REPORT

The goal of this study is to predict the success of telemarketing calls for selling long-term deposits. A Portuguese bank wants to utilise its client details to understand what type of customers are more willing to subscribe to their long-term deposit product. To achieve the aforesaid goal, various machine learning models like logistic regression and trees were developed to predict the whether the customer would subscribe to the product. The models also gave insights into the impact of various features on the outcome.

The bank dataset consists of 4521 observations and 17 variables, like age, job type, marital status, education, present balance, has housing loan or not etc.

Data Preparation:

The following variables need to be removed from the dataset as they are not useful for analysis purpose:

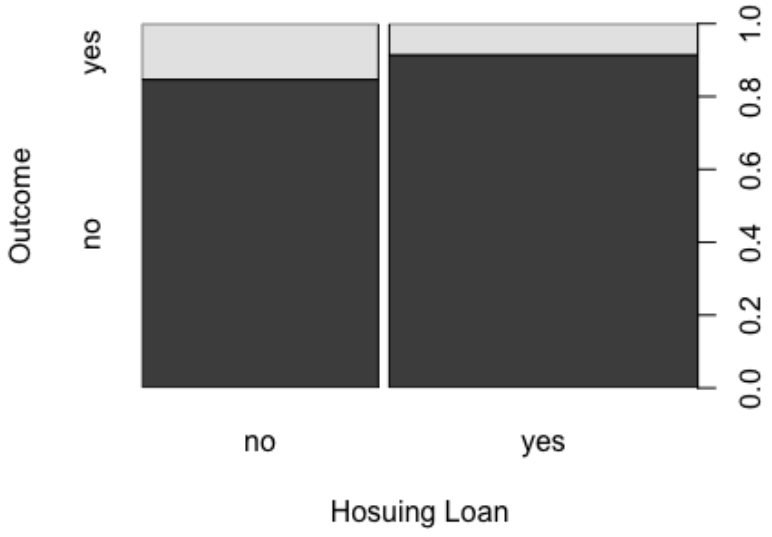
- pdays: 75% of the values are -1 (not previously contacted)
- previous: 75% of the values are 0 (75% of clients never contacted before)
- poutcome: 87% of the observations fall in unknown/other category
- durartion: can be known only after making the call - not useful for prediction purposes

The reason for removing pdays, previous & poutcome variables is that there are too many observations that have the same value which would prevent us from making any concrete conclusion on the basis of these variables. The duration of the call can be known only after making the call thus it is not helpful in predicting if a customer would subscribe to the product or not.

Exploratory Data Analysis (EDA):

Following are the EDA inferences:

- Customers who do not have any housing loan are more likely to invest in the long-term deposits than those who have a housing loan. 57% of the customers who subscribed to the long-term deposit do not have any housing loans.



- Retired professional and students are more likely than others to subscribe to long-term deposits. 23.5% of the retired and 22.62% of students contacted eventually bought the bank's product.

- Divorced and single category customers are more likely than married customers to invest in long-term deposits.

- Customers with tertiary level of education are more likely to invest in long-term deposits than customers with primary or secondary level of education.

- Customers with higher balance in bank (Median:710) are more likely to subscribe to the long-term deposits.

- Customers who do not have any housing loan are more likely to invest in the long-term deposits than those who have a housing loan. 57% of the customers who subscribed to the long-term deposit do not have any housing loans.

- Customers who do not have any personal loan are more likely to invest in the long-term deposits than those who have a personal loan. 91% of the customers who subscribed to the long-term deposit do not have any personal loans.

- Half of the customers who subscribed to the long-term deposits did so after just 2 calls. Moreover, half of the customers who did not subscribe to the bank's product had made up their mind by the 2nd call. Thus, it can be concluded that more than 2 or 3 calls will not improve the likelihood of the product being subscribed to.

Modelling Phase:

The data set was split into training (80%) and test (20%) data set before the modelling process could be initiated. The data set contains 4000 instances of the product not being subscribed to and only 521 instances where the product was subscribed – thus the dataset is quite imbalanced. With approximately 88.47% of the observations labelled as "no" – machine learning algorithms tend to give biased results when trained on such imbalanced data. Thus, the models were first trained on original training data and then on modified training data (generated using oversampling) and then were used to predict for test data (unchanged).

Following is the performance of models trained on original data:

Model	Accuracy	Area Under Curve (AUC)
Logistic Regression	87.98%	0.506
Classification Tree	87.03%	0.566
Random Forest	87.91%	0.537

With 88.47% observations belonging to the "no" class – all the models have misclassified most of the "yes" class (minority class) observations and thus the accuracy is high.

For example, following is the confusion matrix of logistic regression:

	actual	
predicted	no	yes
no	1192	163
yes	0	2

Basically, (almost) all the unseen observations have been classified as “no” and hence even such a biased model is getting good accuracy. The metric to judge the model performance in the current scenario should not be accuracy but the AUC.

Following is the performance of models on test data (unchanged) that were trained on modified train data:

Model	Accuracy	Area Under Curve (AUC)
Logistic Regression	68.31%	0.666
Classification Tree	74.64%	0.642
Random Forest	80.69%	0.703

Even though the accuracy has dropped but there is a massive improvement in the AUC values. What that means is that with the models trained on modified data, the “yes” class (minority class) predictions on unseen test data are much more accurate. Moreover, from business perspective, it is better to have more “yes” class observations classified correctly as every “yes” class observation classified as “no” would mean loss in business.