# Project Proposal: Predicting Bank Turnover with Machine Learning Techniques

Sixue Liu, Jiewen Luo, Yuanxiaoyue Yang

January 2020

## 1 Introduction

The banking industry in the United States is highly competitive. There are currently 5294 domestic and international banks providing banking services in the United States. Given such an intensively competitive market, the demand for banking services becomes highly fragmented. This brings a great challenge for banks to maintain and expand their customer base.

Provided that relatively higher costs are associated with acquiring new customers and re-acquiring deflected customers, retaining current customers can help banks save their customer managing budget (Verbeke et al., 2011). Besides, in the banking industry, long-term account holders are more likely to generate high profits for the bank (Keramati et al., 2016). Therefore, focusing on existing customers appears to be the most effective way of managing customer relationships in the banking industry.

However, because of low shifting costs and highly available alternatives, customers have high flexibility in shifting from one bank to another. Various reasons can cause customers to leave a bank, including the accessibility of frontier technology and products, service satisfaction, interest rate, geographical location, and the variety of products and services (Kumar and Ravi, 2008). This makes managing and maintaining existing customers even harder.

Knowing of this pressing issue, our study aims at assessing the likelihood of customer churn using available bank customer dataset. We will utilize various machine learning techniques, including logistic regression, decision tree and random forest, LASSO, and support vector machine, to identify the hidden patterns of customer churning behavior and establish a relationship between customer churn and associated customer features. By establishing an optimized ML model to form a better understanding of customer churning behavior, our goals are to detect customers that are at risk of leaving the bank in the future, to analyze causes that drive customers' churning behavior, and to eventually help banking firms to develop effective strategies to improve customer retention in the future.

# 2 Literature Review

This section provides a brief review of existing machine learning literature on various models that have been used in prediction of customer churn. Ngai et al. (2009) classified 900 articles relevant to applications of data mining techniques in customer relationship management from academic literature and identified that classification and association models are the two most commonly used.

Widely used models for solving classification problems include Decision Tree (DT), Logistic Regression, and support vector machine (SVM). Keramati et al. (2016) used a DT method to identify features of churners from electronic banking services and the results show that the model successfully identified features of 5 groups of churners. There are also studies that use DT to study churn prediction problems in telecommunication industry (Keramati et al., 2014; Huang et al., 2012). Nie et al. (2011) used both DT and logistic regression to predict customer churn using credit card data collected from a real Chinese bank and found that results achieved from logistic regression are slightly better than results from DT. Support vector machine (SVM) is also extensively used for customer churn prediction tasks. Huang et al. (2012) compared the performance of DT and SVM in predicting telecommunication customer churn and concluded that the preference between the two methods depends on the decision makers' objectives. Other methods, such as Neural Networks, Random Forest, Naive Bayes, and K-Nearest Neighbors, have also been used in various similar studies (Zhao and Tai, 2014; Rajamohamed and Manokaran, 2018; Zoric, 2016).

Many recent studies have started employing methodology that combines different techniques. A 2008 study developed an ensemble system based on majority voting that constituted of several machine learning techniques such as random forest, decision tree, logistic regression, support vector machine, multilayer perception, and radiao basis function (Kumar and Ravi, 2008). They tested their method using data of credit card costumers from a Latin American bank and concluded that the majority voting achieved good overall accuracy in prediction. Farquad et al. (2014) proposed a modified SVM method incorporating rules from Naive Bayes Tree and tested it using a bank credit card dataset. It turned out that this comprehensive hybrid approach improved the performance of previous SVM models. Keramati et al. (2014) also proposed a hybrid methodology that employed Artificial Neural Network, K-Nearest Neighbors (KNN), DT, and SVM and found that the hybrid method achieves a considerably higher than 95% accuracy for precision and recall measures, which outperforms any single method.

# 3 Machine Learning Methods

We will use multiple machine learning models in our project to predict whether bank customers will churn or not based on their demographic characteristics and other related variables. Our initial conceive is to split the data set into training set, validation set, and test set. We would like to use training set to train different models, validation set to select the best performance model and tune the hyper-parameters, and test set to test our model's accuracy. In this part, we will briefly discuss the mechanism of each model we will use. Since whether the customer churn or not is a binary variable, we will mainly use logistic regression and classification methods.

## 3.1 Logistic Regression

The first approach is logistic regression. A typical approach for Logit model of customer $j$ in group $h$ at time $t$ can be expressed as

$$Y_{jht} = f(X, D, O)'\beta + \epsilon_{jt} \tag{1}$$

In this case, $f$ gives us the interaction between the observables ($X$), demographics ($D$), and other related variables ($O$).

## 3.2 LASSO

LASSO is a penalized regression method. In Patrick et al. (2015), the regression is given by

$$min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{p} |\beta_j|) \tag{2}$$

In LASSO regression, $t$ is the tuning parameter governing how strictly additional regressors are penalized. In this form, LASSO is expressed in least square approach, and we will explore more on how it fits logit model. Besides, we will also test which simple penalized regression method, namely ridge or LASSO, will be more suited in our bank customer churning model.

## 3.3 Support Vector Machines

Support Vector Machines (SVM) is another penalized method of regression. It is intensively used for classification tasks. SVM uses a binary classifier to divide observations into two classes with a hyperplane. The main goal is to find the optimal separated hyperplanes through nonlinear mapping to derive correct classifications (Farquad et al., 2014). Therefore, SVM has great ability to model nonlinearities. The regression equation is:

$$\beta = min_{\beta} \sum_{i=1}^{n} V(y_i - X_i'\beta) + \lambda|\beta| \tag{3}$$

SVM gives us another way to analyze the loss differing from the least square approach, where the hinge loss function is

$$V_{\epsilon}(r) = \begin{cases} 0 & if|r|<\epsilon, \\ |r|-\epsilon & otherwise. \end{cases} \tag{4}$$

The tuning parameter, $\epsilon$, controls which errors are included in the regression.

### 3.4 Regression Trees

Regression trees approximate functions by partitioning the characteristic space into a series of hypercubes and reporting the average value of the function in each of those partitions. Regression tree is suitable for the goal of finding specific features of churners, as it produces easily understood and visualized classification rules Keramati et al. (2014). Another reason why regression tree is particularly suitable for churn prediction problems is that these problems often involve using data of customer's demographic information (age, gender, career, education, etc.), transaction method, the length of the customer association, and other account information. As regression tree performs well with numerical and categorical data, it turns out to be a suitable method (Nie et al., 2011). We will use two regression tree estimators: bagging (Breiman, 1996) and random forests (Breiman, 2001) in our analysis.

## 4 Data

The dataset we will use for our analysis is a free public dataset of bank customer information called "Bank Turnover Dataset" obtained from Kaggle.com[1]. The dataset contains demographic and bank account information including from 10,000 customers. Here we provide a list of all variables and their descriptions:

- Gender: A dummy variable indicating the customer's gender
- Geography: A categorical variable indicating the customer's geographic location
- Has Credit Card: A dummy variable indicating whether the customer owns a credit card through the bank
- Is Active Member: A dummy variable indicating whether the customer is an active member at the bank
- Exited: A dummy variable indicating whether the customer exited from the bank services during a six-month period.
- Age: The customer's current age
- Credit Score: The customer's current credit score
- Tenure: The duration of time the customer has been using the bank services
- Balance: The customer's current account balance at the bank
- Estimated Salary: The customer's estimated salary
- Number of Products: The number of bank accounts or bank account affiliated products the customer has

We provide percentage share of each category among the total number of observations for categorical and dummy variables in Table 1. We also provide summary statistics for numerical variables in Table 2.

[1] The link to the dataset is: https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling.

| **Variables** | | **Percentage Share** |
|---|---|---|
| Gender | | |
| | *Female* | 45.43% |
| | *Male* | 54.57% |
| Geography | | |
| | *France* | 50.14% |
| | *Germany* | 25.09% |
| | *Spain* | 24.77% |
| Has Credit Card | | |
| | *Yes* | 70.55% |
| | *No* | 29.45% |
| Is Active Member | | |
| | *Yes* | 51.51% |
| | *No* | 48.49% |
| Exited | | |
| | *Yes* | 20.37% |
| | *No* | 79.63% |
| Total Number of Observations | | 10000 |

Table 1: Summary Statistics for Categorical and Dummy Variables

| **Variables** | **Mean** | **Standard Deviation** | **Min** | **Median** | **Max** |
|---|---|---|---|---|---|
| Age | 38.92 | 10.49 | 18 | 37 | 92 |
| Credit Score | 650.53 | 96.65 | 350 | 652 | 850 |
| Tenure | 5.01 | 2.89 | 0 | 5 | 10 |
| Balance | 76,485.89 | 62,397.41 | 0 | 97,199 | 250,898.1 |
| Estimated Salary | 100,090.2 | 57,510.49 | 11.58 | 100,193.91 | 199,992.5 |
| Number of Products | 1.53 | 0.58 | 1 | 1 | 4 |
| Total Number of Observations | | | | | 10000 |

Table 2: Summary Statistics for Numerical Variables

# References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5.

Farquad, M. A., Ravi, V., and Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing Journal*, 19:31–40.

Huang, B., Kechadi, M. T., and Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425.

Keramati, A., Ghaneei, H., and Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(1).

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., and Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing Journal*, 24:994–1012.

Kumar, D. A. and Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1):4–28.

Ngai, E. W., Xiu, L., and Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2 PART 2):2592–2602.

Nie, G., Rowe, W., Zhang, L., Tian, Y., and Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12):15273–15285.

Patrick, B., Denis, N., Stephen P., R., and Miaoyu, Y. (2015). Machine learning methods for demand estimation. *The American Economic Review*, 105(5):481.

Rajamohamed, R. and Manokaran, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing*, 21(1):65–77.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, 38(3):2354–2364.

Zhao, S. X. and Tai, Q. Y. (2014). Applied research on data mining in bank customer churn. *Applied Mechanics and Materials*, 687-691:5023–5027.

Zoric, A. B. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems*, 14(2):116–124.