BANK MARKETING ANALYSIS

Project Report

Snekha Duppathi

Executive Summary

Marketing campaigns are characterized by focusing on the customer needs and the factors like which segment of population to target, means of communication, the returns for the customers and many other play an important role in devising the marketing plans. By digging in into such questions we can come up with solutions to build up an effective marketing campaign that would help to target time and money on potential customers. In this project, I would analyse and predict whether the consumers will subscribe to a new term deposit product launched by a Portuguese banking institution or not using various machine learning algorithms. After performing some EDA analysis, resampling techniques and running few models on the data we compare the results of different models and find the power of these techniques to find the potential customers. Logistic regression, Support vector machine, Random forest and classification trees are the models I used, and the logistic regression had the highest performance and performed exceptionally well in predicting the potential customers.

Introduction

Data can be used in an effective way to create a huge impact on businesses, if one doesn't leverage it, they can be left behind in this fast-paced world in no time. An organisation can improve its performance in the market by capturing and analysing the customer data to improve the customer experience.

The bank marketing domain dataset would help the organisation to gain insights about the customer behaviour which would facilitate them to devise marketing strategies for effective product delivery to the customers which would in turn increase the sales and revenue of the organisation. The data used here is related to the telephonic marketing campaign held by the Portuguese banking institution for them to access if the customers would subscribe to their bank term deposit. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be('yes') or not('no') subscribed.

The term deposits allow banks to hold onto a deposit for a specific amount of time (ranging from one month to a few years), so banks can invest in higher gain financial products to make a profit. In addition, banks also hold better chance to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues. Also, these offer higher interest rates than traditional liquid savings accounts to the customers whereby they can withdraw their money at any time only by paying corresponding penalty.

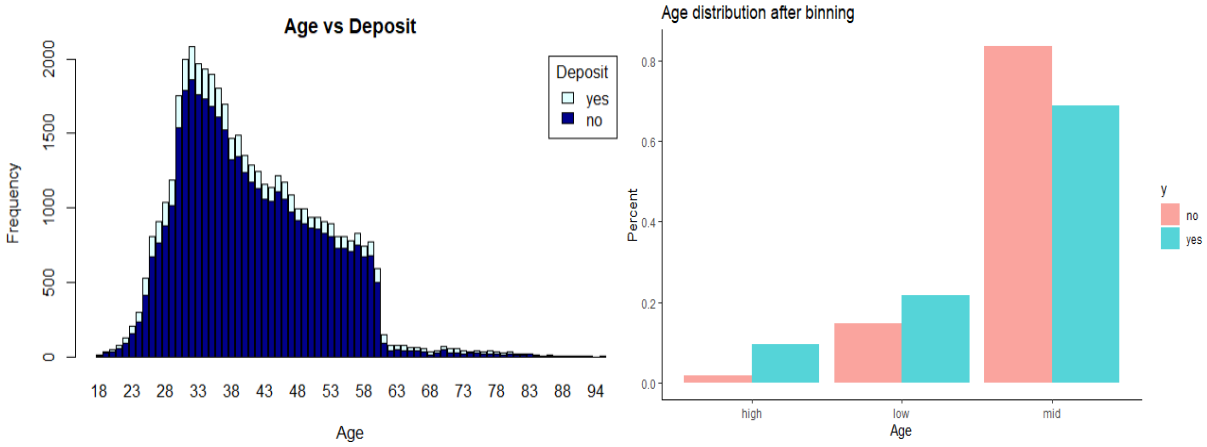
Data Description

The Bank marketing dataset here is from UCI ML repository with 45211 records and 17 features. The goal here is to predict whether the client will subscribe to term deposit (variable y) or not. There are 7 numerical and 10 categorical variables. There are no missing values in the dataset. The dataset is highly imbalanced where the positive class (y= 'yes') account for only 12% of all the records.

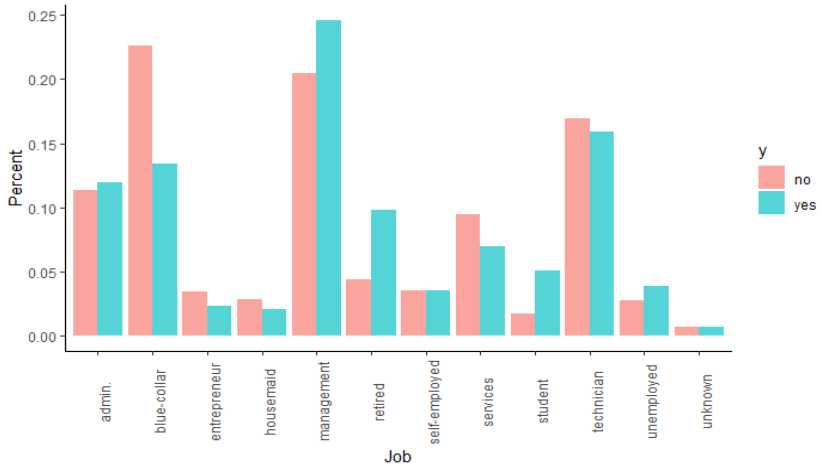
```
'data.frame':   45211 obs. of  17 variables:
 $ age       : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job       : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
 $ marital   : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
 $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
 $ default   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ balance   : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing   : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
 $ loan      : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ contact   : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ day       : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month     : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays     : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y         : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Exploratory data analysis

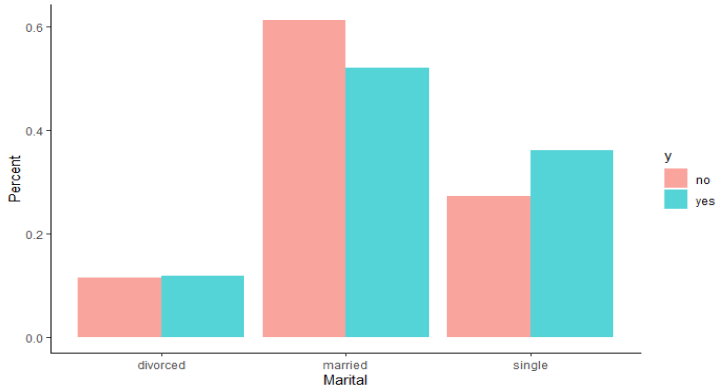
The age variable ranges between 18 to 95 with mean of 41. The distribution of this is right skewed and can be approximated to normal distribution. We are converting this numerical variable to factor based on their distribution. Age above 60 as high, 30-60 as mid and below 30 as low. We observe that the proportion of people subscribing to term deposit is higher in low and high age groups than mid age group.



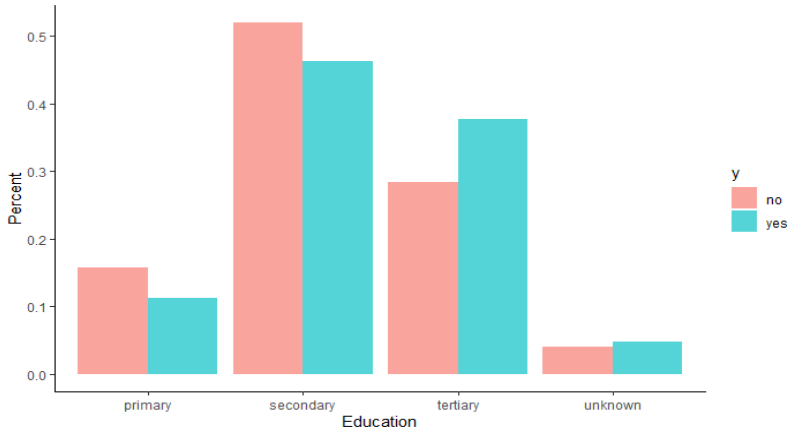
The job variable is a categorical variable with 12 different job types. We can see that the percentage of customers opting for term deposit is less in low paying jobs. The blue-collar, technician, management and admin jobs has comparatively good proportion of term deposit acceptance.



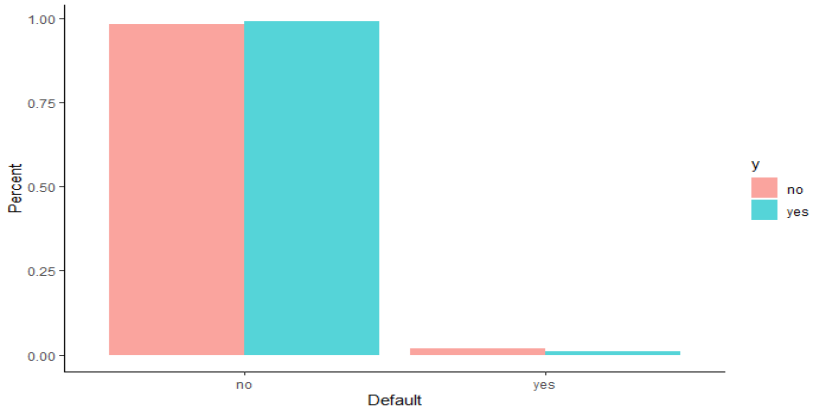
The marital variable is categorical in kind and has one of three following values: 'divorced', 'married', 'single'. Married has the census count with the highest subscribed and unsubscribed to a term deposit, followed by the single while the divorced has the lowest or least subscribed and unsubscribed to a term deposit.



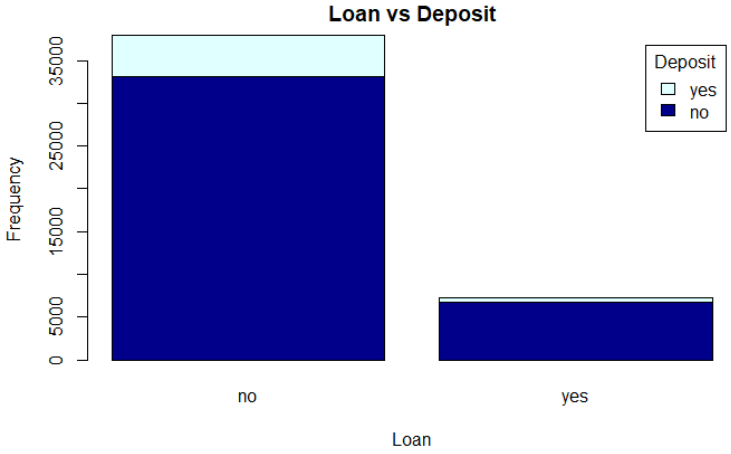
Education variable is of categorical kind and has one of the following values: 'primary', 'secondary', 'tertiary', 'unknown'. The proportion of term deposit acceptance is high in customers with secondary and tertiary education.



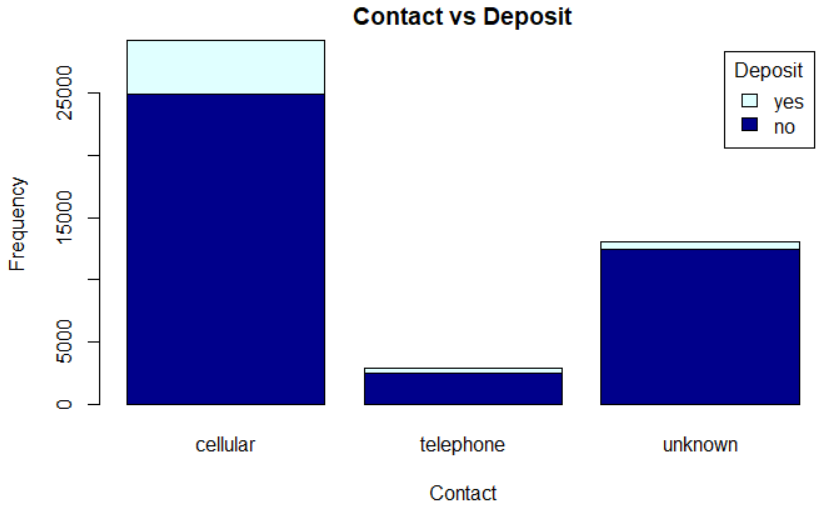
Default variable specifies if the customer has any credit in default. It is a categorical variable. The customers without any credit default are more likely to subscribe to term deposit than customers with credit default.



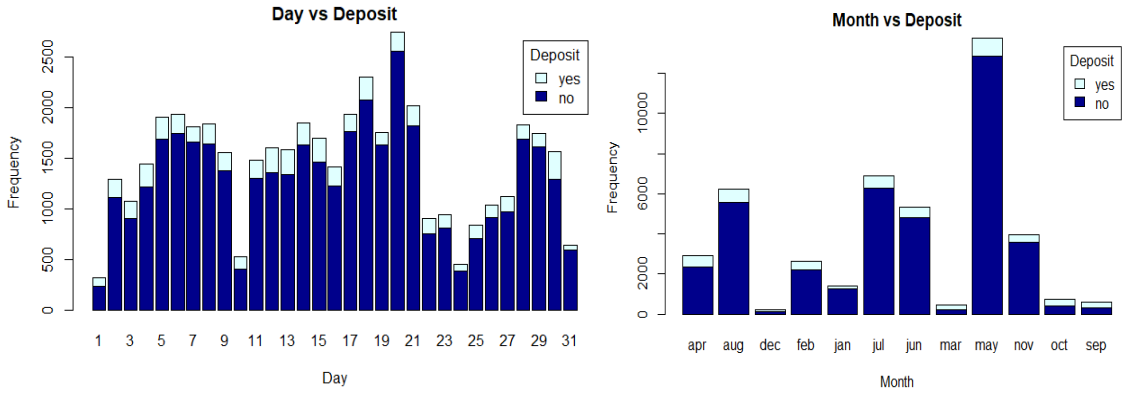
The customers without any personal loan have higher chances of opting for term deposit than those having a personal loan.



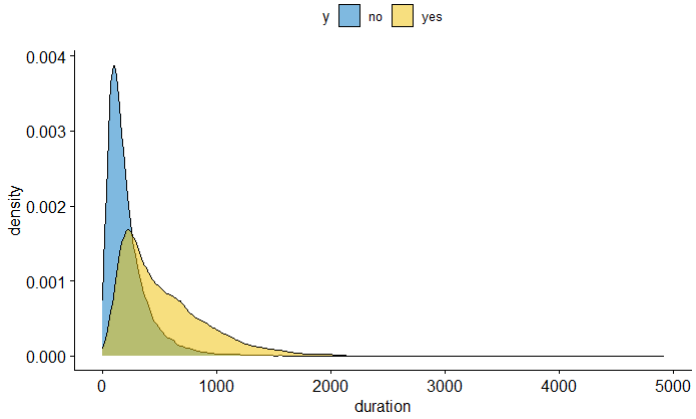
The contact variable has 3 communication types: "unknown", "telephone", "cellular". The cellular has considerably high positive chances for term deposit subscriptions compared to other two.



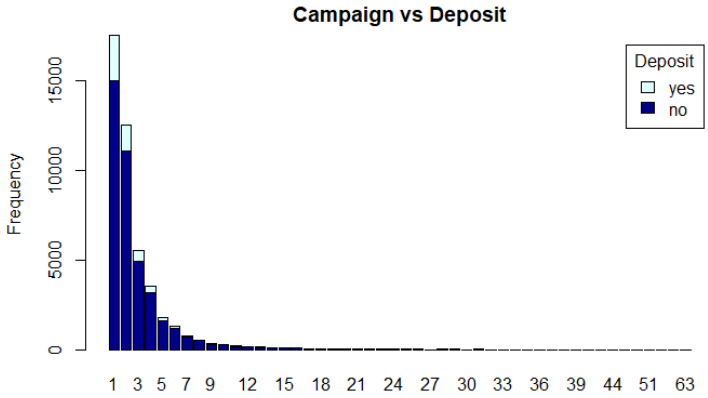
The day and month variables represent when the last contact was made to the customer during this campaign. We can clearly see that the month of May is the month with the highest numbers of subscribed and unsubscribed to the bank term deposit, followed by July and August.



The attribute 'duration' is numeric in kind and has values that range between 0 and 4918. This attribute highly affects the target, for instance, if the duration is equal to zero, it means that no last contact was made within that duration, i.e. y = 'no'. Also, the duration is not known before a call is made but at the end of a call, y is known. So, to build a realistic predictive model this attribute must be dropped.

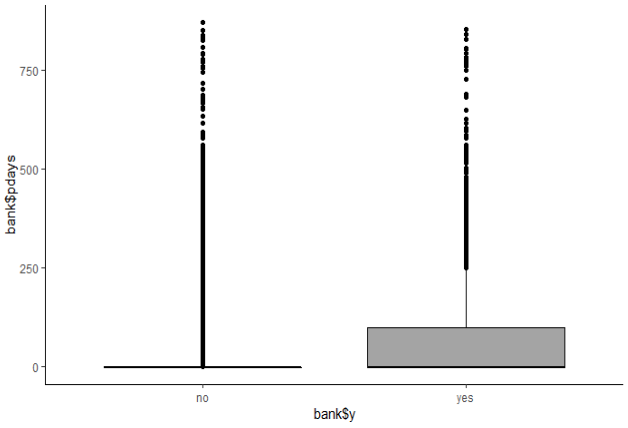


Campaign attribute represents the number of times a specific customer was contacted during this campaign and ranges between 1 to 63. We don't see any significant dependency of customers subscribing to term deposit with the number of contacts made during the campaign.

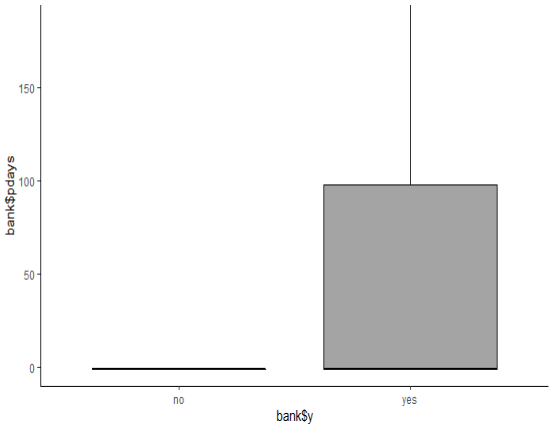


Pdays attribute signifies the days that passed by after the client was last contacted from a previous campaign and it ranges between -1 to 871. -1 represents that they were not contacted before.

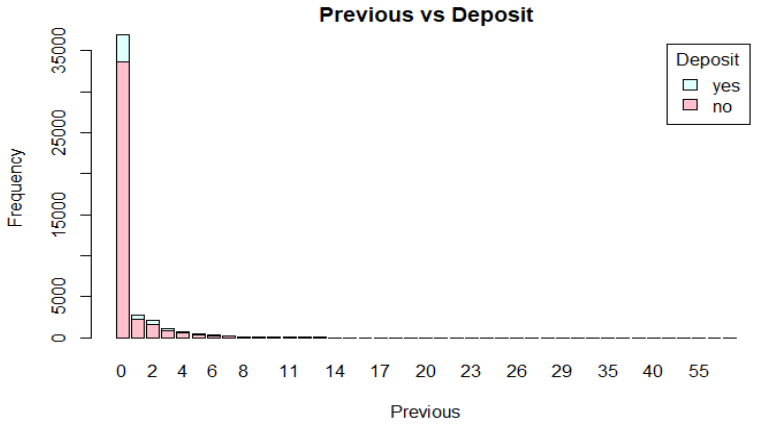
Full data



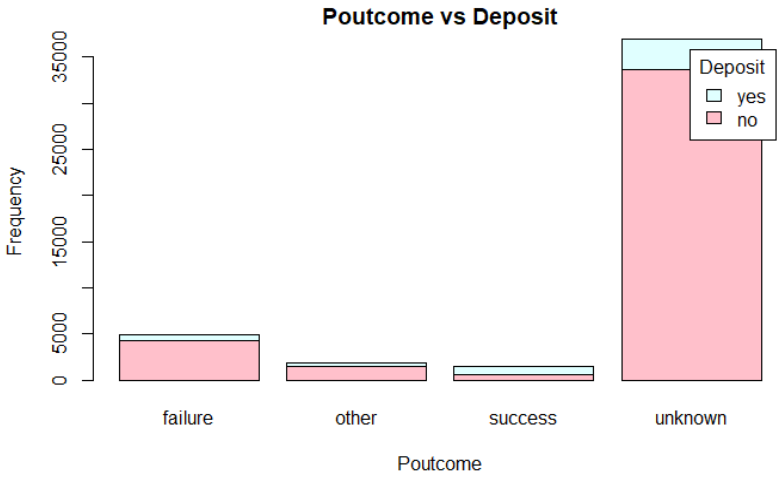
Removing outliers



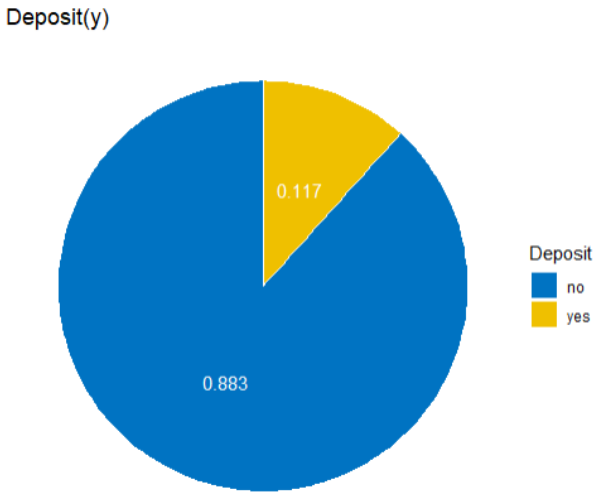
The previous attribute shows the number of contacts made before the campaigns began and values range between 0 and 275. The number of contacts made before the campaign was very small and those who gave their intention to subscribed to the bank term deposit is very less than those who did not intention to participate or subscribe to the bank term deposit policy.



P-outcome attribute has the outcome of the previous marketing campaigns of Portuguese banking institution having the values of failure, other, success and unknown. The proportion of customers who subscribed to term deposit is less dependant of this attribute.



The dataset is highly unbalanced with very less records for the class of interest in the target variable y. Approximately 88% of data has term deposit as 'no' and remaining 12% has customers who subscribed to term deposit. So, we perform resampling techniques to mitigate this issue. The resampling techiques like downsampling, upsampling and smote sampling is performed on the dataset.



Empirical Analysis

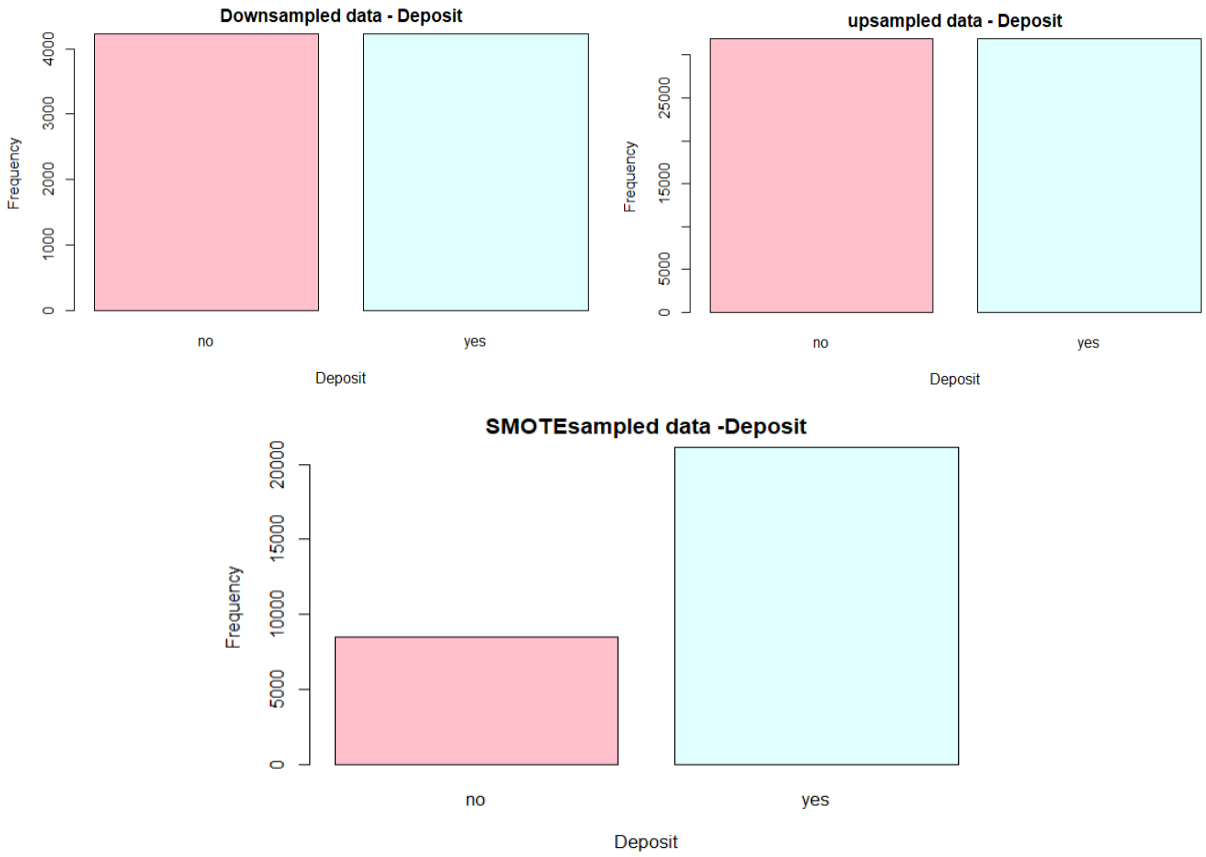
I am going to predict the potential customers of the new term deposit product using the following 4 models.

- Logistic regression
- Random forest
- Support vector machine
- Classification tree

Since the dataset is highly unbalanced the sensitivity, precision is used to compare across models. Sensitivity is used as a main measure to compare across the models and to choose the best one. With imbalanced data sets, an algorithm doesn't get the necessary information about the minority class to make an accurate prediction. So, to deal with this we modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance. This is done by the resampling techniques.

The down-sampling will randomly sample a data set so that all classes have the same frequency as the minority class by randomly removing samples from majority classes. The up-sampling will do sampling with replacement to make the class distributions equal. The SMOTE (Synthetic Minority Over-Sampling Technique) sampling oversamples the rare event by using bootstrapping and k-nearest neighbour to synthetically create additional observations of that event.

We split the data into 80% training and 20% test data and they are scaled to reduce the variance. Then, training data is resampled. In SMOTE sampling, I have added more weights to the "yes" prediction than "no" prediction because the profits are much greater than the cost of obtaining a potential subscriber. Firms care more about the "Yes" prediction.



Using logistic regression model the sensitivity using raw data was 19.48% which jumped to 78% when using resampled data using downsampling and upsampling technique. The SMOTE sampled produces a sensitivity of 86.5%. I have used a cut-off probability of 0.4 in this model.

model <fctr>	accuracy <dbl>	Precision <dbl>	Sensitivity <dbl>	Specificity <dbl>	F1score <dbl>	AIC <dbl>
logistic_regression - Raw data	0.8908307	0.6023392	0.1948912	0.9829659	0.2944961	21681.22
logistic_regression - Downsampled data	0.5865502	0.1910578	0.7842952	0.5603707	0.3072646	9596.90
logistic_regression - Upsampled data	0.5899790	0.1922898	0.7833491	0.5643788	0.3087824	71992.68
logistic_regression - SMOTE sampled data	0.4058179	0.1489015	0.8656575	0.3449399	0.2540961	26410.22

The Random forest model is an ensemble method over decision trees. They are popular for reducing overfitting and providing good estimates. Using this model, the sensitivity using raw data was 24% which jumped to 65% when using downsampled data.

model <fctr>	accuracy <dbl>	Precision <dbl>	Sensitivity <dbl>	Specificity <dbl>	F1score <dbl>	AIC <dbl>
Random_Forest - Raw data	0.8930428	0.6081731	0.2393567	0.9795842	0.3435166	21681.223
Random_Forest - Downsampled data	0.7930539	0.3138152	0.6490066	0.8121242	0.4230651	9538.183
Random_Forest - Upsampled data	0.8642849	0.4286913	0.4834437	0.9147044	0.4544242	72245.879
Random_Forest - SMOTEsampled data	0.7860856	0.2972723	0.6083254	0.8096192	0.3993789	26397.598

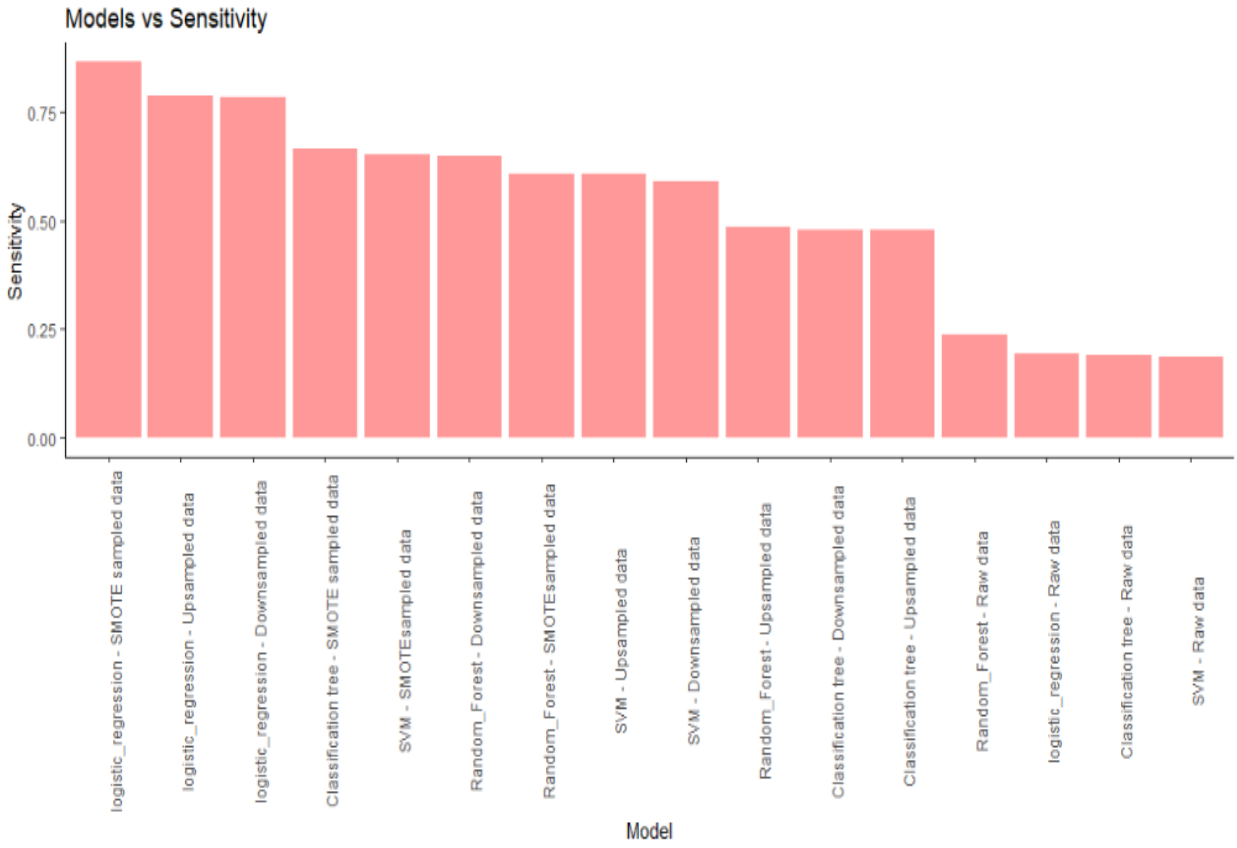
The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems. Using this model, the sensitivity using raw data was 18.8% which jumped to 65% when using resampled data by SMOTE sampling technique.

model <fctr>	accuracy <dbl>	Precision <dbl>	Sensitivity <dbl>	Specificity <dbl>	F1score <dbl>	AIC <dbl>
SVM - Raw data	0.8929322	0.6440129	0.1882687	0.9862224	0.2913616	21681.223
SVM - Downsampled data	0.8230284	0.3485778	0.5912961	0.8537074	0.4385965	9538.183
SVM - Upsampled data	0.8163920	0.3402226	0.6073794	0.8440631	0.4361413	72245.879
SVM - SMOTEsampled data	0.7094348	0.2333560	0.6499527	0.7173096	0.3434141	26397.598

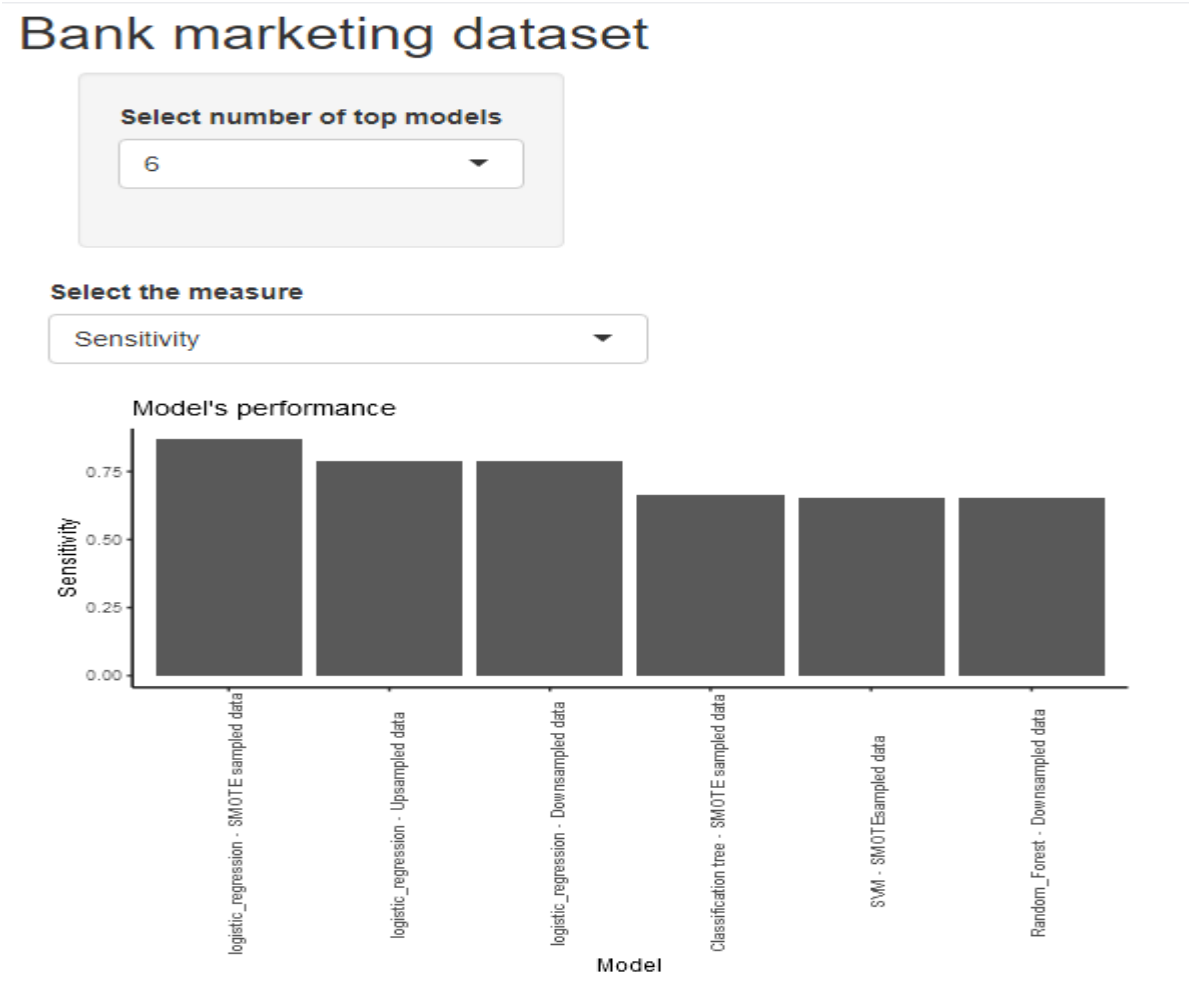
Decision Trees are an important type of algorithm for predictive modeling. Using classification trees algorithm, the sensitivity of raw data was 19% which jumped to 66.5% when using resampled data by SMOTE sampling technique.

model <fctr>	accuracy <dbl>	Precision <dbl>	Sensitivity <dbl>	Specificity <dbl>	F1score <dbl>	AIC <dbl>
Classification tree - Raw data	0.8930428	0.6451613	0.1892148	0.9862224	0.2926116	21681.223
Classification tree - Downsampled data	0.8930428	0.3518776	0.4787133	0.8832665	0.4056112	9538.183
Classification tree - Upsampled data	0.8359695	0.3518776	0.4787133	0.8832665	0.4056112	72245.879
Classification tree - SMOTE sampled data	0.6534675	0.2017241	0.6641438	0.6520541	0.3094556	26397.598

Different performance measures like Accuracy, Precision, Sensitivity, Specificity, F1 score and AIC are included in the results. The model's performance using Sensitivity as the main measure is depicted below.



The results of all the models can be interactively compared and seen using a shiny app in which the user can select the number of top performers(models) and the performance measure as they need.



Conclusion

By careful observation and analysis of the results it can be concluded that the logistic regression model has the best performance in predicting the potential customers who would subscribe to the term deposit product of the banking institution by producing sensitivity of 86.5% by sampling the data using downsampling, upsampling and combined methods (SMOTE sampling). The further fine tuning of the parameters in the models can produce more better results.

Sources:

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

https://topepo.github.io/caret/subsampling-for-class-imbalances.html

https://r4ds.had.co.nz/exploratory-data-analysis.html

https://ggplot2.tidyverse.org/reference/index.html

https://community.rstudio.com/t/shiny-reactive-input-for-multiple-input-selectors/35157