SYMBIOSIS INSTITUTE OF OPERATIONS MANAGEMENT

DRISHTI

FactElytics

An Analytics Case Study Competition

INDIAN INSTITUTE OF MANAGEMENT CALCUTTA

Team Name : VPN

PRESENTED BY:

PRANJAL SHARMA— IIM Calcutta

SATYAM NEELMANI— IIM Calcutta

VATAN GOYAL — IIM Calcutta

EDA and Data Preparation | Modeling | Recommendation

8 numerical & 9 categorical Features

Demographical Information

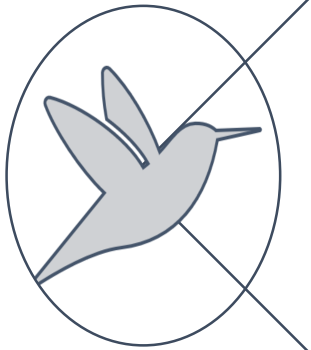
- Customer Age
- Education
- Job Type
- Marital Status

Banking Information

- Balance
- Default
- Personal Loan
- Housing Loan
- Term Deposit Subscribed (Target Variable)

Current and Prior Campaign Data

- Communication Type
- Day of Month
- Month
- Last Contact Duration
- Number of contacts in campaign
- Days passed since last campaign contact
- Number of times client was approached in prior campaign
- Outcome of the previous campaign



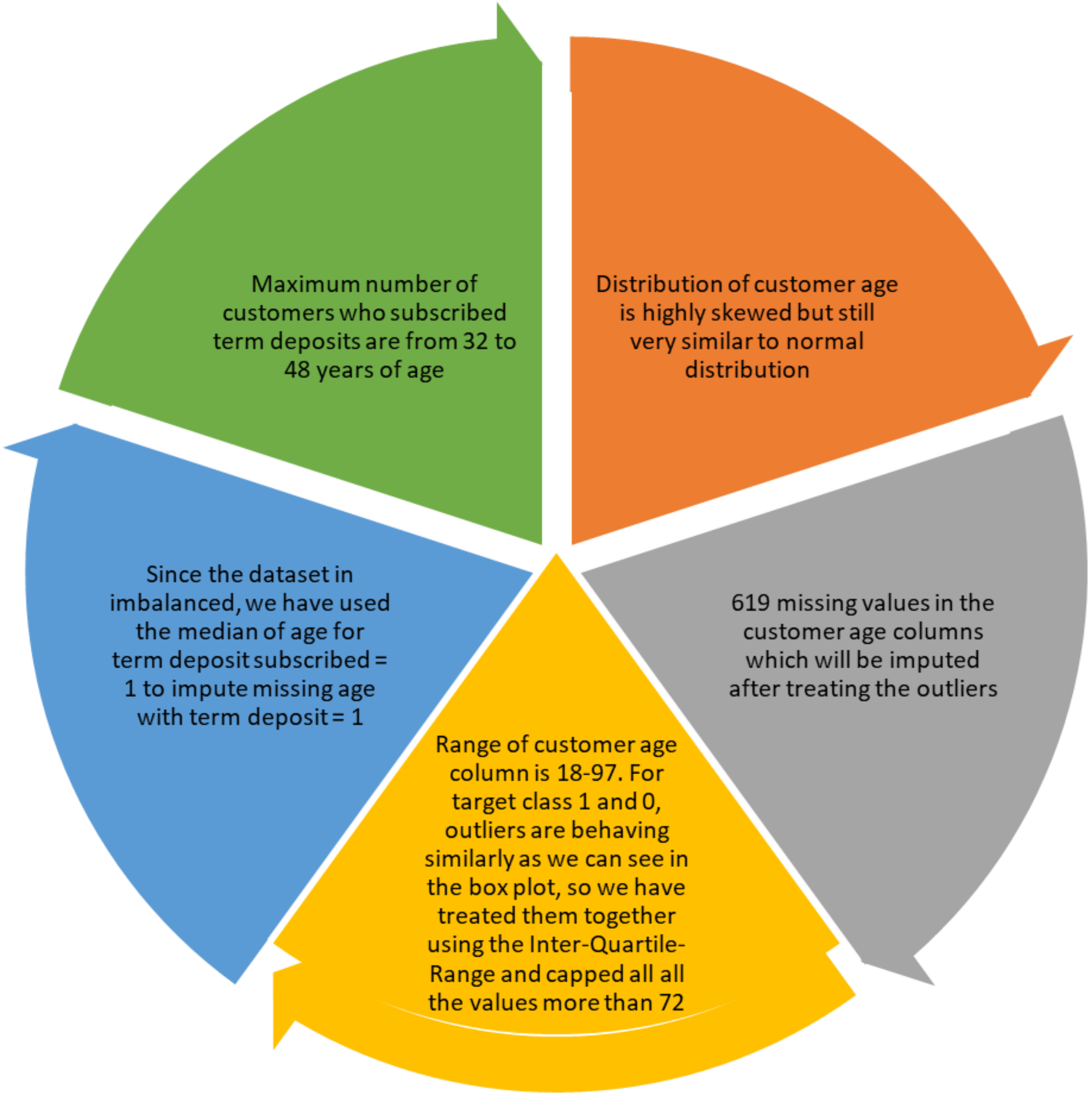
Bird Eye View

- Positive samples:- 3394
- Negative samples:- 28253
- Class Imbalance ratio- 11:89

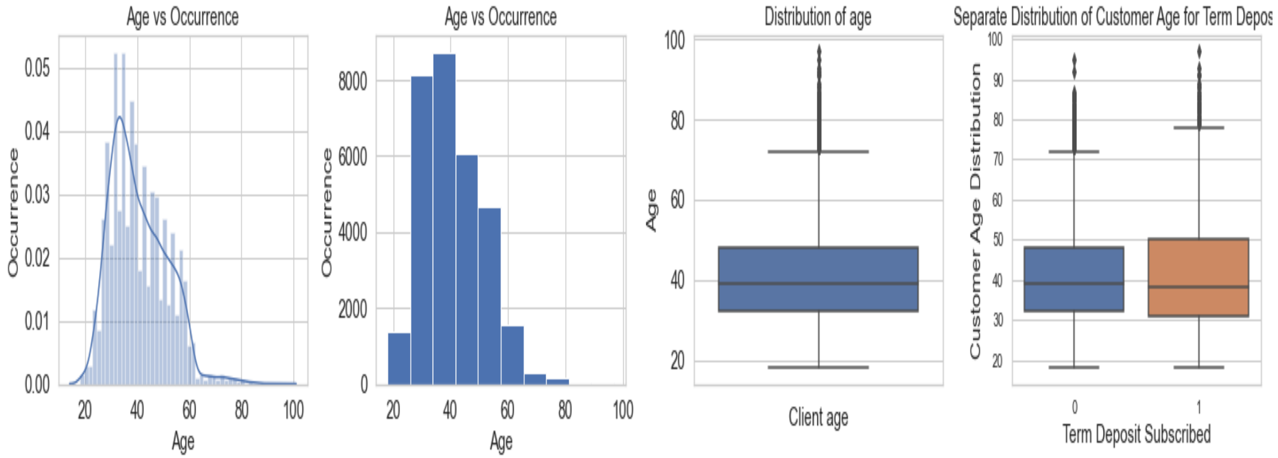
Data Preparation Steps

1. Outlier Treatment using interquartile range for numeric attribute.
2. Imputation of missing values using mean and mode.
3. One hot encoding on categorical data

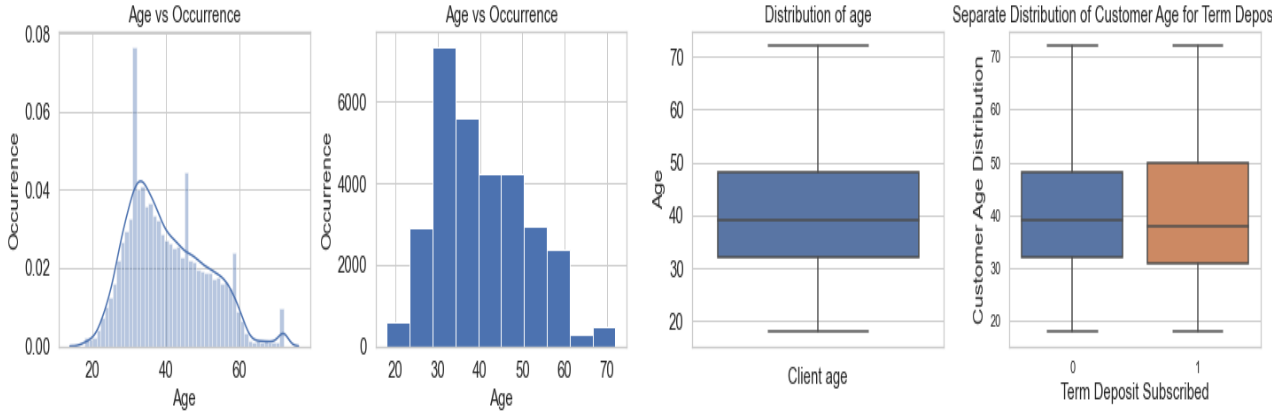
EDA and Data Preparation | Modeling | Recommendation



Customer Age Distribution before outlier treatment and missing value imputation



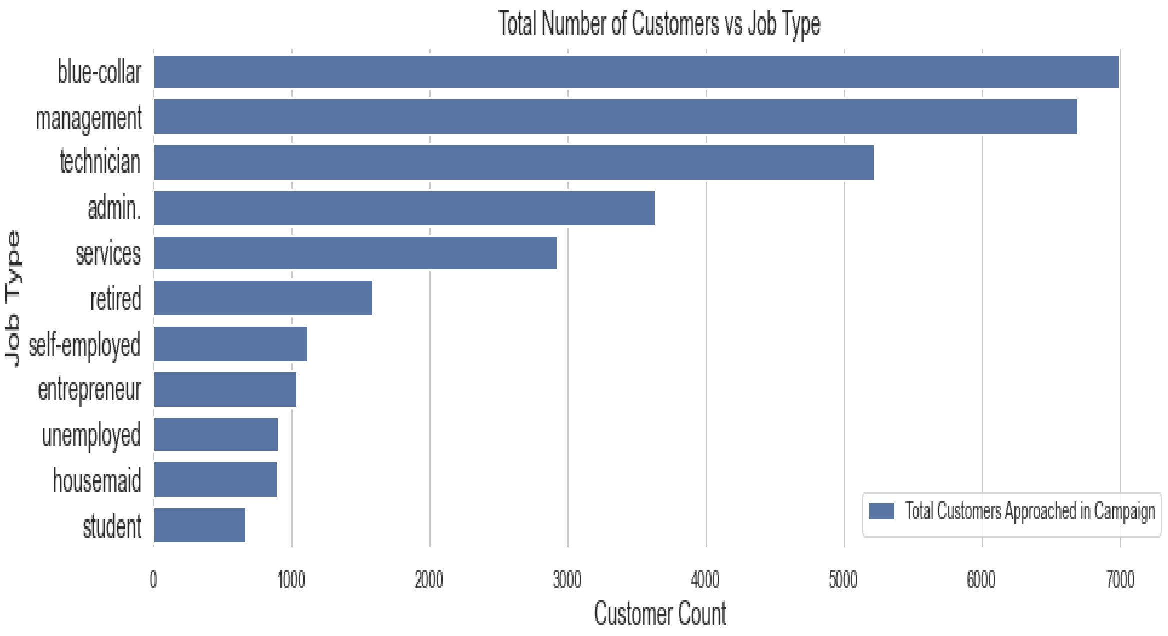
Customer Age Distribution after outlier treatment and missing value imputation

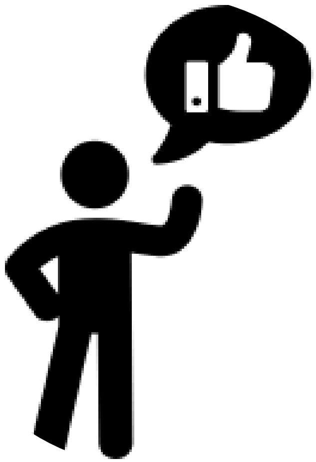


EDA and Data Preparation | Modeling | Recommendation

Inferences

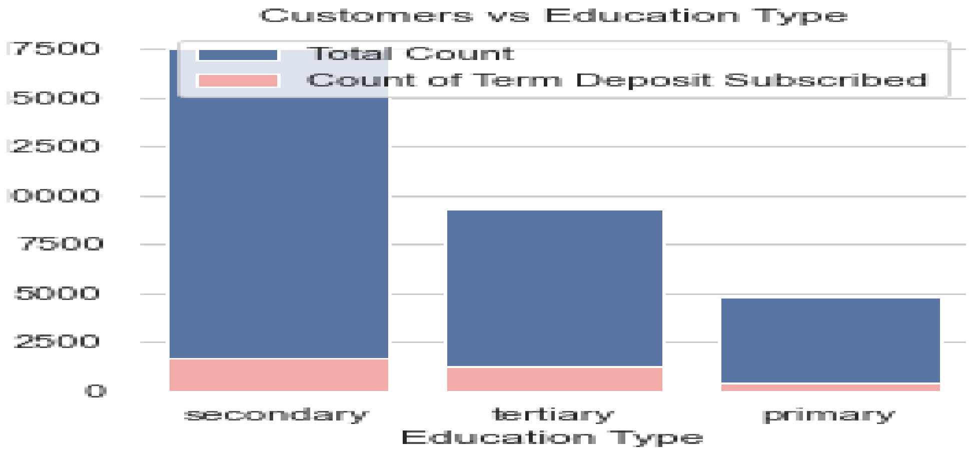
- Term Deposits subscribed heavily by management background clients, then technician and then blue-collar workers
- Blue collar workers were approached more than management background customers
- Tertiary education type customers are more likely to convert than secondary education type



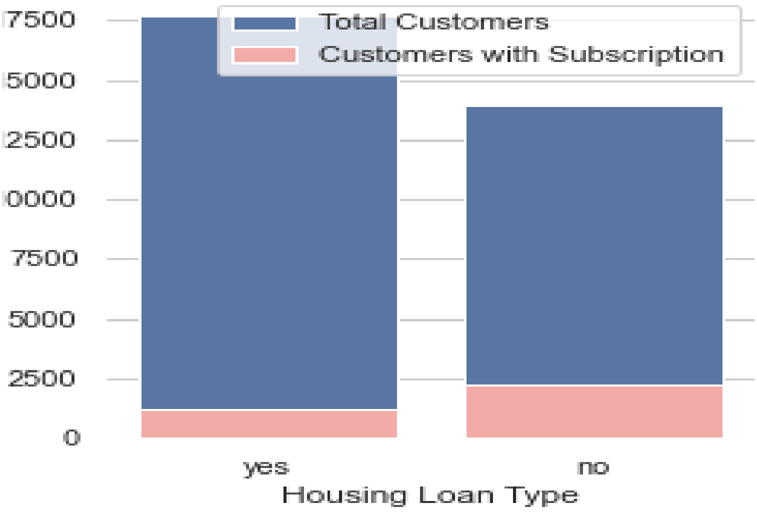


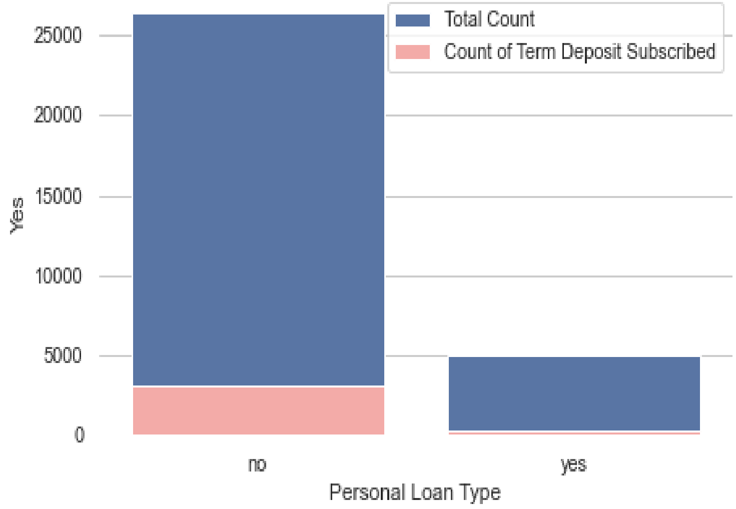
Recommendations

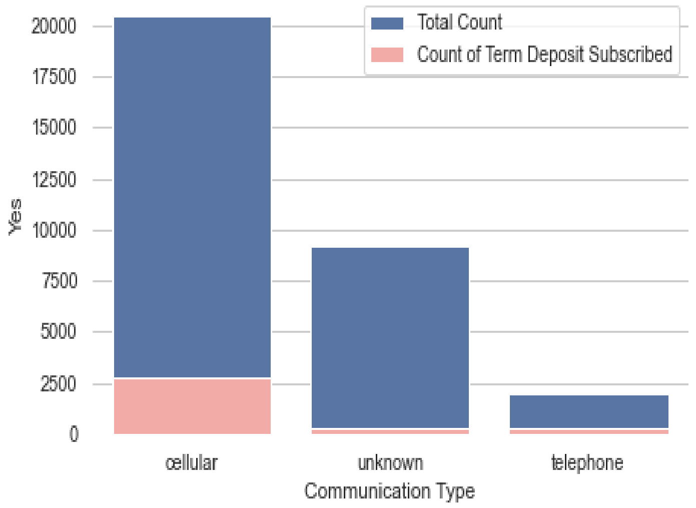
- More chances to increase the term deposit subscription if management people are focused more
- Top 3 recommendations in priority order will be Management, Technician and Blue Collar
- focusing more on tertiary education type customer as their ROI is more and they were not approached very much



EDA and Data Preparation | Modeling | Recommendation



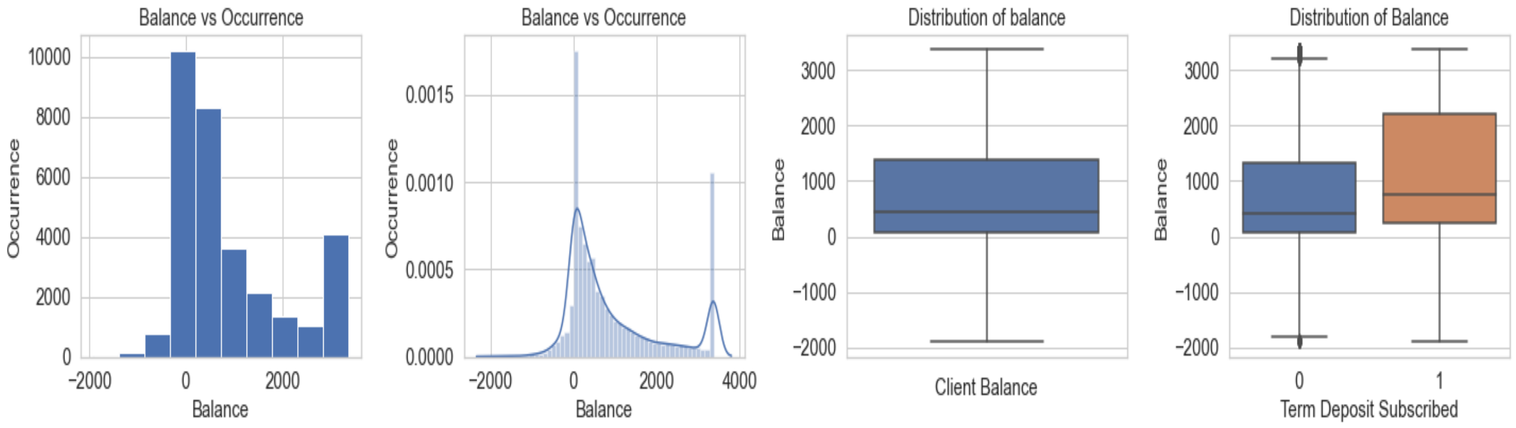




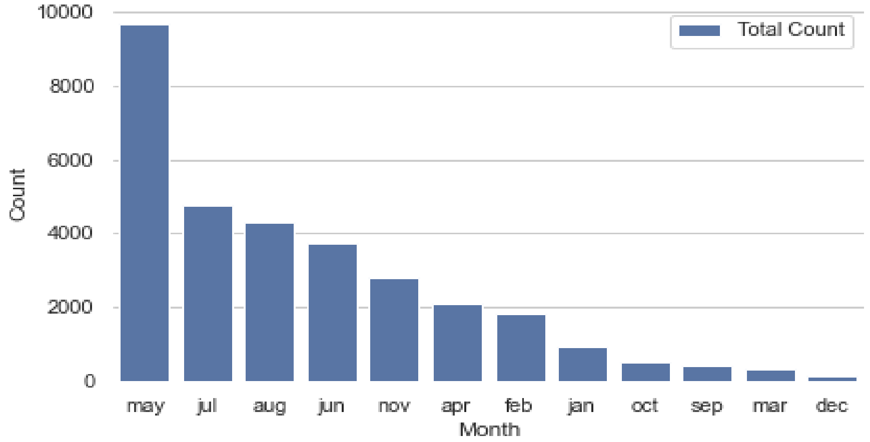
Inference

- Customers which have not taken housing loan are more likely to subscribe. Graph shows customers with housing loan were approached more, but focus should be changed to customers without Housing Loan
- 0 Missing Values. Unknown category accounts for 29% of the observations which is a very large number so we will consider it as a valid category in our analysis
- Communication done via cellular is more likely to engage customer and get term deposit subscription

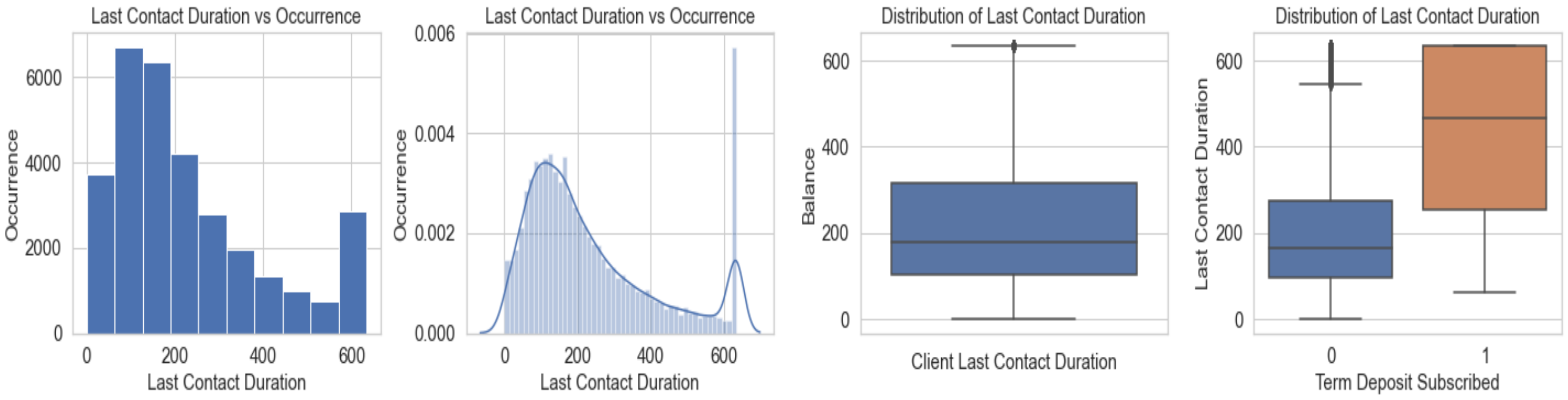
EDA and Data Preparation | Modeling | Recommendation



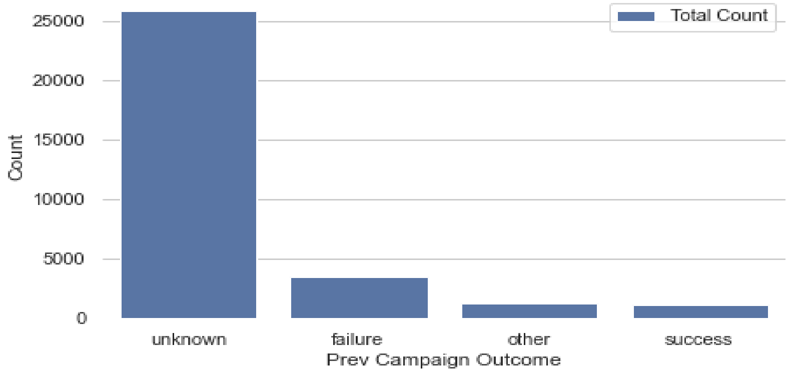
Graph-1 Balance column after treating outlier and missing value imputation



Graph-2 Distribution of month column



Graph-3 Last call duration column after outlier treatment and missing value imputation



Graph-4 Distribution of Previous campaign outcome

EDA and Data Preparation | **Modeling** | Recommendation

Modeling on Imbalance class

- Decision Tree, Bagging, Ada Boosting giving the highest accuracy
- These models will be biased towards the customers who do not have subscribed term deposit as there are 28154 customers who do not have subscribed term deposit
- There are only 3380 customers who have subscribed term deposit. So, these results can't be generalized

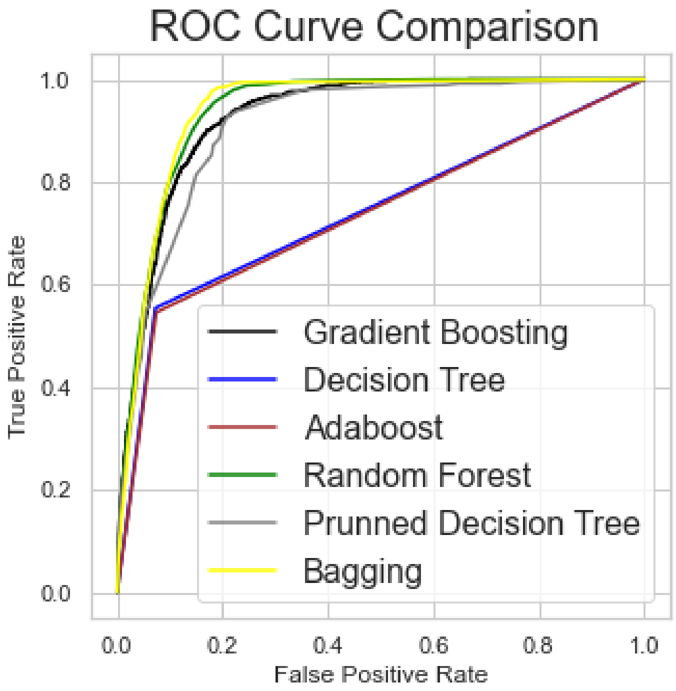
	Method	Binary F1 Score
1	Decision Tree	0.5123
2	Pruned Decision Tree	0.4965
3	Random Forest	0.4852
4	Bagging	0.5118
5	AdaBoosting	0.5180
6	GradientBoosting	0.4779
7	Ensemble	0.2837

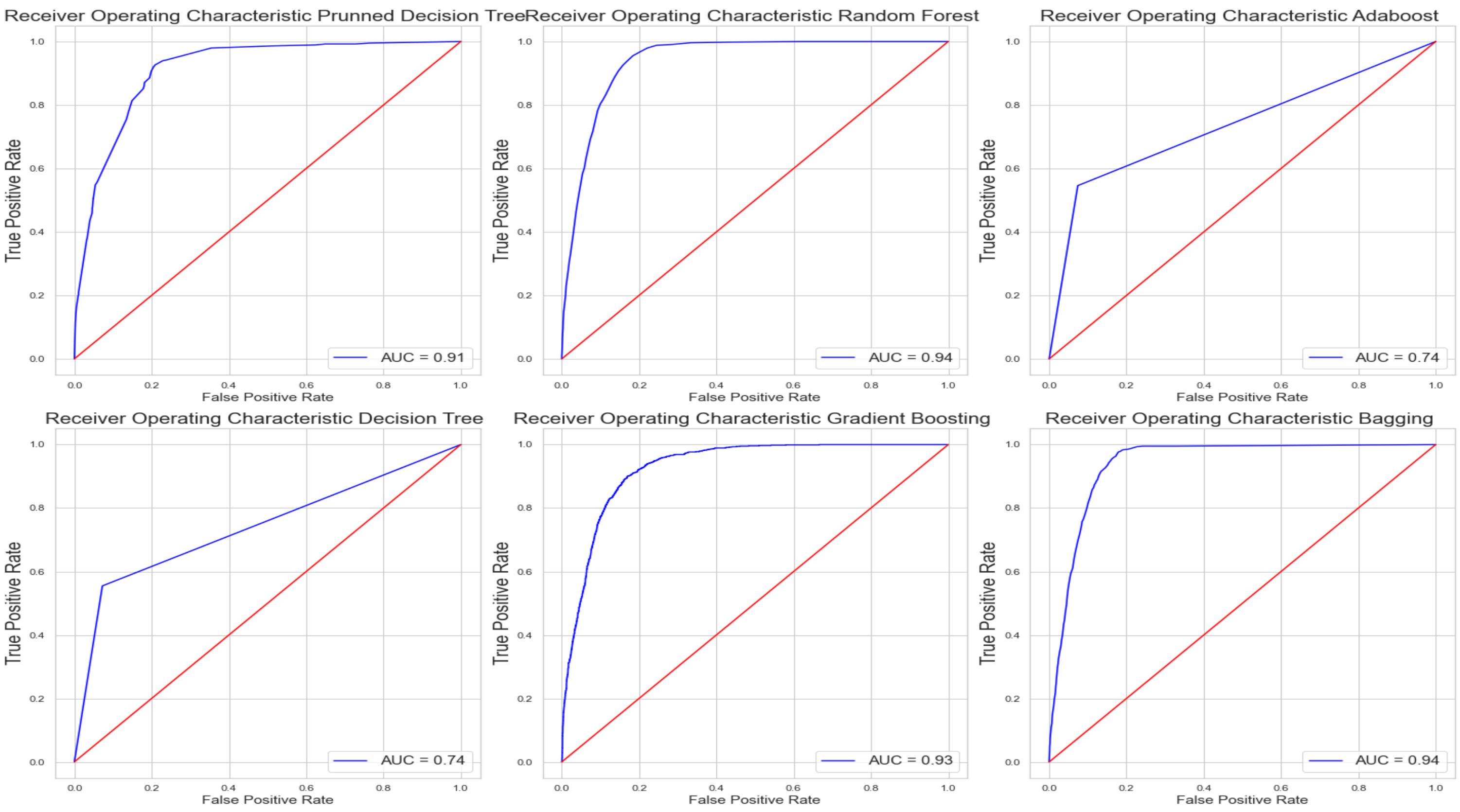
Modeling on balance data

- After managing the imbalancing the target column and scaling the columns, we are getting the highest accuracy using Random Forest and Bagging Models
- These models will perform well as compared to imbalanced class models

	Method	Binary F1 Score
1	Decision Tree	0.5154
2	Pruned Decision Tree	0.5172
3	Random Forest	0.5659
4	Bagging	0.5970
5	AdaBoosting	0.5066
6	GradientBoosting	0.5847
7	Ensemble	0.5930

EDA and Data Preparation | **Modeling** | Recommendation





1. Performance of Bagging and RandomForest is very close, which cover the maximum area under the curve.
2. AUC Bagging is 0.94 and F1-Binary score is 0.5970.

EDA and Data Preparation | Modeling | **Recommendation**

Customer Segment

- Company should focus on Management and technician more than blue collar and increase focus towards students.
- Increase focus on tertiary educated people as they give good ROI.
- Customers without housing loan should be prioritized over customer with housing loan

Top 15 Features business should focus on

- customer_age
- balance
- day_of_month
- last_contact_duration
- num_contacts_in_campaign
- num_contacts_prev_campaign
- job_type_admin.
- job_type_blue-collar
- job_type_entrepreneur
- job_type_housemaid
- job_type_management
- job_type_retired
- job_type_self-employed
- job_type_services
- job_type_student

Thank you