FINAL PROJECT REPORT:

Portuguese Bank Marketing:

"Submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM"

Candidate Name	Student Id
Sri Harshini	900ZPJV65E
Megha	911OTD64HN
Vikas	65X22F7040
Nikhil	MEEWMITSM5
Vaagme Chamarty	AU08JFFU06

Batch: DSE_Bengaluru_July-2019

Mentor: Mrs. C .R. Sowmya

greatlearning
Learning for Life

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

ACKNOWLEDGEMENT

At the outset, we are indebted to our Mentor Mrs. Sowmya Vivek for her time, valuable inputs and guidance.

Her experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are grateful to the course directors PGP– DSE Program.

Their in-depth knowledge coupled with their passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to them for the various sessions they have provided.

We are thankful to Ms Uma Khana, Program Manager PGP-DSE Program for his unflinching and unabated help extended to us always.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics and machine learning.

We would like to thank our respective families for giving us the necessary support to complete this program.

CERTIFICATE OF

I hereby certify that the project titled “Portuguese Bank Marketing ” was undertaken and completed under my guidance and supervision by CHAMARTY VAAGME, SREE HARSHINI,MEGHA,VIKAS,NIKHIL students of July 2019 batch of the Post Graduate Program in Data Science Engineering, Bangalore

Mrs. C.R. Sowmya

Date: November 2019

TABLE OF CONTENTS

TABLE OF CONTENTS

CHAPTER-1 PROJECT OVERVIEW

Problem Statement:

There has been a revenue decline for the Portuguese bank and they would like to know what actions to take. After investigation, we found out that the root cause is that their clients are not depositing as frequently as before.

Knowing that term deposits allow banks to hold onto a deposit for a specific amount of time, so banks can invest in higher gain financial products to make a profit. In addition, banks also hold better chance to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues.

The Portuguese bank would like to identify existing clients that have higher chance to subscribe for a term deposit and focus marketing effort on such clients.

Domain:

Banking

Data Source:

UCI Machine Learning Repository

Data Dictionary:

The Portuguese Banking Institution dataset consists of 45211 records and 17 qualitative and quantitative attributes of the customer. These are Fields which are present in the given dataset are

Age, Job, Marital Status, Education, Credit Default ,Average Yearly Balance, Housing Loan, Personal Loan ,Communication Type ,Last Contact Day ,Last Contact Month, Last Call Duration , No of Campaign, No of days passed after last contact, Previous Contact Before Campaign, Outcome of Previous Campaign , Class Target.

Variable categorization:

There are 9 categorical Fields and 7 Numerical Fields and 1 Target Field

S.NO	Feature Name	Feature Type	Feature Description

1	Age	Numeric, Bank Marketing Data	Age of the client
2	Job	Categorical, Bank Marketing Data (Sub-Categorical: admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)	Type of job of the client
3	Marital Status	Categorical, Bank Marketing Data (Sub-Categorical: 'divorced', 'married‘ , 'single', 'unknown', 'divorced' means divorced or widowed)	Marital Status of the client
4	Education	Categorical, Bank Marketing Data Sub-Categorical: primary, secondary, tertiary and unknown)	Education level of the client
5	Credit Default	Categorical, Bank Marketing Data (Sub-Categorical: 'no', 'yes', ‘unknown’)	Done Credit Default or not
6	Average Yearly balance	Numerical ,Bank Marketing Data	Balance of the Individual Clients.
7	Housing Loan	Categorical, Bank Marketing Data (Sub-Categorical: ‘no‘, ‘yes‘, 'unknown')	If the client has a Housing Loan or not
8	Personal Loan	Categorical, Bank Marketing Data (Sub-Categorical: ‘no‘, ‘yes‘, 'unknown')	If the client has a Personal Loan or not
9	Communication type	Categorical, Bank Marketing Data	
10	Last Contact Day	Categorical, Bank Marketing Data	No of days contacting the customer

11	Last Contact Month	Categorical, Bank Marketing Data	Which month last contact was made
12	Last call duration	Numerical, Bank Marketing Data	The Last Contacting Duration, in seconds
13	No of Campaign	Numerical ,Bank Marketing Data	Number of contacts performed during this campaign with a particular customer
14	No Of Days Passed after last contact	Numeric, Bank Marketing Data	Number of days passed by after the customer was contacted from a previous campaign
15	Previous Contact before campaign	Numeric, Bank Marketing Data	Number of contacts performed before this campaign for a particular customer
16	Outcome of previous Campaign	Categorical, Bank Marketing Data	Number of contacts performed before this campaign For a particular customer
17	Target	Categorical	Subscribed to term deposit

Pre Processing Data Analysis:

Checking for the null values:

Age	0
Job	0
Marital Status	0
Education	0
Credit Default	0
Average Yearly Balance	0
Housing Loan	0
Personal Loan	0

Communication Type	0
Last Contact Day	0
Last Contact Month	0
Last Call Duration	0
No of Campaign	0
No of days passed after last contact	0
Previous Contact before Campaign	0
Outcome of Previous Campaign	0
Target	0

Data type: int64

Observation: There are no null Values and Missing Values

Exploratory Data Analysis:

EDA can be approached in 3 ways

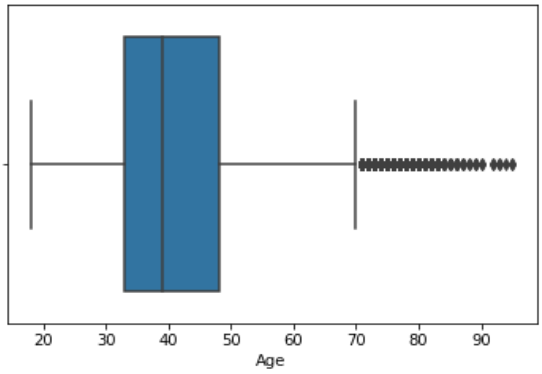
1. Univariate analysis
    a. to understand the central tendency and spread of numerical variables
    b. to understand the proportion of the various levels of categorical variables

2. Bivariate & Multivariate analysis
    a. Understand the relationship between 2 variables
    b. Visualise the impact of various X variables on Y variable, thereby giving cues for feature selection

Univariate Analysis:

Univariate analysis for the Numeric features

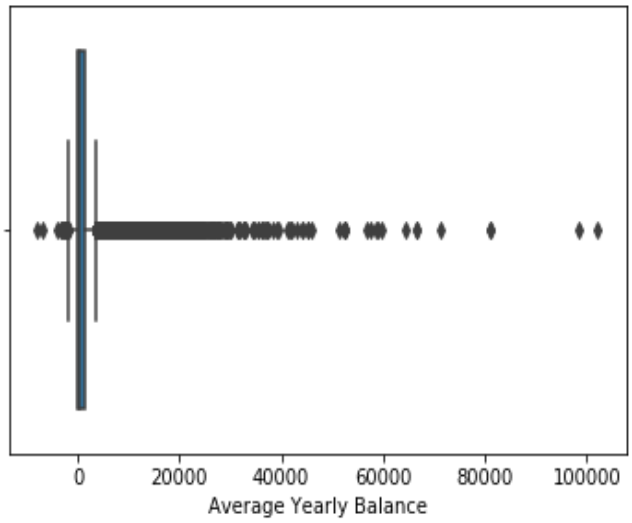
Age



Observation:

- We can identify in the Age column that the mean is greater than the median i.e. it is affected by high outliers.(Right Skewed)
- The median age of customers is around 39 years.

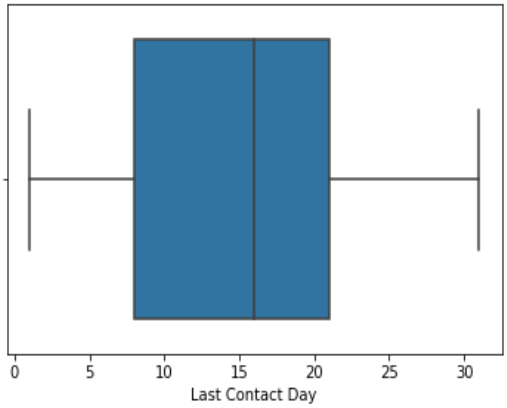
Average Yearly Balance



Observation:

- The Average Yearly Balance is affected by high outliers. (Mean>Median).
- It is right skewed

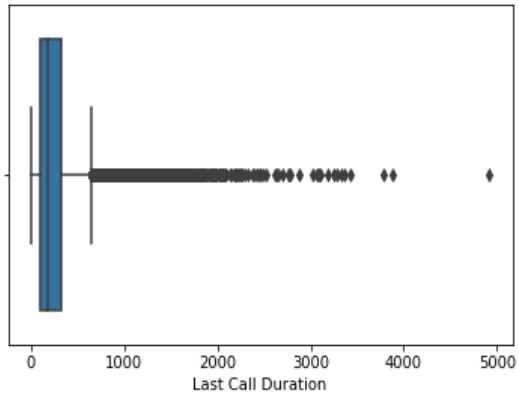
Last Contact Day



Observation:

- Here we can observe that Median > Mean we have low outlier the column is left skewed.
- The median value of last contact day is around 16.
- The maximum value is about 31 which means the latest a customer was contacted was a month earlier.

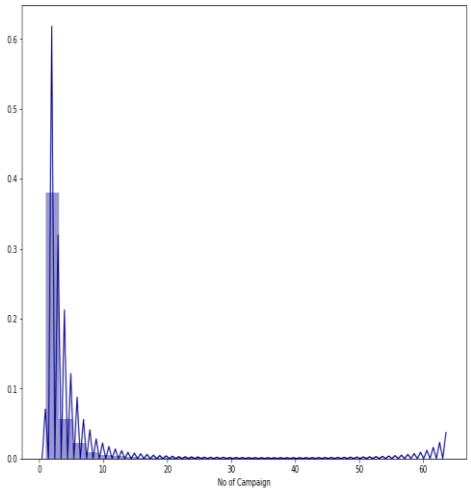
Last Call Duration:



Observation:

- Duration of the call is the feature that most positively correlates with whether a potential client will open a term deposit or not, by providing an interesting questionnaire for potential clients during the calls the conversation length might increase.
- Of course, this does not assure us that the potential client will subscribe to a term deposit! Nevertheless, we don't lose anything by implementing a strategy that will increase the level of engagement of the potential client leading to an increase probability of subscribing to a term deposit, and therefore an increase in effectiveness for the next marketing campaign the bank will execute

No of Campaigns



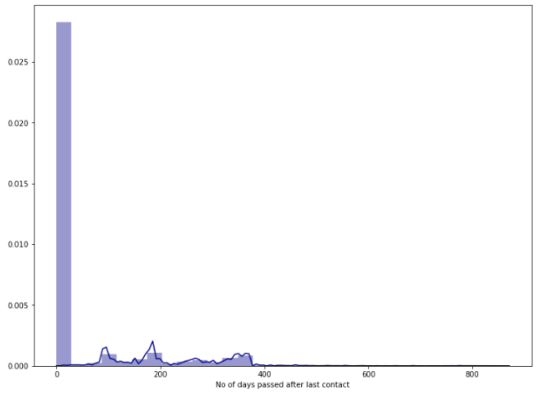
Observation:

Here, we can observe that not much effort is shown by the Portuguese bank in holding more campaigns for the clients.

The value of Mean >Median i.e. there are high outliers and it is right skewed. The Portuguese bank should hold more campaigns in order to get more clients who can subscribe for the term deposit in future.

No of days passed

After last contact



Previous Contact Before

Campaign

```
In [99]:  1  bank["Previous Contact Before Campaign"].value_counts()

Out[99]: 0      36954
         1       2772
         2       2106
         3       1142
         4        714
         5        459
         6        277
         7        205
         8        129
         9         92
         10        67
         11        65
         12        44
         13        38
         15        20
         14        19
         17        15
         16        13
         19        11
         23         8
         20         8
         22         6
         18         6
         24         5
         27         5
         29         4
         25         4
         21         4
         30         3
         28         2
         26         2
         37         2
         38         2
         55         1
         40         1
         35         1
         58         1
         51         1
         41         1
         32         1
         275        1
         Name: Previous Contact Before Campaign, dtype: int64
```

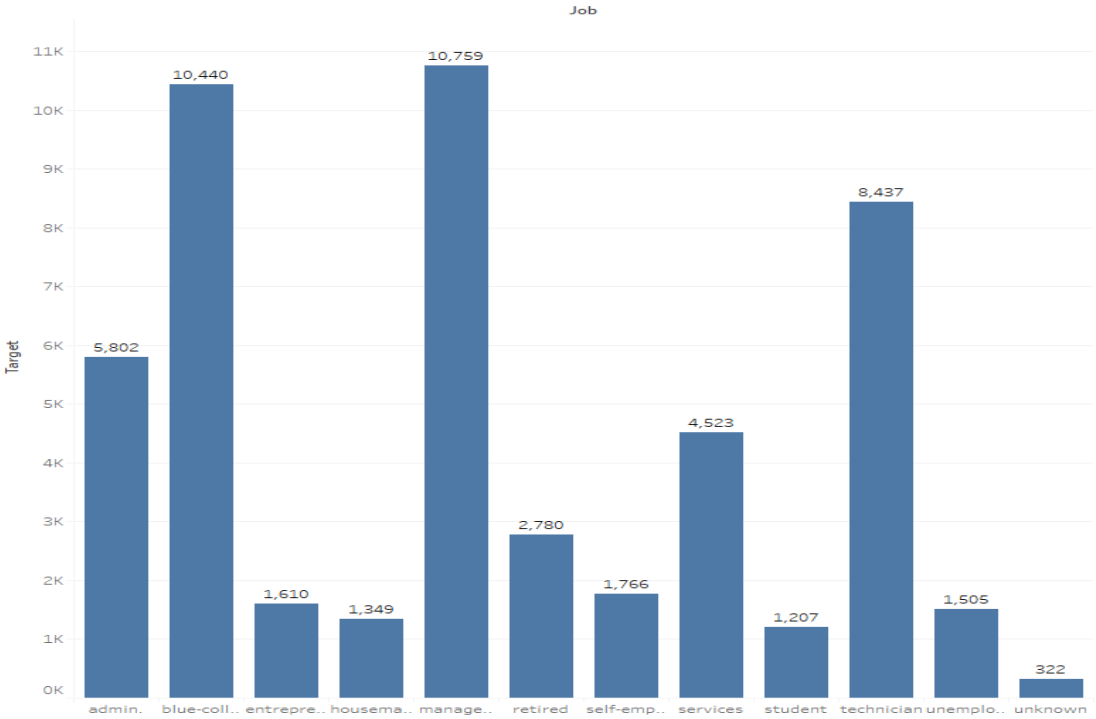
Observation:

Here we can see that Mean >Median i.e. high outliers are present and it's right skewed We can easily infer that after the last day of contact with the client so many days have passed but no other contact was made. They can make their marketing strategies in such a way so that they can reach out to more and more people. Because they are not reachable to more customers with just call. And they should be able to get in touch within a specific time period.

Univariate Analysis:

Univariate analysis for the Categorical features

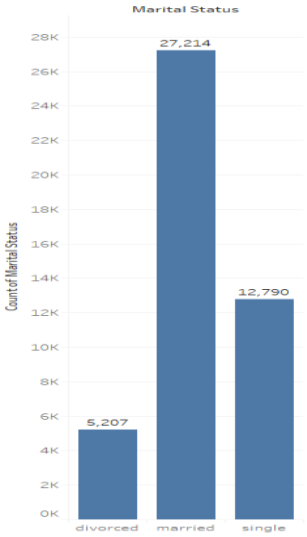
Job:



Observation:

The highest job category of the customers is blue-collared job.

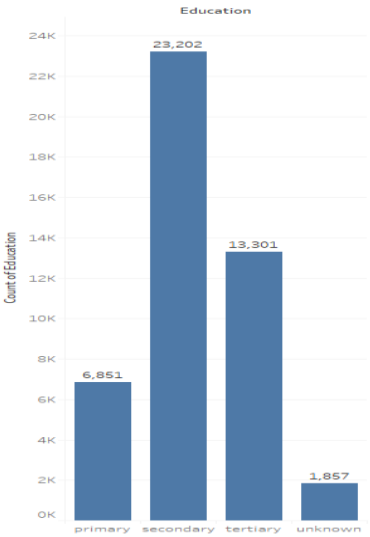
Marital Status



Observation:

- About 50% of the customers are married.

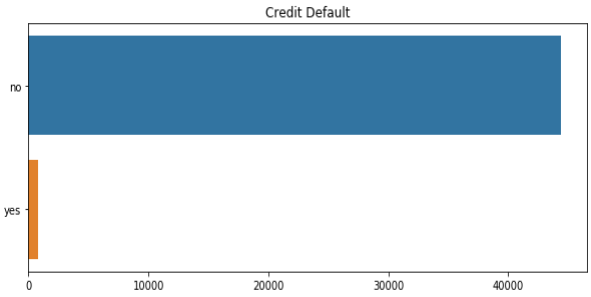
Education



Observation:

- Customers with secondary level of the education constitute the majority of contacted customers

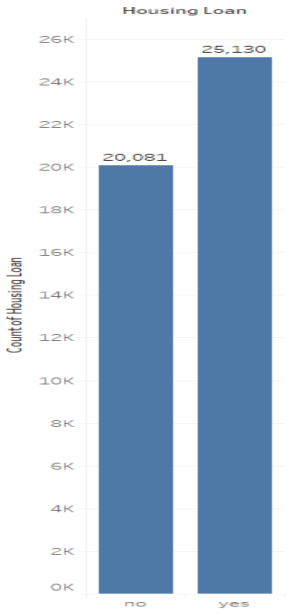
Credit Default



Observation:

- Most customers have not defaulted credit till date.

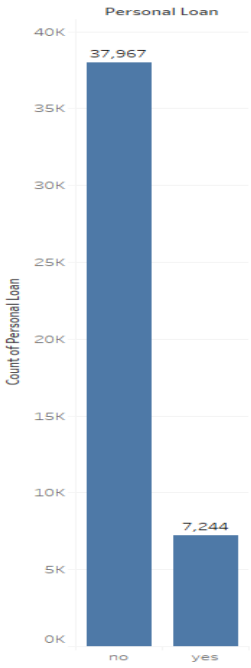
Housing Loan:



Observation:

- Here customers with housing loans received the most number of calls.

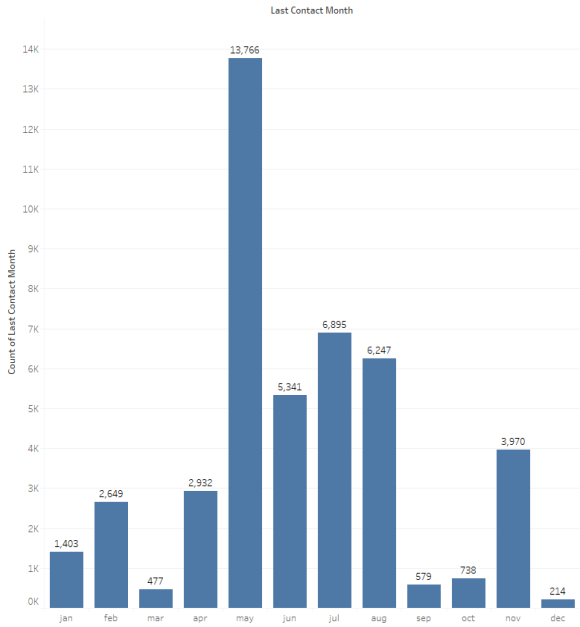
Personal Loan:



Observation:

- Customers with personal loan have received more calls than those who do not have.

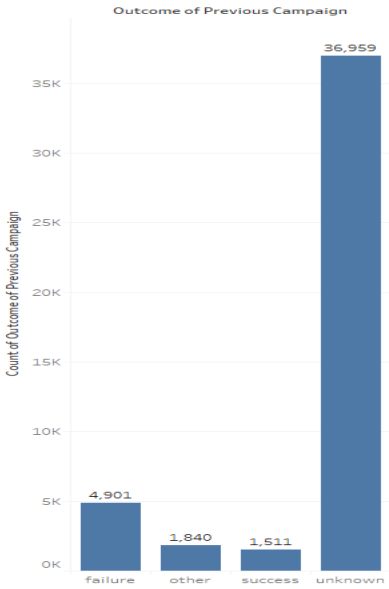
Last Contact Month



Observation:

- May month recorded most of the calls when customers were contacted for the last time.

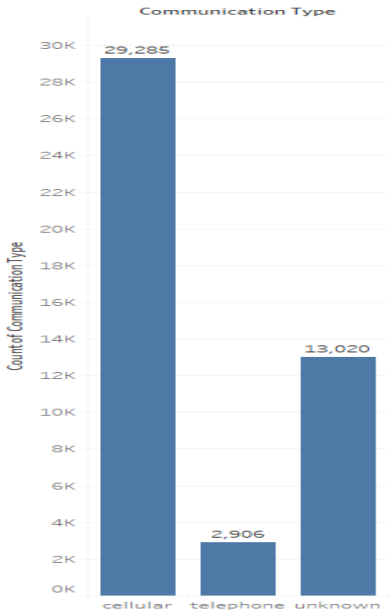
Outcome of Previous Campaign



Observation:

In most cases, the outcome of the previous campaign is not known. This might be a useful information to track in future.

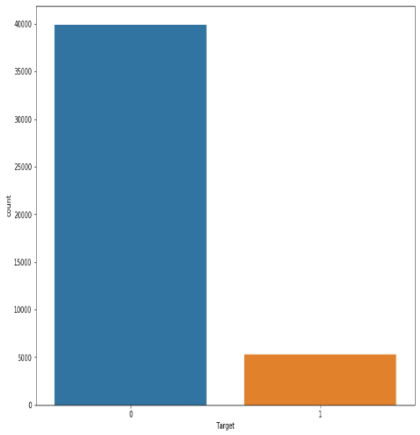
Communication Type



Observation:

- We observe that the medium of communication is mostly through “cellular”.

Target Feature:

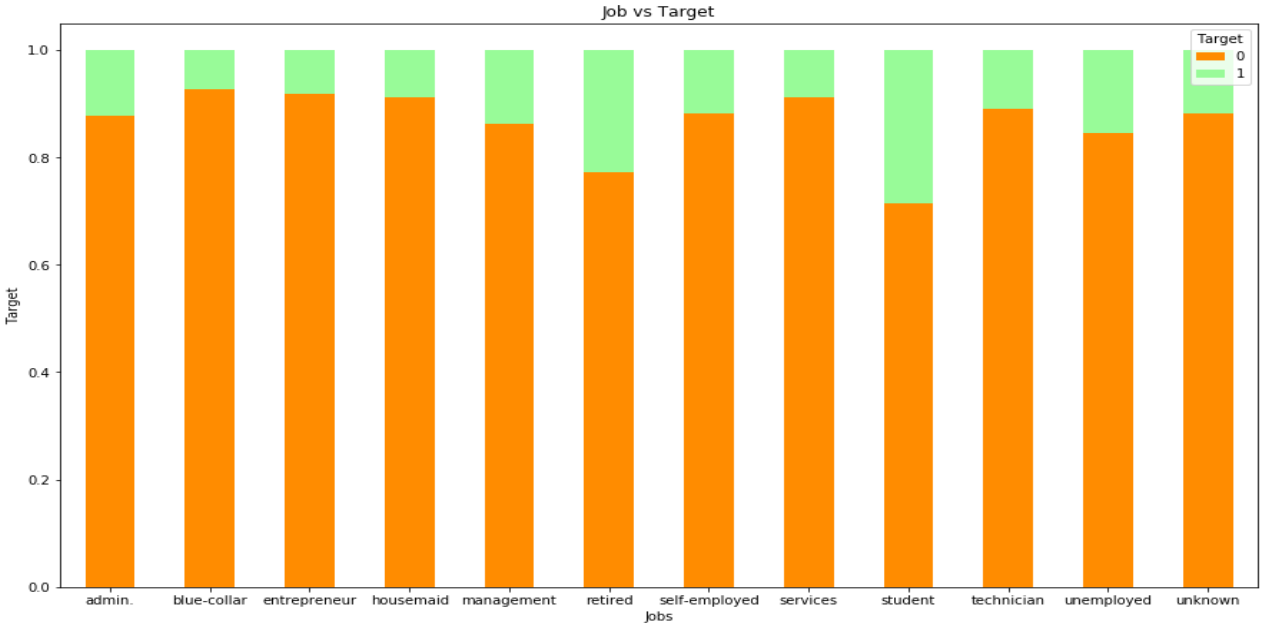


Observation:

- The count of the ‘zeros’ are 88% and ‘ones’ are 12%. This indicates that the dataset has a class imbalance for the Y variable.

Bivariate Analysis

Job vs. Target

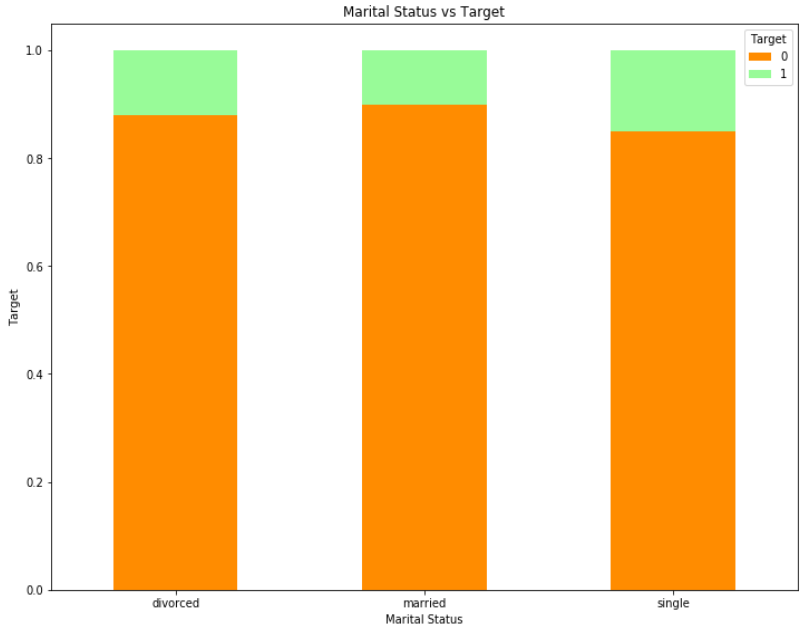


Observation:

- Not surprisingly, clients that were students or retired were the most likely to subscribe to a term deposit.
- Retired individuals tend to have more term deposits in order to gain some cash through interest payments they might agree not to withdraw the cash from the bank until a certain date agreed between the individual and the financial institution.

- After that time the individual gets its capital back and its interest made on the loan. Retired individuals tend to not spend cash so they are more likely to put their cash to work by lending it to the financial institution.
- Students were the other group that used to subscribe term deposits for their future use like higher studies or financial settlement or for marriage purpose

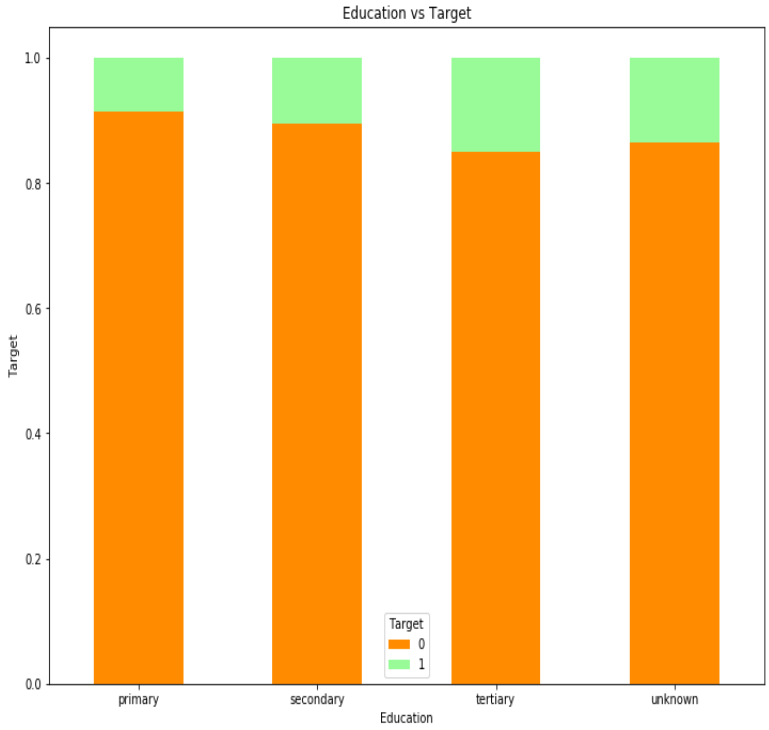
Marital Status vs. Target



Observation:

- Looks like the people who are single are more willing to subscribe for the term deposit followed by people who are married and divorced
- Well in this analysis we can find some significant insights other than most divorced individuals are broke.
- No wonder since they have to split financial assets.
- For married people financial spread is more might be that they cannot subscribe the loan

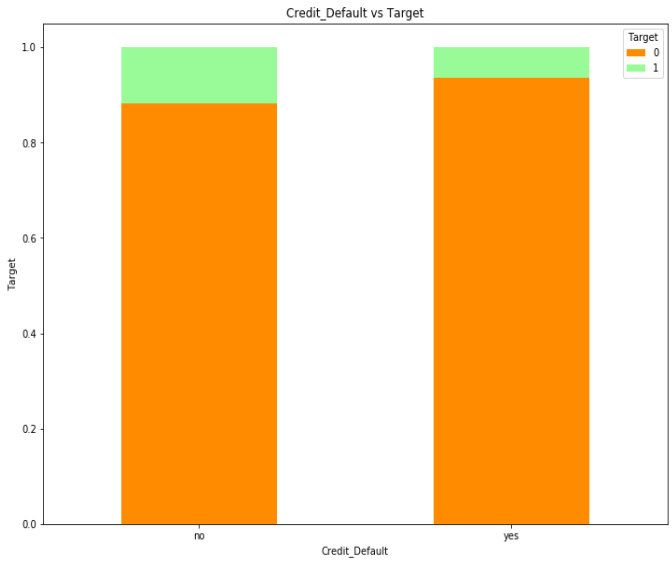
Education vs. Target:



Observation:

- Looks like the people who are tertiary Educated are more willing to subscribe for the term deposit followed by people who are primary, secondary and unknown

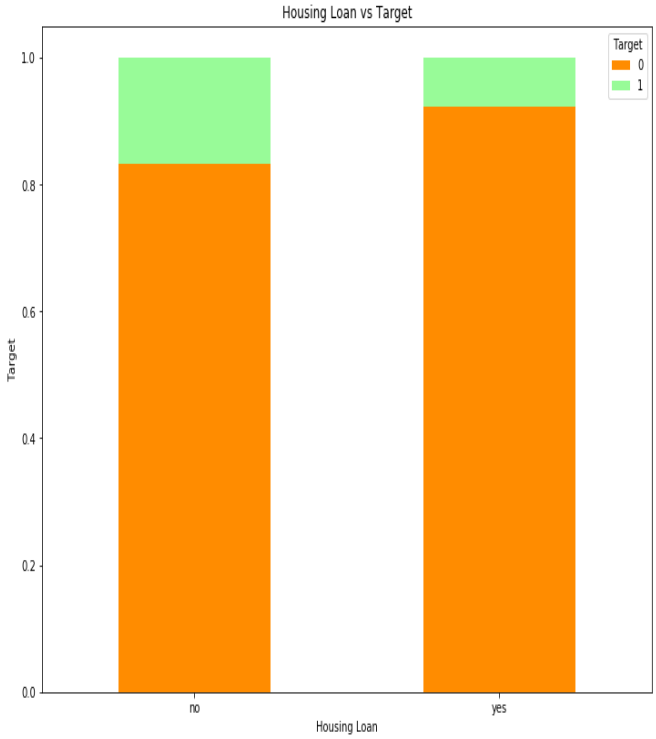
Credit Default vs. Target



Observation:

- The people who are not having any credit default are not subscribing.
- Generally, more people with credit default are subscribing.

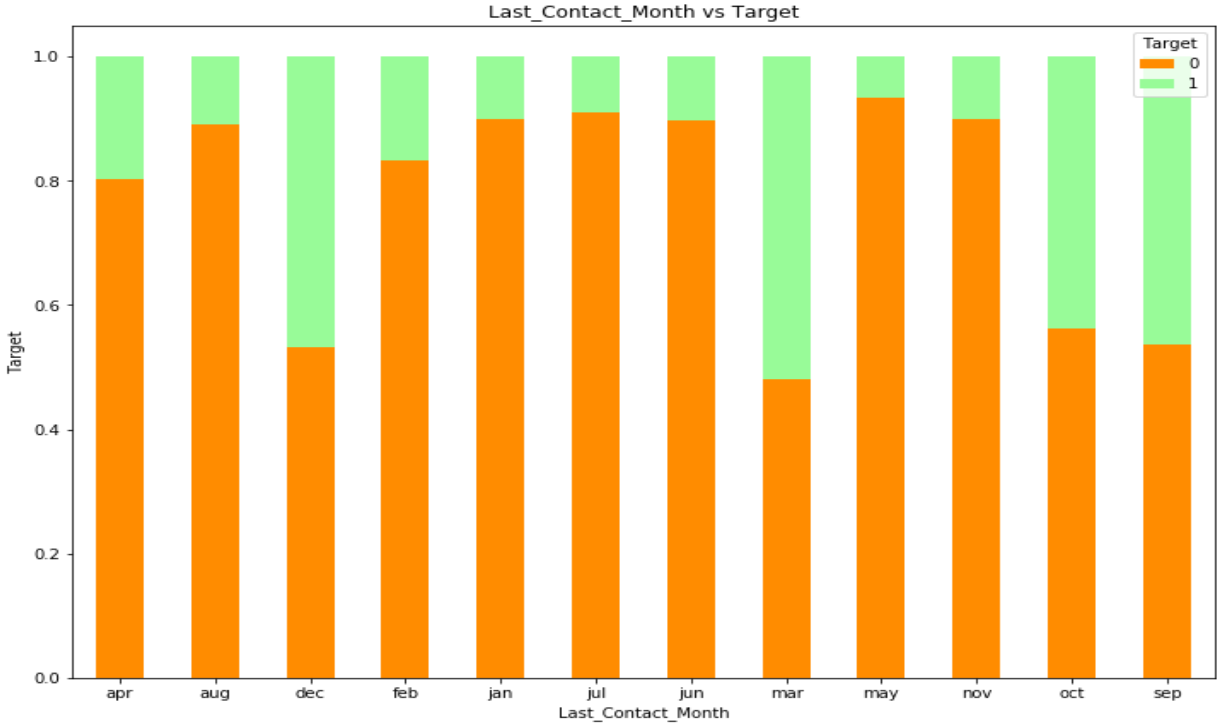
Housing Loan vs. Target



Observation:

- Potential clients in the low balance and no balance category were more likely to have a house loan than people in the average and high balance category.
- This means that the potential client has financial compromises to pay back its house loan and thus, there is no cash for him or her to subscribe to a term deposit account.
- However, we see that potential clients in the average and high balances are less likely to have a house loan and therefore, more likely to open a term deposit.
- Lastly, the next marketing campaign should focus on individuals of average and high balances in order to increase the likelihood of subscribing to a term deposit.

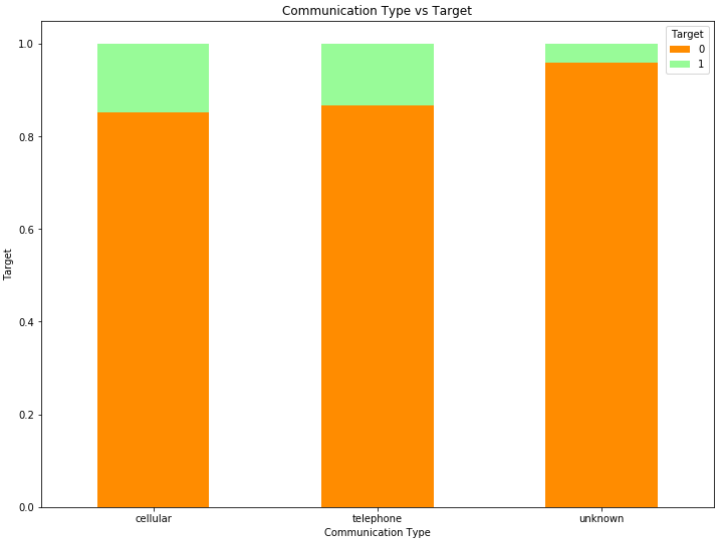
Last_Contact_Month vs. Target



Observation:

- We saw that the month of highest level of conversion was the month of March.

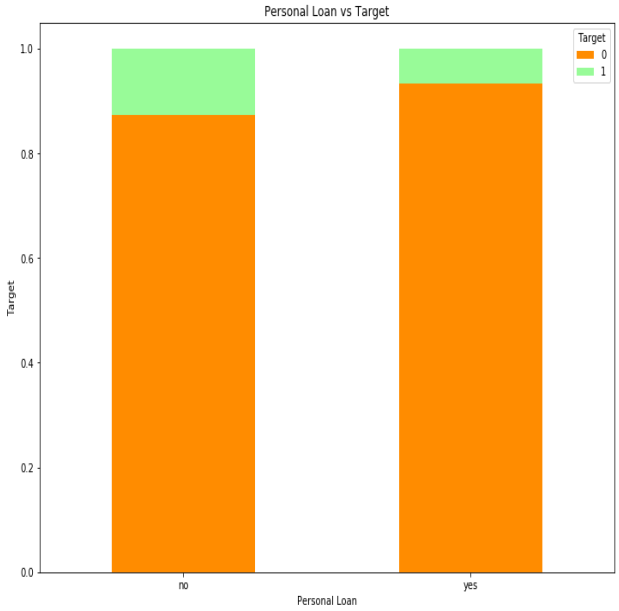
Communication type vs. Target



Observation:

- Conversion % is highest when the customer is contacted through cellular.

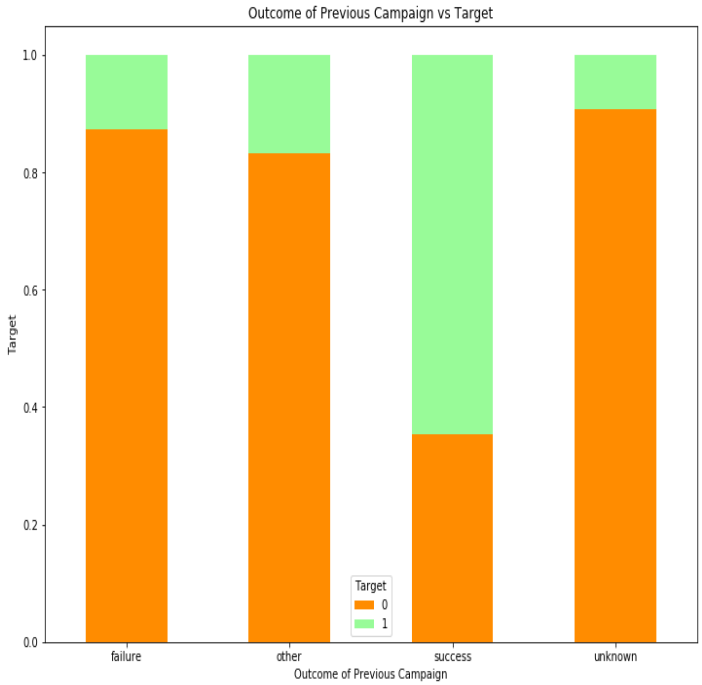
Personal Loan



Observation:

- The people who are not taking personal loans are subscribing for term deposit and the people who have taken personal loans are not subscribing.

Outcomes of the previous Campaign



Observation:

- But, those with previous outcome as success have the highest subscription share to the term deposit.

Encoding:

A data frame is returned with all the possible values after splitting every string. If the text value in original data frame at same index contains the string (Column name/ Spited values) then the value at that position is 1 otherwise, 0.

Get Dummies:

Get dummies is a common way to create dummy variables for categorical features. While it is widely used, there are some drawbacks. First, it modifies your data frame. When you have a categorical feature with hundreds of categories, get dummies adds hundreds of dummy variables to the data frame. And you may need to drop the categorical feature after creating dummies if you want to quickly assign features to X (independent variables). Another thing about get dummies is, if there are many categorical features in the dataset, you will have to do get dummies for every feature. Even constructing a loop to do so is a lot code to write.

CHAPTER 2 :Model Building

Logistic Regression:

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a log it function.

Remember, there are some cases where dependent variables can have more than two outcomes, like married /unmarried/divorced such scenarios are classified as multinomial logistic regression**.** Though they work in the same manner to predict the outcome.

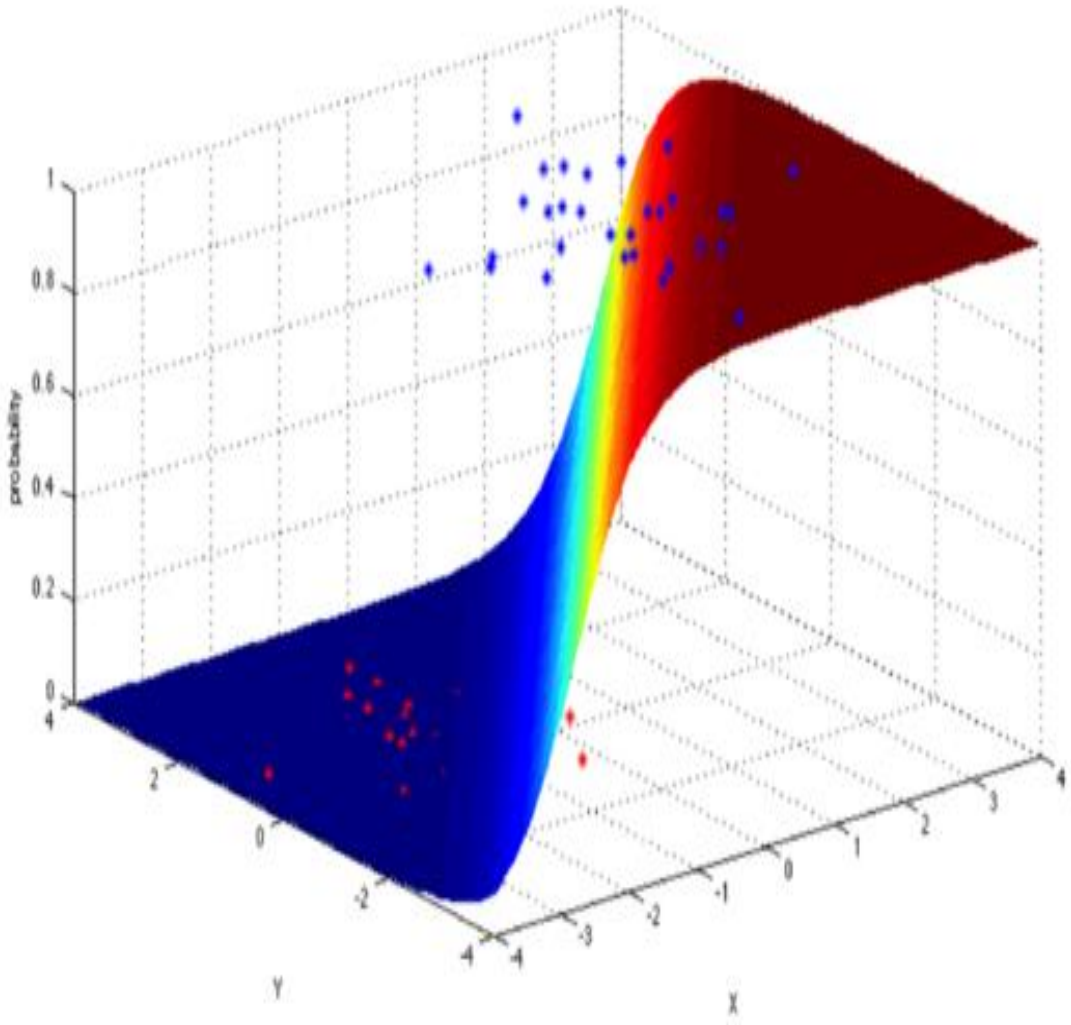
As we are clear that logistics regression majorly makes predictions to handle problems which require a probability estimate as output, in the form of 0/1. Logistic regression is an extremely efficient mechanism for calculating probabilities. So you must be curious to understand how does it comes out with the value of 0 or 1 always. To understand more let's try to decode some math behind Logistic Regression.

Logistic Model: Sigmoid Function

Let us try to understand logistic regression by understanding the logistic model. As in linear regression let's represent our hypothesis (Prediction of Dependent variable) in classification. In classification our hypothesis representation which tries to predict the binary outcome of either o or 1, will look like,

$h\vartheta(x) = g(\vartheta^T x) = 1/ 1 + e^{-\vartheta^T x}$,

Here g (z) = 1/(1 + e ^−z), is called the logistic function or the sigmoid function:



#Summary of logistic regression

Logistic Summary

Logit Regression Results

Dep. Variable:	Target	No. Observations:	31647
Model:	Logit	Df Residuals:	31604
Method:	MLE	Df Model:	42
Date:	Wed, 13 Nov 2019	Pseudo R-squ.:	0.3369
Time:	21:12:24	Log-Likelihood:	-7511.0
converged:	True	LL-Null:	-11327.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5796	0.221	-11.655	0.000	-3.013	-2.146
Age	0.0002	0.003	0.079	0.937	-0.005	0.005
Average Yearly Balance	1.619e-05	6.16e-06	2.628	0.009	4.12e-06	2.83e-05
Last Contact Day	0.0058	0.003	1.943	0.052	-5e-05	0.012
Last Call Duration	0.0041	7.7e-05	53.557	0.000	0.004	0.004
No of Campaign	-0.0741	0.012	-6.308	0.000	-0.097	-0.051
No of days passed after last contact	-0.0004	0.000	-1.101	0.271	-0.001	0.000
Previous Contact Before Campaign	0.0100	0.007	1.435	0.151	-0.004	0.024
Job_blue-collar	-0.2875	0.088	-3.278	0.001	-0.459	-0.116
Job_entrepreneur	-0.3485	0.152	-2.298	0.022	-0.646	-0.051
Job_housemaid	-0.4276	0.159	-2.681	0.007	-0.740	-0.115
Job_management	-0.1232	0.088	-1.399	0.162	-0.296	0.049
Job_retired	0.3164	0.117	2.714	0.007	0.088	0.545
Job_self-employed	-0.2746	0.133	-2.068	0.039	-0.535	-0.014
Job_services	-0.1888	0.100	-1.886	0.059	-0.385	0.007
Job_student	0.4094	0.131	3.119	0.002	0.152	0.667
Job_technician	-0.1911	0.083	-2.303	0.021	-0.354	-0.028
Job_unemployed	-0.1976	0.137	-1.440	0.150	-0.466	0.071
Job_unknown	-0.2419	0.291	-0.832	0.405	-0.812	0.328
Marital Status_married	-0.1576	0.071	-2.223	0.026	-0.297	-0.019
Marital Status_single	0.1891	0.080	2.350	0.019	0.031	0.347
Education_secondary	0.2295	0.079	2.900	0.004	0.074	0.385
Education_tertiary	0.4067	0.092	4.438	0.000	0.227	0.586
Education_unknown	0.2071	0.127	1.629	0.103	-0.042	0.456
Credit Default_yes	0.1166	0.188	0.621	0.534	-0.251	0.485
Housing Loan_yes	-0.6441	0.053	-12.242	0.000	-0.747	-0.541
Personal Loan_yes	-0.4138	0.072	-5.766	0.000	-0.555	-0.273
Communication Type_telephone	-0.2567	0.091	-2.807	0.005	-0.436	-0.077
Communication Type_unknown	-1.6299	0.088	-18.540	0.000	-1.802	-1.458
Last Contact Month_aug	-0.6752	0.094	-7.204	0.000	-0.859	-0.492
Last Contact Month_dec	0.8712	0.205	4.252	0.000	0.470	1.273
Last Contact Month_feb	-0.1727	0.107	-1.619	0.105	-0.382	0.036
Last Contact Month_jan	-1.2416	0.148	-8.411	0.000	-1.531	-0.952

Odds Ratio:

Odds Ratio (OR) is a measure of association between exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

Important points about Odds ratio:

- Calculated in case-control studies as incidence of outcome is not known
- OR >1 indicates increased occurrence of event
- OR <1 indicates decreased occurrence of event (protective exposure)
- Look at CI and P value for statistical significance of value
- In rare outcomes OR = RR (RR = Relative Risk). This applies when the incidence of disease is < 10%

	coef	Odds ratio	probability	pval
Outcome of Previous Campaign_success	2.35	10.45	0.91	0.00
Last Contact Month_mar	1.55	4.71	0.82	0.00
Last Contact Month_oct	0.77	2.16	0.68	0.00
Last Contact Month_sep	0.77	2.15	0.68	0.00
Last Contact Month_dec	0.70	2.01	0.67	0.00
Last Contact Month_jun	0.51	1.67	0.63	0.00
Job_student	0.51	1.67	0.62	0.00
Education_tertiary	0.41	1.50	0.60	0.00
Job_retired	0.38	1.46	0.59	0.00
Education_unknown	0.23	1.26	0.56	0.06
Education_secondary	0.20	1.22	0.55	0.01
Previous Contact Before Campaign	0.03	1.03	0.51	0.03
Last Contact Day	0.01	1.01	0.50	0.00
Last Call Duration	0.00	1.00	0.50	0.00
Average Yearly Balance	0.00	1.00	0.50	0.03

	coef	Odds ratio	probability	pval
No of Campaign	-0.09	0.92	0.48	0.00
Communication Type_telephone	-0.15	0.86	0.46	0.09
Job_technician	-0.19	0.83	0.45	0.02
Last Contact Month_feb	-0.20	0.82	0.45	0.07
Job_services	-0.24	0.79	0.44	0.02
Marital Status_married	-0.28	0.75	0.43	0.00
Job_entrepreneur	-0.30	0.74	0.43	0.05
Job_blue-collar	-0.31	0.74	0.42	0.00
Last Contact Month_may	-0.40	0.67	0.40	0.00
Personal Loan_yes	-0.41	0.66	0.40	0.00
Job_housemaid	-0.44	0.65	0.39	0.01
Housing Loan_yes	-0.71	0.49	0.33	0.00
Last Contact Month_aug	-0.77	0.46	0.32	0.00
Last Contact Month_jul	-0.87	0.42	0.29	0.00
Last Contact Month_nov	-0.92	0.40	0.29	0.00
Last Contact Month_jan	-1.20	0.30	0.23	0.00
Communication Type_unknown	-1.66	0.19	0.16	0.00
const	-2.64	0.07	0.07	0.00

Observation:

As we can see from the previous slide, the Odds ratio of Outcome of Previous Campaign is 10.77 which means the odds that the customer might respond to the campaign is 10 times higher. In other words, If the outcome of the previous campaign is successful, the probability that the customer will respond to the campaign is 92%.

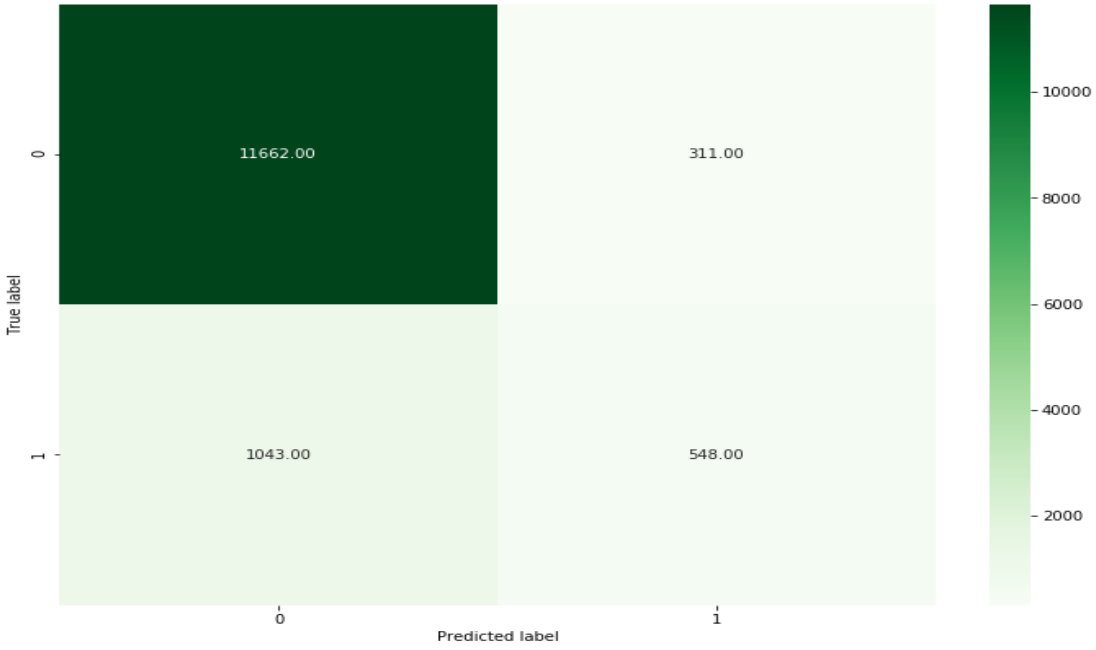
Likewise, if we make a contact to our existing clients in the month of March we have a 83% chance the client may take up the term deposit.

Compared to all the Job Categories, Students have a 60% probability of applying for a term deposit.

Confusion Matrix:

Insights of a Confusion Matrix:
The main purpose of a confusion matrix is to see how our model is performing when it comes to classifying potential clients that are likely to subscribe to a term deposit. We will see in the confusion matrix four terms the True Positives, False Positives, True Negatives and False Negatives.

Confusion matrix



Components of the confusion matrix

Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".

True Negatives (Top-Left Square) 11662	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 548	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 311	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 1043	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.

Model performance measures

Recall:

Is the total number of "Yes" in the label column of the data set. So how many "Yes" labels does our model detect.

Recall = TP/TP+FN

Inference: Recall is 34%, this says that there more misclassified actual positives. This needs to be improved.

Precision:

Means how sure is the prediction of our model that the actual label is a "Yes".
Recall Precision Trade-off: As the precision gets higher the recall gets lower and vice versa. For instance, if we increase the precision from 30% to 60% the model is picking the predictions that the model believes is 60% sure.

Precision = TP/TP+FP

Inference: If there is an instance where the model believes that is 63% likely to be a potential client that will subscribe to a term deposit then the model will classify it as a "No." However, that instance was actually a "Yes" (potential client did subscribe to a term deposit.) That is why the higher the precision the more likely the model is to miss instances that are actually a **"Yes"!** Precision is 0.63%, this says that classes are labelled correctly and are positive

Accuracy:

This is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = TP+TN/TP+FP+FN+TN

Inference: Accuracy is 90%

F1 score:

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Inference: In our case, F1 score is 0.44%, this say that positively classified class are actual positive and correct

Metrics Classification Report

	Precision	recall	f1-score	support
0	0.92	0.97	0.95	11968
1	0.65	0.34	0.44	1596
Accuracy			0.90	13564
Macro avg	0.78	0.66	0.69	13564
Weighted avg	0.89	0.90	0.89	13564

Changing the cut off value for prediction

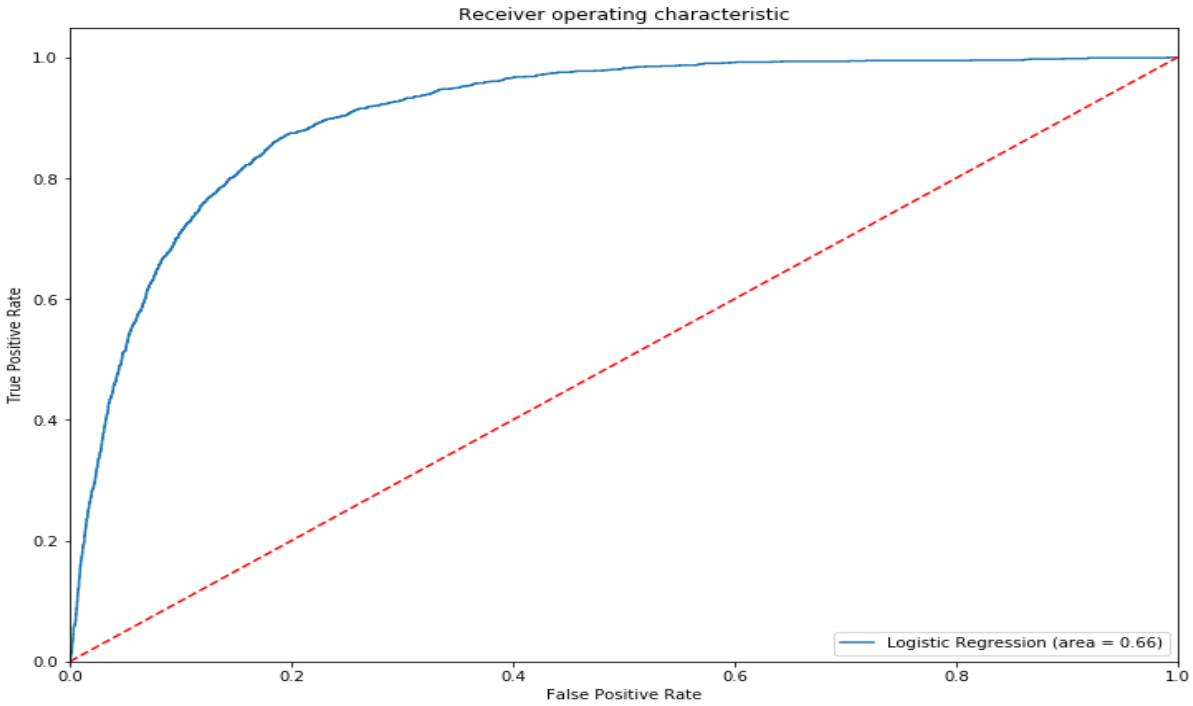
```
******** For i = 0.05 ******
Our testing recall is 0.95
[[7233 4740]
 [75 1516]]


******** For i = 0.06 ******
Our testing recall is 0.94
[[7981 3992]
 [98 1493]]


******** For i = 0.08 ******
Our testing recall is 0.90
[[8979 2994]
 [159 1432]]


******** For i = 0.1 ******
Our testing recall is 0.86
[[9615 2358]
```

```
 [225 1366]]


******** For i = 0.2 ******
Our testing recall is 0.67
[[10894 1079]
 [519 1072]]


******** For i = 0.3 ******
Our testing recall is 0.53
[[11341   632]
 [752   839]]


******** For i = 0.4 ******
Our testing recall is 0.43
[[11545   428]
 [905   686]]


******** For i = 0.5 ******
Our testing recall is 0.34
[[11662   311]
 [1043   548]]


******** For i = 0.7 ******
Our testing recall is 0.20
[[11834   139]
 [1271   320]]

******** For i = 0.75 ******
Our testing recall is 0.17
[[11867   106]
 [1314   277]]


******** For i = 0.8 ******
Our testing recall is 0.13
[[11890    83]
 [1381   210]]
```

We can see that as we change the cut-off, the recall also changes. Based on our use case, we can cho ose a cut-off that provides a relatively higher recall. A cut-off of 20% looks optimal.

ROC Curve (Receiver operating characteristic):

ROC Curve (Receiver Operating Characteristic):
The **ROC curve** tells us how well our classifier is classifying between term deposit subscriptions (True Positives) and non-term deposit subscriptions. The **X-axis** is represented by false positive rates (Specificity) and the **Y-axis** is represented by the True Positive Rate (Sensitivity.) As the line moves the threshold of the classification changes giving us different values. The closer is the line to our top left corner the better is our model separating both classes'
inferences for confusion matrix: False Positive, means the client do NOT SUBSCRIBED to term deposit, but the model thinks he did.
False Negative, means the client SUBSCRIBED to term deposit, but the model said he don't.
The first one its most harmful, because we think that we already have that client but we don't and maybe we lost him in other future campaigns
The second it's not good but its ok, we have that client and in the future we'll discovery that in truth he's already our client



ROC Score:

The Score of Receiver operating characteristic is 0.66

AUC Score:

The Score of Area under the ROC Curve is 0.90

Decision Tree:

Linear regression and logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other. Time to shine for

the decision tree! Tree based models split the data multiple times according to certain cut-off values in the features. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes. To predict the outcome in each leaf node, the average outcome of the training data in this node is used. Trees can be used for classification and regression.

There are various algorithms that can grow a tree. They differ in the possible structure of the tree (e.g. number of splits per node), the criteria how to find the splits, when to stop splitting and how to estimate the simple models within the leaf nodes. The classification and regression trees (CART) algorithm is probably the most popular algorithm for tree induction. We will focus on CART, but the interpretation is similar for most other tree types.

Gini Impurity/Indexing:

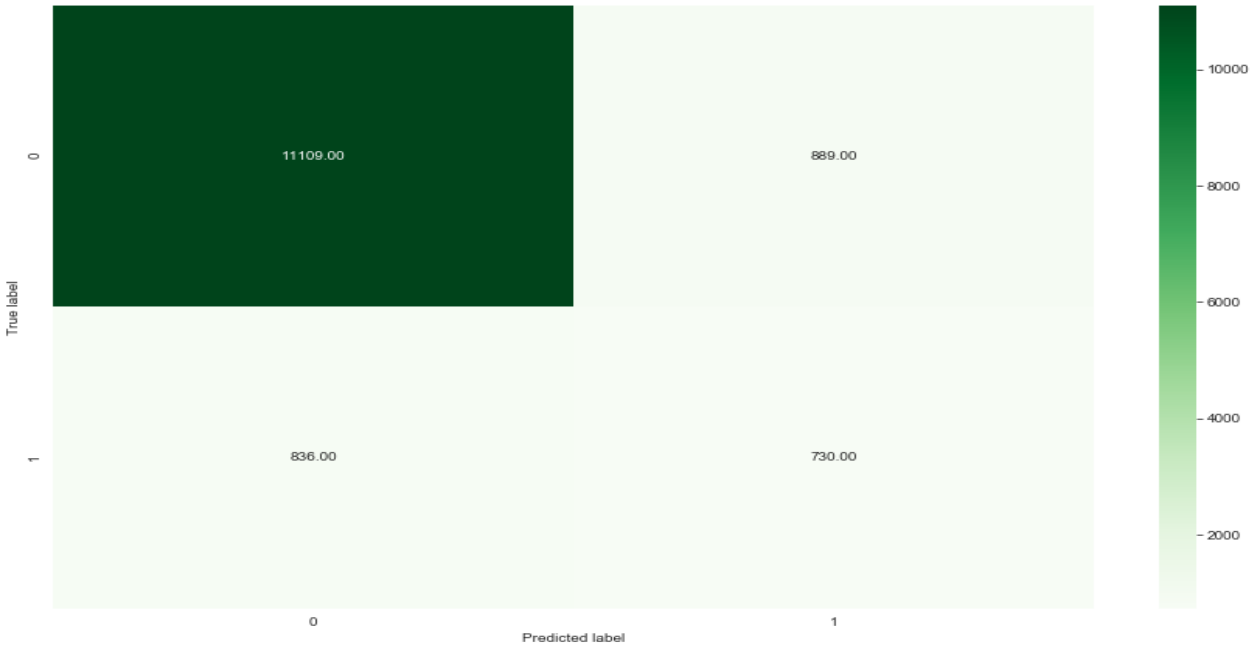
Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. But what is actually meant by 'impurity'? If all the elements belong to a single class, then it can be called pure. The degree of Gini index varies between 0 and 1, where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.

Formula for Gini Index:

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Confusion matrix

Model 1: Criterion='gini'



Evaluation Metrics

Confusion matrix =
[[11109 889]
[836 730]]

Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".
True Negatives (Top-Left Square) 11109	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 730	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 889	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 836	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.

Metrics Classification Report

	Precision	recall	f1-score	support
0	0.93	0.93	0.93	11998
1	0.45	0.47	0.46	1566
Accuracy			0.87	13564
Macro avg	0.69	0.70	0.69	13564
Weighted avg	0.87	0.87	0.87	13564

Inference:

In Precision 0's represent 93% of classes are ladled correctly and are positive

In precision 1's represent 45% of classes are ladled correctly and are positive

In Recall 0‘s represent 93% and this says that there are less misclassified actually positive

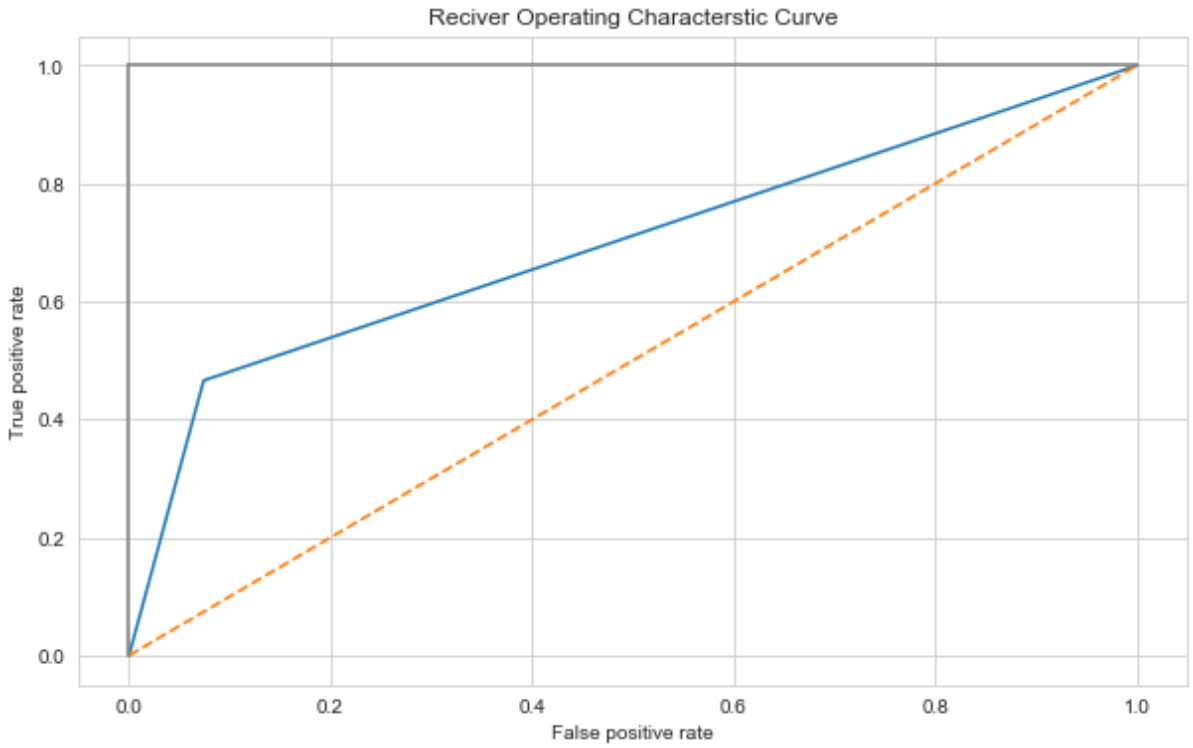
In Recall 1‘s represent 47% and this says that there are more misclassified actually positive

In F1 score 0‘s represent 93% of positively classified classes are actually positive and correct

In F1 score 1‘s represent 46% of positively classified classes are actually positive and correct

Accuracy Score:

The Accuracy Score is 0.87



ROC Score:

The Score of Receiver operating characteristic is 0.69

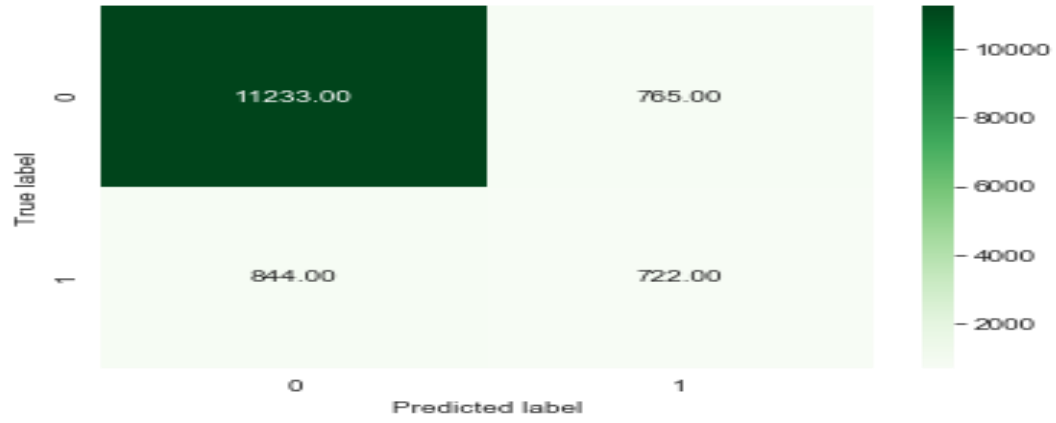
Entropy:

Entropy is a measure of the energy dispersal in the system. We see evidence that the universe tends toward highest entropy many places in our lives.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Confusion matrix

Model 2: Criterion='Entropy'



Evaluation Metrics

Confusion matrix =
[[11233 765]
[844 722]

Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".
True Negatives (Top-Left Square) 11233	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 722	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 765	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 844	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.

Metrics Classification Report

	Precision	recall	f1-score	support
0	0.93	0.94	0.93	11998
1	0.49	0.46	0.47	1566
Micro avg	0.88	0.88	0.88	13564
Macro avg	0.71	0.70	0.70	13564
Weighted avg	0.88	0.88	0.88	13564

Accuracy Score:

The Accuracy Score is 0.88

Inference:

In Precision 0‘s represent 93% of classes are ladled correctly and are positive

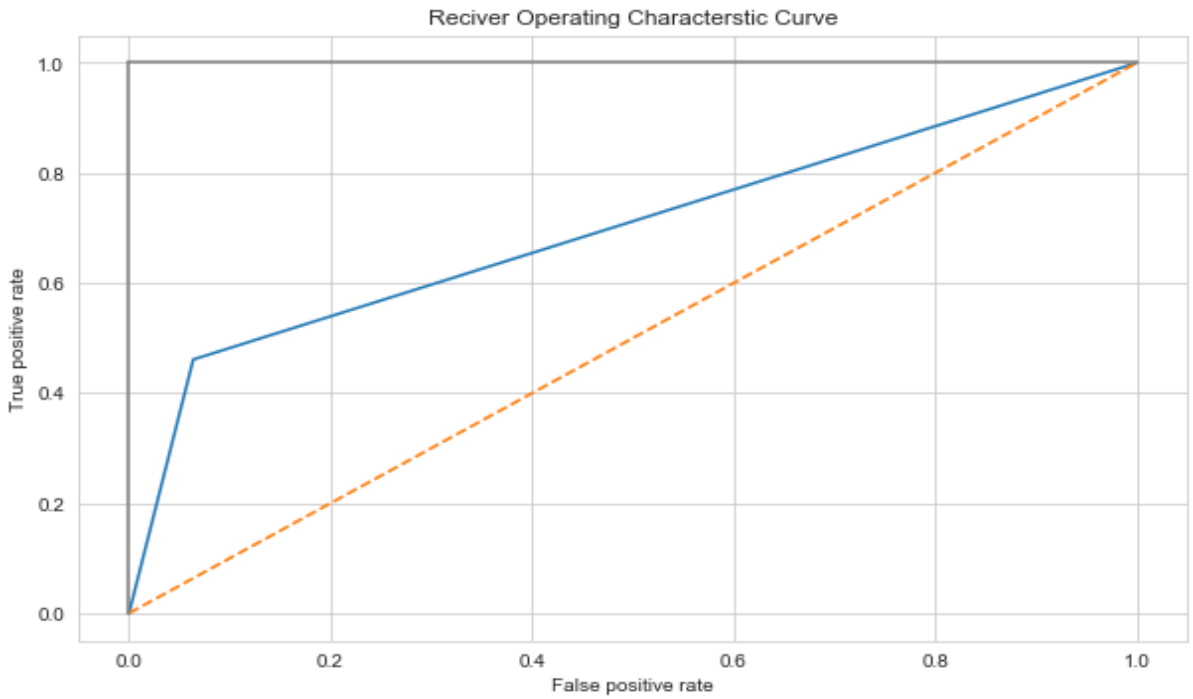
In precision 1’s represent 49% of classes are ladled correctly and are positive

In Recall 0‘s represent 94% and this says that there are less misclassified actually positive

In Recall 1‘s represent 46% and this says that there are more misclassified actually positive

In F1 score 0‘s represent 93% of positively classified classes are actually positive and correct

In F1 score 1‘s represent 47% of positively classified classes are actually positive and correct



ROC Score:

The Score of Receiver operating characteristic is 0.69

Pruning:

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of over fitting.

Evaluation Metrics

```
Confusion matrix =
 [[15567   390]
 [1435   693]]
```

Confusion matrix

Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".
True Negatives (Top-Left Square) 15567	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 693	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 390	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 1435	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.

Metrics Classification Report

	Precision	recall	f1-score	support
0	0.92	0.97	0.95	11998
1	0.62	0.37	0.46	1566
Accuracy			0.90	13564
Macro avg	0.77	0.67	0.70	13564
Weighted avg	0.89	0.90	0.89	13564

Inference:

In Precision 0's represent 92% of classes are ladled correctly and are positive

In precision 1's represent 62% of classes are ladled correctly and are positive

In Recall 0's represent 97% and this says that there are less misclassified actually positive

In Recall 1's represent 37% and this says that there are more misclassified actually positive

In F1 score 0's represent 95% of positively classified classes are actually positive and correct

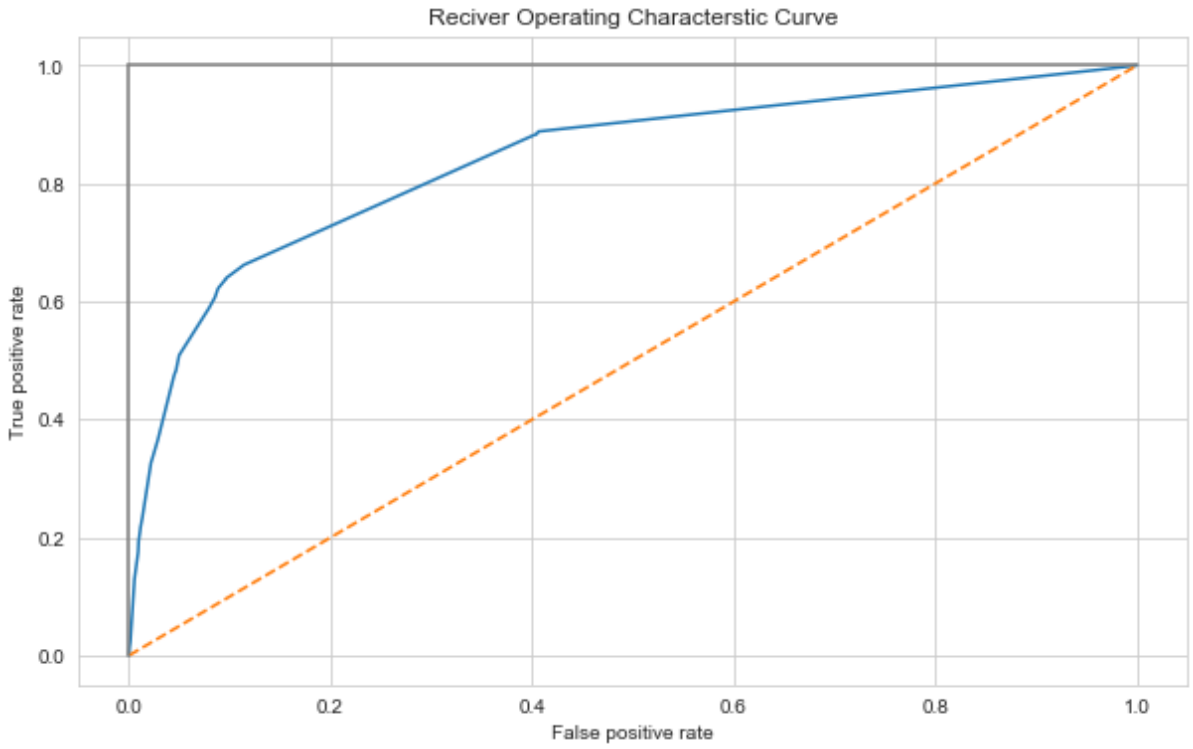
In F1 score 1's represent 46% of positively classified classes are actually positive and correct

Accuracy score

Test score: `0.9014302565614863`

Train Score: 0.9044143204727146

Inference: We can say by seeing the accuracy score for test and train the model is generalized



ROC Score:

The Score of Receiver operating characteristic is 0.66

Calculating feature Importance

	0
Last Call Duration	0.56
Outcome of Previous Campaign success	0.30
Last Contact Month mar	0.05
Age	0.04

	0
Communication Type unknown	0.02
Last Contact Month_oct	0.01
No of days passed after last contact	0.01
Last Contact Month may	0.00
Average Yearly Balance	0.00
Job management	0.00
Marital Status single	0.00
No of Campaign	0.00

Changing the cut off value for prediction

```
******** For i = 0.05 ******
Our testing recall score is 0.88
[[7162 4836]
 [183 1383]]


******** For i = 0.1 ******
Our testing recall score is 0.66
[[10629 1369]
 [529 1037]]


******** For i = 0.15 ******
Our testing recall score is 0.64
[[10840 1158]
 [564 1002]]


******** For i = 0.17 ******
Our testing recall score is 0.64
```

```
[[10840 1158]
 [564 1002]]


******** For i = 0.19 ******
Our testing recall score is 0.64
[[10840 1158]
 [564 1002]]


******** For i = 0.2 ******
Our testing recall score is 0.64
[[10851 1147]
 [568   998]]


******** For i = 0.25 ******
Our testing recall score is 0.62
[[10932 1066]
 [590   976]]


******** For i = 0.3 ******
Our testing recall score is 0.59
[[11046   952]
 [644   922]]


******** For i = 0.4 ******
Our testing recall score is 0.48
[[11446   552]
 [813   753]]


******** For i = 0.5 ******
Our testing recall score is 0.37
[[11654   344]
 [993   573]]

******** For i = 0.6 ******
Our testing recall score is 0.21
[[11864   134]
 [1232   334]]


******** For i = 0.7 ******
Our testing recall score is 0.13
[[11934    64]
 [1365   201]]


******** For i = 0.8 ******
Our testing recall score is 0.02
[[11986    12]
 [1536    30]]


******** For i = 0.9 ******
```

```
Our testing recall score is 0.01
[[11991     7]
 [1544    22]]


******** For i = 0.99 ******
Our testing recall score is 0.00
[[11996     2]
 [1565     1]]
```

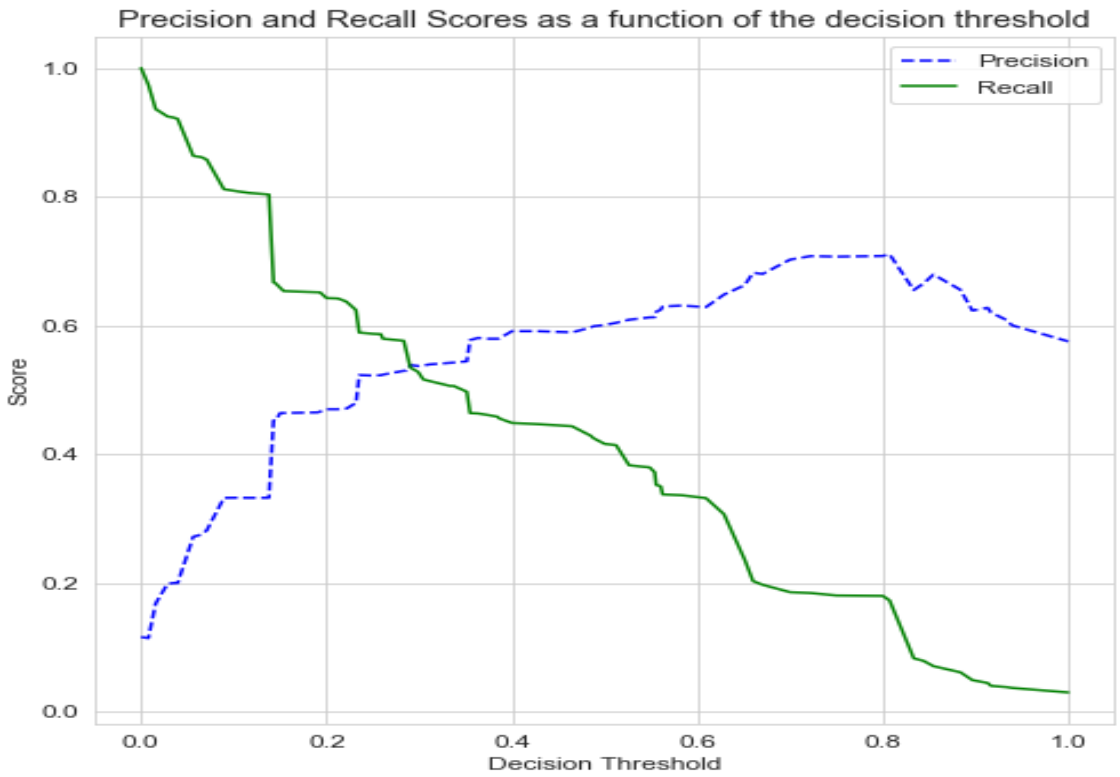
Grid Search CV:

Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

Grid Search CV Result:

	mean_test_precision score	mean_test_recall_ score	mean_test_ accuracy score	Param _max_depth
0	0.66	0.32	0.90	3
1	0.65	0.37	0.90	4
2	0.64	0.36	0.90	5
4	0.63	0.40	0.90	7
3	0.63	0.38	0.90	6
5	0.61	0.41	0.90	8

Precision and Recall Vs. Threshold



Random Forest:

This is a classifier that evolves from decision trees. It actually consists of many decision trees. To classify a new instance, each decision tree provides a classification for input data; random forest collects the classifications and chooses the most voted prediction as the result. The input of each tree is sampled data from the original dataset. In addition, a subset of features is randomly selected from the optional features to grow the tree at each node. Each tree is grown without pruning. Essentially, random forest enables a large number of weak or weakly-correlated classifiers to form a strong classifier.

Metrics Classification Report

	Precision	recall	f1-score	support
0	0.92	0.97	0.94	11998
1	0.61	0.33	0.43	1566
Micro avg	0.90	0.90	0.90	13564
Macro avg	0.77	0.65	0.69	13564
Weighted avg	0.88	0.90	0.89	13564

Inference:

In Precision 0‘s represent 92% of classes are ladled correctly and are positive

In precision 1’s represent 61% of classes are ladled correctly and are positive

In Recall 0‘s represent 97% and this says that there are less misclassified actually positive

In Recall 1‘s represent 33% and this says that there are more misclassified actually positive

In F1 score 0‘s represent 94% of positively classified classes are actually positive and correct

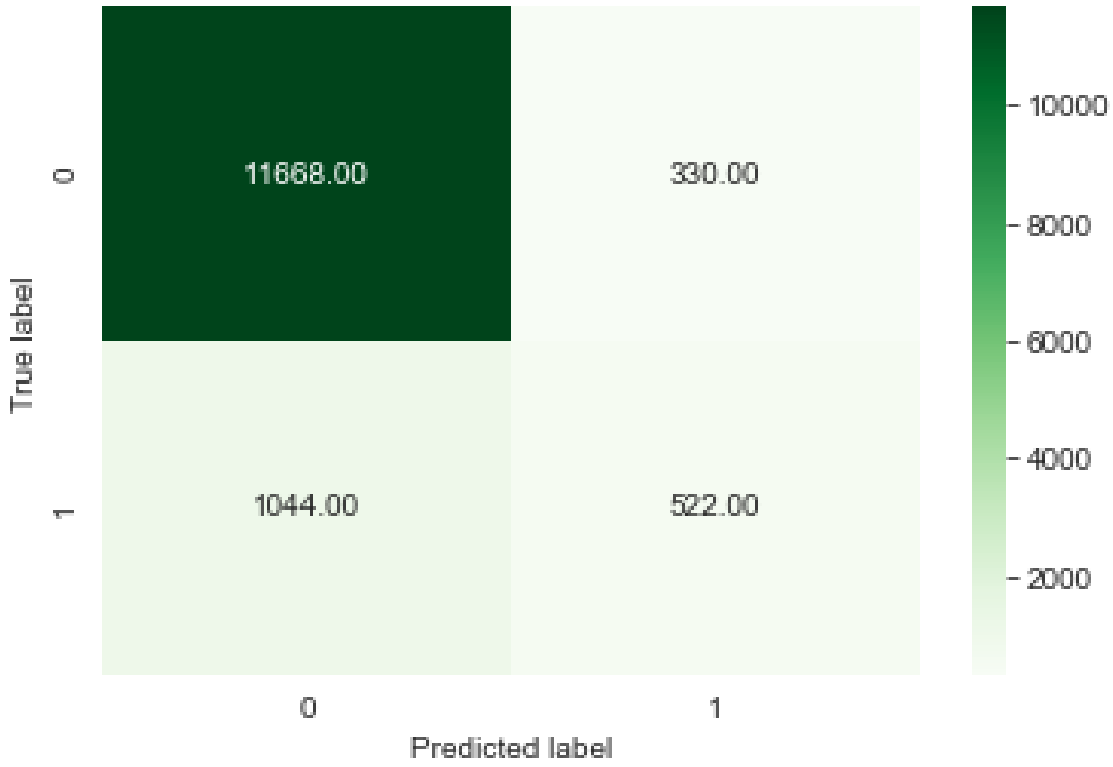
In F1 score 1‘s represent 43% of positively classified classes are actually positive and correct

Evaluation Metrics

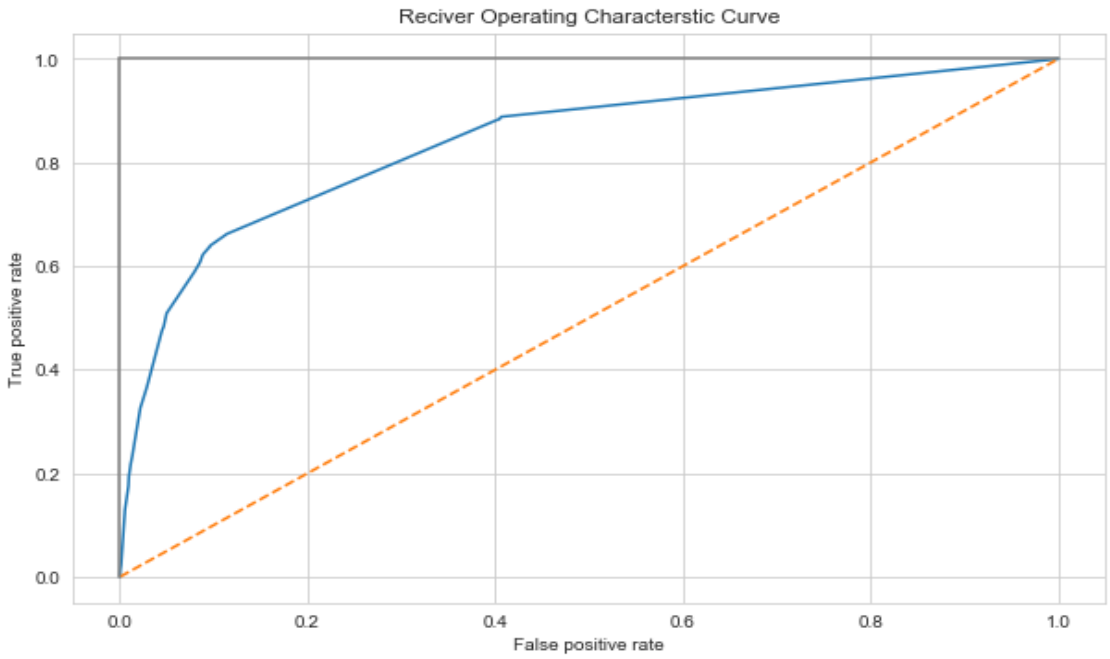
Confusion matrix

```
[[11668   330]
 [1044   522]]
```

Confusion matrix



Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".
True Negatives (Top-Left Square) 11668	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 522	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 330	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 1044	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.



ROC Score:

The Score of Receiver operating characteristic is 0.65

Important Features:

Gini importance of each feature

('Age', 0.097063957418236965)

('Job', 0.098550874785543766)

('Marital Status', 0.085711175613200075)

('Education', 0.27275931521830082)

('Credit Default', 0.039256644004119261)

('Average Yearly Balance', 0.039746094814727115)

('Housing Loan', 0.020774255620850269)

('Personal Loan', 0.0090462390534351254)

('Communication Type', 0.0036067932128967129)

('Last Contact Day', 0.0031890395355900561)

('Last Contact Month', 0.010812713099314098)

('Last Call Duration', 0.0052781449223500499)

('No of Campaign', 0.0046543523129383615)

('No of days passed after last contact', 0.0064090539837337035)

('Previous Contact before Campaign', 0.0050217904762418688)

('Outcome of Previous Campaign', 0.011411043547778171)

CROSS VALIDATION:

Cross validation is a powerful tool that is used for estimating the predictive power of your model, and it performs better than the conventional training and test set. Using cross validation, we can create multiple training and test sets and average the scores to give us a less biased metric.

K-Fold Cross Validation:

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

CHECK K-Fold Cross Validation

Cross validation run 0: 0.899

Cross validation run 1: 0.908

Cross validation run 2: 0.894

Cross validation run 3: 0.889

Cross validation run 4: 0.902

Cross validation run 5: 0.897

Cross validation run 6: 0.899

Cross validation run 7: 0.902

Cross validation run 8: 0.898

Cross validation run 9: 0.900

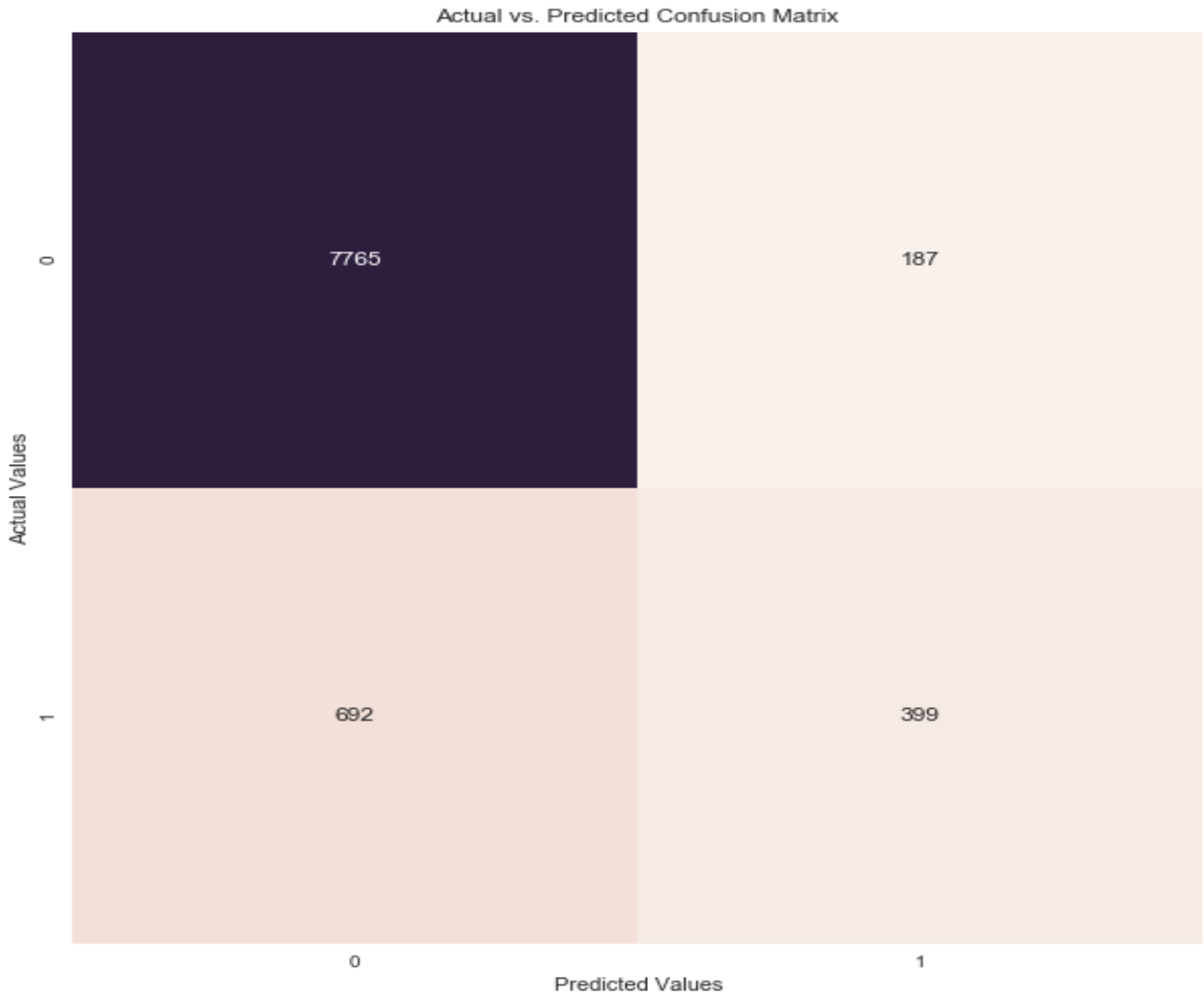
Accuracy: 0.899 (+/- 0.002)

Evaluation Metrics

Confusion matrix

```
[[7765   187]
 [692    399]]
```

Confusion matrix



Positive and Negative class	The class of interest here is customers who subscribe for a term deposit. This is the positive class represented by "Yes".
True Negatives (Top-Left Square) 7765	This is the number of customers who have not subscribed for the term deposit and have been classified correctly by the model
True Positives (Bottom-Right Square) 399	This is the number of customers who have subscribed for the loan and have been correctly classified by the model
False Positives (Top-Right Square) 187	This is an in corrected classification by the model where customers who will not subscribe have been predicted as those who will subscribe for the term deposit.
False Negatives (Bottom-left Square) 692	This is an in corrected classification by the model where customers who will subscribe for the term deposit but the model predicts that they will not subscribe. For our use case, this is the most serious error since this will result in lost customers.

Classification Report for Random Forest:

	Precision	recall	f1-score	support
0	1.00	0.88	0.94	9042
1	0.00	1.00	0.00	1
Micro avg	0.88	0.88	0.88	9043
Macro avg	0.50	0.94	0.47	9043
Weighted avg	1.00	0.88	0.94	9043

Inference:

In Precision 0‘s represent 100% of classes are ladled correctly and are positive

In precision 1’s represent 0% of classes are ladled correctly and are positive

In Recall 0‘s represent 0.88% and this says that there are less misclassified actually positive

In Recall 1‘s represent 100% and this says that there are more misclassified actually positive

In F1 score 0‘s represent 94% of positively classified classes are actually positive and correct

In F1 score 1‘s represent 0% of positively classified classes are actually positive and correct

Accuracy score:

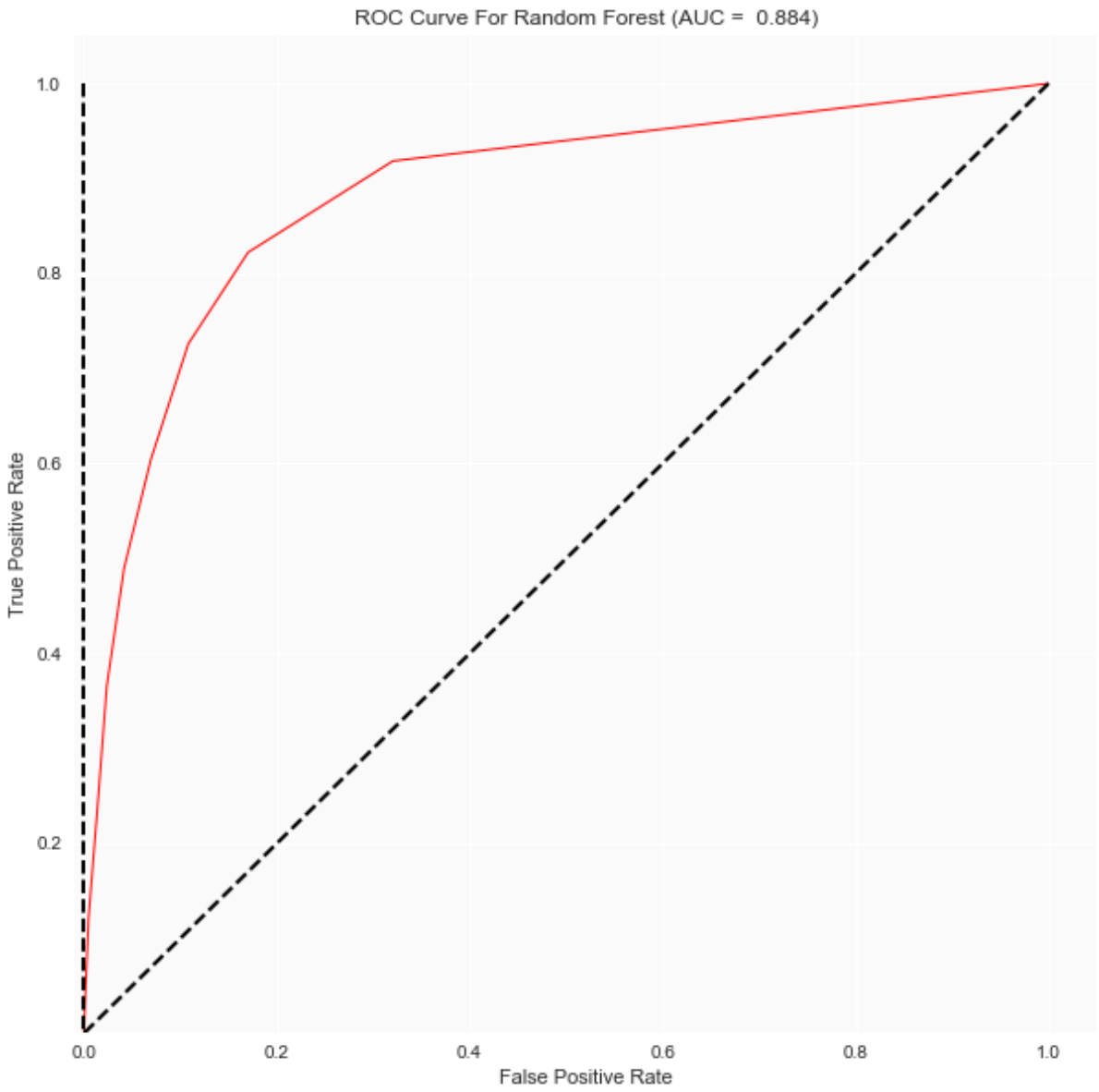
Here is our mean accuracy on the test set: 0.903

Calculate the test error rate

The test error rate for our model is: 0.0972

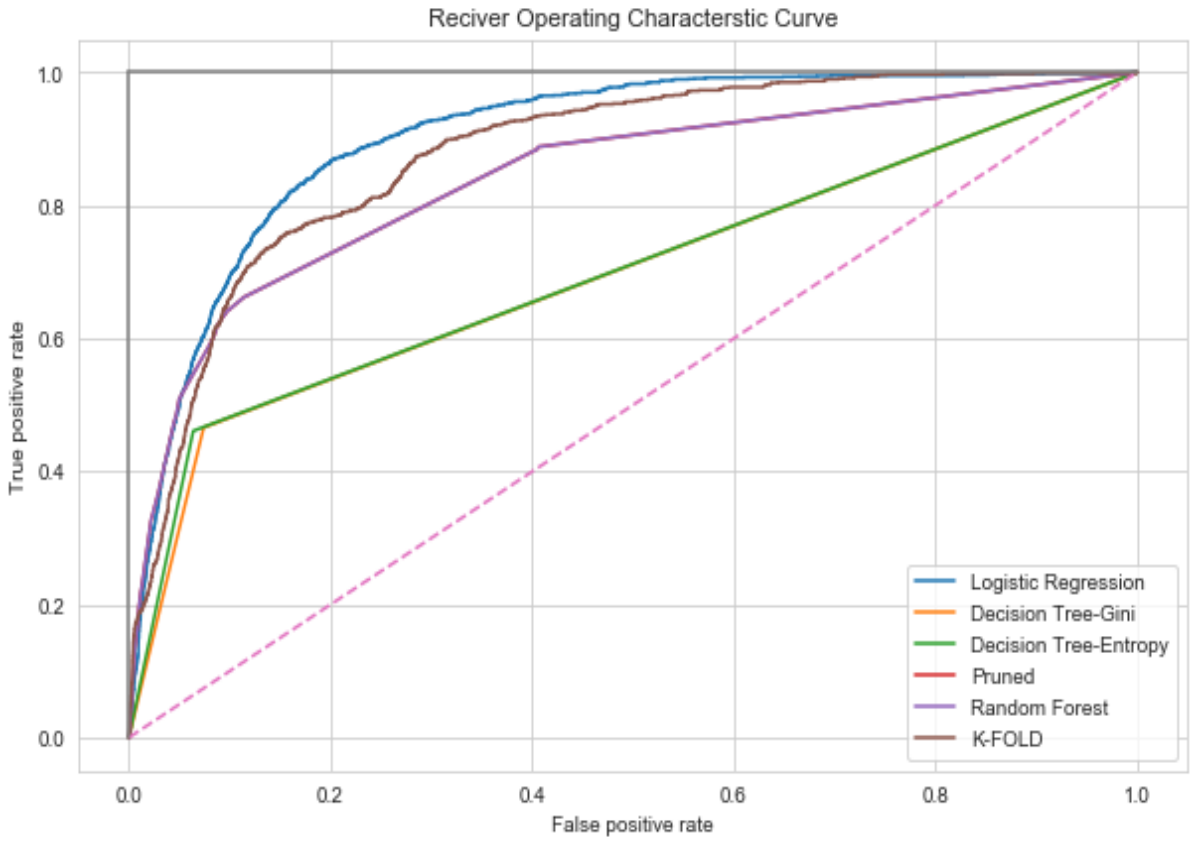
ROC Curve Metrics

A receiver operating characteristic (ROC) curve calculates the false positive rates and true positive rates across different thresholds. Let's graph these calculations.



ROC Score:

The Score of Receiver operating characteristic is 0.50



TEST Model comparison:

Model	Accuracy	Precision	Recall	F1 score	ROC
Logistic regression	0.90	0.64	0.34	0.44	0.66
CART (Decision Tree) Criterion='gini'	0.87	0.45	0.47	0.46	0.69
CART (Decision Tree) Criterion='Entropy'	0.88	0.49	0.46	**0.47**	0.69
Pruning	0.90	0.62	0.37	0.46	0.66
Random forest	0.90	0.61	0.33	0.43	0.65
Random Forest(K Fold=10)	0.88	0	1	0	0.50

FINAL INFERENCES

- For every model the precision is in between 92-93%, here only random forest k fold is over fitted.
  While recall is varying significantly.
  Accuracy of the models are similar
  F1 score is also similar
  ROC scores are slightly varying
  By considering all above metrics we can choose either logistic Regression or Random Forest K fold
  Here logistic regression performs better than the random forest K-Fold because in our case random forest K-Fold precision is over fitted and ROC score is less when compared to Logistic Regression
  So finally we are choosing over random forest k fold as it is giving better F1 score and ROC score

TRAIN Model comparison:

Model	Accuracy	Precision	Recall	F1 score	ROC
Logistic regression	0.90	0.65	0.35	0.46	0.66
CART (Decision Tree) Criterion='gini'	1	0.10	1	1	0.69
CART (Decision Tree) Criterion='Entropy'	1	1	1	1	0.69
Pruning	0.90	0.66	0.39	0.49	0.66
Random forest	0.99	1	0.93	**0.96**	0.65
Random Forest(K Fold=10)	0.88	1	0	0	0.50

Inference:

- Here CART (Decision Tree) Criterion='gini' and 'entropy' is over fitting so to deal with that were using pruning
- Scores for logistic regression and pruning is almost similar
- Random Forest is overfitting
- Random Forest K fold ROC score is less

What type of customers Subscribe term deposits?

Based on the odds ratio from the Logistic regression model and the variable importance from the Decision tree and Random forest models, the following are the potential types of customers who will subscribe for the term deposit:

1. Customers who have subscribed from previous campaigns have a very high likelihood to subscribe for the term deposit.

2. Customers contacted in specific months – March, October, September, December, June are more likely to subscribe to the term deposit.

3. Students and retired people have a higher probability to convert.