ASDS5303 FINAL PROJECT REPORT

TITLE: BANK MARKETING CAMPAIGN

- Presented by

JEYASOORIYA SARAVANAN (1002186838)

HARIHARAN SELVAM (1002174644)

TABLE OF CONTENTS:

ASDS 5303 FINAL PROJECT REPORT

TOPIC: Bank Marketing Campaign

PROBLEM STATEMENT:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

Often, more than one contact with the same client was required, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe to a term deposit or not. (variable y)

Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

The dataset consists of 45211 rows and 17 columns.

The goal of our project is to build a model to predict whether the client is interested in subscribing for the term deposit or not.

Our model will help to predict the future as well.

DESCRIPTIVE DATA ANALYSIS:

Descriptive analysis is a method used in statistics and data analysis to summarize and describe the key features of a dataset. The goal is to provide a clear and concise overview of the data, helping to identify patterns, trends, and key characteristics. Descriptive analysis does not involve making inferences or drawing conclusions about a population; instead, it focuses on organizing and summarizing the available information

Our original dataset looks like this.

```
df_bank_data.head()
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Target
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

```
df_bank_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   age        45211 non-null  int64 
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64 
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64 
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64 
 12  campaign   45211 non-null  int64 
 13  pdays      45211 non-null  int64 
 14  previous   45211 non-null  int64 
 15  poutcome   45211 non-null  object
 16  Target     45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Our dataset doesn't have any missing values.

Here is the information about our features and target variables.

Features:

age --> age

job --> type of Job

marital --> marital status

education --> highest education finished

default --> already has credit in default?

balance --> account balance

housing --> taken housing loan?

loan --> taken personal loan?

contact --> communication via...

day --> day of last contact

month --> month of last contact

duration --> duration of last contact

campaign --> number of contacts made to the client during the campaign

pdays --> number of days that passed by after the client was last contacted from a previous campaign (999 means client wasn't previously contacted)

previous --> number of contacts performed before this campaign and for this client

poutcome --> outcome of the previous marketing campaign

Target variable:

y --> has the client subscribed to a term deposit?

Statistical Parameters:

	count	mean	std	min	25%	50%	75%	max
age	45211.0	40.936210	10.618762	18.0	33.0	39.0	48.0	95.0
balance	45211.0	1362.272058	3044.765829	-8019.0	72.0	448.0	1428.0	102127.0
day	45211.0	15.806419	8.322476	1.0	8.0	16.0	21.0	31.0
duration	45211.0	258.163080	257.527812	0.0	103.0	180.0	319.0	4918.0
campaign	45211.0	2.763841	3.098021	1.0	1.0	2.0	3.0	63.0
pdays	45211.0	40.197828	100.128746	-1.0	-1.0	-1.0	-1.0	871.0
previous	45211.0	0.580323	2.303441	0.0	0.0	0.0	0.0	275.0

From the description of numerical features, it is clearly known that there are several outliers present in these values as the max value varies from the mean value**.**

Checking NULL values:

```
In [261]: df_bank_data.isnull().sum()
Out[261]: age          0
          job          0
          marital      0
          education    0
          default      0
          balance      0
          housing      0
          loan         0
          contact      0
          day          0
          month        0
          duration     0
          campaign     0
          pdays        0
          previous     0
          poutcome     0
          Target       0
          dtype: int64
```

There is no missing values in this dataset.

Unique Values information:

```
In [262]: df_bank_data.nunique()
Out[262]: age          77
          job          12
          marital       3
          education     4
          default       2
          balance    7168
          housing       2
          loan          2
          contact       3
          day          31
          month        12
          duration   1573
          campaign     48
          pdays       559
          previous     41
          poutcome      4
          Target        2
          dtype: int64
```

In our dataset we have 77 unique values in ‘age’, 12 in ‘job’, 3 in ‘marital’, 4 in ‘education’, 2 in ‘default’, 7168 in ‘balance’, 2 in ‘housing’, 2 in ‘loan’, 3 in ‘contact’, 31 in ‘day’, 12 in ‘month’, 1573 in ‘duration’, 48 in ‘campaign’, 559 in ‘pdays’, 41 in ‘previous’, 4 in ‘poutcome’, 2 in ‘Target’.

Missing values and null value observation:

Job has 0.63% unknown value - drop unknown rows

Education has 4.11% unknown values - drop unknown rows

Contact has 13020 unknown value - replace unknown with cellular value.

Day and month have nan value = drop nan

poutcome has 81.75% unknown values - replace unknown with others value.

Age has no null/missing value

Previous has no null/missing value

Marital has no null/missing value

After dropping all unknown values, the shape of dataset is 43193 rows and 17 columns.

Dropping the unnecessary columns:

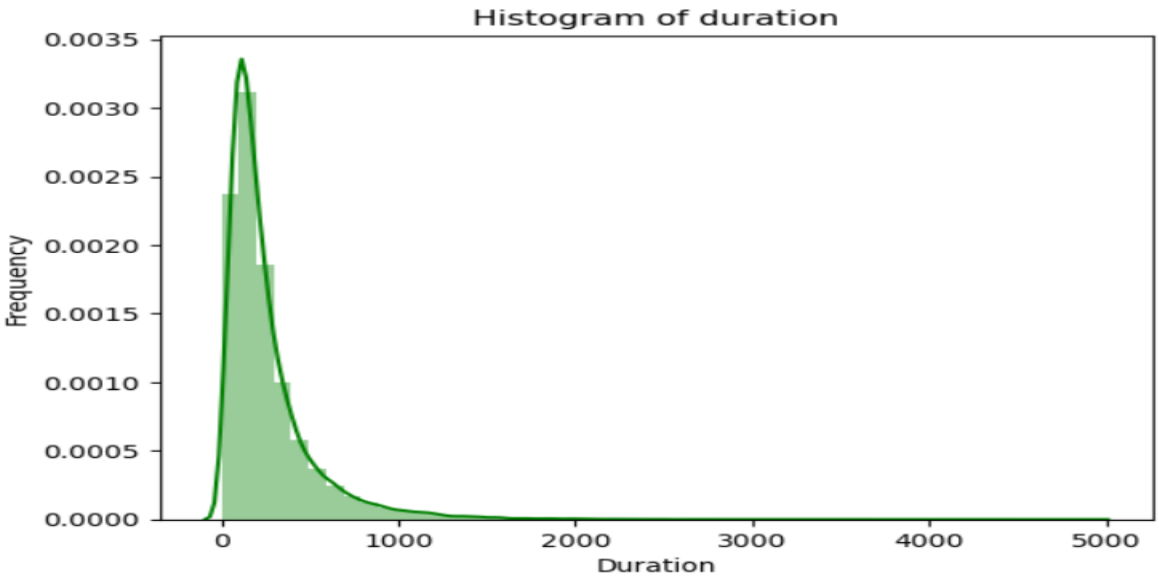
Columns ‘day’ and ‘month’ are removed as both are unsuccessful.

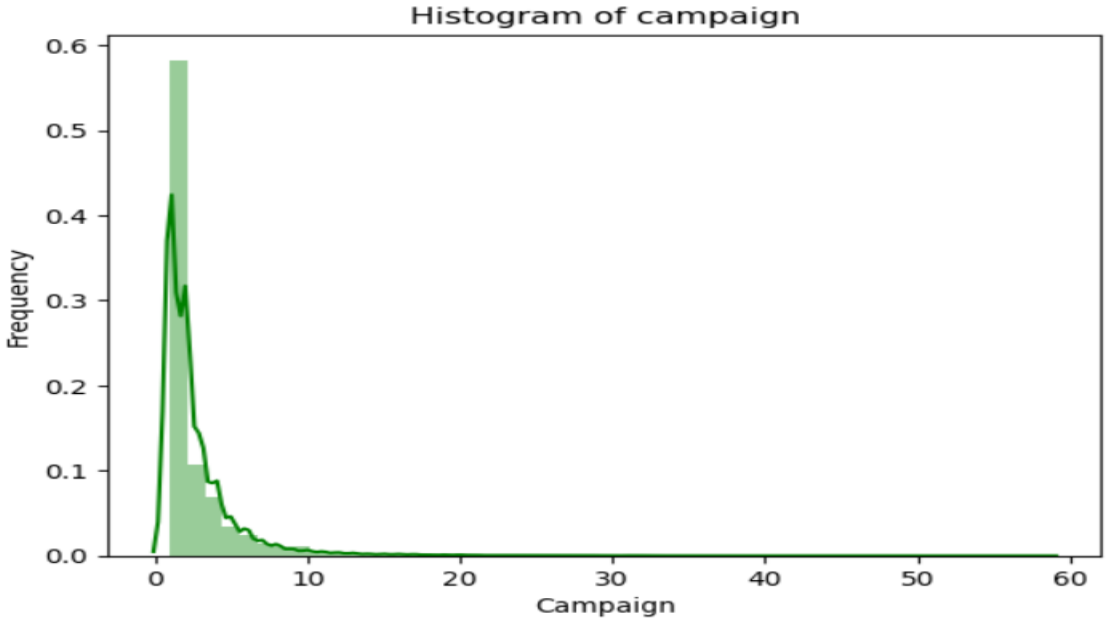
DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS:

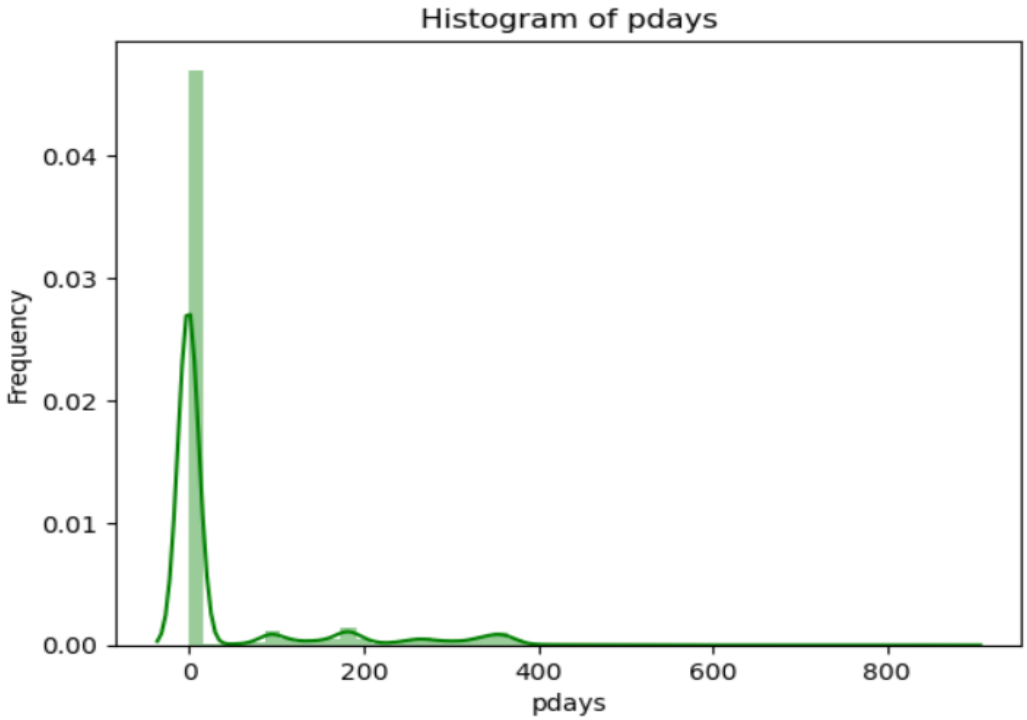
Univariate analysis:

It is used to convert numerical variables into categorical variables with the help of binning technique through histograms.

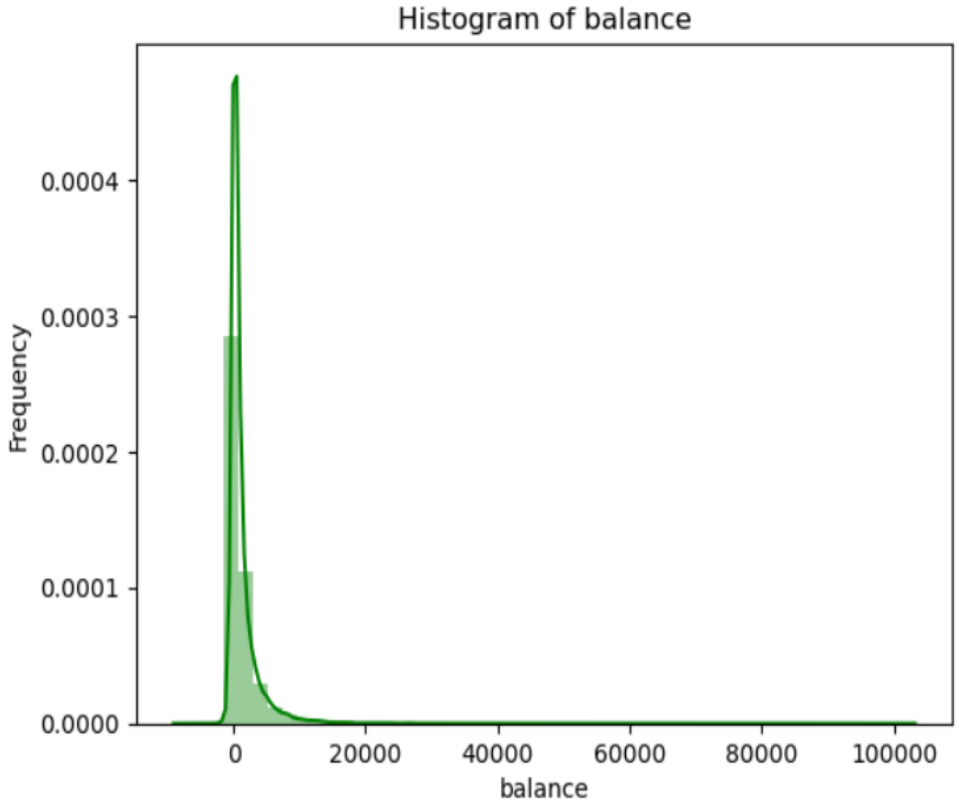
Here is the Histogram of all our numerical variables which are converted to categorical ones.






Histogram of pdays
Frequency
0.04
0.03
0.02
0.01
0.00
0
200
400
600
800
pdays


Histogram of age
Frequency
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
age



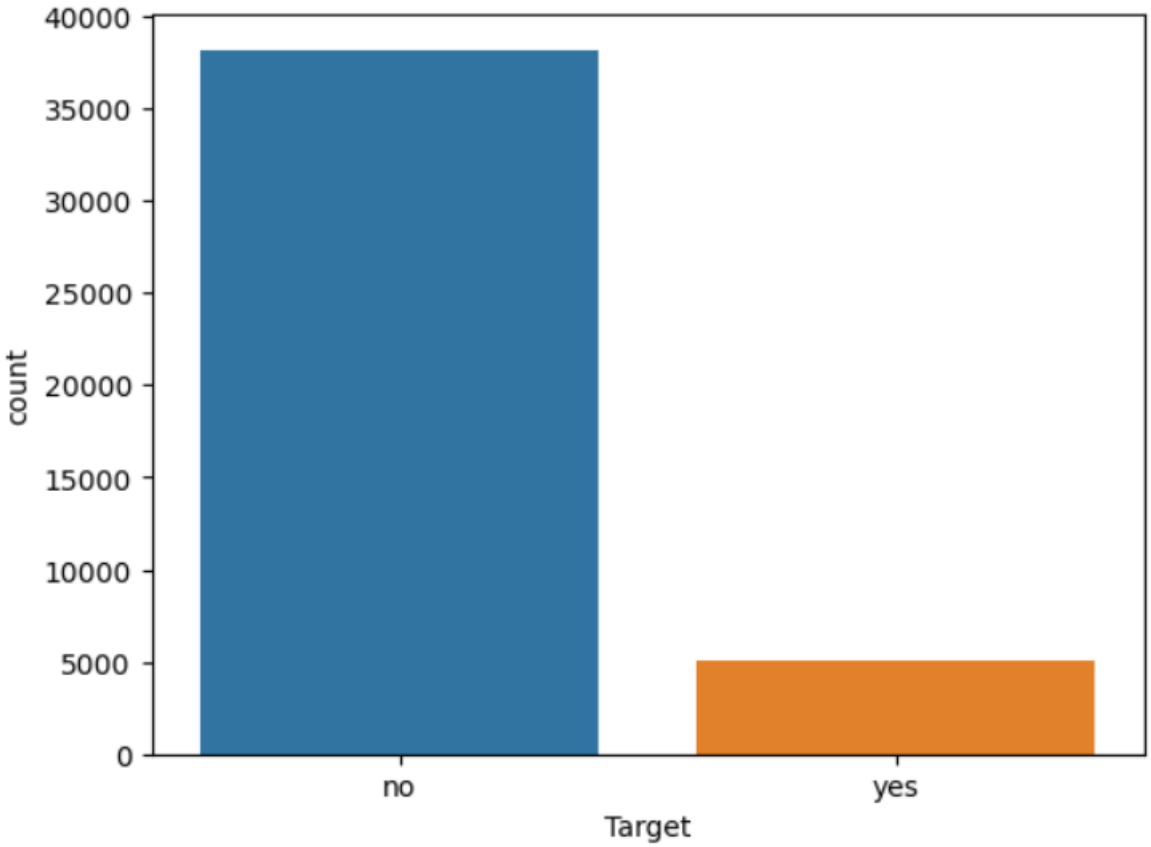
The frequency balance for people with negative balance is more when compared with the people with positive balance.

After completing the univariate analysis, we dropped those columns.

Multivariate analysis:

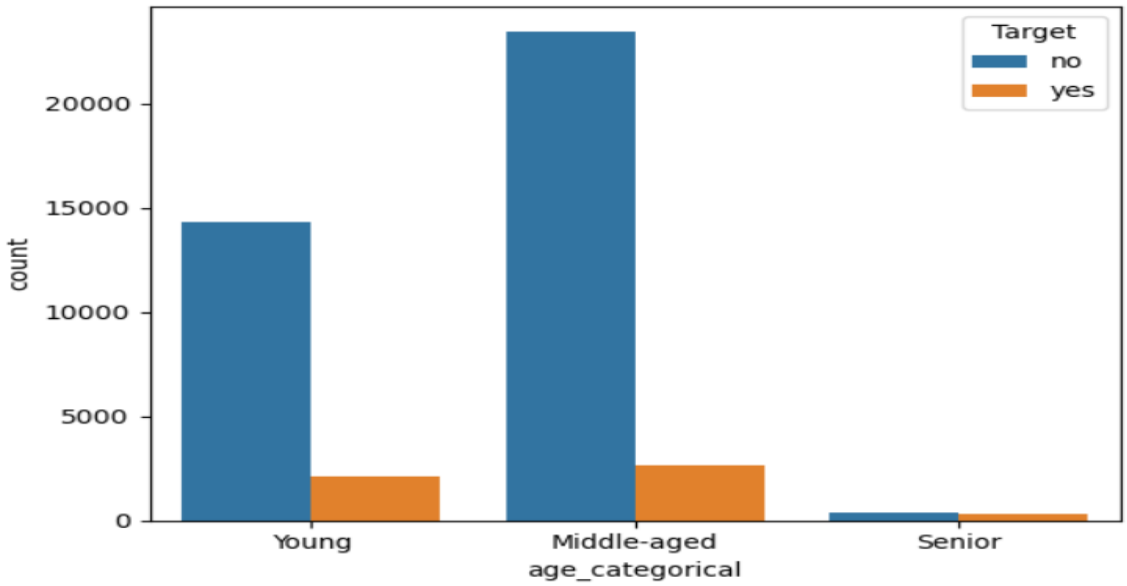
It is the process of exploring the relationship between multiple variables and the target variable.

Here is the countplot for target variable.



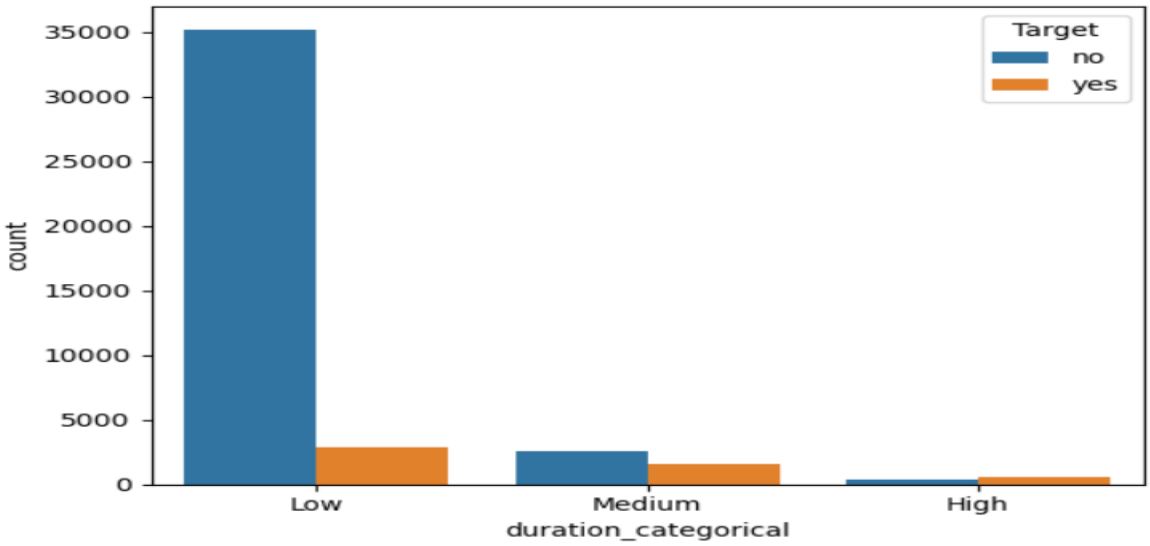
And the following countplots are used to compare the influence of features on our target.

Count plot to show the influence of age_categorical on target.



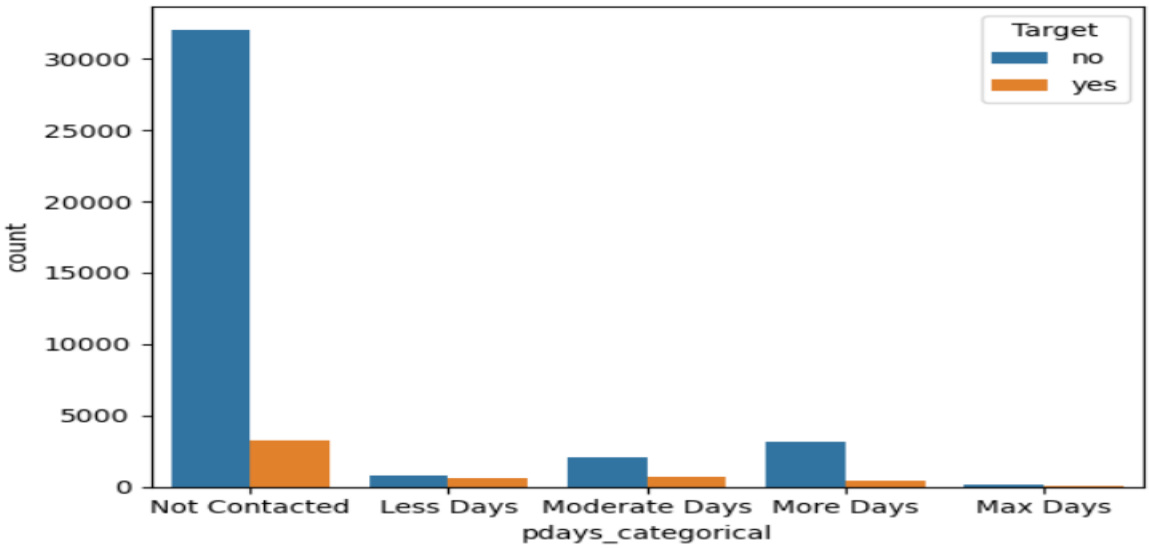
Here it is given that the most no. of clients who haven't subscribed for the term deposit fall under the categories of 'Middle aged' and 'Young'

Count plot to show the influence of duration_categorical on target



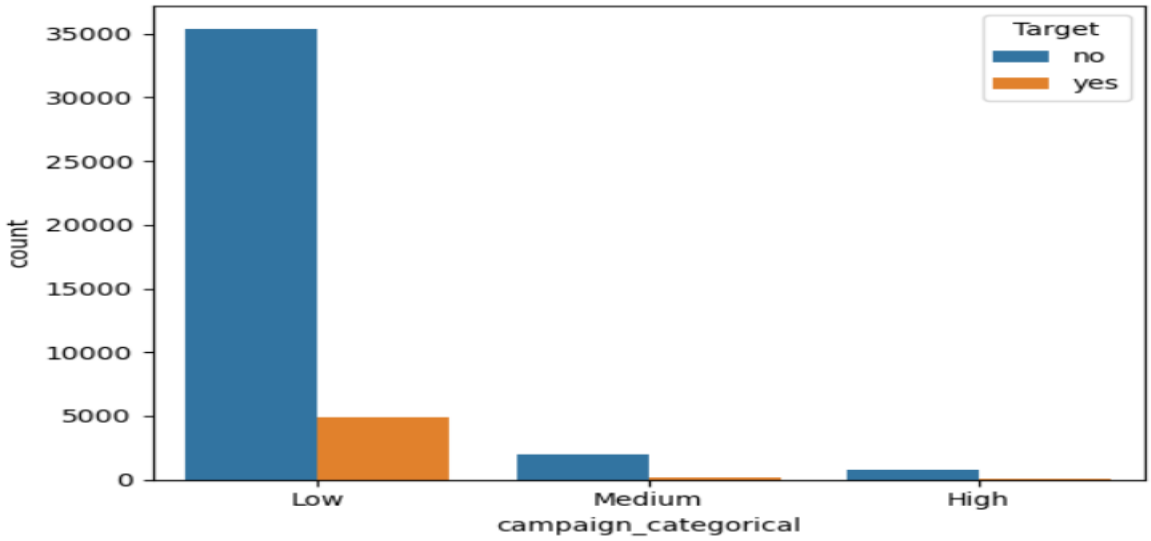
Here it is proved that the most no. of clients who haven't subscribed for the term deposit are the ones who have been interacting with the institution for a short duration of time.

Countplot to show the influence of pdays_categorical on target.



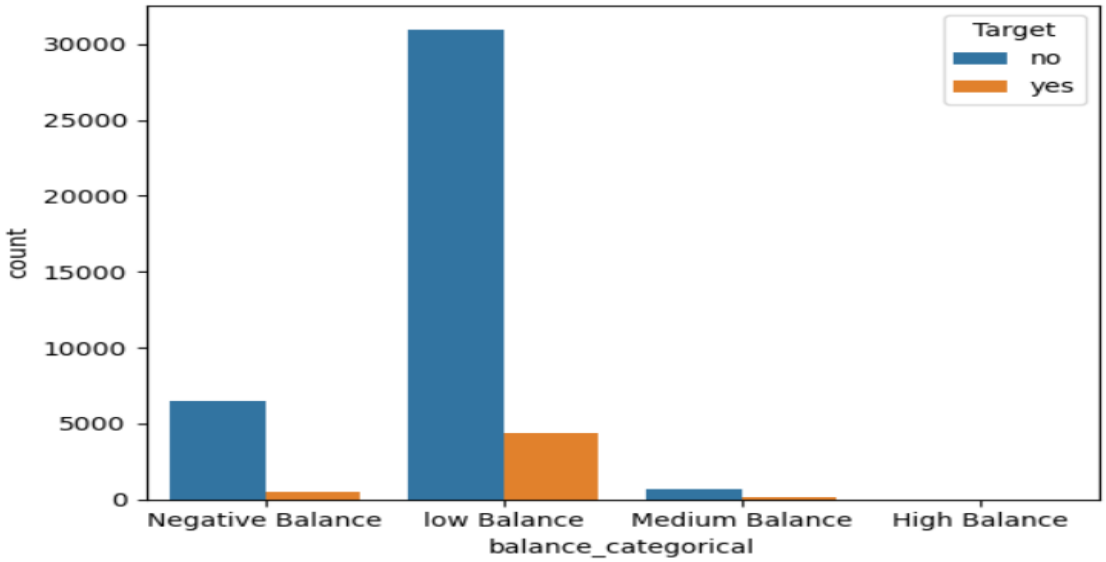
Here it is proved that most no. of clients who haven't been subscribed for the term deposit are those who haven't been contacted by the institution in previous days of the campaign.

Countplot to show the influence of campaign_categorical on target.



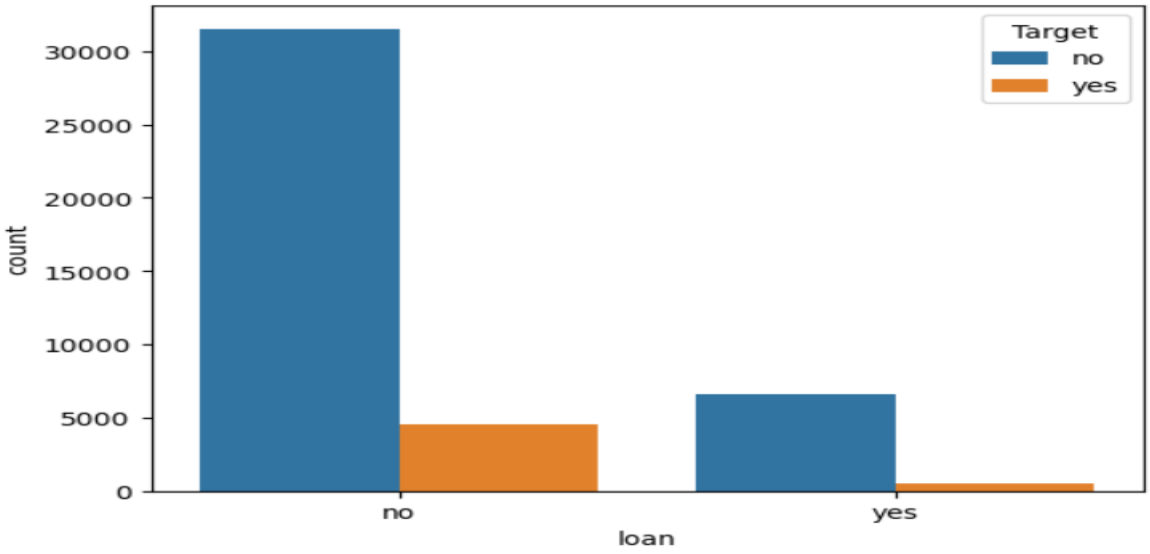
Here it is proved that most no. of clients who haven't been subscribed for the term deposit are those who have been contacted for less no. of time in this campaign.

Countplot to show the influence of balance_categorical on target.



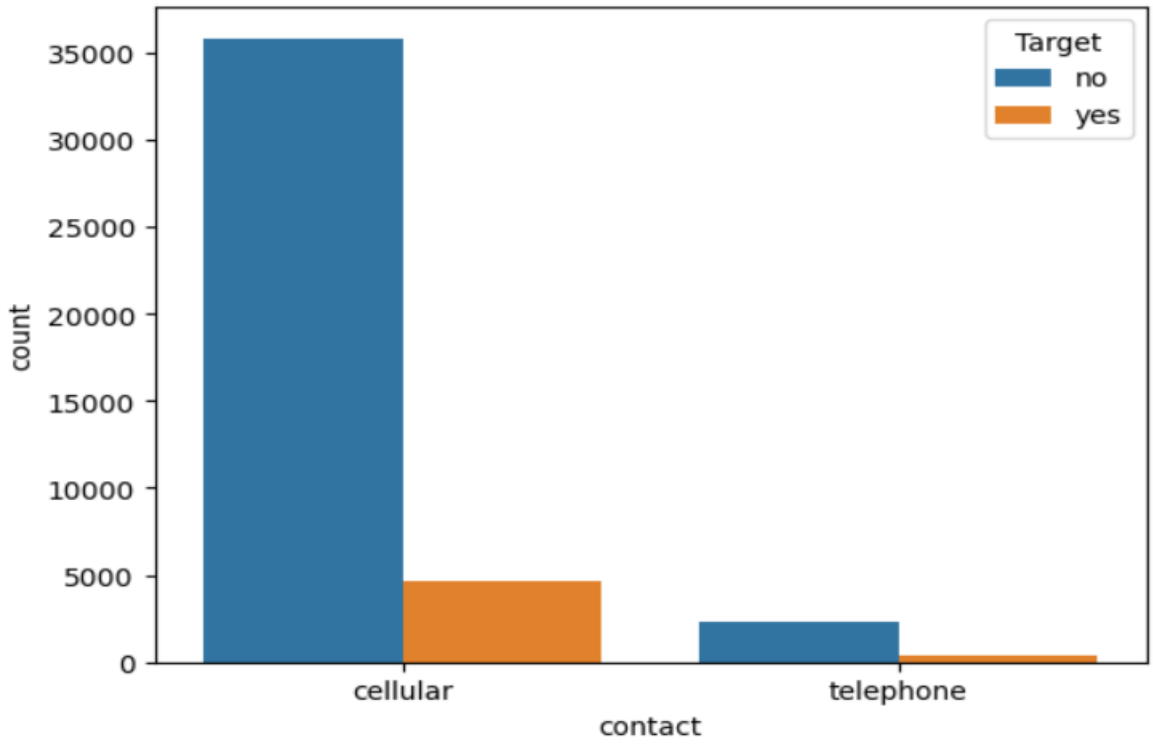
Here it is proved that most no. of clients who haven't been subscribed for the term deposit are those who have low balance and negative balance in the Portuguese banking institution.

Countplot to show the influence of loan_categorical on target.



Here it is proved that most no. of clients who haven't been subscribed for the term deposit are those who haven't taken any loan in this banking institution.

Countplot to show the influence of contact_categorical on target.



Here it is proved that most no. of clients who haven't subscribed for the term deposit have rejected the term deposit on call in cellular phones.

LABEL ENCODING:

We performed Label encoding to categorical features to numerical features.

The output looks as below:

	job	marital	education	default	housing	loan	contact	previous	poutcome	Target	duration_categorical	campaign_categorical	pdays_categorical
0	4	1	2	0	1	0	0	0	1	0	1	1	4
1	9	2	1	0	1	0	0	0	1	0	1	1	4
2	2	1	1	0	1	1	0	0	1	0	1	1	4
5	4	1	2	0	1	0	0	0	1	0	1	1	4
6	4	2	2	0	1	1	0	0	1	0	1	1	4

age_categorical	balance_categorical
0	3
0	3
2	3
2	3
2	3

TRAIN-TEST SPLIT:

Then we split the data into train and test in the ratio of 70: 30 ratios.

DATA STANDARDIZATION :

Standardization of dataset is a common requirement for many machine learning techniques. They might behave badly if the individual features do not look more or less like the standard normally distributed data.

The Standard Scaler method is used to standardize the input values. StandardScaler() and fit_transform() functions are used here.

Model Observation Table:

First, we did model building on imbalanced data.

Then, we did model building on under sampled data and then the oversampled data to compare the models accuracy and get the best model.

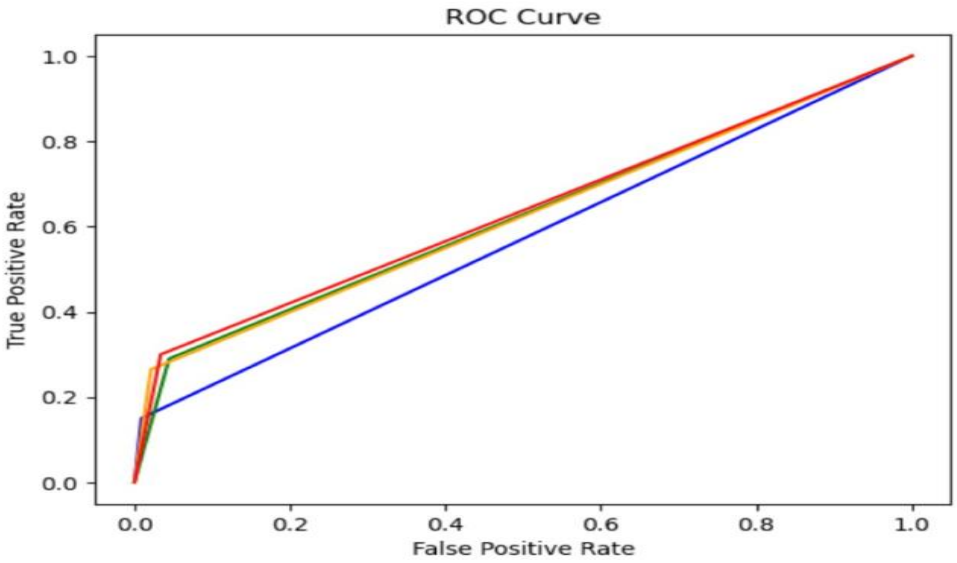
We could observe the below data:

Accuracy was good for unbalanced data but the precision and the recall for the minority class was extremely less. We were more concerned about the precision value as we didn't want to wrongly predict the possibility of a customer accepting a term deposit offer.

Then, we observed better precision values for the under sampled data. But oversampled data had the highest accuracy and good precision and recall values. This is because most of the useful information was lost in under sampling.

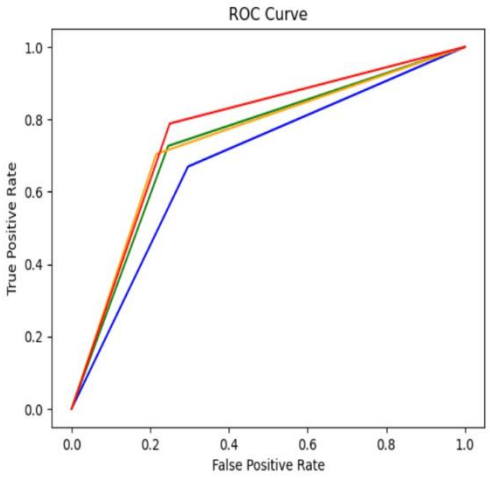
Model	Without sampling				Random undersampling				Random oversampling			
	Accuracy	precision		recall	Accuracy	precision		recall	Accuracy	precision		recall
LR	0.89	0	0.90	0.99	0.69	0	0.68	0.70	0.69	0	0.68	0.73
		1	0.69	0.15		1	0.69	0.67		1	0.71	0.65
DT	0.88	0	0.96	0.93	0.74	0	0.74	0.76	0.83	0	0.83	0.84
		1	0.29	0.36		1	0.75	0.73		1	0.84	0.83
KNN	0.90	0	0.91	0.98	0.75	0	0.72	0.82	0.79	0	0.80	0.77
		1	0.62	0.26		1	0.78	0.68		1	0.78	0.81
Random Forest	0.89	0	0.91	0.97	0.77	0	0.78	0.75	0.83	0	0.83	0.84
		1	0.54	0.30		1	0.76	0.79		1	0.84	0.83

ROC Curve for Imbalanced data:




```
Model AUC score for Logistic regression: 0.5708245142610372
Model AUC score for Decision Tree: 0.6228194684296925
Model AUC score for KNN: 0.6218606189979864
Model AUC score for random forest: 0.6332134496128885
```

ROC curve for undersampled data:



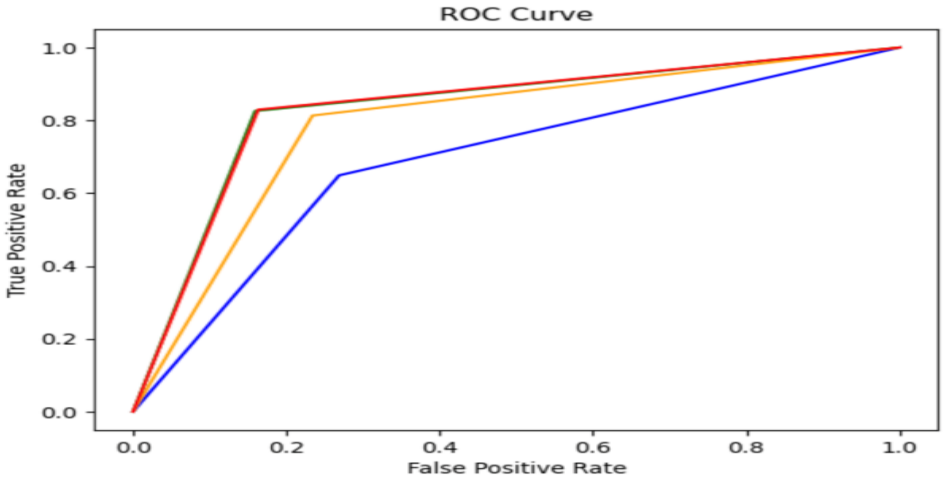
Model AUC score for Logistic regression: 0.6866852431019123

Model AUC score for Decision Tree: 0.7404533601933783

Model AUC score for KNN: 0.7444272897351095

Model AUC score for random forest: 0.7693391441973755

ROC Curve for oversampled data:



Model AUC score for Logistic regression: 0.6900541390150192

Model AUC score for Decision Tree: 0.8329986028641285

Model AUC score for KNN: 0.7895127488648271

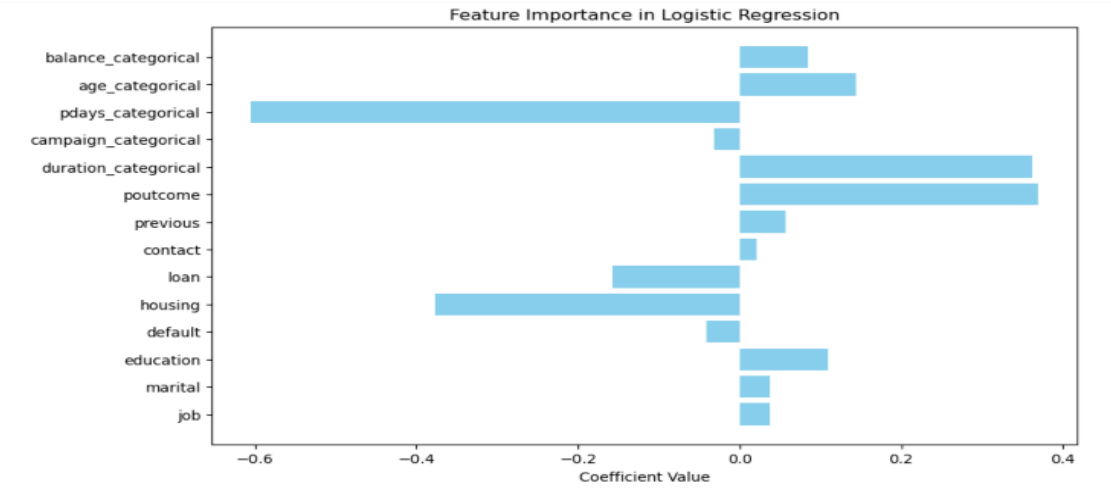
Model AUC score for random forest: 0.8330422633601118

The Random Forest model's ROC curve is slightly better than the Decision Tree's ROC curve as it is closer to the top-left corner of the graph compared to the other Classifiers.

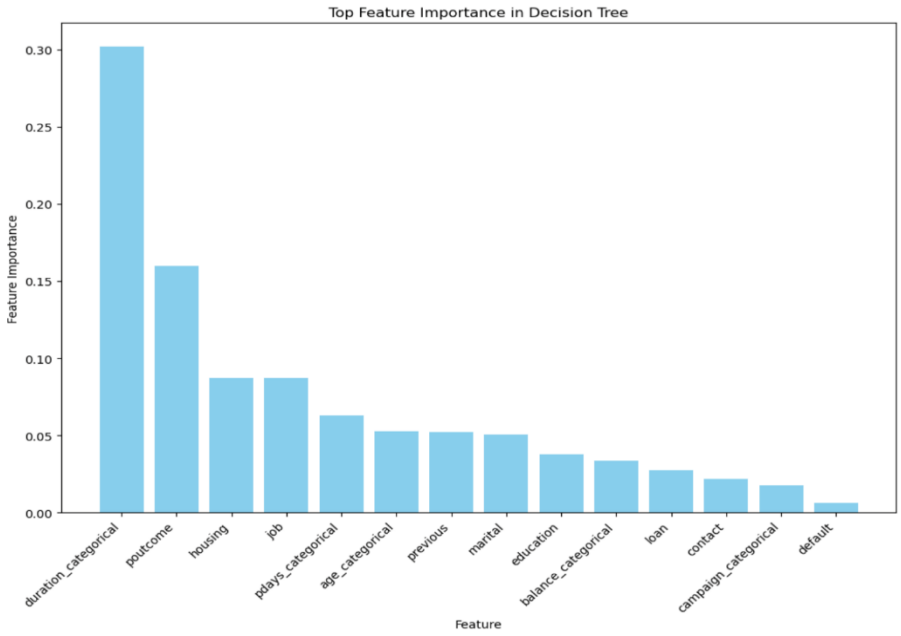
The AUC score for the Random Forest is slightly higher than that of the Decision Tree (0.8330 vs. 0.8329), indicating that both Random Forest and decision Tree perform slightly better in terms of distinguishing between the positive and negative classes based on the dataset.

Feature importance in each of the model:

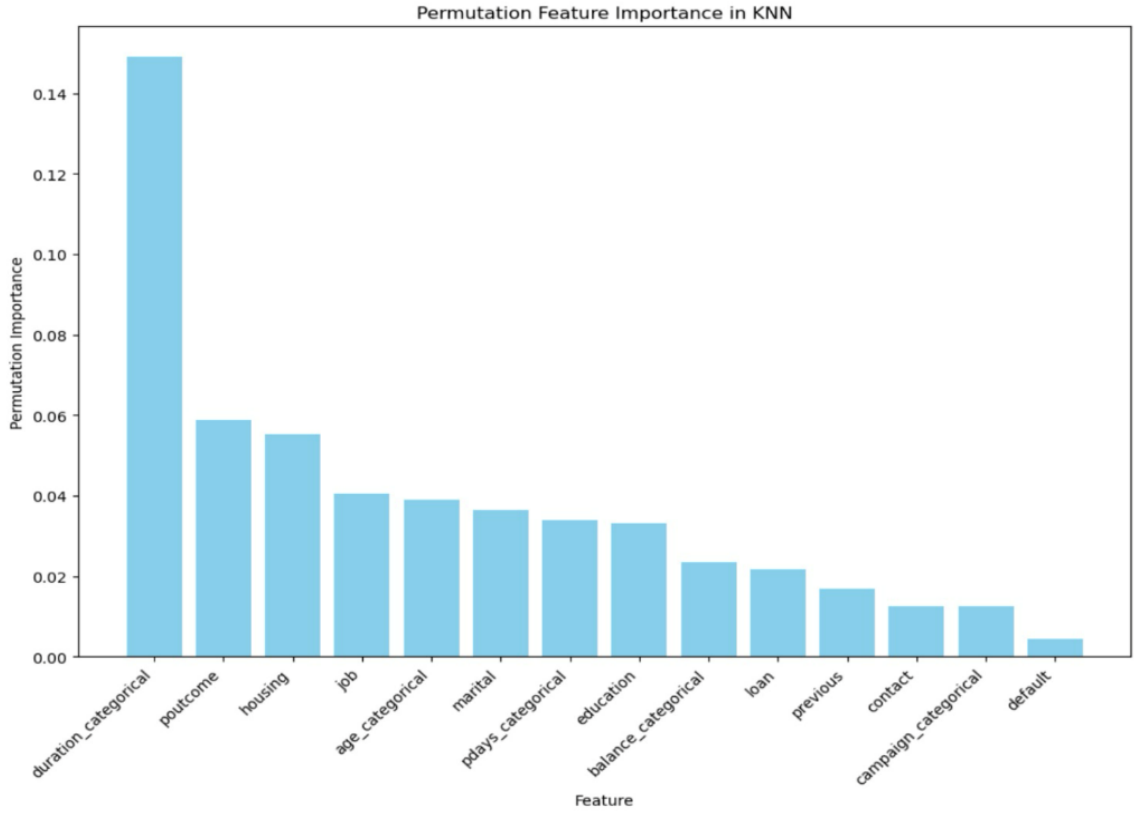
Logistic Regression:



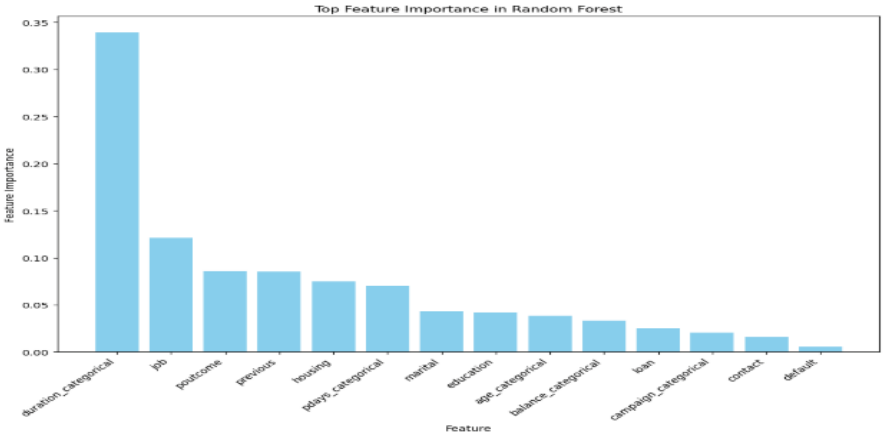
Decision Tree:



KNN:



Random Forest:



We could observe that Duration was the most influential feature for Prediction in all the 4 models followed by poutcome.

Conclusion:

We were able to observe that Both Random Forest and the Decision Tree Model performed better than the other 2 models and had the highest accuracy of 83% amongst other Models. Also, we could observe that Duration was the most influential feature for Prediction in all the 4 models followed by poutcome.So, more the duration of the call, success or others in the poutcome, more the possibility of getting a term deposit offer from the customer.

References:

https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit

https://www.kaggle.com/datasets/krantiswalke/bankfullcsv