# 

# **MH6151-Data Mining**

# 

# **Analysis of Bank Telemarketing Approach**

|  |  |  |
| --- | --- | --- |
| Member Name | Percentage of Contributions | Justification |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# 

# ***Abstract***

# The use of telemarketing campaigns to obtain deposits has become a common approach for banks. In addition, to focus manpower and maximize the output from the telemarketing campaign, research has been done to focus (target) the resources on the customers more likely to have a positive response to the campaign.In this report we will explore the data collected from a Portuguese retail bank pertaining to the success of the telemarketing campaign. Using various machine learning models we will attempt to analyze two situations, that is, after the last contact with the customer and after the campaign (to scout for potential customers for future campaigns).

# **1. Introduction**

Telemarketing is the way of selling goods and/or services using phone call,emails,sms or internet. Recently, phone calls have become the most widely used means to managing the telemarketing services.Using telemarketing helps to gauge the level of interest of the customers as well as helps to provide a more interactive and personal sale service. Due to its cost-effectiveness and high efficiency the use of telemarketing has become widespread in various sectors such as real estate, banks, non-profit organisations etc.

Moreover, in addition to its various advantages this marketing strategy allows the service-providers to focus the resources available in an optimal manner, that is, by focusing on a pool of potential customers highly likely to subscribe to the service/goods based on various evaluations/characteristics of the customer information as well as several other associated factors. However, the service providers cannot determine which customers will subscribe just looking through their database.

To have some idea of interested customers, supervised predictive/classification machine learning algorithms can be utilized to forecast/predict the success of the telemarketing approach, based on the information of the customers. By narrowing the pool of potential customers, resources will be focused more effectively and increase the efficiency of the telemarketing campaigns.

# **2. Problem Definition and Business Understanding**

In this project/report we are going to explore the bank marketing dataset, which contains the information from a telemarketing campaign, conducted by a Portuguese retail bank. The dataset contains the 45,211 observations from years 2008-2013, containing 16 attributes associated with the retail customer,economic attributes,etc. and 1 target class indicating whether the customer subscribed to the term deposit.

A term deposit is a type of deposit account held at a financial institution, where the money is locked up for a specified amount of time. The length of time is usually short (1 month to 1 year) and this type of investment is considered very safe.

In this report, we aim to create conduct two types of analysis based on the available information. One is mid-campaign (micro) analysis where we utilize the information available during the campaign to predict the possibility of the customer subscribing to the term deposit after a call. Another is the post-campaign (macro) analysis where we perform the analysis to identify potential consumers/customers who are more willing to subscribe to the deposits for future campaigns. Henceforth, we refer to the micro analysis as Business Model 1 and macro analysis as Business Model 2.

To perform these analysis we develop various supervised machine learning models, Support Vector Machine (SVM), Logistic Regression, Decision Tree, Naive Bayes and Random Forest. By comparing the performance of the various models, we attempt to improve the efficiency of the telemarketing campaign by selecting a model with high predictive power. Further such analysis such as exploratory data analysis is also performed, to further study the distribution and effect of these variables.

# **3. Experimental Evaluation**

## **3.1 Methodology**

### 3.1.1 Data description

|  |  |  |  |
| --- | --- | --- | --- |
| Response Variable | description | type | e.g. |
| y | has the bank client subscribed to a term deposit? | binary | yes, no |

|  |  |  |  |
| --- | --- | --- | --- |
| Explanatory Variable | description | type | e.g. |
| age | Age of client | numeric | 58,44,33,47 |
| job | type of job |  | admin., unknown, unemployed,management,housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services |
| marital | marital status | categorical | married,divorced,single; note: "divorced" means divorced or widowed |
| education | Highest education level of bank client | categorical | "Primary","secondary",  "tertiary","unknown" |
| default | Does the client have credit in default? | binary | yes, no |
| balance | average yearly balance, in euros | numeric | 0, |
| housing | Does the client have existing housing loan? | binary | yes, no |
| loan | Does the client have existing personal loan? | binary | yes, no |
| contact | contact communication type | categorical | Unknown, telephone, cellular |
| day | last contact day of the month | numeric | 1,2,3,4,5,6,7 |
| month | last contact month of year | categorical | "jan", "feb", "mar", ..., "nov", "dec" |
| duration: | last contact duration, in seconds | numeric | 261,151,76… |
| campaign | number of contacts performed during this campaign and for this client (includes last contact) | numeric | 1,2,3.. |
| pdays | number of days that passed by after the client was last contacted from a previous campaign | numeric | -1 means client was not previously contacted |
| previous | number of contacts performed before this campaign and for this client | numeric | 0,1,2,3… |
| poutcome | outcome of the previous marketing campaign | categorical | Unknown, other, failure, success |

### 

### 

### 3.1.2 Data exploration

1. Overview of numeric variables

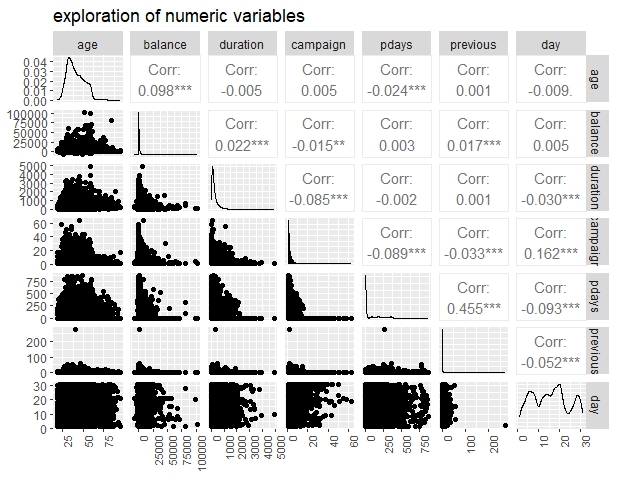
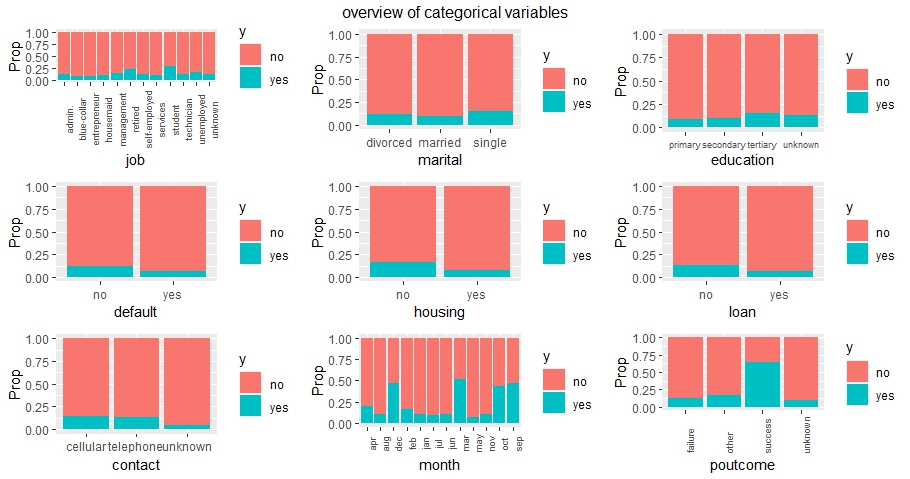


Figure 1: Exploration of numeric variables using a pairplot

The distribution of numeric variables ‘balance’, ‘duration’, ‘campaign’, ‘pdays’, ‘previous’ variables has high positive skewness with most observations in a small range of values. This suggests the presence of a large range of values that are large and positive. Therefore, there could be outliers for some of these numeric variables which have to be addressed. Notably, ‘previous’ has a relatively extreme outlier as seen in all the pair plots involving ‘previous’ in Figure 1.

Most variables are very weakly correlated with the exception of ‘pdays’ and ‘previous’, which are moderately correlated. The moderate correlation of ‘pdays’ and ‘previous’ can be attributed to the nature of the variables, where ‘pdays’ = -1 and ‘previous’ = 0 both suggest no previous contact.

1. Overview of categorical variables



In general, the proportion of class values ‘yes’ and ‘no’ varies across different values of these categorical variables, which suggest that they may be useful as predictors.

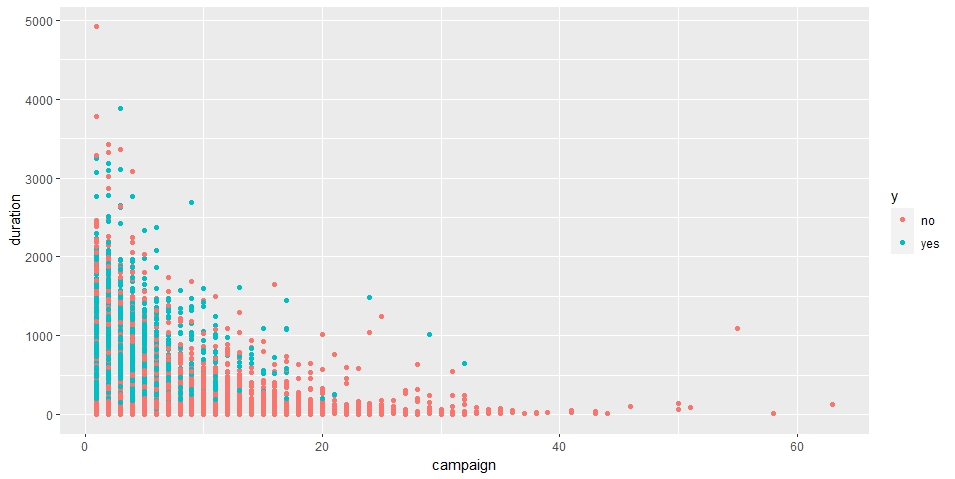
‘Job’, ‘marital’ and ‘education’ provide information about the customer profile. Specifically, it is observed that singles have the highest proportion of subscribing to a term deposit. This could be attributed to singles generally having fewer financial obligations and thus higher disposable income to subscribe to term deposits. Interestingly, customers who are retirees and students make up the highest proportions of customers who subscribe to term deposits. For students, this could be attributed to their non-immediate need for liquidity compared to other working customers. For retirees, this could be due to availability of funds due to past savings and lower expenses. Customers who receive tertiary education have the highest proportion of subscribing to a term deposit. Highly educated customers tend to have greater financial literacy, which could make them relatively more receptive to investments than other groups of customers.

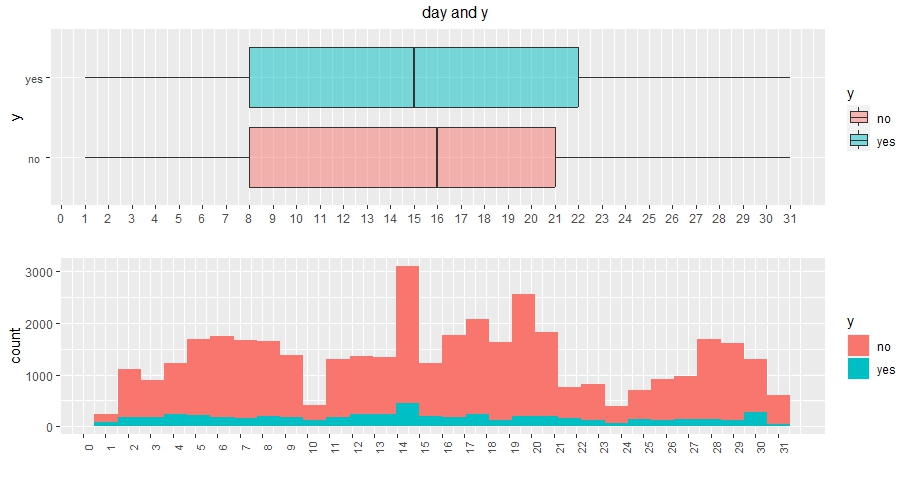
‘Housing’, ‘loan’ and ‘default’ reflect financial liabilities of the customers. For customers with any financial liabilities such as credit in default, existing housing or personal loan(s), it is observed that a relatively lower proportion of them are subscribed to term deposits compared to customers without either of these liabilities. This could be attributed to lower disposable income of customers with financial liabilities.

‘Contact’, ‘month’ and ‘poutcome’ reflect the nature of the telemarketing call. For ‘Contact’, customers who were contacted via telephone have a lower proportion of subscribing to a term deposit compared to those who were contacted via cellular. For ‘poutcome’, a significantly higher proportion of customers who subscribed in the previous marketing campaigns also subscribe to a term deposit, compared to those who did not in previous marketing campaigns. This is not surprising given that previous subscriptions likely indicate higher receptiveness to similar bank products.

For ‘month’, it is observed that ‘Dec’ and ‘march’

1. Micro analysis of notable variables





1. Characteristics of call recipient



### 

### 3.1.3 Data preparation

#### 3.1.3 a. BM1

* Discretization of ‘pdays’ and ‘previous’

‘pdays’ has a unique value of ‘-1’ to reflect cases in which customers were not contacted. 81.7% of total observations take on this value, while the remaining observations fall in a large range of days. Thus, it is more meaningful to divide observations for this variable into 5 bins as seen below. ‘-1’ is a bin on its own

due to its unique meaning, while the rest are divided using a 3-month interval.

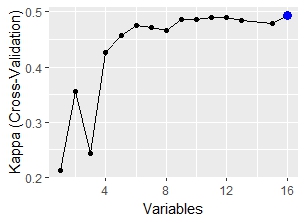
|  |  |
| --- | --- |
| Bins | Interpretation |
| ‘-1’ | Customers who were not contacted previously |
| ‘1\_90d’ | customers who were contacted 1 to 90 days ago(approximately 1-3 months back) |
| ‘91\_180d’ | customers who were contacted 91 to 180 days ago(approximately 3-6 months back) |
| ‘181\_270d’ | customers who were contacted 181 to 270 days ago(approximately 6-9 months back) |
| ‘270d\_plus’ | Customers who were contacted more than 9 months ago |

While the majority of the observations of ‘previous’ takes on the values of 0 to 7, there is still a significant proportion of observations that exist in a large range of values above 7, which can be considered extreme in domain context. However, these observations should not be discarded to avoid information loss. Yet, if ‘previous’ is left as a numeric variable, extreme values may affect predictions. Thus, observations for this variable are divided into 8 bins as follows:

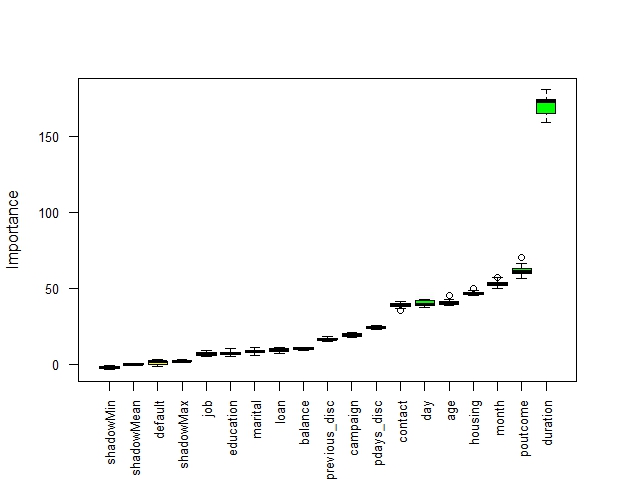
|  |  |
| --- | --- |
| Bins | Interpretation |
| ‘nc’ | Customers who were not contacted previously |
| ‘1c’ | customers who were contacted once previously |
| ‘2c’ | customers who were contacted twice previously |
| ‘3c’ | customers who were contacted thrice previously |
| ‘4c’ | customers who were contacted 4 times previously |
| ‘5c’ | customers who were contacted 5 times previously |
| ‘6c’ | customers who were contacted 6 times previously |
| ‘7c’ | customers who were contacted 7 times previously |
| ‘8c\_plus’ | Customers who were contacted more than 7 times |

* Feature selection

Boruta Algorithm and Recursive Feature Elimination (RFE) are used for feature selection. RFE is performed using 5-fold cross validation and evaluated using accuracy and Kappa. Across both metrics, RFE consistently selects all 16 variables.



Out of the 16 explanatory variables, Boruta Algorithm classifies ‘default’ as tentative and the rest of 15 variables as important. Since RFE requires a decision on the number of features while Boruta does not



Even though one third of the observations in contact have a significant proportion of ‘unknown’ observations, it is classified as an important variable according to Boruta and RFE. will not be discretized due to potential of overfitting.

Since ‘duration’ pertains to the telemarketing call and not the entire telemarketing campaign, and given that BM1 is scoped to analyse the effectiveness of a telemarketing call, ‘duration’ will be retained. The same logic applies for ‘campaign’.

#### 3.1.3 b. BM2

To analyse the effectiveness of a telemarketing campaign, only variables that reflect efforts of a completed telemarketing campaign should be included. Therefore, ‘duration’ and ‘campaign’, which capture ongoing campaign efforts, will be removed. Other than that, selection and treatment of the other variables will be consistent to what is performed in BM1.

3.1.4 Algorithms Used

We use Random Forest, SVM, Decision Tree, Logistic Regression, and Naive Bayes models to predict whether a customer will subscribe to a term deposit or not. We choose these algorithms because they are all commonly used for binary classification algorithms. However, in real life, we are not able to pre-determine which model is the most suitable for solving this problem, different models may have different advantages. Therefore, we choose to experiment with multiple models.

There are some introductions of these algorithms:

Random Forest: Random Forest is a machine learning algorithm based on integrated learning, which improves the performance of classification tasks by combining multiple decision trees. Random Forest improves the performance of classification models by constructing multiple decision trees and combining their results. Output of the Random Forest models is the class that is represented by most trees.

SVM (Support Vector Machine) : SVM is a powerful machine learning algorithm for classification and regression. It aims to find an optimal hyperplane to classify data points for both linear and nonlinear problems. Maximise margin between boundary and closest data (support vectors).

Decision Tree: Decision Tree is a tree-structured classification algorithm. It is easy to understand and interpret, and suitable for both classification and regression tasks. Decision tree builds a tree based on data features to do classification, where each node represents a feature or attribute and leaf nodes represent a category or value.

Logistic Regression: logistic regression is a classical binary classification algorithm for predicting probabilities.

3.1.5 Model Evaluation Methods

To see the performance of different models, we chose to compare MCC, accuracy, sensitivity, specificity, and AUC to evaluate the results. The use of different metrics provides us a more comprehensive understanding of the model performances.

Accuracy measures the percentage of data that the model correctly classifies. Accuracy = (number of correctly classified samples) / (total number of samples). Overall, higher accuracy means better classification. However, accuracy can be affected by an imbalanced dataset. Accuracy can be high if the model predicts all outcomes as the dominant result. For example, in our dataset, many customers do not subscribe, and if the model predicts no subscriptions for the entire test set, the accuracy would still be high, but in fact the model would have very little reference for the prediction.

## **3.2 Results**

### 3.2.1 BM1

The average metrics for various models, built using different algorithms and after appropriate tuning, are summarized in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Support Vector Machine** | **Logistic Regression** | **Decision Tree** | **Random Forest** | **Naive Bayes** |
| **MCC** | 0.4789 | 0.4789 | 0.4991 | 0.5350 | 0.4350 |
| **Accuracy** | 0.8476 | 0.8859 | 0.8941 | 0.9017 | 0.7946 |
| **Specificity** | 0.8460 | 0.5786 | 0.5738 | 0.6055 | 0.8151 |
| **Sensitivity** | 0.8588 | 0.9266 | 0.9365 | 0.9410 | 0.7919 |
| **AUC** | 0.8524 | 0.9015 | 0.9816 | 0.92907 | 0.8796 |

### 

### 

### 

### 

### 3.2.2 BM2

The average metrics for various models, built using different algorithms and after appropriate tuning, are summarized in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Support Vector Machine** | **Logistic Regression** | **Decision Tree** | **Random Forest** | **Naive Bayes** |
| **MCC** | 0.3580 | 0.3084 | 0.3007 | 0.3373 | 0.2606 |
| **Accuracy** | 0.8216 | 0.8664 | 0.8709 | 0.8747 | 0.7082 |
| **Specificity** | 0.8517 | 0.3512 | 0.3179 | 0.3624 | 0.6680 |
| **Sensitivity** | 0.5939 | 0.9346 | 0.9439 | 0.9425 | 0.7136 |
| **AUC** | 0.7228 | 0.7517 | 0.9816 | 0.7769 | 0.7517 |

## **3.3 Discussion**

### 3.3.1 BM1

In Business Model 1, the primary objective is to predict the likelihood of a customer subscribing to a term deposit after a call during the campaign. Given the imbalanced nature of the response variable, relying solely on accuracy can be misleading. Therefore, we'll emphasize other metrics like MCC, Specificity, Sensitivity, and AUC, which provide a more comprehensive view of model performance.

* Random Forest emerges as a strong contender with the highest MCC of 0.5350. In the context of an imbalanced dataset, a high MCC indicates that the model is doing well in both predicting the positive class and the negative class. Its Sensitivity of 0.9410 suggests it's adept at identifying potential subscribers, which is crucial mid-campaign to ensure resources are directed towards the most promising leads.
* Decision Tree boasts the highest AUC of 0.9816. A high AUC indicates that the model has an excellent ability to differentiate between potential subscribers and non-subscribers. This is vital for the bank, especially during the campaign, to ensure that efforts are not wasted on less promising leads.
* Logistic Regression has a high Sensitivity of 0.9266, suggesting it's good at identifying potential subscribers. However, its lower Specificity indicates a higher false positive rate, which might lead to wasted resources on customers less likely to subscribe.
* Support Vector Machine offers a balanced Specificity and Sensitivity, making it a potential choice for banks looking for a balanced approach in their campaign.
* Naive Bayes, while having a lower MCC, has a high Specificity of 0.8151. This means it's good at correctly identifying those who won't subscribe. This can be beneficial if the bank wants to be cautious and avoid approaching less promising leads.

#### 3.3.1.a Business Implications

In a telemarketing campaign, especially mid-campaign, it's crucial to adjust strategies based on insights from data. The bank needs to ensure that efforts are directed towards potential subscribers to maximize the return on investment.

The ***Random Forest*** model, with its high MCC and Sensitivity, seems most aligned with this objective. Implementing this model can help the bank identify and target the most promising leads, ensuring efficient use of resources and maximizing subscriptions.

However, if the bank's strategy leans towards being more cautious and ensuring they don't waste resources on less likely leads, the ***Naive Bayes*** model with its high Specificity might be more appropriate.

In conclusion, the choice of model should be in sync with the bank's mid-campaign objectives. The results presented here offer a solid foundation for making a data-driven decision to enhance the campaign's effectiveness.

### 3.3.2 BM2

In Business Model 2, the focus shifts to a post-campaign (macro) analysis. The objective here is to identify potential consumers who are more inclined to subscribe to the deposits in future campaigns. Given the imbalanced nature of the dataset, we'll again emphasize metrics like MCC, Specificity, Sensitivity, and AUC over Accuracy.

* Support Vector Machine emerges as the model with the highest MCC of 0.3580. This indicates that, despite the challenges, it's the most balanced in terms of binary classification performance for this dataset. Its high Specificity suggests it's adept at identifying true negatives, which can be crucial in post-campaign analysis to avoid targeting the wrong audience in future campaigns.
* Decision Tree has an impressive AUC of 0.9816, indicating its strong ability to differentiate between potential subscribers and non-subscribers. However, its relatively lower MCC suggests that while it can rank order prospects well, its binary classification might not be as reliable.
* Logistic Regression and Random Forest show high Sensitivity, suggesting they are good at identifying potential subscribers. However, their lower Specificity indicates a higher false positive rate, which might lead to wasted resources in future campaigns.
* Naive Bayes has the lowest MCC among the models, suggesting it might be the least reliable for this specific post-campaign analysis. However, its balanced Specificity and Sensitivity might still offer some value.

#### 3.3.2.b Business Implication

In the context of future campaigns, where the objective is to identify potential subscribers for term deposits, the choice of model becomes pivotal. The **Random Forest model**, with its commendable MCC and Sensitivity, appears to be well-suited for this task. Its deployment can aid the bank in pinpointing and engaging the most promising leads, ensuring better resource allocation and enhancing subscription rates.

On the other hand, if the bank's strategy is geared towards a more conservative approach, aiming to minimize resource wastage on less probable leads, the **Support Vector Machine** model, with its high Specificity, will be a viable option.

However, the performance of the models in Business Model 2 (BM2) suggests a need of caution in their application for future campaigns. While they offer some direction, the suboptimal results indicate potential risks in resource misallocation and unintended customer outreach. This could lead to operational inefficiencies and potential disruptions in customer experience. Given these challenges, the bank should consider these models as supplementary tools, complemented by other data-driven strategies. It's imperative for the bank to establish a robust feedback mechanism, comparing model predictions with actual outcomes, to refine and enhance future campaign strategies. This balanced approach will ensure that the bank remains competitive, optimizes resources, and fosters positive customer relationships.

# **4. Related Work**

Answer the following questions for each piece of related work that addresses the same or a similar problem. What is their problem and method? How is your problem and method different? Why is your problem and method better?

# 

# **5. Future Work**

Our models (SVM, DT, LG, RF, NB) are mainly classification models used to classify the probability of whether one will subscribe to term deposit or not. It would be good to use a multi-linear regression model to see which attributes are more positively correlated to y and rank them based on the importance of the attributes by looking at the size of the coefficient of each attribute so that the bank managers can tweak their policies to term deposit. We can also transform those categorical attributes like education level to be numerical values through ordinal encoding so that we can see how strong the relationship is between education level and purchasing of term deposit so that bank managers can focus on that particular group of customers. Although we can see that the higher the level of education, the more likely one will subscribe to term deposit, it is unknown how strong this attribute contributes to y. Thus, deploying a multi-linear regression model would be a great addition since we only have classification models in our analysis.

In this dataset, there are only 16 attributes, moving forward, we can consider including more input attributes such as the socio-economic factors like the interest rate offered by the banks, risk appetite of the customers, inflation rate, etc. All these attributes can influence whether one will take up a term deposit or not. For example, in a high interest rate environment, unless the interest of the term deposit is higher, there is no pull factor for one to subscribe it. Also for risk averse customers who are usually those senior citizens, they might be more inclined to subscribe to term deposits as it is much safer than buying stocks.

Although there are no missing values in the dataset, there are quite a number of “unknowns” which are essentially missing information. For such cases, to improve on the accuracy of the data we can seek the help of the domain experts such as bank managers to help label part of the missing information of maybe education level since they can most likely identify based on other attributes like “job” and “balance”. For those customers who are in managerial positions and have a high average yearly balance, perhaps more than 10,000, their education level might be tertiary education. And use the information to train other “unknown” values so that we can get more accurate predictions.

Also depending on our business objective, if we want to reduce resources, we should use downsampling as it can eliminate irrelevant attributes thus leading to a more focused and accurate model. However if we want to maximize business opportunities, upsampling like what we have done here would be more appropriate as it can improve the model’s performance since it supplies more training data.

Lastly, we can explore the prediction of other attributes such as duration of call as usually if the customer receives such calls from banks - uninterested customers will end their calls early. Thus the longer the customer stays on the call, the higher the chances that this customer is likely to subscribe to a term deposit. If the bank has historical telemarketing behaviour information for those contacted customers, these can complement our existing information. If not, we can conduct qualitative surveys to study if the experience of what the customer had during the call is favourable or not and whether this is one of the factors when subscribing to a term deposit.

# **6. Conclusion**

The models used across Business Model 1 (BM1) are generally better performances than models used across Business Model 2 (BM2) in terms of average MCC, accuracy, sensitivity, and AUC. In the BM1, out of all the models, DT and RF are one of the better performing models since both have relatively higher average MCC and AUC scores. Between DT and RF, RF is superior since it has a higher average sensitivity score which means it can potentially identify those customers who are more likely to subscribe to term deposit. Although in BM2, it is not surprising that RF is reaffirmed to be one of the better performing models, it is surprising that SVM seems like a strong contender as one of the better performing models alongside RF if you just look at average MCC score. Also, SVM has relatively balanced metric scores as compared to DT especially in terms of sensitivity measure.

For every organization, including the bank, the motivation to conduct this analysis is because banks typically want to maximize revenue and minimize cost.

Each model performs well based on the objective of the bank. For SVM, in order to find out which clients are more likely to subscribe to term deposit. Looking at high sensitivity scores, banks can maximize business opportunity because banks do not want to miss out on customers that can potentially subscribe to term deposits as these are beneficial to banks. Banks can use the term deposits to invest in higher yield products and increase revenue for the company in return. In this case, the cost of missing such potential customers is high thus banks should look closely at sensitivity scores.

If banks want to find out which clients are not likely to subscribe so that they can have greater cost savings since banks can reduce the number of false alarms such as following up with uninterested clients which is a waste of resources and time. As such, banks should look at models with high specificity such as DT/RF.

Generally, banks should target singles and those without a loan as they are more likely to subscribe to term deposits.

For future campaigns in term deposits, our research helps to direct the bank's efforts based on their objective so that resources are well-allocated at the right place. In banking, timely decisions are essential, thus our models can be included in future research that focus on developing models that can make real-time predictions taking into account evolving customer behaviour.