Term Deposit Prediction for a Banking Institution

Table of Contents

Problem Setting:

A Banking institution conducted marketing campaigns through phone calls to promote term deposits among existing customers. Because they enable companies to target specific client groups with specialized promotions and offers, direct marketing initiatives like phone call campaigns are crucial for businesses. The bank's goal was to retain customers and increase revenue from term deposit products. The bank may be able to enhance term deposit product sales and client retention by using this kind of promotion. Additionally, the bank can gain valuable insights into customer behavior client behavior, and preferences through data analysis of these campaigns, which can help it make smarter judgments for future marketing initiatives. Overall, running marketing campaigns is essential to the success of a business since it helps to both attract and keep customers.

Problem Definition:

The problem definition states that the objective of the project is to forecast whether a client of a banking institution would subscribe to a term deposit, based on past data gathered from direct marketing efforts that were carried out via phone calls.

In other words, the project aims to build a predictive model that can accurately predict whether a client would subscribe to a term deposit, given certain features or variables that describe the client and the marketing campaign. The term deposit is a financial product offered by the banking institution, where a client deposits a fixed amount of money for a fixed period of time, typically ranging from a few months to a few years.

Data Source:

The dataset has been taken from UCI Machine Learning Repository, an open-source repository for research data (https://archive-beta.ics.uci.edu/dataset/222/bank+marketing). The dataset used in this project is related to direct marketing campaigns carried out by a banking institution via phone calls. The dataset contains information about the clients and the marketing campaign, as well as the outcome of the campaign, i.e., whether the client subscribed to a term deposit or not.

Data Description:

The data provided contains information on 41188 customers and 20 input variables. The input variables include information on the Bank's client data such as customer's age, job, marital status, etc., last contact related information such as last contact month and day of the week, duration of the last call, etc., Social and economic context data such as employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate, and number of employees. The target variable, 'y', is a binary variable indicating whether the client subscribed to a term deposit or not.

Input variables:

Bank client data:

1. age (numeric)
2. job: type of job (categorical: 'admin', 'blue collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4. education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. default: has credit in default? (Categorical: 'no', 'yes', 'unknown')
6. housing: has housing loan? (Categorical: 'no', 'yes', 'unknown')
7. loan: has personal loan? (Categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

8. contact: contact communication type (categorical: 'cellular', 'telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

Social and economic context attributes

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3-month rate - daily indicator (numeric)
20. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21. y - has the client subscribed a term deposit? (Binary: 'yes', 'no')

Data Exploration:

Data Sample: Displaying the sample data

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.pri
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	334	1	999	0	nonexistent	-1.1	
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	383	1	999	0	nonexistent	-1.1	
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	189	2	999	0	nonexistent	-1.1	
41186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	442	1	999	0	nonexistent	-1.1	
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	239	3	999	1	failure	-1.1	

41188 rows × 21 columns

- Data Type Correction

Data Type Correction is an important step in the data pre-processing phase where we check the data types and null count of each variable in the dataset. This is necessary because we need to update the appropriate data type for each variable for mathematical operations.

```
bank_add.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

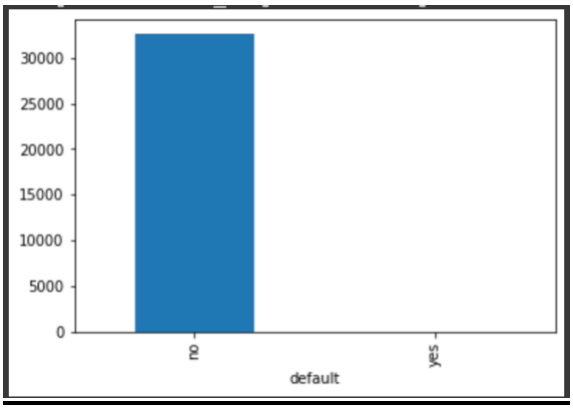
- Handling Missing Values

In Categorical variables some values are “Unknown”. So, replacing all the unknown as NULL and some default values are replaced as NULLS in numerical variables such as 999 in pdays.

Percentage of NULL values in each column:

```
age               0.000000
job               0.801204
marital           0.194231
education         4.202680
default          20.872584
housing           2.403613
loan              2.403613
contact           0.000000
month             0.000000
day_of_week       0.000000
duration          0.000000
campaign          0.000000
pdays            96.321744
previous          0.000000
poutcome          0.000000
emp.var.rate      0.000000
cons.price.idx    0.000000
cons.conf.idx     0.000000
euribor3m         0.000000
nr.employed       0.000000
y                 0.000000
dtype: float64
```

1. Removing the variables which have more than 80% NULL values.
2. We removed ‘default’ column because 99.9% of values are ‘no’.
3. Removed the rows if there is NULL value because its small amount.



- Handling Outlier

  We have outliers in duration column. However, we cannot use duration as a feature because the duration is not known before a call is performed. Also, after the end of the call y is obviously known. So, we remove duration column.

Statistics of the Variables:

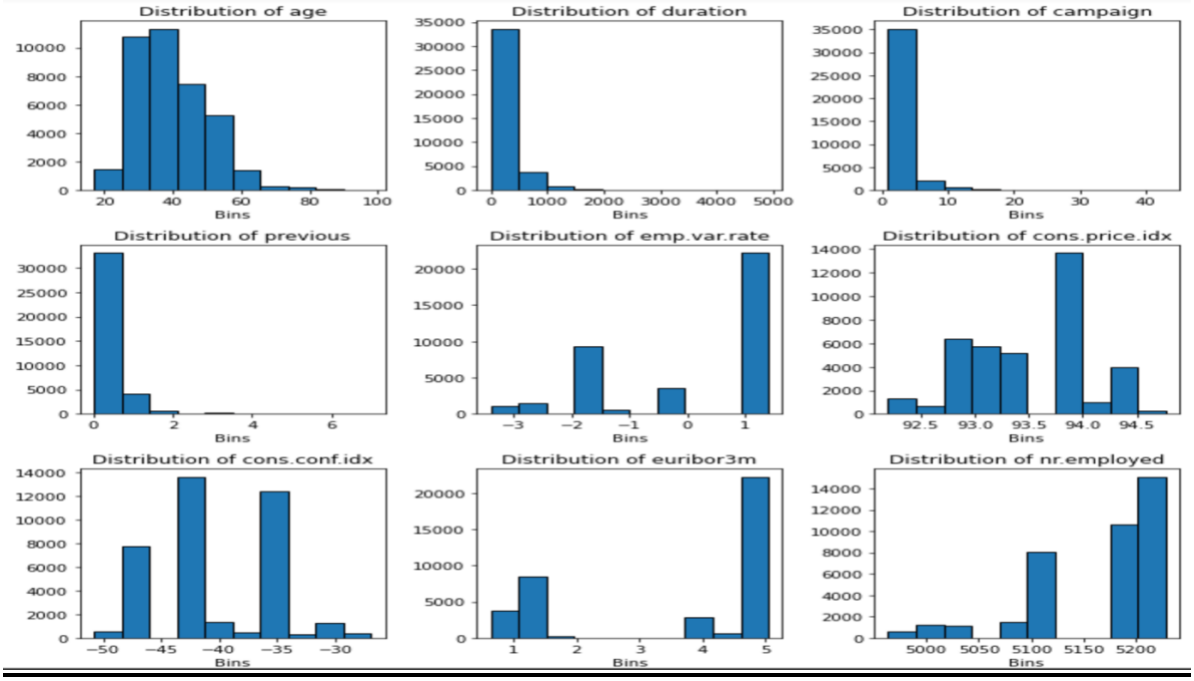
- Descriptive Statistics of numerical variables

  This table gives the detail understanding of every numerical values regarding there mean, median etc. and also gives information on outliers

	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	38245.000000	38245.000000	38245.000000	38245.000000	38245.000000	38245.000000	38245.000000	38245.000000	38245.000000
mean	39.860871	258.207583	2.566662	0.170009	0.082861	93.570313	-40.541164	3.623298	5167.432566
std	10.289488	259.792638	2.767473	0.487169	1.565945	0.576367	4.623200	1.730226	71.760333
min	17.000000	0.000000	1.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.000000	102.000000	1.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.000000	180.000000	2.000000	0.000000	1.100000	93.444000	-41.800000	4.857000	5191.000000
75%	47.000000	319.000000	3.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.000000	4918.000000	43.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

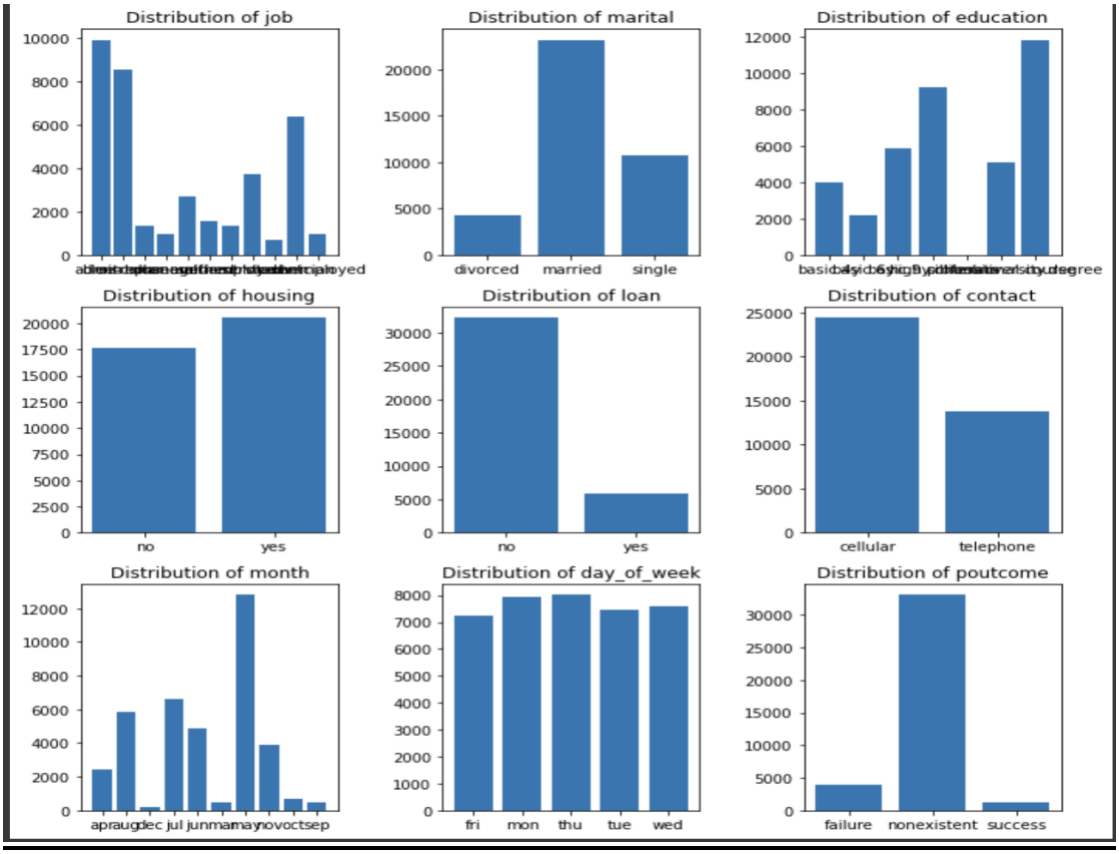
- Distribution of numerical variables

  These plots help us understand how the values are distributed across different ranges. They will give information on how skewed or symmetric the distribution spread.



- <u>Distribution of categorical variables</u>

  These plots help us understand how each category is distributed in every variable. It will be of no use if there is only one category in almost entire data.

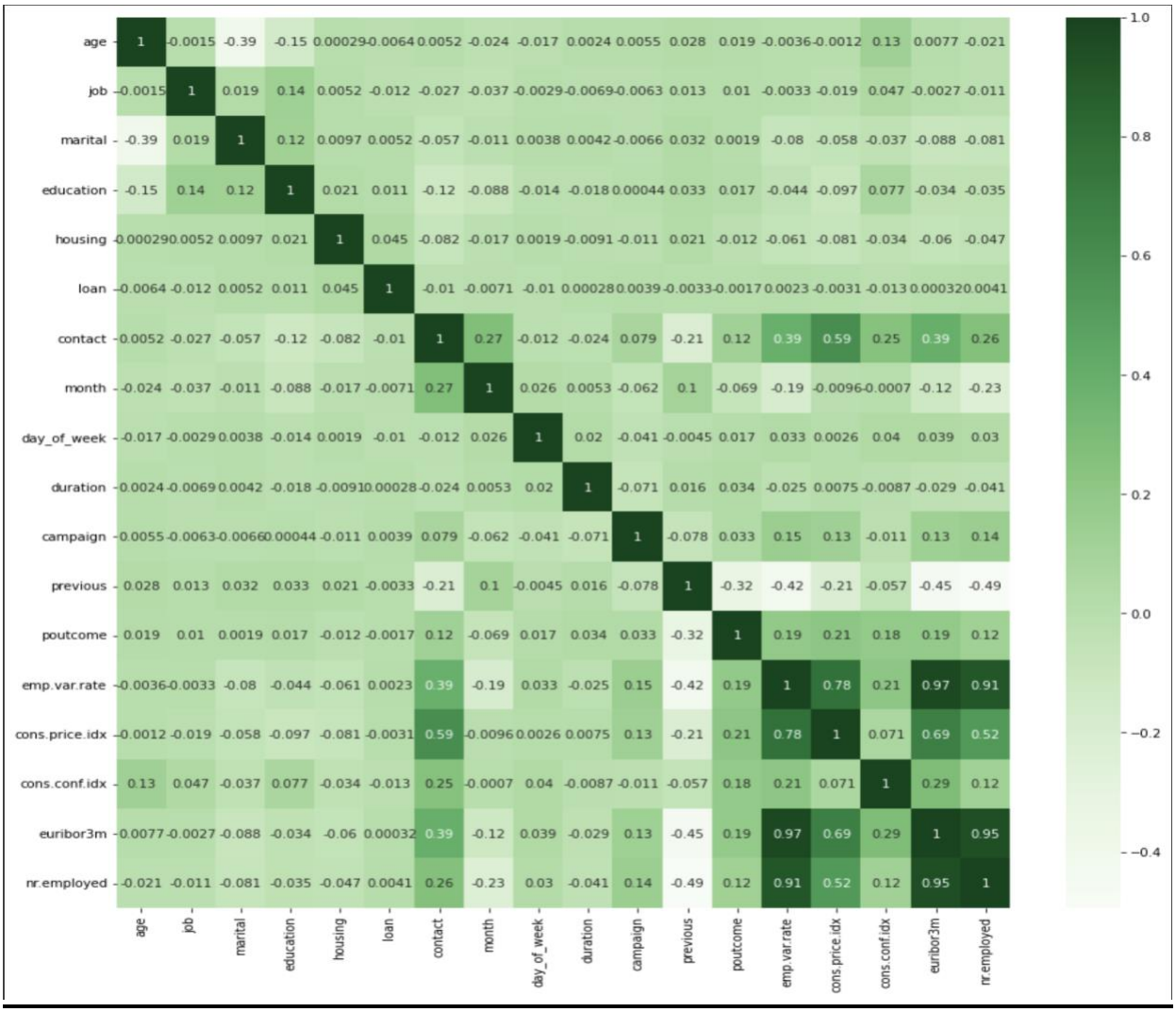


Correlation Analysis:

We perform correlation analysis to remove redundant variables. Performed ‘Label Encoding’ on categorical columns to check the redundancy between them.

Plotted the correlation heat map which clearly helps us in finding highly correlated variables.

As we can see, 'euribor3m' & 'nr.employed' are highly correlated with ‘nr.employed’. So, we are dropping those two columns.



One Hot Encoding:

Performed one hot encoding on categorical variables so that they cn be used as features while building the model.

Exploration of Data Mining Models:

The target variable in our problem statement is predicting whether or not a customer is going to take term deposit or not. So, it's a classification model. We are exploring the possible candidate models that are suitable for our analysis.

The models are:

Logistic Regression:

It's a classification model that is used in cases where the outcome variable can take only one of two possible values. It is also efficient and can handle large datasets well. Since our outcome variable has only two possibilities, we can use this model.

Advantages:

- It is very easy to intrepret. The coefficients of this model can be easily intrepreted and provide insights into the magnitude and direction of the relationship between independent and dependent variables
- This model can handle wide range of variables, including both continuous and categorical.

Disadvantages:

- This model is very prone to overfitting when there are too many variables or fewer records and sensitive to outliers.
- It assumes the relation between independent variables are completely linear or independent. This may not always hols true in real life scenarios.

Support Vector Machines (SVM):

This model is good for high-dimensional data and can handle both linear and non-linear relationships. It uses a hyperplane to separate data into different classes and tries to maximize the margin between the hyperplane and the closest data points. It is also efficient and can handle large datasets.

Advantages:

- SVM works well when there is a clear margin of separation between the two possibilities.
- It works well with high dimension data.

Disadvantages:

- SVMs can be sensitive to the choice of hyperparameters and may require significant preprocessing of the data.
- It is not suitable for large datasets.
- It might not perform well when there is too much noise in the dataset.

Decision Trees:

This model is good for creating intuitive and interpretable models that can handle both numerical and categorical data.

Advantages:

- Decision trees are better for categorical data.
- It deals collinearity better than SVM.
- Feature selection and data preprocessing can be done using Decision Trees.

Disadvantage:

- They have a tendency to overfit the data and may not generalize well.

Random Forest:

This is an ensemble method that uses multiple decision trees to improve accuracy and reduce overfitting.

Advantages:

- The random forest method can also handle large amounts of data with numerous different variables.
- They are typically more accurate than single decision trees.
- By using multiple decision trees, it reduces the chance of overfitting.

Disadvantages:

- The main disadvantage of random forest is that having too many decision trees an make the algorithm slow in real time.

Gradient Boosting:

This model trains many models in a gradual and sequential manner. It combines decision trees at the beginning of the process unlike Random Forest that combines at the end.

Advantages:

- It is highly accurate and can achieve good results with complex data.
- It identifies the most important features, which can be used for feature selection.

Disadvantages:

- It has several hyperparameters, so it is time consuming to find the optimal performance
- The model results are difficult to intrepret and very prone to overfitting.

We train the above-mentioned models for our classification problem and select the best-performing model based on the performance metrics performed in test data.

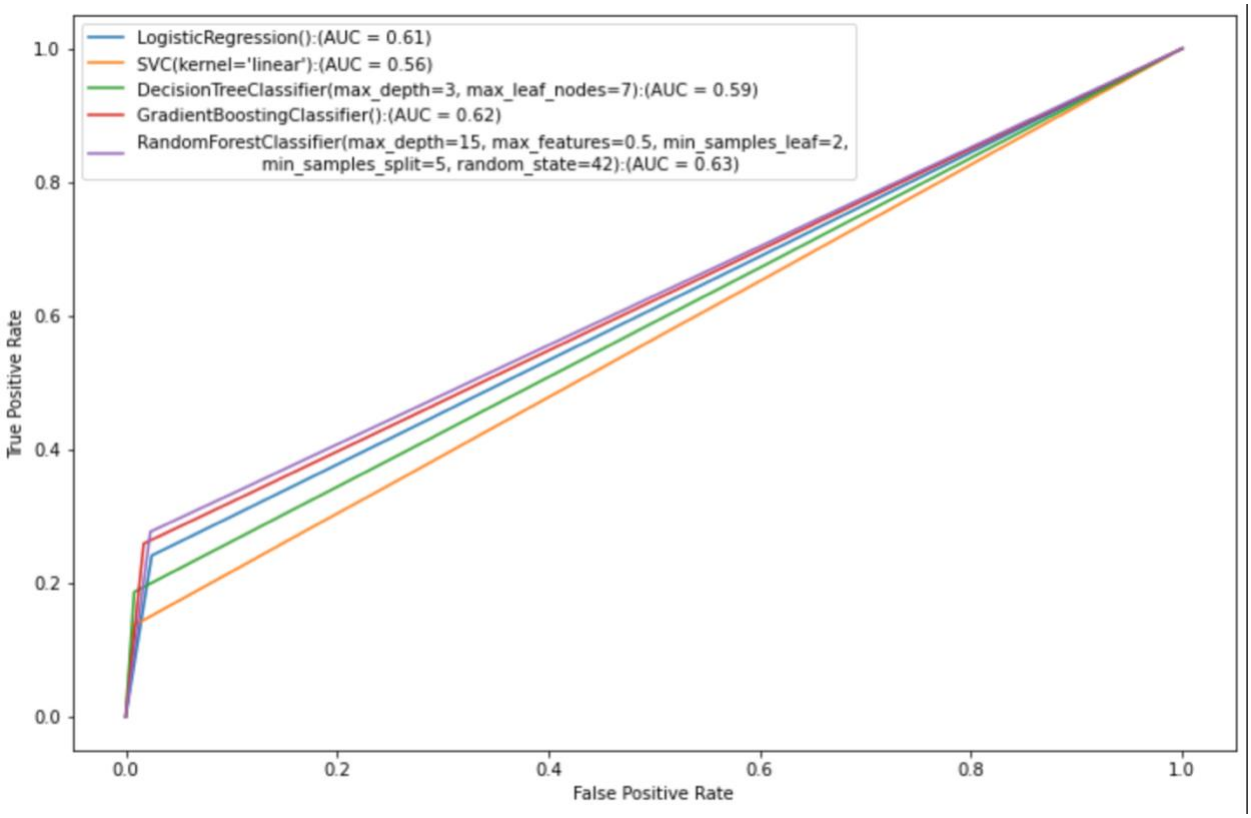
Model Performance Evaluation and Interpretation :

The target variable in our problem statement is predicting whether or not a customer is going to take term deposit or not. So, it's a classification model. After exploring the possible candidate models that are suitable for our analysis and building the models, we are evaluating the performace to choose the best fit model for this problem.

The models and their performance metrics:

Models	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.893537	0.551502	0.241088
SVM	0.897511	0.710784	0.136023
Decision Tree Classifier	0.902217	0.745318	0.186679
Gradient Boost	0.902322	0.657143	0.258912
Random Forest	0.898661	0.598377	0.276735

ROC and AUC comparison for all algorithms:



Project Result:

Looking at the table, we can see that the Decision Tree Classifier has the highest accuracy (0.902217) and precision (0.745318) score compared to other models. So, we can confirm that Decision Tree Classifier best fit model for this problem.

Conclusion:

Based on our analysis, we recommend that the banking institution use this model to predict whether a client would subscribe to a term deposit or not. Additionally, the bank can gain valuable insights into customer behavior and preferences through data analysis of these campaigns, which can help it make smarter judgments for future marketing initiatives. By identifying the most important factors that influence customer satisfaction and loyalty, businesses can better target their marketing efforts and resources towards customers who are most likely to engage with their brand and products. By identifying the most important factors that influence customer satisfaction and loyalty, businesses can focus their efforts on addressing these areas, potentially reducing customer churn and the costs associated with acquiring new customers. Overall, running marketing campaigns is essential to the success of a business since it helps to both attract and keep customers.