970 | 3.031 | 74.536 |
| 0.962 | 3.007 | 77.544 |
| 0.959 | 2.995 | 80.539 |
| 0.951 | 2.973 | 83.512 |
| 0.948 | 2.962 | 86.474 |
| 0.911 | 2.847 | 89.321 |
| 0.885 | 2.766 | 92.087 |
| 0.788 | 2.461 | 94.548 |
| 0.597 | 1.866 | 96.414 |
| 0.409 | 1.278 | 97.692 |
| 0.215 | 0.673 | 98.365 |
| 0.210 | 0.655 | 99.020 |
| 0.165 | 0.515 | 99.535 |
| 0.149 | 0.465 | 100 |

## Results and Discussion

We divide the entire dataset into two sets, i.e. training set and test set in which the former constitutes 70% original data as rule of thumb. Furthermore, the stratified random sampling is employed, which is a method of sampling that involves the division of a population into smaller sub-groups known as strata. In this method, or stratification, the strata are formed based on members' shared attributes. The reasons we use stratified sampling rather than simple random sampling include (i) it gives smaller error in estimation if measurements within strata have lower standard deviation (ii) reserving the empirical population proportion of the target variable.

Benchmark model is just a *Naive* logistic regression without any other data processing techniques accompanied. Here we employ stepwise to select best feature combination with the aim of minimizing AIC. Table 6 demonstrates that the benchmark model selects 11 variables, namely Gender, Dependent_count, Marital_Status, Total_Relationship_Count, Months_Inactive_12_mon, Contacts_Count_12_mon, Total_Revolving_Bal, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct and Total_Ct_Chng_Q4_Q1 as contributing factors for discrimination. All variables are significant at 5%. Also, no multicollinearity has been reported. The final AIC is reported at 3,339.

To interpret this output, we should first begin with the intercept. Taking the exponential of the intercept, we have the mean odds to churn individuals in the reference category. So according to this model, the chance of churn is very little as around 0.4 percent. For continuous variables, such as Total no. of products held by the customer or No. of inactive months in the last 12 months, we can see the opposite sign in the coefficients. With the former, as total number of holding products increases, the probability of churn increases. This's kinda strange as holding more products in the bank, the loyalty should be higher and churn rate should decrease. The increase of churn rate implies either (1) there is something wrong with the products of bank, making their customers less satisfied and they leave or (2) there might be another competitive banks with more attractive products to the customers, making them leave. However, this result seems appropriate to the result of exploratory analyses at the first time where the correlation between total no of products and Total Transaction Count is negative, -0.24. In other words, as the total of holding product increases, the number of transaction decreases, signalizing that customers feel the credit card services less attractive. On the other hand, it is normal if number of inactive month increase, it is more likely that customer will leave the credit card service. However, our logistic regression shows the reverse. The negative coefficient shows that if a customers does not use the credit card for a long time, she more likely stay with credit card service. This might be the warning signal that the bank's credit card need an innovation in service so as to re-fire the trigger of credit card consumption needs. We know that the more customer use credit cards, the bank will earn more profits as service fees and interest rates. If the credit card holders do not use for a long time,

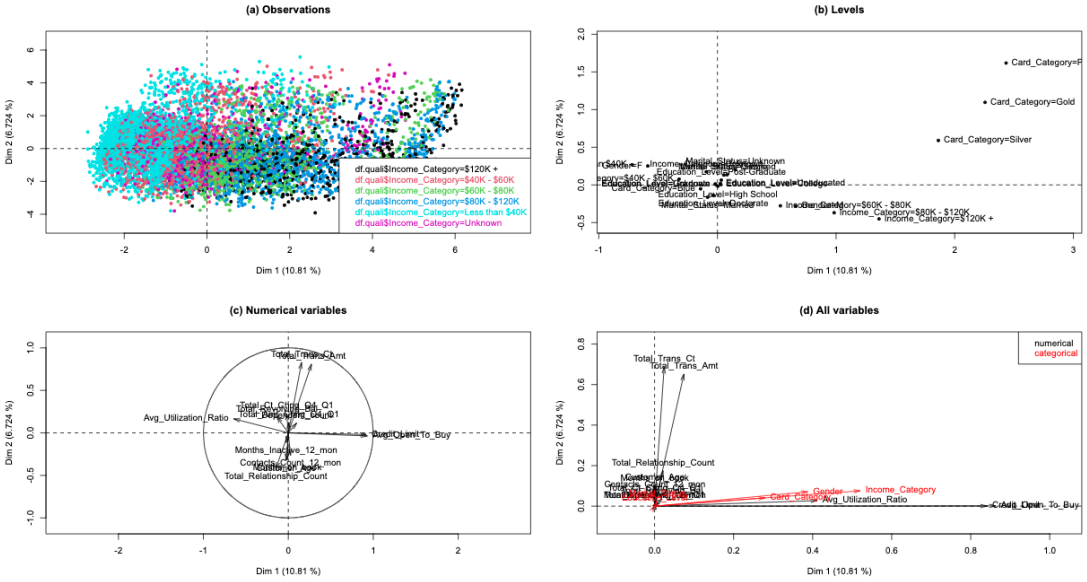


**Figure 5:** PCA result

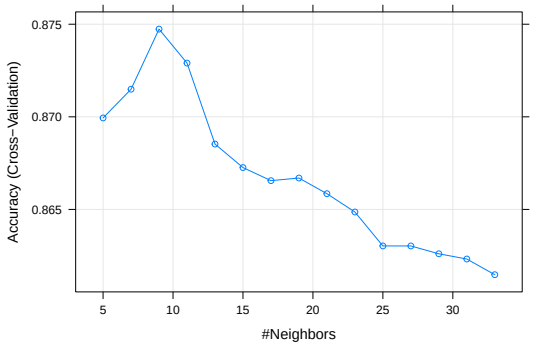


**Figure 6:** KNN scree plot

it might create a financial distress. The positive coefficient of usage behavior variables confirms our viewpoint: if customer uses credit card more, so that total transaction and change in total amount spending last year increases, the rate of churning will increase. So, there must be something wrong with credit card services. Interestingly, positive coefficient in Gender M shows that males will more likely leave the service than females and if number of dependent persons increases, the clients more likely stay with the credit card.

SMOTE technique requires KNN to select the best K. Following result shows K should be 25 for the elbow point, see Figure 6. According to Table 6, the SMOTE logistic regression selects 13 variables, beside the same 11 variables as benchmark output, the SMOTE technique complements further 2 variables: Customer_Age and Card_Category. Also, no multicollinearity has been reported. With the enhance of SMOTE, the reported AIC improves to 2,965. It means that if customer is getting older, he is more likely to leave the service and if the customer uses the blue credit card, he will have more reasons to leave. This suggests a potential solution for the churning problem of the bank. The bank should invest more in the younger customers with higher credit card grade such as Gold, Platinum, and Silver. The interesting point here is that investing in Gold class is more strategic and better than Platinum class because the absolute value of coefficient of Gold is higher. The signs of other variables maintain the same as Naive versions.

Also, according to Figure 7, Gender, Dependent_count, Education_Level, Marital_Status, Income_Category, Card_Category and Months_on_book are generally unpredictive. This result seems conflicted to previous regression. Figure 8 reports the trend in WoE of variables. As can be seen, there is a negative linear relationship between the odd of $P(y = 1)$ and the WoE of Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1, Total_Relationship_Count, Total_Trans_Ct. Some several variables show non-linearity between odd of $P(y = 1)$ and WoE, namely Avg_Utilization_Ratio, Credit_Limit, Total_Revolving_Bal, Total_Trans_Amt. We employ isotonic regression to attempt to implement the monotonic binning, see Figure 9. We next implement the WoE transformation for training set and run

**Table 6:** Model summaries

| | *Dependent variable:* | | |
|---|---|---|---|
| | Attrition_Flag | | |
| | Naive | SMOTE | WOE |
| | (1) | (2) | (3) |
| Customer_Age | | 0.016** | |
| | | (0.006) | |
| GenderM | 0.716*** | 0.342*** | |
| | (0.092) | (0.095) | |
| Dependent_count | −0.146*** | −0.097** | |
| | (0.035) | (0.038) | |
| Marital_StatusMarried | 0.521*** | 0.540*** | |
| | (0.188) | (0.189) | |
| Marital_StatusSingle | −0.069 | 0.079 | |
| | (0.189) | (0.190) | |
| Marital_StatusUnknown | −0.080 | −0.019 | |
| | (0.237) | (0.240) | |
| Card_CategoryGold | | −0.778** | |
| | | (0.377) | |
| Card_CategoryPlatinum | | −0.700 | |
| | | (0.621) | |
| Card_CategorySilver | | −0.578*** | |
| | | (0.178) | |
| Total_Relationship_Count | 0.476*** | 0.400*** | −2.320*** |
| | (0.033) | (0.033) | (0.112) |
| Months_Inactive_12_mon | −0.506*** | −0.561*** | −1.013*** |
| | (0.045) | (0.053) | (0.088) |
| Contacts_Count_12_mon | −0.507*** | −0.577*** | −0.844*** |
| | (0.044) | (0.048) | (0.090) |
| Total_Revolving_Bal | 0.001*** | 0.001*** | |
| | (0.0001) | (0.0001) | |
| Credit_Limit | | | −1.108*** |
| | | | (0.129) |
| Total_Amt_Chng_Q4_Q1 | 0.493** | 0.656*** | −0.569*** |
| | (0.225) | (0.245) | (0.064) |
| Total_Trans_Amt | −0.0005*** | −0.001*** | −0.325*** |
| | (0.00003) | (0.00003) | (0.054) |
| Total_Trans_Ct | 0.119*** | 0.140*** | −0.774*** |
| | (0.004) | (0.005) | (0.053) |
| Total_Ct_Chng_Q4_Q1 | 2.699*** | 3.004*** | −0.424*** |
| | (0.224) | (0.240) | (0.044) |
| Constant | −5.588*** | −8.430*** | 1.536*** |
| | (0.377) | (0.528) | (0.046) |
| Observations | 7,089 | 4,556 | 7,089 |
| Log Likelihood | −1,655.623 | −1,464.505 | −1,800.547 |
| Akaike Inf. Crit. | 3,339.246 | 2,965.011 | 3,619.094 |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

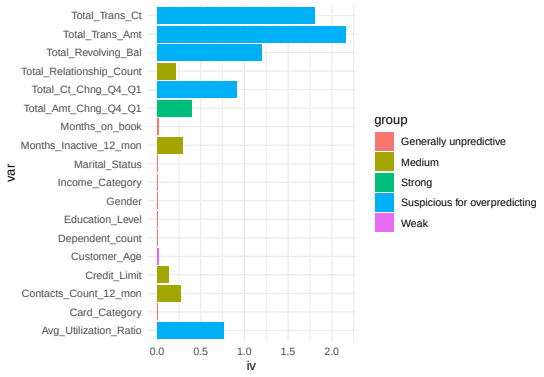


**Figure 7:** Information value

**Table 7:** PCA model summary

| | *Dependent variable:* |
|---|---|
| | Attrition_Flag |
| | PCA |
| 'dim 2' | 0.677*** |
| | (0.032) |
| 'dim 4' | 0.905*** |
| | (0.035) |
| 'dim 5' | 0.278*** |
| | (0.038) |
| 'dim 9' | −0.128*** |
| | (0.037) |
| 'dim 12' | 0.108*** |
| | (0.037) |
| 'dim 13' | −0.361*** |
| | (0.039) |
| 'dim 14' | −0.183*** |
| | (0.038) |
| 'dim 16' | 0.083** |
| | (0.037) |
| 'dim 17' | −0.167*** |
| | (0.038) |
| Constant | 2.353*** |
| | (0.051) |
| Observations | 7,089 |
| Log Likelihood | −2,270.159 |
| Akaike Inf. Crit. | 4,560.317 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

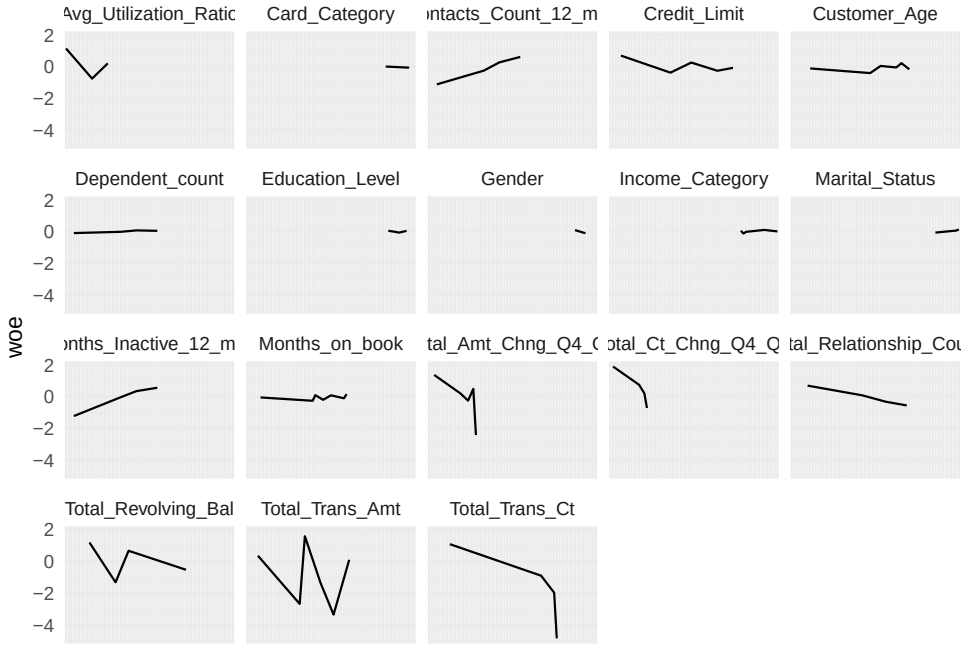


**Figure 8:** Weight of evidence

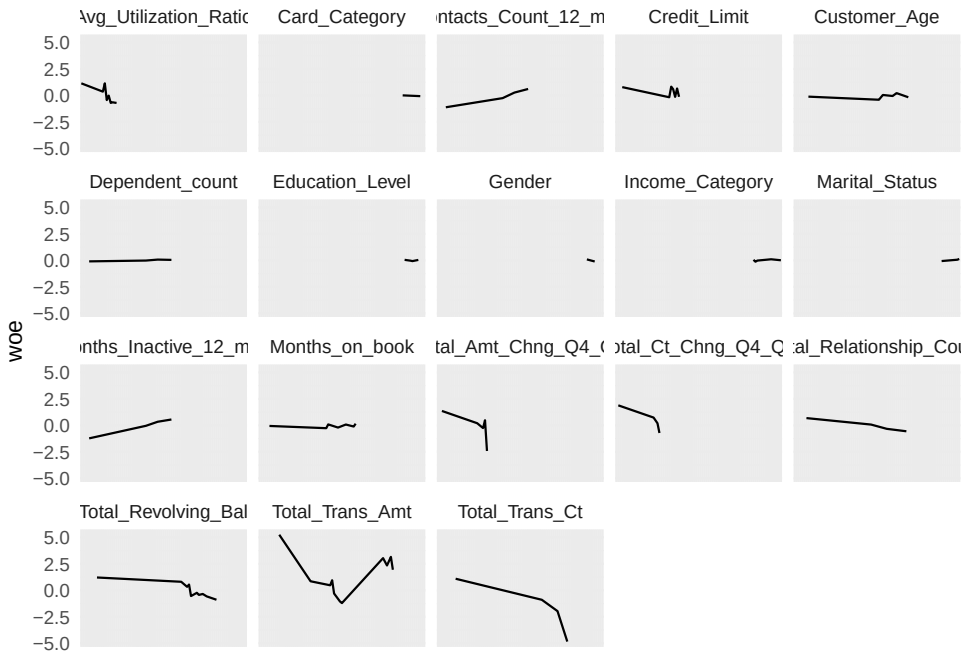


**Figure 9:** Weight of evidence after using isotonic regression

logistic regression. Accordingly, the WOE transformation helps to reduce the AIC of the best stepwise logistic model to 3092.3 as opposed to 3339.2 of benchmark. Interestingly, there are no categorical variables in original training set due to their general unpredictive power based on IV rules. Checking multicolinearity checks suggest Avg_Utilization_Ratio and Total_Revolving_Bal should be excluded from the model. Also, check for significance of individual variable in new models, Customer_Age should be exculded due to its p-vale less than 0.05. Final version WoE-based Logistic regression is reported in Table 6.

Table 6 reports the WoE transformation unfortunately make the training estimation less efficient as the AIC increase a little bit to 3,619 as compared to Naive model. Furthermore, the WOE-based variables is unitless, making this method is harder for interpretation. A better improvement for is representation of the WOE model is using scorecard, as an idea which is commonly used in the financial credit risk management, see Table 8. The idea is simple: a customer will start with the base score, at 277, and when her features fall in any bins of the variables, her score will add or subtract to corresponding points. And, if the score is higher, her probability of churn is also lower. For example, we can see in the no of inactive months, as it increases, the aligned points increases too, making the churning rate lower with higher score. Another example is as total no of holding products increases, the score decreases, making the churn probability increases. These results confirm the results of 2 aforementioned models. Therefore, the scorecard is a better idea for WOE-based model representation. The WOE transformation Logistic regression just recognizes only 8 variables with the new appearance of Credit_limit. The scores aligned to each bin of Credit_limit also shows non-monotonic, leading another problem for the bank about its customer segmentation. This is because the higher credit limit, the higher class of the customer. If banks target accurately true customer segmentation, the segmentation members should have lower churn rate. We can see in the summary, the low churn rate class is mainly low credit limits, less perform transactions with little money amount and holding less products of the bank. Overall, if this circumstance continues, the bank will end up with future business and finance distress on credit card service.

In terms of PCA-based logistic regression, we select 17 as the number of principal components. The final PCA model contains only 2nd, 4th, 5th, 9th, 12th, 13th, 14th, 16th, and 17th components, see Table 7. Interestingly, The first component which constitutes the most of total variation of dataset is excluded. Actually in full model, this component is insignificant at 5%. VIF check shows there are no multicollinearity as a result of PCA technique. So according to this model, card information does not contain any discrimination power, as opposed to aforementioned model. This is a misleading result because the prior analyses show the reverse.

A no-skill classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases. A model with no skill is represented at the point (0.5, 0.5) in the plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0.A model with no skill at each threshold is

**Table 8:** Scorecard representation

| variable | bin | points |
|---|---|---|
| basepoints | | 277 |
| Total_Relationship_Count | [-Inf,3) | 115 |
| Total_Relationship_Count | [3,4) | 12 |
| Total_Relationship_Count | [4,6) | -55 |
| Total_Relationship_Count | [6, Inf) | -94 |
| Months_Inactive_12_mon | [-Inf,2) | -90 |
| Months_Inactive_12_mon | [2,3) | -3 |
| Months_Inactive_12_mon | [3,4) | 25 |
| Months_Inactive_12_mon | [4, Inf) | 40 |
| Contacts_Count_12_mon | [-Inf,2) | -68 |
| Contacts_Count_12_mon | [2,3) | -16 |
| Contacts_Count_12_mon | [3,4) | 16 |
| Contacts_Count_12_mon | [4, Inf) | 37 |
| Credit_Limit | [-Inf,1764) | 63 |
| Credit_Limit | [1764,1819) | 66 |
| Credit_Limit | [1819,1900) | 47 |
| Credit_Limit | [1900,1931) | -11 |
| Credit_Limit | [1931,1952) | 52 |
| Credit_Limit | [1952,15219) | -11 |
| Credit_Limit | [15219, Inf) | -14 |
| Total_Amt_Chng_Q4_Q1 | [-Inf,0.5) | 56 |
| Total_Amt_Chng_Q4_Q1 | [0.5,0.6) | 8 |
| Total_Amt_Chng_Q4_Q1 | [0.6,0.95) | -11 |
| Total_Amt_Chng_Q4_Q1 | [0.95,1.1) | 20 |
| Total_Amt_Chng_Q4_Q1 | [1.1, Inf) | -99 |
| Total_Trans_Amt | [-Inf,891) | 123 |
| Total_Trans_Amt | [891,929) | 71 |
| Total_Trans_Amt | [929,947) | 55 |
| Total_Trans_Amt | [947,974) | 74 |
| Total_Trans_Amt | [974,1000) | 46 |
| Total_Trans_Amt | [1000,2899) | 20 |
| Total_Trans_Amt | [2899,2932) | 11 |
| Total_Trans_Amt | [2932,2935) | 22 |
| Total_Trans_Amt | [2935,3121) | -7 |
| Total_Trans_Amt | [3121,3151) | -25 |
| Total_Trans_Amt | [3151, Inf) | -28 |
| Total_Trans_Ct | [-Inf,58) | 61 |
| Total_Trans_Ct | [58,76) | -50 |
| Total_Trans_Ct | [76,94) | -110 |
| Total_Trans_Ct | [94, Inf) | -271 |
| Total_Ct_Chng_Q4_Q1 | [-Inf,0.45) | 58 |
| Total_Ct_Chng_Q4_Q1 | [0.45,0.55) | 22 |
| Total_Ct_Chng_Q4_Q1 | [0.55,0.6) | 6 |
| Total_Ct_Chng_Q4_Q1 | [0.6, Inf) | -22 |

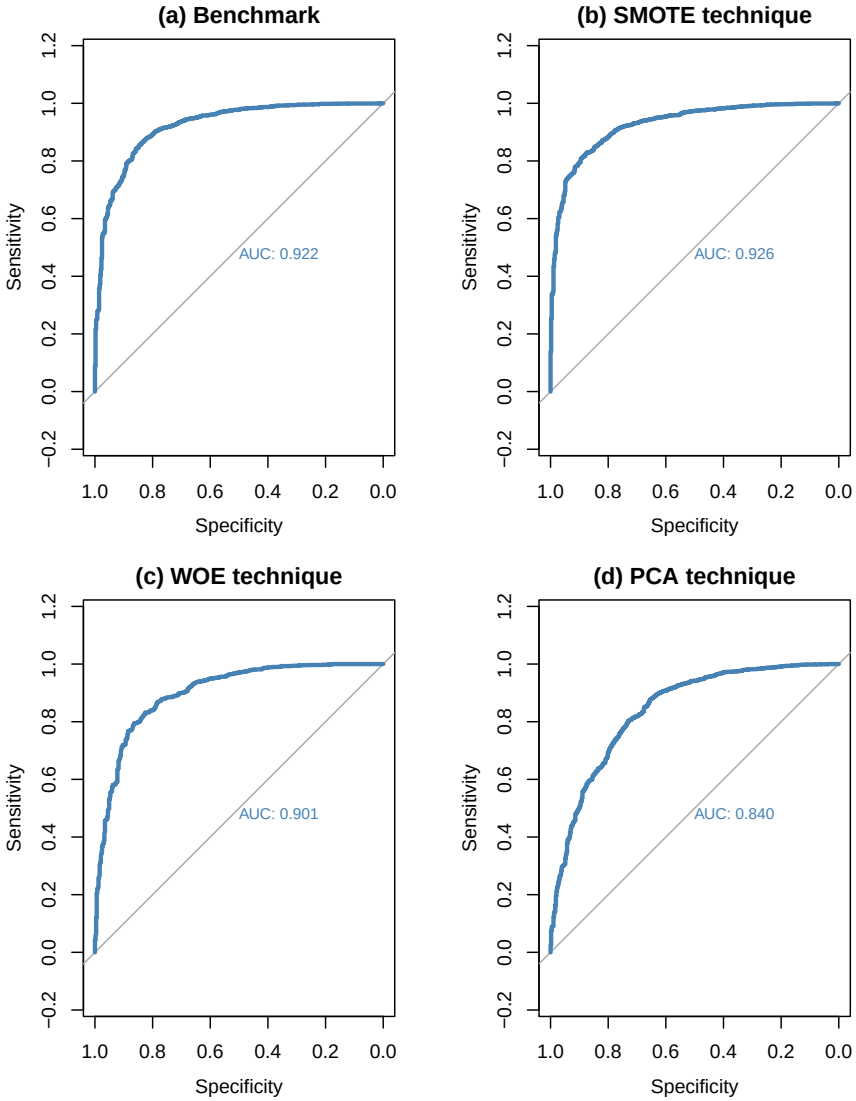


**Figure 10:** ROC curves and corresponding AUC

represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. A skillful model will assign a higher probability to a randomly chosen real positive occurrence than a negative occurrence on average. This is what we mean when we say that the model has skill. Generally, skillful models are represented by curves that bow up to the top left of the plot. A model with perfect skill is represented at a point (0,1) and AUC reaches 1.0.

We compare the performance of several models based on the test set. From Figure 10 and 11 Some highlights are:

- SMOTE does improve the benchmark performance in terms of both AUC and Recall.
- WOE improves the benchmark performance in terms of only Recall.
- PCA reduces the benchmark performance in terms of both AUC and Recall.

## Conclusion

To answer our first research question, the predictors that we used in our model to predict credit card customer churn were: Gender, number of dependents, marital status, relationship count, months inactive, count of contacts, revolving balance, amount change from quarter 1 to 4, total transaction amount, total transaction count and total count changed from quarter 1 to 4.

For our second question, SMOTE turned out to be the best processing technique to improve the logistic regression. Our hypothesis was incorrect since we thought that PCA with logistic regression

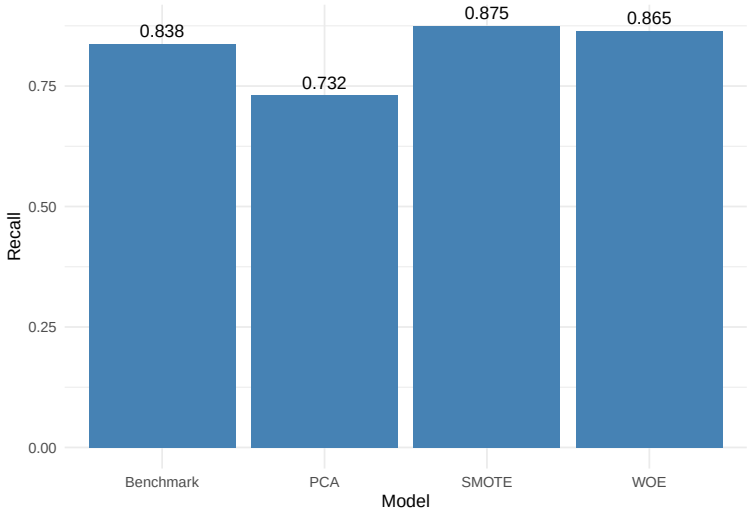


**Figure 11:** Recall

will yield the prediction model, but it turned out to be SMOTE since it resolved our imbalance in our dataset.

Overall, with our analysis, the bank manager can now predict future clients from churning!

Our work is yet another example of how data science and analytics helps improve performance in many different industries including that of the financial services.

## Acknowledgements

We would like to thank our supervisor Stephanie Besser for her guidance throughout this project.

## Bibliography

2014. URL https://digital-library.theiet.org/content/conferences/10.1049/cp.2014.1576. [p1]

2019. URL https://doi.org/10.1145/3330204.3330220. [p1]

H. Abbasimehr and S. Alizadeh. A novel genetic algorithm based method for building accurate and comprehensible churn prediction models. 2013. [p1]

H. Abdi, L. J. Williams, and D. Valentin. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, 5(2): 149–179, 2013. ISSN 1939-5108. [p4]

A. K. Ahmad, A. Jafar, and K. Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):28, 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0191-6. URL https://doi.org/10.1186/s40537-019-0191-6. [p1]

W.-H. Au, K. C. Chan, and X. Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, 7(6):532–545, 2003. ISSN 1089-778X. [p1, 2]

V. Barnett and T. Lewis. Outliers in statistical data. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*, 1984. [p2]

I. Brandusoiu and G. Toderean. *NEURAL NETWORKS FOR CHURN PREDICTION IN THE MOBILE TELECOMMUNICATIONS INDUSTRY*. 2020. ISBN 978-973-720-778-4. [p1]

N. Chater and M. Oaksford. *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, USA, 2008. ISBN 0199216096. [p3]

M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Multivariate analysis of mixed data: The r package pcamixdata. *arXiv preprint arXiv:1411.4911*, 2014. [p4]

M. Chavent, V. Kuentz, A. Labenne, B. Liquet, and J. Saracco. Pcamixdata: Multivariate analysis of mixed data. *R package version*, 3, 2017. [p4]

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. ISSN 1076-9757. [p3]

K. Chen, K. Zhu, Y. Meng, A. Yadav, and A. Khan. Mixed credit scoring model of logistic regression and evidence weight in the background of big data. In *International Conference on Intelligent Systems Design and Applications*, pages 435–443. Springer. [p4]

K. Coussement and D. Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327, 2008. ISSN 0957-4174. [p5]

B. Culbert, B. Fu, J. Brownlow, C. Chu, Q. Meng, and G. Xu. Customer churn prediction in superannuation: A sequential pattern mining approach. In J. Wang, G. Cong, J. Chen, and J. Qi, editors, *Databases Theory and Applications*, pages 123–134. Springer International Publishing. ISBN 978-3-319-92013-9. [p1]

K. Dahiya and K. Talwar. Customer churn prediction in telecommunication industries using data mining techniques-a review. *Int. J. Adv. Res. Comput. Sci. Softw. Eng*, 5(4):417–433, 2015. [p2]

A. Dijendra and J. Sisodia. Telecomm churn prediction using fundamental classifiers to identify cumulative probability. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–5. doi: 10.1109/ICCICT50803.2021.9510111. [p1]

O. Dorokhov, L. Dorokhova, L. Malyarets, and I. Ushakova. Customer churn predictive modeling by classification methods. *SERIES III - MATEMATICS, INFORMATICS, PHYSICS*, 13(62):347–362, 2020. doi: 10.31926/but.mif.2020.13.62.1.26. [p1]

B. Escofier and J. Pages. Multiple factor analysis (afmult package). *Computational statistics & data analysis*, 18(1):121–140, 1994. ISSN 0167-9473. [p4]

H. Faris, B. Al-Shboul, and N. Ghatasheh. *A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry*, volume 8733. 2014. ISBN 978-3-319-11288-6. doi: 10.1007/978-3-319-11289-3_36. [p1]

I. J. Good. Probability and the weighing of evidence. Report, C. Griffin London, 1950. [p4]

S. G. Heeringa, B. T. West, and P. A. Berglund. *Applied survey data analysis*. chapman and hall/CRC, 2017. ISBN 1315153270. [p2]

D. W. Hosmer. Model-building strategies and methods for logistic regression. *Applied logistic regression*, 1989. [p3]

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013. ISBN 0470582472. [p3]

S. Höppner, E. Stripling, B. Baesens, S. v. Broucke, and T. Verdonck. Profit driven decision trees for churn prediction. *European Journal of Operational Research*, 284(3):920–933, 2020. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.11.072. URL https://www.sciencedirect.com/science/article/pii/S0377221718310166. [p1]

H. Jain, G. Yadav, and R. Manoov. Churn prediction and retention in banking, telecom and it sectors using machine learning techniques. In S. Patnaik, X.-S. Yang, and I. K. Sethi, editors, *Advances in Machine Learning and Computational Intelligence*, pages 137–156. Springer Singapore. ISBN 978-981-15-5243-4. [p1]

H. Jain, A. Khunteta, and S. Srivastava. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 167:101–112, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.03.187. URL https://www.sciencedirect.com/science/article/pii/S1877050920306529. [p1]

I. T. Jolliffe. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):160–173, 1972. ISSN 0035-9254. [p8]

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958. ISSN 0033-3123. [p8]

Y. Khan, S. Shafiq, A. Naeem, S. Ahmed, N. Safwan, and S. Hussain. Customers churn prediction using artificial neural networks (ann) in telecom industry. *International Journal of Advanced Computer Science and Applications*, 10, 2019. doi: 10.14569/IJACSA.2019.0100918. [p1]

A. Kracklauer, D. Mills, and D. Seifert. Customer management as the origin of collaborative customer relationship management. *Collaborative Customer Relationship Management - Taking CRM to the Next Level*, pages 3–6, 2004. ISSN 978-3-642-05529-4. doi: 10.1007/978-3-540-24710-4_1. [p1]

C. M. Laha. A, & krishna p.(2006). modeling churn behavior of bank customers using predictive data mining techniques. In *National conference on soft computing techniques for engineering application (SCT-2006)*. [p1]

B. Larivière and D. Van den Poel. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27(2):277–285, 2004. ISSN 0957-4174. [p2]

T.-S. Lee, C.-C. Chiu, Y.-C. Chou, and C.-J. Lu. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, 50(4):1113–1130, 2006. ISSN 0167-9473. [p5]

Y. Li and G. Xia. The explanation of support vector machine in customer churn prediction. In *2010 International Conference on E-Product E-Service and E-Entertainment*, pages 1–4. doi: 10.1109/ICEEE.2010.5660501. [p1]

R. Ling and D. C. Yen. Customer relationship management: An analysis framework and implementation strategies. *Journal of Computer Information Systems*, 41(3):82–97, 2001. ISSN 0887-4417. doi: 10.1080/08874417.2001.11647013. URL https://www.tandfonline.com/doi/abs/10.1080/08874417.2001.11647013. [p1]

N. Luo and Z. Mu. Bayesian network classifier and its application in crm. *Computer Application*, 24(3):79–81, 2004. [p2]

J. Manďák and J. Hančlová. Use of logistic regression for understanding and prediction of customer churn in telecommunications. 2019. ISSN 0322-788X. [p1]

J. Mount and N. Zumel. *Practical data science with R*. Simon and Schuster, 2019. ISBN 1638352747. [p2]

E. W. T. Ngai. Customer relationship management research (1992-2002). *Marketing Intelligence & Planning*, 23(6):582–605, 2005. ISSN 0263-4503. doi: 10.1108/02634500510624147. URL https://doi.org/10.1108/02634500510624147. [p1]

E. W. T. Ngai, L. Xiu, and D. C. K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592–2602, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.02.021. URL https://www.sciencedirect.com/science/article/pii/S0957417408001243. [p1]

G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12):15273–15285, 2011. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2011.06.028. URL https://www.sciencedirect.com/science/article/pii/S0957417411009237. [p1]

A. Rodan, H. Faris, J. Al-sakran, and O. Al-Kadi. A support vector machine approach for churn prediction in telecom industry. *International journal on information*, 17, 2014. [p1]

M. Saghir, Z. Bibi, S. Bashir, and F. H. Khan. Churn prediction using neural network based individual and ensemble models. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 634–639. ISBN 2151-1411. doi: 10.1109/IBCAST.2019.8667113. [p1]

C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira. *Noise Versus Outliers*, pages 163–183. Springer International Publishing, Cham, 2016. ISBN 978-3-319-43742-2. doi: 10.1007/978-3-319-43742-2_14. URL https://doi.org/10.1007/978-3-319-43742-2_14. [p2]

H. T. SEbASTIAN and R. Wagh. Churn analysis in telecommunication using logistic regression. *Orient. J. Comp. Sci. and Technol*, 10(1):207–212, 2017. [p2]

T. Shah. Putting a quality edge on digital wireless networks: The echo canceller is at the center of complex digital networks, holding the key to a variety of voice enhancement opportunities. *Cellular Business*, 13:82–91, 1996. ISSN 0741-6520. [p1]

F. Stojanovski. Churn prediction using sequential activity patterns in an on-demand music streaming service. [p1]

E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, and M. Snoeck. Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, 40:116–130, 2018. ISSN 2210-6502. doi: https://doi.org/10.1016/j.swevo.2017.10.010. URL https://www.sciencedirect.com/science/article/pii/S2210650216301754. [p1]

D. L. Weed. Weight of evidence: a review of concept and methods. *Risk Analysis: An International Journal*, 25(6):1545–1557, 2005. ISSN 0272-4332. [p3]

G.-e. Xia and W. Jin. Model of customer churn prediction on support vector machine. *Systems Engineering - Theory & Practice*, 28:71–77, 2008. [p1]

Y. Xie, X. Li, E. W. T. Ngai, and W. Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1):5445–5449, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.06.121. URL https://www.sciencedirect.com/science/article/pii/S0957417408004326. [p1]

Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren. Customer churn prediction using improved one-class support vector machine. In *International Conference on Advanced Data Mining and Applications*, pages 300–306. Springer. [p1]

*Hieu Phi Nguyen*
*College of Computing and Digital Media, DePaul University*
*243 South Wabash Avenue, Chicago, IL 60604*
pnguye40@depaul.edu

*John Vishwas Nelson*
*College of Computing and Digital Media, DePaul University*
*243 South Wabash Avenue, Chicago, IL 60604*
pnguye40@depaul.edu

*Seif Abuhashish*
*College of Computing and Digital Media, DePaul University*
*243 South Wabash Avenue, Chicago, IL 60604*
pnguye40@depaul.edu

*Mariana Tollko*
*College of Computing and Digital Media, DePaul University*
*243 South Wabash Avenue, Chicago, IL 60604*
pnguye40@depaul.edu

*Haodong Wu*
*College of Computing and Digital Media, DePaul University*
*243 South Wabash Avenue, Chicago, IL 60604*
pnguye40@depaul.edu