Marketing Campaign Outcome

A reference hackathon presentation

Problem Statement

- Predict if the client would subscribe to a term deposit based on a marketing campaign.

Potential Business Problems

- Run optimized campaigns to bring in more customers, and thereby increase the bank revenue ?
- Increase long-term holdings which can be further invested in different financial instruments?
- Stakeholders
  - Chief Marketing Officer ?
  - Campaign Strategy Manager ?
  - Who else?

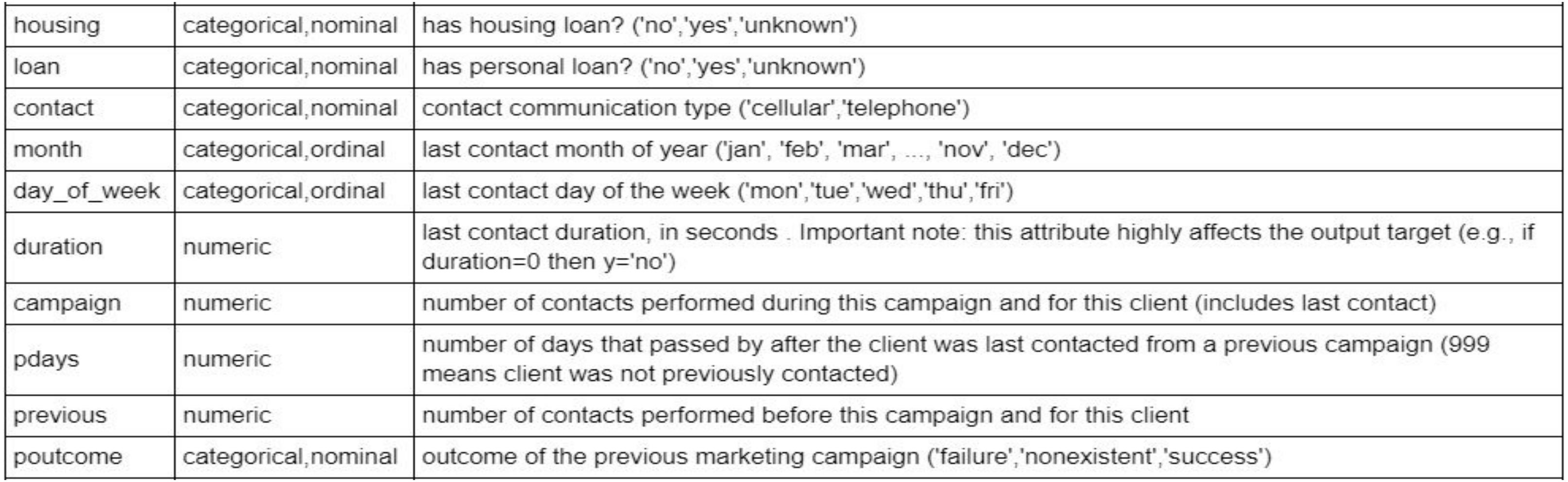
Why solve this problem?

- Business Impact
  - Improve prediction -> identify common features of subscribing customers -> targeted campaigns
  - Improve prediction -> identify right target audience-> efficient budget for marketing
  - Improve prediction-> identify right frequency interval for campaign -> optimum campaign
-

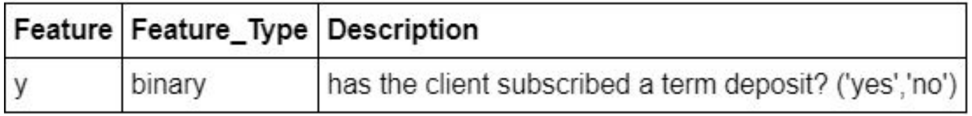
Data

Dataset Information : The data consists of records of roughly 41000 clients and 21 features. There are 20 predictors and 1 target that describes whether the client will subscribe or not.

Below are some of the features and the target variable



housing	categorical,nominal	has housing loan? ('no','yes','unknown')
loan	categorical,nominal	has personal loan? ('no','yes','unknown')
contact	categorical,nominal	contact communication type ('cellular','telephone')
month	categorical,ordinal	last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec')
day_of_week	categorical,ordinal	last contact day of the week ('mon','tue','wed','thu','fri')
duration	numeric	last contact duration, in seconds . Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no')
campaign	numeric	number of contacts performed during this campaign and for this client (includes last contact)
pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
previous	numeric	number of contacts performed before this campaign and for this client
poutcome	categorical,nominal	outcome of the previous marketing campaign ('failure','nonexistent','success')



Feature	Feature_Type	Description
y	binary	has the client subscribed a term deposit? ('yes','no')

Evaluation Metric

The evaluation metric for this project is **AUC_ROC_score.**

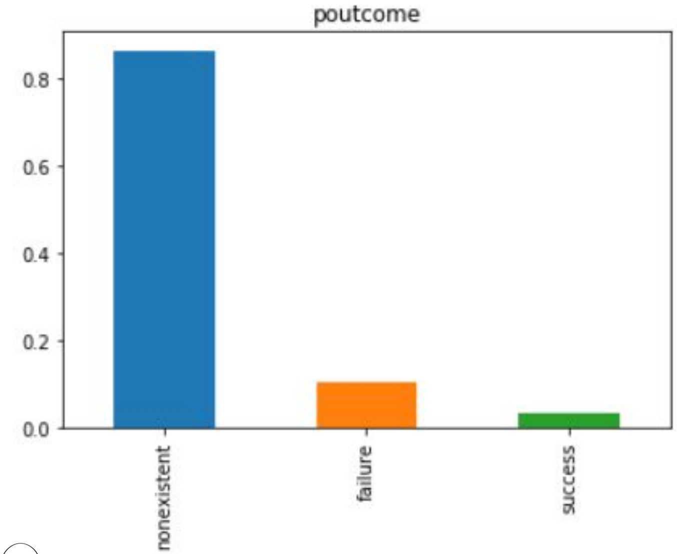
False Positive - predicted subscribe to a term deposit, but actually not subscribed.

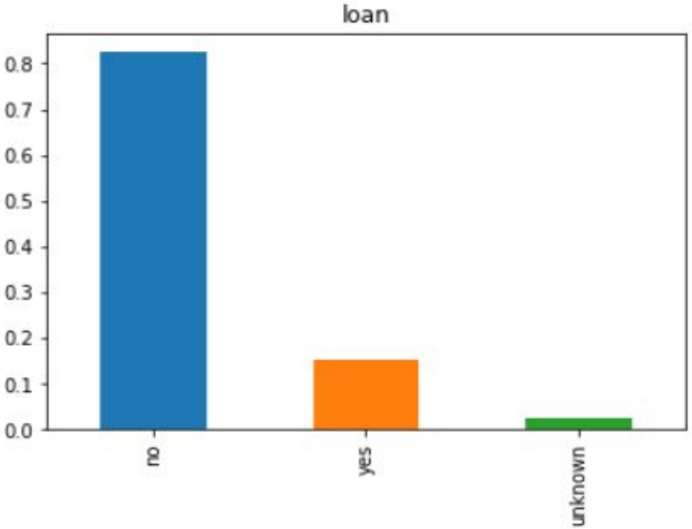
False Negative - predicted not subscribe to term deposit, but actually subscribed.

For the use case, False negatives must be reduced. So recall to be given more importance.

First steps - EDA

On the left is the univariate analysis of the feature **poutcome** and on the right is that of **loan**.

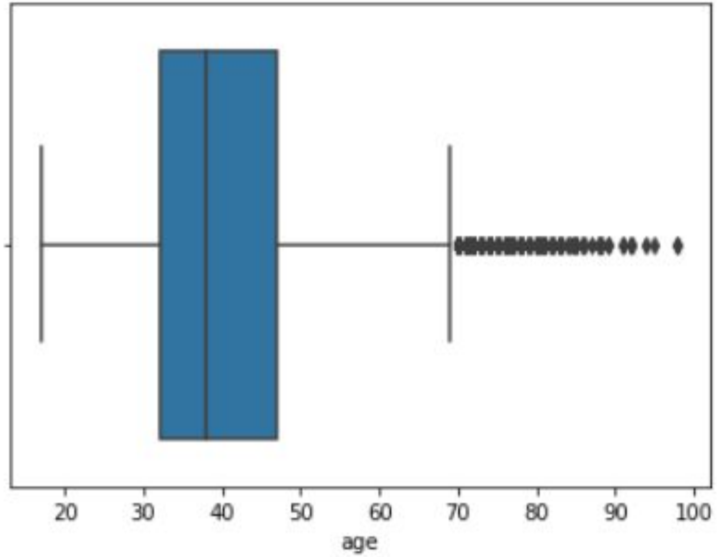


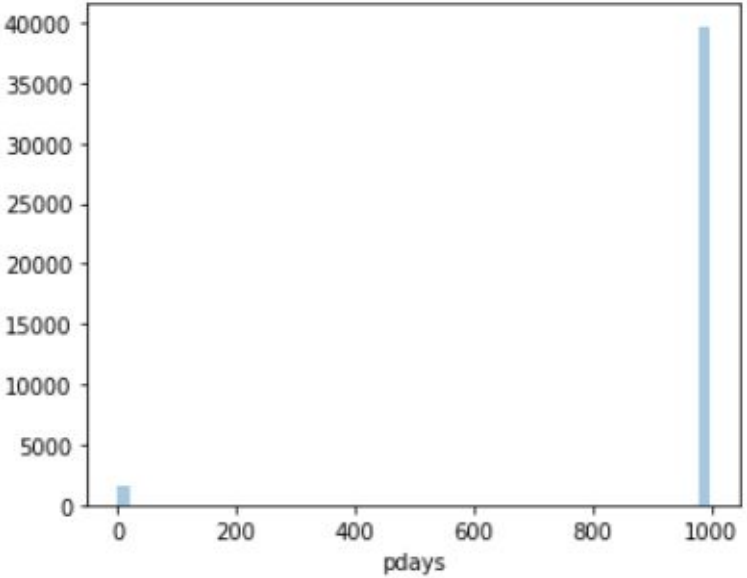


Most of the data is of customers where we are not sure about the outcomes of the previous campaign.

Most people already have a personal loan. (Will they be willing to commit to a term deposit?)

EDA - continuous - age, pdays - Bring out Key Insights



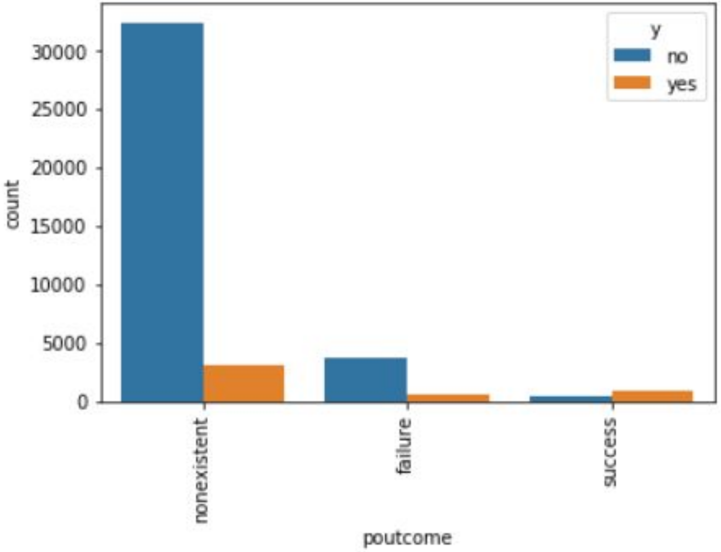


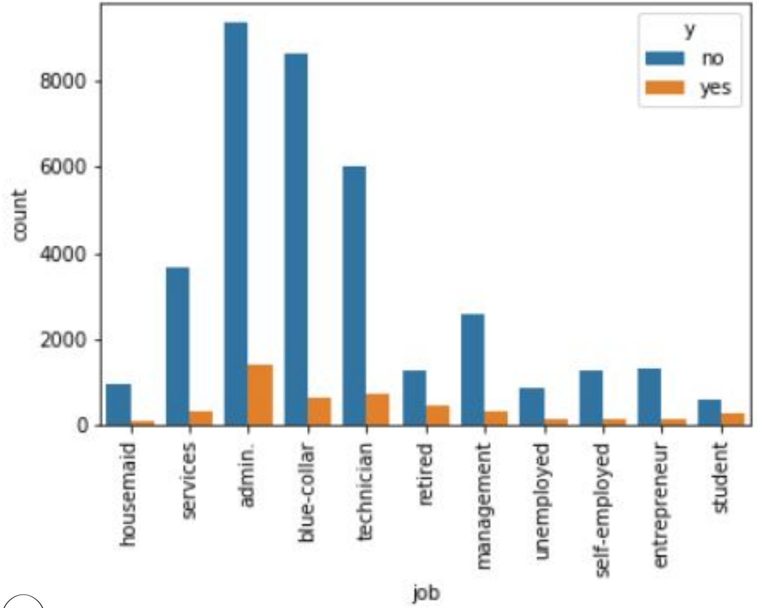
Age - many in the age bracket 30-40 i.e working population

Pdays - lot of previously non-contacted customers.

EDA - bivariate

Below are the bivariate analysis of features **poutcome** and job w.r.t the **target**.





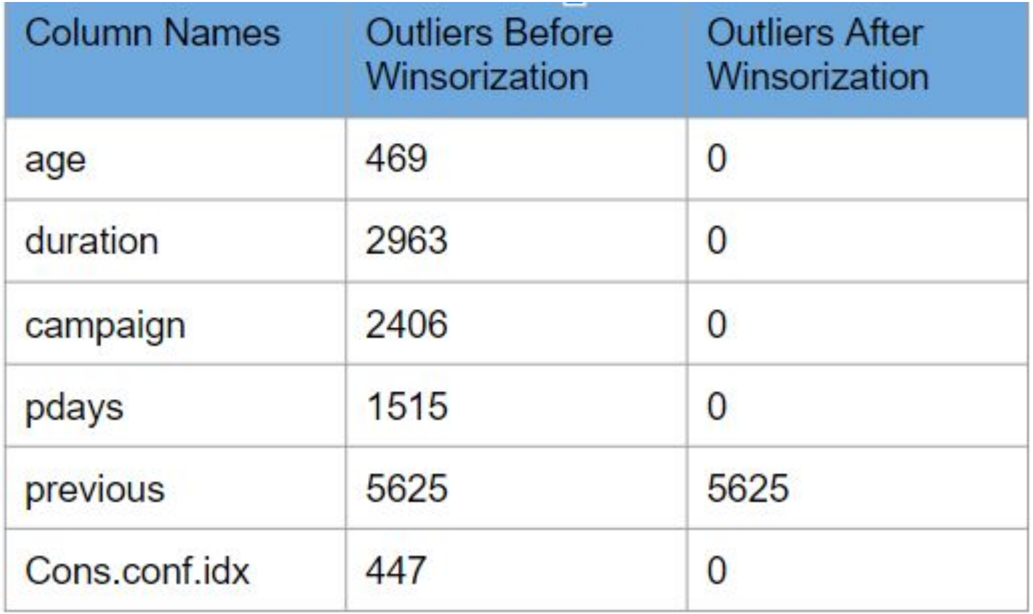
Customers who have successfully connected in previous campaign tend to subscribe. But we are not capturing those learners.

Lot of people in admin tend to subscribe to a term deposit.

Pipeline

Outlier treatment :

The Outliers in the continuous features were detected and treated using a method called **Winsorization**.



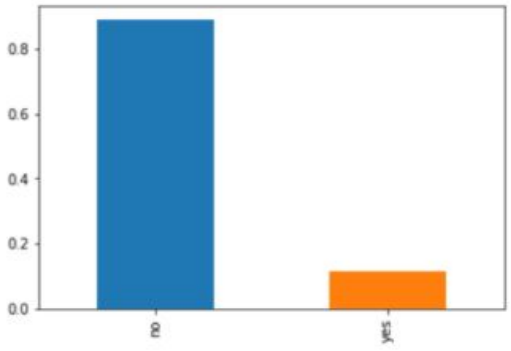
Column Names	Outliers Before Winsorization	Outliers After Winsorization
age	469	0
duration	2963	0
campaign	2406	0
pdays	1515	0
previous	5625	5625
Cons.conf.idx	447	0

Pipeline

Missing Values :

1. There were no missing values in the continuous features
2. The categorical features had missing values represented as **‘unknown’.** These were imputed using the mode values of the respective columns

Class Imbalance : The distribution of the target below shows a clear imbalance in the two classes.



Target variable : *y*

Pipeline

Feature Selection :

Following methods were used for feature selection :

- Correlation
- RFE

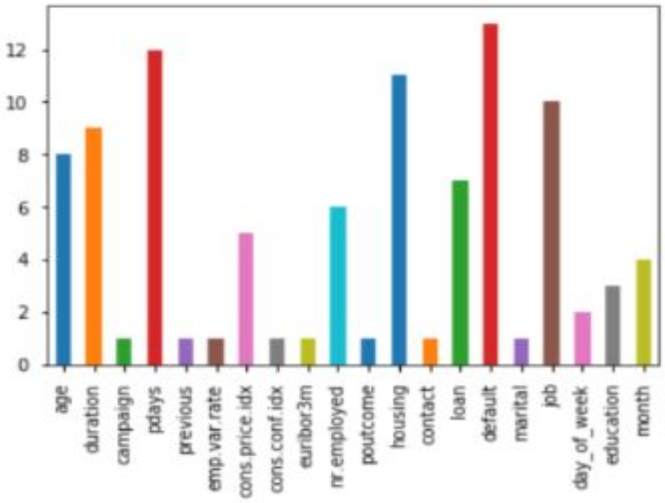
After estimating Pearson Correlation coefficients between continuous features, following features were dropped

- euribor3m
- nr.employed

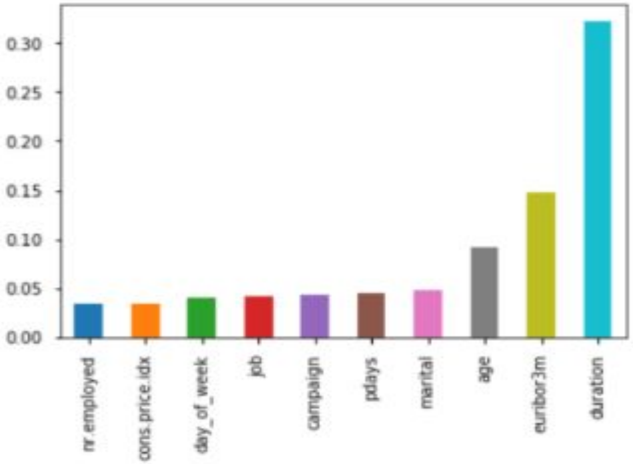
Pipeline

Feature Selection :

Recursive feature elimination was performed using Random forest and Logistic Regression as the estimators. Below are the feature importances obtained using both the methods



RFE



RFC

In both methods, duration looks like an important feature and captured accurately.

Models and Approaches

Three vanilla models were assessed without performing any hyperparameter tuning and without treatment of class imbalance of the target. The models were

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier

None of the three models were able to give an ROC_AUC score above 70%.

This called for performing hyperparameter tuning using Grid Search and also treatment of class imbalance using SMOTE for further improvement of the ROC_AUC score.

Models and Approaches

Models Assessed : The vanilla models used yielded the following results below.

Modelling Method	Precision	Recall	AUC_ROC
Logistic Regression	● 0 - 0.92 ● 1 - 0.64	● 0 - 0.97 ● 1 - 0.37	67.17 %
Random Forest Classifier	● 0 - 0.92 ● 1 - 0.61	● 0 - 0.98 ● 1 - 0.39	68.01 %
XGBClassifier	● 0.92 ● 0.69	● 0.98 ● 0.39	68.19 %

Model Tuning

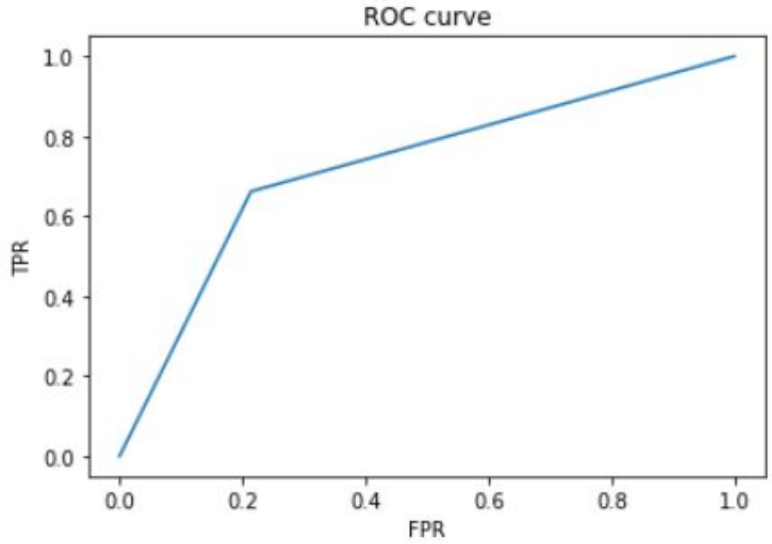
After performing hyperparameter tuning using Grid Search and treating imbalanced classes, the following results were observed on the features selected using RFE method. Also an Ensemble model of Logistic Regression and Random Forest Classifier was used as a hard voting classifer and yielded the following results:

Modelling Method	Precision	Recall	ROC_AUC
Logistic Regression	● 0 - 0.70 ● 1 - 0.75	● 0 - 0.79 ● 1 - 0.66	72.4 %
Random Forest Classifier	● 0 - 0.97 ● 1 - 0.89	● 0 - 0.88 ● 1 - 0.97	92.48 %
XGBoost Classifier	● 0 - 0.96 ● 1 - 0.91	● 0 - 0.91 ● 1 - 0.97	93.81 %
Ensembling	● 0 - 0.88 ● 1 - 0.95	● 0 - 0.96 ● 1 - 0.87	91.34 %

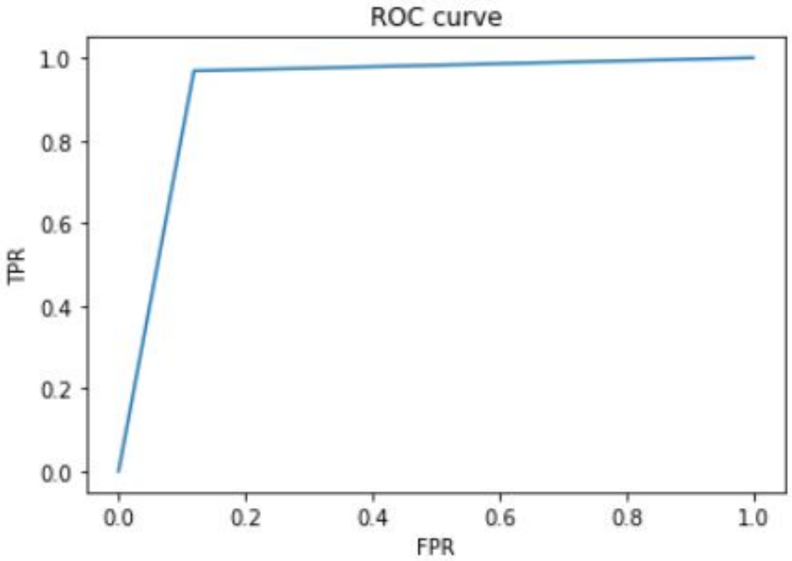
Evaluation & Results

Below are the AUC_ROC plots for the after hyperparameter tuning.

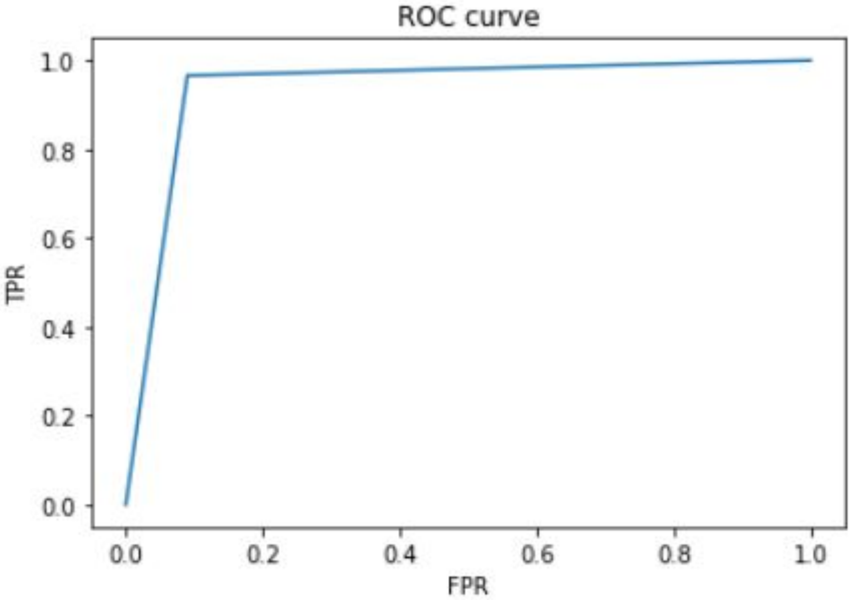
Logistic Regression



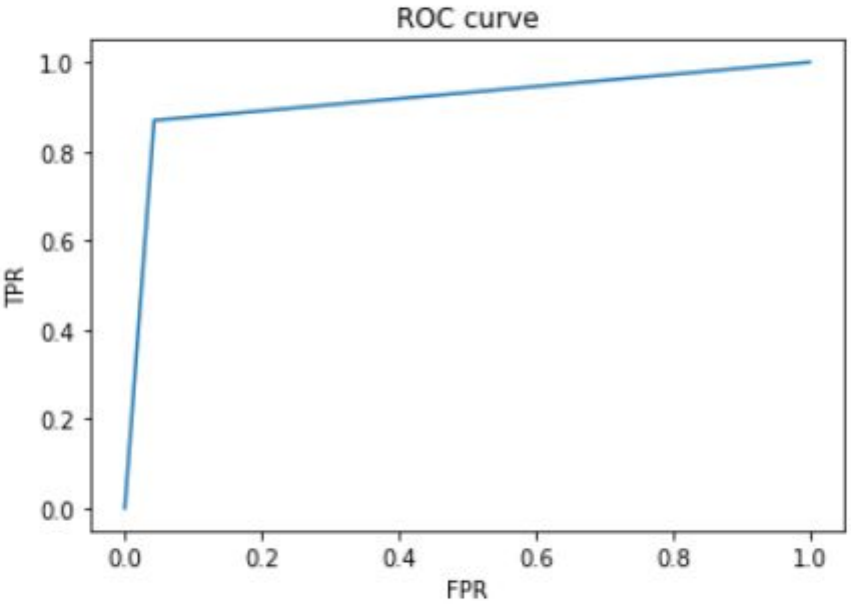
Random Forest



XGBoost Classifier



Ensemble model



Final Results

From the above observations and plottings it can be inferred that the best performing model was XGBoost giving an AUC_ROC score of 93.81 %. While XGBoost is used a lot, it is always prudent to start from simpler algorithms and then go to complex ones.

Confusion Matrix :

	Predicted Positive	Predicted Negative
Actual Positive	9981	1024
Actual Negative	398	10526

Insights & Decisions

- Customers to be targeted
    - Age : 30 – 50
    - Education : University, High School, Professional Courses
    - Job : Admin, Blue-collar, Technician
- Campaign Targets
    - Customers who were not targeted before
    - Customers successful in previous campaigns
    - Plan campaigns from May through August

Next Steps

If time permitted, could have tried the following :

- Better feature engineering
- An ensemble of different models
- A UI for a real user

Things to Remember

Max 20-25 slides for the entire team

Time: 10 mins for the team.

Put bullet points and pictures. (No code)

Think of it as a short pitch to the stakeholder.