Raheela Charania, Anmol Anchala, Emmanuel Wediko

BDA 620: Data Mining Section A20

Quantifying Churn Risk in Banking: A Comparative Study of Predictive Models

Dr. Eshan Ahmadi

Mercer University

December 12, 2024

Background

Customer churn is when a customer no longer utilizes the services of a specific company and closes their account. In most industries, it is much more cost effective to prevent customers from leaving the company than to attract new customers. Specifically, in the banking industry, understanding reasons for customer churn and ways to prevent it is important not only for financial purposes but also for the bank's reputation. In 2019, it was observed that "66.8% of current banking customers have already used or intend to use a bank account from a non-traditional company (big tech or fintech) in the next three years" (de Lima Lemos, Renato Alexandre et al.). In order to prevent customers from churning, banks as well as other institutions have begun to analyze factors that indicate a customer may churn. In this analysis, we aim to figure out which factors contribute the most in predicting whether a customer is more likely to churn. By identifying customers that are at risk of churn, banks can get ahead of the curve and provide incentives to retain customers and increase profits.

Determining whether a customer is at risk of churning has a few challenges. It will require thorough analysis of various tangible factors and even then, there may be other factors at play that are unable to be measured, such as human interaction that may or may not affect the model. Another potential challenge is conducting analysis on the large customer database as well as having a large number of predictor variables that may or may not be significant in determining churn rate.

In this paper, we will begin by reviewing the various predictors that can affect the churn rate. The full list of variables are in the data dictionary (Chart1). We will then partition the dataset, create graphs and conduct regression analysis on the training dataset to determine which predictors have a significant contribution to the target variables. The insignificant variables or the variables that indicate multicollinearity will be removed to optimize the model via stepwise selection and a regression equation will be created. This model will be tested against the test dataset and will be used to calculate the specificity, sensitivity, AUC, and RMSE. Our goal is to decrease the number of customers churning by providing incentives to the customers who are at risk for churning. We will use the model to do a 6-month pilot program identifying at-risk customers and providing them with incentives to remain with the bank. We plan on reviewing the percentage of customers before and after this program to determine the effectiveness of our model and implement at all locations once the model is further optimized.

Exploratory Data Analysis

We ran some exploratory stats to understand data distribution, to identify any outliers and to learn patterns in the dataset.

Cleaning Dataset

We cleaned the dataset to ensure its quality for analysis. After identifying outliers, we treated them using the IQR method for key numerical columns (Age, NumOfProducts and CreditScore). This helps in maintaining the integrity of our data by ensuring that extreme, unrepresentative values do not skew the analysis results. Fortunately, the initial dataset had no missing values, so there was no need for removal of records or imputation.

Preprocessing

In order to prepare our dataset for analysis, we first removed the outliers. Originally we had 10,000 records. After removing the outliers, we had 9,404 records remaining. We then converted our categorical variables to dummy variables. We standardized our numerical variables via z-score standardization. We then partitioned the datasets into 70% train and 30% test. The test dataset was then set aside and the models were created on the training dataset.

Methodology

For this dataset, we aim to create a simple model with a high accuracy rate that can predict whether a customer will churn based on their financial history and relationship or interaction with our bank. In conducting our analysis, we partitioned the dataset into two parts, including 70% for training and 30% for testing purposes. We decided to run 4 models, including the Logistic Regression, CART, Random Forest and K-nearest neighbors (KNN). We will use the training subset to create a logistic model via the Forward Selection, Backward Elimination and Stepwise Procedure. We used logistic regression since our target variable is categorical.

A key measure in these three variable selection methods is the F-ratio. The F-ratio is "a measure of the ratio of variances" (Kissell et. al. 2017) . It is a statistical test to determine if there is support to reject the null hypothesis, which in this case is that there is no relationship between the predictor variable and the target variable (Kissell et. al. 2017) . If the F-ratio is high enough, then the null hypothesis can be rejected and the predictor variable remains in the model (Kissell et. al. 2017) .

We first started running the dataset through the Forward Selection method. In this method, the model starts off without any predictor variables. These are then added one-by-one based on if the F-ratio is above a threshold. They are added in order of the amount contributed to the fit. The Backward Elimination method starts off with all the predictor variables, which are then removed one-by-one, again based on the F-ratio.

Lastly, we ran the Stepwise Procedure, which combines the aforementioned methods by adding and removing variables based on their significance as determined by the F-ratio.

To determine which model is more suitable for our dataset, the AIC of each variable selection method is compared. The lowest AIC indicates a better model selection. In our case, all three had the same AIC of 115.03 so any model selection method can be utilized (Greenwood et al. 2017).

The model we selected was the result of the stepwise selection is as follows:
Model 1: Exited = 0.13231*Age + (-2.76154*IsActiveMember) + 14.02229*Complain - 11.94459

In Figure 1, we can see that the ROC AUC is 99.97% indicating that our model is a great fit.

In addition, the model has an accuracy rate of 99.8%, a specificity of 99.8% and a sensitivity of 99.7% (Table 2). These rates are very high, indicating that the model is a good fit for the training dataset; however, we are concerned that this model overfits the dataset. In the next section, we will apply this model to the test dataset and if we see some overfitting, we will revisit the model and perhaps make some changes to better the model.

Predictive Performance

When running our first model with all predictor variables, we had an AUC of 0.998. After thoroughly looking through our model we realized that the predictor variable "Complain" was highly correlated with the outcome variable "Exited" as seen in the heat map (Figure 10). This is causing our model to overfit. After removing "Complain" from the full model our accuracy rate decreased by 19% (Table3, Chart3). There was also a decrease in all other aspects including specificity and sensitivity.

We reran the Logistic Regression and came up with the model below with an AUC of 0.795. This proves that the model has good discerning ability and can accurately differentiate between clients who have churned and those who have not. The difficulties of real-world datasets with overlapping customer behaviors and noise are reflected in this AUC, despite its imperfections.

Model 2: glm(formula = Exited ~ Age + Tenure + Balance + NumOfProducts + IsActiveMember + Point.Earned + Geography_Germany + Geography_Spain + Card.Type_GOLD + Card.Type_SILVER + Gender_Male, family = "binomial", data = train_data_std)

We then ran the Random Forest, CART and KNN models. With consistent accuracy and AUC values, the Random Forest and Decision Tree models showed strong generalization. While the Decision Tree offered comprehensible guidelines for churn prediction, Random Forest most likely obtained a little higher AUC because of its ability to capture intricate linkages.

By providing consistent performance across training, validation, and test datasets, cross-validation confirmed these models' stability. In order to prevent forecasts from being unduly impacted by a variable that had a perfect correlation with the target, the Complain variable was eliminated during preprocessing.

We further evaluated Random Forest by identifying which of the variables affect the model more. Importance is a function within the random forest package which we used to do this. The output of importance is two graphs, one is Mean Decrease Accuracy and the other is Mean Decrease Accuracy. It is a tool that helps rank which variables are more important to the target variable. Specifically, the Mean Decrease Accuracy measures the difference in accuracy if a variable is taken out. Mean Decrease Gini, on the other hand, looks at how important the variables are based on a Gini impurity index which comes from how the Random Forest tree is split and how homogenous the sections are. For both of these, if the value is high, the variable is important (Martinez-Taboada et. al.). We identified Age, Balance and NumofProducts as the top three variables that are most important in this model (Figure 11).

Conclusion

All things considered, the Random Forest most is suitable for deployment and offers useful insights for churn identification. Although the AUC of 0.795 indicates that there is potential for enhancement, the existing models effectively balance generalizability with prediction accuracy while preventing overfitting.

Business Recommendation

Based on our analysis we would suggest the following for our top three variables of Age, Balance, and NumofProducts. As age increases the log odds of leaving the bank increases.The customer is 2.98 times more likely to exit. Specifically customers between the age of 40 and 52 are more likely to churn (Figure 4). We suggest offering restaurants, groceries, and travel rewards to those customers only. We do not recommend opening this promotion to customers who are less likely to churn since that will result in the business losing money. Secondly, as balance increases the log odds of leaving the bank increases. A customer is 15.81 times more likely to exit. To prevent customers from leaving the bank, the bank could offer extra rewards for every three months that are paid off earlier. For NumofProducts, when customers with a higher number of products through the bank are approximately 6.78 times more involved, customers are less likely to churn compared to those without such a plan (Figure 3). We can offer customers to pick industries of their liking to earn double points for each month.

References

1. de Lima Lemos, Renato Alexandre et al. “Propension to customer churn in a financial institution: a machine learning approach.” Neural computing & applications vol. 34,14 (2022): 11751-11768. doi:10.1007/s00521-022-07067-x
    a. Stats originally sourced from: Capgemini E (2019) World retail banking report (last accessed on 03/28/2020)
2. Greenwood, Mark. “8.13: AICS for Model Selection.” *LibreTexts Statistics*, Libretexts, 16 Dec. 2022.
3. Keramati, Abbas, Hajar Ghaneei, and Seyed Mohammad Mirmohammadi. "Developing a prediction model for customer churn from electronic banking services using data mining." Financial Innovation 2 (2016): 1-13.
4. Kissell, Robert, and James Poserina. Optimal sports math, statistics, and fantasy. Academic Press, 2017.
5. Long, Hoang Viet et al. “A New Approach for Construction of Geodemographic Segmentation Model and Prediction Analysis.” Computational intelligence and neuroscience vol. 2019 9252837. 13 May. 2019, doi:10.1155/2019/9252837
6. Martinez-Taboada, Fernando; Redondo, Jose Ignacio (2020). Variable importance plot (mean decrease accuracy and mean decrease Gini). PLOS ONE. Figure. https://doi.org/10.1371/journal.pone.0230799.g002
7. Prasad, U. Devi, and S. Madhavi. "Prediction of churn behavior of bank customers using data mining tools." Business Intelligence Journal 5.1 (2012): 96-101.
8. Singh, Pahul Preet, et al. "Investigating customer churn in banking: A machine learning approach and visualization app for data science and management." Data Science and Management 7.1 (2024): 7-16.

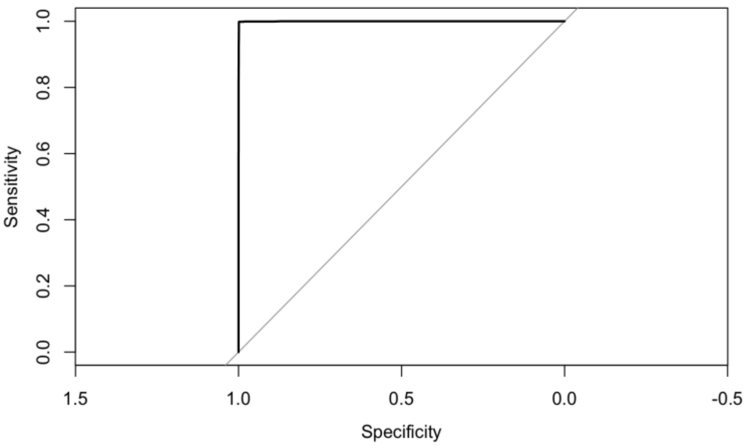
Appendix:

Data Dictionary (provided by Dr. Ahmadi with the dataset)

(Table 1).

	Definitions
RowNumber	Corresponds to the record (row) number and has no effect on the output.
CustomerId	Contains random values and has no effect on customers leaving the bank.
Surname	The surname of a customer has no impact on their decision to leave the bank.
CreditScore	Can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
Geography	A customer's location can affect their decision to leave the bank.
Gender	It's interesting to explore whether gender plays a role in a customer leaving the bank.
Tenure	Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
Balance	Also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
NumOfProducts	Refers to the # of products that a customer has purchased through the bank.
HasCrCard	Denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
IsActiveMember	Active customers are less likely to leave the bank.
EstimatedSalary	As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
Exited	Whether or not the customer left the bank. 0 = customer stays. 1 = customer leaves.
Complain	Customer has complaint or not.
Satisfaction Score	Score provided by the customer for their complaint resolution.
Card Type	Type of card held by the customer.
Points Earned	The points earned by the customer for using credit card
Age	This is certainly relevant, since older customers are less likely to leave their bank than younger ones.

Full Model (With Complain)

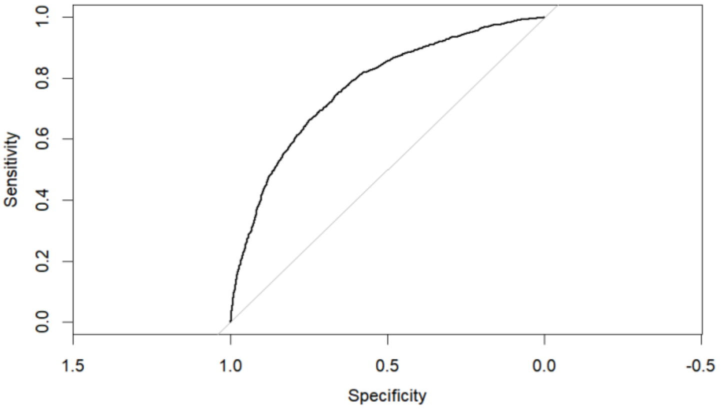


(Figure 1)

(Table 2)

Specificity	Sensitivity	Accuracy
99.8%	99.7%	99.8%

Adjusted Model (Without Complain)



(Figure 2)

(Table 3)

Specificity	**Sensitivity**	**Accuracy**
29.4%	96.5%	83.4%

(Table 4)

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Logistic	71.7	73.50	69.87	79.59
CART	79.73	81.89	71.21	83.22
RF	83.8	89.26	62.24	85.48
KNN	83.53	68.41	98.96	83.68

Descriptive Statistics

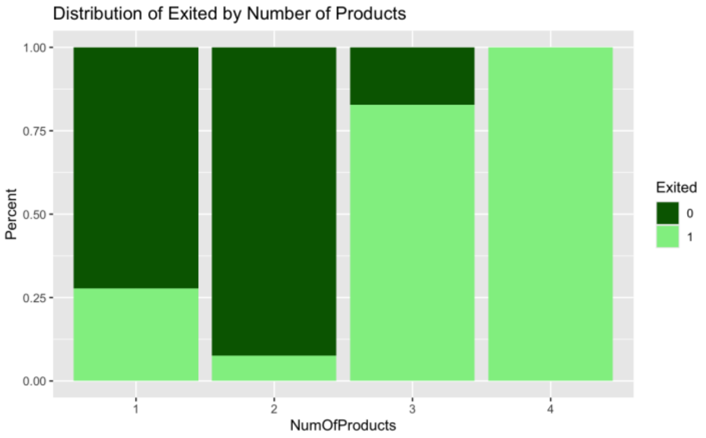
(Table 5)

	Count	**Percentage**
Gender Female: Male:	 4332 5237	 0.45 0.55
Has Credit Card 0: No 1: Yes	 2821 6748	 0.29 0.71
Card Type Diamond: Gold: Platinum: Sliver:	 2404 2376 2382 2407	 0.25 0.25 0.25 0.25
Exited 0: No 1: Yes	 7676 1893	 0.80 0.20

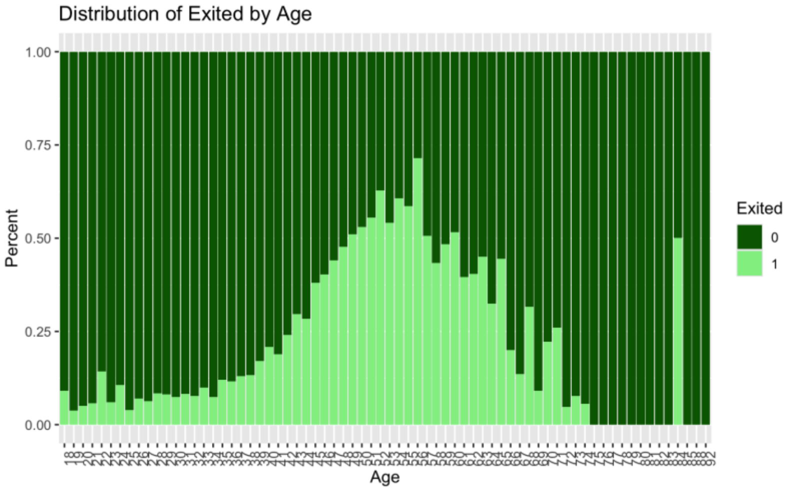
(Table 6)

	Mean	Standard Deviation
Age	37.74	8.78
Points Earned	605.92	225.87
Balance	76426.07	62417.82
Credit Score	650.66	96.13
Tenure	5.01	2.89

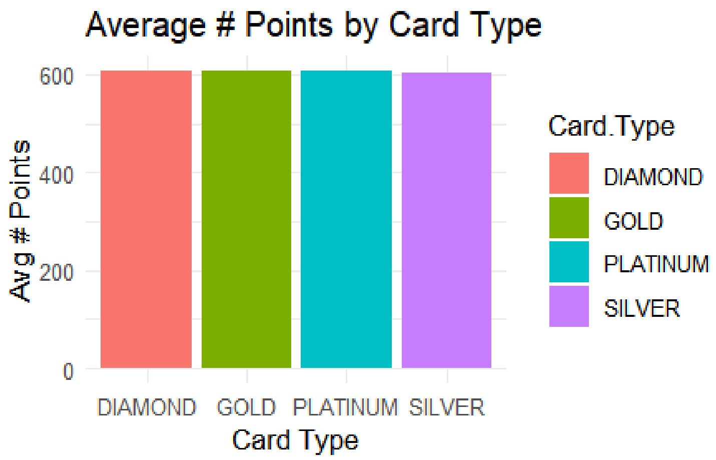
Exploratory Statistics



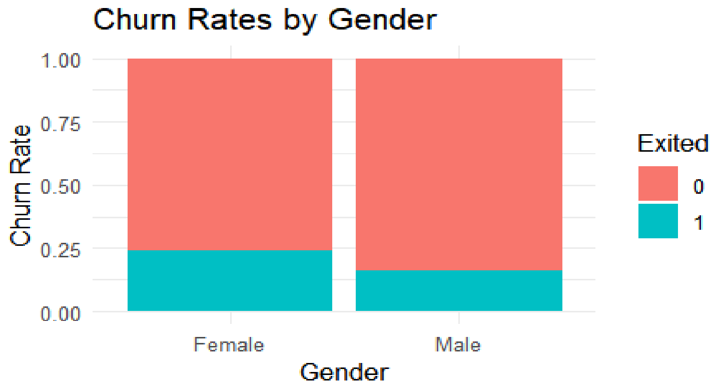
(Figure 3)



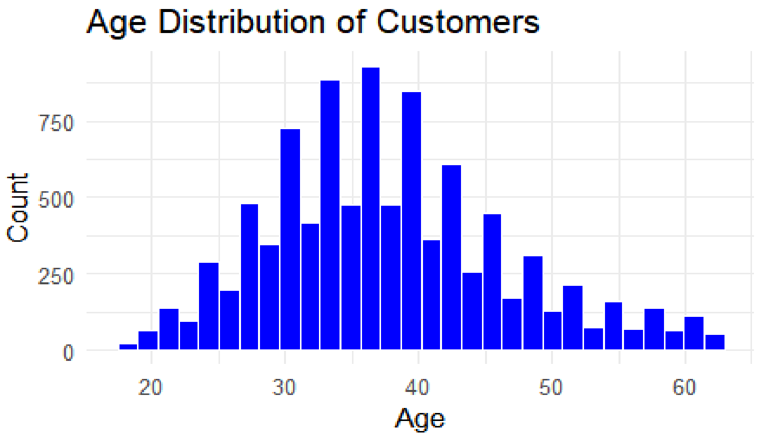
(Figure 4)



(Figure 5)



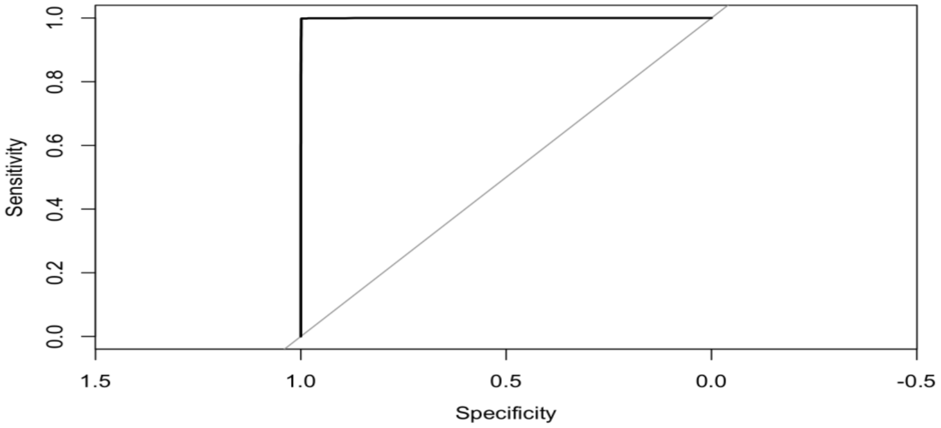
(Figure 6)



(Figure 7)

(Table 7)

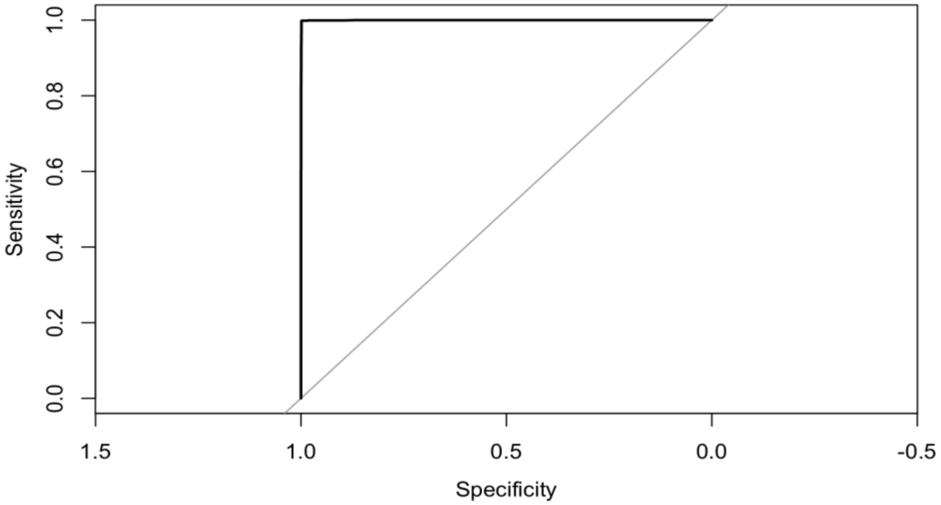
Accuracy	Sensitivity	Specificity
99.8%	99.7%	99.8%



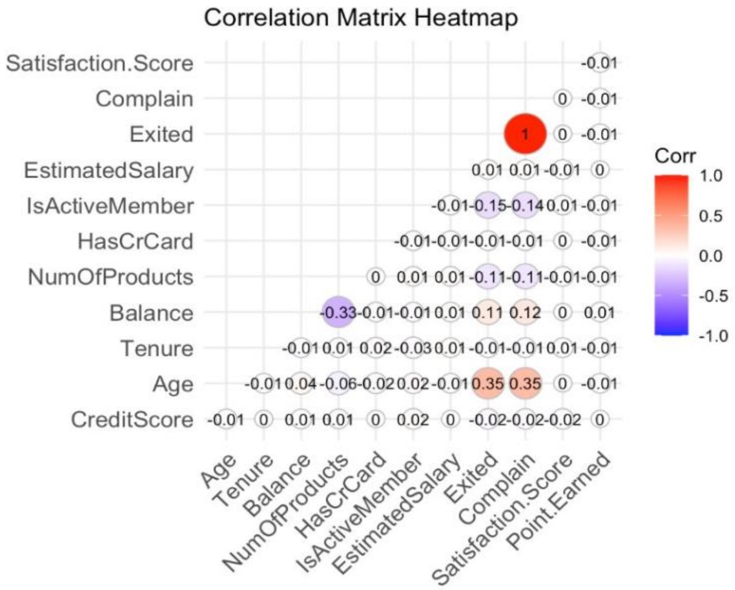
(Figure 8)

(Table 8)

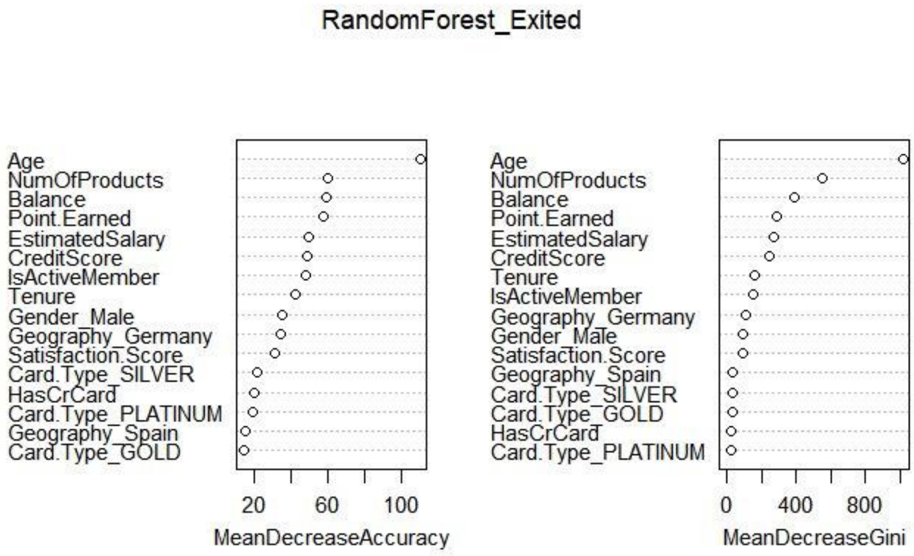
Accuracy	Sensitivity	Specificity
99.8%	99.7%	99.8%



(Figure 9)



(Figure 10)



(Figure 11)