Retail Bank Customer Relationship Management

Data Science for Business Analytics

Alsan Ali
Sajan Bang
Hitesh Laxmichand Patel

1. Introduction

Retail and consumer banking institutions have multiple product and service offerings to customers. The typical line of products which such banks offer their customers include checking accounts, savings accounts, credit cards, Certificates of Deposit ("CDs"), mortgages, and automobile loans. Most customers can benefit from many of these offerings simultaneously, and banks can gain cost efficiencies if the same customer participates in multiple service offerings. Since each customer may engage with the bank through multiple channels, Customer Relationship Management ("CRM") can be complex, but also serve as a powerful tool in driving continued growth of the business.

In this project, we analyze a particular use case where data mining and analytics models applied to customer data can help a bank target products to customers. Specifically, we analyze data for customers who currently customers of the bank, and identify those who currently do not have a credit card at the bank but should be targeted as potential credit card customers. We also examine data on customers who have exited the bank to discern patterns of customer churn. This analysis takes advantage of existing customer relationships to grow different areas of the bank's business. From a CRM standpoint, this is an effective strategy since the bank can deploy customer data collated from different channels and disseminate information that can provide actionable insights in product sales and customer retention.

2. Business Understanding

Retail banks can be distinguished from other financial institutions as a deposit-taking entity primarily targeting regular consumers as customers. Unlike investment banks and commercial banks, which may seek business opportunities and growth through relationships with a (relatively) small number of large corporate clients, retail banks may have hundreds of thousands or even millions of customers. For example, Bank of America, which is the second largest bank by assets in the United States, has approximately 50 million customers (Silady 2018). Since it is not feasible to develop for a bank to develop deep financial relationships with such a large number of individual customers, a data-driven approach to relationship management is a necessity.

We expect most retail bank customers to hold a checking account at that institution, and a subset of those customers may also avail the bank's other product offerings such as credit cards or mortgages. From the bank's perspective, customer deposits in checking accounts may be lent out by the bank to other counterparties in the form of interest-bearing loans. The bank earns the spread between the rate it lends money and the rate it pays for customer deposits. Since checking accounts are typically not interest-bearing or have a low interest rate, the bank earns a positive spread. However, the bank is exposed to liquidity risk since customers are given the option to withdraw their deposits at any time. In addition to general operational costs (such as technology support, middle and back office operations, and customer service), this exposure reduces the bank's risk-adjusted spread. Thus, in the current low-interest rate environment, the ability to maintain and increase profits purely through the lending of customer deposits may be limited.

An option for increasing spreads, and, therefore, profit, is increase the rate the bank lends to its customers. While economic factors prevent banks from offering loans at non-competitive interest rates, different products can charge vastly different rates. For example, mortgage rates in May 2019 were 4.14% (Orton 2019). On the other hand, the average credit card interest rate in early 2019 was 19.24% APR, which is a large spread over other categories of consumer lending (McCann 2019). Thus, it would be beneficial for a retail bank to have more credit card customers in its lending portfolio relative to other loan categories. Furthermore, credit cards may also charge annual fees, which provides an additional revenue stream from a single product and single customer.

With this motivation, it becomes clearer why data analytics should be used to attract additional credit card customers. While one source new credit card holders could be those who do not currently have accounts at the bank, it would first be prudent to cross-sell credit cards as a product offering to those who are already customers of the bank but do not hold credit cards. According to industry surveys, selling products to new customers is 68% more expensive than cross-selling to existing customers, the sales cycle for an existing account holder can be 25% shorter than the cycle for a new account holder, and only about 5-30% of revenue normally comes from initial sales (Hellmann 2018). Based on this research, targeting existing bank customers is likely to be a more effective path to increasing the number of credit card account holders.

Since we have institutional data on existing customers - and no or little structured data on non-customers - it may be possible to develop a model which can help target special offers to our customer base. These offers should be designed to entice these customers who currently do not hold a credit card to open a credit card account. It is important to further consider the types of offers that should be made to customers. Not only will such decisions inform the underlying assumptions of the model and its development, but also may have a material impact on longer term financial returns which should be considered carefully.

One offer we can consider to attract credit card customers is an introductory 0% APR for a period one year. This essentially provides the customer interest-free cash forwards. From the perspective of a utilitarian profit-maximizing rational agent, this incentive should always be sufficient for a customer to open a new credit card account and use that credit card for, at a bare minimum, that one year offer period. While in reality, there may be practical constraints which limit customers who actually accept the offer, for the purposes of our model and its application, it is still a reasonable assumption that most customers will take the “free” offer. This is an important assumption since as we develop model, we do not need to predict or estimate the efficacy of the offer. Instead, the model can focus on predicting the customers who will continue to hold the credit card once the offer expires. As we shall see, this better corresponds to the data set we have available so the assumption reinforces consistency between the model inputs and model outputs.

It should be noted that one of the risks of this strategy is offering the free trial period to customers who may have opened a credit card on their own, but from the bank’s perspective, the cost of targeting a customer is very low relative to gains (the expected present value of a customer’s interest fees for a single year is negligible compared to that same customer’s compounding fees over their lifetime). Thus, under such simplifying assumptions, it may be reasonable to deploy a marketing campaign using a model which can only predict probabilities without intervention.

Another customer targeting strategy which should be analyzed is offering a bonus credit of some amount if the customer spend on the credit card exceeds some threshold over a period of time. While further analysis could be conducted to better calibrate these amounts and thresholds (and potentially calibrating differently for different customers), for the sake limiting the model parameter space, we will set this offer to a bonus $200 credit if the customer spends $1,000 in the first three months. If coupled with the 0% APR offer described above, we can still employ the 100% efficacy assumption (that is, all customers who receive the offer will open a credit card account) while allowing for different cost scenarios to be considered. While we do not expect any of these assumptions to have a direct impact on the underlying model machine learning itself, the various cost, revenue and efficacy assumptions we make will a significant impact on the deployment of the offer strategy, and this additional scenario will be useful in our analysis.

In addition to analysis on customer credit card data, we can also examine the data to model customer churn behavior. This analytics model will help identify current customers that are most at risk of leaving the bank, which could be tremendously beneficial. A loss of a customer means not only a loss of their deposits, but also losses from other products and services they may potentially use in the future. For example, a customer is more likely to have a credit card at a bank where he or she has a checking account. This may be because a bank is more easily able to target special offers to customers who already have accounts and because the customer gains benefits from synergies between the two products (for example, a credit card reward program could provide higher cash back rewards if deposited directly to the customer’s checking account at the same bank). Therefore, the closure of a checking account implies a reduction in the bank’s other services.

There are many applications of a customer churn model such as targeting retention programs or offers or re-focusing marketing for new customers to ones less likely to exit. For the purposes of this project, we do not intend to evaluate the full range of business applications for this model, but instead consider in context of the credit card analysis. One use case of this is by overlaying the customer churn model with the credit card model, we can minimize the cost of targeting credit card offers to customers who are likely to exit anyway. From a business context, at this point such customers should be targeted for a more focused retention offer, but we consider that to be outside of the scope of this project. Still, if the churn model has sufficient explanatory power, we can improve the deployment of the credit card model and the customers who are targeted.

The objectives and strategies outlined in this section forms the foundation of the analysis performed. The assumptions derived from the business problem directly inform the modeling choices which need to be made. This is must be considered in conjunction with the available customer data. Therefore, a deeper review and discussion of this data is necessary before development of the model.

3. Data Understanding

The data set used for this analysis is called “Churn Prediction of bank customers” and has been downloaded from Kaggle (Dasgupta 2018). Although the name indicates that the primary intent of the uploader was for users to build predictive churn models for the bank, we also believe there is sufficient data available for additional meaningful insights; specifically, the attributes included should allow us to estimate the likelihood of a given customer having a credit card at the bank. The data set contains information on 10,000 accounts for customers at the bank. Included are both attributes about the account itself (such as account balance and tenure) as well as attributes about the customer who is the primary account holder (such as age, gender, geography, and salary).

The Kaggle data set page does not provide further details on the banking institution itself, but based on the data contained with, we can assume that it is a smaller European retail bank which provides basic consumer baking services such as checking accounts and credit cards. These assumptions based on the data help inform the assumptions and construct behind the business problem statement described in the previous section. The fields in the data set also drive and define the types of predictive models that can be developed, which also further refines the business problem statement.

In order to draw the connection between the data, the business problem, and model development, we must first gain a more thorough understanding on the data attributes available. The names of the fields contained within the data set are listed below along with a short description of each as it applies to the modeling problem.

- Customer ID:Unique identifier for each customer. This field does not contain any information pertaining to the customer or his/her account, so it is not likely to be pertinent to the final model.
- Surname:Last name of the account holder. While this is a descriptive attribute of the customer, it is unlikely for this field to have any bearing on the likelihood to hold a credit card. Therefore, it is not likely to be pertinent to the final model.
- CreditScore: Represents the credit score of the primary account holder. Although the data set does not specify the source or the agency providing these scores, the range of values in the data indicate scores can be between a low of 350 and high of 850. The

higher the credit score, the higher the creditworthiness of the customer. Individuals may have low credit scores either because they have not accumulated sufficient credit history or they have made late interest payments. While it may not be clear what the relationship between credit score and likelihood to have a credit card is, we should consider it as a feature to be used as a predictor in the model.

- Geography: Geographical residence of the primary account holder. The data seems to indicate that the bank is a European bank, with customers exclusively in Germany, Spain and France. Approximately 50% of the current customers are in France, while about 25% are in Spain and Germany each. While there is no clear indication that customers from one country are more likely to have a credit card that customers of another, the bank's prevalence in different markets could have an impact. In future investigations, this field may even help identify if any geographical market is underserved by the bank, but for the purposes of this analysis, we can benefit from any regional differences that may exist.
- Gender:Gender of the primary account holder, either male or female. While there is no obvious reason why one gender would be more likely to hold a credit card at the bank than another, this should be a feature of the model in case some sort of relationship does indeed exist.
- Age: Age of the primary account holder. Ages for current customers range from 18 to 92, with an mean of 38.9 and standard deviation of 10.5 years. It seems likely that there is a relationship between age and the likelihood to hold a credit card, so this feature should be included as a variable in the model. One possible hypothesis is that younger customers are less likely to have a credit card since they simply may not have yet opened one, while an older customer is more likely to have opened a credit card at the bank over his/her life.
- Tenure:Length of time in years the account holder has held this account. The maximum tenure in the dataset is 10, so we assume that any accounts with longer tenure have been bucketed into the 10 year category. The mean tenure is approximately 5 years and standard deviation is 2.9. Similar to age, it seems like a reasonable hypothesis that a customer who has held an account for a longer period of time is more likely to hold a credit card at the bank. Even if this hypothesis does not hold true, some sort of relationship is likely, so this feature should be included as a predictor variable used in the model.

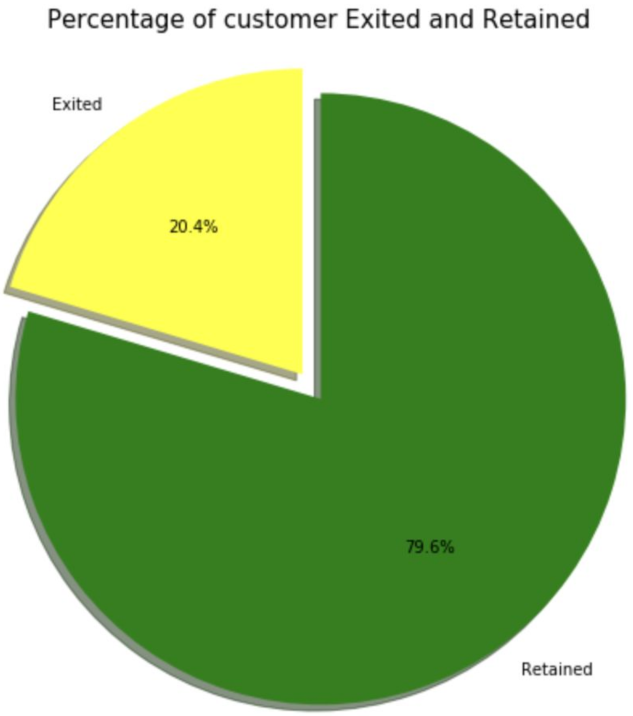
- Balance: Account balances denominated in an unspecified currency. Balances range from 0 to about 250k, with a mean of 76.4k and a standard deviation of 62.4k. The account balance may be an indicator of financial sophistication of the account holder which may have a relationship with likeliness to hold a credit card, but there are numerous other confounding variables. For example, a sophisticated account holder may have multiple other bank accounts, so the deposits held at this bank could be low. Regardless, since some sort of relationship is feasible, the account balance is another feature which should be included in the model.
- NumOfProducts: Denotes the number of financial products or services availed by the primary account holder. Although the full offerings of the bank are not clear, such products could include mortgages, automobile loans, and Certificates of Deposit. A credit card could also be considered as a product offered by the bank, but as shown below, there is a separate field in the data which flags customers who have credit cards, so we assume that this field does not include credit cards. The minimum value of the field is 1 so we can also assume that the checking account itself is counted as a product. The maximum value is 4, so it is likely that no more than 4 products are offered by the bank. The mean is 1.53 products and the standard deviation is 0.58. Since the bank only has four product offerings, it appears to be a less sophisticated institution when comparing to some other large retail banks. With that said, a possible hypothesis is that an account holder with many products at a bank may be more likely to have a credit card as well. Because of this possible relationship or any other potential relationship, the number of products should be a feature used in the model.
- HasCrCard: Flag which indicates whether the primary account holder has a credit card at the bank or not. This is the target variable for the credit card model. About 70.6% of account holders currently have a credit card.
- IsActiveMember: Flag which indicates whether the primary account holder is active or not. About 51.5% of accounts are currently denoted as active. It is not immediately clear how this impacts probability of having a credit card, it it is possible that there is some relationship, so further investigation is needed to understand if this feature should be considered as a predictor variable.
- EstimatedSalary: Approximate annual salary of the primary account holder, denominated in an unspecified currency. Salaries range from about 0 to 200k, with an average of 100k and standard deviation of 57.5k. The data set does not detail how these salary figures

were derived (for example, through disclosure from the account holder), but, if accurate, the field may indicate the financial sophistication of the customer more accurately than account balance. As mentioned above, an individual can split cash holdings across different bank accounts which could obfuscate his or her actual financial worth, but salary is an indivisible figure, and could be a stronger indicator of financial sophistication. If the hypothesis described previously where financial sophistication increases the probability of having a credit card, estimated salary is a feature that should be included in the model design.

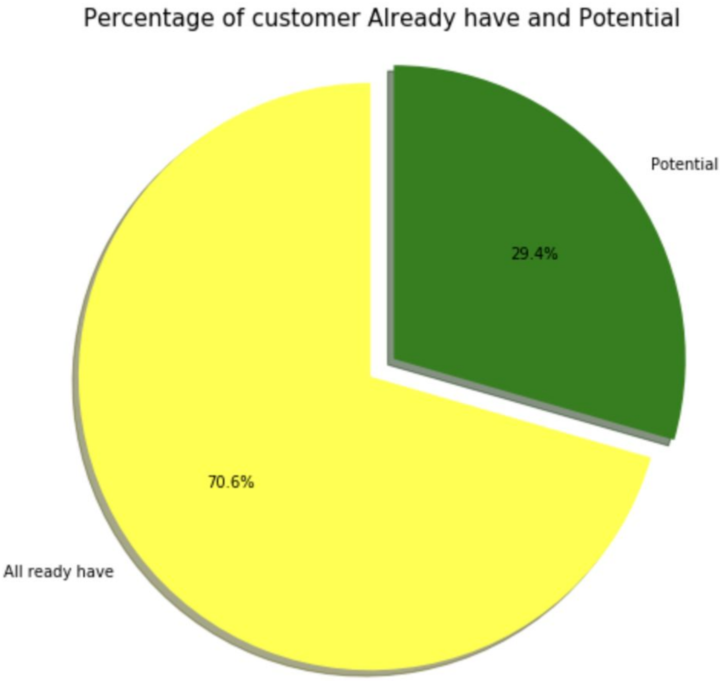
- Exited: Flag indicating whether the primary account holder has closed out his or her account at a given point in time. This is the target variable for the customer churn model. Currently, about 20% of the account holders in this data set have exited.

It is important to note that, in the context of the business problem, the bank customer data available can be used to construct a predictive model which determines the likelihood of a bank account holder also having a credit card account with the bank. Such a model is informative from a marketing standpoint since it can forecast customers most likely to have a credit card *without* intervention (that is, targeted offers like 0% APR and bonus credits described in the previous section). While an ideal model would be one which can predict likelihood both conditional with intervention and conditional without intervention, we currently do not have sufficient data. For a more thorough analysis, the bank could make an investment in obtaining such data by offering incentives to select random customers to develop training data on the likelihood of opening a credit card account after invention. This could be extended even further by using different types of incentive schemes or incentive schemes parameterized with different values and thresholds.

Before we move on to next phase, let's get some basic insights from our data. We have visualized data to get some more insightful understanding. Pie chart 1 below demonstrates that 20.4% customers have churned. So our model needs to predict that 20% of the customers will churn. Given 20% is a small number, we need to ensure that the chosen model does predict this with great accuracy as it is of interest to the bank to identify and keep this population as opposed to accurately predicting the customers that are retained. Pie chart 2 below demonstrates that potential customer for credit card account are coming as 29.4%.



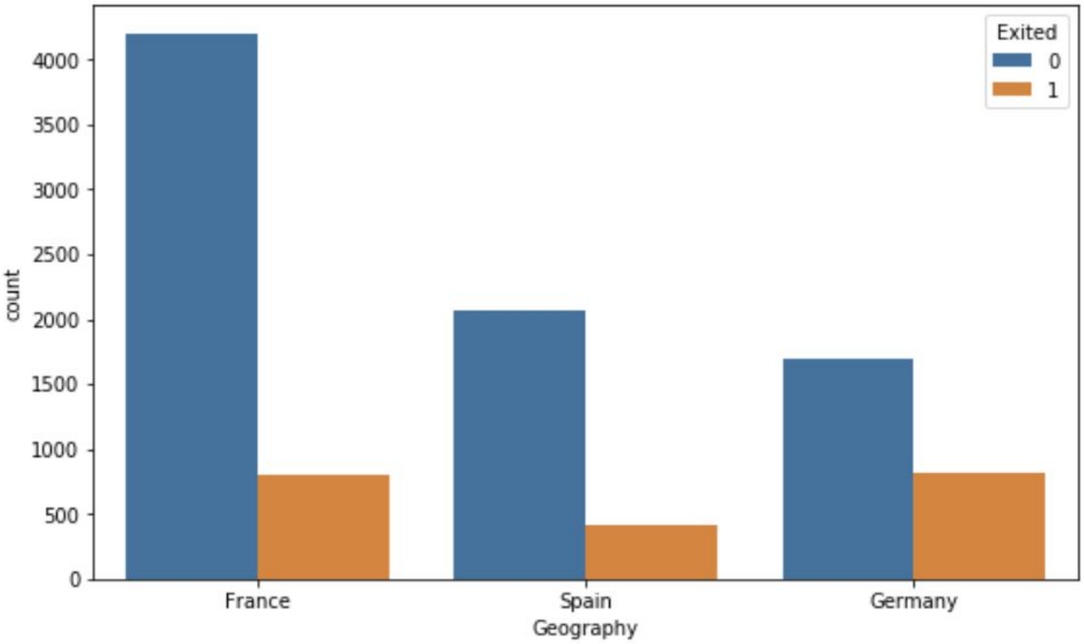
Pie Chart 1



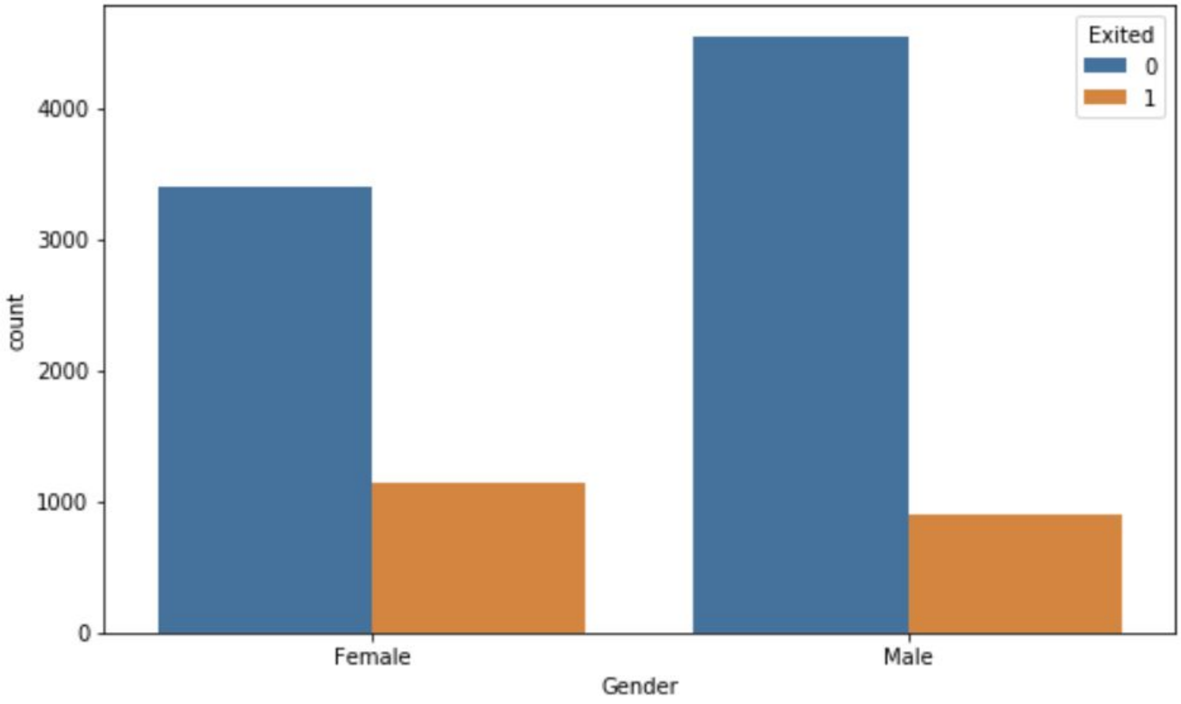
Pie Chart 2

Furthermore when we dig inside the data we got some more clarity, when we started analysing the columns and how they are affected with our target variable i.e. “Exited” for churn prediction.

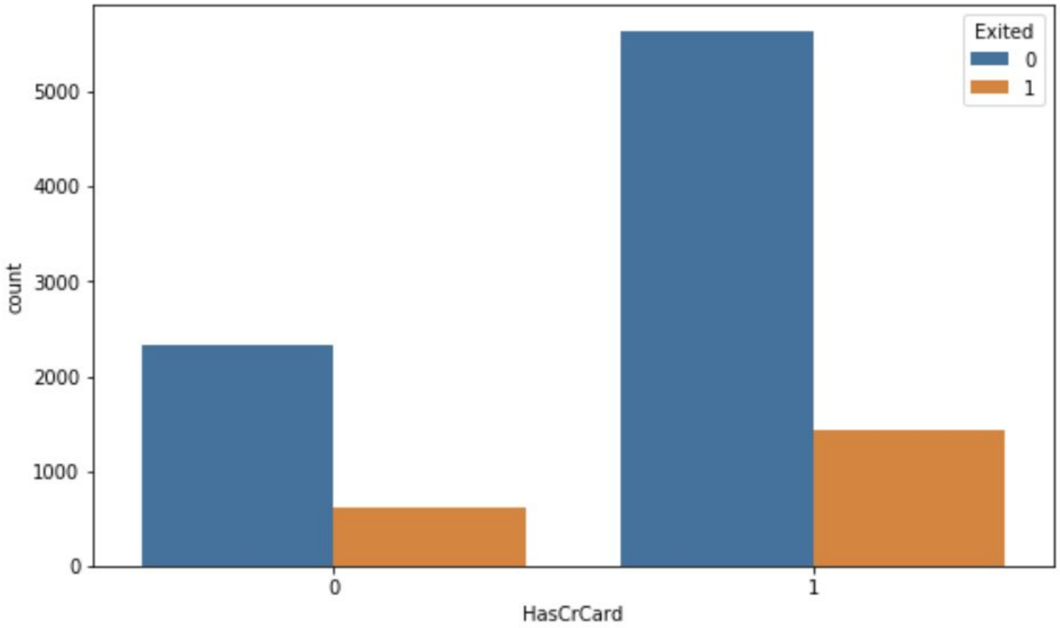
1. **<u>Geography:</u>** Maximum data is of people from France. But if we look at the ratio of Retained vs Churned, results are very different. However, the proportion of churned customers is with inversely related to the population of customers alluding to the bank possibly having a problem (maybe not enough customer service resources allocated) in the areas where it has fewer clients.



2. **<u>Gender:</u>** Female customer are churning in more proportion than that of male customers.

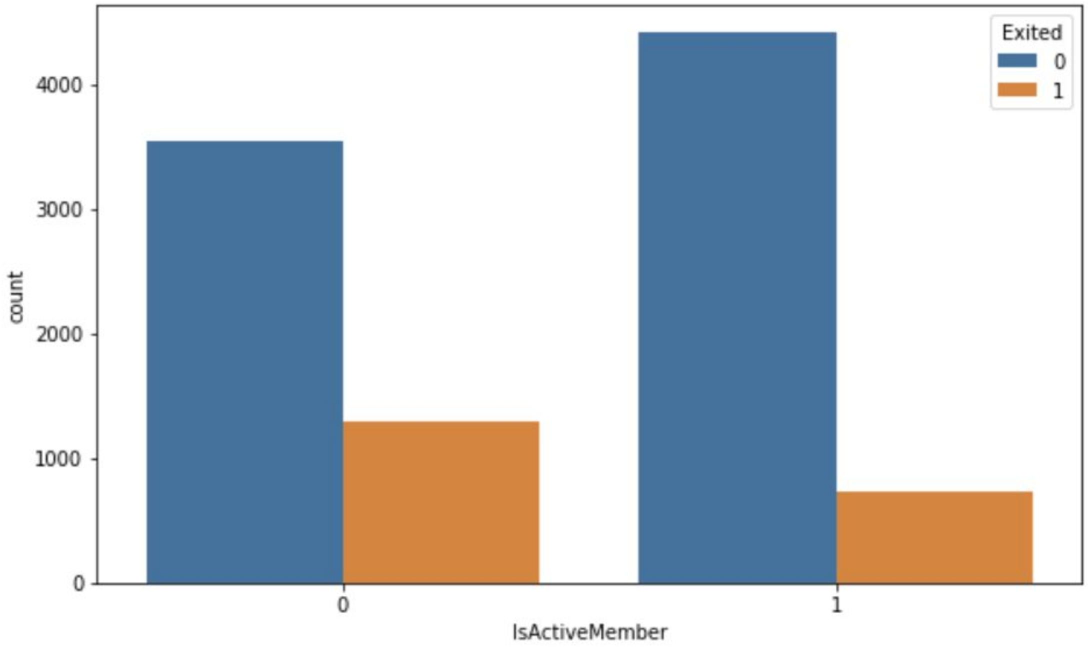


3. **HasCrCard:** Credit Card account holder customers are churning more, given that majority of the customers have credit cards could prove this to be just a coincidence.



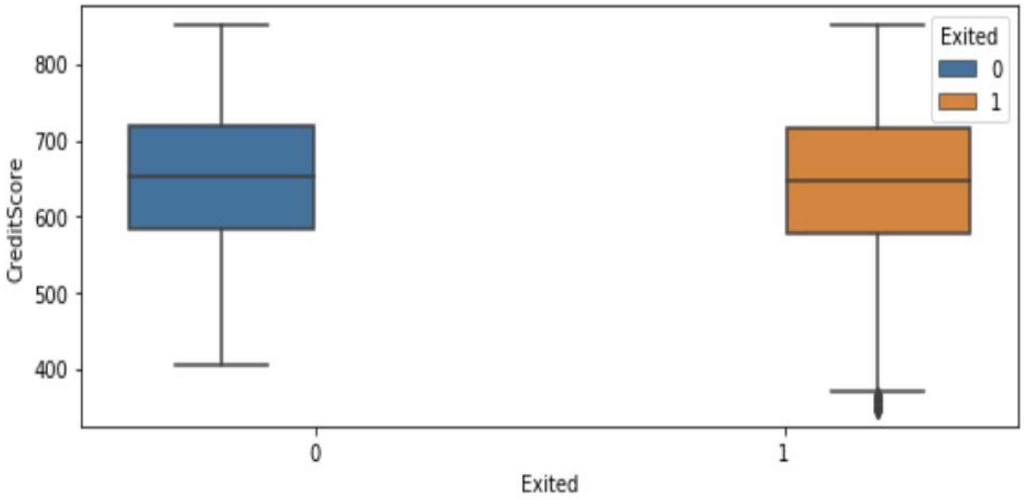
4. **IsActiveMember:** Inactive members have a greater churn portion. Overall proportion of inactive members is quite high suggesting that the bank may need a program implemented

to turn this group to active customers as this will definitely have a positive impact on the customer churn.

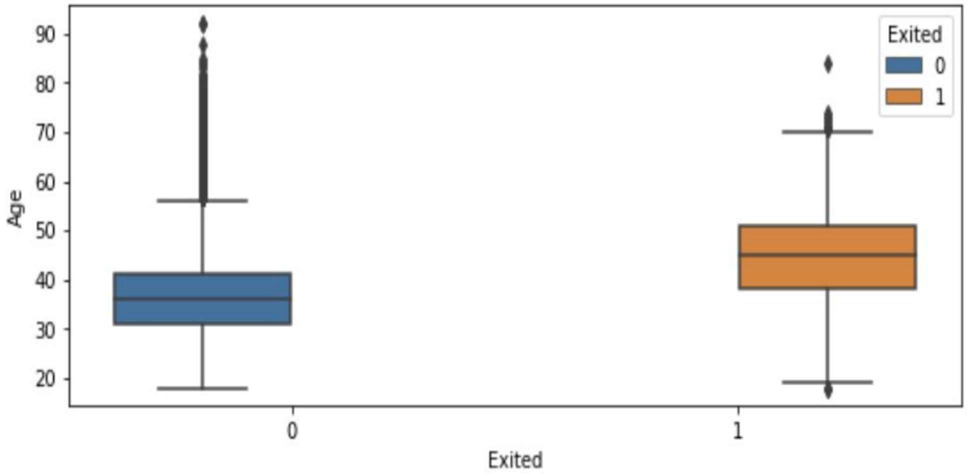


Now let's get some more insight from our continuous attributes:

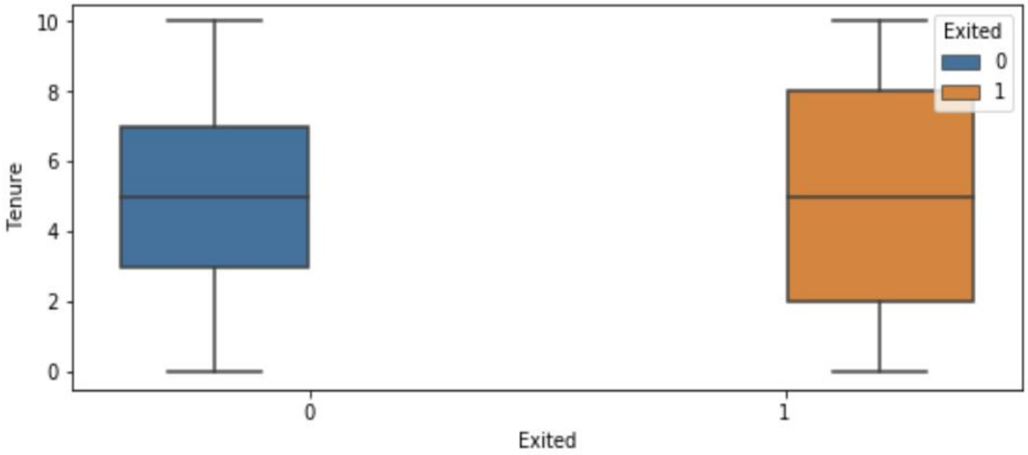
- Credit score distribution does not show any vital changes .



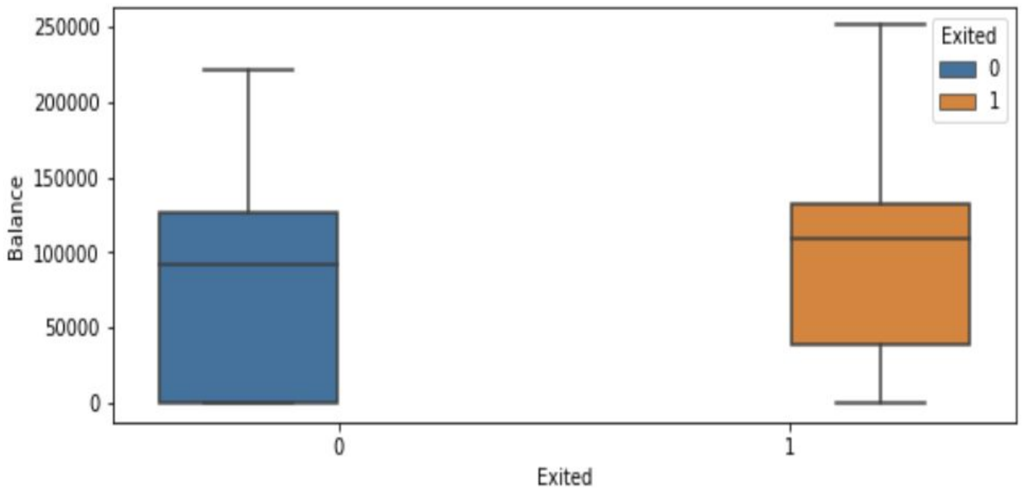
- Older customers are churning at more than the younger ones in the age categories.



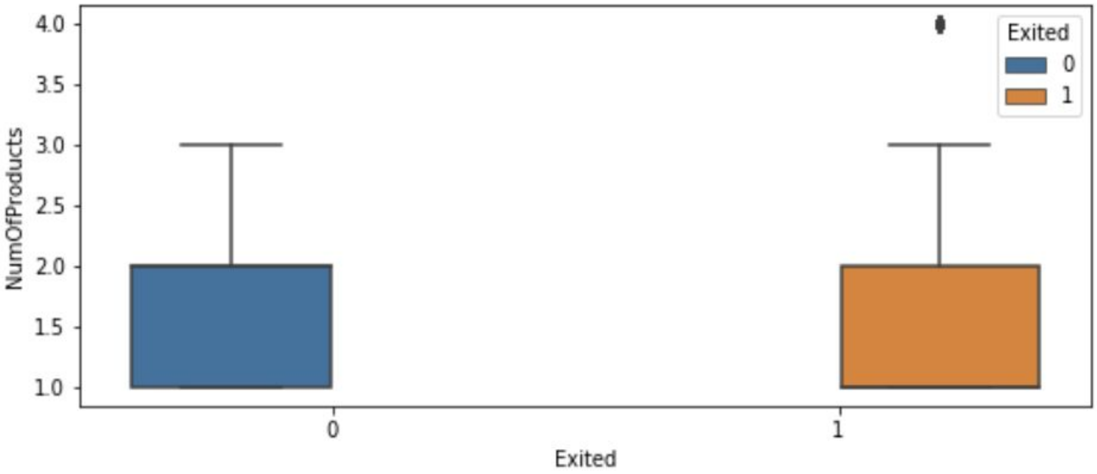
- Extreme tenure clients are churning more i.e. very new or very old with respect to tenure.

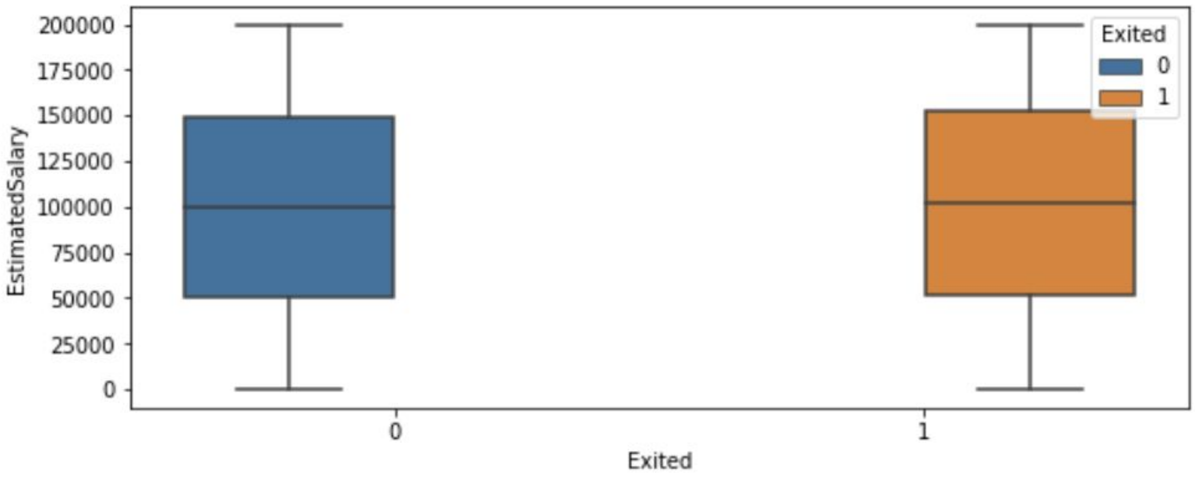


- Losing customers with good bank balances will affect more as it will limit the capital for loans and lending.



- Attributes like Product and Salary are also not showing any significant variations.



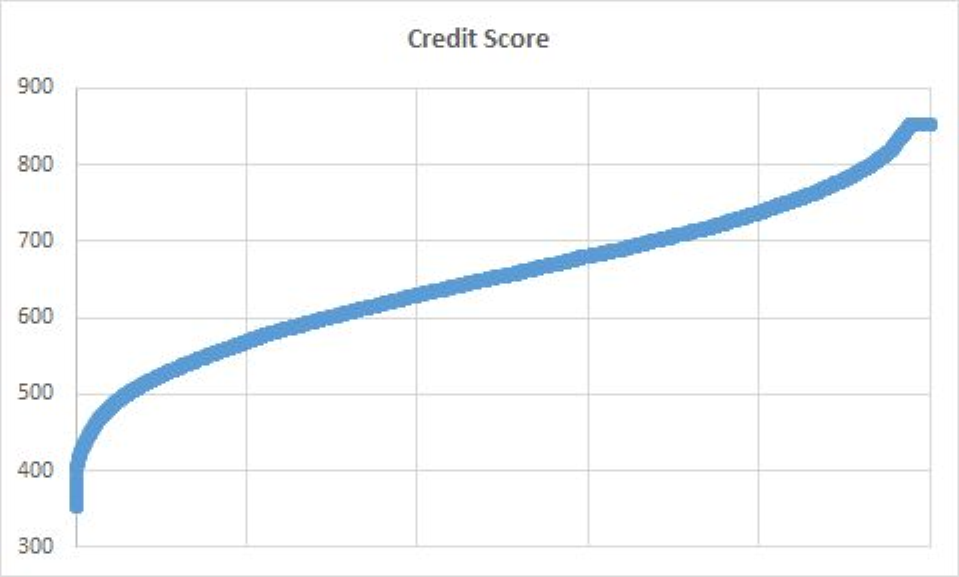


4. Data Preparation

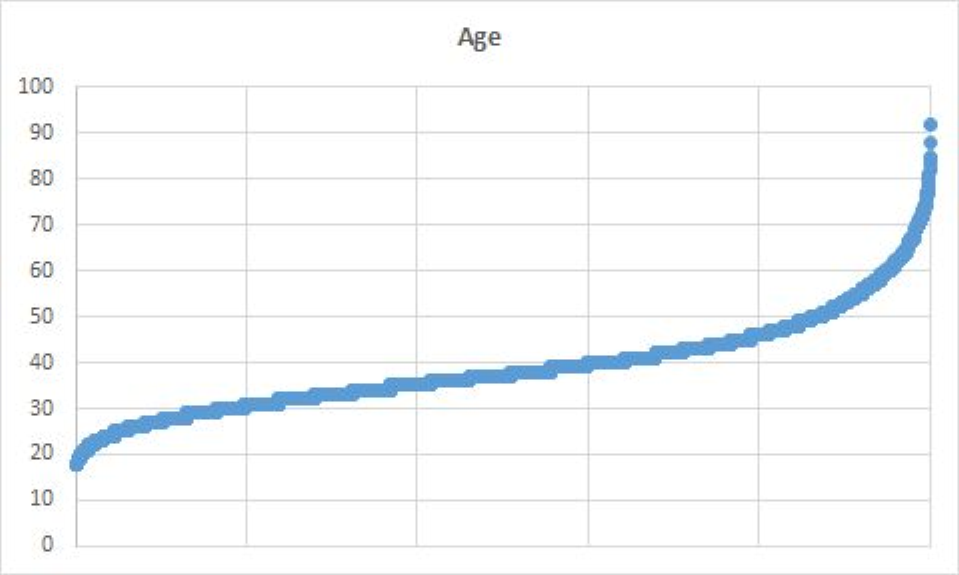
The quality of the “Churn Prediction of bank customers” data set in general appears to be in good condition, so the steps involved in preparation of the data for model training are straightforward and minimal. Notably, there are 10,000 customer accounts in the data set, and none of them appear to be missing values. Thus, data preparation tasks fall into one of three varieties: conversion of categorical variables to numeric; dropping data columns which do not represent features relevant for modeling; and transformation of numerical values (for example, to remove outliers or change distribution) where applicable.

Descriptions of each data field are provided in the Data Understanding section above. The following bullets describe what preparation actions, if any, have been performed on each data field.

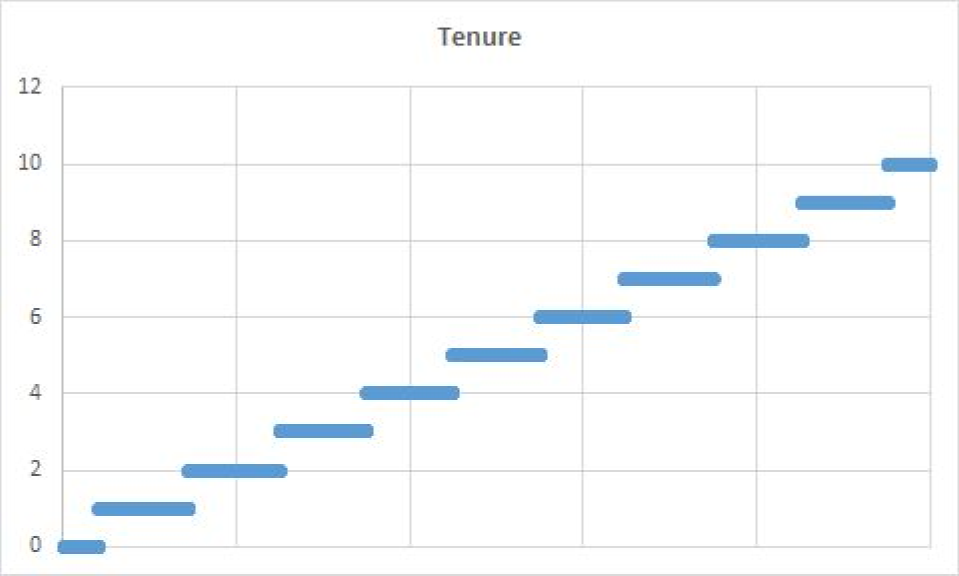
- Customer ID:This column is not a feature relevant to the model, so it can be dropped for training and analysis purposes.
- Surname:This column is not a feature relevant to the model, so it can be dropped for training and analysis purposes.
- CreditScore: This column has numeric values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



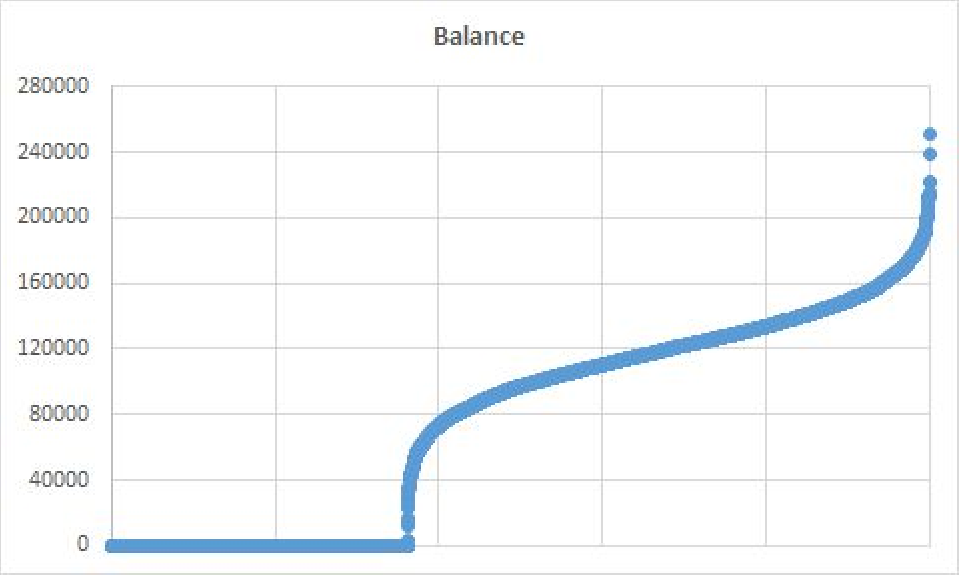
- Geography: This column has categorical values. In the data set, there are only three possible values represented (France, Germany, Spain). Therefore, the column needs to be converted into three different columns with boolean values, with a value of 1 in the column which represents a given account holder's country. Once this transformation is performed, the original categorical value column can be dropped from the data. New columns are "Is_Spain", "Is_France" and "Is_Germany"
- Gender: This column has categorical values. In the data set, there are two possible values represented (Male, Female). Therefore, the column needs to be converted into into a column with boolean values, with a value of 1 for females and 0 for males. Once this transformation is performed, the original categorical value column can be dropped from the data.
- Age: This column has numeric values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



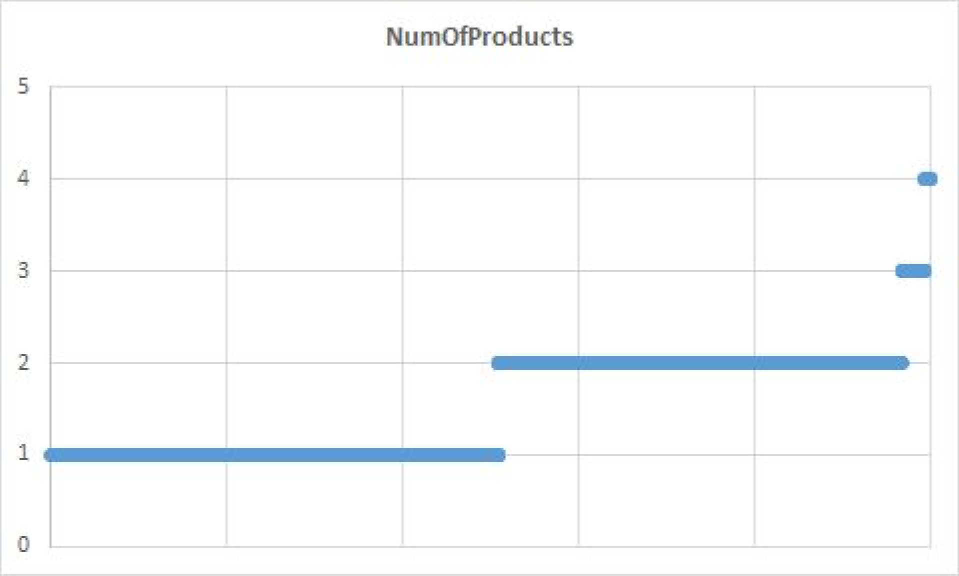
- Tenure: This column has numeric integer values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



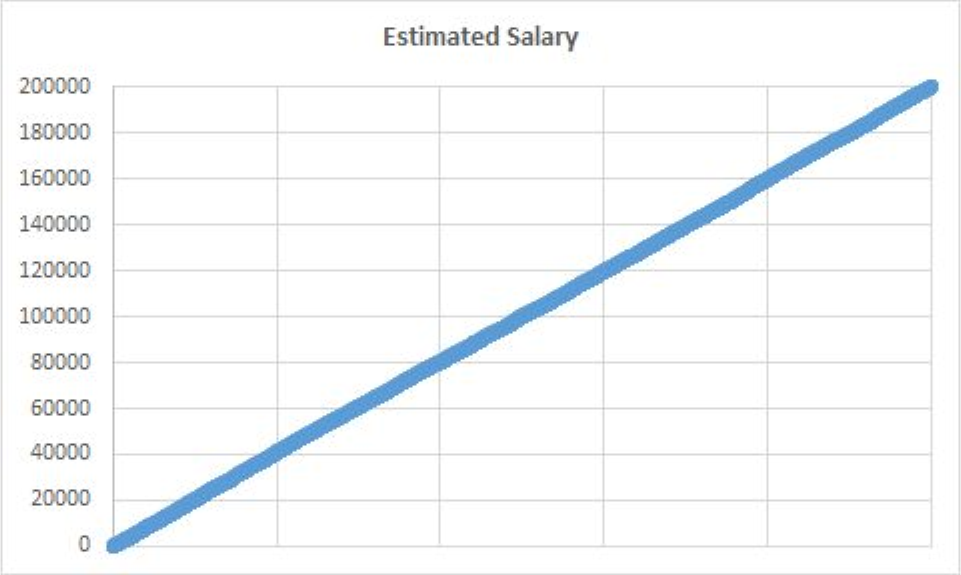
- Balance: This column has numeric values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



- NumOfProducts: This column has numeric integer values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



- HasCrCard: This column has boolean 1 or 0 values. No data preparation or transformation is necessary.
- IsActiveMember: This column has boolean 1 or 0 values. No data preparation or transformation is necessary.
- EstimatedSalary: This column has numeric values. As shown in the plot of ranked values below, there do not appear to be any outliers or abnormalities. Therefore, no data preparation or transformation is necessary.



- Exited: This column is a boolean 1 or 0 value. No data preparation or transformation is necessary.

As anticipated, data preparation and transformations have been minimal since the original data set was in generally good condition. In order to build our machine learning models, we must divide the data into a training set for model development and a testing set for model validation. Using the `train_test_split` function from the `model_selection` package within the Scikit-learn Python machine learning library, we can easily split the data set to use 80% of the total observations (8,000 observations) as training data and the remaining 20% (2,000 observations) as testing data. Using this training data prepared as described in this section, as well as the business problem and underlying data described in the previous two sections, we can now begin developing our analytics models.

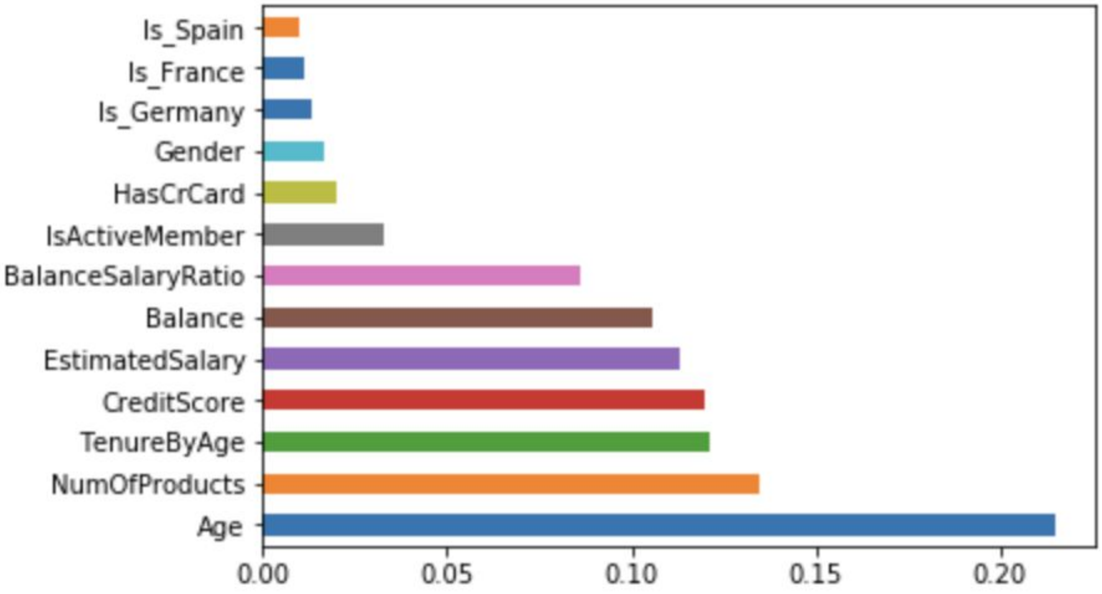
Apart from this we also performed feature engineering where we added features that are likely to impact probability of churning as we have seen above in initial analytics part. Balance, Salary, Tenure and Age are creating impact on churning; thus we created new features using them**.**

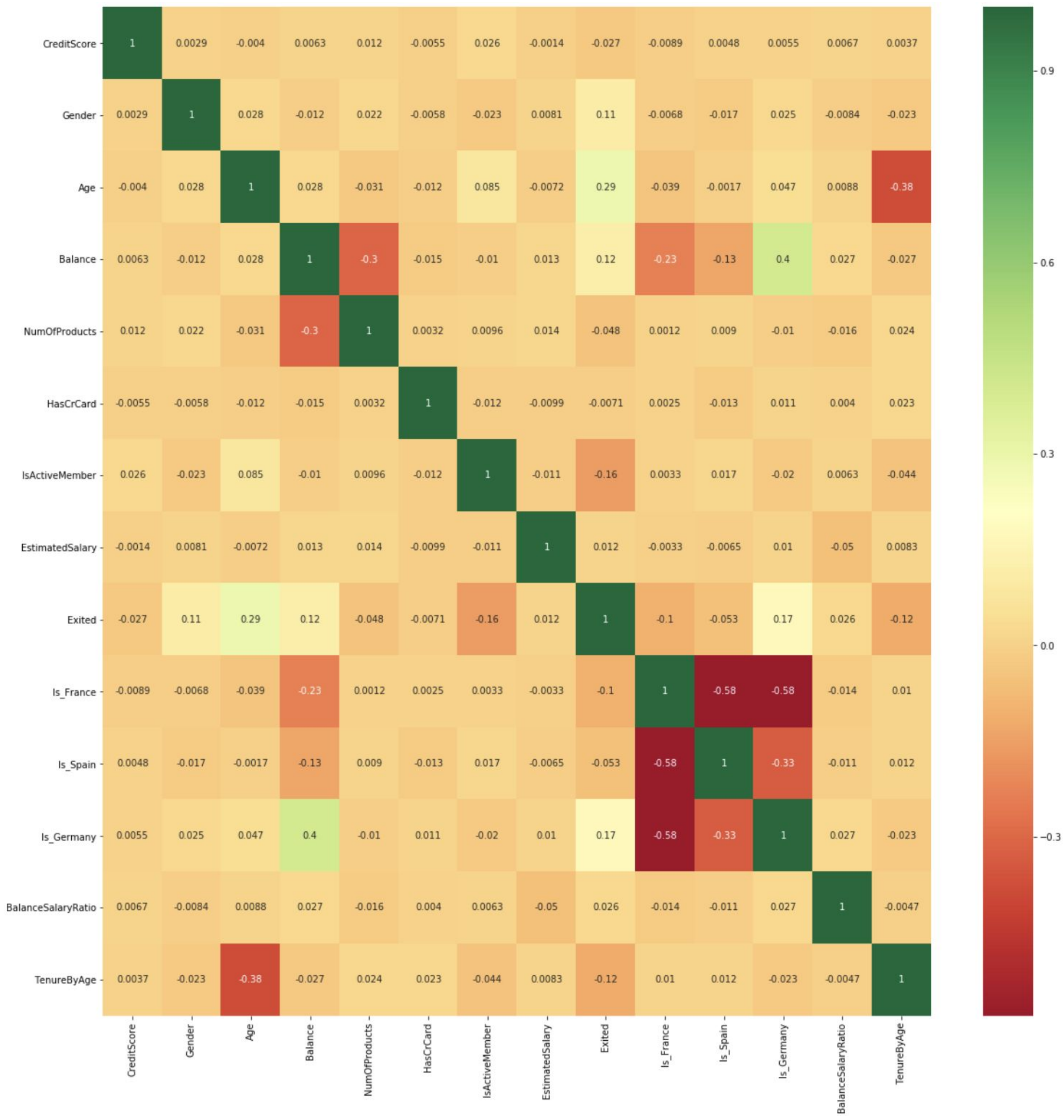
- **BalanceSalaryRatio:** In this feature we divided balance with estimated salary column to come up with this feature. Ratio of the bank balance and the estimated salary indicates that customers with a higher balance salary ratio churn more which would be worrying to the bank as this impacts their source of loan capital.

- **TenureByAge:** As tenure is function of age we introduced this variable to standardize tenure over age. In this feature we divided Tenure with Age column to come up with this feature.

5. Modeling

We have developed a predictive supervised learning model to forecast customer conversion probability. We consider various modeling approaches including Random Forest, Decision Tree and Support Vector Machine. We first derived initial accuracy of all the models but the results were not meeting our expectation, so we decided to build correlation matrix with heat map which helped us to find how features are related to each other and with target variable. Then on the basis of that we removed variables which were not having significant relation with target variable. We also implemented feature selection process using Extra Tree classifier model which provided us with the important feature.





	CreditScore	Gender	Age	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Is_France	Is_Spain	Is_Germany	BalanceSalaryRatio	TenureByAge
CreditScore	1	0.0029	-0.004	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027	-0.0089	0.0048	0.0055	0.0067	0.0037
Gender	0.0029	1	0.028	-0.012	0.022	-0.0058	-0.023	0.0081	0.11	-0.0068	-0.017	0.025	-0.0084	-0.023
Age	-0.004	0.028	1	0.028	-0.031	-0.012	0.085	-0.0072	0.29	-0.039	-0.0017	0.047	0.0088	-0.38
Balance	0.0063	-0.012	0.028	1	-0.3	-0.015	-0.01	0.013	0.12	-0.23	-0.13	0.4	0.027	-0.027
NumOfProducts	0.012	0.022	-0.031	-0.3	1	0.0032	0.0096	0.014	-0.048	0.0012	0.009	-0.01	-0.016	0.024
HasCrCard	-0.0055	-0.0058	-0.012	-0.015	0.0032	1	-0.012	-0.0099	-0.0071	0.0025	-0.013	0.011	0.004	0.023
IsActiveMember	0.026	-0.023	0.085	-0.01	0.0096	-0.012	1	-0.011	-0.16	0.0033	0.017	-0.02	0.0063	-0.044
EstimatedSalary	-0.0014	0.0081	-0.0072	0.013	0.014	-0.0099	-0.011	1	0.012	-0.0033	-0.0065	0.01	-0.05	0.0083
Exited	-0.027	0.11	0.29	0.12	-0.048	-0.0071	-0.16	0.012	1	-0.1	-0.053	0.17	0.026	-0.12
Is_France	-0.0089	-0.0068	-0.039	-0.23	0.0012	0.0025	0.0033	-0.0033	-0.1	1	-0.58	-0.58	-0.014	0.01
Is_Spain	0.0048	-0.017	-0.0017	-0.13	0.009	-0.013	0.017	-0.0065	-0.053	-0.58	1	-0.33	-0.011	0.012
Is_Germany	0.0055	0.025	0.047	0.4	-0.01	0.011	-0.02	0.01	0.17	-0.58	-0.33	1	0.027	-0.023
BalanceSalaryRatio	0.0067	-0.0084	0.0088	0.027	-0.016	0.004	0.0063	-0.05	0.026	-0.014	-0.011	0.027	1	-0.0047
TenureByAge	0.0037	-0.023	-0.38	-0.027	0.024	0.023	-0.044	0.0083	-0.12	0.01	0.012	-0.023	-0.0047	1

After selecting important features and features with significant relation with target variable, we again tried all the model and did hyper parameter tuning to rectify problem of choosing a set of optimal hyper parameters (parameter whose value is used to control the learning process) to get the best accuracy for our models.

Model	Train Accuracy	Test Accuracy
Decision Tree	89.725%	81.60%
Support Vector Machine	91.1%	83.4%
Random Forest	91.3	85.3

For the second part of the CRM, where we are trying to find the potential customer for credit card account, we split the data into train and test set by ensuring that the labels are equally distributed in train and test set. Our model is trained on the training data by adopting a cross validation approach where “HasCrCard” column is our Target variable and our trained model is tested on our test set to evaluate our models generalization ability. For this use case we used Decision tree classifier and our accuracy of our model on the test set was 82.7%.

6. Evaluation

Model evaluation will be contingent on business criteria. For our understanding of model effectiveness, we have applied techniques such as ROC/AUC analysis and confusion matrix. However, for communication to business stakeholders, we may also consider more intuitive measures of effectiveness such as a Cumulative Response Curve for future analysis.

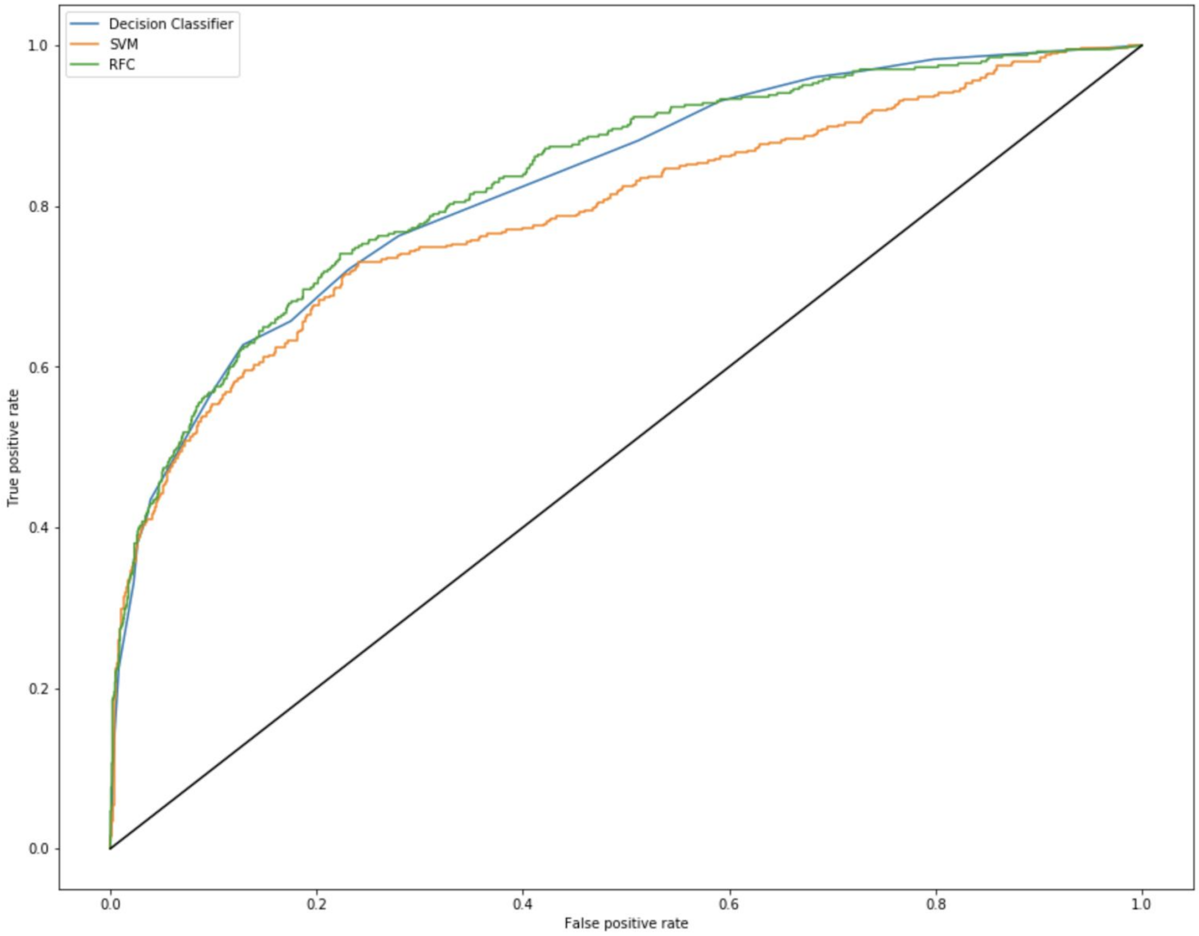
Based on our ROC/AUC curve and recall values we went ahead with Decision Tree classifier as our Primary model. Since, from a business perspective we are more interested in capturing people who are likely to churn, we decided our final model based on how well our model is able to identify the churn customers.

Actual/Predicted	Exited	Not Exited
Exited	183	146
Not Exited	222	1449

Precision= 90.84%

Recall=45.1%

AUC-ROC Curve



7. Deployment

With the final machine learning models developed, trained, and validated, it is now possible to deploy them to address the business problem introduced in the Business Understanding section of this paper. In order to understand how the credit card model can be used most effectively, let us first consider the value proposition of targeting potential credit card customers in the context of the expected value framework.

We can express the value of targeting, VT, as the expected benefit of targeting customer x, $EB_T(x)$ and the expected benefit of not targeting customer x, $EB_{notT}(x)$:

$$VT = EB_T(x) - EB_{notT}(x)$$
$$VT = [p(CC|x,T) - p(CC|x,notT)]u_{CC}(x) - c(x)$$

where $p(CC|x,T)$ is the probability of customer x continuing to use the credit card after being targeted, $p(CC|x,notT)$ is the probability of customer x continuing to use the credit card after not being targeted, $u_{CC}(x)$ is the credit card profit from customer x, and $c(x)$ is the customer targeting cost.

Based on the discussion the Business Understanding section, $p(CC|x,T)$ can be approximated as 1 since we assume that all customers will use the card if given the introductory 0% APR offer. Thus, the value of targeting can be reduced to:

$$VT = [1 - p(CC|x,notT)]u_{CC}(x) - c(x)$$

This implies that the model developed through the data set, which predicts the likelihood of obtaining a credit card without targeting, is sufficient under this construct to obtain the value of targeting. It should be noted that this strategy (under constant profits and costs) will implicitly target customers who are *least* likely to have credit cards without intervention. This is consistent with our business thesis and objectives since the bank will avoid incurring the cost of forgone interest payment receipts from customers who are likely to independently open credit card accounts at some point in the future.

The profit function, $u_{CC}(x)$, is the difference between the bank's cost of funding the customer's credit card balance and the revenue generated by the customer's interest payments over some window after one year to t years. For the sake reducing complexity, we can represent this simply as the spread between the bank's funding rate and the customer's interest rate. Normally, financial institutions use sophisticated models to construct forward funding and interest rate curves, but using a simple approximation should be sufficient for our purposes. We can gain much of the utility of the model using a static spread estimate on a representative balance amount.

The Effective Federal Funds Rate (EFFR) is a rate published daily by the Federal Reserve which represents the rate of overnight (short-term) federal funds transactions. On May 9, 2019, this rate 2.38% ("Federal Funds Data" 2019). As stated previously, the average credit card APR is 19.24%. The spread between these rates is 16.86%; ignoring the mechanics of term rates, the evolution of rates over time, and various other technical subtleties in interest rate products and markets, this spread is a sufficient proxy for the profit yield on deposits.

To estimate dollar value of profits, we must make an assumption on the average credit card balance held by the bank's customers. The average French citizen charges less than $300 to credit cards annually while the average German charges $152 ("How Does Our Credit Card Debt Compare to the Rest of the World?" 2012). Although we do not have data on Spanish credit card charging habits, we can take a weighted average (in our data set, there are twice as many accounts in France than in Germany) to obtain an estimate of $250. Ignoring compounding and the time value of money, at the spread derived above, this implies annual profit of $42.15. Note that this ignores customer prepayments of credit card debt (that is, in our scenario, the customer always pays only the minimum amount due). The average customer tenure in our data set is 5 years; if we assume that applies to credit card tenures as well, then over a 5 year period, total profit will be $168.6. Note that this does *not* include the cost of funding for the first year, which is reflected separately in the $c(x)$ value.

The value $c(x)$ is the expected cost to the bank for funding the customer's credit card balance over one year. Using the funding rate and average balance values described above, the cost of funding a $250 at 2.38% over one year (ignoring compounding) is approximately $5.95. Since a critical assumption to our modeling approach is that every customer that receives the 0% APR offer will accept it, that means the cost of each individual customer that is targeted is $5.95

Note that in obtaining these profit and cost values, a number of very significant assumptions have been made. It is possible to enhance our approach with more sophisticated methods, but we consider this to be out of scope for our current purposes. Some of these enhancements include: using more robust models of term funding curves and term interest rate curves (for example, credit card interest can be a spread on LIBOR or a risk--free benchmark rate), and applying the the appropriate rates to expected balances; applying different credit card balance assumptions for customers in different regions or even more sophisticated balance projections for different customers; addressing significant prepayment risk where customers pay more than the minimum interest rate payments; incorporating standard practices in valuation at financial institutions such as the time value of money and compounding; and forecasting the tenure of each customer rather than apply a static value.

We can use these estimates to construct the following cost matrix:

	Gets CC	Does Not Get CC
Targeted	$u_{CC}(x) - c(x) = \$162.65$	$- c(x) =- \$6.95$
Not Targeted	$u^*{}_{CC}(x) = \$212.50$	\$0

Note that $u^*{}_{CC}(x)$ here is higher than the \$168.60 amount discussed above because in the scenario where the non-targeted customer obtains a credit card, the bank will earn a spread on deposits for 5 years rather than 4 since no introductory APR offer is being made. Under the assumptions we have applied to the expected value framework, the cost matrix highlights that the cost of targeting a customer with the introductory offer is very low. This suggests that targeting all customers would be an effective strategy since we can guarantee a future yield on their deposits. However, there are other considerations that must be made. One important dimension not reflected above is time, since cash flows from costs and profits do not occur at once. It may be the case that the bank is not willing to forgo a large number of potential interest payments for an entire year - for such a year, the bank's profitability would fare very poorly, which may not be an admissible cost to management, the board of directors, or shareholders. Therefore, an initial cost constraint is necessary to effectively deploy the model using this framework.

As mentioned earlier, we can enhance the deployment of the credit card model with insights from the churn model. Revisiting the framework and cost matrix above, it is apparent that this discussion and analysis is true conditional on the customer *not* exiting the bank. Let us assume, however, that the customer does intend to exit, but the introductory credit card offer is sufficient to entice him or her to stay at the bank for one more year to take advantage of it. In this case, the bank would still incur the funding cost for that year without interest rate receipts, but once the customer leaves, the bank would no longer earn the spread on deposits in subsequent years. The cost matrix conditional on *the customer leaving after one year* is below:

	Gets CC	Does Not Get CC
Targeted	$- c(x) =- \$6.95$	$- c(x) =- \$6.95$
Not Targeted	$u^{**}{}_{CC}(x) = \$42.5$	\$0

Note that the value $u^{**}{}_{CC}(x)$ reflects the spread earned by the bank for a single year on the customer's deposits. Although we do populate a value for customers who are not targeted and get a credit card in the customer exit scenario, in reality we can consider the probability of this to be negligible since it is unlikely that the customer obtains more of the bank's services if he or she will be leaving soon. Thus, as can be easily seen in the matrix, the expected value of targeting such a customer in the context of credit card profits with the introductory offer is negative, and we should avoid doing so if we can. Therefore, the customer churn model can effectively filter out these customers to ensure that we limited our offers to customers who are likely to stay and be profitable in the future. It should be noted, however, that the customers we expect to exit should not be *ignored*. Rather, retention strategies for these customers need to be considered closely. However, for the purposes of this paper, we largely limit the scope and discussion of customers to the context of credit card accounts.

8. Conclusion

As shown through this paper, machine learning models can be powerful tool in Customer Relationship Management at retail banks and other institutions. Although we have focused our example on a single use case relatively limited in scope, these approaches can be translated to a wide range of other business problems in the context of CRM. Since retail banking institutions and organizations in other sectors interact with their customers through numerous channels, data mining through one such channel may reveal informative insights on another. In this case, we have seen how customer checking account data can be used to extend the bank's relationship with the customer into the credit card space. However, the number of permutations of product offerings we can consider is large, and grows larger with the complexity of the organization. With further research and analysis, it may be possible to extract even more meaningful insights from customer data which can benefit not only the entity serving the customers, but also improve the customer experience entirely.

References

Dasgupta, Sonali. 2018. “Churn Prediction of Bank Customers.” Kaggle. October 11, 2018. https://kaggle.com/sonalidasgupta95/churn-prediction-of-bank-customers.

“Federal Funds Data.” 2019. Federal Reserve Bank of New York. May 9, 2019. https://apps.newyorkfed.org/markets/autorates/fed%20funds.

Hellmann, Rebecca. 2018. “Are You Maximizing Your Primary Account Holders? Why Cross-Sell Matters.” BAI. August 23, 2018. https://www.bai.org/banking-strategies/article-detail/are-you-maximizing-your-primary-account-holders-why-cross-sell-matters.

“How Does Our Credit Card Debt Compare to the Rest of the World?” 2012. The Penny Hoarder. June 11, 2012. http://www.thepennyhoarder.com/smart-money/how-does-our-credit-card-debt-compare-to-the-rest-of-the-world/.

McCann, Adam. 2019. “What Is the Average Credit Card Interest Rate?” WalletHub. January 9, 2019. https://wallethub.com/edu/average-credit-card-interest-rate/50841/.

Orton, Kathy. 2019. “Mortgage Rates Pull Back on Weak Inflation Data.” Washington Post. The Washington Post. May 2, 2019. https://www.washingtonpost.com/business/2019/05/02/mortgage-rates-pull-back-weak-inflation-data/.

Silady, Alex. 2018. “The Top 10 Banks in America | SmartAsset.” SmartAsset. SmartAsset. November 20, 2018. https://smartasset.com/checking-account/the-top-ten-banks-by-assets-held.

Appendix 1: Team Contributions

Team Member	Contributions
Alsan Ali	Write-ups for: Introduction, Business Understanding, Data Understanding, Data Preparation, Deployment, Conclusion Coding: Data Preparation, Modelling, Evaluation.
Sajan Bang	Write-ups for:Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation. Coding: Data Preparation, Modelling, Evaluation.
Hitesh Laxmichand Patel	Write-ups for: Data Understanding, Data Preparation, Modelling, Evaluation, Conclusion. Coding: Data Preparation, Modelling, Evaluation.