Data Mining Techniques for Bank Direct Marketing

Aniketh Kulkarni, Roshan Lalwani, Xuan Dong

ABSTRACT

Modern businesses are always looking for leaner and more efficient ways of marketing. The rate of conversion achieved via traditional marketing strategies such as mass marketing are just not good enough for certain industries. Banking is one such industry.

In banking, direct marketing techniques have long replaced more traditional ones. Direct marketing is a method of contacting customers and potential customers personally, rather than having an indirect medium between the company and the consumer, such as magazine ads or billboards that are seen by the general public. Direct marketing can take many forms, including mail, telephone calls, emails, brochures, and coupons. For our project, we are using a publicly available dataset. containing real time data about one such direct marketing campaign (phone calls) of a Portuguese banking institution.

INTRODUCTION

According to Oleksandra Onosova in his research, "A well-executed direct advertising campaign can provide a positive return on investment by motivating customers to respond to a call to action. One of the preparatory stages ensuring a high return on investment consist of carefully querying a marketing database to generate a target list of respondents". The writer agrees that using modeling technique with marketing database will lead to a successful direct marketing campaign with maximum return. He also points out in his article that measuring negative impact of unsuccessful direct marketing can also help banks to narrow down customer list, which in turn will reduce the over-all cost. (Onosova, 2016)

It is clearly stated in the Dejana Pavlovic' research that "the use of data mining models is important for accomplishing goal of the companies such as to gain competitive advantage on the market". Being able to understand customers' characteristic is important for the successful campaign. (Pavlovic, 2015)

In our case, the business problem that inspired this initiative is that the bank needs to find a model that helps them create a selection of an affordable set of potential buying customers who are quantifiably more likely to respond positively to their marketing campaign by subscribing to a term deposit. This needs to be done in order to better manage available resources such as human effort, phone calls and time. The business decision makers are also pushing for a reduction in costs and an increase in efficiency such that fewer phone contacts need to be made while retaining the same number of successes. Since this is a public dataset from the UCI Machine Learning repository, the name of the banking institution has been left anonymous by the bank and hence the stakeholders are unknown, however these most likely constitute key decision makers at the

executive level who shape the marketing strategy of the company. It is also safe to assume that they are on board with this modeling project since they made their dataset publicly available for research. Since the advent of targeted marketing, banks collect vast amounts of data about their customers. This makes it very easy to convert this business problem into a highly feasible analytics challenge.

From an analytics standpoint, this is a classification problem. The target variable which is: "has the client subscribed to a term deposit?" is a 0/1 target variable where 0 indicates that the marketing phone calls did not lead to the customer making a term deposit, i.e., a failure, and 1 indicates that the customer did make a term deposit, i.e., success. To achieve what the business needs, we must build a prediction model that can predict whether any given customer, based on their attributes, will make a term deposit in response to the bank's telemarketing or not.
In order to build an accurate model, we will be trying several modeling techniques that are conducive to classification type problems such as logistic regression, random forests, neural networks, decision trees, etc.

The rest of the paper is organized as follows: in the following section, we review the existing literature around the subject. We found some research papers in the same area that we are working in. From those prior researches, we identified what aspects have been already known and what aspects need further investigation. In the section 3, we briefly introduced data we are using for our model. In the section 4, we described what modeling methodologies we implemented for our predictive models. Subsequently, we introduced all those predictive models we built for the business problem in depth. In the section 6, we showed results from our model to evaluate the performance of each model based on various parameters and to identify the best model. We then conclude the study by summarizing our findings and discussing about the future scope of studies.

LITERATURE REVIEW

The business problem and the solution exist in the intersection of the Banking industry, the Marketing domain and the solution involves predictive modelling using different machine learning algorithms. In this section, we investigate the existing research in each of these areas and summarize the work of various scholars so that we can build upon them.

The term direct marketing was first suggested by Lester Wunderman in 1967 and he is considered to be the father of direct marketing [1]. For some industries, including the banking industry the traditional marketing methods prove to be inefficient and often expensive. Hence, banking institutions have quickly adopted more direct approaches to reach out to their customers through multiple channels like e-mails, direct mails and telephone calling. With the advent of technology and the increasing digitization of data, banks have been able to leverage data mining techniques in various ways to improve operations in many domains such as Attrition Reduction, Delinquency Management, Cross-Selling etc. [2]. Data mining has been used widely in direct marketing to identify prospective customers for new products. By using purchasing data, a

predictive model could be built to measure if a customer is going to respond positively to the promotion or the offer [3]. The three data variables that have the largest impact on odds ratios are job, default, and loan. The baseline level for job is 'unemployed'. All but 3 job categories that varied from unemployed were 2.5 to 5 times more likely to make a term deposit purchase than an unemployed person. Housemaids, entrepreneurs, and self-employed persons were no more likely to make a purchase than an unemployed person. For Client Data variable default, a person who reports they have no credit in default is 3 times more likely to purchase a term deposit than a person who reports that it is unknown whether they have credit in default. For Client Data variable loan, a person who reports that they don't have a personal loan is 3 times more likely to purchase a term deposit that a person that reports they have a personal loan [14].

Methods for analyzing and modelling data can be split into two groups: supervised learning and unsupervised learning. The supervised learning requires input data that has both predictor (independent) attributes and a target (dependent) attribute whose value is to be estimated. In addition, the process learns how to model (predict) the value of the target attribute based on predictor attributes. The famous examples of supervised learning are decision trees, and neural networks. Actually, the supervised learning is suitable for analysis dealing with the prediction of some attribute [4]. One key observation to note about the dataset is that only 11% of all calls result in a sale (i.e. classified as "yes"), and the remaining 89% do not (i.e. "no"). This is an example of an imbalanced data, as the classes here ("yes", "no") are not represented equally in the dataset. The imbalanced nature of the dataset may result in model accuracy being incorrectly measured, depending on the metric used. As an example, a model could be correct 9 times out of 10 simply by predicting "no" on every test record, but such a model would provide no practical value [8].

One of the methods to achieve these predictions is to build a multi-layer perception neural network. Multi-layer perception neural network is a mutually dependent group of artificial neurons that apply a mathematical or computational model for information processing using a connected approach to computation [4]. Another popular technique used for classification problems is Decision Trees. It can generate understandable rules, and to handle both continuous and categorical variables [5]. One of the famous recent techniques of the decision tree is C5.0, which was applied in the paper by Hany A. Elsalamony [6]. Logistic Regression is also another popular method used in classification problems. It is a largely simplistic approach with very little flexibility, but it is usually robust and reproducible. Logistic regression is a mathematical modelling approach that can be used to describe the relationship of several X's to a dichotomous dependent variable, such as D. The fact that the logistic function f(z) ranges between 0 and 1 is the primary reason the logistic model is so popular. The model is designed to describe a probability, which is always some number between 0 and 1 [7]. The paper by Alving choong et all. [8] also use more advanced methodologies like Multivariate Adaptive Regression Splines (MARS), Stochastic Gradient Boosting (GBM) and Random Forest. The models built with less flexible algorithms like Decision Trees and Logistic Regressions, achieve low values of accuracy and area under the ROC Curve. For example, in the paper by Sergio Moro et all. [9], Logistic Regression model was only able to

achieve an area under the curve of 0.715 whereas the support vector machine model built by Oleksandra Onosova achieved a direct increase of 41% in the profits through the same marketing effort [10]. In our case, we seek to build three models: 1) Random Forest, 2) Stochastic Gradient Boosting (GBM) [8] and 3) Multi-layer Perception Neural Network [9]. Although an SVM model was attempted to be trained on the dataset on the lines of Cédric Archaux et all [11], this proved to be computationally infeasible with the available resources.

The model performance measures chosen appropriate to evaluate these models were chosen to be Accuracy and Area under the ROC Curve. The receiver operating characteristic (ROC) curve shows the performance of a two-class classifier across the range of possible threshold (D) values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [12]. The overall accuracy is given by the area under the curve ($AUC = \int_0^1 ROC.dD$), measuring the degree of discrimination that can be obtained from a given model. AUC is a popular classification metric [13] that presents advantages of being independent of the class frequency or specific false positive/negative costs. The ideal method should present an AUC of 1.0, while an AUC of 0.5 denotes a random classifier. Some of the models built have also taken into account a cost factor to measure the direct financial impact of such modelling on the business [10]. This method is more relevant than the other statistical performance measures because this metric essentially links the model results directly to the business objectives. However, it is often not feasible to arrive at a single cost factor as there are numerous variables involved and there is no single cost that can be applied to all instances. Therefore, for our model, we have ignored the cost factors due to unavailability of data.

DATA

A first look at the data indicated that the dataset comprised of 41188 observations across a target variable and 20 predictor variables. Further, the predictor variables comprised of unexplained categories, such as unknown and others, at multiple instances, and while there were no missing values within the same, negative values did exist for certain variables.

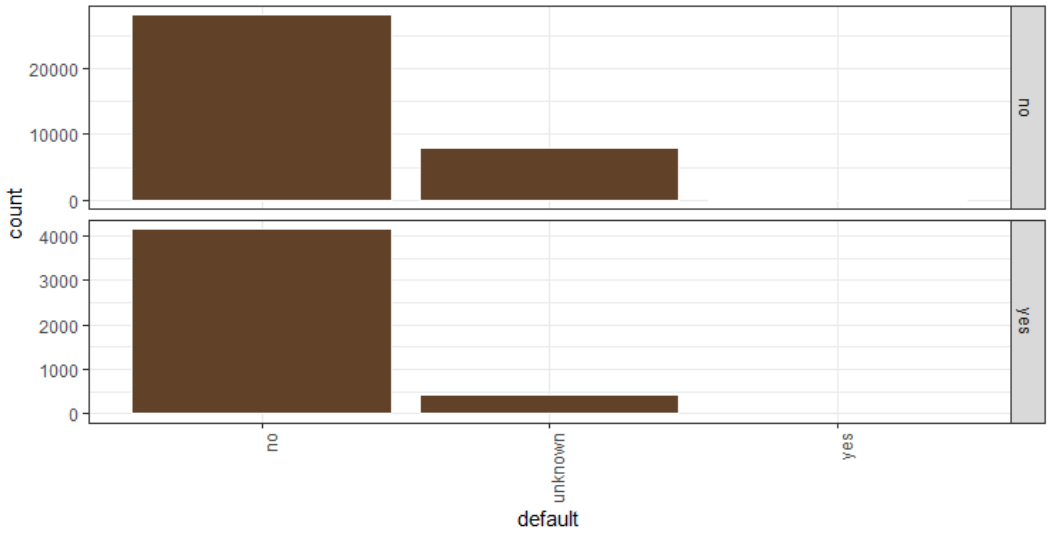
To understand the significance of such variables within the context of the business and analytical problem, we further analyzed each variable separately:

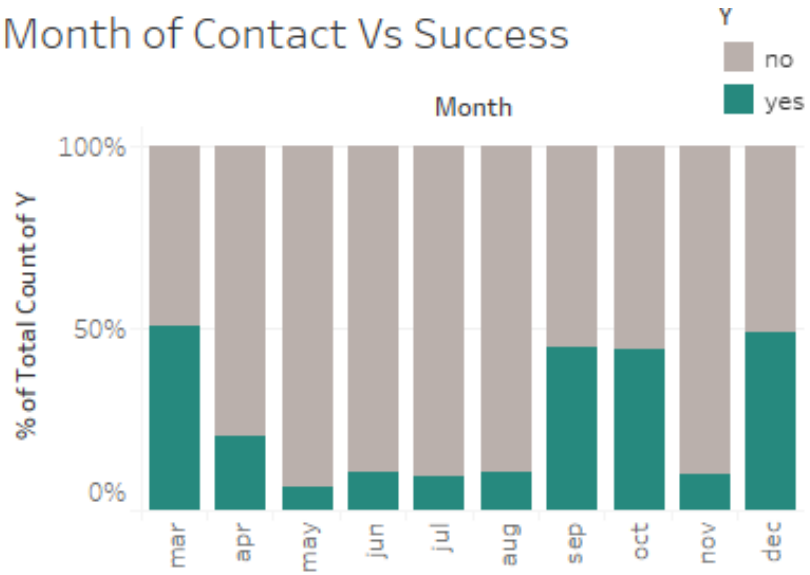
1. The **age** variable reflects a customer's age in years. A plot of the related values suggested a minimal number of occurrences for ages below 18 years and above 90 years, and the same were, hence, removed.
2. **Job** variable provides the work profile of the customer under 11 specified and 1 unspecified category. The unknown category within this head comprised of only 330 observations and was excluded from the analysis.
3. The marital status of a customer is reflected within the **marital** column. The unknown category within the head was removed since it comprised of a small number of observations.
4. The **education** variable comprised of a significant number of observations within the unknown category, which seemed to have an equal division within the two instances of the target variable. Since such category seemed to contribute to the determination of the target variable,

the same was retained. On the other hand, the illiterate category only had a small few values and was removed.

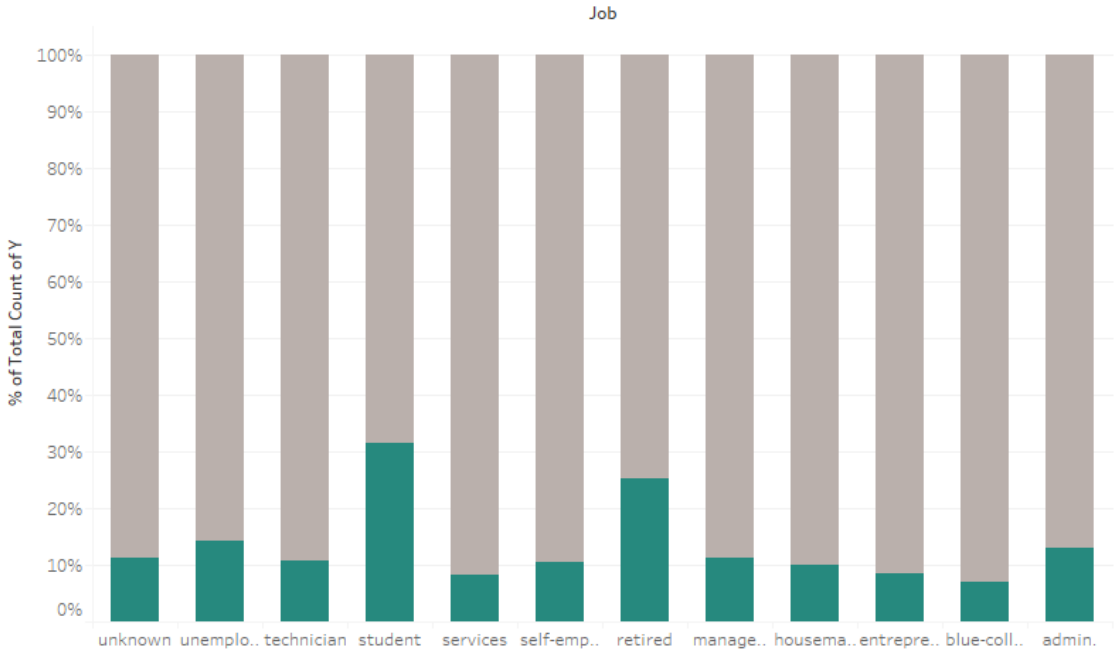
5. The **default** variable, indicating whether a customer had credit in default, had a considerably high number of observations marked unknown, which were, hence, retained. Alternatively, since only 3 instances of 'yes' existed for this column, the same were removed being insignificant for the analysis.
6. The **housing** variable indicates whether a customer has a housing loan or not, with the unknown category claiming certain observations. Being small in number, the unknown category was removed from the analysis.
7. The **loan** variable answers the question whether a customer has a personal loan or not, further including a small number of unknown observations that were eventually removed.
8. The **contact** variable shows the mode of contact for a customer.
9. The **month, day of week** and **duration** variables reflect the last contact time and duration of call for a particular customer. While the first two variables were retained as is, the third variable was dropped. For the dropped variable, while outcome was certainly unknown at duration = 0, it would be certainly known at the end of the communication.
10. The number of contacts performed for a customer during a campaign were recorded under **campaign**. Assuming that contacts above 10 may be atypical and possibly an error, the same were dropped from the variable.
11. The number of contacts performed for a customer during a previous campaign and the related outcome are recorded under **previous** and **poutcome**.
12. Certain variables within the data served as macroeconomic indicators and were removed being unnecessary for the prediction model.

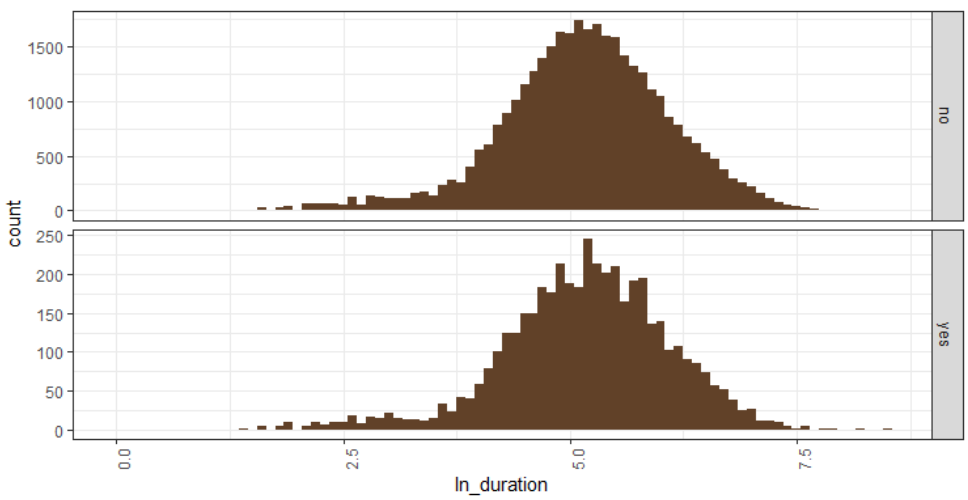
Subsequently, categorical variables were split into separate columns based on their specific values, and the original columns were removed to avoid duplication. The dimensions of the dataset, following the cleaning process, are 38717 observations over 55 variables. Below plots show some of the significant insights obtained through the EDA.





Job vs Success - Normalized

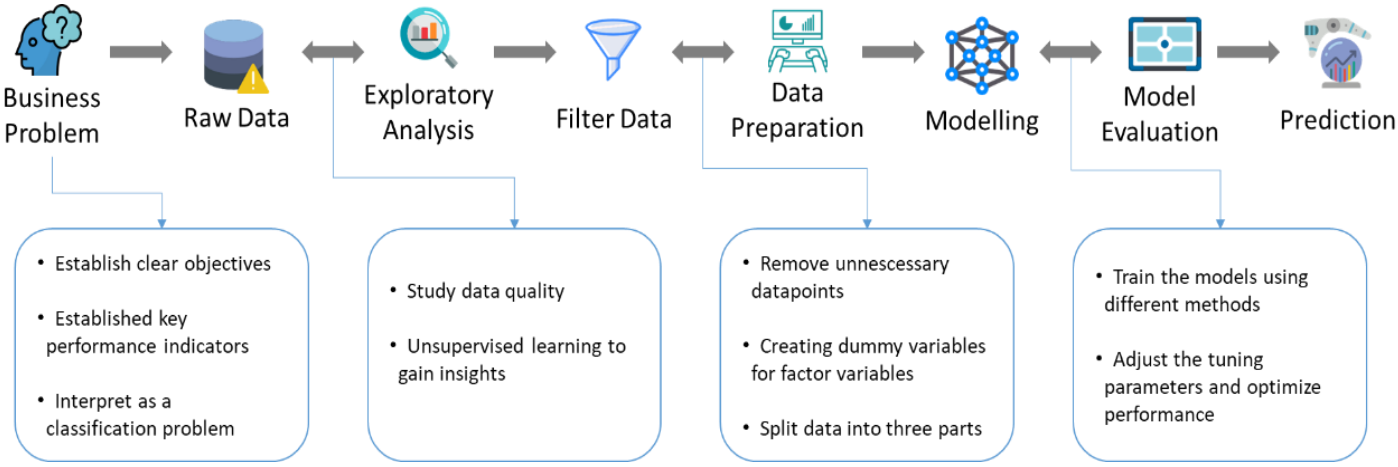




METHODOLOGY

For the chosen business problem, we adopted the following workflow methodology based on the CRISP DM framework:

1. Business Understanding: The first step was to establish clear objectives for the task and to evaluate the nuances specific to the domain. We also established key performance indicators to measure the impact. This step also involved studying the existing literature in the field of direct marketing. Subsequently, this business understanding was then interpreted as a classification problem.
2. Data Understanding: The data was read into the R environment and basic tests on the data were carried out to determine the quality of data. We also carried out the exploratory data analysis to uncover hidden insights in the data as part of our unsupervised learning. We discovered that the likelihood of a customer subscribing to the term deposit was highly influenced by variables like the month of contact, age, jobs etc.
3. Data Preparation: The data did not contain any missing values but many variables such as marital status, previous outcome etc. contained values that were unknown. These were mostly not helpful in explaining variance in the target variable and hence were dealt with on a case by case basis. After all the data cleaning, we created the dummy variables for the factor variables and removed the original columns. Since the data consists of only 20 features, we retained most of these. Finally, after the data cleansing, we were left with 38717 observations over 55 variables.
4. Following the data cleaning and pre-processing, we split the dataset into train, test and validations sets. Generally, a dataset is split into train and test sets, where the former is used to train the model, which is subsequently used to predict the outcome for the latter. In such case, it is likely that the training error is significantly different from the testing error, and the former may underestimate the latter. Accordingly, for the instant scenario, validation set approach is used which holds out a portion of the train set from the fitting process and then applies the trained model to this subset. The validation set error provides an estimate of the test error, i.e., an unbiased evaluation of the model fit on the training dataset. As a next step, for the model training portion, we have used the remote H2O connection for enhanced efficiency and uploaded the three datasets thereon. Our three datasets are evaluated using three models: gradient boosting model (“GBM”), neural network (“NN”), and random forest (“RF”). The efficiency of the models is determined using the area under the curve (“AUC”) measure noticed for all three mechanisms.
5. The key performance indicators for these models were chosen to be confusion matrix-based parameters like the area under the ROC curve, Accuracy etc. as these measures are suitable for classification problems like this one. Measures like RMSE and R^2 values are more suitable for regression type problems.



MODEL(s)

Three algorithms were chosen for the analysis based on the best performance. These models provide a mix of varying flexibility and interpretability.

GRADIENT BOOSTING MODEL

GBM is a machine learning technique for regression and classification models that produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The method allows us to update our predictions based on a specified learning rate that can minimize the mean squared error (MSE) for the model. It is an algorithm that repetitively leverages the patterns in residuals and strengthens a model with weak predictors to make it better. The tuning parameters for the GBM model were chosen to be the following: ntrees = 500, max depth = 15, learning rate = 0.03 and nbins = 100. These values were arrived at after several iterations to balance the computational feasibility, accuracy and the fit.

NEURAL NETWORK

The mechanism of NN processes data information in a parallel manner, which enables it to infer meaning and detect patterns from complex datasets. The first layer of a NN receives the raw input, processes it and passes the processed information to the hidden layers. The hidden layer passes the information to the last layer, which produces the output. The mechanism is adaptive; it trains itself from the data with known outcomes and optimizes its weight for better predictions from data with unknown outcomes. The neural network was tuned with 2 hidden layers with 200 nodes each and 100 epochs to ensure sufficient accuracy without over-fitting.

RANDOM FOREST

RF if a regression and classification tool that generates a forest of regression or classification trees given a dataset. Such trees are weak learners built on a subset of rows and columns, and the increase in the number of the trees reduces the variance. Both classification

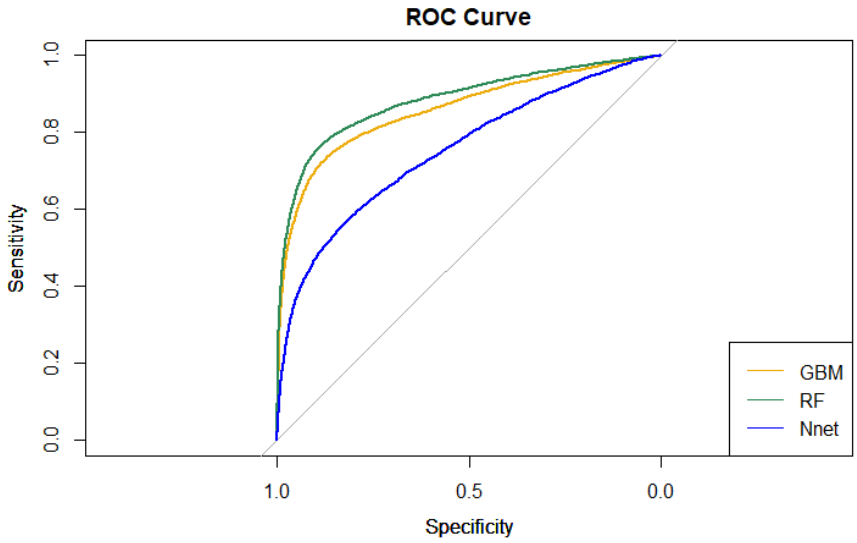
and regression take the average prediction over all of their trees to make a final prediction, whether predicting for a class or numeric value. For the current dataset, the model achieves the highest accuracy under the gradient boosting model; a higher AUC and lower MSE is recorded. The Random Forest was grown with a maximum depth of 20 and number of trees as 500 and number of bins as 200 to ensure that the computation is feasible to run and to ensure that the model does not over-fit.

RESULTS

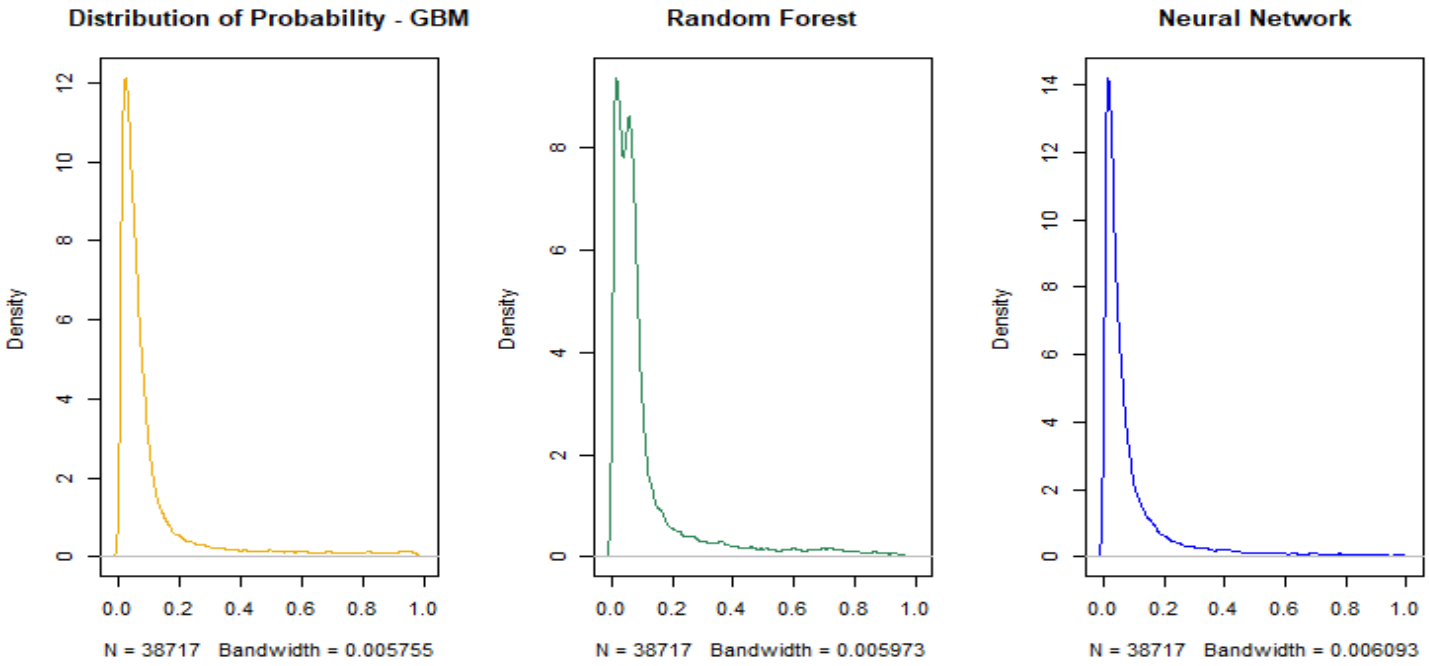
The models were evaluated based on their performance on various metrics like RMSE, MSE, Accuracy and the Area under the ROC curve. Overall, the Random Forest model performs best on the validation dataset with an AUC of 0.7685 and the highest accuracy of 0.8661. The various performance measure values for each of the models are given in the table below:

Model	Random Forest		Neural Network		GBM	
Dataset	**Train**	**Validation**	**Train**	**Validation**	**Train**	**Validation**
Accuracy	0.8648	0.8661	0.8675	0.8671	0.9529	0.8585
AUC	0.7556	0.7685	0.7811	0.7291	0.9772	0.7452
RMSE	0.2901	0.2895	0.2835	0.3012	0.2018	0.2956
MSE	0.0841	0.0838	0.0804	0.0907	0.0407	0.0874
LogLoss	0.2981	0.2981	0.2911	0.3328	0.1523	0.3145

Based on the difference in the accuracy values of GBM model on the train and the validation set, it can be seen that the model is nearly over-fit. Using this model for the final predictions could be potentially problematic. The Random Forest model performs best in terms of the area under the ROC curve as can be seen from the ROC curves below:



All three models perform similarly in terms of the calibration of probabilities with the Random Forest model giving a higher density in the high probability values region. This is useful



CONCLUSION

Through this exercise, we were able to develop models with varying degrees of flexibility, accuracy and interpretabilities to predict with considerable accuracy, the probability of a customer positively responding to a direct marketing campaign. This information can be used by various stakeholders to make data driven decisions on the time and effort to be invested in pursuing a particular customer. Overall, the organization can expect a significant increase in the efficiency and decrease in the customer acquisition cost of their direct marketing campaign with the right amount of effort reaching the customers who are more likely to avail a term deposit.

This modelling effort could be followed up by designing an application that could allow a user to input customer information to generate a prediction for the probability. This would enable the model to be deployable in the business and drive necessary impact for the business.

The model could be improved upon by increasing the volume and/or by inculcating more input features as and when they are available. With sufficient computing power, we could also try more computationally complex algorithms such as support vector machines to improve the accuracy further. Some further investigation could also be carried out upon the impact of clustering algorithms on the prediction accuracy of the models.

REFERENCES

1) Wikipedia contributors. (2018, July 29). Direct marketing. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:08, October 13, 2018, from - *https://en.wikipedia.org/w/index.php?title=Direct_marketing&oldid=852554263*
2) Disha Budale, Dashrath Mane (2013, June). Predictive Analytics in Retail Banking. In *International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013*

3) Eniafe Festus Ayetiran, "A Data Mining-Based Response Model for Target Selection in Direct Marketing", *I.J.Information Technology and Computer Science, 2012, 1, 9-18*
4) T. Munkata, "Fundamentals of new artificial intelligence," *2nd edition, London, Springer-Verlag, 2008.*
5) Su-lin PANG, Ji-zhang GONG, C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks, *Systems Engineering - Theory & Practice, Volume 29, Issue 12, Pages 94–104, December 2009.*
6) Hany A. Elsalamony, Bank Direct Marketing Analysis of Data Mining Techniques, *International Journal of Computer Applications (0975 – 8887) Volume 85 – No 7, January 2014*
7) D.G. Kleinbaum and M. Klein, Logistic Regression, Statistics for Biology and Health, *DOI 10.1007/978-1- 4419-1742-31, Springer Science Business Media, LLC 2010*
8) Alvin Choong, David Menezes, Frank Devlin, Mudit Gupta, Tan Wei-Chyin and Kate Chen Predictive Analytics in Marketing - A Practical Example from Retail Banking, *Singapore Actuarial Society - SAS Big Data Committee, Research Note #1 (Oct 2017)*
9) Sérgio Moro, Paulo Cortez, Paulo Rita - A data-driven approach to predict the success of bank telemarketing, *ISCTE-IUL, Business Research Unit (BRU-IUL), Lisboa, Portugal*
10) Oleksandra Onosova, Maximizing Return On Direct Marketing Campaigns in Commercial Banking
11) Cédric Archaux, Hicham Laanaya, Arnaud Martin, Ali Khenchaf, An SVM based Churn Detector in Prepaid Mobile Telephony, *Bouygues Telecom, 20 quai du point du jour, 92640 Boulogne Billancourt, France*
12) Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett., 27(8):861– 874, 2006.*
13) David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. Performance of classification models from a user perspective. *Decis. Support Syst., 51(4):782–793, 2011.*
14) Gina Colaianni, Jonas Magdangal, Matthew Mitchell, Factors Determining Term Deposit Purchases, December 31, 2016
15) Ayetiran, E. F. (2012, Jan). *A Data Mining-Based Response Model for Target Selection in Direct Marketing*.
16) Contributors, W. (2018, Jun 13). *Direct marketing*.
17) Flici, A. (2016). *A Direct Marketing Framework to Facilitate Data Mining Usage for Marketers: A Case Study in Supermarket Promotions Strategy* .
18) Mane, D. B. (2013, June). *Predictive Analytics in Retail Banking. International journal of Engineering and Advance Technology*.
19) Onosova, O. (2016). *MAXIMIZING RETURN ON DIRECT MARKETING CAMPAIGNS IN COMMERCIAL BANKING* .
20) Pavlovic, D. (2015). *Application of data mining in direct marketing in banking sector*.