Capstone project on:

Bank Marketing Effectiveness

Presented by:

Subhadip Ghosh

Journey Roadmap:

- Problem Statement
- Discussing our dataset
- Basic Data Inspection
- Exploratory Data Analysis
- Conclusions from EDA
- Feature Engineering
- Feature Selection
- Handling Imbalance of our Response Variable
- Fitting and Evaluating our model
- Discussing metrics for imbalanced classification problems
- Conclusion

Problem Statement:

Our dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaign were based on phone calls. The product of the banking institution is 'Bank Term Deposit'. Our response feature is named 'subscribed' and it has two class labels:

- Yes: meaning the client has subscribed
- No: meaning the client has not subscribed to the product

We have modified our class labels to 0 and 1 corresponding to 'No' and 'Yes' respectively.

Our aim of this model is to predict the 'subscribed' column to predict whether a client will subscribe to the 'Bank Term Deposit' or not.

Discussing our dataset:

- Our dataset has 45211 records, and we have 15 features which determines our response variable 'Subscribed'.

Discussing some of our features which are quite hard to interpret:

- **Month,Day :** Corresponding Date when the customer was last contacted
- **Campaign:** Number of times the client has been called for selling the product during this campaign.
- **Pdays:** Number of days that have passed by after the client was last contacted.
- **previous:** number of times the client has been called before this campaign.
- **poutcome:** outcome of the previous marketing campaign

Basic Data Inspection:

- Our dataset did not have any null values
- Our response variable is heavily imbalanced. That is the number of clients who have subscribed to the 'Bank Term Deposit' are much lesser in counts than those who haven’t.
- Some features like ‘pdays’ have very less variance, we will be inspecting these features in the further slides.
- For the ease of understanding our features through Exploratory Data Analysis(EDA) we divided our features into:
  - Categorical Features
  - Numerical Features
    - Continuous Features
    - Discrete Features

Exploratory Data Analysis

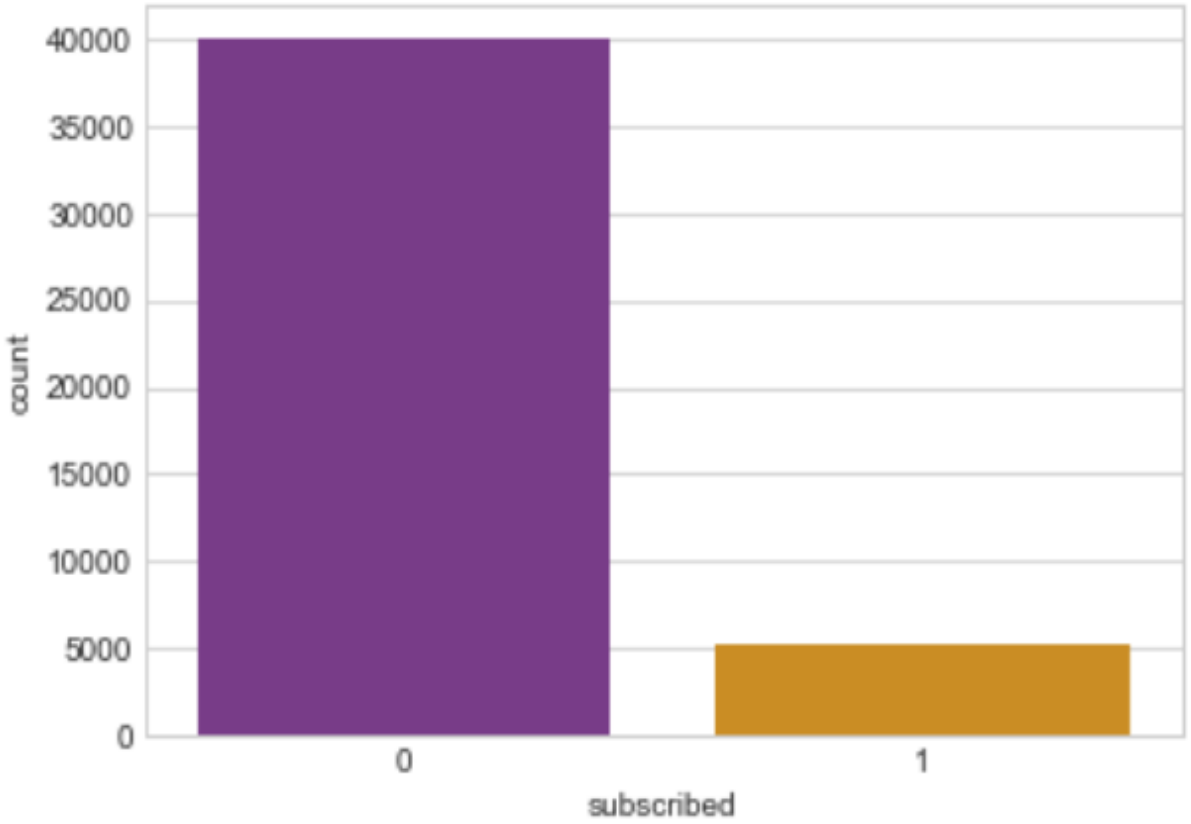
EDA (Continuous and Discrete Features)

EDA (Categorical Features)

Exploratory Data Analysis(EDA):

Counts of our class labels

- **0 : Not Subscribed**
- **1 : Subscribed**

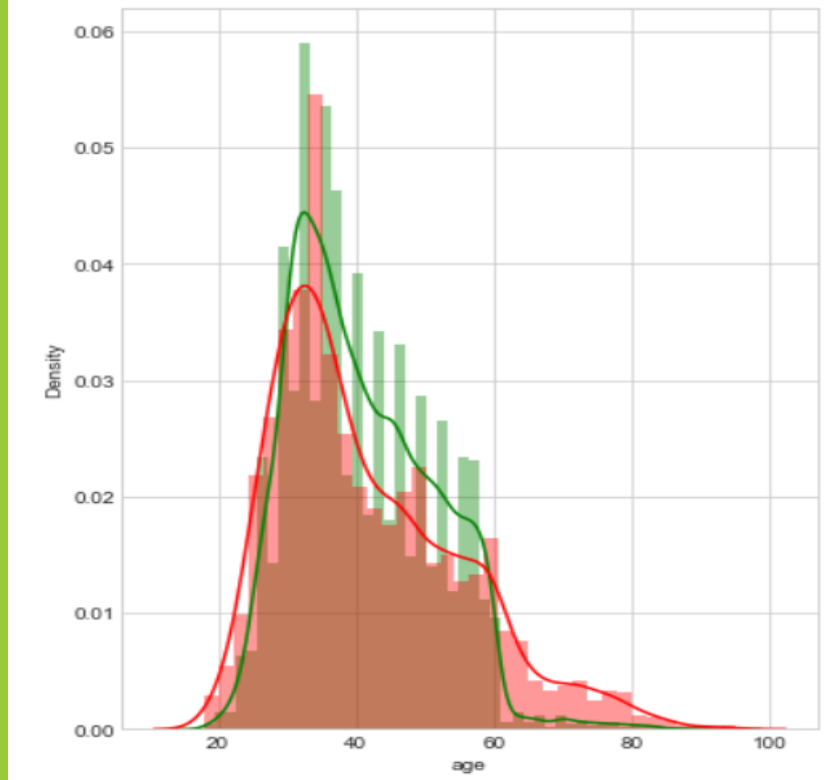


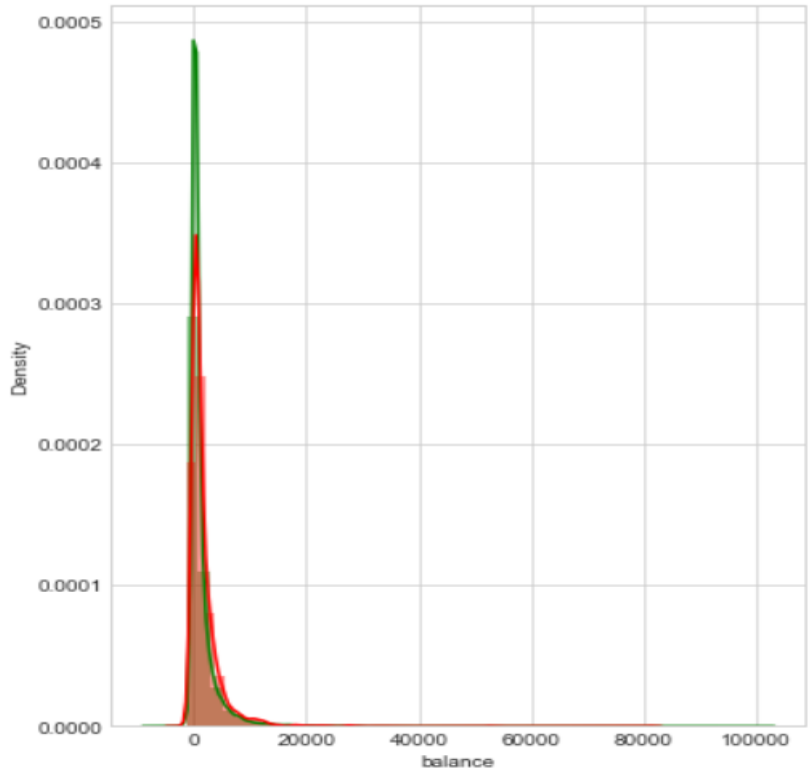
►Conclusion:

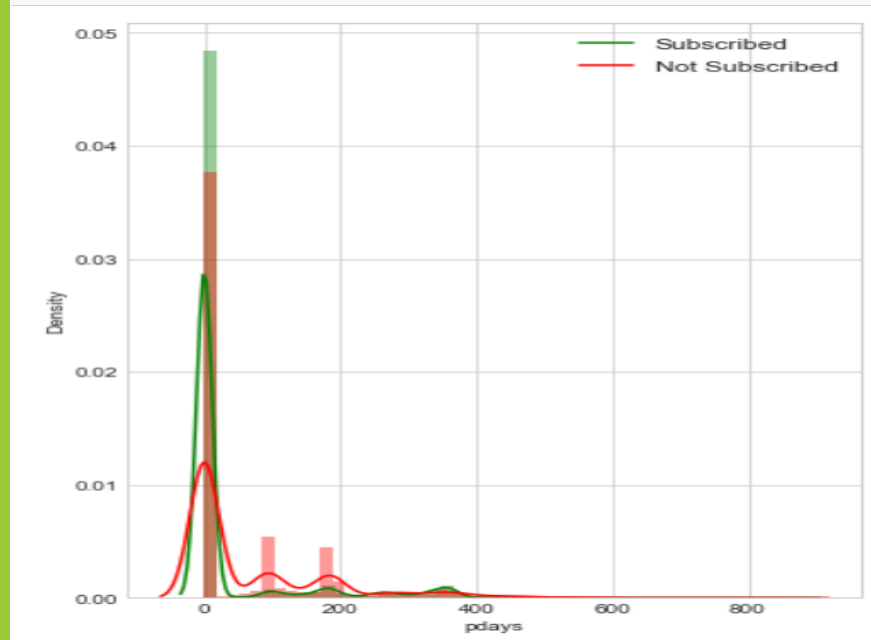
Our response variable is heavily 'imbalanced' with almost 40000 clients corresponding to label 0, meaning they haven't 'subscribed' and only a little over 5000 clients have subscribed to the 'Bank Term Deposit'

Exploratory Data Analysis(EDA):

- **Numerical Features (Continuous)**
- **Distribution Plot**



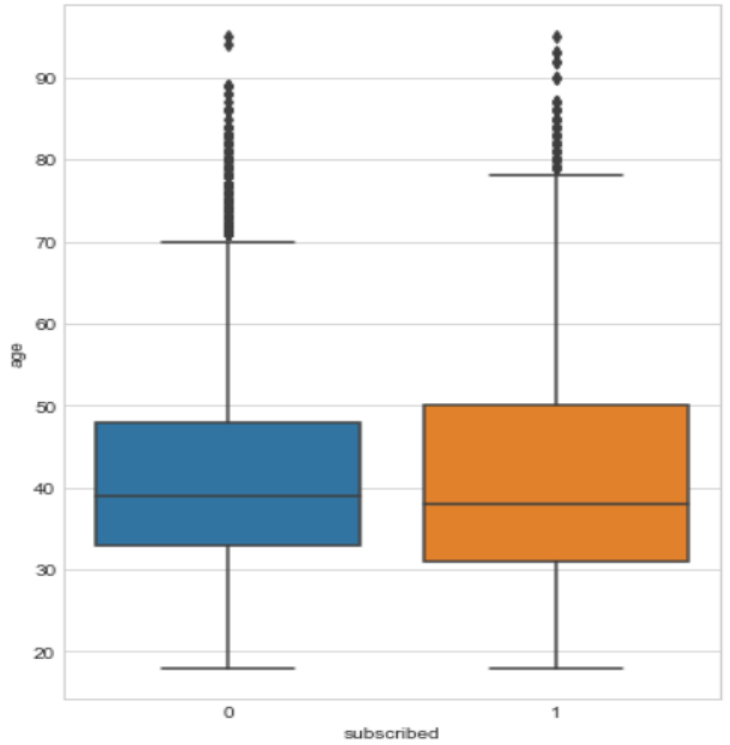


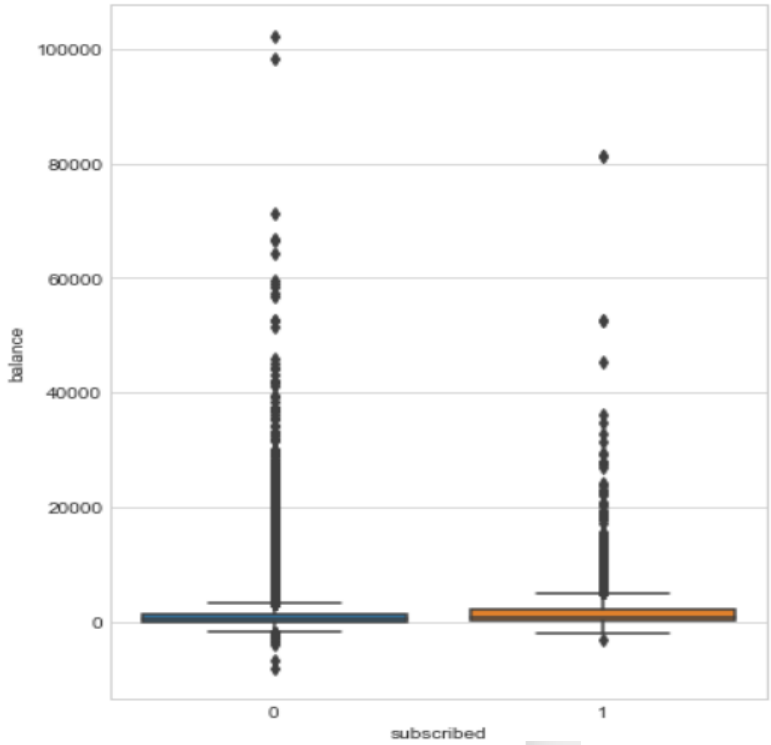


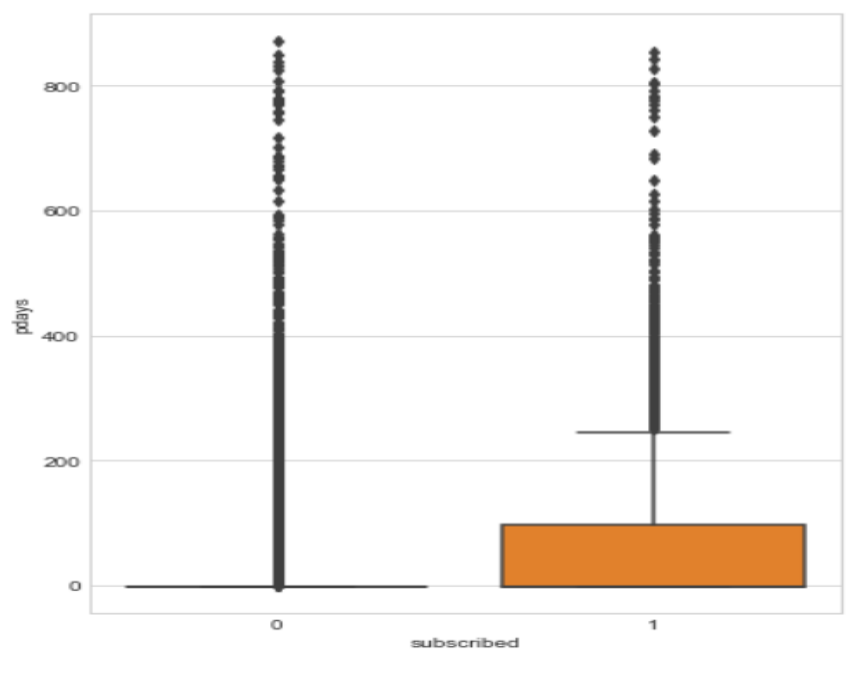
- **Not Subscribed**
- **Subscribed**

Exploratory Data Analysis(EDA):

- **Numeric Features (Continuous)**
- **Boxplot Distribution**

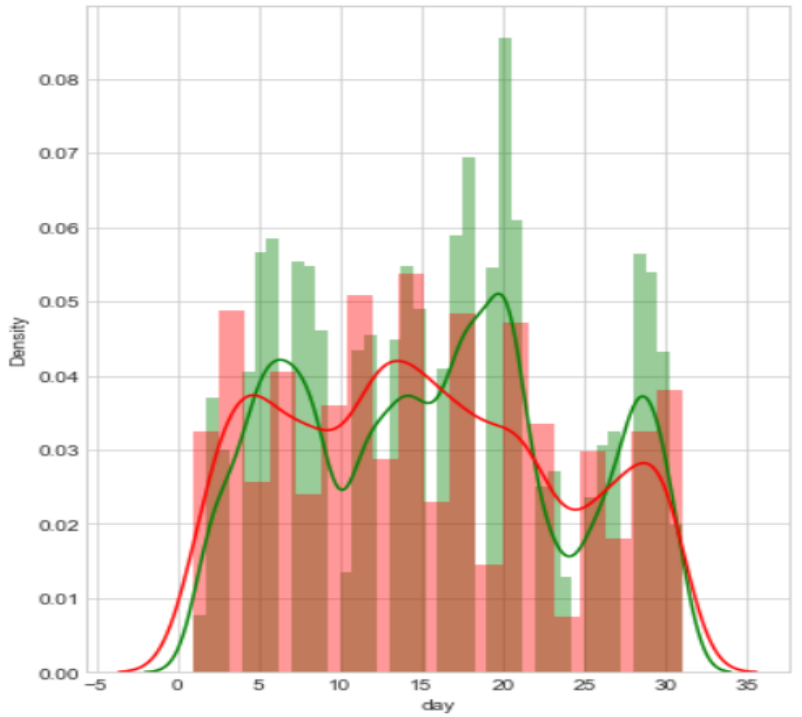


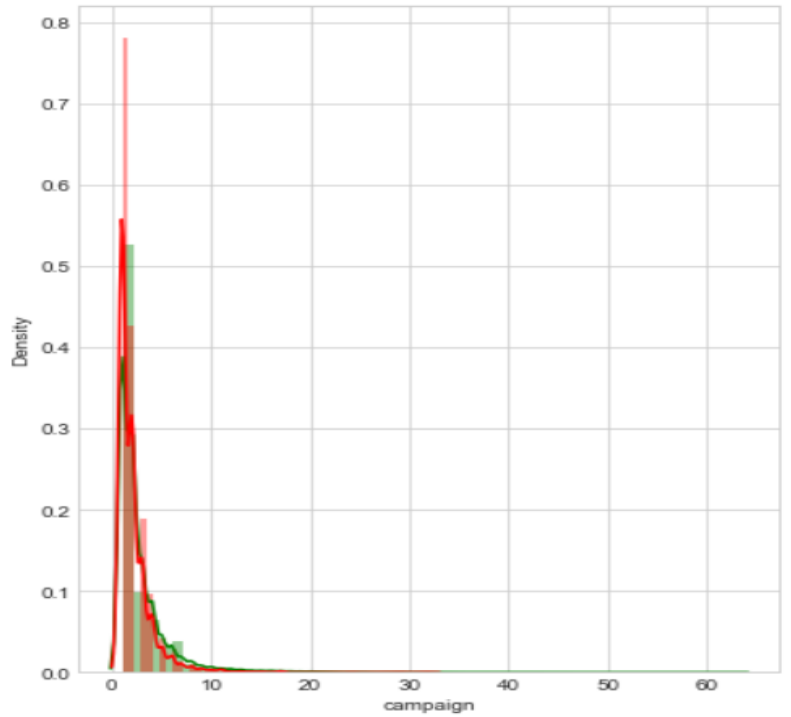


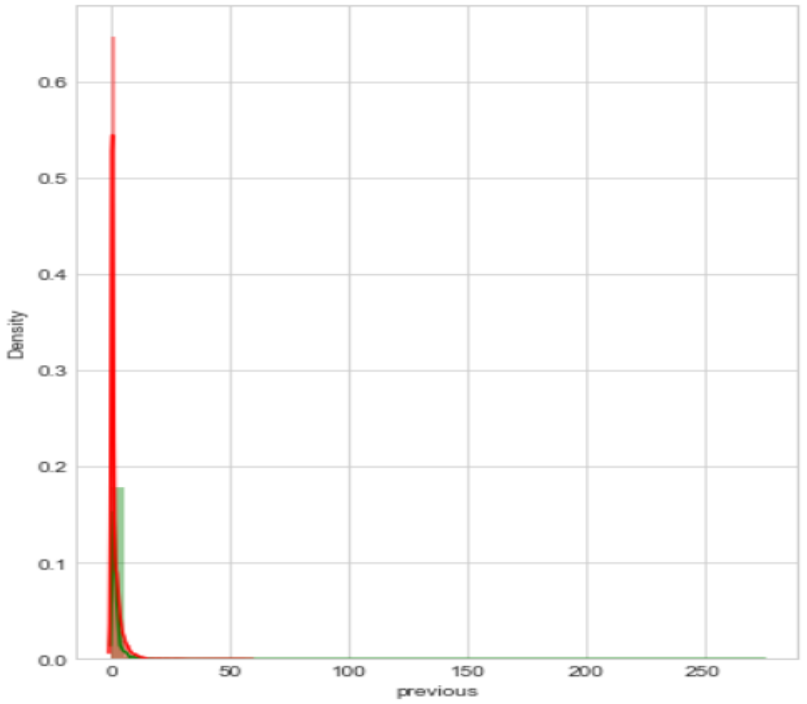


Exploratory Data Analysis(EDA):

- **Numerical Features(Discrete)**
- **Distribution Plot**



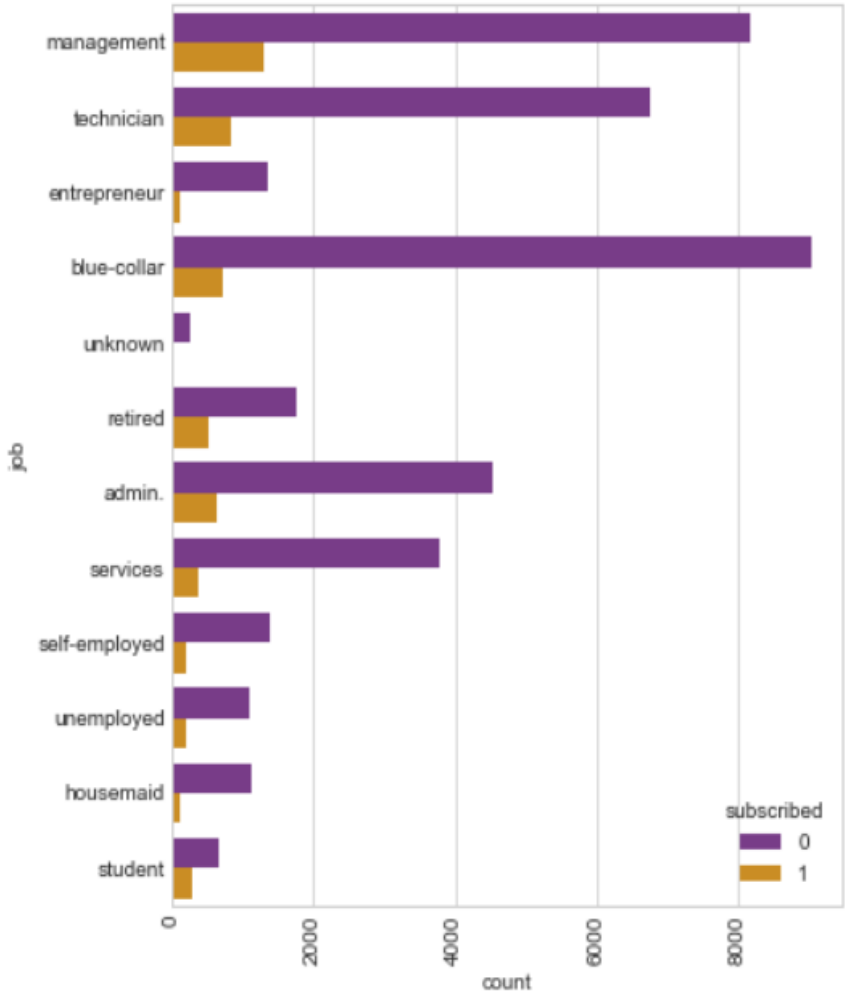


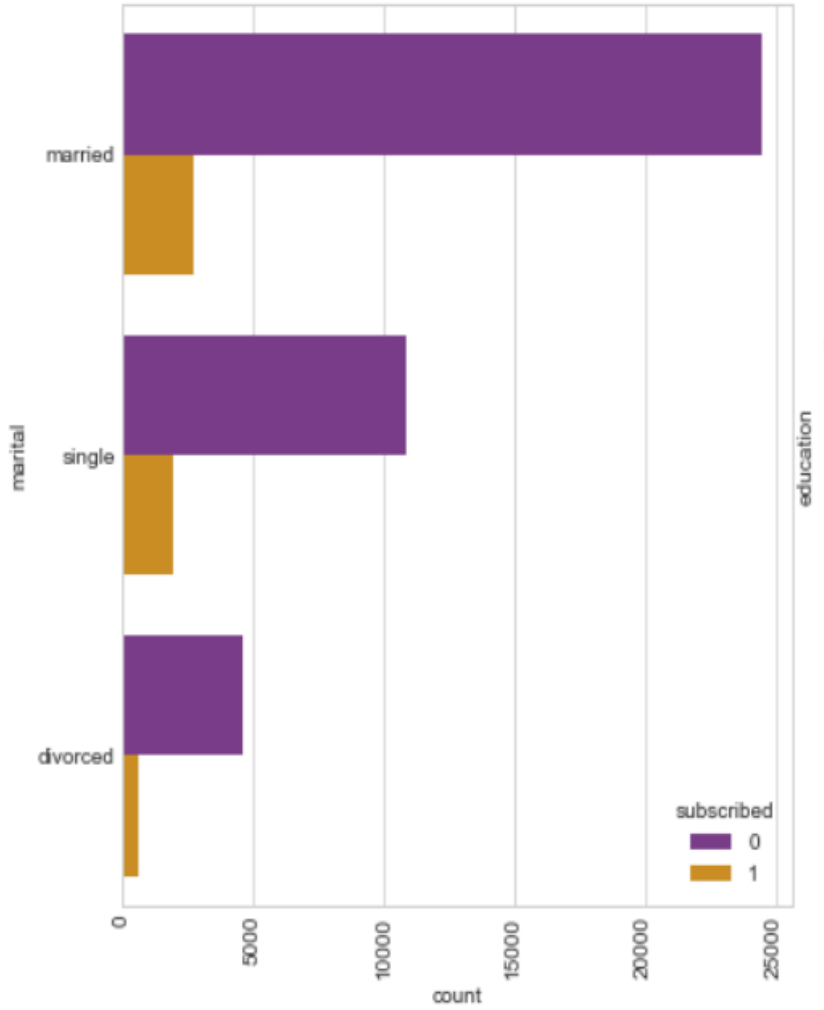


- **Not Subscribed**
- **Subscribed**

Exploratory Data Analysis(EDA):

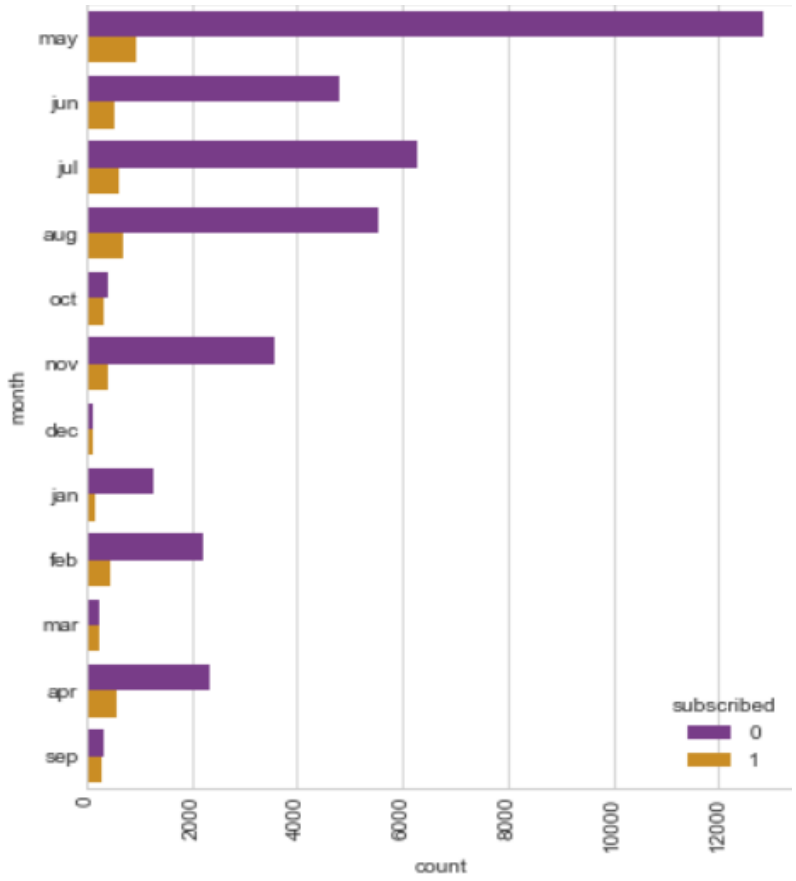
- **Numerical Features(Discrete)**
- **Countplot**

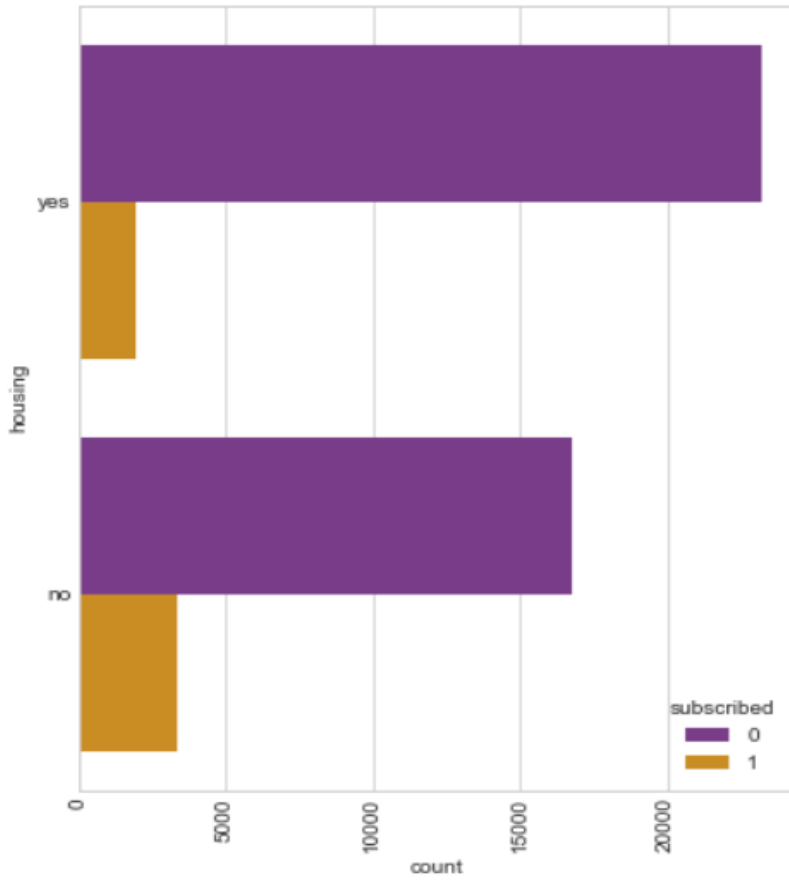




Exploratory Data Analysis(EDA):

- **Categorical Features**
- **Countplot**





Some questions still needs to be answered!!

Let’s take a look at some of the questions that still need more research:

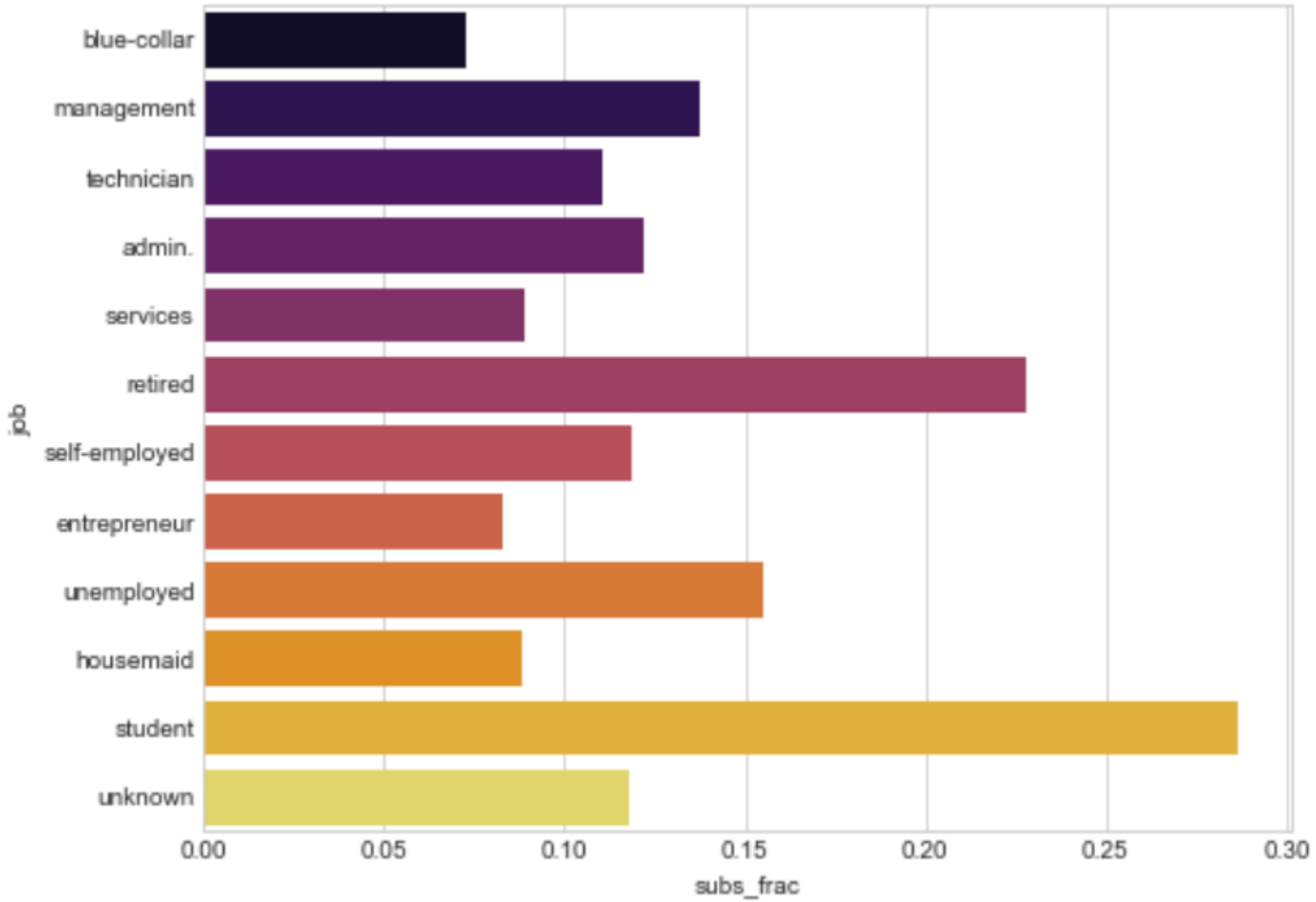
- Which job sector have the banks targeted the most in order to have better number of subscriptions?
- What are the average ages of clients who have subscribed the most?
- Why is our feature ‘pdays’ having abnormal distributions in our dataset? Is it even a continuous feature?
- What is the success rate of the marketing campaign corresponding to each job sector?

Exploratory Data Analysis (EDA):

The number of subscriptions by clients of each job sector and their average ages:

Exploratory Data Analysis (EDA):

Success Rate of subscriptions corresponding to clients from each job sector



Feature Engineering Section:

- Fixing Irregularities in features
- Encoding our categorical features
  (One hot encode, label encode, target encode)
- Checking Statistical Significance of features
  (Using Chi-Square Test)
- Outlier Treatment

Fixing irregularities of certain features:

- As we noticed in the distribution plots, the features ‘campaign’, ‘pdays’ and ‘previous’ along with some categorical features showed properties of abnormal distribution.

- We would be operating on these columns by grouping all the values which contribute to less than 1 percent of our total number of samples, and then putting them in one value naming it 'Rare value’. We would do the same for the feature ‘Month’ as well.

Target Encoding on ‘Job’ Feature:

Giving a sample of how we have target encoded features:

	job	total_subs	avg_age	subscribed	subs_frac
0	blue-collar	9732	40.044081	708	0.072750
1	management	9458	40.449567	1301	0.137556
2	technician	7597	39.314598	840	0.110570
3	admin.	5171	39.289886	631	0.122027
4	services	4154	38.740250	369	0.088830
5	retired	2264	61.626767	516	0.227915
6	self-employed	1579	40.484484	187	0.118429
7	entrepreneur	1487	42.190989	123	0.082717
8	unemployed	1303	40.961627	202	0.155027
9	housemaid	1240	46.415323	109	0.087903
10	student	938	26.542644	269	0.286780
11	unknown	288	47.593750	34	0.118056

$$\text{Success Ratio} = \frac{\text{Subscribed}}{\text{Total\_subs}}$$

- We have calculated the total number of times clients from each profession has been contacted and stored it in ‘total_subs’ column.
- Calculated the number of times clients from each profession have subscribed and stored in ‘subscribed’ column.
- Calculated the success ratio and stored in ‘subs_frac’ column

Checking Statistical Significance of discrete features using Chi-Square Test:

	f value	p value
job	13.653570	2.198230e-04
education	35.069469	3.181501e-09
day	6.436125	1.118223e-02
month	182.626780	1.293928e-41
campaign	2.355764	1.248207e-01
previous	33.222131	8.220955e-09
mar_married	65.352757	6.262225e-16
mar_single	130.835717	2.689699e-30
def_yes	22.313875	2.315277e-06
house_yes	388.949715	1.401285e-86
loan_yes	176.516137	2.793375e-40
cont_telephone	8.342166	3.873539e-03
cont_unknown	733.354934	1.669907e-161
pout_other	44.287113	2.835768e-11
pout_success	4113.000571	0.000000e+00
pout_unknown	230.279723	5.180119e-52
pdays_Rare_val	1033.864210	7.825795e-227

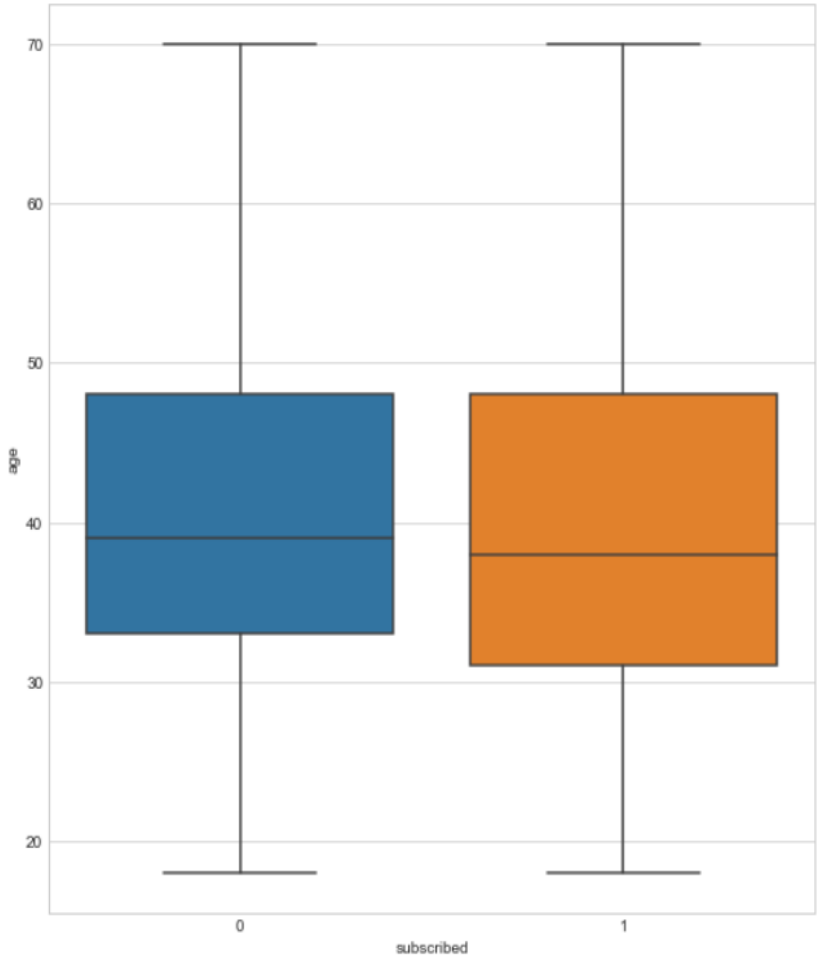
- We imported ‘chi2’ from the scikit learn library.
- ‘Chi2’ returns two values for each feature with respect to the response feature. Those are the ‘f-values’ and ‘p-values’ respectively.

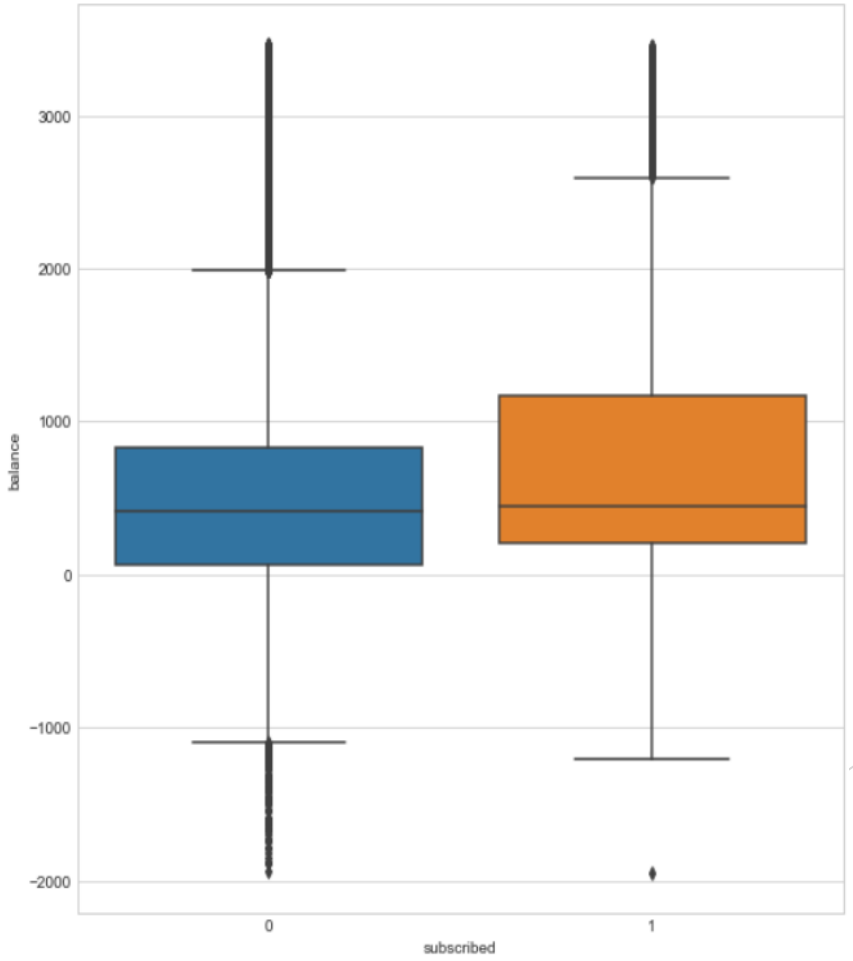
Null Hypothesis and Alternate Hypotheis Assumption:

- Null Hypothesis(Ho) assumes there’s no relation between the features and the response feature ‘subscribed’.
- Alternate Hypothesis(Ha) assumes that the relation between the features and the response feature is statistically significant.
- We will reject the null hypothesis if p-value of a feature is less than alpha (0.05).

Outliers Treatment:

- Used the IQR (Inter-Quartile Range) method.
- Replaced Outliers of a feature by the value of their respective medians.





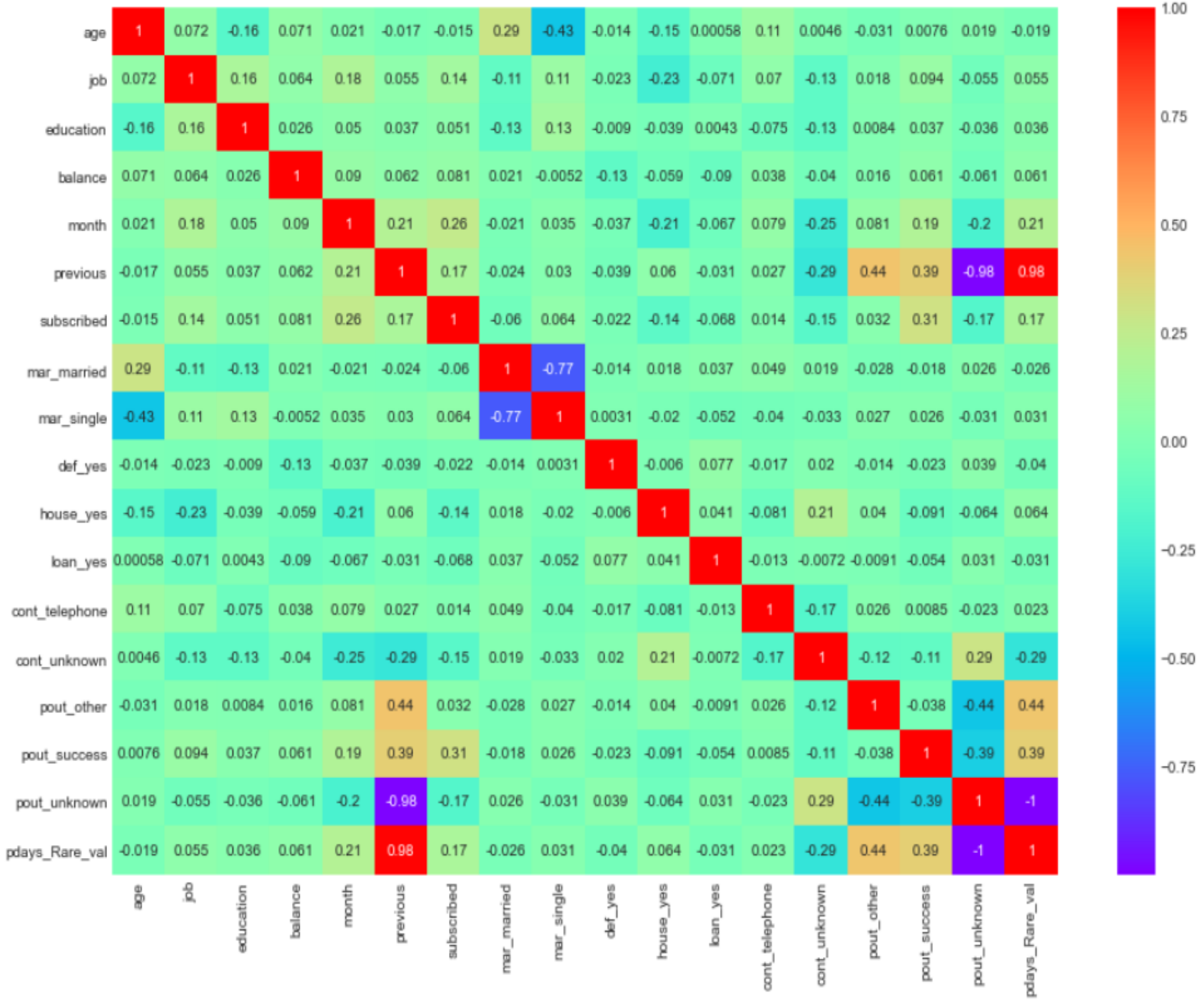
Feature Selection

Correlation Heatmap

ExtraTrees Classifier

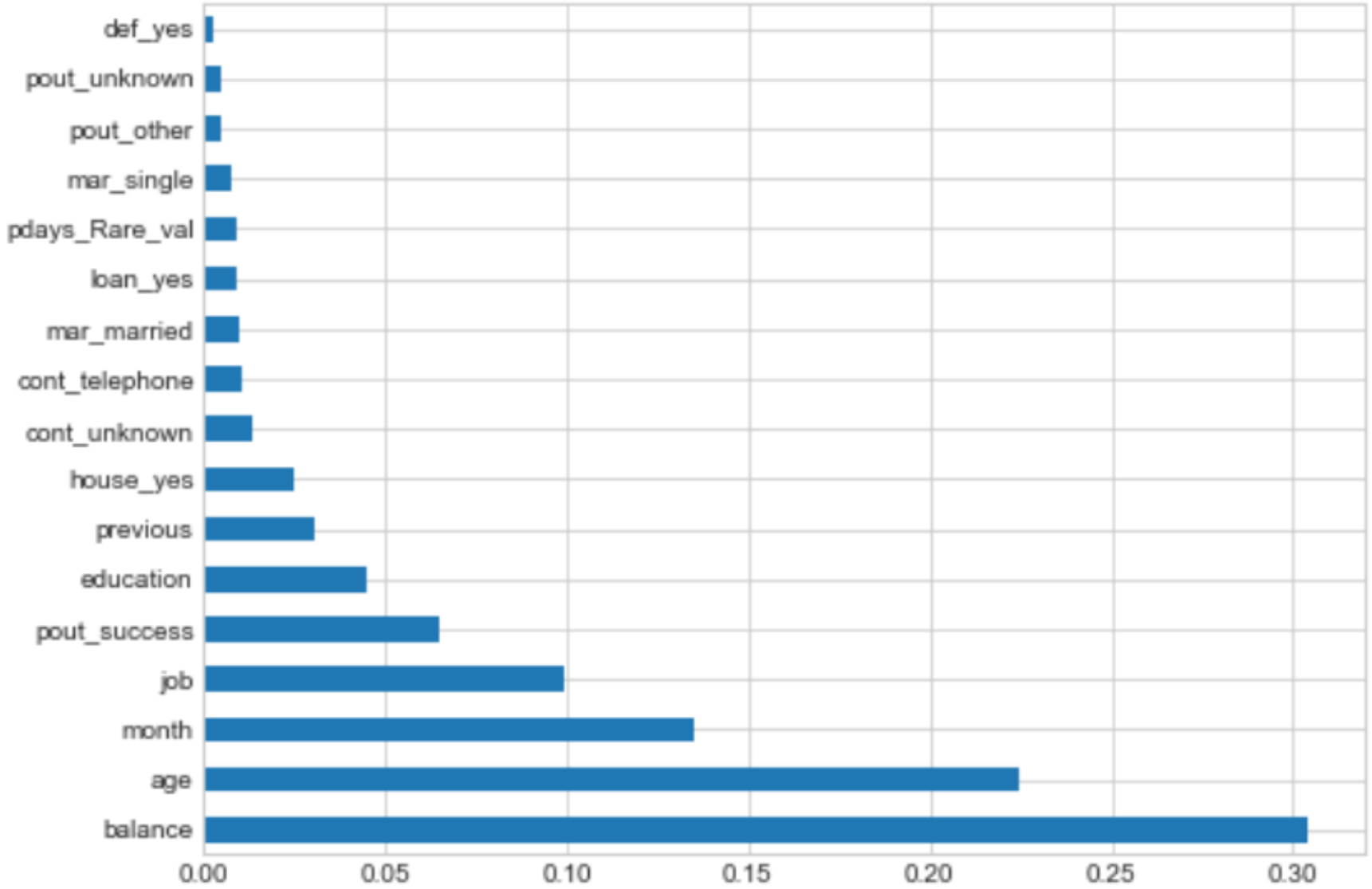
Correlation Heatmap

- There are some strong correlations between in our dataset
- Pdays_rare_val and ‘previous’ are some of the features showing the strongest correlations



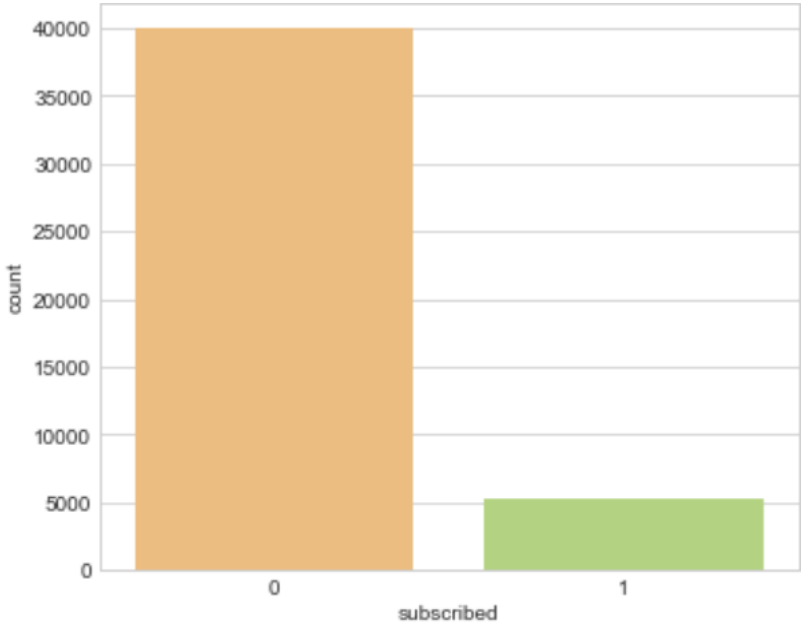
Feature Selection using ExtraTreesClassifier:

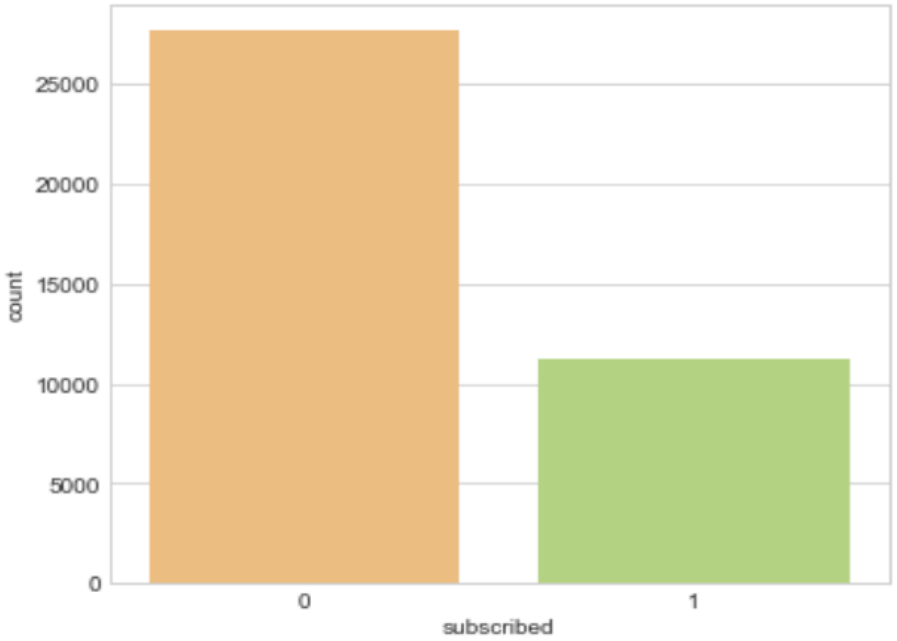
- Considered the top eight features of our dataset



Dealing with class Imbalance using SMOTEENN:

Counts of our class labels initially





Counts of our class labels after applying SMOTEENN

Fitting and Evaluation of our Model

Using Random Forest Classifier

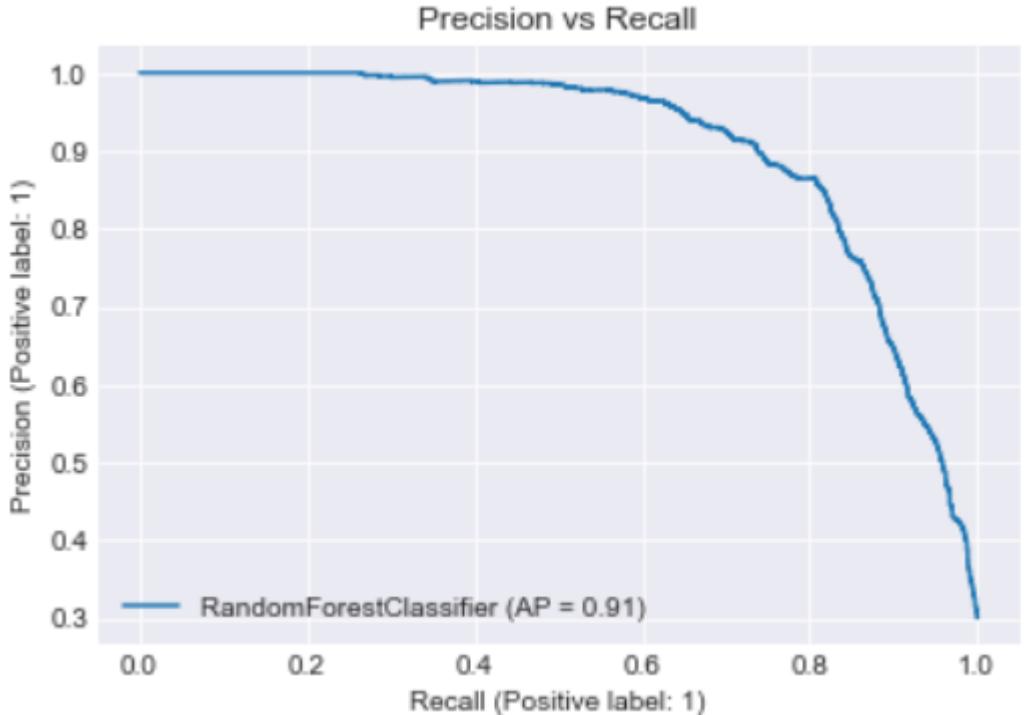
Using K-Nearest Neighbors Classifier

Using Logistic Regression

Evaluating our models at default threshold value (0.5)

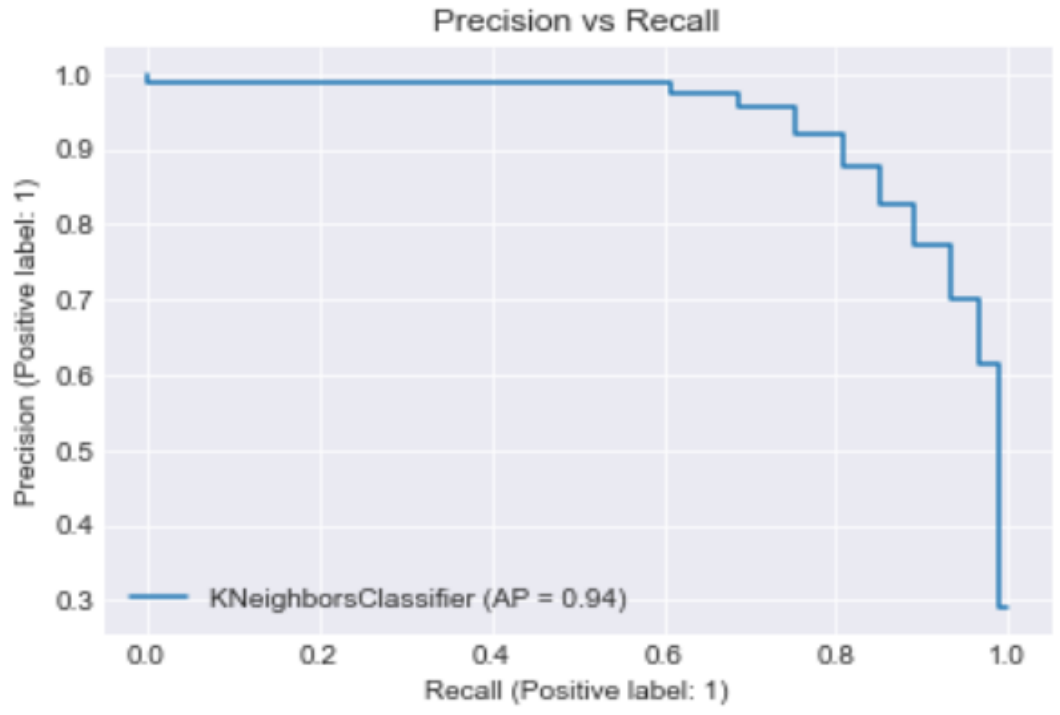
Random forest Classifier

	precision	recall	f1-score	support
0	0.90	0.98	0.94	5525
1	0.93	0.74	0.83	2255
accuracy			0.91	7780
macro avg	0.92	0.86	0.88	7780
weighted avg	0.91	0.91	0.91	7780



K-Nearest Neighbors Classifier

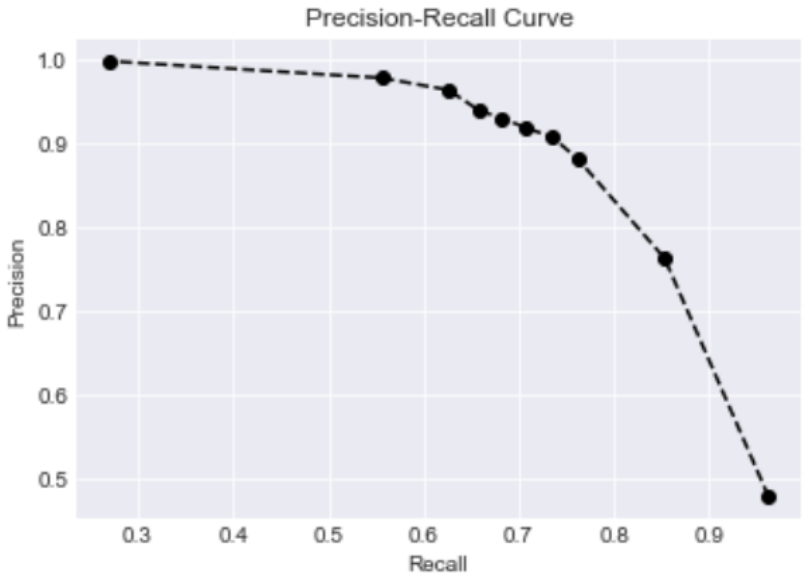
	precision	recall	f1-score	support
0	0.95	0.94	0.95	5588
1	0.85	0.88	0.87	2192
accuracy			0.92	7780
macro avg	0.90	0.91	0.91	7780
weighted avg	0.92	0.92	0.92	7780



Evaluating our models at different threshold values:

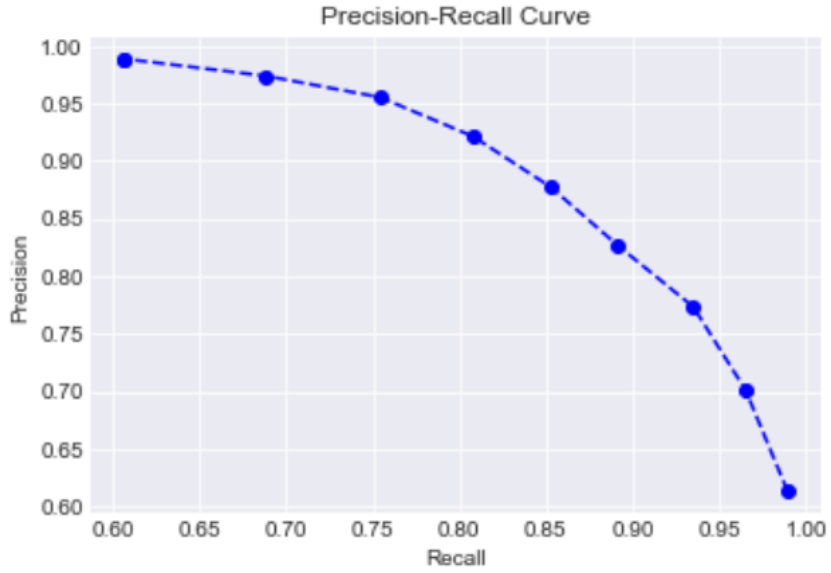
RandomForest Classifier

Threshold	Recall	Precision	F1 score
0.100000	0.956098	0.477202	0.636646
0.198889	0.848780	0.800167	0.823757
0.297778	0.782262	0.900919	0.837408
0.396667	0.766297	0.916711	0.834783
0.495556	0.742794	0.930556	0.826141
0.594444	0.705543	0.942536	0.807000
0.693333	0.670953	0.952771	0.787406
0.792222	0.623503	0.967653	0.758360
0.891111	0.540133	0.973621	0.694809
0.990000	0.268293	0.995066	0.422634



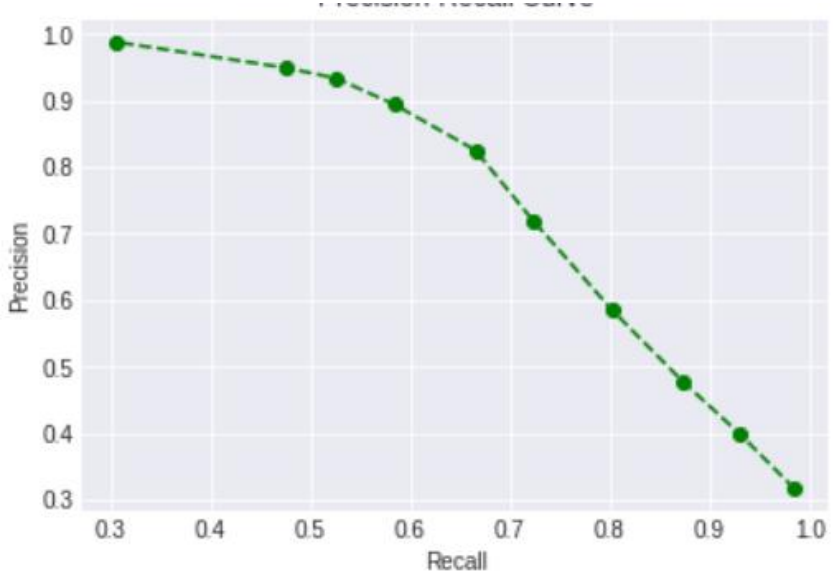
K-Nearest Neighbors

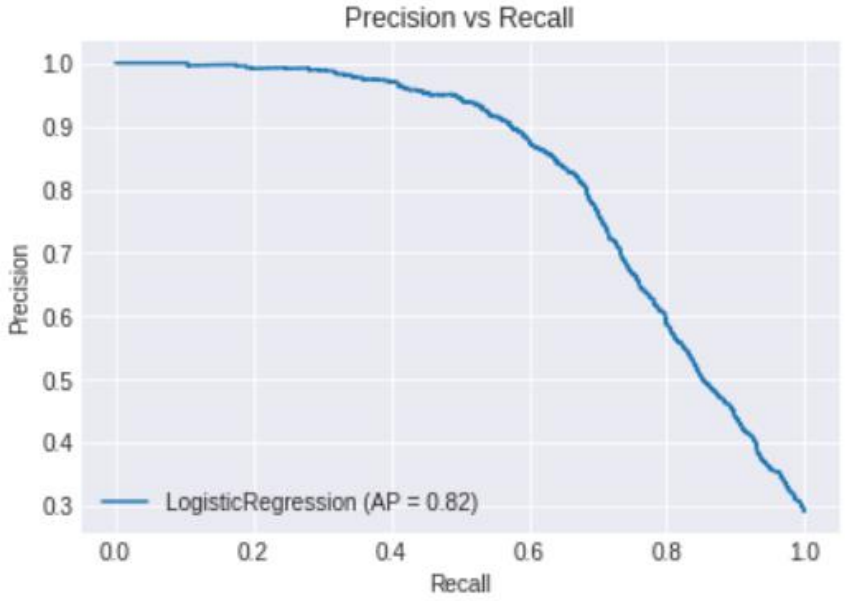
	Recall	Precision	F1 score
0.100000	0.988914	0.625000	0.765928
0.198889	0.975166	0.710042	0.821749
0.297778	0.944568	0.779363	0.854050
0.396667	0.902882	0.832039	0.866014
0.495556	0.854545	0.879106	0.866652
0.594444	0.803991	0.922177	0.859038
0.693333	0.760532	0.946468	0.843373
0.792222	0.697561	0.968000	0.810825
0.891111	0.624390	0.984615	0.764179
0.990000	0.624390	0.984615	0.764179



Logistic Regression:

Threshold	Recall	Precision	F1 score
0.100000	0.984922	0.315483	0.477891
0.198889	0.929047	0.399504	0.558741
0.297778	0.871840	0.478462	0.617850
0.396667	0.801330	0.585548	0.676652
0.495556	0.722395	0.719841	0.721116
0.594444	0.664745	0.825441	0.736428
0.693333	0.583149	0.894558	0.706040
0.792222	0.525055	0.933754	0.672154
0.891111	0.474945	0.949468	0.633166
0.990000	0.303769	0.988456	0.464722





Metrics Discussion

When to prefer Precision over Recall?

- Precision should be used when we want to predict the minority class with greater confidence. Suppose the bank wants to reduce the FP such that none of the clients who haven't subscribed to our bank term deposit, should be classified incorrectly. The Bank Manager does not want any clients who haven't subscribed to our plans, take benefits from the bank. In such cases, the threshold is to be increased, such that even if the recall decreases, our model is correctly able to classify the minority class properly.

When to prefer Recall over Precision?

- Suppose the bank wants to reduce the FN such that none of the clients who have subscribed to our bank term deposit, be classified incorrectly. The Bank Manager does not want any clients who have subscribed to our plans be misclassified because of which they may not be able to avail our returns on the principal amount. This may lead to a bad reputation of the bank. In such cases, the threshold is to be decreased, such that even if the precision decreases, our model is correctly able to classify the clients who have subscribed to our plans

Conclusion:

- We had a good overall Precision-Recall score at varying thresholds using both RandomForest Classifier and KNN Classifier. Depending on our problem statement, we will be using Recall as our metric.

- Interpreting the results from EDA, we found out that the bank had targeted clients mostly from management professionals, blue-collared and administrators. However, we found that the percentage of success of a bank in subscribing a client is much more in students, retired clients and even unemployed clients.

- Bank authorities have previously contacted their clients more frequently in the months of May, however the percentage of clients subscribing were much less. The percentage of success was much higher when the clients were last contacted in September, however the counts of calls were much less in September.

- Age, Job, Balance, Education and month are some of the most important features in predicting our model