DSA3101 Project Report 1

National University of Singapore

DSA3101 Team 16

Bank Marketing Prediction

Author:
Chan Zhen Hao Benny (A02047758Y)
Christal Woon Xinyi (A0205579W)
Evan Ang Jia-Jun (A0200591R)
Ho Kah Hsin Carine (A0205168H)
Joel Choe Yee Hsien (A0149842W)
Sanraj Mitra (A0200075X)
Trisha Ajay Mehta (A0205894W)

September 9, 2021

Contributions

1 Chan Zhen Hao Benny

Report writing (Introduction, EDA, Citations and References), slides (Introduction, EDA)

2 Christal Woon Xinyi

Report writing (Introduction, EDA). Compiled all EDA diagrams into a Tableau Packaged Workbook. Organised slides and presentation materials.

3 Evan Ang Jia-Jun

Code (Logistic regression model and review). Presentation and slides preparation (Modelling and recommendation). Report (Logistic Regression)

4 Ho Kah Hsin Carine

Report writing (abstract, EDA, model evaluation, recommendations, conclusion), prepared slides and presentation materials (introduction, EDA and recommendation).

5 Joel Choe Yee Hsien

Proposed and implemented an ensemble model and SVM model. Trained classifiers for final code. Compiled and tidied model code. Report writing (Data Pre-processing, EDA, model evaluations, Modelling - ensemble and SVM).

6 Sanraj Mitra

Report Writing (Preprocessing, EDA, Modelling, Recommendations). Presentation Slides (Modelling). Code (Data preprocessing, Modelling- Tree Family, EDA)

7 Trisha Ajay Mehta

Report writing(EDA) and slides(EDA)

DSA3101 Project 1 Report

Bank Marketing Prediction

DSA3101 Team 16
Chan Zhen Hao Benny (A0204758Y), Christal Woon Xinyi (A0205579W),
Evan Ang Jia-Jun (A0200591R), Ho Kah Hsin Carine (A0205168H),
Joel Choe Yee Hsien (A0149842W), Sanraj Mitra (A0200075X),
Trisha Ajay Mehta (A0205894W)

September 2021

Abstract

This paper covers the applications of Machine Learning in helping banks make better business decisions. We utilized a Portuguese Bank Marketing data set in this project with the hopes of showcasing how Machine Learning can be used in predicting whether a client subscribes to a term deposit service with the bank. Through our analysis, we have synthesized data-driven recommendations that could potentially help banks maximize their term deposit subscription rates.

Introduction

Term deposits are a key source of capital for banks [1]. Banks often offer a lower interest rate for consumer deposits, and can loan out these deposits at a much higher rate, which creates an interest rate spread that they can profit from. For several years, Portuguese banks have been over-lending [2] and have been at risk of having insufficient reserves [3] available for unexpected contingencies such as economic downturns. Attracting more subscribers to their term deposits is therefore important for banks, as that would reduce the chances of severe financial crises, such as bank runs [4] or liquidity crunches [5], from occurring.

This report aims to analyse a Portuguese bank marketing data set - deriving insights using exploratory data analysis and Machine Learning frameworks such as Tree-based models, Logistic Regression and Support Vector Machines. Through our findings, we hope to make insightful recommendations for the banks so that they can successfully identify potential clients and advertise their products efficiently. This not only helps banks save cost and time, but also generates revenue.

While our data is based on a traditional Portuguese bank, the scope of our project extends across other types of banks as well. For instance, in recent years, digital banks have become increasingly prominent, but many of them are relatively new. One way these new banks can build up liquidity and their customer base is by attracting term depositors. Over the years, banks have been increasing their budget allocation for marketing significantly [6]. As marketing campaigns are costly, banks should allocate marketing resources efficiently i.e., on people who are more likely to subscribe to their term deposits.

This paper is structured into 7 sections. Section 1 is the Introduction, Section 2 gives a brief overview of the Data Pre-processing process, Section 3 focuses on our Exploratory Data Analysis, Section 4 describes the Feature Selection process, Section 5 illustrates the Modelling process, Section 6 details the Recommendations and after which Section 7 presents the Conclusion.

Data Pre-processing

Data Source

Our data is obtained from the UCI Machine Learning Repository - a large, public collection of databases, domain theories, and data generators that are often used by the Machine Learning community for the empirical analysis of Machine Learning algorithms. In the context of our project, we have chosen to work on the Portuguese Bank Marketing data set - a data set which contains real data from the direct marketing campaigns conducted over the phone by a Portuguese retail bank between May 2008 and December 2010.

The data set consists of 41,188 observations and 20 input variables, of which 10 are categorical and the remaining 10 are numerical (Table 1). Each observation contains the details relating to a particular client that has been approached by the bank. Some of these information includes the age of the client, marital status of the client, education level of the client, number of days passed after last contact with client and even industry metrics such as the employment variation rate.

Variable	Details	Variable Type
age	Age of client	Numerical
job	Job of client	Categorical
marital	Marital status of client	Categorical
education	Education level of client	Categorical
default	Client has credit in default	Categorical
housing	Client has housing loan	Categorical
loan	Client has personal loan	Categorical
contact	Contact communication type with client	Categorical
month	Last contact month with client	Categorical
day_of_week	Last contact day of the week with client	Categorical
duration	Last contact duration with client (in seconds)	Numerical
campaign	Number of contacts performed during campaign on client	Numerical
pdays	Number of days passed after last contact with client	Numerical
previous	Number of contacts performed before this campaign with client	Numerical
poutcome	Outcome of previous marketing campaign with client	Categorical
emp.var.rate	Quarterly indicator of employment variation rate	Numerical
cons.price.index	Monthly indicator of consumer price index	Numerical
cons.conf.index	Monthly indicator of consumer confidence index	Numerical
euribor3m	Daily indicator of 3 months EURIBOR rate	Numerical
nr.employed	Quarterly indicator of number of employees	Numerical
y	Whether client has subscribed to term deposit	Categorical

Table 1: Attribute Information for Bank Marketing Data Set

Data Governance

Missing Data

Prior to our analysis, we ensured that there was no missing data in the data set. Some problems that might arise due to missing data includes a reduction in the statistical power of prediction, increased bias in the estimation of parameters in models, or even a reduction in the representation of samples.

Mislabelled Data

We also renamed some values in the data set which were previously mislabelled. For example, the instance ”admin” under the job category had been mislabelled as ”admin.”. Hence, we took the pre-emptive step to tidy the data by replacing all the instances of “admin.” in job to “admin”.

Duplicated Rows

When checking the data set for duplicated rows, we identified 12 rows that contained similar attributes. Since there were no unique identifiers for each client, we had no reason to drop them as these duplicated rows could belong to different customers who shared the same attributes.

Unknown Values

While analyzing our data, one concern was the presence of ’unknown’ values within the data set. This was not unexpected as we understand that clients often value their privacy and would sometimes choose not to disclose their complete personal information. Regardless, to ensure that our final models were meaningful, we decided to interpolate the 1,930 ’unknown’ values for the columns ’job’ and ’education’ prior to building our models. We predicted these ’unknown’ values based on the columns that did not contain any ’unknown’ values, using a *RandomForestClassifier* model from the *sklearn* library. By not dropping these rows, we retained as much information from the original data set as possible, without significantly changing the distribution of features.

Exploratory Data Analysis

Response Variable - y

As seen from Figure 1, our data set is highly imbalanced. There are approximately 8 times as many 'No' (did not subscribe to deposit) to 'Yes' (subscribed to deposit) responses.

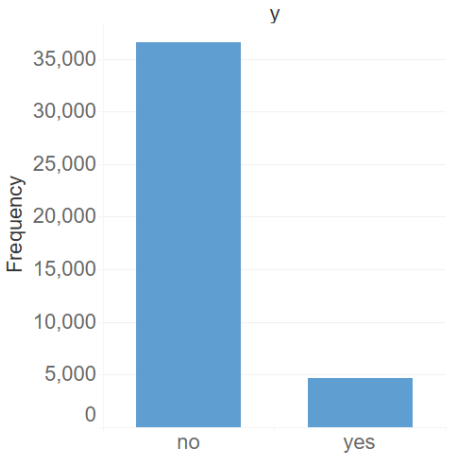
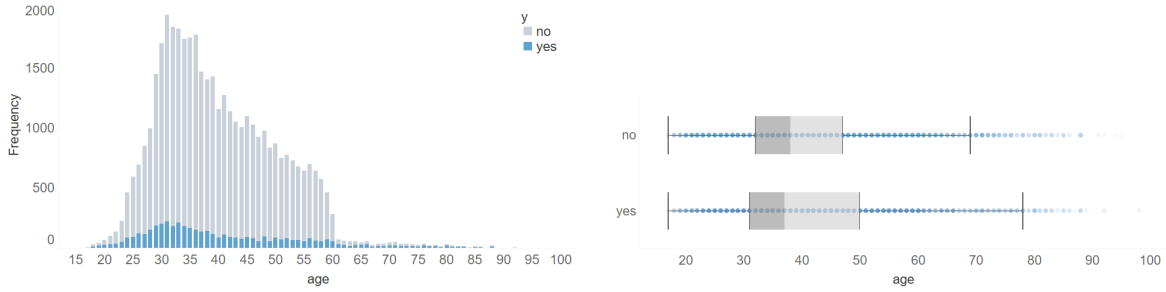


Figure 1: Response Variable

Numerical Variables

Age

As seen from the histogram and box plot (Figure 2), the continuous variable 'age' did not provide much insight into which age groups were more receptive to term deposits. As such, we binned the ages into subgroups (Table 2) as a means to identify any possible relationship between different age baskets of clients and their receptiveness towards term deposits. Each bin contains a sizable number of observations (>500).



(a) Histogram of Age

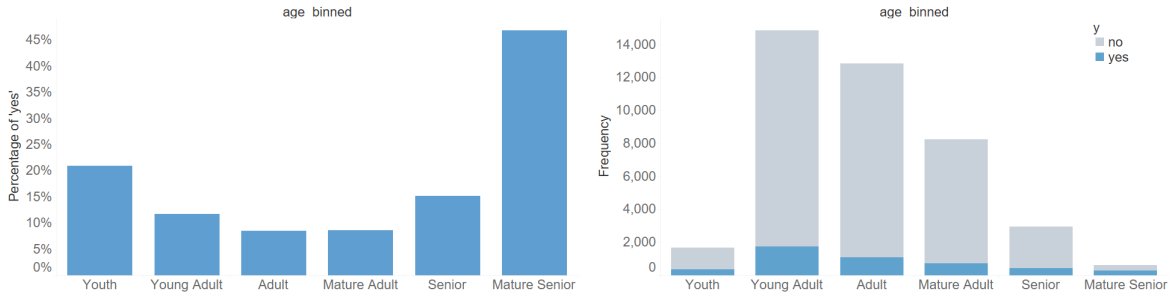
(b) Box plot of Age

Figure 2: Age

Bins	Group	Group Size
17 - 25	Youth	1,666
26 - 35	Young Adults	14,847
36 - 45	Adult	12,844
46 - 55	Mature Adults	8,249
56 - 65	Seniors	2,963
>65	Mature Seniors	619

Table 2: Age groups

Upon further analysis of the different age groups (Figure 3a), we found that mature seniors, seniors and youth tend to have a higher subscription rate. Another observation is that young adults make up a significant sample size (Figure 3b).



(a) Subscription rate

(b) Subscription count

Figure 3: Binned Age

Campaign

'Campaign' refers to the number of contacts performed during the current campaign for a client. From the histogram plot (Figure 4), we observed that between the clients who subscribed to term deposits and the clients who did not, both groups have a mode of zero. While those who subscribed to a deposit have been contacted fewer times than those who did not subscribe during the campaign, the difference in margin is small. Furthermore, since the correlation between 'Campaign' and the response was low (Figure 21), we decided to drop this feature.

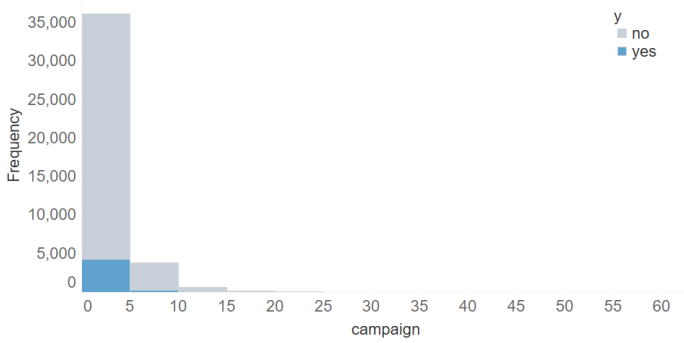
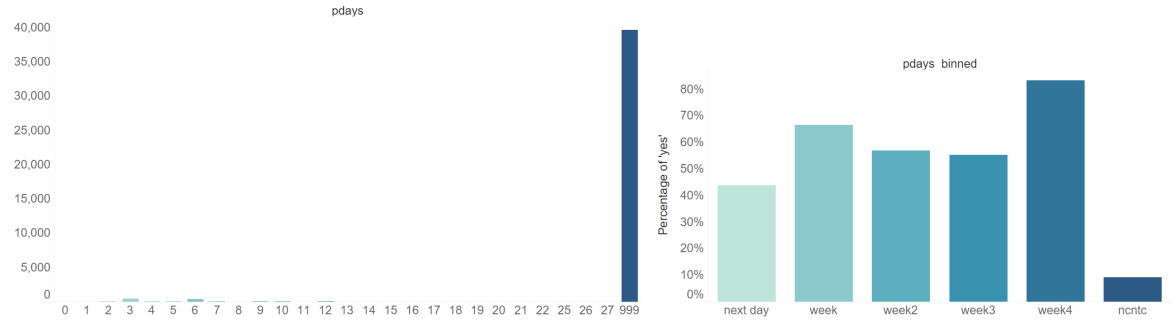


Figure 4: Histogram of Campaign

Pdays

'Pdays' refers to the number of days since the client was last contacted from a previous campaign. The value 999 denotes the clients who were not previously contacted. This could either mean they are new clients or they have never been contacted during the last campaign. From the bar chart (Figure 5a), it can be observed that majority of clients fall under the value 999. Rather than considering 'pdays' as a numerical variable, we took the initiative to bin 'pdays' into various categories (Table 3) in an attempt to make our analysis more meaningful. From the bar chart of the binned 'pdays' (Figure 5b), customers who had been contacted in the fourth week had the highest percentage of subscribing, but there are only 6 people who fall under this category (Table 3). The general observation is that as the number of days increased, people were less likely to subscribe to a term deposit.



(a) Frequency of Pdays

(b) Subscription rate by pdays groups

Figure 5: Pdays

Bins	Group	Group Size
0	Contacted the next day	41
1 - 7	Contacted within one week	1,136
8 - 14	Contacted in second week	276
15 - 21	Contacted in third week	56
22 - 31	Contacted in fourth week	6
999	Not contacted previously / New client	39,673

Table 3: Pdays Groups

Previous

'Previous' refers to the number of times a client was contacted by the bank before the campaign. A 'previous' value of 0 indicates that the bank did not contact the client prior to the campaign. From the bar chart (Figure 6), the distribution of values is highly right-skewed, with 85% of the values being 0. This is indicative that majority of the customers were new. Rather than considering the 'previous' value, we decided to engineer the variable into a binary feature, 'previously_contacted'. The new feature takes on the value of 'no' when 'previous' is 0, and takes on the value of 'yes' otherwise.

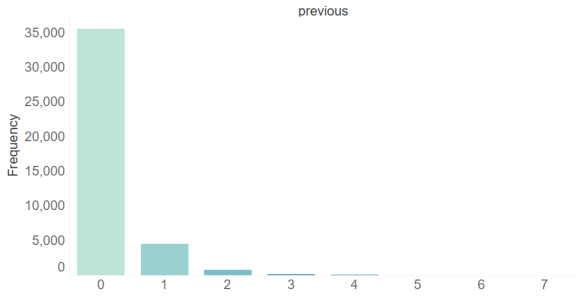


Figure 6: Frequency of Previous

Employment Variation Rate

Employment variation rate, or employment rate dispersion [7], is the variation of regional employment rates in a country. It is calculated by the coefficient of variation of regional employment rate in a country, weighted by the working age population. A high positive variation rate occurs with increasing employment while a high negative rate occurs with increasing unemployment. From the box plot (Figure 7), the median employment variation rate of people who subscribed to term deposits was lower and of a negative value. A possible reason could be due to people cutting back on consumption and increasing savings by subscribing to term deposits in economic downturns. To further understand the possible impact of employment variation rate on a particular customer's decision, one would need to consult industry experts prior to making concluding statements.

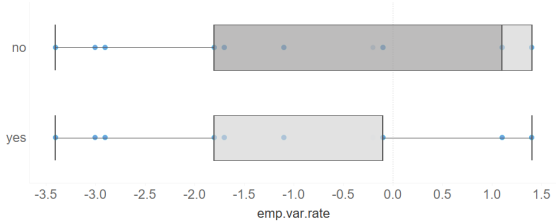


Figure 7: Box plot of Employment Variation Rate

Consumer Price Index

Consumer Price Index (CPI) [8] is a measure of the average change in prices over time for a basket of goods and services. It is frequently used to identify inflation or deflation in an economy. A rising CPI indicates inflation while a decreasing CPI indicates deflation. From the box plot (Figure 8), the median CPI of people who did not subscribe was about 93.9. This differs significantly from the median of people who subscribed at 93.2, considering that CPI values in the data set range from 92.2 to 94.8. A possible reason could be since inflation (rising CPI) erodes the value of savings, people might be less likely to subscribe to term deposits. The CPI variable is included in our model building since the median differs greatly for both classes, which suggests this feature may be useful in predicting subscription.

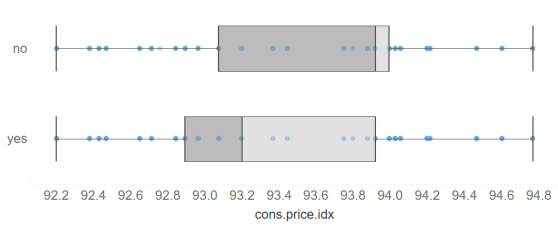


Figure 8: Box plot of Consumer Price Index

Consumer Confidence Index

Consumer Confidence Index [9] is an economic sentiment indicator that is obtained by surveying households, measuring their confidence in the state of the economy currently and in the near future. A negative, decreasing index suggests that people have a pessimistic outlook on the economy, which influences their spending and saving decisions. From the box plot (Figure 9), the consumer confidence index of majority of people who subscribed and did not subscribe lies between -35 and -43. This suggests that most people had a negative outlook of the economy. The medians of both groups who subscribed and did not subscribe differ by around 10%. As we believe that confidence in the economy could affect people's decision to subscribe, we kept this feature in the building of our models.

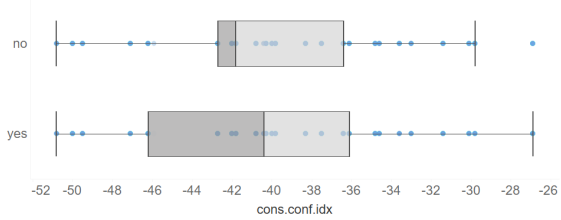


Figure 9: Box Plot of Consumer Confidence Index

3 Months EURIBOR Rate

EURIBOR refers to the Euro Interbank Offered Rate [10]. It denotes the basic rate of interest used in lending between banks on the European Union Interbank Market. This is the rate which banks may borrow from other banks to make up for any short term shortage in liquidity, or loan out any excess cash that they have to other banks [11]. From the box plot (Figure 10), the median Euribor Rate of the group who did not subscribe was much higher than the median rate of those who subscribed. While this variable might be beyond our control, it is still often used by many banks as a guide for setting the interest rate on other loans [12]. Since interest rate has an impact on the subscription decision to a term deposit by a client, it is included as a feature in our models.

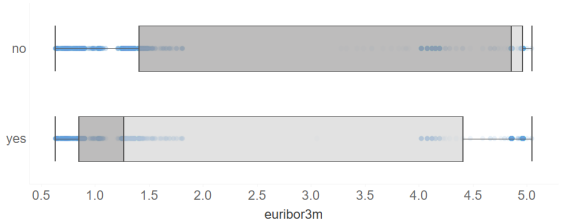


Figure 10: Box Plot of Euribor Rate

Number of Employees

From the box box plot (Figure 11), the median number of employees of the group who subscribed is 5100, which is significantly lower than the median of those who did not subscribe at 5200. As this feature is likely to be useful in predicting whether a customer subscribes, we kept it in our data set for the time being.

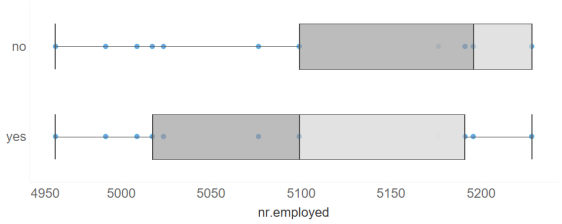


Figure 11: Box Plot of Number of Employees

Duration

Duration is the time spent on the call with a client during the campaign. As seen from Figure 12, the longer the bank spends on the call with a client during the campaign, the more likely they will subscribe to the deposit. This makes sense as those who subscribed to the deposit must have been on the call longer than those who did not subscribe. However, we excluded this variable moving forward as realistically, we wouldn't know the duration of a call until after the call is over.

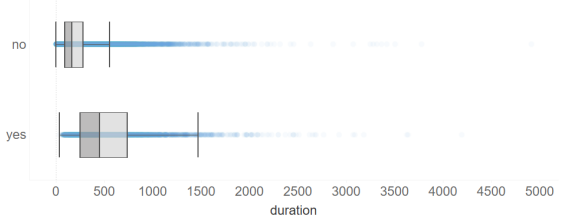
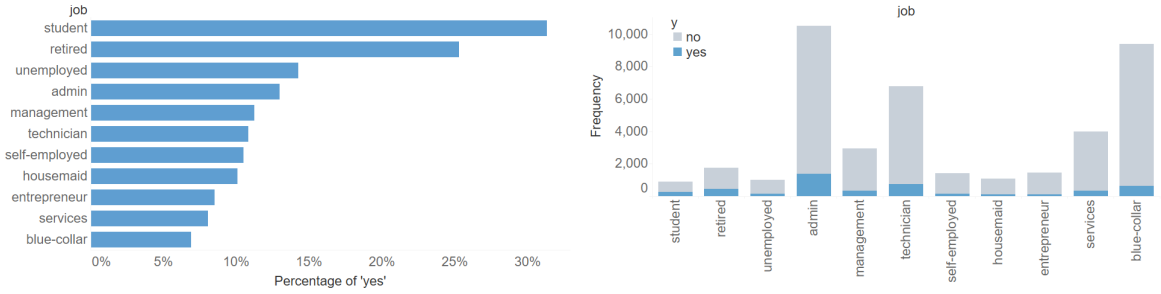


Figure 12: Box Plot of Duration

Categorical Variables

Job

From the bar chart (Figure 13a), two categories that stood out due to their high subscription rate are students and retirees. It can be observed that 31% of students and 25% of retirees signed up for a term deposit. This is significantly higher compared to the other occupations. This might be due to students and retirees having lesser financial commitments (e.g., mortgage payments, family expenses). However, it can also be observed that the group sizes of students and retirees are relatively smaller compared to the group sizes of admin, technician and blue-collar occupations (Figure 13b).



(a) Subscription rate

(b) Subscription count

Figure 13: Job

Education

It can be observed that conversion rate tends to increase with one's education level (Figure 14). As such, we have included 'education' as a field in the building of our models.

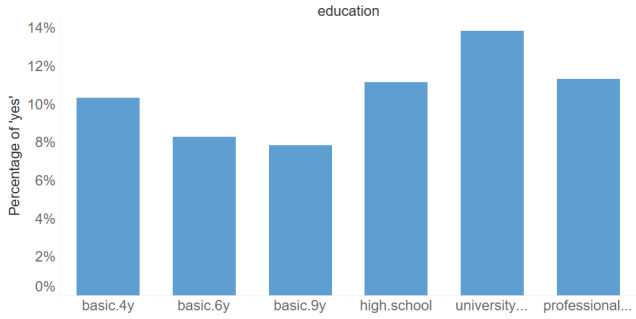


Figure 14: Subscription Rate by Education

Marital

From the bar chart (Figure 15), we see that the distribution between the 'divorced' and 'married' categories are fairly similar, while 'single' is slightly higher than both. Together with a low correlation value with the target response (Figure 21), we chose to drop this feature as it seems unhelpful in explaining the target response.

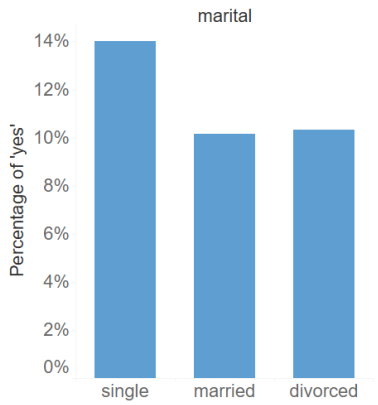


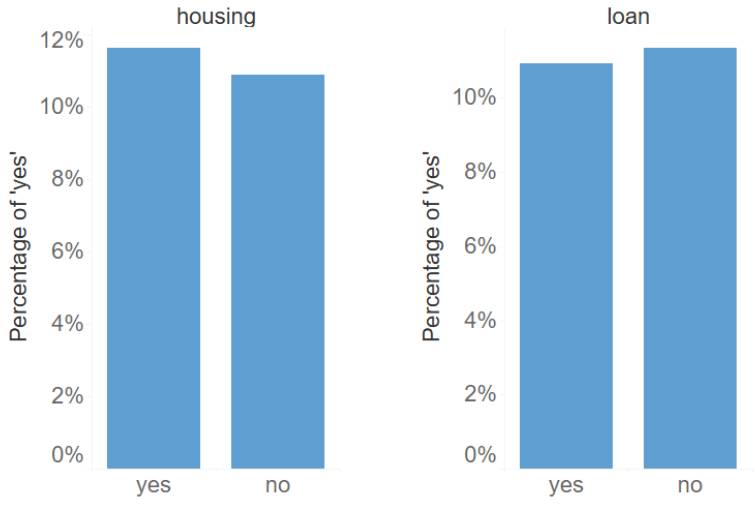
Figure 15: Subscription Rate by Marital

Default

An individual with 'default' value as 'yes' has failed to repay the bank a loan or make interest payments on time. Since the data set consists of only 3 such individuals, this variable offers little to no information in explaining the response. Furthermore, it has a low correlation value with the response (Figure 21). Hence, we have chosen to drop the 'default' feature.

Housing / Loan

The conversion rates for individuals with housing loans or personal loans are largely similar (Figure 16) and have low correlation values with the response (Figure 21). Since knowing whether a potential client has a prior loan is unlikely to help us determine the likelihood of them taking up a term deposit, we decided to drop the variables 'housing' and 'loan' from our data set.



(a) Subscription rate

(b) Subscription rate

Figure 16: Housing / Loan

Contact

'Contact' refers to the communication type with clients, and take on values 'cellular' or 'telephone'. From the bar chart (Figure 17), the rate of subscription among those who were contacted via cellular was 14.7%. This was significantly higher than the rate of 5.2% for those contacted via telephone. The significant difference between these two rates suggests that this feature might be useful in making predictions.

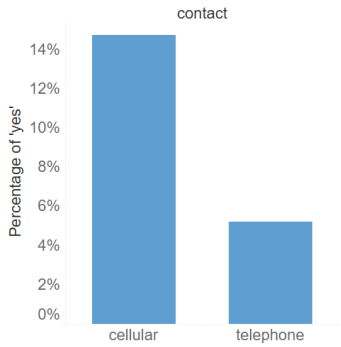
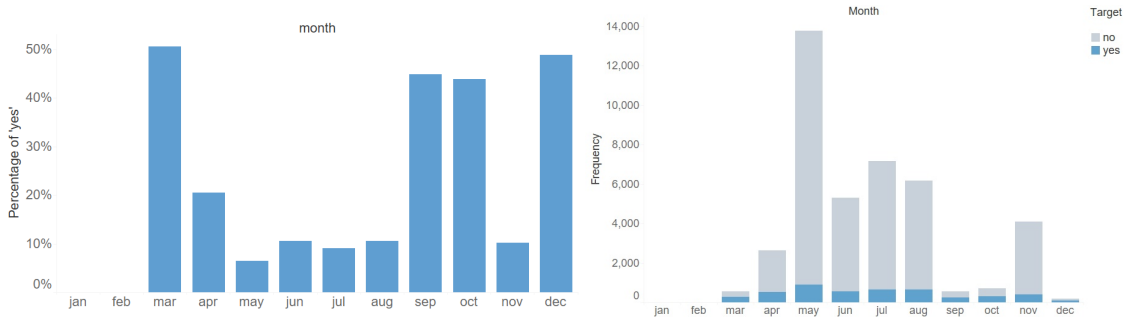


Figure 17: Subscription Rate by Contact

Month

'Month' refers to the month a client was last contacted. From the bar chart (Figure 18a), we observe that March, September, October and December have a subscription rate higher than 40%, with March having the highest at 51%. This is much higher than the remaining months, where the subscription rates are below 25%. We noted that months with higher subscription rates generally have smaller client counts (Figure 18b).



(a) Subscription rate

(b) Subscription count

Figure 18: Month

Day of Week

'Day of Week' refers to the last contact day with the client. From the bar chart (Figure 19), the percentage of clients who subscribed are approximately similar across all days. Furthermore, since this variable have a low correlation value with the response (Figure 21) and might not provide us with much information in predicting whether a client subscribes to the bank's service, we decided to drop this column in the building of our models.

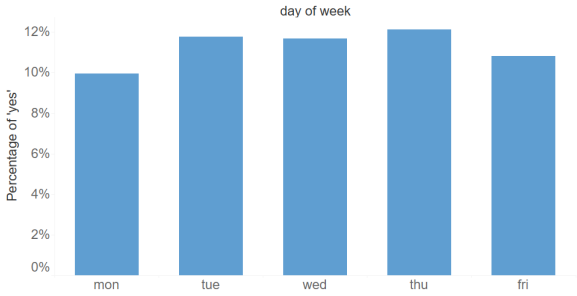
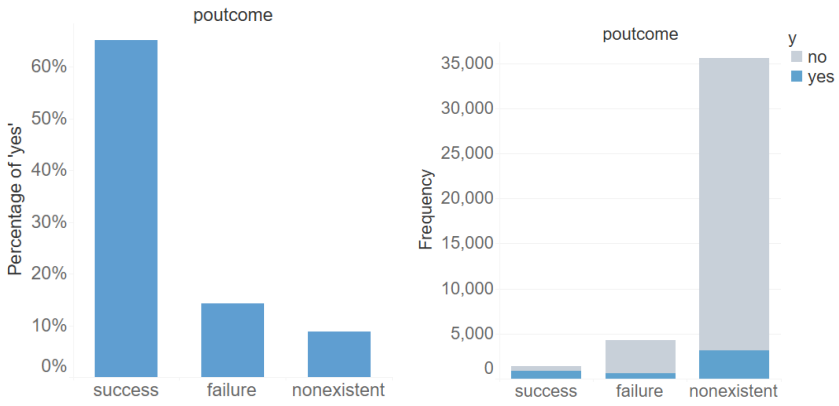


Figure 19: Subscription Rate by Day of Week

Poutcome

'Poutcome' refers to the outcome of the previous marketing campaign. The possible values are 'failure', 'nonexistent' and 'success'. From the bar chart (Figure 20a), clients who subscribed to a term deposit in the previous marketing campaign were significantly more likely to subscribe again in the current campaign. We also noted that majority of clients were new, and had nonexistent poutcome (Figure 20b). This column is kept in our analysis, and we further studied it in later sections.



(a) Subscription rate

(b) Subscription count

Figure 20: Poutcome

Feature Selection

To understand the relationships between variables, we utilised three different metrics: Pearson Correlation Coefficient which identifies correlation between two continuous variables, Cramér's V which identifies correlation between two categorical variables, and Correlation Ratio which identifies the correlation between a categorical variable and a continuous variable[13].

A correlation matrix representing the correlations between different variables can be observed in Figure 21.

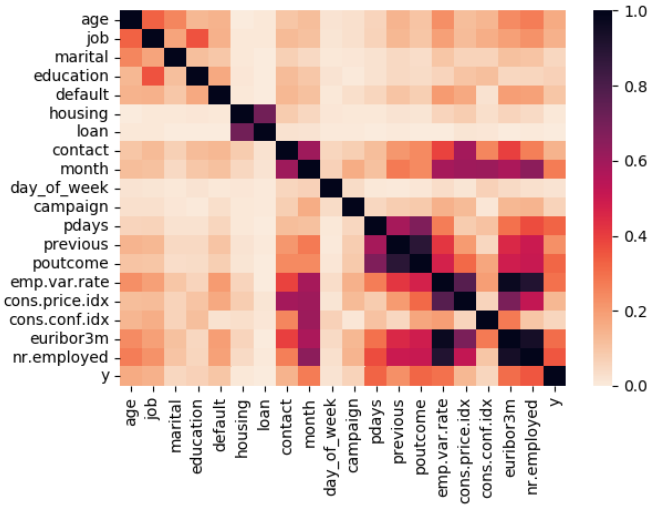


Figure 21: Correlation Matrix

Selection of Variables

From the correlation plot (Figure 21), we observe that the correlation between employment variation rate, 3 months Euribor rate and number of employees employed exceeds 0.9. This indicates that these three variables are highly correlated and keeping them will introduce multicollinearity into the models we build. As such, we have chosen to keep only one of these variables when building our models. In the context of our project, we have chosen to keep the Euribor 3 months rate. As a daily indicator, it should provide us with a stronger illustration of the economy as compared to the other two variables - employment variation rate and current employees headcount, which are both quarterly indicators.

Additionally, we also observe a strong correlation between 'previous' and 'poutcome'. Since 'previous' is highly skewed and provides less information as compared to 'poutcome', we chose to drop 'previous' and keep 'poutcome' in our models.

In summary, we kept the following numerical variables: Consumer Price Index, Consumer Confidence Index, 3 Months Euribor Rate; and the following categorical variables: 'Age', 'Job', 'Education', 'Contact', 'Month', 'Poutcome'.

Modelling

Encoding

In dealing with the categorical features, we applied label encoding to the features 'age', 'education' and 'contact' and one-hot encoding to the remaining categorical features - 'job', 'month', 'pdays' and 'poutcome'. Label encoding is applied to 'age' and 'education', as these features contain categories that can be ordered; and 'contact', as it is a binary variable. For one-hot encoding, it is clear that the categories in 'job', 'month' and 'poutcome' are nominal and hence, label encoding cannot be utilised for these three variables. As for 'pdays', the distributon of 'pdays' suggests that the categories may or may not be ordinal, so using one-hot encoding would be a safer choice.

Stratified Sampling

Stratified Sampling was used as the chosen technique to split the data set into training data and testing data sets. We used Stratified Sampling so as to ensure that the features that have the greatest influence on our response are equally distributed in the training and testing data set. Doing so would also ensure that both the training and testing dataset contain the same ratio of classes. Using the *StratifiedShuffleSplit* class from the *sklearn* library in python, we split the original data set into a 80:20 ratio, where 80% of the data set is used for training, and the rest for testing.

Oversampling

To combat the class imbalance in our data set, we implemented an oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE). By oversampling the minority class, this reduces the bias in our training data set which can influence many Machine Learning algorithms to focus their learning on the majority class, causing some to ignore the minority class entirely.

Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) generates synthetic data points using a K-Nearest Neighbour approach. We used SMOTE to create synthetic data points of the minority class, till the total number of the minority class matches the total number of the majority class. The SMOTE works by first selecting a random instance from the minority class, x_1. Then, k (usually k=5) of the nearest neighbours are identified. A random neighbour, x_2, from the k neighbours is chosen to be connected with x_1 to form a line segment in the feature space. Then, a randomly selected point on this line is chosen and its attributes are used to form a new synthetic data point of the minority class [14]. Since we are creating new synthetic points, we are adding in new information into the data set, which is more effective than the random oversampling technique. This is because random oversampling does not introduce any new data into the data set. Furthermore, randomly oversampling may also cause over-fitting of the model, because the model is learning examples that it has seen before. Under-sampling was also considered, but because under-sampling reduces the number of the majority class to match the number of minority class, we would be losing more than 85% of our data which could be detrimental to the learning process for our models. We used the python library *imbalanced-learn* to implement the SMOTE technique. The oversampling results are summarized in Table 4.

Before SMOTE	
Response	Count
'yes'	3712
'no'	29238

After SMOTE	
Response	Count
'yes'	29238
'no'	29238

Table 4: Smote Count

Proposed Models

Decision Trees

We used the class *DecisionTreeClassifier* to implement the decision tree classifier from the *sklearn* library. Decision trees are used for classification problems, and is a suitable model for our problem. Decision trees classify the examples by sorting them down the tree from the root node to a leaf node, with the leaf node providing the classification of the example. Each node in the tree contains a test case for a feature that is used to split the data set into subsets, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and repeated for every sub-tree rooted at the new node until all leaves are assigned a class. The criterion for splitting the data set at every node other than the leaf nodes is based on the purity of the node. The two ways to measure the purity of a node are via the Gini index and the entropy of a node. These measures evaluate the the quality of a split of a node at a particular feature.

The purer a node is, the more information is gained by traversing its different edges. Information gain measures how well a given feature separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain. Hence, the root node must be the most pure, i.e. where the maximum amount of information is gained when traversing down its edges. This rule is then followed for each sub-tree after traversing down the edges of the root node. A high information gain indicates that the feature provides more information about the target variable. Information gain is given by the following formula: $InformationGain = E(Y) - E(Y|X)$ where Y and X are features. To measure information gain, we use entropy. Entropy is a measure of the randomness in the information being processed. Entropy is given by the formula: $Entropy = \sum_{i=1}^{c} -p_i log_2 p_i$ where p_i is the probability of the different outcomes from using a feature to split the node. The higher the value of entropy, the harder it is to draw any conclusions from that information. A set with higher impurity will have higher entropy, while a set with higher purity will have lower entropy.

The Gini index measures the probability of assigning a wrong label to a sample by picking the label randomly. It is also used to measure feature importance in a tree. The Gini index is given by the formula: $Gini = 1 - \sum_j p_j^2$ where p_j is the probability of an outcome from using a feature to split the node. The lower the Gini index is for a node, the more pure the node is. Hence, the decision tree algorithm chooses a split that minimizes the Gini index.

By default, the decision tree grows deep and complex until every leaf is pure and hence it is prone to over-fitting. We shall control this using the $maxdepth$ parameter during hyperparameter tuning.

Random Forests

We used the class *RandomForestClassifier* to implement a Random Forest Classifier from the *sklearn* library. The Random Forest Classifier is an ensemble of decision trees for classification, where the decision trees make up the 'forest' in Random Forest. The output of the Random Forest uses the majority rule, where the mode of the predictions from each decision tree is used as the final output. The 'Random' in Random Forest follows two rules: first, the bagging technique deployed by the model, and second, the random choosing of features to teach each decision tree.

The bagging technique deployed by the Random Forest generates m new subsets of the original data set, D, using random sampling with replacement to form $D_1, D_2, ..., D_m$ data sets of size n. Then, m decision trees are fitted to each sub-data set $D_1, D_2, ..., D_m$.

A subset of features are randomly chosen to be learned for by each decision tree, which increases the diversity of each decision tree and keeps the trees short. This also helps to solve the problem of over-fitting that arises when using a single decision tree to do classification.

eXtreme Gradient Boosting (XGBoost)

We used the class *xgboost* to implement an eXtreme Gradient Boosting (XGBoost) Classifier from the *sklearn* library. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models that predict the residuals or errors of prior models are created and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. The 'eXtreme' in eXtreme Gradient Boosting means that the Gradient Boosting model is optimized using various methods such as regularization and tree pruning to avoid over-fitting [15].

Logistic Regression

The class *LogisticRegression* was used to implement a Logistic Regression model from the *sklearn* library. The Logistic Regression model uses a logistic function, generally the sigmoid function, rather than a linear function to output a probability distribution for a binary target variable. By having a threshold value, the Logistic Regression Model can be transformed into a binary or OneVsRest classifier.

In the Logistic Regression model, input variables are assumed to have a linear relationship with the log-odds probability of a positive output: $l = ln\frac{p}{1-p} = \sum_0^k \beta_k x_k$. For optimisation, the Logistic Regression Model make use of the *LogLoss* metric as well as regularization such as 'L1' or 'L2' to optimise the gradient descent. To ensure that the regularization is not affected by the scale of the features, scaling is applied to the features using *StandardScalar* from the *sklearn* library.

In addition, we used *f regression* from the *sklearn* library as well to conduct individual F-tests for each regressor in the model. By doing so, we can measure the p-values of the contribution of each regressor and ensure the robustness of the model by dropping regressors with low to no contribution.

Ensemble

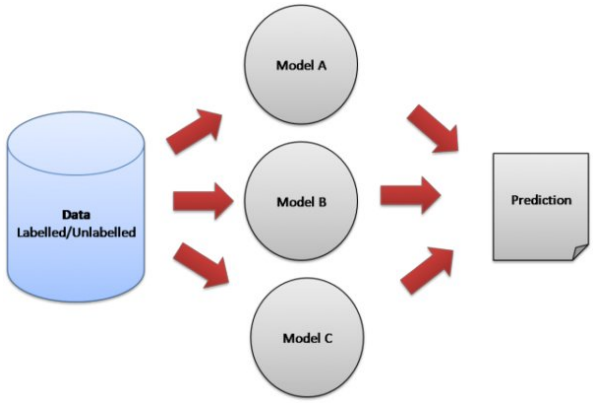


Figure 22: Ensemble

To increase the predictive power of our models, we considered building an ensemble model as well. Ensemble methods is a Machine Learning technique that combines several base models in order to produce one optimal predictive model (Figure 22). The intuition behind an ensemble model is that rather than just considering one fixed model to produce a prediction for a particular client, an ensemble model allows us to take a sample of models into account, making a final predictor based on the aggregated results of the base models. The base models that we have considered in the building of our ensemble model are the Decision Tree Model, Random Forest Model, XGBoost Model and Logistic Regression Model.

Support Vector Machines (SVM)

As an alternative to Logistic Regression and Tree-Based models, we built a SVM model as well. We used the SVM implementation available in the *sklearn* library, with the Radial Basis Function (RBF) as the kernel.

An SVM model is a model typically used for classification problems. It works, in our case, by creating a hyper-plane which attempts to separate the data into classes. In general, SVMs are faster to train as training only involves 1 pass of solving a quadratic optimisation problem, for which there are many efficient algorithms available, such as Sequential Minimal Optimisation (SMO). The learning process and output are also transparent to the user, making the results easier to interpret and problems easier to diagnose.

However, SVMs often require selected features from the data set. This would imply that SVM models might be less flexible and a new set of features may need to be selected for different data sets. Furthermore, SVMs also require a carefully selected kernel as the separability of the data points depends on choosing an appropriate kernel. This will also involve additional experimentation or specialized domain knowledge to identify.

Hyper Parameter Tuning

To optimise our models, we used the class *GridSearchCV* from the *sklearn* library to find the best hyper-parameters for our models. The *GridSearchCV* was optimised on the Precision Macro score with 5-Fold Cross Validation. Due to the imbalance present in our data set, we chose to optimise our models on the Precision score, which focuses on reducing the number of False Positives. Macro-precision measures the average precision per class [16], and this ensures that the model can identify both the Positive and Negative classes well. Table 5 shows the hyper parameters used to tune each of the models. GridSearchCV exhaustively searches through each combination of the hyper parameters, and returns the best hyper parameters that were able to optimise the precision macro score.

Decision Tree	**Random Forest**	**XGBoost**	**Logistic Regression**
criterion	criterion	learning_rate	penalty
max_depth	max_depth	max_depth	C
min_samples_split	max_features	colsample_bytree	-
min_samples_leaf	n_estimators	n_estimators	-

Table 5: GridSearchCV Tuning Parameters

Model Evaluation

To evaluate the performance of our Machine Learning models, we used the following six metrics:

1. Precision

Precision measures the proportion of positive cases that were correctly identified by the model. It has the formula $Precision = \frac{TruePositives}{TruePositives+FalsePositives}$.

2. Recall

Recall measures the proportion of actual positive cases which were correctly identified by the model. It has the formula $Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$.

3. Accuracy

Accuracy measures the proportion of the total number of predictions that were correct. It has the formula $Accuracy = \frac{TruePositives+TrueNegatives}{TruePositives+FalsePositives+TrueNegatives+FalseNegatives}$.

4. Specificity

Specificity measures the proportion of actual negative cases which were correctly identified by the model. It has the formula $Specificity = \frac{TrueNegatives}{TrueNegatives+FalsePositives}$.

5. F1 Score

F1 Score is the harmonic mean of Precision and Recall, which gives a better measure of incorrectly classified cases as compared to the Accuracy metric. It has the formula $F1Score = 2 * \frac{Precision*Recall}{Precision+Recall}$.

6. AUC

AUC stands for Area Under the Receiver Operating Characteristic (ROC) Curve. It provides an aggregate measure of performance across all possible classification thresholds.

The results of our models are shown in Table 6 and Figure 23. With the SVM, we are able to get 4 metrics above 0.7, meeting the requirement for this project. However, the trade-off of implementing a SVM is the lower precision that comes with it. The next best model is the Ensemble model, which scored higher than all other models in terms of AUC, Specificity, Accuracy and most importantly, Precision.

Model	Decision Tree	Random Forest	XGBoost	Logistic Regression	Ensemble	SVM
Precision	0.4748	0.4596	0.5225	0.3415	0.5242	0.2543
Recall	0.5690	0.6067	0.5129	0.6595	0.5011	0.7328
Accuracy	0.8806	0.8753	0.8923	0.8184	0.8925	0.7278
Specificity	0.9201	0.9094	0.9405	0.8386	0.9423	0.7272
F1 Score	0.5176	0.5230	0.5177	0.4500	0.5124	0.3776
AUC	0.7860	0.8161	0.8255	0.7981	0.8440	0.7763

Table 6: Performance Metrics of Models

From the Ensemble model, we derived the various models' importance and feature importance (Figures 24 to 28). This provides valuable insight in understanding which models and features contributed most towards the Ensemble's performance, allowing us to develop possible recommendations for the bank.

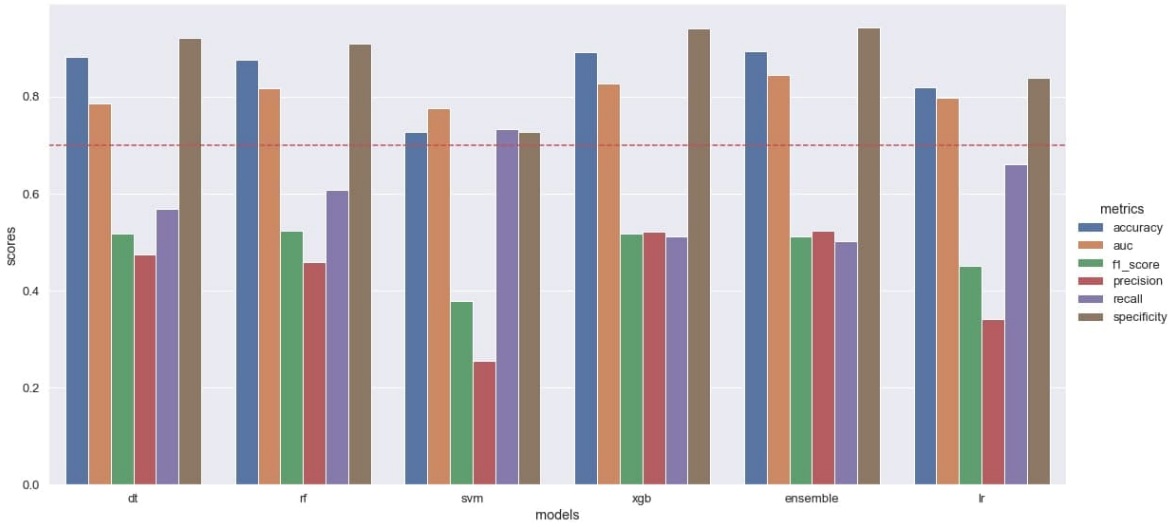


Figure 23: Model Scores

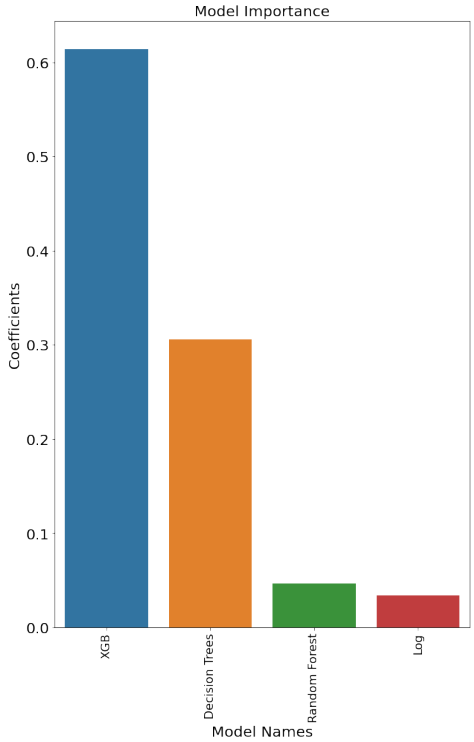


Figure 24: Model Importance in Ensemble

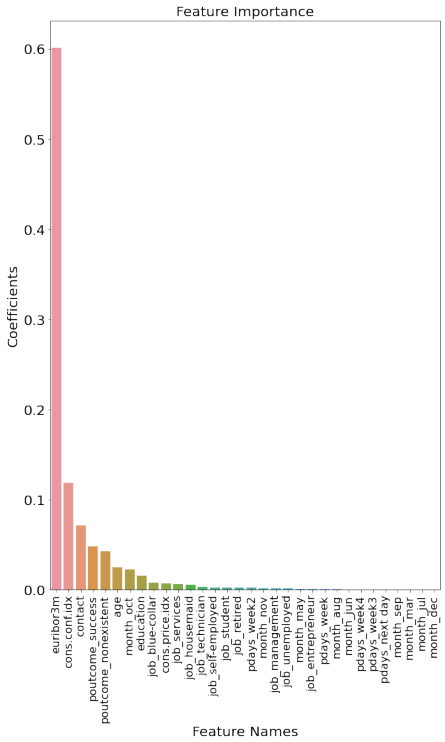


Figure 25: Feature Importance in Decision Tree

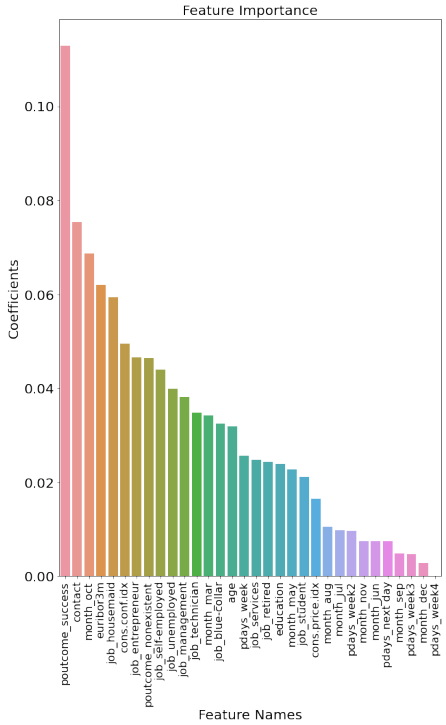


Figure 26: Feature Importance in XGBoost

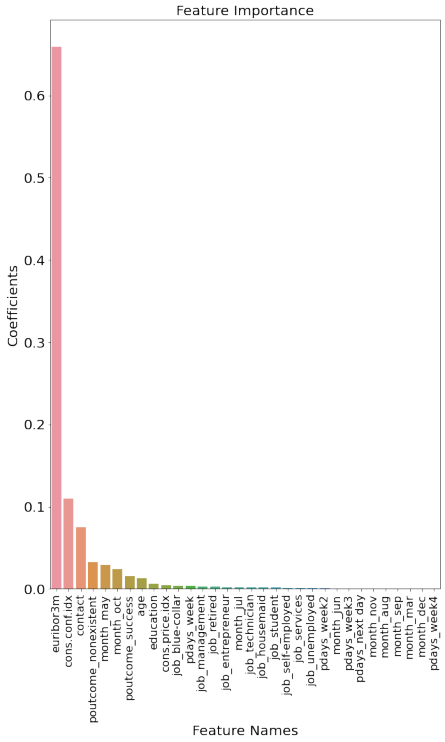


Figure 27: Feature Importance in Random Forests

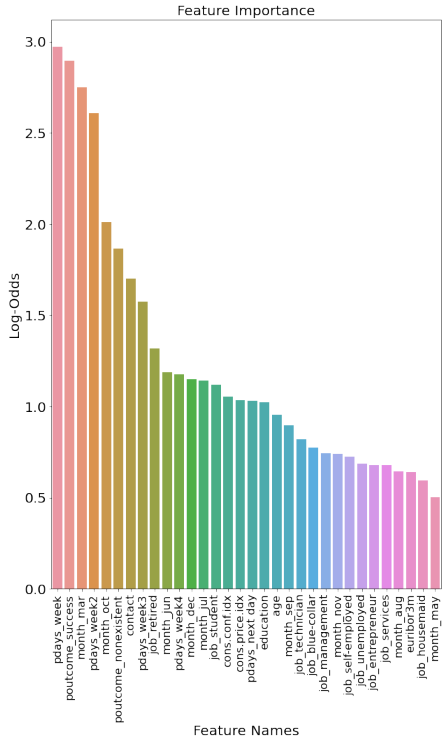


Figure 28: Logs-Odd Ratio in Logistic Regression

Discussion & Recommendations

Models

We have implemented various models that would enable banks to adopt a more targeted approach for their clients. Our models allow banks to predict the likelihood that a customer would subscribe to their services, based on some foreknown attributes about the customer. Used with the right domain expertise, this could allow banks to be more intentional in the customers they reach out to, reducing time and resources wasted by avoiding customers that might have a low predicted probability of subscribing to the bank's term deposits.

The ensemble model worked best in our project. The model was able to achieve a high accuracy, AUC and specificity, and a relatively strong precision as well. Given that our motivation was to reduce the amount banks spend on their marketing efforts, it is important that the ensemble model obtains a low false positive rate. Compared to the metrics of the other models, this model proves best in cutting down on wasted resources by banks when contacting clients.

Another model that seemed to do relatively well was the SVM model. Although it was able to achieve a high accuracy, AUC, specificity and recall score, the model appeared to do relatively worse in precision. While one can still run both models alongside each other, when considering the motivation behind our project, the ensemble model seems to be most reliable in achieving our goal.

The ensemble model built consists of four base models: an optimised Decision Tree model, an optimised Random Forest Model, an optimised XGBoost Model and an optimised Logistic Regression model.

From Figure 28, one can observe that the XGBoost model contributes the most to the ensemble model (60%), followed by the Decision Tree model (30%). Upon closer inspection of these two base models, some important features include Euribor Rate, Consumer Confidence Index, Previous Outcome, Contact and Age. As these features have the largest effect in influencing the subscription probability of a customer, banks should focus on demographics surrounding these features when rolling out their marketing campaigns.

Customer Demographics

Firstly, banks could consider rolling out marketing strategies that focus on mature seniors, seniors and youth, to increase sales (Age). Aside from the aforementioned groups, banks might not want to exclude the young adult demographic. We have selected these age groups based on our exploratory data analysis findings. Practically, since young adults are more likely to be employed and earn a higher income, banks might want to devise certain marketing strategies for them. Banks could target these particular age groups by launching marketing campaigns to students, retirees, admin, technicians and blue-collar workers before other job groups. We noted that these few occupations make up the bulk of the age groups of interest in the data set. Additionally, age was determined by our ensemble model to be a secondary feature of importance in influencing subscription rates. Hence, marketing to members of these demographics would naturally increase the bank's chances of converting a potential client. Banks could also consider targeting customers with a higher education level, as they generally have a disposable income available (Job). Our initial data exploration showed that a higher education level was associated with a higher subscription rate. Reaching out to individuals who have college education or higher, is one way banks might be able to raise their chances of onboarding a potential client (Education).

Secondly, banks could target loyal customers who had subscribed in previous marketing campaigns, as it usually costs less to retain existing customers than to acquire new customers, and existing customers are also more likely to subscribe again in the current campaign. Banks can then devise separate strategies to target new customers and prior customers (Pdays). For clients who have been previously contacted, banks could schedule follow-up meetings within one week, and in general, maintain an ongoing relationship with them. On the other hand, for clients who have not been previously contacted, banks could offer more incentives for subscriptions made on the

spot. Since new customers already form the majority of the client demographic, banks should focus on raising the success rate of their marketing campaign (Poutcome).

Communication

To communicate with clients, banks can contact clients via cellular phone, as clients contacted via cellular phone were much more likely to subscribe than those contacted via telephone, and the group size of clients who use cellular phone is larger than the telephone group (Contact).

Time Period

Banks may want to focus on conducting their next campaign on the following months - March, September, October and December. While these months showed higher subscription rates, the bank had reached out to fewer clients. A possible way to increase their subscription rates would be to to increase their targets outreach for the aforementioned months (Months).

Economic Factors

Furthermore, we advise that the bank keep track of the daily Euribor (3 Month) (euribor3m) rates, so that it can be forecast. This allows the bank to predict more suitable periods to conduct their next campaign. As seen from our exploratory data analysis, the lower the Euribor rate, the higher the rate of subscription. Virtually every model we have built has suggested that Euribor is one of the most important features in determining if a person will subscribe to a term deposit. As Euribor is a daily indicator calculated using a 3-month average, it is theoretically possible to forecast these rates into the future using recurrent neural networks, such as as the Long-Short Term Memory Model (LSTM) which is a powerful model often used in time-series applications. We strongly recommend that the bank should keep track of the daily Euribor rates, so as to forecast it into the future with the help of experts in time-series analysis.

Banks may also account for other external economic factors such as the Consumer Confidence Index to plan their marketing campaign and tailor their term deposit plans to encourage higher subscription. By having a high interest rate, this will serve as a great incentive to encourage people to invest in them to hedge against inflation.

Conclusion

Through this project, we are able to achieve our objective of identifying certain demographics that are key to a high subscription rate. Using this information, banks can maximise their subscription rate through focusing their efforts on clients that match those demographics. Furthermore, using our models, they can identify clients better, and can now plan when to conduct their next campaign based on economic factors and favourable time periods. However, model maintenance needs to be carried out periodically to ensure Machine Learning models are trained with updated data, and continue to make useful predictions. As banks continue to collect more data on clients and update the models, we anticipate our models would help banks improve their marketing campaign strategies.

References

[1] James Chen. *Term Deposit*. 2020. URL: https://www.investopedia.com/terms/t/termdeposit.asp.

[2] Trading Economics. *Portugal - Bank Credit To Bank Deposits*. 2017. URL: https://tradingeconomics.com/portugal/bank-credit-to-bank-deposits-percent-wb-data.html.

[3] Chris B. Murphy. *Loan-to-Deposit Ratio (LDR)*. 2020. URL: https://www.investopedia.com/terms/l/loan-to-deposit-ratio.asp.

[4] Adam Hayes. *Bank Run*. 2021. URL: https://www.investopedia.com/terms/b/bankrun.asp.

[5] The Investopedia Team. *Liquidity Crisis*. 2020. URL: https://www.investopedia.com/terms/l/liquidity-crisis.asp.

[6] ActiveWin. *How Much Do Banks Spend On Marketing?* 2020. URL: https://www.activewin.co.uk/blog/how-much-do-banks-spend-on-marketing/.

[7] Eurostat. *Glossary: Employment rate dispersion*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Employment_rate_dispersion.

[8] Jason Fernando. *Consumer Price Index (CPI)*. 2021. URL: https://www.investopedia.com/terms/c/consumerpriceindex.asp.

[9] Akhilesh Ganti. *Consumer Confidence Index (CCI)*. 2020. URL: https://www.investopedia.com/terms/c/cci.asp.

[10] Euribor-rates.eu. *What is Euribor?* URL: https://www.euribor-rates.eu/en/what-is-euribor/.

[11] Adam Hayes. *Interbank Rate*. 2021. URL: https://www.investopedia.com/terms/i/interbankrate.asp.

[12] Holborn Assets. *How Euribor Affects Your Savings and Loans*. 2015. URL: https://holbornassets.com/blog/finance/how-euribor-affects-your-savings-and-loans/.

[13] John T.E. Richardson. *Eta squared and partial eta squared as measures of effect size in educational research*. 2011. URL: https://www.sciencedirect.com/science/article/pii/S1747938X11000029.

[14] Yunqian Ma Haibo He. *Imbalanced Learning: Foundations, Algorithms, and Applications*. 2013. URL: https://www.wiley.com/en-us/Imbalanced+Learning%3A+Foundations%2C+Algorithms%2C+and+Applications-p-9781118074626.

[15] Jason Brownlee. *A Gentle Introduction to XGBoost for Applied Machine Learning*. 2016. URL: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.

[16] peltarion. *Macro-precision*. 2021. URL: https://peltarion.com/knowledge-center/documentation/evaluation-view/classification-loss-metrics/macro-precision.