Data Science Term Deposit Project Report
Portuguese Banking Institution

Term Deposit Predictive Model “Predict whether client subscribes long term deposit through telemarketing”

Project Synopsis

Team Alpha has analyzed the data related to the direct marketing campaign of the client, a local Portuguese bank. Team Alpha has created different predictive models (Logistic Regression, Random Forest, Decision Tree, and Linear SVC) while applying the best practices of data science. The project identifies the critical features of data (key performance indicators - KPIs) that can help increase the term deposit subscriptions. Team Alpha has put forward recommendations and conclusions based on the stringent evaluations of predictive models and their performances. The machine learning model will help the bank officials—VP Marketing, VP Sales, Digital Marketing Manager, Marketing Manager, Financial Advisor, call center managers, call center agents, customer sales analyst, customer sales analyst, business analysts, regional marketing manager, salesperson, sales manager, customer business manager, and shareholders—understand the factors that are playing a critical role in converting new clients as subscribers to the term-deposit campaign. The marketing officials can improve their campaigns while considering crucial factors and circumstances. The call center managers can improve the efficiency of their call agents by finding out the key performance indicators in the telemarketing campaigns.

OVERVIEW

1. Problem Statement

The Portuguese bank wants to sell its term deposits to its new and existing clients. Team Alpha wants to help the bank officials —VP Marketing, VP Sales, Digital Marketing Manager, Marketing Manager, Financial Advisor, call center managers, call center agents, customer sales analyst, customer sales analyst, business analysts, regional marketing manager, salesperson, sales manager, customer business manager, and shareholders—generate more revenue through term deposits, higher return on investment (ROI) as a result of marketing campaigns, increase the bank's customers base, and forecast who can be the ideal potential customers, and predict the likelihood of their conversions as subscribers to the term-deposits.

2. Research Methodology & Ethics

Team Alpha thinks of data science as a multidisciplinary field dealing with technology, process, and systems to extract in-depth knowledge and actionable insight from data. We tackle two aspects of data science: the management and processing of data and analytical methods and theories for descriptive and predictive analysis and prescriptive analysis and optimization. We used Jupyter notebook to understand data (numerical and categorical), find out missing values for imputation, unique values, imbalanced data, outliers, key performance indicators, correlations, etc.

3. Visualization with Dashboard

Team Alpha uses Jupiter notebook seaborn library and Tableau to create a dashboard that can help the bank officials and marketing decision-makers to understand the relationship between different features, their impact on the outcomes of the campaign, model features, and new features and factors that can help overcome the marketing campaign problems and target the right potential clients who meet the certain criteria. Following is the link to the Tableau dashboard: https://public.tableau.com/profile/hassan6741#!/vizhome/BankTermDepositHassanGMarch16/BankInstitutionTermDepositPredictiveModel?publish=yes

4. Model Creation

Team Alpha has built different machine learning models like Random Forest, Decision Tree, Linear SVC, and Logistic Regression. Models have been evaluated based on the ROC-AUC curves. We compared the four models (Random Forest, Logistic Regression, Decision Tree, and Linear SVC) based on the training score which are 96.3%, 90.4%,96.3% and 90.0% respectively. We picked the Random Forest classifier, considering performance metrics like accuracy, ROC Curve, and AUC.

5. Recommendations & Conclusion

Team Alpha has concluded that many features played a role in term deposit subscription. The client should add more features to the dataset to understand competitors' rate offerings, economic conditions, etc.

Problem statement

The banks are using marketing strategies to grow client subscriptions to investments resulting in increased customer retention. Telemarketing has proven one of the best-selling techniques; Phone calls made by banks gain more investment and enlarge company profits. Our client, the Portuguese bank, wants to sell its term deposits to its new and existing clients. The bank is running a marketing campaign targeting different people hailing from different financial backgrounds, educational levels, job titles, etc., throughout the year. The bank gathered the users' data during that campaign. Even though this seems a successful working strategy, this can still be improved to maximize profits. Team Alpha suggests that these marketing strategies coupled with statistical techniques can help predict outcomes, and the client can gain a competitive edge in the industry. While using data analysis, modeling, machine learning, and classification algorithms, Team Alpha can help the client make the predictions which can help refine the marketing strategies and customize them appropriately for its different customers' base.

The Team Alpha looks at the dataset to figure out the likelihood of a customer getting subscribed to a term deposit and assess the overall performance of the campaign, especially in terms of key performance indicators (KPIs), critical features in data, the total number of successful subscriptions, marketing expense, high productivity, time efficiency, and return on investment (ROI).

Our client has the dataset about the banking customers living in the same geographical location. The bank is spending a considerable amount, time, and resources on marketing campaigns to identify and win customers to sign up for term deposits. The predictive modeling can optimize the response of term deposit subscriptions when the marketing campaign is run.

The bank officials can learn how to make marketing campaigns more effective while targeting the right customers to get more term deposit subscriptions. The bank officials including VP Marketing, VP Sales, Digital Marketing Manager, Marketing Manager, Financial Advisor, call center managers, call center agents, customer sales analyst, customer sales analyst, business analysts, regional marketing manager, salesperson, sales manager, customer business manager, and shareholders can generate additional revenue through term deposits, higher return on investment (ROI), increase bank's customer base, and forecast the number of financial advisors required to meet the required targets.

Team Alpha handles the project to help the client achieve the business objectives. Some of the key questions that can be answered include the following: determination of the critical factors contributing to the success of the marketing campaign; finding out the additional features and their relationship with the term deposit subscriptions; building machine learning models to predict the outcomes of future campaigns; and identification of the customers most likely to subscribe to a term deposit.

Research Methodology & Ethics:

The key factors in determining how well a machine learning algorithm can learn to include the quality of the data and the amount of useful information that the data contains. Therefore, we ensure that we examine and process the dataset before we use it to feed it for modeling. The sole objective was to achieve these objectives: finding, removing, and imputing the missing values from the dataset; shaping up categorical data for machine learning algorithms; and zeroing in on the relevant features for the model building.

In a broader sense, Team Alpha thinks of data science as a multidisciplinary field dealing with technology, process, and systems to extract in-depth knowledge and actionable insight from data. We help the client understand the business problems leading to sound reasoning and effective decision-making. We tackle two aspects of data science: the management and processing of data and analytical methods and theories for descriptive and predictive analysis and prescriptive analysis and optimization.

Major Elements:

The methodology process that we followed in the data analysis can be classified into the following categories: data exploration, imputation, and feature engineering.

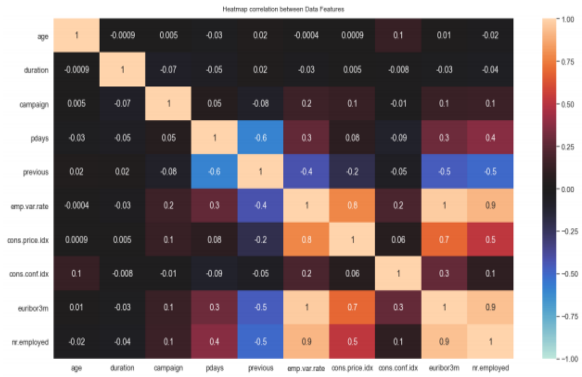
Exploratory Data Analysis:

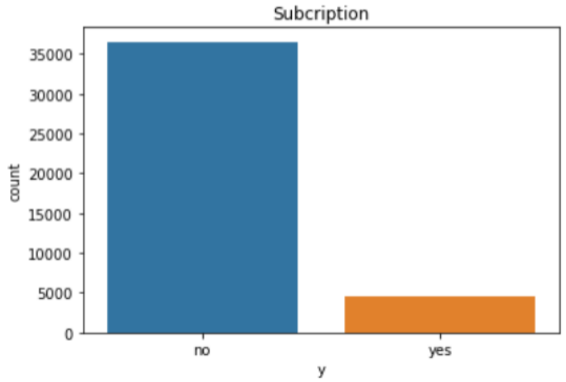
We have performed exploratory data analysis to gather patterns and insights that can help our client attract more customers for its 'term deposit' campaign, spend marketing budget judiciously, and extract actionable insights helping us in feature engineering. The UCI machine learning repository is used to download the banking data for analysis. Dataset is downloaded from https://archive.ics.uci.edu/ml/datasets/bank+marketing# . The repository has 4 data files. However, bank-additional-full.csv with 41188 observations and 21 inputs is randomly selected to upload in Jupyter notebook for further analysis.

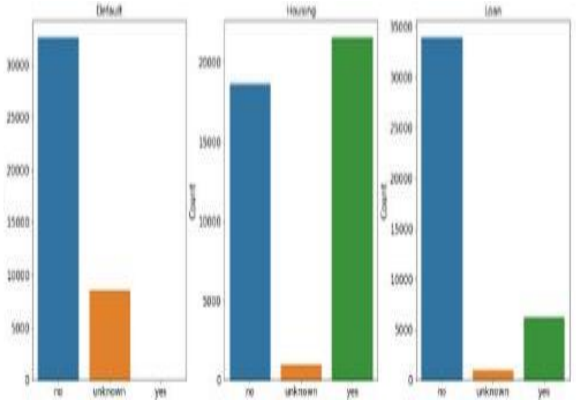
We decided to separate the data into 3 parts using binning method:

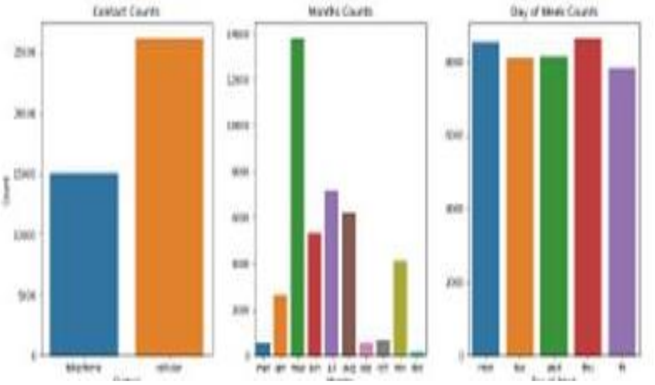
- Client Data: 1-7 columns/variables
- Marketing Data: 8-15 columns/variables
- Economic Data: Bucket with remaining features

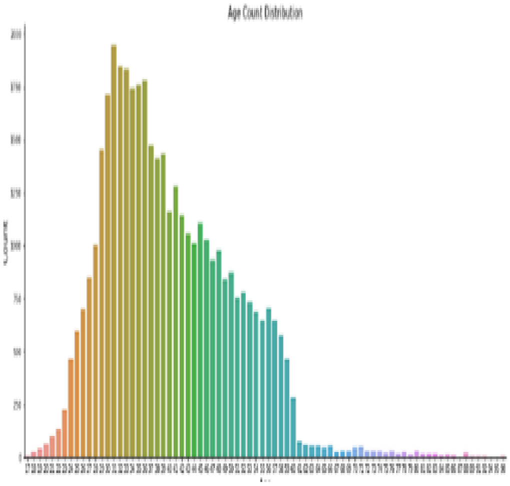
The methodology process that we followed in the data analysis can be classified into the following categories: data exploration, imputation, and feature engineering. We analyze the raw data as “a good step is to analyze the variety of values the different columns in the raw data take.[1] In our data exploratory analysis, we have streamlined the following attributes of the data: shape and size of the dataset, information and statistics summary, unique values, missing values, correlation of the dataset variables, outliers, etc. Descriptive statistics such as mean, media, mode, extremes (max and mean) are helpful at this stage.[2] We used Jupyter and Tableau to analyze, visualize, and find out correlations among the data features. Tableau also helped us figure out how the dependent variables were impacting the output variable.

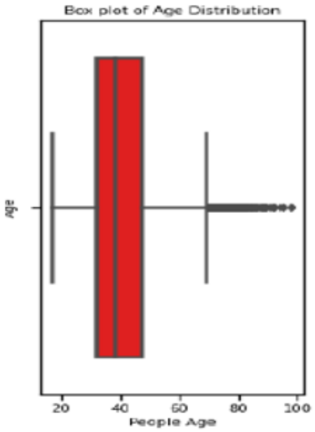


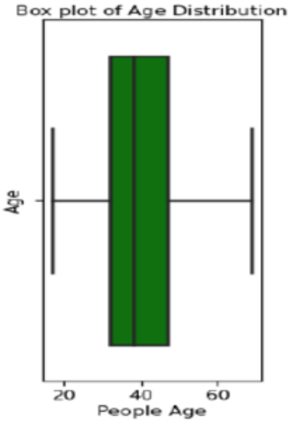












Imputation:

Imputation is the process used to replace missing data with substituted values in a given dataset. There are no missing continuous values in this data set. Thus, no imputation is necessary. We used imputation with mean value dealing with People Age outlier and duration in our key finding

Feature Engineering:

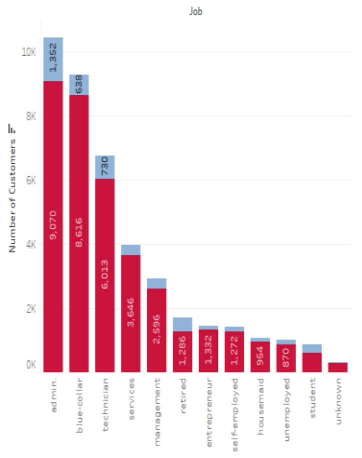
Feature Engineering is classifying features such as numerical and categorical into groups to deeply section and analyze the data for results in machine learning algorithms. In this section, we will create features for our predictive model. For each section, we will add new variables to the data frame and then keep track of which columns of the data frame we want to use as part of the predictive model features. We break down our dataset into numerical and categorical features.

Machine learning algorithms only read numerical values, which is why we need to change our categorical values to numerical values. We made use of label encoder on marketing data set and pandas get dummies method to one-hot encode the columns onto client bank data set. In one-hot encoding, a new column for each unique value in that column is created. Then the value of the column is 1 if the sample has that unique value or 0 otherwise. We created 34 features, including 1 numerical feature and 33 categorical features. We are sharing some of the visualizations of the feature engineering process.

Job vs Term Deposit Subscription:

What occupation should we target more?

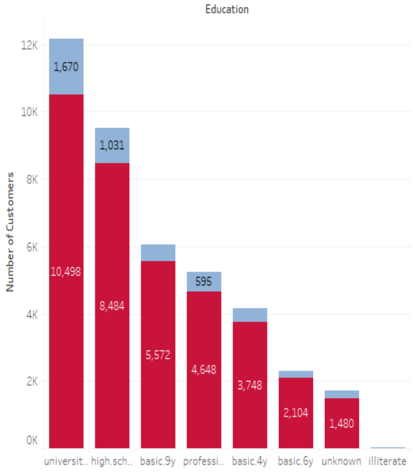
Analysis - Admin and Blue-Collar Individuals are more likely to sign up for the Term deposits as compare to other occupations.



Education Vs Term Deposit Subscription:

What educational level should we target more?

Analysis – University Degree holders are more likely to sign up for the Term deposits as compare to other occupations. Around 36% of the total that signed up held university degrees. More likely, they have more awareness and know the benefit.

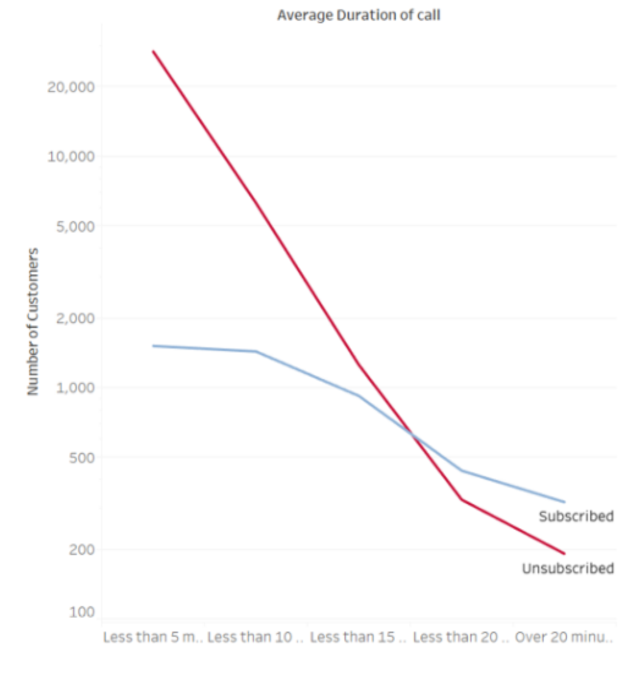


Duration Vs Term Deposit Subscription:

Does call duration has any significance?

Analysis – Longer call duration plays an important role in term deposit subscription.

Focus more on getting clients to talk for longer duration.

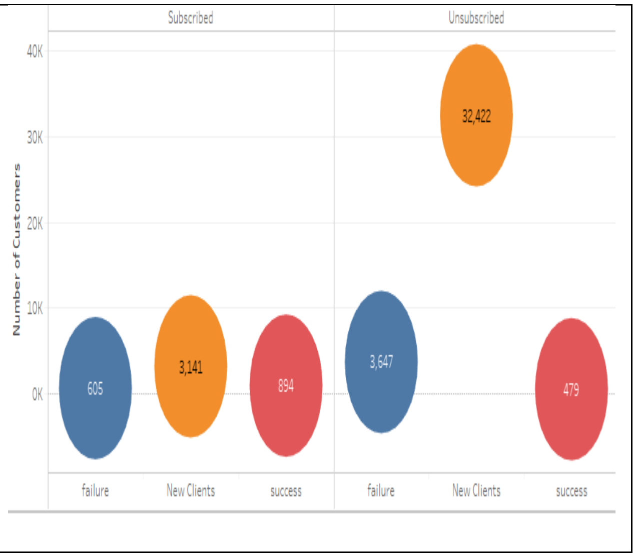


Changes from Previous Campaign to New Campaign:

Did the customer changed their decision related to term deposit?

- 605 customers who did not subscribe in the last campaign did subscribe this time, but 3647 clients still did not subscribe.

- 894 customers who did subscribe in the last campaign also subscribed this time but 479 clients who were successful last time did not sign up.

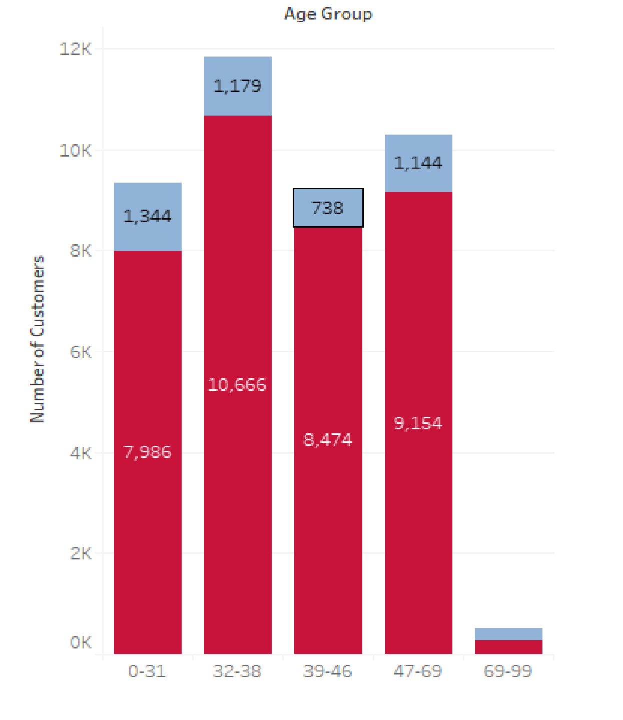


Age Group vs Term Deposit Subscription:

Does age play a role in term deposit subscription?

-Most clients are aged less than 38 should be targeted to get higher numbers of subscription. However, ages from 39-69 also showed interest.

-Less focus on age greater than 69 as it is our outlier

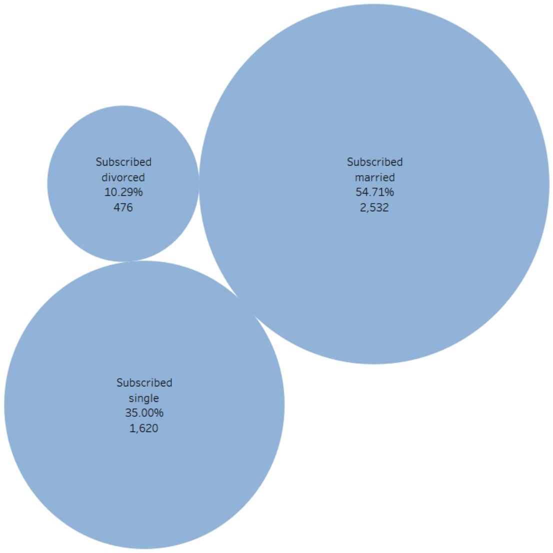


Marital vs Successful Term Deposit Subscription:

What Marital Status are more inclined towards term deposits?

The marital status plays a critical role in converting the new customers into subscribers to the term-deposit campaign.

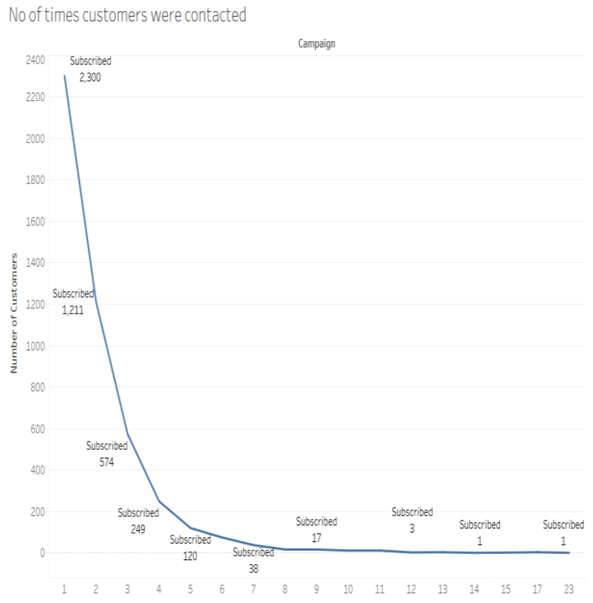
Our analysis shows that almost 55% who signed up for term deposit are married.



Number of times contacted for the term deposit campaign:

How many times should we approach the clients for a deposit?

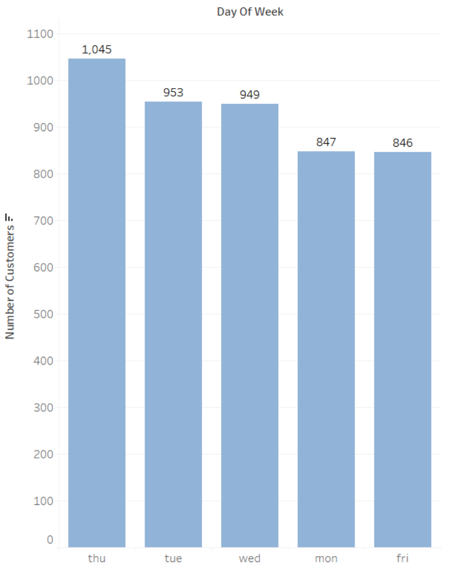
If the client is not convinced till the 12th time when the customer was contact, that customer should be dropped from the marketing list as there is not a likelihood of that customer to be a subscriber to the term deposit. The client can save the resources and budget which can be utilized to target the new potential clients. The calling agents can have a better understanding of approaching to the new potential clients.



Days of the week:

What day should we do the campaign?

Our analysis is showing that the client should run term deposit campaign on Thursday as we think that Thursday is the best day for carrying out term deposit campaign.



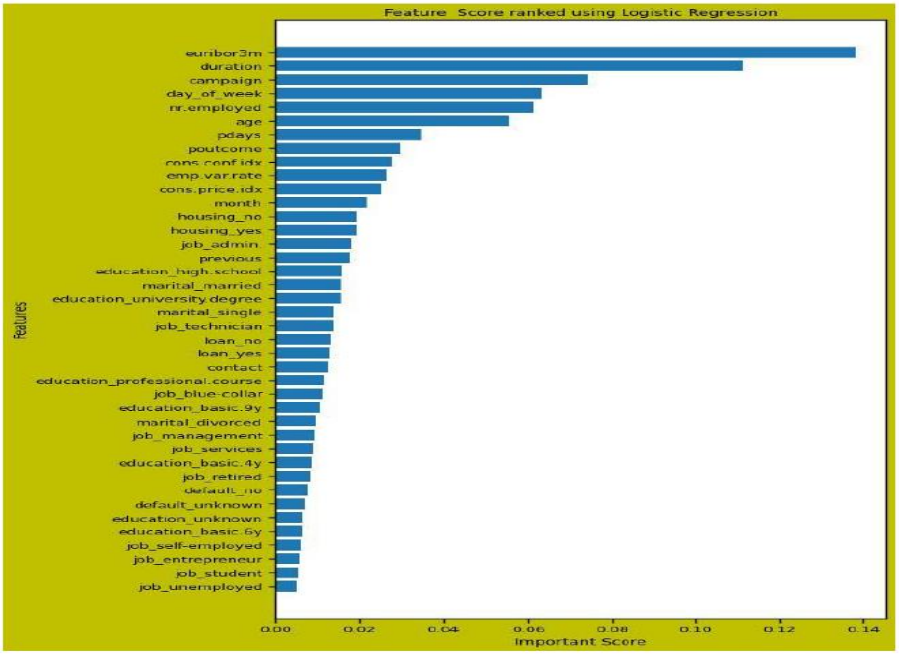
Link to the Tableau Dashboard:

https://public.tableau.com/profile/hassan6741#!/vizhome/BankTermDepositHassanGMarch16/BankInstitutionTermDepositPredictiveModel?publish=yes

Model Creation:

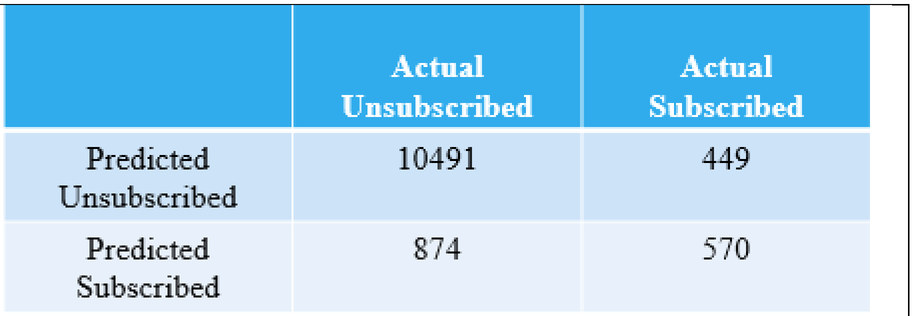
We have created machine learning models that can predict how likely clients will subscribe to a bank term deposit, along with efficient use of marketing dollars spent. The four models that we created are: 1-Logistic Regression 2- Decision Tree 3- Support Vector Machines (Linear SVC) 4- Random Forest. We have analyzed confusion matrix, ROC and AUC results of each models.

The best model that stood out is the Random Forest classifier, considering performance metrics like accuracy, ROC Curve, and AUC. Our model's test performance is ended with a 96.3 % accuracy score with the highest AUC. After going through the Random Forest theories used for the feature importance, modeling, and comparison of various performance metrics from the other relevant models, we can conclude that random forest will be the best fit model for this project. The Random Forest model builds multiple decision trees according to the features available. While using random forest, we have incorporated new features based on their significance to the output variable y. After reviewing all the features, their characteristics, correlation with response variable, we suggest our client should target the potential customers when Euribor: 3month and duration are high. It would be worth noting that unemployed and students carry less weightage towards subscribing to a term deposition as compared to other categorical features.

Euribor3m Duration Campaign Days of the week	

Random Forest Classification

We have used Random Forest Classification to get an optimized model that can reduce the cost of marketing and increase the subscriptions rate. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.[4]

Training Score: 96.3%
Precision = .95
Recall = .93
F1 Score = .96
Accuracy = 89%
	

	Actual Unsubscribed	Actual Subscribed
Predicted Unsubscribed	10491	449
Predicted Subscribed	874	570

Logistic Regression

We have created a logistic regression model to predict the clients with subscriptions and non-subscriptions to the bank term-deposit. Logistic regression works on the principle of maximum likelihood estimation by transforming the dependent variable into logit variable with respect to independent variable. [3]

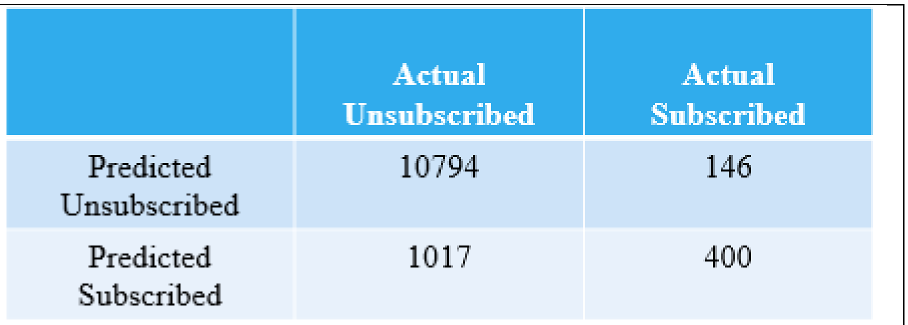
Training Score: 90.4%

Precision = .96

Recall = .91

F1 Score = .95

Accuracy = 90.6%



	Actual Unsubscribed	Actual Subscribed
Predicted Unsubscribed	10794	146
Predicted Subscribed	1017	400

Linear SVC

We have created a Linear SVC model to predict the best fitting model for gaining the best output resulting in an increased subscriptions rate. The Linear SVC classifier tries to find a line (a line here, more generally a hyperplane) that separates the True labels from the False labels.[5]

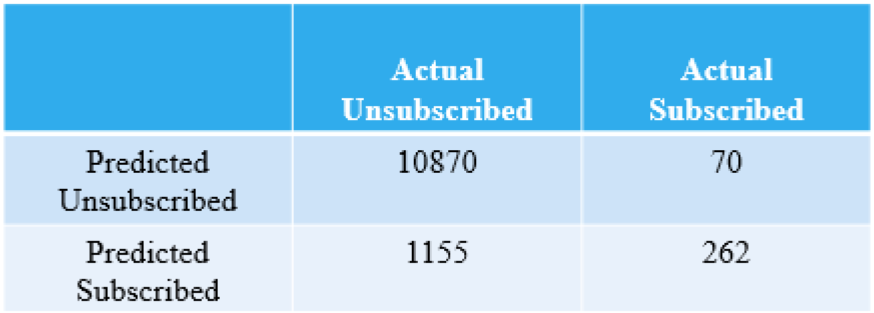
Training Score: 90.0%

Precision = .98

Recall = .90

F1 Score = .95

Accuracy = 90%



	Actual Unsubscribed	Actual Subscribed
Predicted Unsubscribed	10870	70
Predicted Subscribed	1155	262

Decision Tree

We have also developed a model based on decision tree. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).[6]

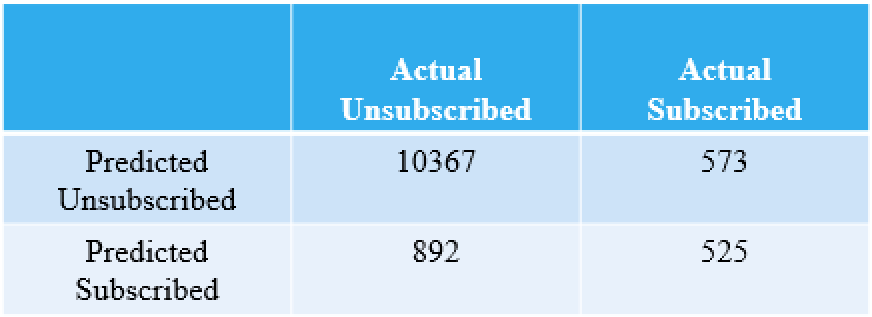
Training Score: 96.3%

Precision = .95

Recall = .92

F1 Score = .96

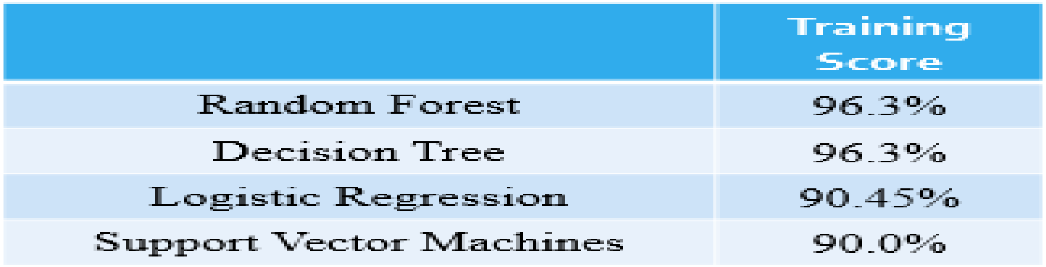
Accuracy = 88%



	Actual Unsubscribed	Actual Subscribed
Predicted Unsubscribed	10367	573
Predicted Subscribed	892	525

Model Comparison:

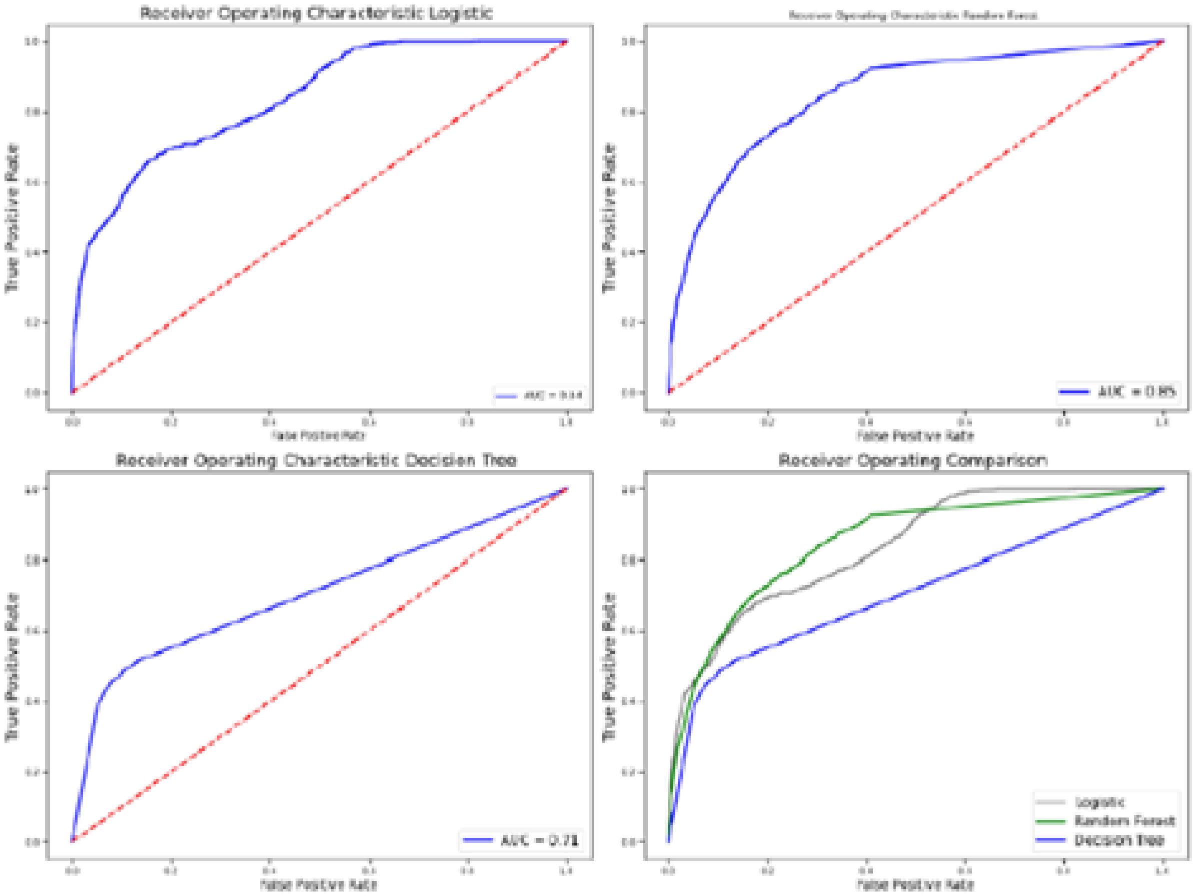
We have compared the models we built by using the given data from our client to find out the best model that can yield more revenue and reduce the cost of marketing campaigns beside highlighting the factors that can improve the efficiency of the call agents, marketing strategies, budget allocations, etc. We compared the four models (Random Forest, Logistic Regression, Decision Tree, and Linear SVC) based on the training score which are 96.3%, 90.4%,96.3% and 90.0% respectively. We picked the Random Forest classifier, considering performance metrics like accuracy, ROC Curve, and AUC.



	Training Score
Random Forest	96.3%
Decision Tree	96.3%
Logistic Regression	90.45%
Support Vector Machines	90.0%

Best Fit Model Selection based on ROC-AUC Curve:

Following is the visual presentation of the best fit model based on ROC-AUC curves.



Recommendations:

- We highly recommend that banks should talk to the customers for longer time in order to clarify customers queries and so to get a term deposit subscription. The bank can apply natural language processing (NLP) tasks to have a better understanding of the communications held between the clients and the call center agents. The NLP machine learning techniques can help the back understand what the clients expect and how convincing or persuasive the call center agents are during the whole interaction.

- When Euribor3m is high, the bank should put more efforts as customers will earn higher returns and will be highly motivated to sign up for a term deposit.
- Keep on updating all the model for better predictions to keep fresh and current.

Conclusions:

Team Alpha concludes that several features played a critical role in term deposit subscriptions.

- Many features have played a role in term deposit subscription.
- Although Euribor3m should influence the customers to purchase term deposit but the data shows Euribor3m does not excite the customers.
- To make the campaign more successful, the client should add more features in the data sets like competitors rate offerings, economic conditions (recession, boom).
- After plotting the ROC and AUC Curve, we could conclude that Random Forest is the best model among all three models which we have chosen due to maximum area under the curve.

References/ Appendices:

1. Duboue, Pablo, *The Art of Feature Engineering, Essentials for Machine Learning*, Cambridge University Press, 2020.
2. Duboue, Pablo, *The Art of Feature Engineering, Essentials for Machine Learning*, Cambridge University Press, 2020.
3. Dangeti, Pratap, *Statistics for Machine Learning*, Packt Publishing, 2017.
4. Surhone Lambert M., et al., Random Forest, VDM Publishing, 2010.
5. Rossant, Cyrille, *IPython Cookbook*, Packt Publishing, 2014.
6. Chauhan, Nagesh Sign, *Decision Tree Algorithm, Explained*, KDnuggets, https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. Accessed 12 April 2020.