Report on Customer Churn

This assignment aims to develop a machine-learning model to predict customer churn for a Portuguese banking institution. The data analysis aims to predict whether or not a customer will subscribe to a term deposit.

Data Analysis

The data consists of 4521 observations and 17 attributes. Out of the 17 features, 10 of them were observed to be categorical variables and 7 to be numerical variables. The below table shows the data frame of our dataset.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no

Features description of the Bank Marketing Data set

Description of Numerical variables

	age	balance	day	duration	campaign	pdays	previous
count	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000
mean	41.170095	1422.657819	15.915284	263.961292	2.793630	39.766645	0.542579
std	10.576211	3009.638142	8.247667	259.856633	3.109807	100.121124	1.693562
min	19.000000	-3313.000000	1.000000	4.000000	1.000000	-1.000000	0.000000
25%	33.000000	69.000000	9.000000	104.000000	1.000000	-1.000000	0.000000
50%	39.000000	444.000000	16.000000	185.000000	2.000000	-1.000000	0.000000
75%	49.000000	1480.000000	21.000000	329.000000	3.000000	-1.000000	0.000000
max	87.000000	71188.000000	31.000000	3025.000000	50.000000	871.000000	25.000000

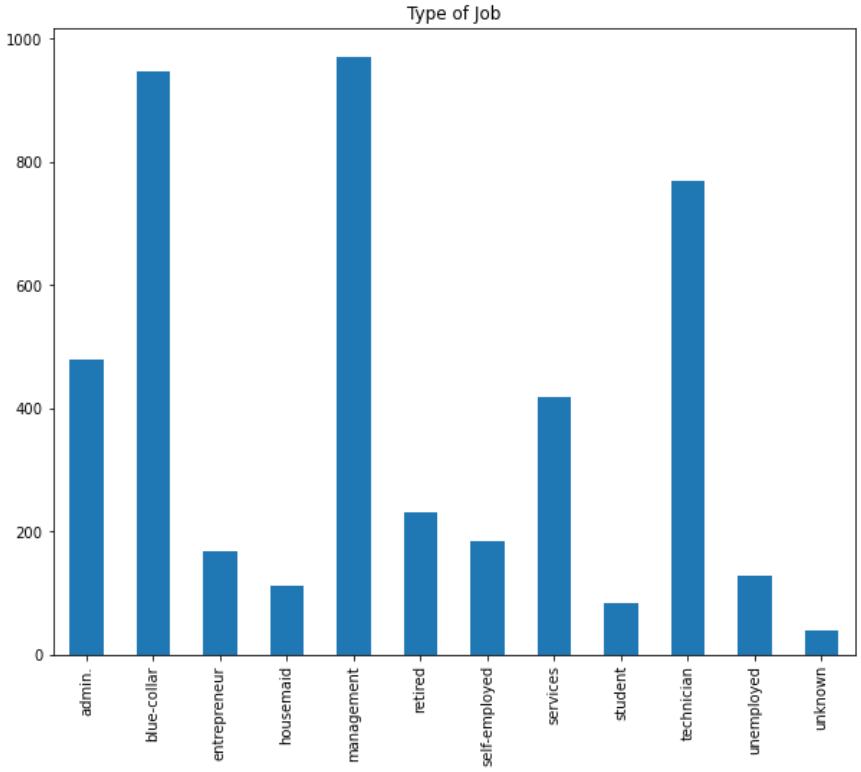
Age, balance, day, duration, campaign, pdays and precious are the numerical features in our dataset. A brief description of the count, mean, std. Deviation etc…has been provided in the above table.

Categorical features include job, marital, education, default, housing, loan, contact, month, poutcome and y.
No missing values were found in the data nor were duplicates.

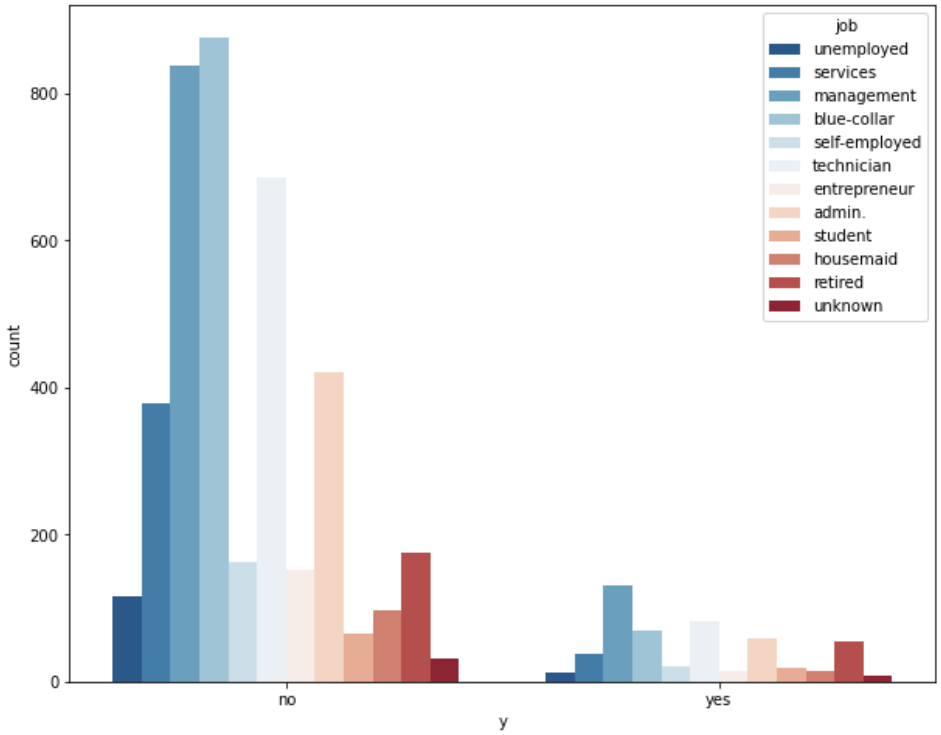
Correlation among the numerical variables was found and not much correlation could be found among the variables. A heatmap of the correlation between the variables was plotted.

Exploratory data analysis was done to know the trends in the dataset.

Among the jobs, management jobs were found to be the most. From the below bar chart, 11 categories of jobs were observed. In the later analysis, we will treat 'unknown' variables as missing values.

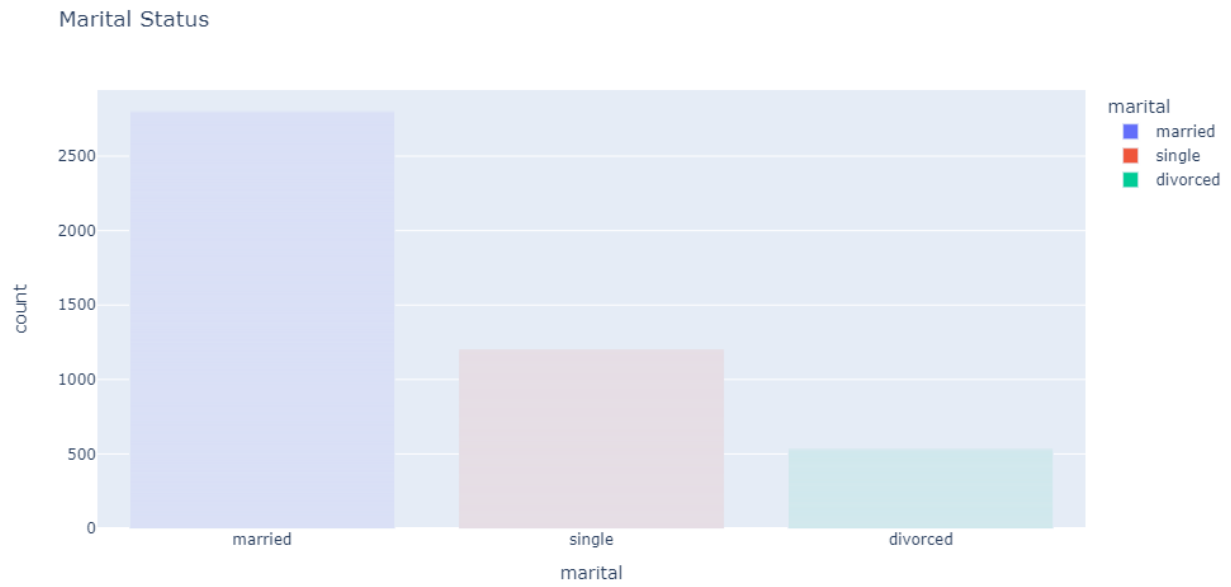


The below table shows the type of employees who has and hasn't subscribed for a term deposit.

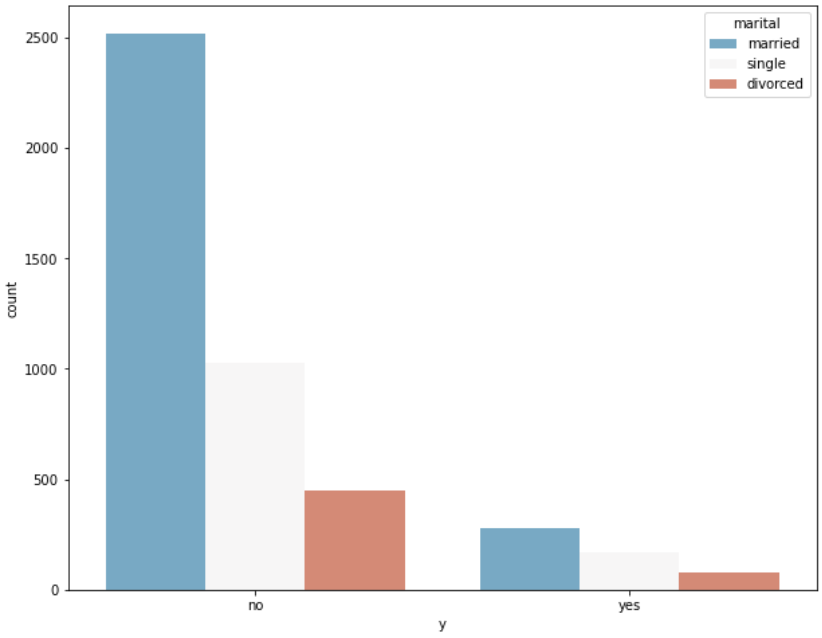


Overall, the clients haven't subscribed to the term deposit and most of them were blue-collar job clients.
Clients with management jobs subscribed the most.

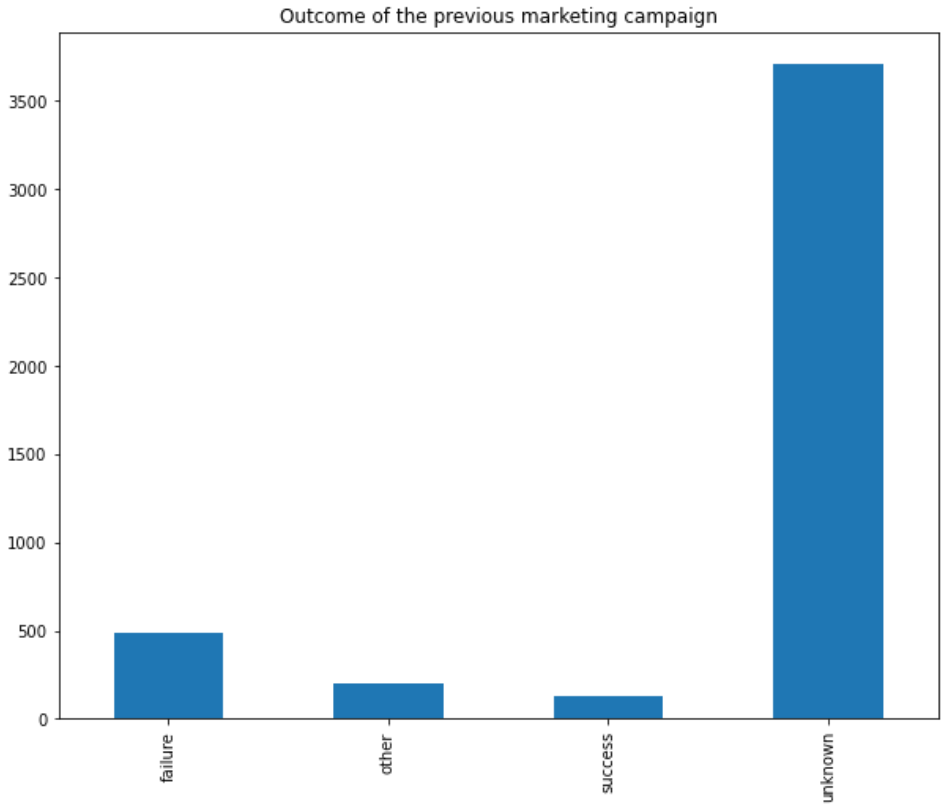
When focusing on marital status which is categorical in kind and has a range of the following values: divorced, married and single. Married ones were found the most.



Subscription, based on marital status has been plotted below.



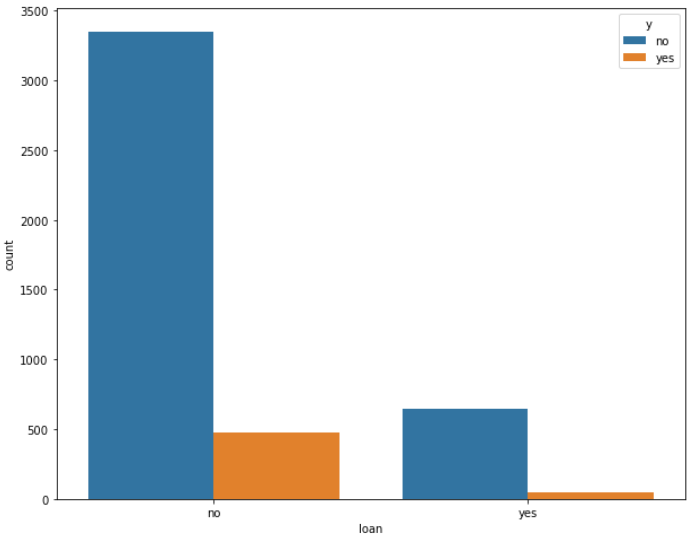
Outcome of the previous marketing campaign was plotted to know the success and failures of the campaign.



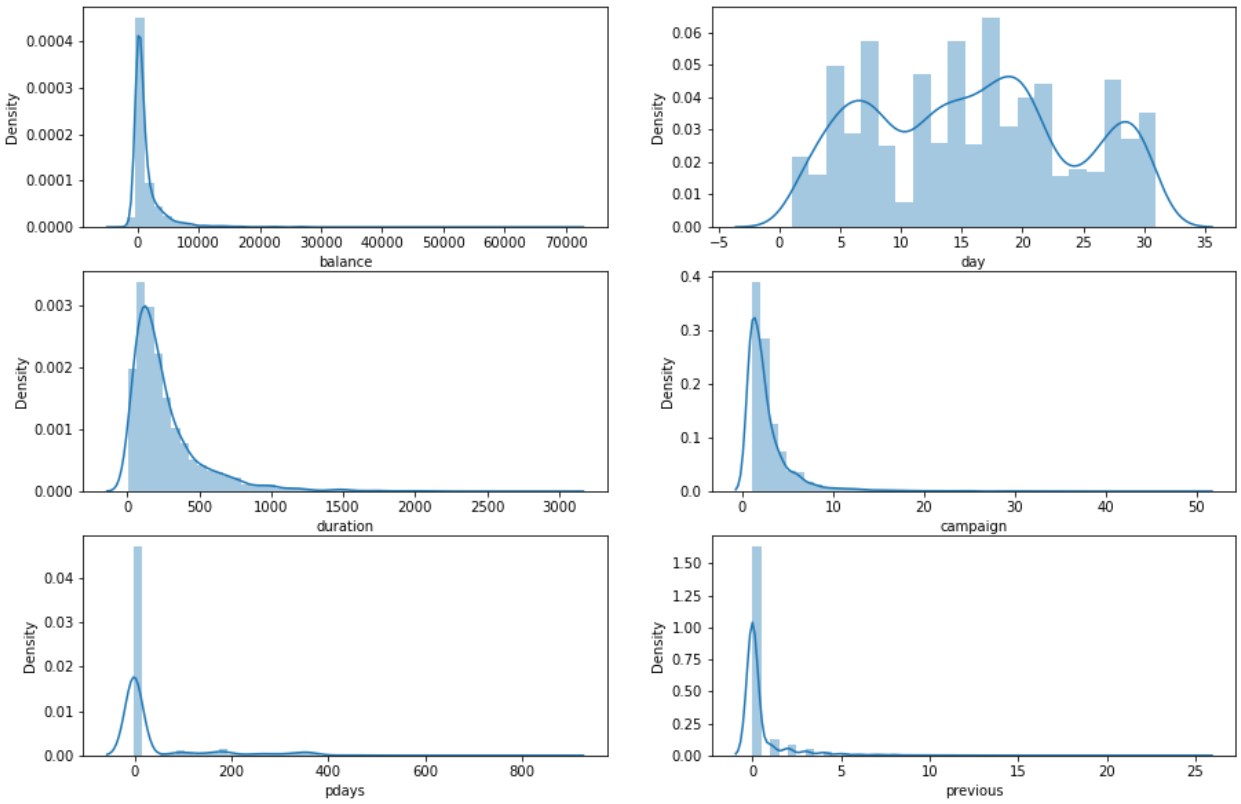
Unknown values were found to be the most. The unknown value will be considered null in our further analysis.

Clients who have subscribed to a term deposit were just 521 i.e. 11.52%.

Clients who have a loan and a subscription is listed below.



The distplot for various numerical features was plotted as shown.



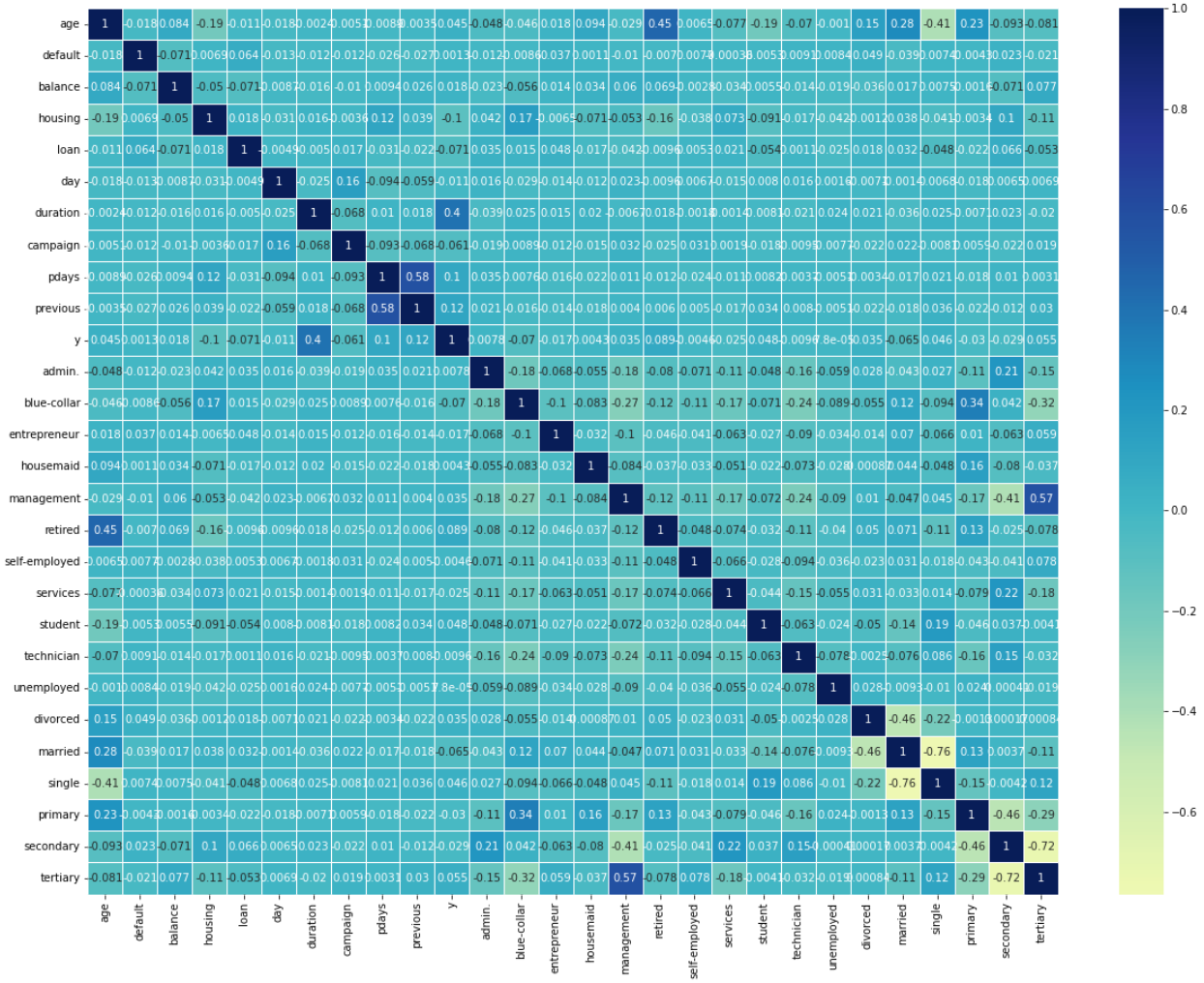
Data Processing

Features which had values unknown were converted to NaN so that numerically they are 0 or null. This would make our further analysis better. It was observed that variables such as p-outcome and contact had a huge amount of null values. Therefore these variables were dropped from our dataset. Features such as day and month were not important for our analysis since p-days give the number of days that passed by after the client was last contacted from a previous campaign. Moreover, there were few unknown values in education and job, the unknown values were replaced using the 'ffill' method where it fills the missing values with the previous value in the column.

Binary mapping was done to attributes which have values of yes or no. Yes was mapped as 1 whereas no as 0. Features such as default, housing, loan and y had this implementation.

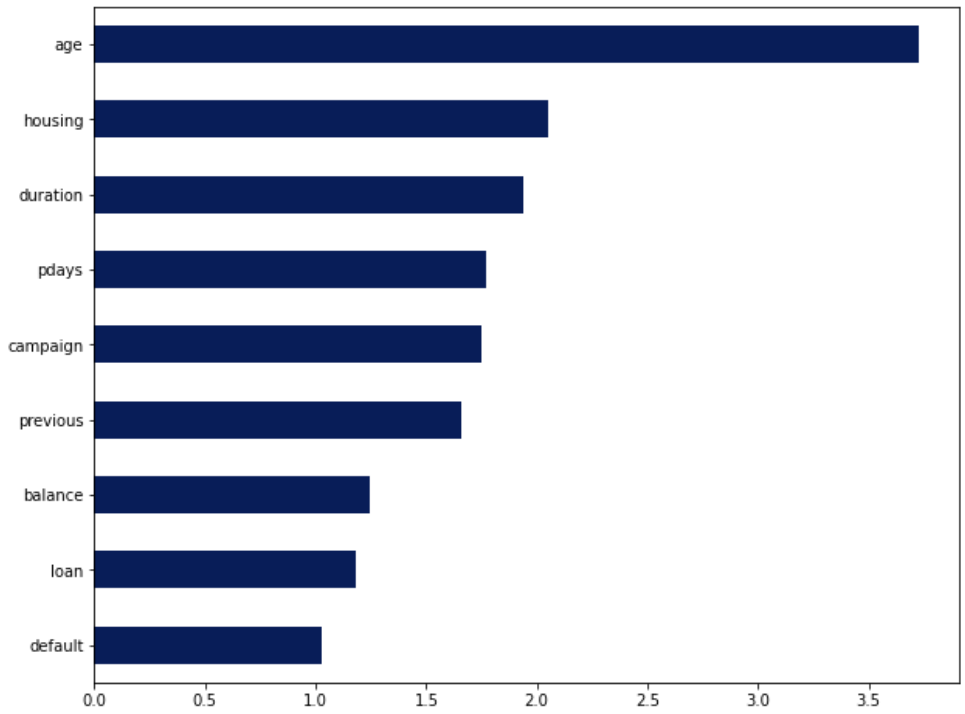
One-hot encoding was done to features such as job, education and marital.

Now our main dataset had numerical values which is now easy for analysis and model building. A new correlation heatmap was plotted based on the new data frame.



Before getting into model building, multi-collinearity among the variables was checked. We couldn't find any multi-collinearity among the variables.
Therefore every variable was considered for model building.

All the features are independent of each other.



Model Building

Necessary libraries such as sklearn, xgboost, imblearn etc…were imported first. For input, variables include all the variables except y which is the output variable. 80 percentage of the data was split for training the data and the remaining for testing purposes. Logistic regression, Random Forest and Gradient boosting were done to know which model performs better.

Logistic Regression

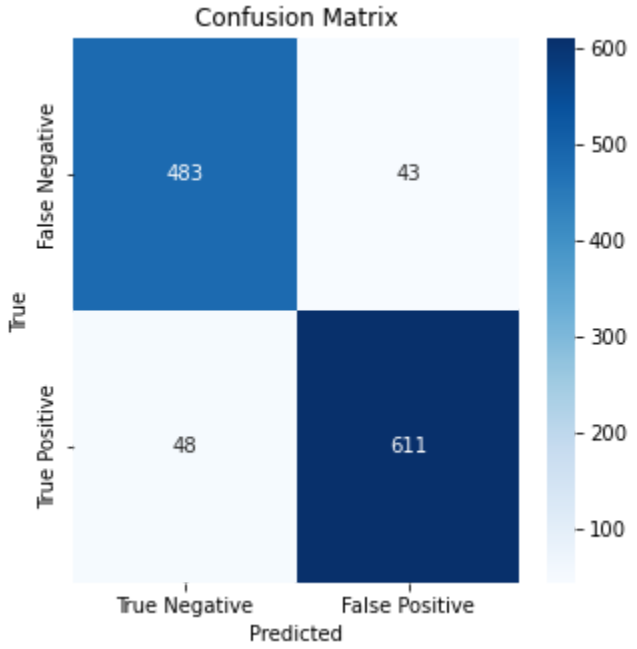
The input variables include One-Hot encoded variables, Binary mapped variables and other numerical variables. The output variable is y.

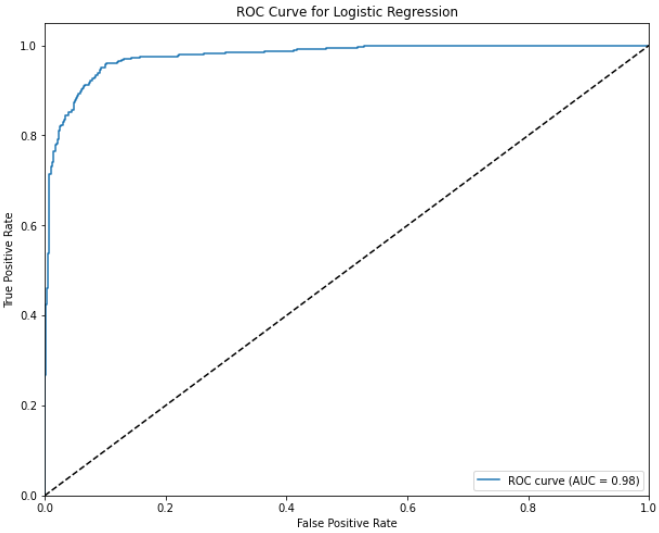
Data were split into 8:2 ratios for training and testing. After training the model and testing on testing data, an accuracy of 0.875 was obtained. Since our data was imbalanced, we used upsampling to make the data balance with the help of SMOTEENN. It fits the resampling of the input data x and target variable y using the fit_resample() method of the sm instance, and returns the resampled data x_resampled and the corresponding resampled target variable y_resampled. Now the resampled data was used to run the model and it was observed that the model accuracy was improved to 0.923. Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score and AUC-ROC.

Model Score: 0.9232067510548523

	precision	recall	f1-score	support
0	0.91	0.92	0.91	526
1	0.93	0.93	0.93	659
accuracy			0.92	1185
macro avg	0.92	0.92	0.92	1185
weighted avg	0.92	0.92	0.92	1185

A confusion matrix was plotted for the same.





A ROC curve for Logistic regression was plotted with an AUC Score of 0.98. The AUC (Area Under the Curve) score is a performance metric used for evaluating binary classification models. It measures the ability of a model to discriminate between positive and negative classes. In other words, it measures the

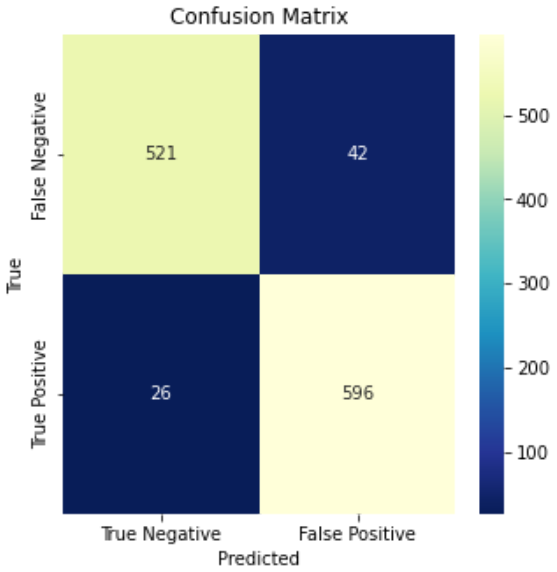
model's capability to correctly predict the positive class as positive and the negative class as negative.
The AUC score ranges from 0 to 1, where a score of 1 represents a perfect classifier, i.e., one that correctly classifies all positive and negative examples, and a score of 0.5 represents a random classifier, i.e., one that makes predictions with no skill. A score greater than 0.5 represents a model that is better than random, and a score less than 0.5 represents a model that is worse than random.
In binary classification problems, the AUC score is calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different thresholds. The TPR is the proportion of positive examples that are correctly classified as positive, while the FPR is the proportion of negative examples that are incorrectly classified as positive. The AUC score is the area under the ROC (Receiver Operating Characteristic) curve, which is a plot of the TPR against the FPR.

Random Forest
Data were split as earlier in the ratio of 8:2 and then the model was trained. The same resampled data was used here to improve the accuracy. A model accuracy of 0.942 was obtained here which was better than Logistic regression. Model matrices were also found.

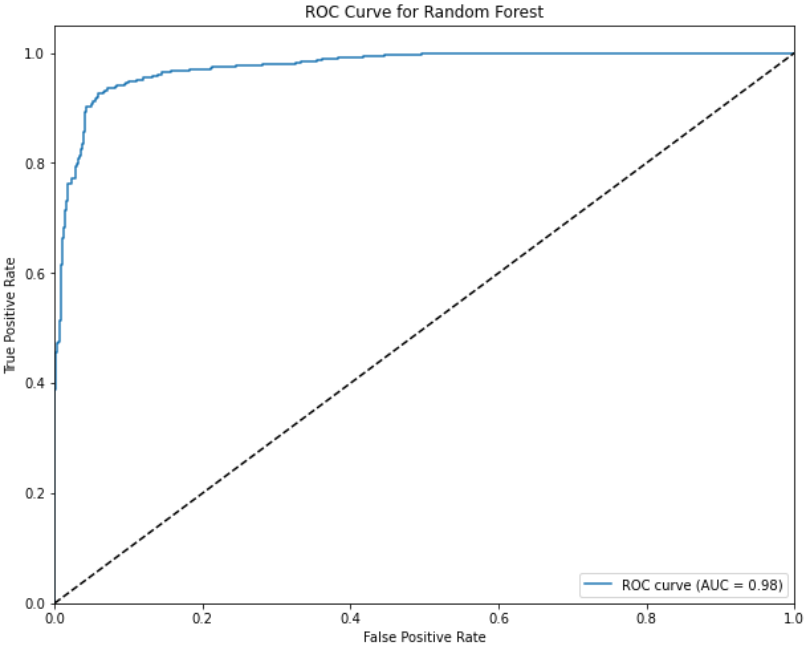
	precision	recall	f1-score	support
0	0.95	0.93	0.94	563
1	0.93	0.96	0.95	622
accuracy			0.94	1185
macro avg	0.94	0.94	0.94	1185
weighted avg	0.94	0.94	0.94	1185

A confusion matrix was plotted for the model. A confusion matrix is a table that is used to evaluate the performance of a classifier. It is a representation of the accuracy of the classifier in predicting the correct class for a set of data points. The matrix is usually 2-dimensional, with the rows representing the actual class of the data points and the columns representing the predicted class.
Each cell in the matrix represents the number of data points that belong to a certain actual class and were predicted to belong to a certain predicted class.

The diagonal elements of the matrix represent the number of data points that were correctly classified, while the off-diagonal elements represent the number of data points that were misclassified.
Based on the values in the confusion matrix, several evaluation metrics can be computed, such as accuracy, precision, recall, F1-score, etc. These metrics provide different perspectives on the performance of the classifier and can help in understanding the strengths and weaknesses of the classifier.



A ROC curve with an AUC value was also plotted for the same.

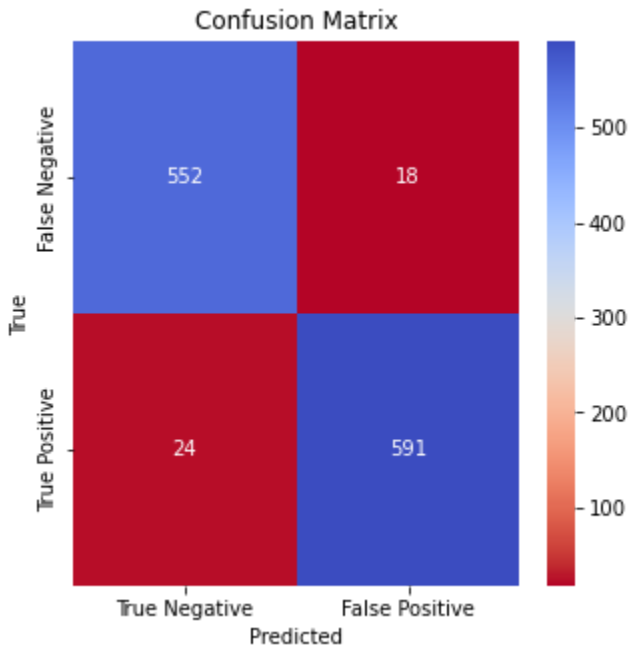


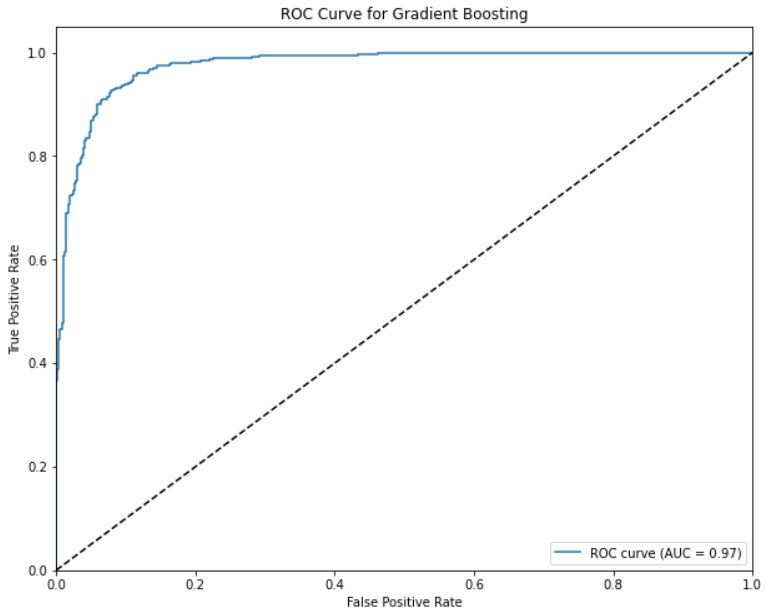
Gradient Boosting

A model accuracy of 0.96 was obtained which was even better than Random Forest. Model metrics has been shown below.

	precision	recall	f1-score	support
0	0.96	0.97	0.96	570
1	0.97	0.96	0.97	615
accuracy			0.96	1185
macro avg	0.96	0.96	0.96	1185
weighted avg	0.96	0.96	0.96	1185

The confusion matrix and ROC curve were also plotted for a better understanding of the model performance.





To further enhance the model performance of Gradient boosting, cross-validation was done. Cross-validation is a technique for evaluating the performance of a machine learning model by splitting the data into multiple folds, training the model on a portion of the data, and evaluating it on the remaining portion. After cross-validation, we obtained a model score of 96.14% which was a little improvement.

Further, a Random search technique for hyperparameter tuning was done to know the best parameters to improve the model performance. This improved the model's performance better by giving an accuracy of 96.45%.

Conclusion

I could analyze the Portuguese bank marketing dataset and come up with various models which could improve the analysis. I could plot various charts to better understand the data. EDA was done to know the various underlying factors in our data.