BANK TERM DEPOSIT

SUBSCRIPTION ANALYSIS

Team : Aadithya Anandaraj, Sadhana Koneni, Satya Naren Pachigolla
UNDER THE GUIDANCE OF PROF.JASON DUAN| MARKETING ANALYTICS II

TABLE OF CONTENTS

1. Introduction

1.1 Background

Banks today are operating in a competitive and dynamic environment. The problem we would like to look at is the subscription of term deposits. Term deposits are short term deposits held at banks where the money is locked for a time period. The interest rates offered are higher than ordinary savings accounts and so there is incentive for the customer to opt for this. On the other hand, the banks also benefit when more customers subscribe for this. The bank can invest the money in other financial products that pay a higher rate of return or lend the money to borrowers and earn a higher interest. So, it is in the best interest of the bank to market this service to customers [5].

1.2 Dataset Overview

This dataset is based on "Bank Marketing" UCI dataset (please check the description at: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#).

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required to assess if the term deposit would be subscribed ('yes') or not ('no'). The classification goal is to predict if the client will subscribe to a term deposit (variable y).

1.3 Business Problem and importance

With information on marketing campaign for term deposits, the business questions we are trying to answer are:

1) Would a customer subscribe to a term deposit from the bank?
2) What is the profile of a customer who is more likely to convert?
3) If the bank is buying information from external sources for their campaigns, how valuable is the data?
4) When and where should the marketing efforts be aligned for maximum returns in future campaigns?

Targeting the right audience for a marketing campaign can save the bank thousands of dollars. An analytical approach to decision making would ensure that the bank sells the term deposit service to the right set of customers. Answers to these business questions would help the bank save money by assessing the data source and develop more targeting marketing campaigns.

1.4 Objective

The objective of this analysis is to provide a reliable and feasible recommendation algorithm to predict client propensity to subscribe based on client type and campaign type. The target value is the binary “yes” or “no” regarding the subscription of term deposit by the client. Hence the task could be solved by classification models. We plan to use look at F-scores in the end to judge which model generated more accurate results and then look at the key variables that are used by the model.

1.5 Approach

To understand the customer propensity to opt for a term deposit with the help of the dataset, we plan to take the following steps:

1) Import data and perform initial exploratory analysis: the size of the data, explore presence of missing values, outliers. In addition, study the effect of various features on the campaign outcome and possible interactions within the variables.
2) Clean the data by removing insignificant variables, handle missing values and outliers , convert categorical columns into dummy variables.
3) Use machine learning techniques to predict the outcome and to find out the factors which affect the success of the campaign.

2. Data Summary and Exploratory Data Analysis

The secondary dataset on marketing campaign information by a bank had columns of both categorical and numerical variables along with the binary response variable (‘yes’ or ‘no’) indicating whether a given client had subscribed to the bank term deposit. The variables can be broadly described as the *client related information* (age, job, marital status, education, default, housing loan, personal loan), *contact related information* (contact mode, month, day of week, duration of the last contact, number contacts made for current campaign, number of days since the last contact for a previous campaign, previous contacts made, previous campaign’s outcome)

and *external indicators* (five social and economic features/attributes of nationwide indicators - employment variation rate, consumer price index, consumer confidence index, euribor 3-month rate, number of employees). We had 41188 rows of data with each row representing one client's data for the success of the marketing efforts for the subscription.

From Table 1 with the summary statistics of the numerical columns, we observed that the 'pdays' column has a value of 999 which upon inspection we found indicates that the customer has not been contacted for any previous campaign. We noticed that the age column needed to be either scaled or split into categories as the range of values is high.

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Table 1: The summary statistics of the quantitative columns

The missing values have been filled with 'unknown' as the value for the categorical columns. It was observed that the percentage of missing values was below 5% of all columns except the 'default' column. The 'duration' column was dropped from analysis. Although this variable highly affects the output target, the duration is not known before a call is performed i.e. it would not be available for predicting the test data. The numerical columns did not have any missing values.

We then moved onto exploratory data analysis. First, we explored the relationship between the categories in the column with the campaign success. We wanted to understand if there is a significant difference in the success percentage based on the category the client belongs to.

The bar chart in Figure 1 shows the distribution of campaign success based on job, marital status, and education of the client. The likeliness of subscription is high if the clients are students or retired. There is not a significant difference among the clients of different marital status although

single customers are slightly more likely to subscribe. From the third plot, we observed that the illiterate clients are more likely to subscribe followed by the ones that hold a university degree.

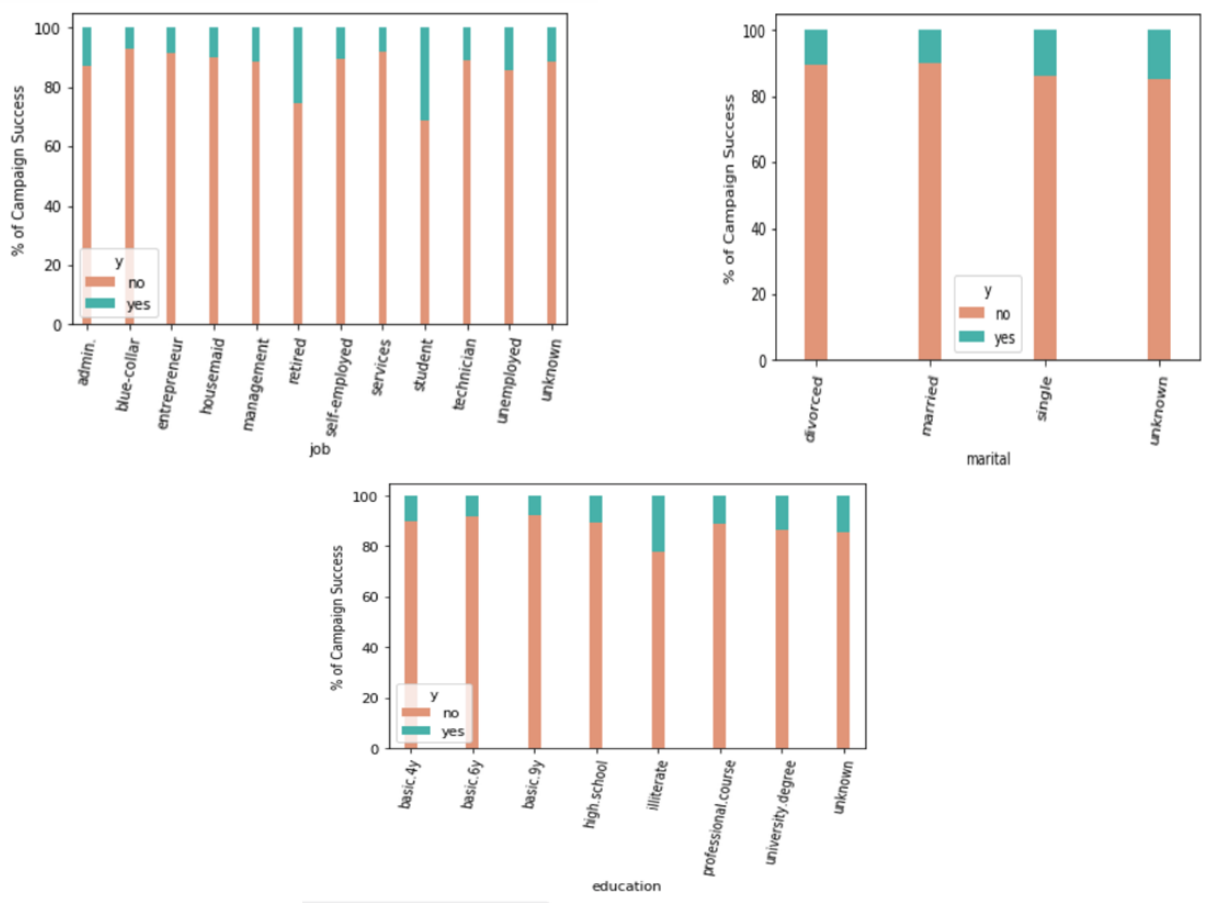

Figure 1: Percentage distribution of campaign success based on job, marital status, and education

From the plots in Figure 2, we noticed that there is no significant difference in the campaign success based on whether the client has a housing loan or a personal loan.

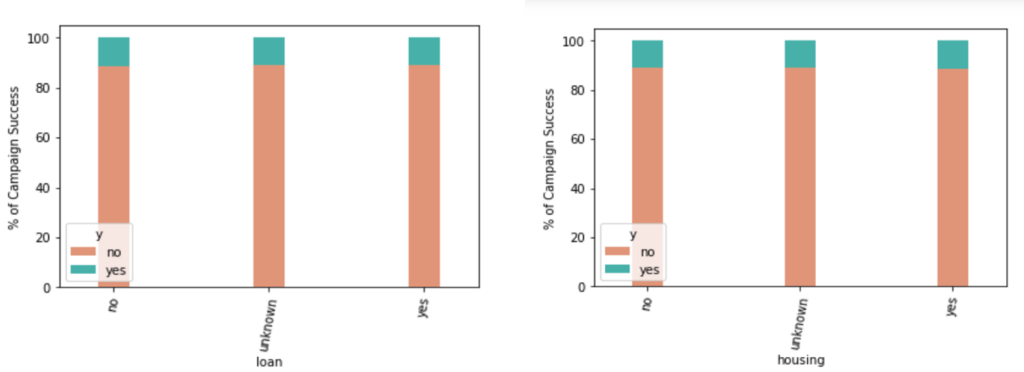

Figure 2: Percentage distribution of campaign success based on personal loan and housing loan

From the plots in Figure 3, we observed that the likelihood of subscription is higher if the client is contacted in the months of March, September, October, and December compared to other months. On the other hand, the day of the week of contact for the campaign does not significantly affect the likeliness of subscription.

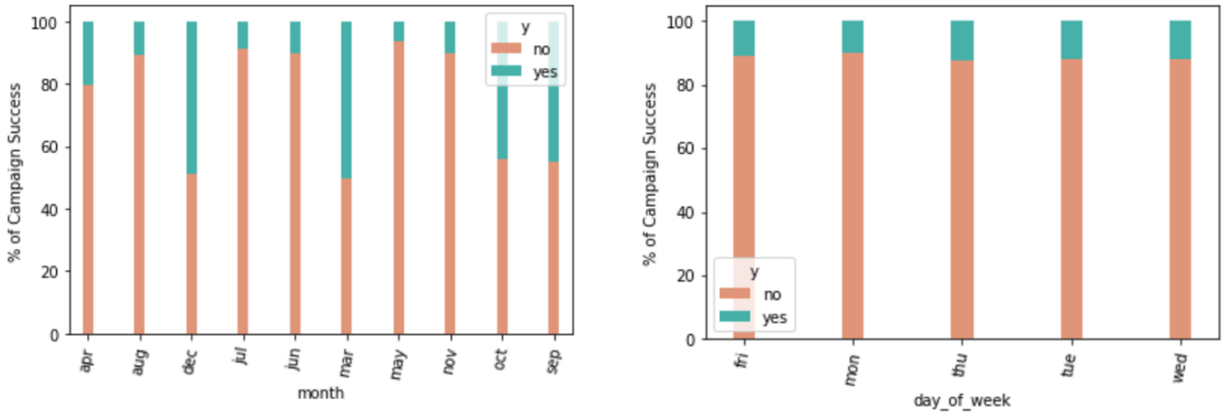

Figure 3: Percentage distribution of campaign success based on the month and day of the week

From the plots in Figure 4, it is evident that the clients who responded positively to the previous campaign are more likely to subscribe with the current campaign. The customers who were contacted on their cellular phone were more inclined to subscription compared to the ones reached out to through telephone.

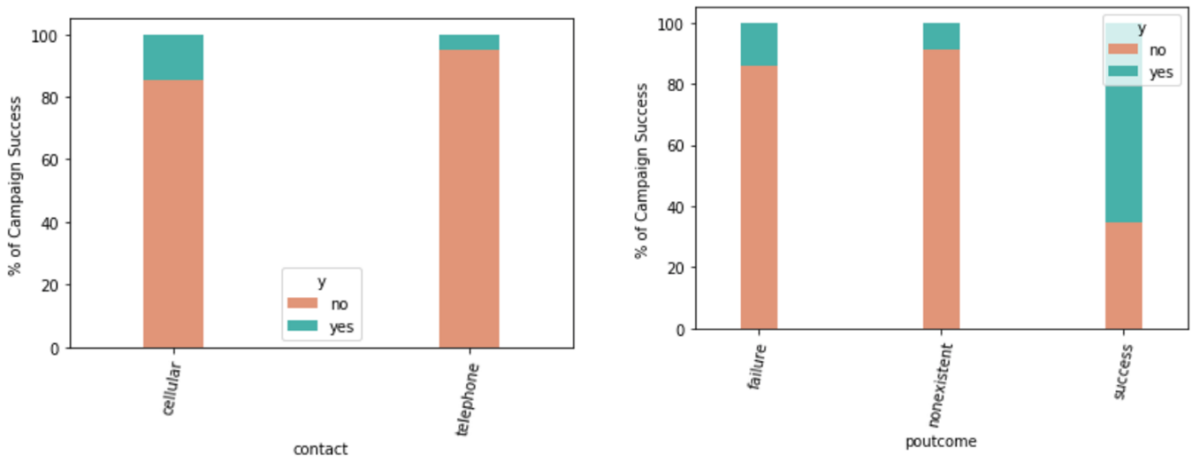

Figure 4: % distribution of campaign success based on outcome of previous campaign & mode of contact

From the box plot in Figure 5, it can be concluded that the older clients are more likely to subscribe which falls in line with the previous observation about the higher campaign success when the clients are retired.

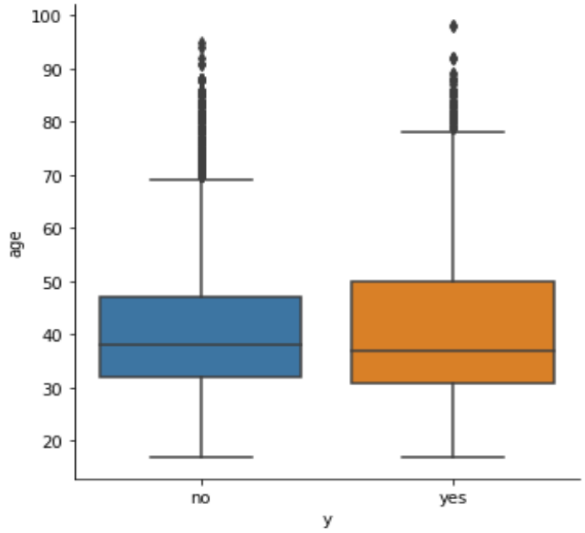

Figure 5: Distribution of age based on campaign success

3. Data Analyses, Key Findings and Conclusions

Given that the goal was to predict the likelihood of subscription, we treated this as a binary classification problem. Before stepping into the model building stage, the data warranted us to handle class imbalance, format categorical variables and check multicollinearity. We did the following:

- **One Hot Encoding :** The categorical features were one hot encoded with one dummy column created for each of the N-1 categories (where N = number of categories in the feature). The reason for considering N-1 categories is to avoid multicollinearity. While tree-based models would handle categorical variables, this step is important for models such as logistic regression that require numerical independent variables.

- **Reduce Multicollinearity :** We dropped variables with correlation scores of 90% and higher.

● **Remove Zero variance variables :** The features with constant values, implying zero variance and hence no information were dropped.

● **Handle class imbalance:** The dataset was imbalanced implying that the label had a smaller proportion of customers who subscribed (37%) when compared to the other class - customers who did not subscribe (63%). The class of interest to us was the subscription class and to create an effective and accurate model, the data needed to be balanced. We used this project to study the performance of sampling methods and intended to choose the method that performs best on the model for this data set.

3.1 Sampling methods

3.1.1 Oversampling

This is a method of sampling where the minority class is resampled with repetition to match the size of the majority class. The result is a data set with balanced classes. It is important to note that this was done after splitting the dataset into train and validation sets. The reason is that if it is done before splitting the data could allow for exact data points to be present in the train and test set. This would cause overfitting and poor generalization to the test data[2].

3.1.2 SMOTE

SMOTE(Synthetic Minority Oversampling Technique) as the name suggests is an oversampling technique where the minority class examples are oversampled. Instead of the naive method of duplicating the minority samples in the training data, SMOTE synthesizes new examples from the minority class. It works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen, and a synthetic example is created at a randomly selected point between the two examples in feature space. A general drawback of the approach is that synthetic examples are created without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes[1].

3.1.3 Under sampling

This is a resampling technique where the size of the majority class is reduced to match the minority class. This is done by randomly selecting a few points from the majority class, while removing the other data points. A potential drawback in this method is that we are removing information that may be valuable to the model. So, this could lead to underfitting and poor generalization, possibly for the majority class.

3.1.4 Comparison of sampling methods

This being a classification problem, we used the Logistic regression model and recall score to evaluate the performance of the sampling methods. The reason for the choice of model and metric will be explained in detail in the later section.

Comparing the performance of the sampling methods on a Logistic Regression model in Table 2, we observed that Down sampling gave the best results. Considering that in addition to improvement in scores, this would also reduce computation time and resources, going further, we used the down sampled data.

Logistic	No sampling	Up sampling	Down sampling	SMOTE
Recall	0.23	0.63	0.64	0.34
F1-score	0.34	0.43	0.43	0.38

Table 2: Comparison of logistic regression results from different sampling and no sampling methods

We now had an under sampled data set with 6494 rows and 17 features. We wanted to further use other models for prediction, starting with Probit and clog-log links and comparing it with the Logit model.

3.2 Choice of metric

The metric considered to evaluate the performance of the model was Recall. It is defined as,

Recall = True Positive / (True Positive + False Negative)

where, True Positive is when the model correctly predicts the actual positive class (subscription class here). False Negative is an outcome when the model misclassified the positive class. This score is high when true positive is high and false negative is low.

With more emphasis on the banks being able to identify potential subscribers, we wanted the recall to be high as the cost of misprediction of subscription is much higher than misprediction of non-subscription. At the same time, we wanted to keep a tab on the precision value to maintain the trade-off.

3.3 Model Building

3.3.1 Logistic Regression

We wanted to obtain probability values and the log link bounds the value between 0 and 1 making it suitable for a classification problem. It is a simple model with an interpretable equation and the coefficients give us the direction of influence on the dependent variable. So, for these reasons we used this model. This model assumes linearity between log odds and the features.

Using the logit model from the stats model package, the coefficients values along with their p-values were observed. With the threshold for significance set at 0.05, we removed the insignificant features. We also observed the following results:

Metric	Log-likelihood	LL-Null	Pseudo - R2	LLR- p value
Value	-3422.9	-4501.3	0.2396	0.000

Table 3: Results from the logistic regression model

Log-likelihood is the maximized value of the log-likelihood function for a full model. **LL-Null** is the result of the maximized log-likelihood function when only an intercept is included. **Pseudo-R2** is a substitute of the familiar R2 available under least squares. It is computed based on the ratio of the maximized log-likelihood function for the null model and the full model. The value of 0.2396 for this model here shows that the fit is not good and the explainability of the label by the features considered is very less. The null hypothesis of the LLR test is that the restricted model performs

better. The low **LLR - p value** here suggests that we can reject this hypothesis and prefer the full model over the null model.

3.3.2 Probit Regression

This model is widely used in binary classification problems and is very similar to logit in the sense that it estimates the probability of a value belonging to one of two classes. Both logit and probit take a linear model and feed it through a function to yield a nonlinear relationship. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors.

3.3.3 Clog-log regression

This model is also an alternative to the logit and probit model in case of binary classification problems. It gives the likelihood of occurrence of a class.

3.3.4 Comparison of Logit, Probit and Clog log

Comparing the BIC values of the models in Table 4, we observe that the logit link model has the lowest BIC and hence is the best model for this data set.

Model	Logit	Probit	Clog-log
BIC	7373.600	7377.765	7378.691

Table 4: Comparison of Logit, Probit and Clog-log models

So, the recall score for the logit model with the significant features selected was calculated to be 0.64 in the validation set. This was used as the baseline model. The important features were obtained based on the coefficient values. The plot in Figure 6 shows the relationship between the significant features and the response variable. The bars that are pointing towards the left have a negative relationship indicating that a higher value of that variable leads to lower probability of subscribing to the term deposit and the ones towards the right with higher values increase the likeliness of the subscription.

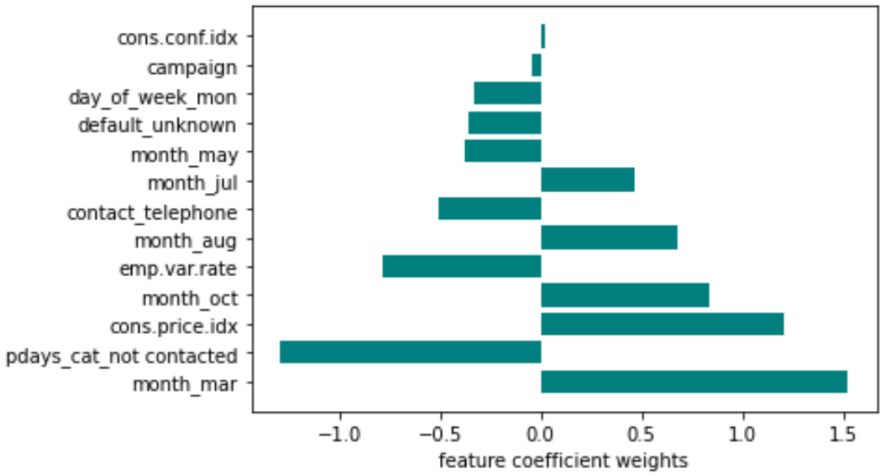

Figure 6: The coefficients of the predictors that indicate their relationship with the response variable

3.3.5 Random Forest

This tree model gives more accurate results as it is an ensemble of many individual models. The random selection of features to build each model makes the trees built less correlated and so this could improve the results. With this model, we observed a recall score of 0.69. Also, it gives the important features by using information gain. The variable importance plot in Figure 7 shows the relative importance of the features for determining whether a client is more likely to subscribe. We can observe from the plot that the most important features are the socio-economic indicators - employment variation rate, consumer price index and consumer confidence index, the number of days since last contact for a previous campaign, the mode of contact being telephone followed by month/day indicators.

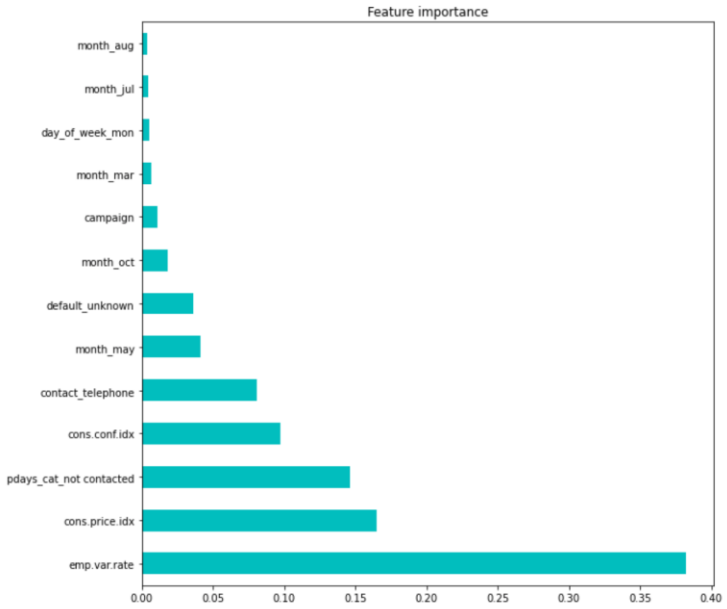

Figure 7: Variable importance plot from the Random Forest model

3.3.6 SVM

SVM is an effective model in high dimensional spaces. As we had a high number of predictors, we decided to try this algorithm for our dataset. It uses a subset of points for the decision function and hence is memory efficient. It is a versatile model as it provides different Kernel functions that can be specified for the decision function. Common kernels are linear, radial basis function (rbf) and polynomial although custom kernels can also be defined. After using grid search, we arrived at the best parameters and the best kernel turned out to be 'rbf'. This model gave us a recall of 0.70 for the threshold on 0.48 for the validation set [4].

3.3.7 Light GBM

After trying different models, we wanted to test the results using a gradient boosting model - Light GBM. This is a fast, high performing algorithm which is based on decision trees. It uses only leaf-based growth strategy for trees and splits the leaf that reduces the loss the most. While Random Forests are an ensemble of strong learners, Light GBM is an ensemble of weak learners. We observed reduction in computation time when compared with Random Forest.

This model gave us a recall of 0.62 for the threshold of 0.48 on the validation set. The feature importance in Figure 8 suggests that 'Campaign' is the most important predictor followed by the social and economic context attributes (consumer confidence index, consumer price index, employment variation rate).

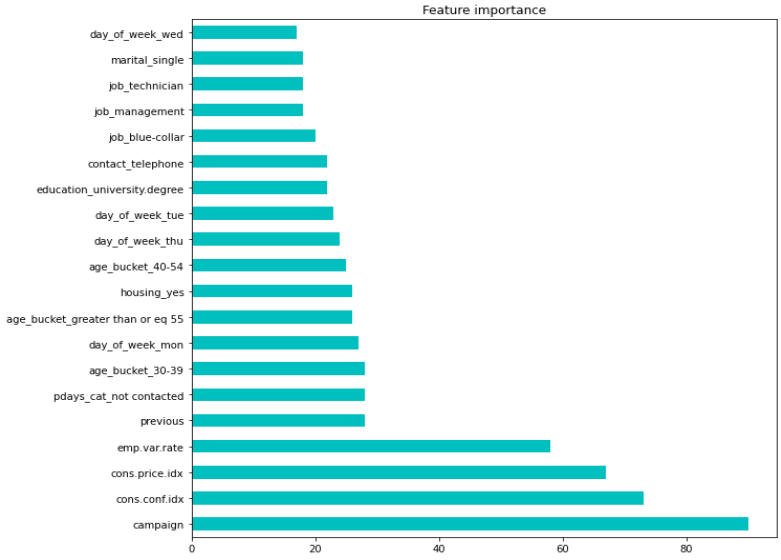

Figure 8: Variable importance plot: Light GBM

3.3.8 Cat Boost

Cat Boost is a machine learning model based on gradient boosting principle which is specialized in handling categorical data. Since most of the columns in our dataset are categorical, we wanted to see if Cat Boost can give superior results.

Our Cat Boost model gave us a recall of 0.62 for the threshold of 0.48, which is slightly below the Light GBM models. The variable importance plot in Figure 9 shows that 'employment variation rate' is the most important factor followed by the dummy variable for the month May.

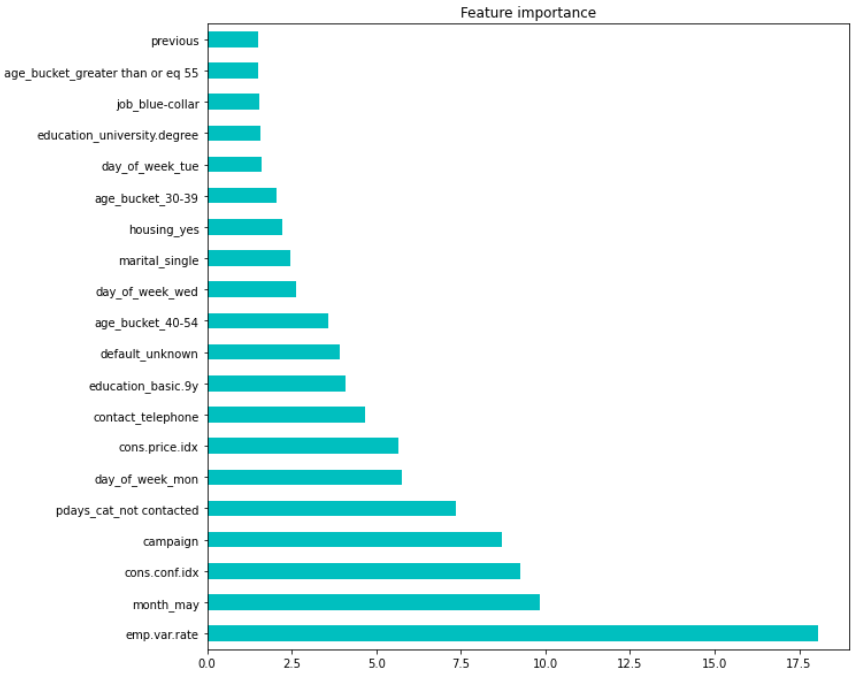

Figure 9: Variable importance plot: Cat Boost

3.3.9 Comparison of models

Comparing the scores on the validation set for the various models, we see that SVM Classifier has the highest recall score of 0.70, but Random forest and Logistic Regression have comparable recall scores and are more interpretable with feature importance. So, we select the Random Forest and Logistic model to base our recommendations on.

Model	Logistic	Random Forest	SVM	Light GBM	Cat Boost
Recall	0.64	0.69	0.70	0.62	0.64

Table 5: Recall scores from models

4. Recommendations, Limitations and Future Research

Studying the results of the best model and the inferences we got using the feature importance values, we recommend the following, we recommend the following to the bank staff:

- The Employment variation rate is a strong indicator of the probability of a customer subscribing for term deposit. Decrease in this macroeconomic factor leads to a higher likelihood of subscription. So, it is important that the bank records the changes in this metric to estimate if the period is good for contacting customers in this regard. This analysis would help the bank anticipate the cost involved and save it in case of economical stressed situations.
- The Consumer price index and confidence index are also indicators of performance of the economy. These metrics being high shows that there is a higher probability of subscription. These metrics can be used by the bank to also decide on good performing markets in addition to the period to focus their marketing efforts. It is important to note that these are lagged factors of performance in the sense that they indicate the performance of the market in the past quarter.
- The feature indicating whether a customer has been contacted in the previous campaign shows that the probability of a customer subscribing to a term deposit is lower if he is being contacted for the very first time. This intuitively makes sense as if the customer has been in the marketing lead database of the bank, it shows that he is of interest to the bank owing to his propensity of buying or earlier purchases he has made with the bank. So, it is recommended that while the bank reaches out to new customers, they also contact their repeat customers for subscriptions.
- Observing the feature month, it is seen that contacting the customers in the months of March, October, august, July gives the bank a higher chance of subscription while the month of May reduces the chance of subscription. So, they could plan their marketing activities in the months that increase the chance of subscription.
- In addition to the month, it is seen that contacting a customer on Monday leads to lower probability of subscription. So, the bank could schedule their staffing needs accordingly and re-allocate staff for other activities.
- It is also important to decide as to how many times a customer needs to be contacted. The campaign feature indicating the number of times a customer is contacted in the present campaign shows that the more one is contacted, the lesser is their chance of conversion. So, it is important for the bank to design their calling outreach programs accordingly.

The results have helped us build recommendations to the bank on indicators to decide on the socio-economic factors, time of contact and the previous contact pattern for the marketing campaign to the bank. However, it would be interesting to have variables regarding potential revenue in case of subscription, cost of marketing programs and indicators on the performance of the calling agents. This would help us study the cost-revenue analysis of the program and help navigate the calling agents to increase their productivity and customer experience.

5. Appendix

References:

1. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
2. https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28
3. https://www.statsmodels.org/stable/index.html
4. https://scikit-learn.org/stable/modules/svm.html
5. https://www.investopedia.com/terms/t/termdeposit.asp

The code book and data set used has been submitted.