Deposit- Analysis

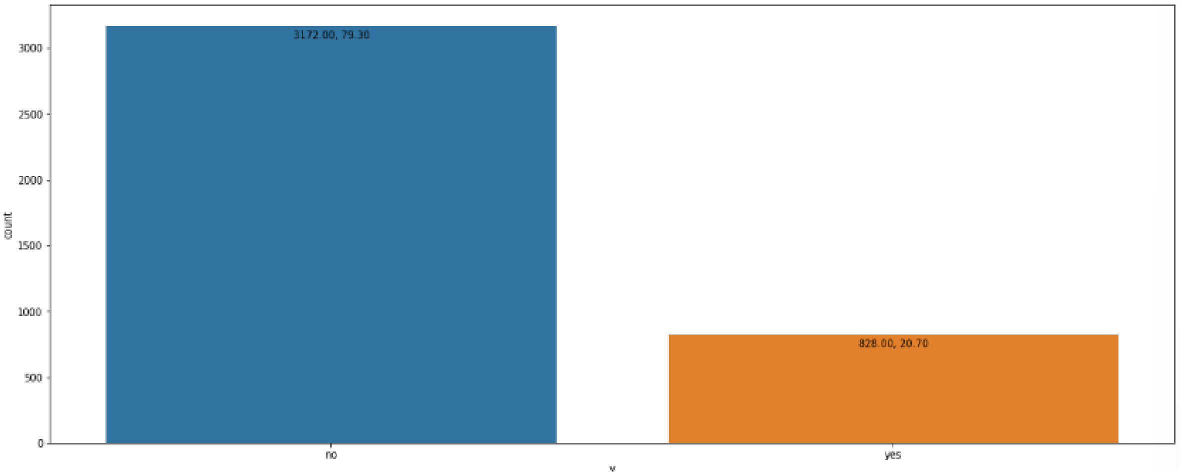
The data handed over by the previous company has provided us with some useful insights. These insights, analysis and predictions are going to be presented in this following report.

The provided dataset contains 4000 records,15 independent features and one dependent feature ‘y’ which indicates if the client invested in the term deposit or not. As time progresses, more data will be generated by the cold calling team, the model will be trained on a bigger dataset to increase the accuracy.

Section A: Summarization

Univariate analysis of categorical variables-

The data provided by the previous company contains 4000 rows and 16 columns.



The dataset is highly imbalanced as 79.30% of the dependent feature is ‘no’ and 20.70% is ‘yes’.

There are no null values and NA values, but a lot of unknown values which cannot be dropped due to limited number of records. Thus, during pre-processing dummy variables are created which gives the flexibility to drop the categorical column containing the unknown data.

The dataset provided contains continuous and categorical data.

Distribution analysis of categorical data:

Job: The two groups that were approached the most were management (21.20%) and blue-collar workers (19.98%). The least approached were students (2.02%) and housemaids (3.10%). 0.65% of those contacted had unspecified work types.

Marital: Married(58.25%) individuals make up most of the dataset, whereas less data has been received for divorced (11.65%)individuals.

Education: 50.70% of individuals have had a secondary education and 15.25% individuals have highest education till primary, No data is found for 4.03% individuals.

Default: The data for default is highly skewed towards ‘no’ by 98.22% and should be dropped as it provides no value.

Housing: Housing loan data is balanced, with 53.15% "yes" responses and 46.85% "no" responses.

Loan: 15.25% of people have taken out a personal loan, compared to 84.75% of people who haven't.

Contact: 67.38% were contacted via cell phone, 5.73% via telephone and no data is found for 26.90% of population.

Poutcome: 80.73% of data is unknown but 4.45% of individuals were successfully converted in the previous campaign show high interest on the term deposit. Thus, this feature cannot be dropped.

Key takeaways between categorical input and output features-

Note- The percentages below indicate the percentage of individuals who were interested in the term deposit to the total approached individuals.

Job: 36.91% of retired and 41.97% of students approached showed interest in the deposit, but the total individuals approached in this category is low. Blue-collar, management and technician were less keen on the deposits as only 14.87%, 23.46$ and 19.53% showed interest.

Marital: Married individuals(18.02%) were less inclined for deposit compared to single individuals (25.33%).

Default: Default data is highly skewed for any insight and will be dropped.

Education: Difference in education shows no significant difference in investing in term deposits. Individuals having tertiary education were seen to show slightly more interest.

Housing and loan: People with no housing loan and personal loan(27.37%) (22.09%) were slightly more interested in deposits as compared to individuals with housing loan and personal loan(14.81%) (12.95).

Contact: The method of contact didn’t affect the output feature significantly.

Poutcome: 77.5% of individuals who subscribed to the previous product also subscribed in the current campaign.

In the initial stages of analysis, Poutcome seems to be the most important variable, but the high number of unknown values may result into less weight.

Numerical Features analysis-

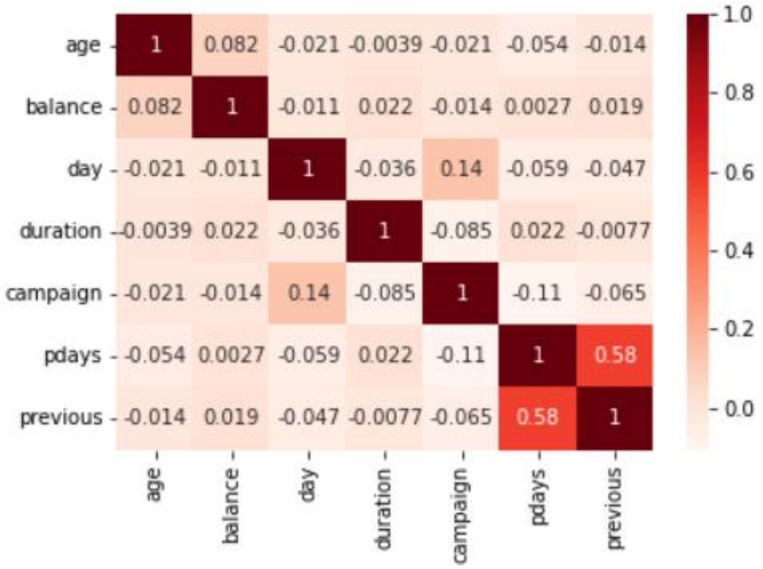
Age and day are distributed normally.

Balance, duration, campaign, pdays and previous are heavily skewed towards left and also have outliers.

No outlier removal is to be performed for age as individuals are between the age of 18 and 95, which is expected. People having higher bank balance are more interested in term deposits, thus an outlier removal is not performed on balance.

Clients who have had a longer duration seems to show more interest in the term deposit. Duration variable cannot be used in our predictive model as we will not have this information before contacting the client. Thus, duration is dropped.

No feature is seen to be correlated.



Section B: Exploration

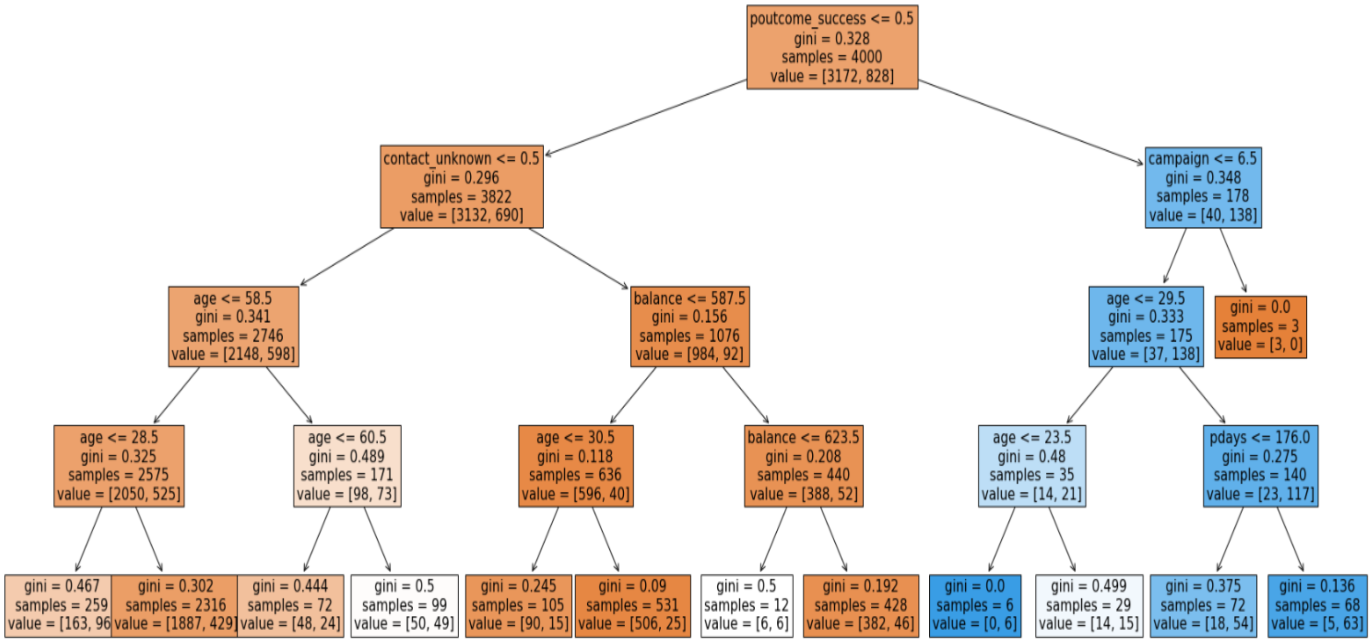
When the maximum depth is not set, thus decision tree accuracy comes out to be quite low, thus it is decided to limit the maximum depth of decision tree.

The decision tree was giving the best accuracy when the max_ depth was limited to 4. An accuracy of 81% along with 67% precision was obtained (without duration variable). Thus, for data exploration and finding the important features we will limit the depth to 4.

Duration is the most important feature as per our analysis but we cannot use it for model implementation as we will not have any idea of the duration before contacting the individual.

Data exploration was performed with and without duration and the following features were found using the 'feature_importances_ ' method. Only the most important features were considered as the max depth of decision tree was limited to 4.

Our initial analysis was proven correct by the decision tree. Previous customers showed high interest and thus higher weightage. The higher the duration of the call the higher the chances of conversion are. Certain age bands of individuals were seen more keen to invest.



Most important features for *analysis* (duration considered)

Feature	Importance
duration	0.6575
poutcome_success	0.2531
contact_unknown	0.0714
pdays	0.0059
age	0.0049
previous	0.0037
marital_married	0.0032

Most important features for *model Implementation* (duration not considered)

Feature	Importance
poutcome_success	0.6028
age	0.2063
contact_unknown	0.1355
balance	0.0259
campaign	0.0184
pdays	0.0109

The important features are selected by how the output variable is influenced by a feature. An feature which highly affects the output variable but have less occurrences will be assigned lesser weight. As you increase the depth the weight of individual features get distributed.

Observations made from the decision tree-

It is seen that, when the previous outcome is successful and the campaign exceeds more than 6.5 the outcome is always no.

It is seen that for previous outcome as successful, campaign is less than 6.5 and age is less than 23.5, the outcome is always successful.

As per the decision tree, if previous outcome was unsuccessful, the bank balance is less than 587.5 and age is greater than 30.5 only 4.7% people subscribed to our product.

Output variable is highly dependent on the poutcome_success. Overall, people with higher age and high bank balance invest more in term deposits.

Section 3 : Model Evaluation

Data Preprocessing

To get the best results on any model the data first has to be pre-processed. The following steps were taken to get the best accuracy in the selected models.

Outlier adjusting- For campaign, pdays and previous instead or completely removing the outlier the top 1% was adjusted according to the max_threshold as 99 percentile. This number was chosen as it gave the best accuracy. Adjusting a higher quantile was resulting into lesser accuracy for the models.

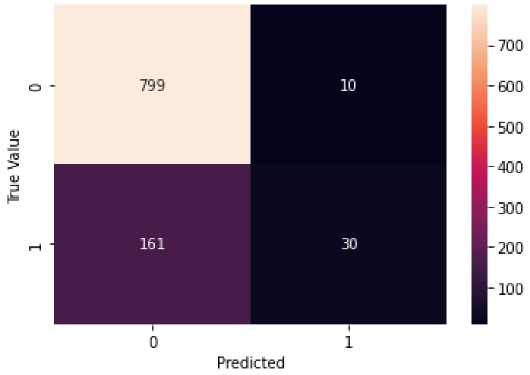
Scaling/Standarizing- MinMaxScaler was used to normalize the data to boost the performance of the model. Numerical data was scaled between 0-1. Categorical data converted to numerical data was not scaled.

Encoding / Categorical data to numerical data- There were a lot of categorical data which was converted to numerical data using the 'get_dummies' method. For each feature a redundant column was dropped using the drop_first method. Get_dummies was used instead of label encoding due to the creation of separate columns which makes it easier to drop the less important columns.

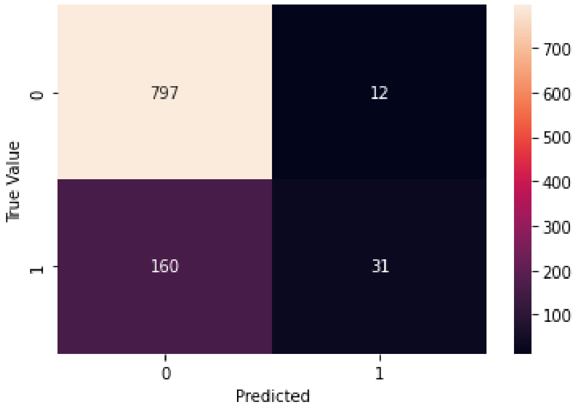
Dropping insignificant columns- The columns containing original categorical data are dropped. Along with that insignificant columns decided by the decision tree are dropped too.

Cross validation testing- Our model is performed on multiple splits to see how it performs on unseen data. An average of 5 scores is taken for which a mean is generated which is taken as final score.
Accuracy, precision and recall for each model is calculated and final model is chosen.

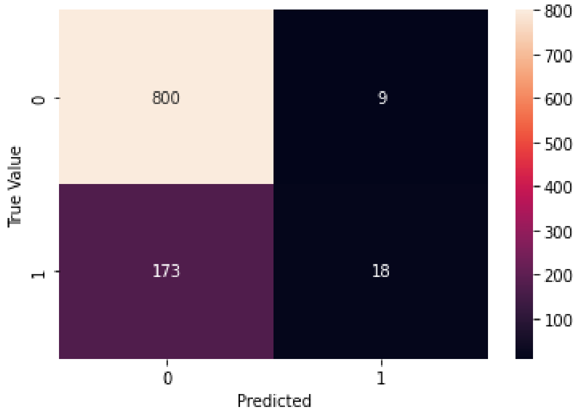
The models which have been evaluated are as follows-
Random Forest Classifier- Random Forest Classifier tend to have high accuracy for large number of features or when relationships are complex. It is a combination of multiple decision trees, thus generally giving higher accuracy than decision tree. Random forests are also considered to be resistant to overfitting.



Logistic Regression- Logistic Regression classifier is preferred due to its simplicity and easy understandability. In case of Binary classifications, logistic regression model is preferred. The time complexity and data required for Logistic regression model is less, making it highly efficient.



K-NN- K-NN is a non parametric, supervised classifier which groups the individual data points to make predictions. The closer datapoints are put in one class. It doesn't make any assumption about the underlying data and follows a lazy learning method. K-NN is preferred in cases where there are high number of features and small dataset.



```
Classification Report: for Random Forest
              precision    recall  f1-score   support

           0       0.83      0.99      0.90       809
           1       0.75      0.16      0.26       191

    accuracy                           0.83      1000
   macro avg       0.79      0.57      0.58      1000
weighted avg       0.82      0.83      0.78      1000

Classification Report: for Logistic Regression
              precision    recall  f1-score   support

           0       0.82      0.99      0.90       809
           1       0.67      0.09      0.17       191

    accuracy                           0.82      1000
   macro avg       0.74      0.54      0.53      1000
weighted avg       0.79      0.82      0.76      1000
```

```
Classification Report: for  K-NN
              precision    recall  f1-score   support

           0       0.83      0.99      0.90       809
           1       0.72      0.16      0.26       191

    accuracy                           0.83      1000
   macro avg       0.78      0.57      0.58      1000
weighted avg       0.81      0.83      0.78      1000
```

Model Efficiency

From the above confusion matrix and classification report, It is seen that all the models struggle to predict the target output class, this is due to the low occurrences of the term deposit subscribed. The accuracy and predictions are considerable due to the accurate predictions of people not subscribing to the term deposit. Only way to solve this problem is having a balanced dataset with high number of rows. For now, overall accuracy and precision is considered to check the efficiency of the model.

As seen in the classification report, precision and recall value for Logistic regression is low and is thus eliminated. K-NN and Random forest performed very similarly with Random forest having slightly better precision for the target class.

Due to the current situation, we will consider the overall accuracy and precision of the model as the parameter of model selection. In further analysis, where the dataset get's balanced, sensitivity will be the parameter of choice as the cost of failing to predict the target class is very high compared to a false positive. In our case false positive are not expensive, but not calling a high potential client can cause significant loss to the company.

Section D: Final Assessment

As informed in the previous section, due to the dataset being highly imbalanced and low number of observations, all the models tested (Decision tree, Random forest, K-NN, Logistic regression) have a high number of false negative.

Random forest will be chosen due to familiarity, robustness and the efficiency can be increased using the gridsearchcv method which gives the best parameters for the model.

The best parameters turned out to be criterion= 'entropy', max_depth= 3, max_features= 'auto', n_estimators= 100

In case of the above tested models, we got the highest accuracy, precision and f1 score for Random forest with K-NN being close 2nd. Highest accuracy is required for the model. False positives are not an issue as cost of calling is bearable as compared to not contacting a high potential client.

Thus we require

High- True positives , Low- False Negatives, False positive-no impact (cost is low)

In future, when there will be sufficient data to balance the dataset and get more observations, Random forest will be the best performing model compared to others due to it's resistance to overfitting and high accuracy for complex relationships.

Section E: Model Implementation:

Testing the model

1. Open the test_model.ipynb file on any chosen platform like jupyter or Google Collab (Collab preferred). This ipynb file is available with this report.
2. For Google Collab, please upload your dataset and name it as 'test.csv' along with the model which is uploaded along with this report named 'finalised_model.sav'.
3. For Jupyter Notebook, please upload the files in the same directory.
4. All the necessary requirements are now complete, Go to runtime and click on run all.
5. The model's classification report along with the confusion matrix will be available at the bottom.

Predict data

1. To predict the data, open the predict.ipynb file uploaded along with this report on Google Collab or Jupyter notebook.
2. Please upload the 'finalised_model.sav' file on the Google Collab file section, for Jupyter Notebook keep the sav file in the same directory.
3. Upload the dataset whose data is to be predicted in the same section and directory.
4. All the necessary requirements are now completed. Please click on the run all tab in the runtime section.
5. 5. The dataset with the predicted value is now stored in the data variable.

Section F: Business Case Recommendations

1. Existing clients who have had a term deposit with our bank show very high interest in N/LAB platinum deposit. All the previous clients should be contacted on priority basis.
2. People with a high bank balance show high interest and should also be contacted.
3. As the duration of call increases the client becomes more interested in our products. Thus, the client on call should be explained more about the benefits and perks, increasing the duration ultimately increasing the client's interest.
4. Retired people should be targeted as they were seen to be interested in the term deposits.
5. As more data is gathered, the model will be trained on that to increase the accuracy, recall and give out better predictions.