FINAL REPORT

PROJECT TITLE: Predicting Bank Marketing Campaign Success Using Machine Learning

PROBLEM STATEMENT

A Term Deposit is a deposit held at a financial institution that has a fixed term. These are generally short-term with maturities ranging anywhere from a month to a few years. When a term deposit is purchased, the customer understands that the money can only be withdrawn after the term has ended or by giving a predetermined number of days. Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk investors. Instead of mass marketing, the bank has chosen to be more proactive in identifying potential buyers and communicate straight to the customer via telephone calls. Direct marketing is useful here because its positive results can be measured directly.

The goal of this project is to perform post-campaign analytics to identify the potential subscribers of the term deposit product for future campaigns. The data mining task is to create a classification model identifying potential subscribers by using supervised learning algorithms.

DATA DESCRIPTION

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, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed**.** It is publicly available in the UCI Machine learning Repository, which can be retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

We used Python 3 and JupyterNotebook to upload and analyze the data. The Data is consisted of 41,188 customers on direct marketing campaigns (phone calls) of a Portuguese banking institution, with 21 variables. 11 of these variables such as job, marital status, and education are categorical. We will use categorical variable ‘y’ as a target output. Other variables are numerical variables such as number of employees, Euribor rate, and consumer price index. The below tables show the all variables, their description and types:

Categorical Variables:

Column Name	Description
job	Client's occupation
marital	Marital status
education	Client's education level
default	Indicates whether the client has credit in default
housing	Indicates whether the client has a housing loan
loan	Indicates whether the client has a personal loan
contact	Type of contact communication
month	Month that last contact was made
day of week	Day that last contact was made
poutcome	Outcome of the previous marketing campaign
y	Indicates whether the client has subscribed for a term deposit

Numerical Variables:

Column Name	Description
age	Age of the Client
duration	Duration of last contact in seconds
campaign	Number of contacts performed during this campaign for this client (including last contact)
pdays	Number of days since the client was last contacted in a previous campaign
previous	Number of contacts performed before this campaign for this client
empvarrate	Employment variation rate (quarterly indicator)
conspriceidx	Consumer price index (monthly indicator)
consconfidx	Consumer confidence index (monthly indicator)
euribor3m	Euribor 3-month rate (daily indicator)
nremployed	Number of employees (quarterly indicator)

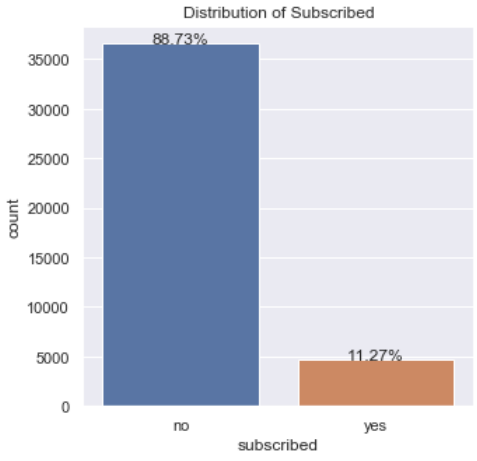
EXPLORATORY DATA ANALYSIS

To create our machine learning model, we need to understand which variable is target variable and which variables are input variables that target variable are dependent of these variables. We analyzed output and input variables statistically, visualized their distributions and found relationships between each other.

Target Variable

We changed the column name ‘y’ to ‘subscribed’ to understand the meaning of target variable. From the below table, we can say that %88.73 clients has subscribed for a term deposit. Other %11.27 of the clients did not accept a term deposit.

EXPLORATORY DATA ANALYSIS



Input Variables

We built box plots and histograms for each numerical column to understand their distributions and detect outliers. We also summarized important statistics all variables such as mean, standard deviation, minimum and maximum values. Then, for numerical variables, we made hypothesis test to compare two independent means of ‘yes’ and ‘no’ for each numerical variable in order to decide whether the variables are statistically significant for the output variable. The significance

level is 0.05 for the tests. We found p value for each variable hypothesis tests. The hypothesis tests we used is:

$$H_o: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

We compared our p-value to a stated level of significance.

- If the P-value ≤ 0.05, we reject the null hypothesis in favor of the alternative hypothesis.
- If the P-value > 0.05, we fail to reject the null hypothesis. We do not have enough evidence to support the alternative hypothesis.

A Chi-square test is designed to analyze the categorical variables. We created contingency table for each categorical variable. That means that the data has been counted and divided into categories in terms of subscribing yes or no. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent. The chi square statistic is commonly used for testing relationships between categorical variables.

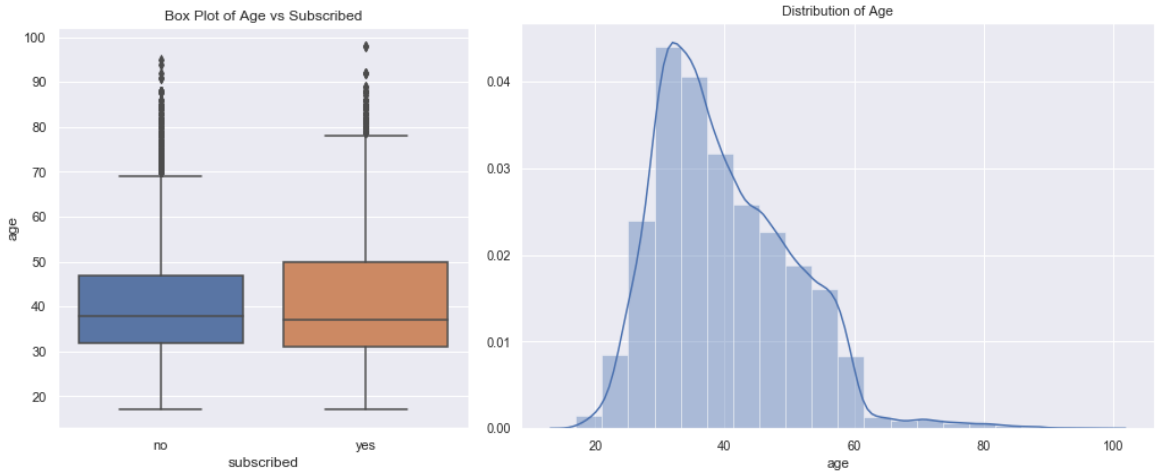
H_o: No relationship between variables

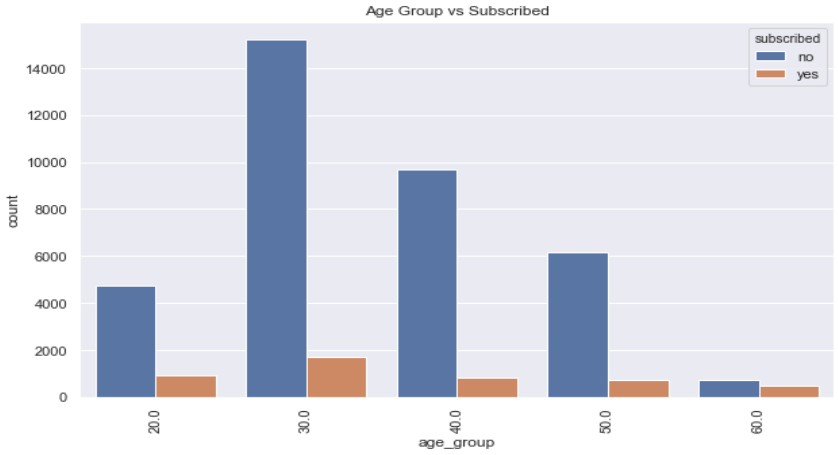
H_a: There is a relationship between variables

The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables; they are independent. An example of a research question would be: Is there a significant relationship between job categories in terms of subscribed yes and no? Again, we compared their p values while the significance level is 0.05. If p values are less than the significance level, we reject the null hypothesis.

Features Related with Clients Information

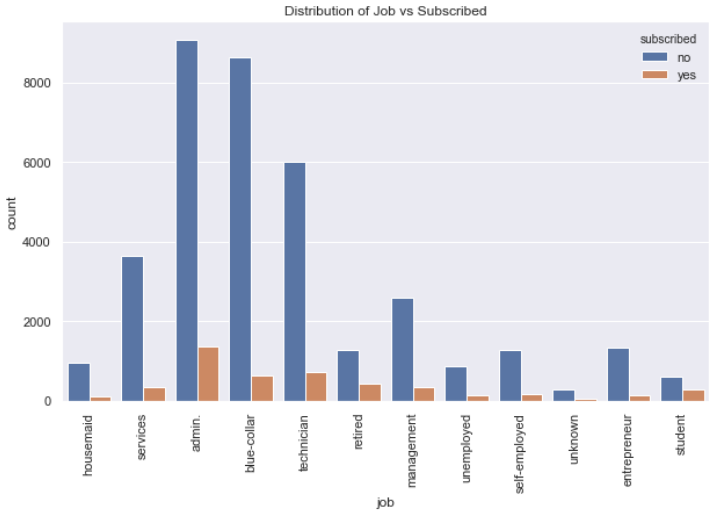
Age:





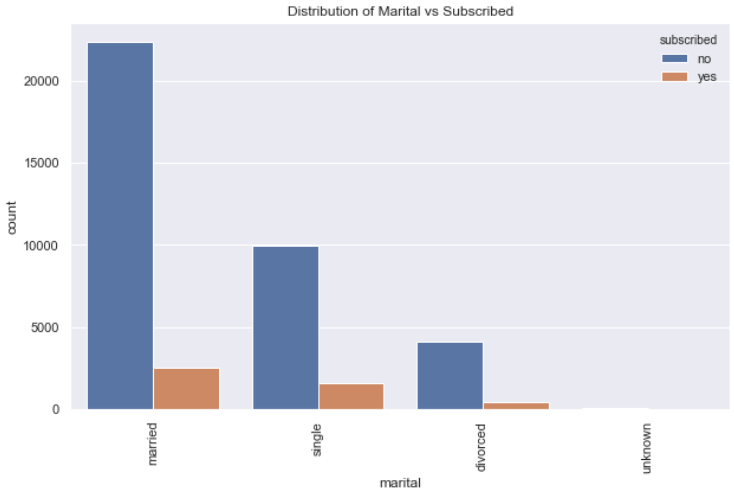
There are outliers in age column which are greater than age 70. The %50 of the age values are distributed between age 32 and 47. There is a difference between means of age subscribing 'yes' and 'no'. The age variable is statistically significant at 95% confidence interval in terms of subscribing term deposits. To understand which ages group has more likelihood to subscribe term deposit, we created age groups column. We renamed age values as age groups. For example, if the ages are less than 30, we named it 20 and ages between 30 and 40, we named it 30. As a result, we created a histogram of age groups versus subscribed. As a result, ages of more than 60 has higher success rate to subscribe term deposit. Ages less than 30 comes second to positive subscribed. Ages between 30 and 40 has the less success rate comparing with other ages group.

Job:



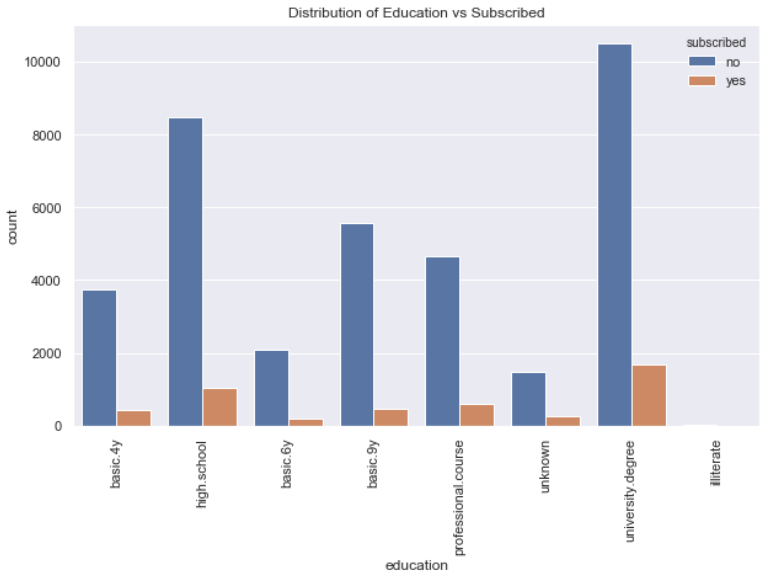
There are unknown values in the job column, but the number of unknowns is not significant. We can drop these missing values. Administrative, blue-collar and technician jobs are observed more than other type of jobs in the data. The success rate for subscribing term deposit is higher if the clients are student or retired. On the opposite side, clients in blue-collar, entrepreneur or services jobs tend to not subscribe term deposit comparing to other clients. After applying t test for job types regarding with the subscription, p value is smaller than 0.05 so the job variable is statistically significant at 95% confidence interval.

Marital Status:



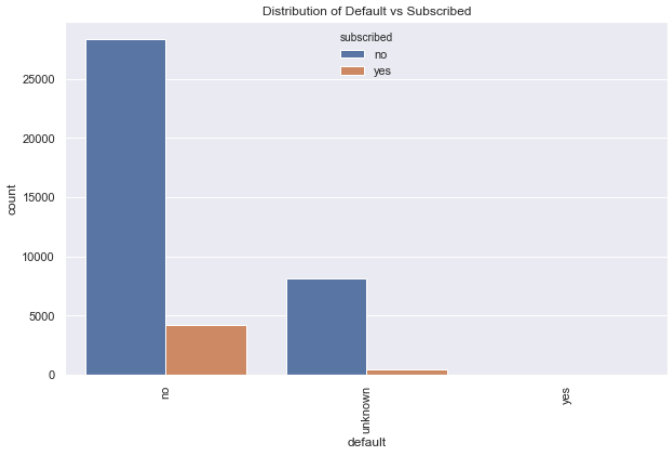
Most of the clients in data are married. There are a few unknown values in marital status. We can ignore and delete these missing values. The single clients' success rate is higher than other marital status to response positive to subscribe the term deposit. The marital variable is statistically significant at 95% confidence interval because p value is less than 0.05 in t test.

Education:

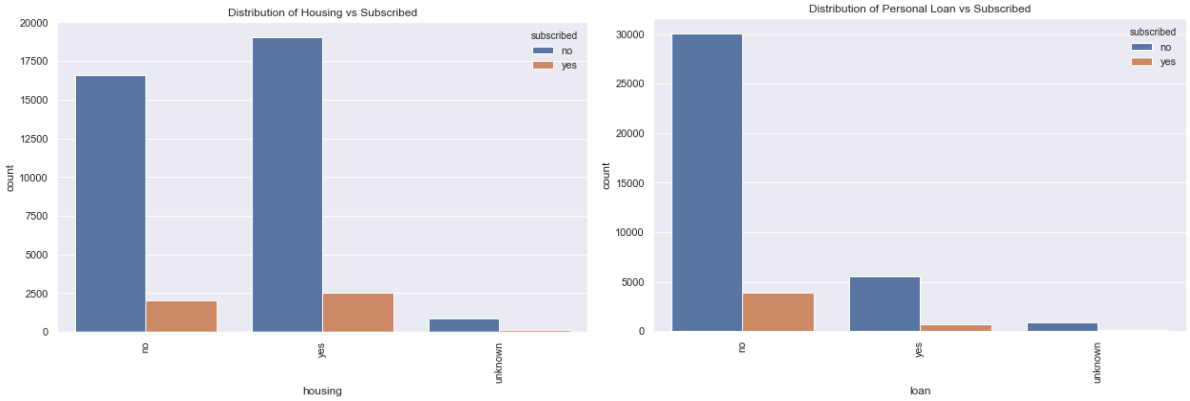


Most of the clients in the data have university of high school degree. There are unknown education levels of the clients. The illiterate clients have more positive rate to subscribe term deposit comparing to other education levels. Then, clients who have university degree tend to have higher success rate. On the other hand, the clients in basic education levels have less success rate. The p value of the test is less than 0.05 so the education variable is statistically significant at 95% confidence interval.

Having Credit in Default, Housing Loan and Personal Loan:



There are only 3 clients having credit in default. Other %20 of the clients do not known. The default variable will be dropped because of not having meaningful values.

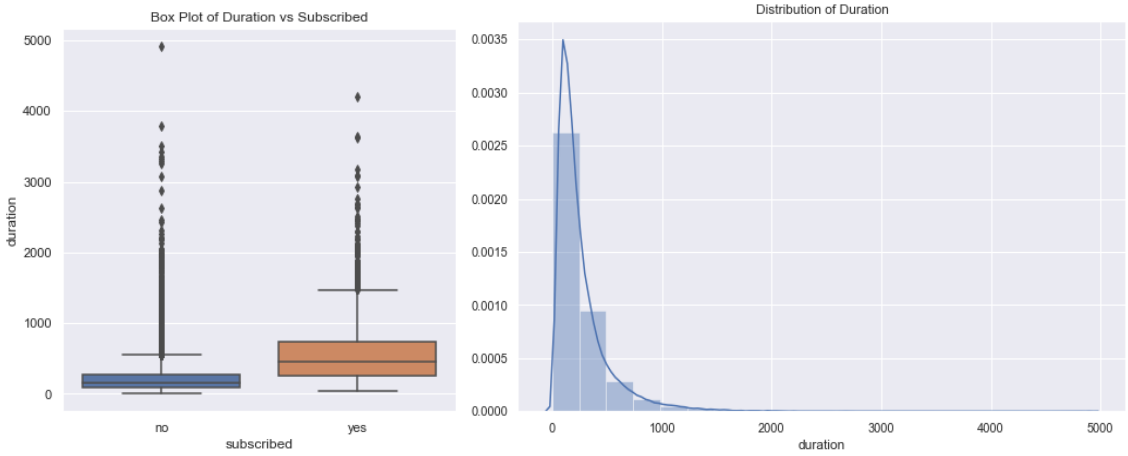


There a few unknown values for both housing and personal loan variables. When we analyze the hypothesis tests of these two variables regarding with subscribing the term deposit, p values for the tests are greater than 0.05. Thus, the clients having personal or housing loan have not meaningful difference from other clients not having these loans. We can also drop these variables because of not having effect to the target variable.

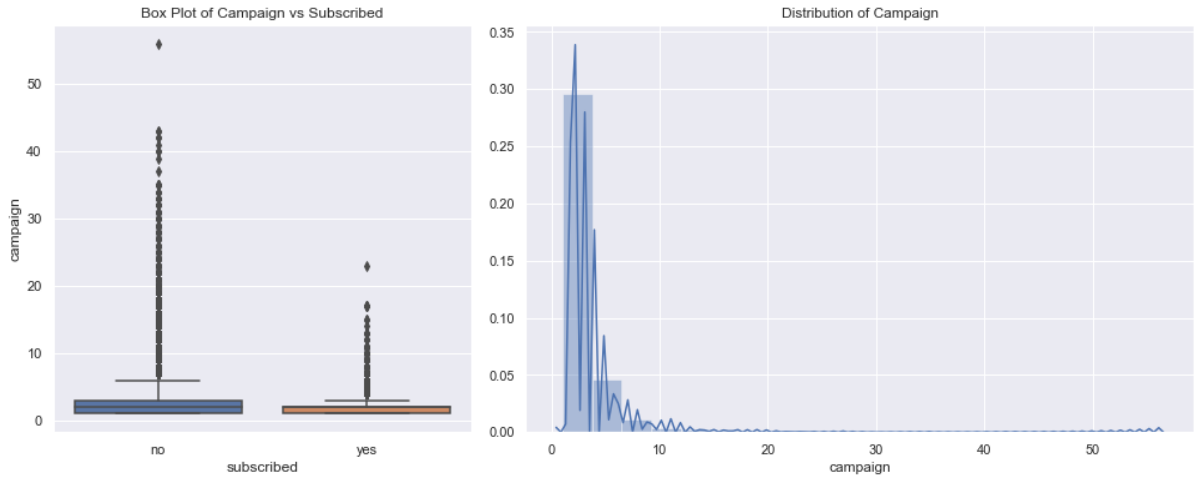
Features Related with Campaign

Duration:

The duration of last contact in seconds variable highly affects the target variable. However, the duration is not known before a call is performed. After the end of the call y is obviously known. Thus, this input variable should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. As a result, the duration variable will be dropped.

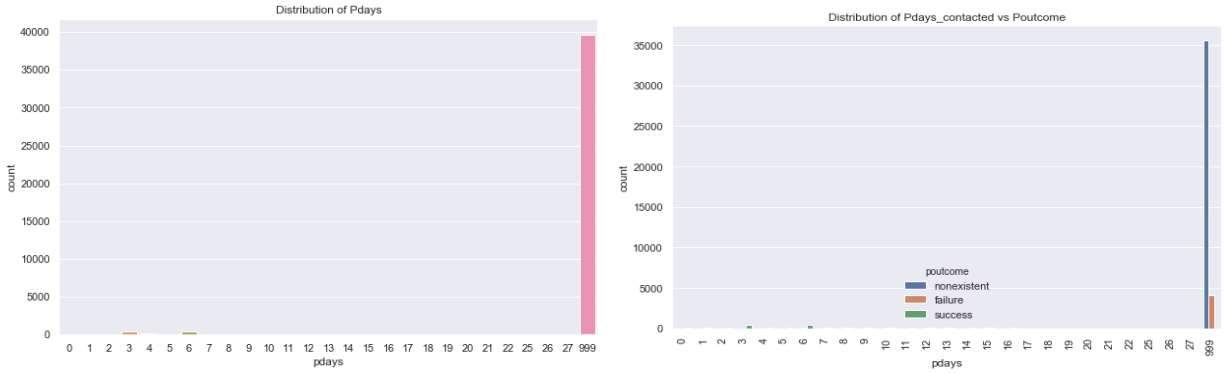


Campaign:



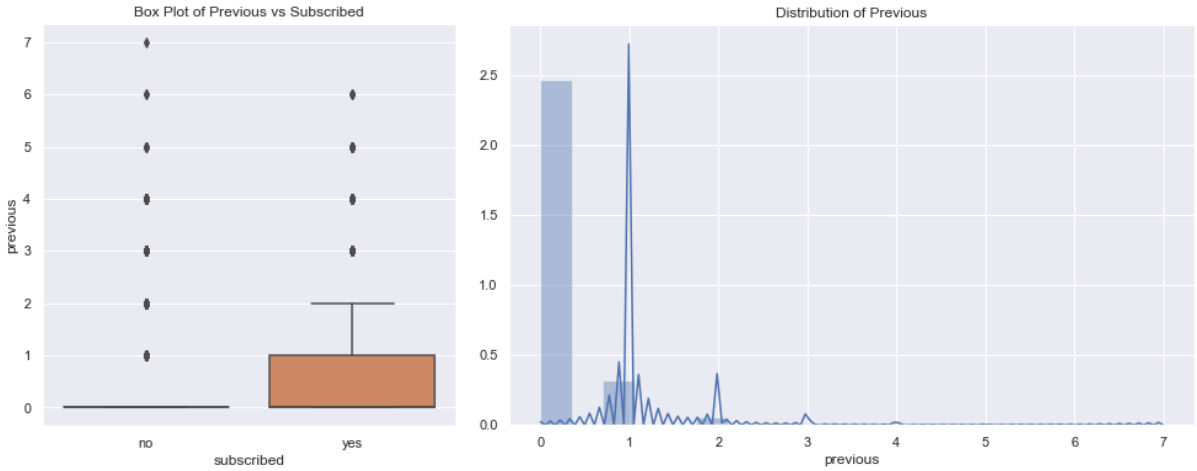
Number of contacts performed during the campaign for the client variable has outliers. The outliers are almost greater than 20 number of contacts. %50 of the variable consist of 1 and 2 number of contacts. There is a difference between means of campaign subscribing 'yes' and 'no'. The campaign variable is statistically significant at 95% confidence interval.

Pdays:



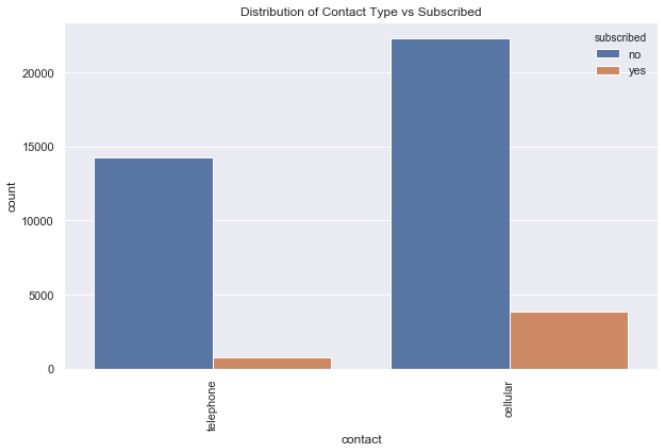
The boxplots of pdays variable (number of days since the client last contacted in a previous campaign) looks very strange so we investigated their values in detailed. We can see majority of the values in pdays are 999 and its stated in the variable description that 999 means clients were not previously contacted. When we analyze other variables, we see that poutcome variable also has the same meaning value 'nonexistent'. All nonexistent values match with 999. Thus, we will drop this column in our model.

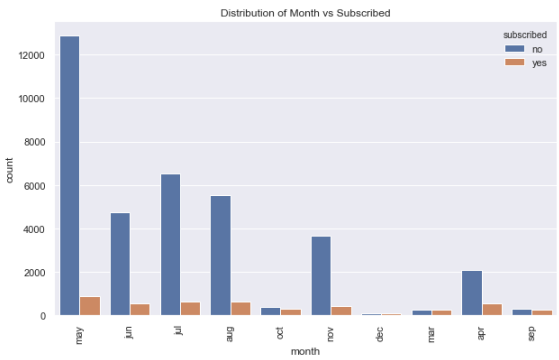
Previous:

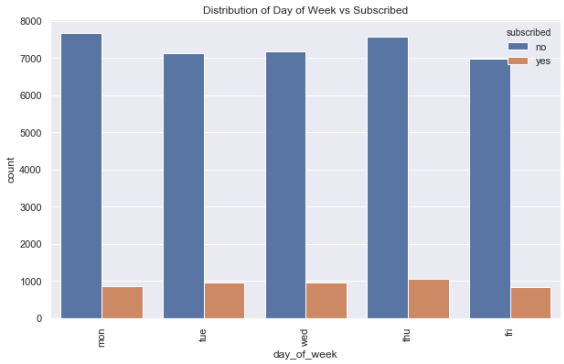


There are outliers in previous variable (number of contacts performed before this campaign). Almost more than 3 number of contacts can be outlier. The major number of contacts is zero. It means that most of the clients was not contacted before this campaign. The p value of the test is less than zero, so we rejected the hypothesis. It means that there is a difference between the means of previous variable in terms of subscribing term credit. As a result, precious variable is statistically significant at 0.95 confidence level.

Contact Type, Month, Day of Week:

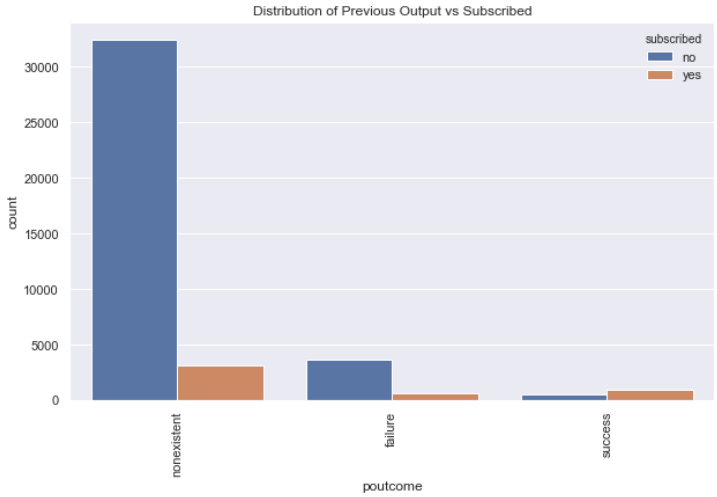






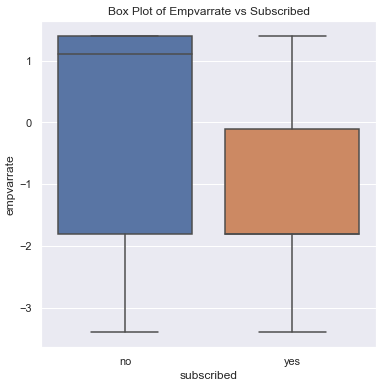
The rate of responding the campaign positively is higher for the clients contacted via cellular. Also, most of the clients were reached via cellular. The p value of the t test is less than 0.05. The contact type variable is statistically significant at 95% confidence interval so we will use this variable in our model. The rate of response 'yes' is higher for the months of March, December, October and September. The months May and July have negative affect on subscribing term deposit. The days of Monday and Friday have more 'yes' response to subscribe term deposit comparing to other days of the week. The month and day of week variables are also statistically significant at 95% confidence interval.

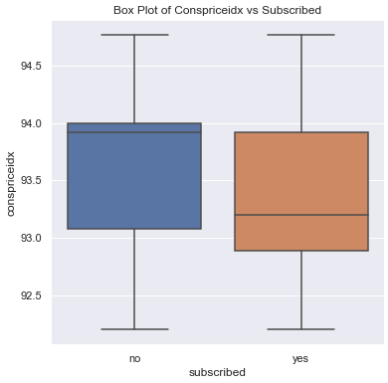
Previous Output:

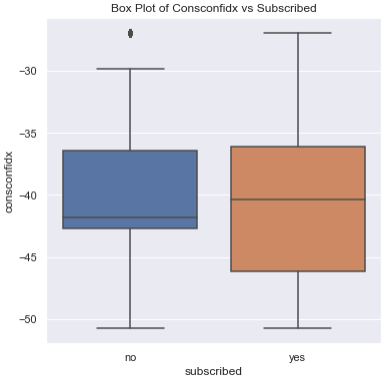


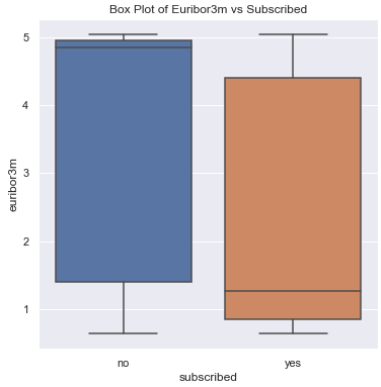
If the output of previous campaign is success, the probability of saying 'yes' to subscribing new campaign is higher. The clients not contacted for previous marketing campaign tend to say 'no' for new campaign comparing to other clients. In addition to that, most of the clients in this campaign did not contacted in the previous campaign. The previous outcome variable is statistically significant at 95% confidence interval.

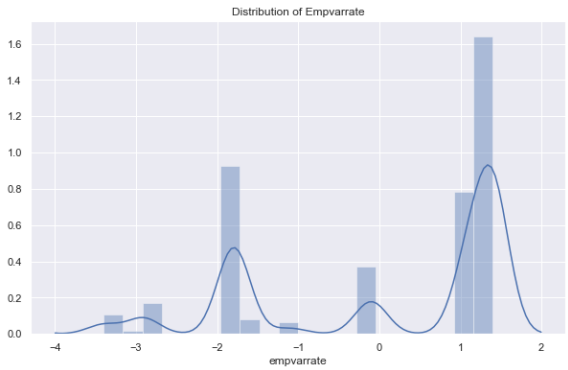
Features Related with Social and Economic Context

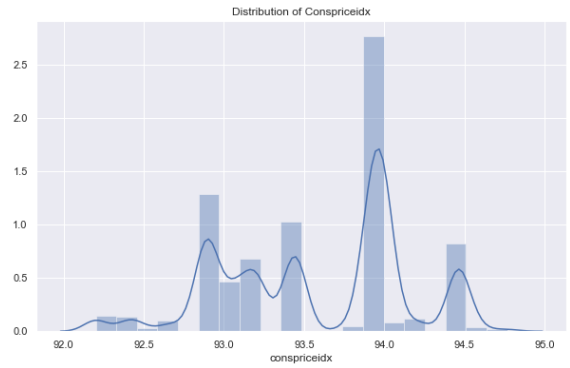


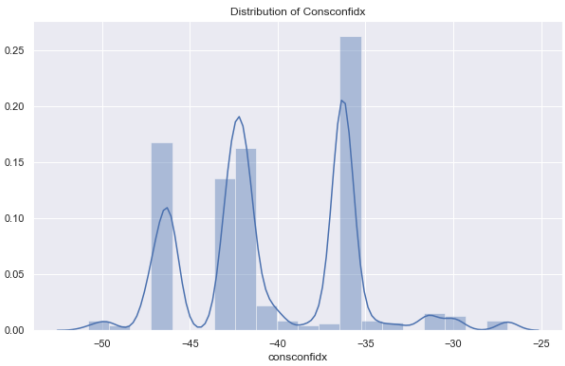


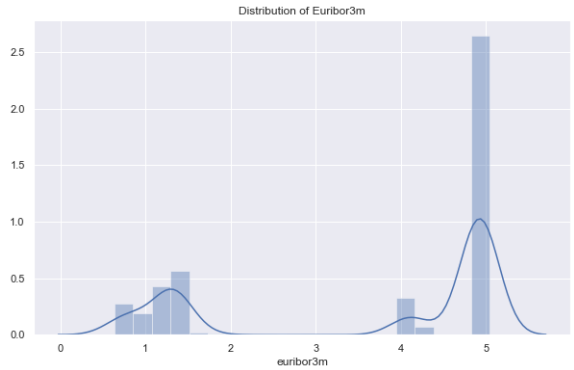


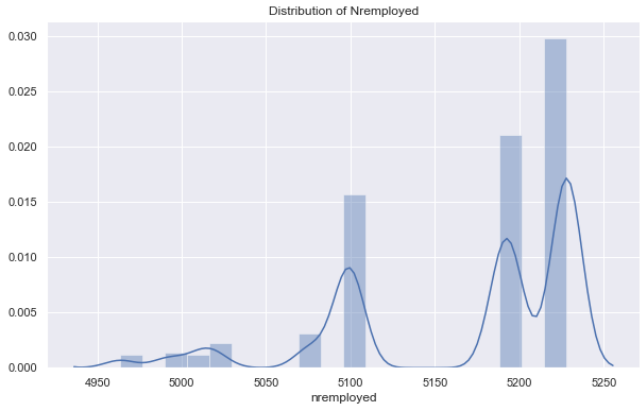










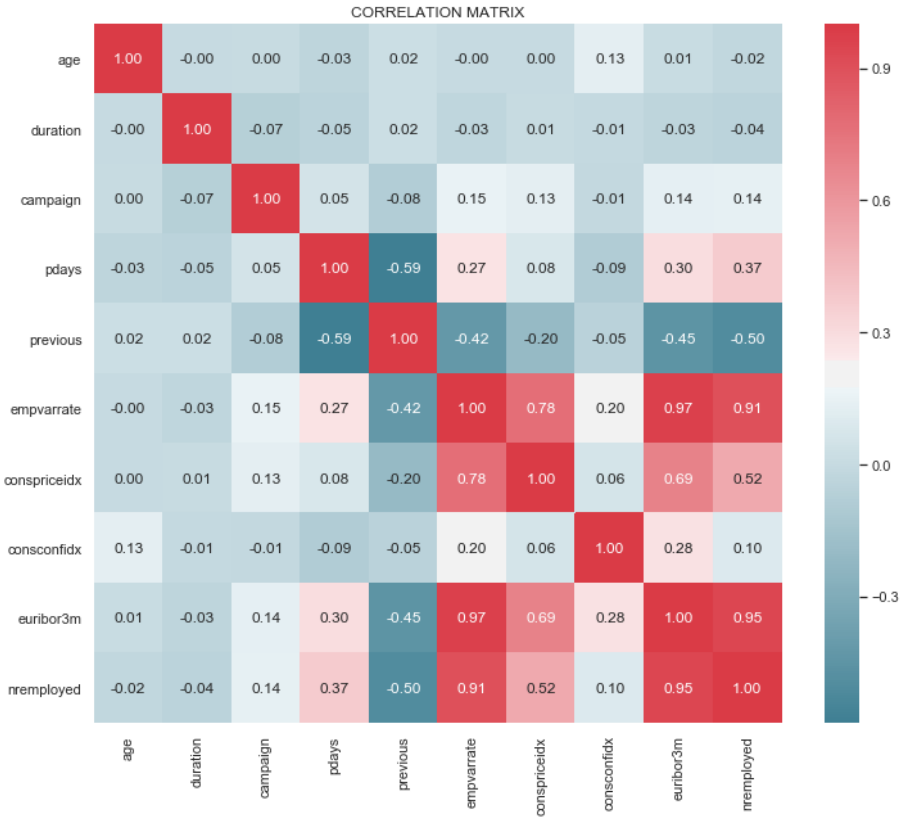


The variables of emp.var.rate (employment variation rate), cons.price.idx (consumer price index), cons.conf.idx (consumer confidence index), euribor3m (euribor 3 month rate) and nr.employer (number of employees) are social and economic context features in the data. We built box plots and distribution plots of these variables. These features do not have outliers. There are negative and positive employment variance rates. The mean of the employment variation rate is 0.08. The consumer price index changes between the values 92.2 and 94.76. The mean of this index is 93.57. The consumer confidence index changes between the values -50.8 and -26.9. The mean of the index is -40.5. All values in this index are negative. The Euribor rate changes between the values 0.63 and 5.04. The mean of this variable is 3.6. The number of employees changes between the values 4963 and 5228. The mean of the number of employees is 5167. We built hypothesis tests whether these variables mean are different in terms of subscribing term deposit. After calculating their t tests, p values of all these social and economic features are smaller than zero. As a result, we can say that these variables are statistically significant at 95% confidence interval.

Correlation Between Numerical Features

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases. In addition, a correlation matrix is a table showing correlation coefficients between sets of variables. Each numerical variable in the table is correlated with each of the other values in the table. This allows us to see which pairs have the highest correlation.

We built the correlation matrix for the numerical variables to see their relationship. From the below table, we can say that number of employees, employee's variance rate and euribor rate are very high positive correlated with each other.



DATA PREPROCESSING

Unnecessary columns, missing values and outliers

We will drop a couple of columns that are not statistically significant for our output. In the exploratory data analysis part, we discovered the unnecessary columns from data. After the inference from our hypothesis test for each variable, we found that 'duration', 'pdays', 'default', ' 'housing', 'loan' variables are not meaningful for our model. Also, we will not use ‘age_group’ column that we created to analyze ages in detail. Thus, we removed these columns from the data.

In our exploratory data analysis part, we discovered that there are unknown values in some variables. After deleting unnecessary columns, only the variables 'job', 'marital' and 'education' have unknown values. The percentage of unknown values for each column is very low so deleting these missing values does not affect our model very much. We also dropped the missing values. Currently, the data has 39,191 rows.

After cleaning data by dropping unnecessary columns and missing values, only 'age' and ' campaign' features have outliers. Age variable has 15 and campaign variable has 2 outliers. The clients' ages change between 17 and 98. There is no anormal age in the data. Also, the number of contacts performed during the campaign change between 1 and 56. These values are also very normal in real model. Thus, we will not remove outliers to create a realistic model.

Feature Scaling and Train-Test Split

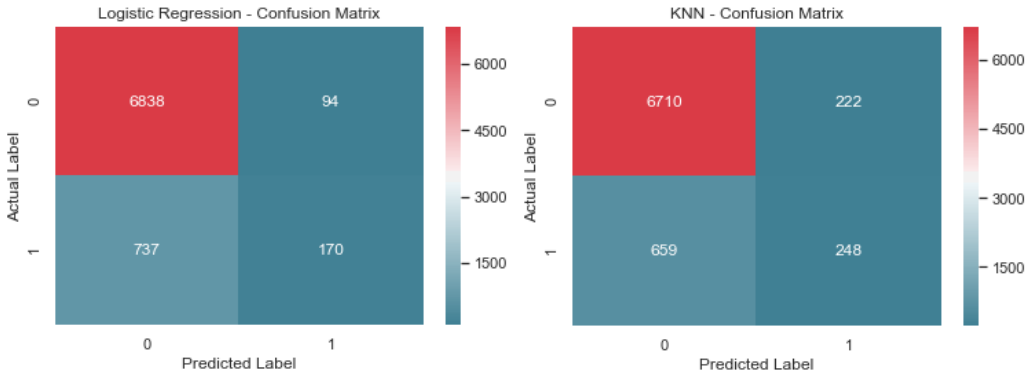
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. We used standardization to scale the variables. Standardization is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1. For each value of variables, we calculated their new values using Python Scikit-Learn library.

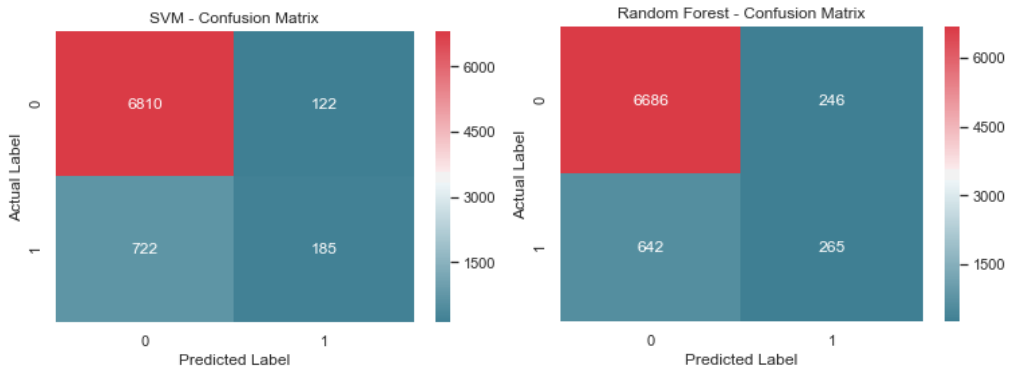
After scaling the data, we built two different data sets as training and test. To evaluate the performance of a machine learning algorithm is to use different training and testing datasets. We took our original dataset, split it into two parts. We will train the algorithm on the first part, make predictions on the second part and evaluate the predictions against the expected results. The size of the split is 80% of the data for training and the remaining 20% for testing.

MACHINE LEARNING

We can speed up the fitting of a machine learning algorithm by changing the optimization algorithm. A more common way of speeding up a machine learning algorithm is by using Principal Component Analysis (PCA). If learning algorithms are too slow because the input dimension is too high, then using PCA to speed it up can be a reasonable choice. Thus, we use PCA to reduce the number of variables. In order to protect %90 of variance (information), we chose first 29 variables. We don't want to lose %90 variance of the data.

Classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. We applied the classification algorithms, logistic regression, k nearest neighbors, support vector machine and random forest, respectively. Then, we calculated their performance metrics such as accuracy score and auc score. We created confusion matrix and ROC curves for each algoritms. AUC - ROC curve is a performance measurement for classification problem at various thresholds settings.



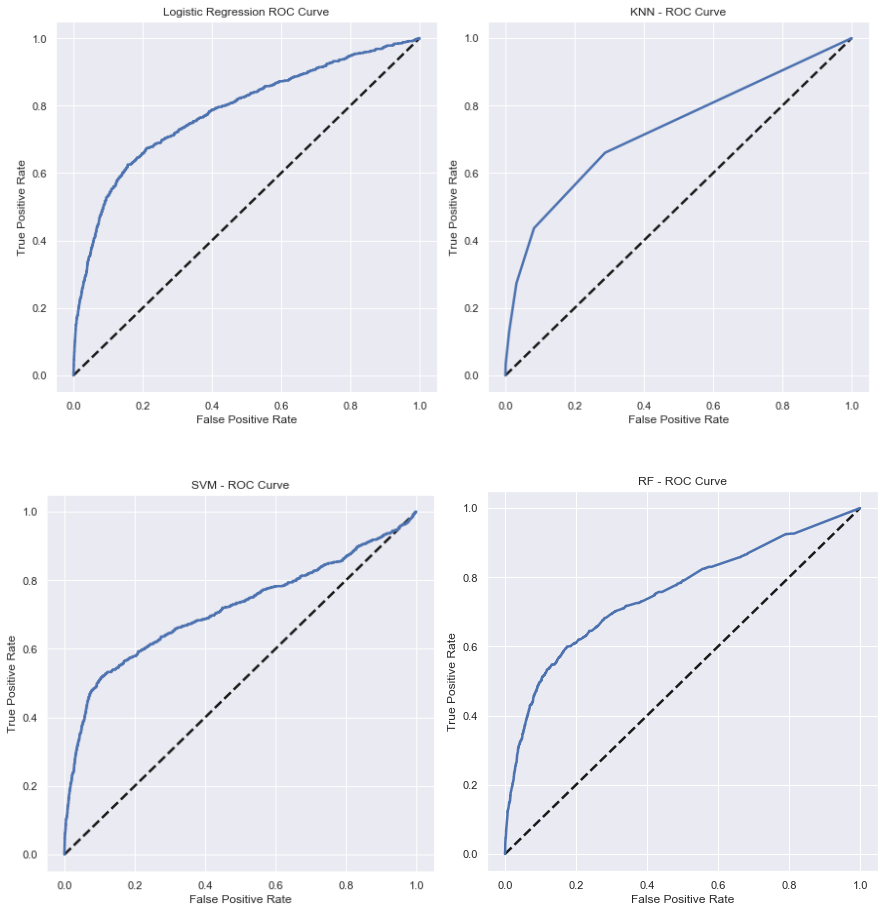


ROC Curve

A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, True Positive Rate and False Positive Rate.

$$TPR=TP/ (TP+FN)$$
$$FPR=FP/ (FP+TN)$$

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figures show typical ROC curves of each models.



AUC

AUC stands for "Area under the ROC Curve." AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. As a result, we compared the AUC scores of the models, Logistic regression has the higher AUC score. This model is the best for our classification problem.

Model	AUC Score
Logistic Regression	0.782
KNN	0.727
SVM	0.715
Random Forest	0.752

CONCLUSION

The main objective of this project is to increase the effectiveness of the bank's telemarketing campaign, which was successfully met through data analysis, visualization and analytical model building. A target customer profile was established while classification models were built to predict customers' response to the term deposit campaign.

According to previous analysis, a target customer profile can be established. The most responsive customers possess following features. Age of customers should be greater than 60 and less than 30. The job types should be student and retired. The clients should have university degree and single.

The classification and estimation model were successfully built by applying logistic regression algorithm. The bank will be able to predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits and call less to those who are unlikely to make term deposits. Thus, it will increase the efficiency of the bank’s telemarketing campaign, saving time and efforts. On the other hand, it prevents some clients from receiving undesirable advertisements and increases customer satisfaction.