Project Overview

Term deposit is an important source of income for banks which can be used to disburse loans at a higher interest rate to their customers. Building a new customer base from term deposit can also improve the reputation, customer engagement and other source of income (i.e., investment fund) as well for the banks. Hence, the bank uses a lot of marketing techniques to connect and acquire new customers for their campaigns. Telephonic marketing (i.e., phone calls) remains an effective way to acquire term deposit customers, especially if enabled with machine learning. Banks will be using data and machine learning informed alert models to identify customers who are more likely to subscribe a term deposit and to inform a telephonic marketing campaign accordingly.

In this project, a classification task is expected to be performed to find out which customers are more likely to subscribe to the bank's term deposit via telemarketing campaign. The classification goal is to predict if the customer will subscribe to a term deposit (Term Deposit = 1) based on customer characteristics and macroeconomic indicators.

Summary of project outcome

The model was able to predict the customers who subscribe to a term deposit with a 73% successful rate. Macroeconomic indicators such as Euribor 3-month rate, employment variation rate holds a higher influential power when customers decide on subscribing a term deposit. When it comes to customer demographics, type of contact, ethnicity and age group have a higher power on dictating customer decisions.

Recommendations:

- Launch campaigns during periods of low Euribor 3-month interest rates and low employment variation rate.
- Focusing on customers who are between 26 to 36 as they subscribe the most for term deposit.

Data Overview

The dataset from this project is related to the direct telemarketing campaigns (phone calls) of a Portuguese banking institution and was collected from Kaggle.

Before Preprocessing:

The raw dataset consists of over 40,000 records and 22 variables, including customer demographic details like age, job, marital status, macroeconomic indicators such as Euribor 3-month rate, employment variable rate and marketing-related variables such as previous campaign outcomes.

After Preprocessing:

The processed data remains around 30,000 rows and 15 features. Outliers within numerical columns were imputed using its upper limit and some columns like month and day of week were removed because they are not predictive columns by themselves. Some columns such as number of employees were also removed due to its interpretability towards the outcome variable.

The processed data was then split into 75% training data and 25% testing data, stratified by the target variable to ensure the data is balanced. The training and test data was encoded individually to ensure the test data would not be exposed to the model during the model training process with training data. Categorical variables like marital status and jobs were one-hot encoded, binary columns like default, contact communication type and columns with hierarchy like education level were also encoded into numerical values.

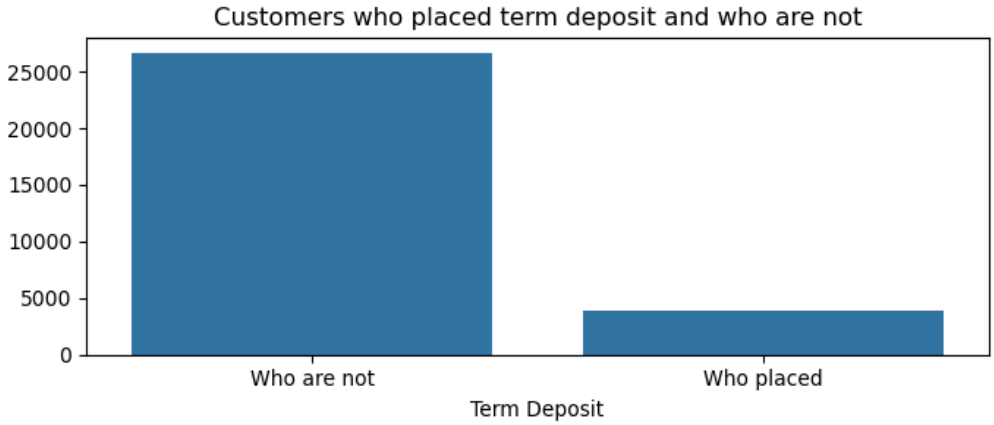
Training data was then split again into 75% training data and 25% validation data, stratified once again by the target variable. Having a separate subset of validation data can validate and evaluate the model performance with unseen data after training the model which results will be more realistic to real-world scenarios.

Lastly, the training data and validation data was scaled with StandardScaler to ensure the data using the same scale when fitting into the model.

Exploratory Data Analysis

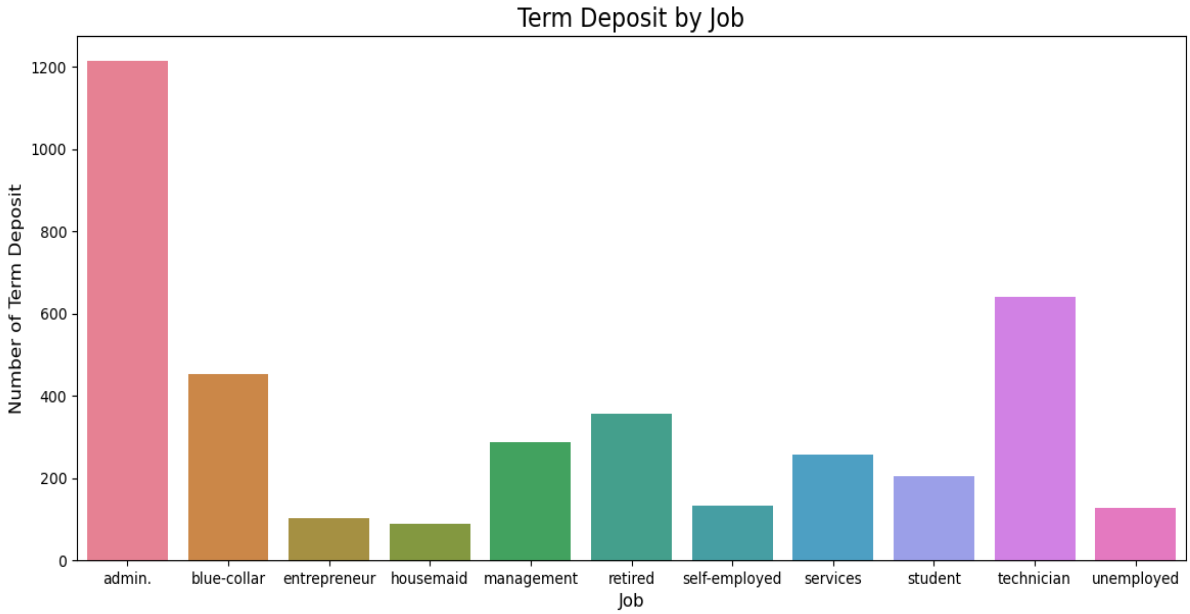
Graph 1: Proportion of Term Deposit

Only 12% of approximately 4,640 of the customers subscribed to a term deposit (Graph 1) which is crucial in deciding the model evaluation focus during modelling.



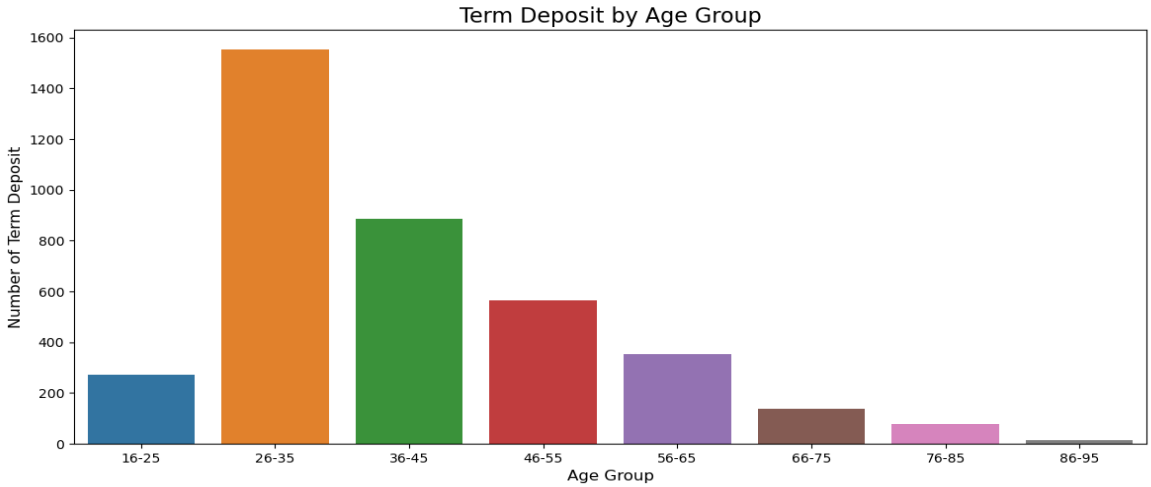
Graph 2: Term Deposit by different Jobs

Bigger proportion of the term deposit subscribers are admins. Housemaid and entrepreneur subscribe the least on term deposit.



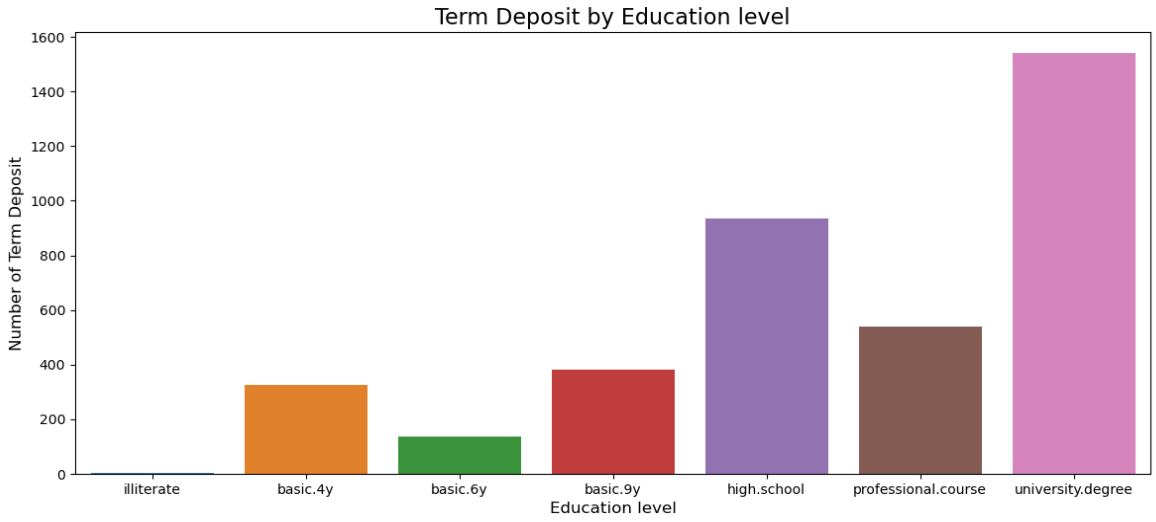
Graph 3: Term Deposit by different Age Groups

Customers between 26 to 35 have the highest subscribers to a term deposit and the age group between 86 to 95 has the least subscribers compared to other age groups. As people grow older, people are less likely to subscribe to a term deposit which has a pattern of decline starting from age 26 to age 95.

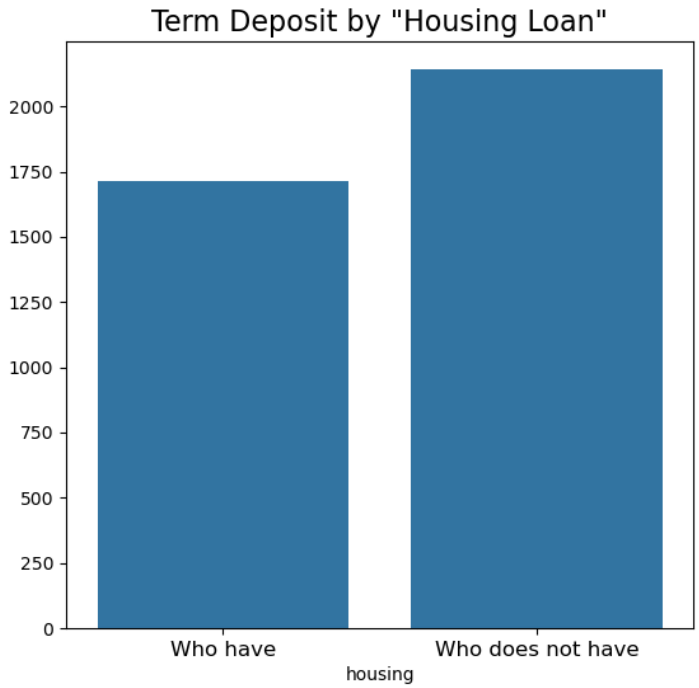


Graph 4: Term Deposit by different Education Level

Bigger proportion of the subscribers are university degree graduates. There is an incline pattern on term deposit as the education level increases which shows that education level is positively correlated with term deposit.

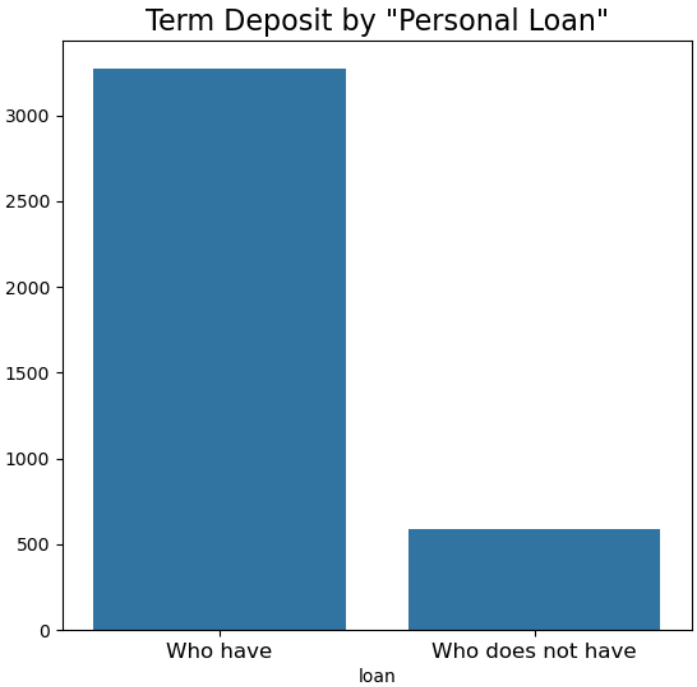


Graph 5: Term Deposit by Housing Loan



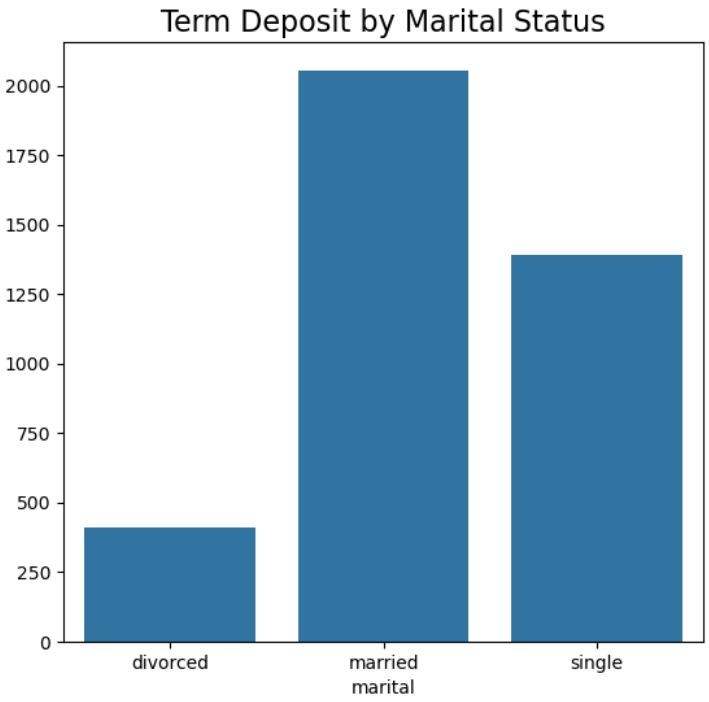
Customers with housing loans are less likely to place a term deposit compared to customers who don't have a housing loan. This can be due to more financial pressure in which they don't have extra income to place a term deposit.

Graph 6: Term Deposit by Personal Loan



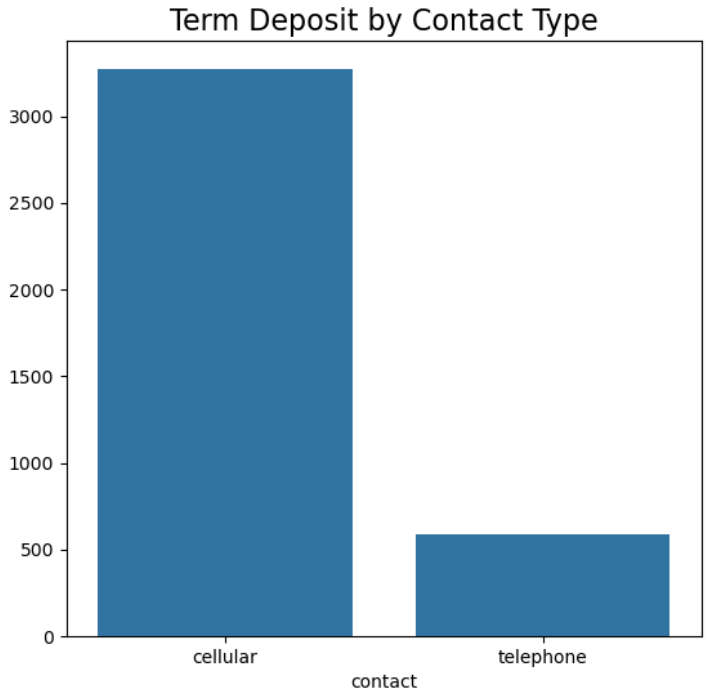
There are more customers with personal loans who subscribe to a term deposit which can be due to their proactive approach to wealth-building.

Graph 7: Term Deposit by Marital Status



Married couples or families are more likely to subscribe to term deposits compared to other groups.

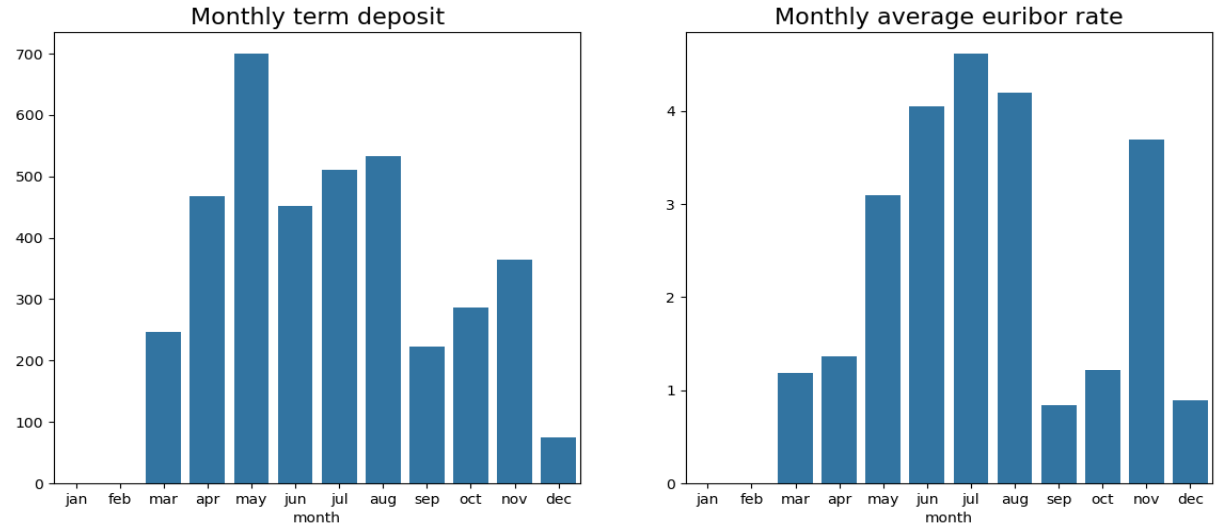
Graph 8: Term Deposit by Contact Type



Customers who have wireless cellular phones are more likely to place a term deposit, which may be due to the fact that they are more easily able to receive contact from the bank during the campaign compared to landline telephone.

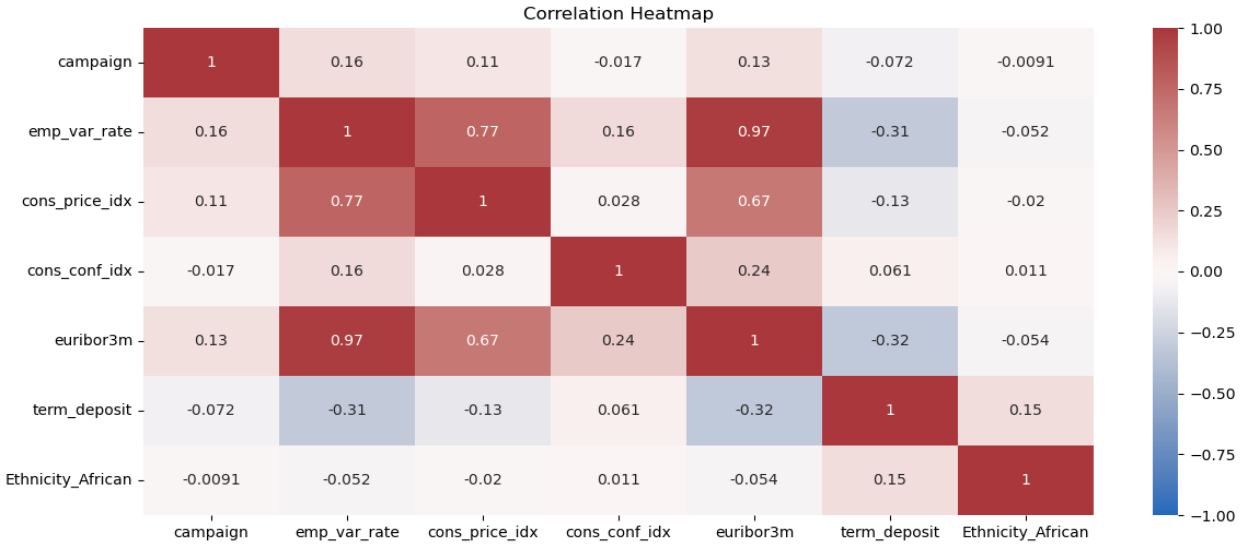
Graph 9: Monthly average Euribor Rate

From graph 7, monthly term deposits from April to August and November are higher than other months which are slightly identical to higher monthly Euribor Rate from May to August and November.



Graph 10: Correlation Heatmap

The correlation heatmap shows that employment variation rate and Euribor 3-month rate have the highest negative correlation with term deposit which have more predictive power than other features on predicting term deposit. When the employment variation rate is low, people might prefer safer investments like term deposits.



Methodology

Throughout the modelling process, 4 models were used to perform classification task on predicting customer term deposit subscriptions which are Decision Tree**,** Random Forest, XGBoost and Logistic Regression.

Recall metric was being prioritized in the modelling process as there are only a small portion of customers who place term deposit (true positive) which reduces false negatives, missed opportunities become much more important and valuable for the bank.

There will be a tradeoff between precision and recall, focusing on reducing the false negatives will make the model to classify more positive predictions which can lead to more false positives. The higher the recall, the lower the precision. In my opinion, false positives are just a waste of time and efforts for the bank which are less important in comparison to missed opportunities so recall should be focused on.

Within the modelling process, k-fold cross validation of 5 and balanced class weight were used for all the models to counter the imbalanced class within term deposit and provide much more reliable results. A random state of 42 was used to reproduce the same results. GridSearchCV was fitted into the model during the model training process to search for the best hyperparameters. After the training process is completed, validation data was fitted into the tuned model with the best hyperparameters to evaluate the model performance.

Model performance on validation data:

Model	Recall	Precision	F1	Accuracy	AUC
Decision Tree	0.68	0.37	0.48	0.81	0.80
Random Forest	0.75	0.26	0.39	0.70	0.81
XGBoost	0.69	0.37	0.48	0.81	0.82
Logistic Regression	0.70	0.31	0.43	0.76	0.80

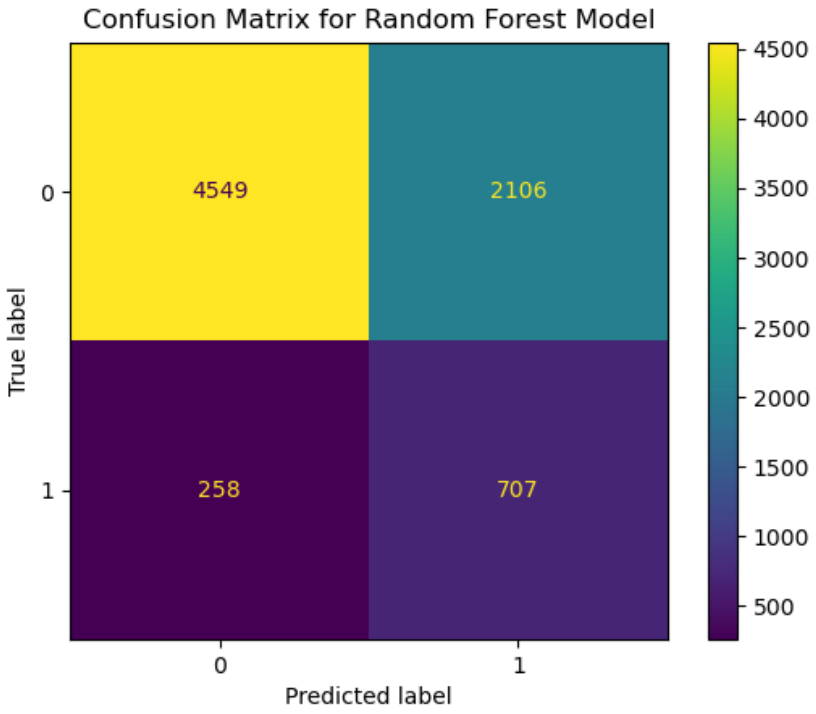
Random Forest was chosen to be the champion model to run on test data as the model produces the highest recall and least false negatives among other models.

Project Test Results

Random Forest model performance on validation data and test data:

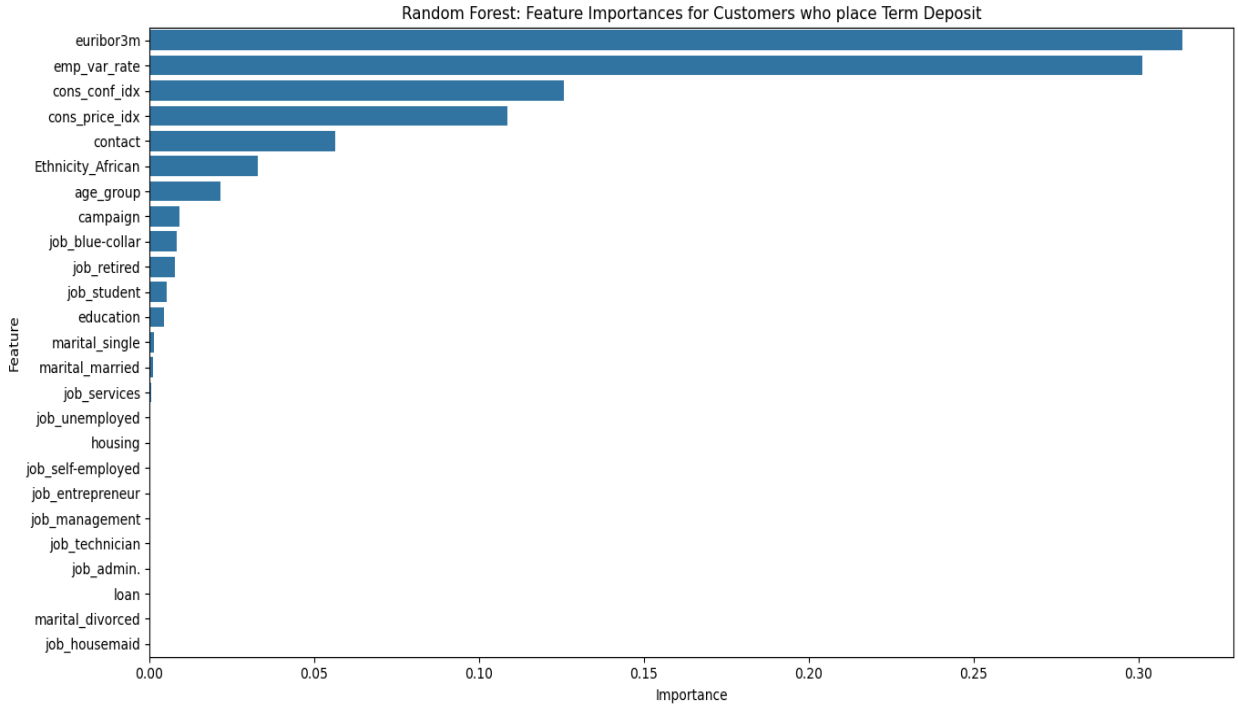
Model	Recall	Precision	F1	Accuracy	AUC
Random Forest on validation data	0.75	0.26	0.39	0.70	0.81
Random Forest on test data	0.73	0.25	0.37	0.69	0.71

Random Forest Confusion Matrix on test data:



The model was able to predict 707 true positives out of 965 positives which gave a recall of 0.73.

Random Forest Features Importance on test data:



From the graph, the top 4 most important features based on Random Forest Gini Importance are actually macroeconomic indicators which dictate whether a customer decides on subscribing a deposit.

Euribor 3-month rate and employment variation rate have the highest feature importances, bank can launch campaigns when both of the rates are low which is believed will lead to more term deposit subscription because both rates are negatively correlated to the subscription.

Telemarketing contact type is one of the features that is important in dictating customer decisions. People with cellular phones are more likely to place a deposit which the bank can focus on.

Other than that, age group is also one of the important features which bank can be focusing on. Customers who are between 26 to 35 subscribed the most term deposit compared to other age groups.