Abstract

The aim of the project was to predict the customers who are going to churn from banks credit cards services and give the prediction to the bank manager in order to tackle the problems they face and thus retain the customers. The project followed the following steps defining the problem, data preparation, data visualization and finally applying Machine Learning techniques on the prepared data. The data provided by the bank manager at open source platform kaggle was very clean with no missing or repetitive values. After successfully visualising the data and finding out the relationship between the 21 (out of 23 since 3 columns were removed) columns we proceed with machine learning model selection. Deciding which model to use was one of the most important parts. We decided to start with Naive Bayes since it's used mostly for classification problems. The reasonings for all model selections are detailed below in the Model part of the report. The model is very good for starting but accuracy plays a huge role and that's why in order to increase accuracy number development of the model is mandatory. Decision tree and Random forest models are the best accuracy demonstrated models. Data has been taken from kaggle.com which is a popular website for datasets. Information about this data was reliable and established business problem - to find out the rate of churn among customers, can be solved through this data and above mentioned models.

Introduction

Business analytics (BA) task overview and comprehending the underlying importance is essential for its application. Business analytics is applied as a part of a decision-making process.[1] One of the definitions is that it is "the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions".[2] The integration of the analytics ensures that the data, the information, is well-understood and

[1] Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal, 22*(4), 6-17.

[2] Evans, J. R. (2016). *Business Analytics: Methods, Models, and Decisions* (2nd ed.). Boston, MA: Pearson Education Limited.

intelligently used.[3] It can be implemented to a wide-range of industries that have a business, medicine, astrophysics and public policy nature.[4] For instance, in the implemented industries, its application improves the organizations such as marketing and finance activities, supply chain and customer relationships management and many others.[5] BA's provided value is driven from the fact that it corresponds to hindsight, insight and foresight over its application process. This value is due to its wide scope of descriptive, predictive and prescriptive analytics.[6] During the integration, it basically centers around the data and statistical analysis in order to derive business decisions by foreseeing the future states.[7] Thus, it is observed that the impact areas of BA are "business decisions, big data, data-driven solutions and forecasting."[8]

In this project report, the proposal is to examine the business issue of a churn analysis. For this purpose, following the CRISP-DM Methodology played a crucial role. The steps of the CRISP-DM generates the organization of the rest of the proposal and the report. In the methodology, as the business analytics application, decision tree, random forest, naive bayes are used. The nature of the project is constructing a classification model with the bank's data for predicting. Briefly, in the project, the top priority is to identify whether the credit card customer will attrite or not. In this way, the harm of the churn can be prevented through the retain of the relevant customers.

Problem Understanding

In the project report the dataset is provided by a bank manager who is worried that more and more people keep opting out of the bank's credit card services. The aim is to predict whether

[3] Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal, 22*(4), 6-17.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*(Vol. 103). New York: Springer.

[5] Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal, 22*(4), 6-17.

[6] Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal, 22*(4), 6-17.

[7] Robin. (2018, March 19). Difference between Business Analysis and Business Analytics. Retrieved January 16, 2021, from https://thebusinessanalystjobdescription.com/difference-between-business-analysis-and-business-analytics/

[8] Robin. (2018, March 19). Difference between Business Analysis and Business Analytics. Retrieved January 16, 2021, from https://thebusinessanalystjobdescription.com/difference-between-business-analysis-and-business-analytics/

the credit card customer’s behavior will be attrition or not. The aim is driven from the problem statement itself. Defining a clear problem statement is a must before solving the problem. The problem statement is investigation of the customer attrition data set to predict customer’s situation (churn or not). The investigation of the following question will be done:

“Is the behaviour of the relevant customer is to attrite (churn) or not?”

Under the aforementioned problem statement and aim, the task that is conducted is performing an exploratory analysis of the data as a solution to the problem. The analysis is done by utilising visualization and calculative analysis tools. Based on the exploratory analysis, a predictive model is built in the purpose of determining the behaviour of the customers. This is achieved by applying various machine learning algorithms in order to predict the people who will leave. In the end, the model with the highest accuracy is selected in predicting the customer who will be leaving the bank’s credit card services. With both these predictions, the bank will be able to provide improved services which will lead retaining current customers.

Data Understanding

The Credit Card Customers Dataset that is utilised can be seen in the following link: https://www.kaggle.com/sakshigoyal7/credit-card-customers. The dataset selected serves the business issue of churn analysis. There exists 10127 rows of data indicating the credit card customers of the bank, and 23 columns of data indicating the variables. The columns contain 10 integers, 7 double (float, decimal) and 6 categorical data types. Although it should be taken into consideration that the last two columns of the dataset, Naive_Bayes_Classifier decimal typed columns, should be extracted before any examination on the data.

This dataset consists of approximately 10,000 customers mentioning their age, gender, education_level, marital_status, salary, credit card limit, credit card category, months in the bank etc. There are a total of 18 attributes. The data contains the following information:

1. Existing customers in bank 84%, Attrited Customers 16%,
2. The age group of customers ranges from 26 - 65, with the majority around 44,
3. Male to female ratio is 53:47,
4. 30% of the customers are college graduate while 20% are still in high school,

5. 46% are married whereas 39% are single,
6. 35% of consumers lie within the income group of less than 40k$ per annum,
7. A whopping 93% of the customers are blue card users,

and it has only 16.07% of customers who have churned.

The Attrition_Flag is the target attribute while the other attributes play a role of predictive attributes. The dataset is not a complex dataset, it can be said that it is a somewhat clean dataset since there exist no missing values, duplicated, thus there are no challenges that occurred during the work.

Data Preparation

The data preparation section is done after acquiring a detailed understanding of the dataset, and before the modeling part. This part should be conducted with caution so that the following modeling part will occur in a smooth way.

In the dataset, the duplicated rows, missing values are checked. The duplicated function is utilised and it is observed that there are no duplicated rows occurred. Also, to detect any indication of missing values, the glimpse function is utilised and it is concluded that there are also no missing values occurred. Furthermore, for the continuous variables of Credit_Limit, Total_Revolving_Bal, and Months_Inactive_12_mon, both the outliers are checked and also their correlation with the Attrition_Flag is checked. It is seen that there is no correlation that is observed among these variables, however it is seen that there are some outliers. To not change the nature of the dataset, the outliers are not removed. Plus, in the target attribute, Attrition_Flag column, it is observed that there is an imbalance of data. Ratios being equal in the target attribute column is important because it may cause data memorizing instead of data learning since it creates a bias at the beginning. Therefore, SMOTE technique is used as an oversampling method to solve the imbalance problem. Furthermore, in order to observe and to make further implications, data is also visualised using several graphs which are explained in the following paragraphs.

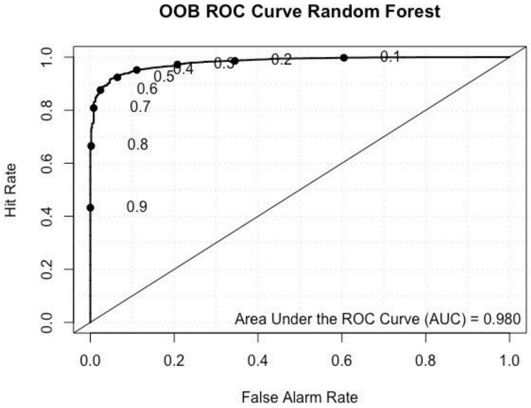


Figure 1

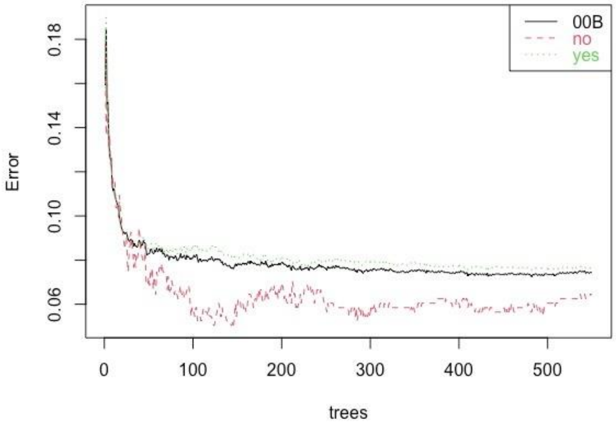


Figure 2

Figure 1 shows the OOB ROC Curve for this model. The diagonal line is representing a random classification. The area under it is 0.6. As the distance between the diagonal line and the OOB ROC curve increases it indicates that the model performs well. The area under the curve is 0.980, as it is an indicator of high performance.

Secondly, Figure 2 shows the information about Errors in the number of trees. Prediction of negative and positive outcomes according to our model shows true results. Meanwhile, negative results show higher volatility than positive ones.

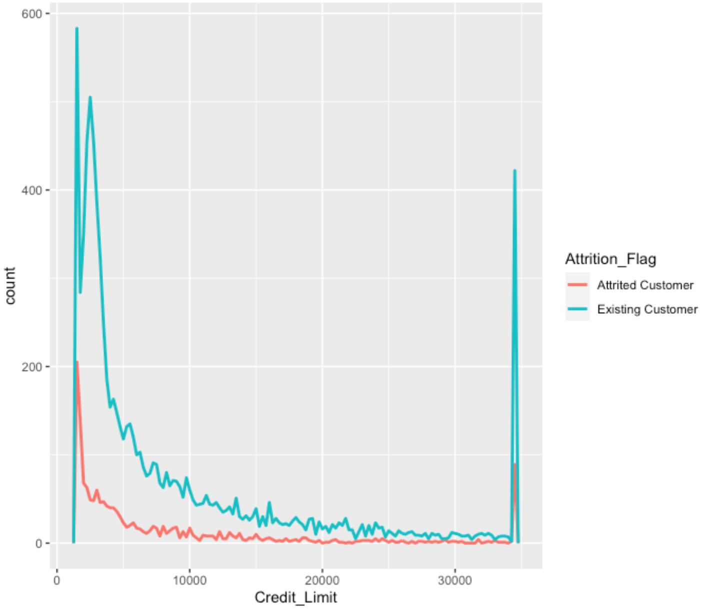


Figure 3

There are few different plots that could analyze and explain the whole model and its preparations. Figure 3, shows distribution of feature - ***Attrition Flag*** in terms of credit limit feature. Continuous variables over there show that the credit limit is actually quite an influential factor for customer satisfaction and decision of either keeping or leaving the customer status at the bank.

The boxplot in Figure 4 demonstrates the existence of customers according to their inactivity within 12 months. There are outliers near the scale of 5 and 6 which means the possibility of customers staying with the company after 6 months of inactivity is higher than others which can also be analyzed through filtering data on Excel or R.

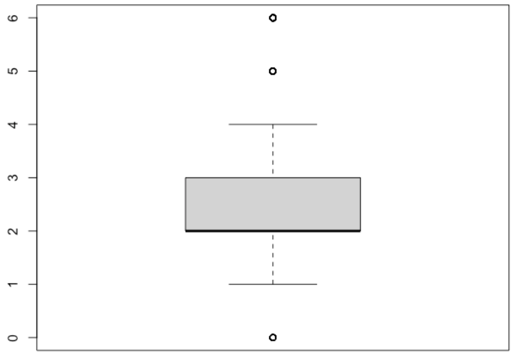
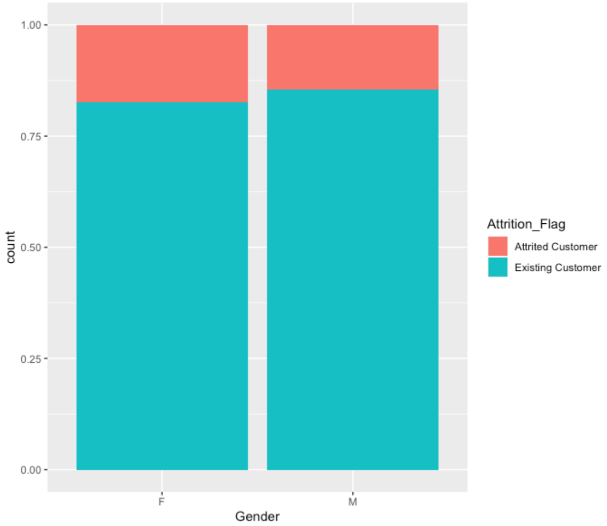


Figure 4

Describing data visualization with graphs always brings column charts to minds. It's a more well-known and highly efficient method of data visualization. As can be seen below 3 different column charts explains the visualization of data with different variables. Firstly, differentiation churn as regards gender is not changing objectives as the rate between them is slightly different where males tend to stay with banks more than females. On the other hand, academic life and income of the customer shows more volatility than gender for their attrition decision. Existence of customers who are doctorate level holders are lower which creates different questions that can be investigated by the bank to find out why especially this share of their audiences tend to leave the bank rather than others.



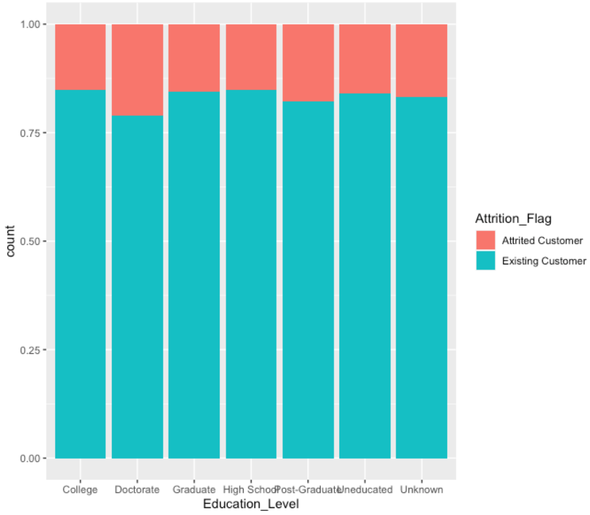


Figure 5

Figure 6

Modeling

1. Naive Bayes Algorithm: Naive Bayes model is a type of classification machine learning algorithm. It enables an easy way to create accurate models with a very good performance given their simplicity. Naive Bayes model was chosen since it allows to use both numerical and continuous variables in the prediction. By using Naive Bayes Model, the first column as a target attribute and other columns as predictive attributes were chosen. After that using the confidence table the accuracy of the model was calculated. Also average probability values are important performance metrics. Therefore to find their predicted outcomes, maximum probability values were found and their mean value was found. To visualize the performance metrics on Naive Bayes, which are the accuracy and average probability values, their plots were visualized.

2. Decision Tree Algorithm: Decision Trees are data mining techniques for classification and regression analysis. The algorithm finds different ways to split the data into partitions. First, the real data were initialized into tree data. Then all the character variables are transformed into factor variables since categorical variables are not allowed to be used in the Decision Tree Algorithm. Bank data were splitted with 0.75 and 0.25 ratios into training and test data sets. DTree1 was created and it had variables for "is it going to churn or not" situation since class prediction will be used with the target attribute Attrition_Flag. Then the prediction of class and probability values were found for both training and test data sets. By using the class predictions, a confusion matrix was set up and the accuracy of the model was calculated. Also to find the AUC value, probability values were used because AUC requires real continuous or real numerical values. And finally, ROC was found to show the performance of the prediction model.

3. Third candidate model selection is Random Forest Classification. The reason for choosing the decision tree model is that they provide rules that are easy to interpret and it also handles

missing data. The data set is partitioned into two groups of training and testing dataset. There are 7520 observations in the training dataset, and the rest are in testing dataset.

While constructing the random forest classification tree model in R, the target variable is Attrition_Flag. Based on the target variable and predictive the written decision tree algorithm provides predictions whether customers will churn or not.

The training dataset is used for the model. Based in class and probability predictions were conducted. It is followed by a confusion matrix which was found by using prediction based on classes, which is an indicator of how successful the class predictions are.

In order to produce an ROC curve, the real number values were used. For finding AUC curve values continuous values are needed so probability values were used.

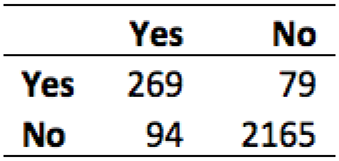
Evaluation

1. Model with Naive Bayes Algorithm: To create data partition, stratified sampling was used. Stratified sampling is a type of hold-out sampling and it enabled to obtain a sample population that well represents the entire data that was examined and it made sure that each subgroup was represented. As a result of sampling, training and test sub data were created. Naive Bayes algorithm was applied afterwards. Predictions of the model were found and by using confusion matrix results were obtained. According to the observations, the accuracy value for the BankChurn prediction model is 0.80. Furthermore, after finding the accuracy value, the average probability of predicted outcomes was found. In addition to accuracy metric, also probabilities are meaningful performance metrics, and it can be said that they are even more valuable than accuracy values. So the average value of predicted outcomes was found as 0.89.

2. Model with Decision Tree Algorithm: Before starting to apply the Decision Tree algorithm, random sampling was implemented to create data partition for test and training. Random sampling is a simple random sample generated by a design, which provides that each subgroup of the population with the size n has an equal probability of being picked as the sample. After that, cross-validation was applied for resampling and it

was used to obtain confusion matrix and AUC values. By using the confusion matrix, accuracy of both test and training data set was found 0.93 separately. On the other hand, by using the probability values, performance of the AUC metric was found as 0.90.

Figure 7



	Yes	No
Yes	269	79
No	94	2165

3) Figure 7 shows the confusion matrix. “Yes” refers to an Attrited Customer observation, and “No” refers to Existing Customer one. The total number of correct classifications is 2434, which is obtained by adding the True Positive (269) and the True Negative (2165) values. Precision, also called Positive Predictive Value, is 0.7730. Another performance measure, sensitivity, is 0.7410, which means 74.10% of observations have been correctly classified as Attrited Customer and the probability of a Type II error is 25.9%. Specificity refers to the ability of the model to correctly predict Existing Customer observations, and it is 0.9648 for this model, which shows that 96.48% of observations have been correctly classified as Existing Customer and the possibility of a Type I error is 3.52%. The accuracy measure shows the percentage of observations the model has classified correctly. For this model, accuracy is 0.9336, which means the model has classified 93.36% of the observations correctly. These values would suggest that this model is successful in all aspects. Also OOB ROC curve is used in the evaluation part as discussed in the data visualization the area under the OOB ROC curve is 0.980, it is suggesting that this model is a successful predictor.

Discussion and Conclusion

Accuracy was chosen as a comparison metrics in the assessment part. When the results of all three models which are Naive Bayes, Decision Tree and Random Forest Algorithms are compared, it can be observed that all of them have accuracy values higher than 0.80. Therefore it is possible to say that all algorithms have good performance levels. However Random Forest algorithm has 0.96 accuracy which is very high and almost perfect fit for data. Decision Tree and Naive Bayes have relatively lower accuracy values that are 0.93 and 0.80 consecutively. Also some other prediction performance metrics were found to analyze the model. The Random Forest and Decision Tree algorithm has AUC values as 0.99 and 0.91. Also since it is a significant performance metric, average probability values were calculated for Naive Bayes Algorithm and it was found as 0.89 as well. Accuracy is the performance metric that is common and suitable for all models therefore it is possible to choose the best model according to their accuracy values. Hence Random Forest algorithm is the best option since it has the highest accuracy.

Bibliography

- Evans, J. R. (2015). Modern Analytics and the Future of Quality and Performance Excellence. *Quality Management Journal, 22*(4), 6-17. doi:https://doi.org/10.1080/10686967.2015.11918447
- Evans, J. R. (2016). *Business Analytics: Methods, Models, and Decisions* (2nd ed.). Boston, MA: Pearson Education Limited.
- Robin. (2018, March 19). Difference between Business Analysis and Business Analytics. Retrieved January 16, 2021, from https://thebusinessanalystjobdescription.com/difference-between-business-analysis-and-business-analytics/
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*(Vol. 103). New York: Springer. doi:DOI 10.1007/978-1-4614-7138-7 1
- Das, S. (2018, June 06). Data Sampling Methods in R - DZone AI. Retrieved January 21, 2021, from https://dzone.com/articles/data-sampling-methods-in-r
- Decision Tree Algorithm Examples in Data Mining. (2021, January 18). Retrieved January 21, 2021, from https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/#:~:text=Decision%20Trees%20are%20data%20mining%20techniques%20for%20classification%20and%20regression%20analysis.&text=These%20algorithms%20find%20different%20ways,machine%20learning%20and%20pattern%20analysis.
-