Bank Marketing Prediction

Bhargav Kethineni

06/29/2022

## Abstract:

Now a days, organizations are showing more interest in online and direct database marketing as part of the marketing strategy to increase sales and keep track of their customers. One of primary methodology by which organizations advance their products is through making direct campaign’s, In our case we campaign only specific group of individuals through phone calls. The problem is whether consumers of Portuguese bank will subscribe to the bank term deposit or not through this campaign. This project is based on the bank marketing data set that is collected by Portuguese banking institution. The main goal of this project is to perform classification and build classification algorithms which could predict based on the set of input variables that whether a customer will subscribe the term deposit or not in this case [1]. This model will help the banking institution to plan better marketing strategies to attract more customers who would then subscribe the term deposit. The data is taken from publicly available data sets which are available on Kaggle and could be downloaded easily to perform analysis and build classification algorithms.

## Introduction:

Identifying the behavior of customer is an important aspect of business and banking sectors. This helps the management to plan and produce new schemes which could help the general public, they get to know their customers, their level of understanding and what is beneficial to them. This project is also based on a marketing data set where the management is interested in finding out that whether a person is going to subscribe the term deposit or not. The goal is to build a model which could accurately predict this result and if the model predicts that the customer is not going to subscribe, the bank management could give the customer some more information / deals to increase their customer base.

## Literature Review:

In the study of Olatunji Apampa (2016)[1], they performed various classification algorithms with the balanced and unbalanced datasets. Firstly, they performed Random forest classifier on the Unbalanced dataset which consists of unequal numbers of “no” and “yes”. Moreover, they compared with other classification algorithms. In this 34% of the dataset is considered for the analysis i.e. only 15,375 instances were selected and then normalized. Later, the three classification algorithms Logistic regression, Decision tree( CART), Naive bayes and the random forest ensemble performed cross validation using 10-fold cross validation method. As a result the Area under the curve(AUC) value returned 0.576 for the random forest ensemble which is lower than the other classification algorithms. All the three classification algorithms showed similar AUC of 0.657 and the highest Classification Accuracy showed on CART algorithm with 0.99. Precision which shows the relationships which is tells positive predictive analysis for the logistic regression found to be 0.651 which is highest among all other algorithms. Whereas, recall which refers to the percentage of relevant instances that were correctly retrieved by classification algorithm. this resulted in the unbalanced dataset. In addition,by performing classification techniques results can be improved in the balanced bank dataset which has equal instances of “No” or “Yes” in the response class.

According to the study of Justice Asare-Frempong[2], the problem specify predicting customer response to bank via direct marketing. In this he focused on the customers likelihood of subscriptions to their products. This study consists of two objectives firstly, to predict customer response to bank direct marketing and the second is to identify the key features of customers who subscribed to term deposits. Three classification techniques are performed Random Forest, Decision tree and logistic regression. As a result, among all other classifiers Random Forest classifier portrayed with high accuracy of 92.7%. Moreover, to identify the key features of the customers to subscribe, a cluster analysis was performed.In my opinion, using classification analysis we can find customers with more call durations are spent has higher chances on subscribing to the term deposit.

## Reserach Problem:

The research problem in this case is: " Perform exploratory data analysis to understand the data set, build two different machine learning classifiers for binary classification, compare the performances of both classifiers on the test set and finally conclude the best classifier among both."

## Theory:

To know whether clients will subscribe to the bank term deposit which was initiated as a marketing campaign via phone calls. Below is the following hypothesis.

H1: By observing the bank data and previous activities performed by the customer. We get to know whether the customer is active or inactive to the bank term desposit subscriptions.

H2: Applying classification techniques to bank data, we can observe past term deposit subscriptions of the clients and get to know whether they are willing to subscribe to this term or not.

## Dataset:

The data set may be accessed at <https://www.kaggle.com/datasets/kukuroo3/bank-marketing-response-predict?select=train.csv>, which was downloaded from the Kaggle website. The data collection was compiled using information from the bank’s studies. In this scenario, we have a total of 12870 observations and 17 variables [1]. There is a combination of numerical and category data provided. Before beginning exploratory data analysis, the data set must be cleansed and prepared.The data is checked for null values, the negative values from the balance column are removed, the outliers are removed using IQR and cut off method and finally useless columns are also removed. The descriptive statistics of the final cleaned data is attached below:

## age job marital education default balance housing loan contact  
## 1 29 technician single tertiary no 18254 no no cellular  
## 2 26 services single secondary no 512 yes yes unknown  
## 3 30 management single secondary no 135 no no cellular  
## 4 41 technician married unknown no 30 yes no cellular  
## 5 27 admin. single secondary no 321 no yes unknown  
## 6 22 student single secondary no 185 no no cellular  
## campaign pdays previous poutcome y  
## 1 2 -1 0 unknown no  
## 2 3 -1 0 unknown no  
## 3 2 -1 0 unknown no  
## 4 1 -1 0 unknown no  
## 5 1 -1 0 unknown no  
## 6 1 -1 0 unknown yes

## Methodology

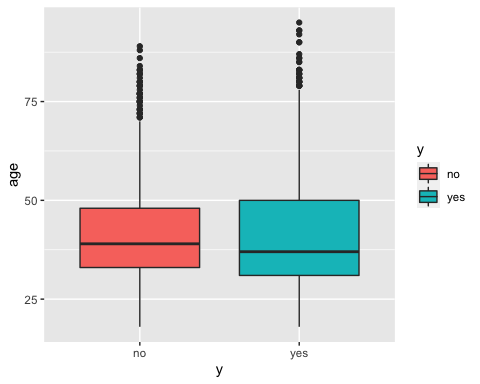
The flow of our project includes the following steps which are performed one by one to ensure smooth pre-processing and cleaning of the data set:

1. Data Acquisition & data cleaning.
2. Removal of outliers using Inter Quartile Range method with a cut off of 25% and 75% for balance column [3].
3. Exploratory data analysis to understand the data and get some key insights from data.
4. Applying machine learning algorithms which are Logistic Regression & Random Forest classifier.
5. A comparison of their results.
6. Conclude the findings

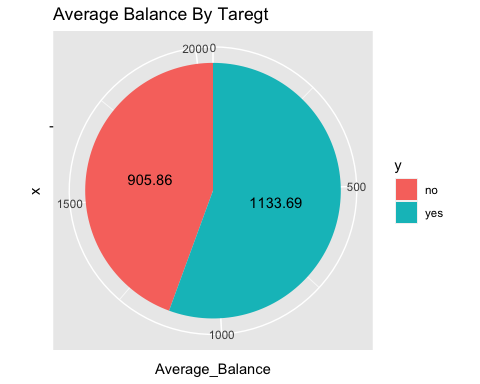
## Results:

We discovered that there are no null values in the data frame while verifying null values. There are 12870 rows and 17 columns in the data frame. Initially, we have 6 numeric columns and 11 categorical columns. We’ve removed the extraneous columns and double-checked the data set’s structure; our data set now just has 14 columns.

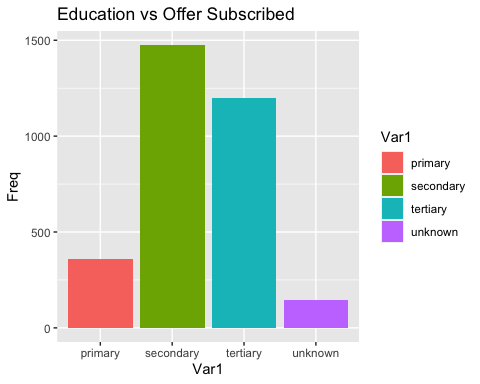
The first graph developed during exploratory data analysis is between age and the goal variable. The box plot [4] below demonstrates that persons who subscribe to a term deposit have a younger average age than those who do not. This suggests that younger people are more likely to subscribe than older people. The gap in average ages between the two groups isn’t that great, and they’re very close.



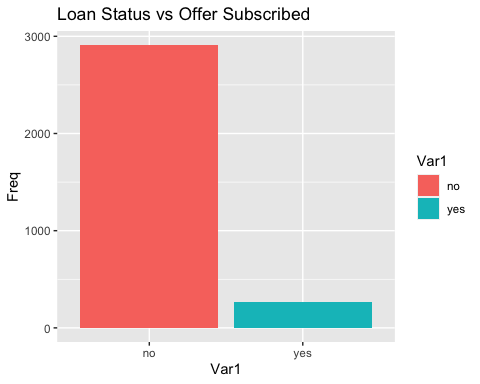
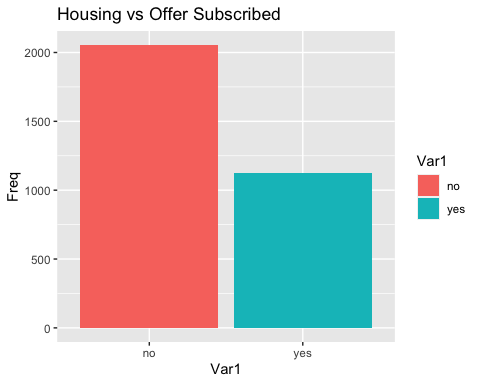
When it comes to balances, we can observe that individuals who subscribe to the term deposit have a larger average balance of 1133 than those who do not, which is around 905. This means that a person’s financial situation has a significant impact on the final decision of whether or not to subscribe to a term deposit in this circumstance.



Moving on to the factor of education, the majority of persons who subscribe to term deposits have only a secondary education, tertiary education comes in second, and primary education comes in third. The likelihood of subscription increases with more education, while the likelihood of subscription decreases with lower education. This is the conclusion that can be drawn from the attached result.



When it comes to the impact of home ownership and loan status, we can see that a large percentage of clients who have enrolled do not own a home and do not have any outstanding loans.



In the final step, we have trained two different classifiers which are random forest classifier [5] and logistic regression [6]. Before training the classifiers, the data is divided into training and testing sets. A split of 70% and 30% is used respectively for the training and testing sets.

Below is the result of random forest classifier on the testing set. It could be seen that the accuracy of our random forest classifier on the test set is 75%, from CI, we can say that we are 95% confident that the accuracy of our model on the test set lies between 73% to 76% approximately. The kappa statistics is also good in this case. The higher the value the better the model. The specificity and sensitivity are also aligned well and they are not biased. From confusion matrix we can see that false positive and false negatives are high which needs to be reduced.

## Confusion Matrix and Statistics  
##   
## predict\_unseen  
## no yes  
## no 1882 145  
## yes 609 349  
##   
## Accuracy : 0.7474   
## 95% CI : (0.7314, 0.7629)  
## No Information Rate : 0.8345   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3356   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.7555   
## Specificity : 0.7065   
## Pos Pred Value : 0.9285   
## Neg Pred Value : 0.3643   
## Prevalence : 0.8345   
## Detection Rate : 0.6305   
## Detection Prevalence : 0.6791   
## Balanced Accuracy : 0.7310   
##   
## 'Positive' Class : no   
##

Below is the result of logistic regression classifier on the testing set. It could be seen that the accuracy of our random forest classifier on the test set is 74%, from CI, we can say that we are 95% confident that the accuracy of our model on the test set lies between 72% to 75% approximately. The kappa statistics is also good in this case. The higher the value the better the model. The specificity and sensitivity are also aligned well and they are not biased. From confusion matrix we can see that false positive and false negatives as compared to random forest classifier are high which needs to be reduced.

## Confusion Matrix and Statistics  
##   
## predicted.classes  
## no yes  
## no 1898 129  
## yes 647 311  
##   
## Accuracy : 0.74   
## 95% CI : (0.7239, 0.7557)  
## No Information Rate : 0.8526   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.3044   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.7458   
## Specificity : 0.7068   
## Pos Pred Value : 0.9364   
## Neg Pred Value : 0.3246   
## Prevalence : 0.8526   
## Detection Rate : 0.6358   
## Detection Prevalence : 0.6791   
## Balanced Accuracy : 0.7263   
##   
## 'Positive' Class : no   
##

From the comparison performed on the above two classifiers, based on accuracy of the model on the test set, false positive and false negatives of the model, we can say that the performance of random forest classifier on the test set is better as compared to the performance of logistic regression model.

## Implications:

Future study that could be done in this area is to gather more information about customers and their behaviors such as number of times they have subscribed in the past, why they have unsubscribed in the past and having more data about customers and types of customer would definitely improve the results of this analysis and will provide the results at granular level.

## Conclusion:

Concluding the project we found out that the random forest classifier out performs the logistic regression model in terms of performance on the test set hence we can use the RF model for prediction of whether a customer will subscribe the term deposit or not. It will give us more accurate result.High population is of those customers who do not own a house and who do not have any loan on them but they have subscribed the term deposit. Highest number of people who subscribe the term deposit have a degree of secondary education only and the average balance of those who subscribe is higher than those who do not subscribe.

## References:

1. Apampa, O. (2016). Evaluation of Classification and Ensemble Algorithms for Bank. Journal of International Technology and Information Management, 85-100.
2. Asare-Frempong, J. (2017). Predicting Customer Response to Bank Direct. International Conference on Engineering Technology and Technopreneurship (ICE2T).
3. BOXPLOT in R 🟩 [boxplot by GROUP, MULTIPLE box plot, …]. R CODER. (2022). Retrieved 6 June 2022, from <https://r-coder.com/boxplot-r/>.
4. Random Forest In R. Medium. (2022). Retrieved 6 June 2022, from <https://towardsdatascience.com/random-forest-in-r-f66adf80ec9>.
5. How to perform a Logistic Regression in R | R-bloggers. R-bloggers. (2022). Retrieved 6 June 2022, from <https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/>.