Assignment-06-Logistic Regression

In [5]:

```python
# Import the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
```

In [6]:

```python
# Importing the dataset
bank=pd.read_csv('bank-full.csv', sep=';')
bank
```

Out[6]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3	-1	0	unknown	yes
45207	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2	-1	0	unknown	yes
45208	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5	184	3	success	yes
45209	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4	-1	0	unknown	no
45210	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2	188	11	other	no

45211 rows × 17 columns

EDA

In [7]:

```python
bank.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  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

In [8]:

```
# One-Hot Encoding of Catergrical varaiables
data1=pd.get_dummies(bank,columns=['job','marital','education','contact','poutcome'])
data1
```

Out[8]:

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	...	education_secondary	education_tertiary	education_unknown	contact_
0	58	no	2143	yes	no	5	may	261	1	-1	...	0	1	0	
1	44	no	29	yes	no	5	may	151	1	-1	...	1	0	0	
2	33	no	2	yes	yes	5	may	76	1	-1	...	1	0	0	
3	47	no	1506	yes	no	5	may	92	1	-1	...	0	0	1	
4	33	no	1	no	no	5	may	198	1	-1	...	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
45206	51	no	825	no	no	17	nov	977	3	-1	...	0	1	0	
45207	71	no	1729	no	no	17	nov	456	2	-1	...	0	0	0	
45208	72	no	5715	no	no	17	nov	1127	5	184	...	1	0	0	
45209	57	no	668	no	no	17	nov	508	4	-1	...	1	0	0	
45210	37	no	2971	no	no	17	nov	361	2	188	...	1	0	0	

45211 rows × 38 columns

In [9]:

```
# To see all columns
pd.set_option("display.max.columns",None)
data1
```

Out[9]:

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	previous	y	job_admin.	job_blue-collar	job_entrepreneur	job_housemaid	j
0	58	no	2143	yes	no	5	may	261	1	-1	0	no	0	0	0	0	
1	44	no	29	yes	no	5	may	151	1	-1	0	no	0	0	0	0	
2	33	no	2	yes	yes	5	may	76	1	-1	0	no	0	0	1	0	
3	47	no	1506	yes	no	5	may	92	1	-1	0	no	0	1	0	0	
4	33	no	1	no	no	5	may	198	1	-1	0	no	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
45206	51	no	825	no	no	17	nov	977	3	-1	0	yes	0	0	0	0	
45207	71	no	1729	no	no	17	nov	456	2	-1	0	yes	0	0	0	0	
45208	72	no	5715	no	no	17	nov	1127	5	184	3	yes	0	0	0	0	
45209	57	no	668	no	no	17	nov	508	4	-1	0	no	0	1	0	0	
45210	37	no	2971	no	no	17	nov	361	2	188	11	no	0	0	1	0	

45211 rows × 38 columns

In [10]:

```
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 38 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   age                  45211 non-null  int64 
 1   default              45211 non-null  object
 2   balance              45211 non-null  int64 
 3   housing              45211 non-null  object
 4   loan                 45211 non-null  object
 5   day                  45211 non-null  int64 
 6   month                45211 non-null  object
 7   duration             45211 non-null  int64 
 8   campaign             45211 non-null  int64 
 9   pdays                45211 non-null  int64 
 10  previous             45211 non-null  int64 
 11  y                    45211 non-null  object
 12  job_admin.           45211 non-null  uint8 
 13  job_blue-collar      45211 non-null  uint8 
 14  job_entrepreneur     45211 non-null  uint8 
 15  job_housemaid        45211 non-null  uint8 
 16  job_management       45211 non-null  uint8 
 17  job_retired          45211 non-null  uint8 
 18  job_self-employed    45211 non-null  uint8 
 19  job_services         45211 non-null  uint8 
 20  job_student          45211 non-null  uint8 
 21  job_technician       45211 non-null  uint8 
 22  job_unemployed       45211 non-null  uint8 
 23  job_unknown          45211 non-null  uint8 
 24  marital_divorced     45211 non-null  uint8 
 25  marital_married      45211 non-null  uint8 
 26  marital_single       45211 non-null  uint8 
 27  education_primary    45211 non-null  uint8 
 28  education_secondary  45211 non-null  uint8 
 29  education_tertiary   45211 non-null  uint8 
 30  education_unknown    45211 non-null  uint8 
 31  contact_cellular     45211 non-null  uint8 
 32  contact_telephone    45211 non-null  uint8 
 33  contact_unknown      45211 non-null  uint8 
 34  poutcome_failure     45211 non-null  uint8 
 35  poutcome_other       45211 non-null  uint8 
 36  poutcome_success     45211 non-null  uint8 
 37  poutcome_unknown     45211 non-null  uint8 
dtypes: int64(7), object(5), uint8(26)
memory usage: 5.3+ MB
```

In [11]:

```
#Custom Binary Encoding of Binary o/p variables
data1['default'] = np.where(data1['default'].str.contains("yes"), 1, 0)
data1['housing'] = np.where(data1['housing'].str.contains("yes"), 1, 0)
data1['loan'] = np.where(data1['loan'].str.contains("yes"), 1, 0)
data1['y'] = np.where(data1['y'].str.contains("yes"), 1, 0)
data1
```

Out[11]:

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	previous	y	job_admin.	job_blue-collar	job_entrepreneur	job_housemaid	jo
0	58	0	2143	1	0	5	may	261	1	-1	0	0	0	0	0	0	
1	44	0	29	1	0	5	may	151	1	-1	0	0	0	0	0	0	
2	33	0	2	1	1	5	may	76	1	-1	0	0	0	0	1	0	
3	47	0	1506	1	0	5	may	92	1	-1	0	0	0	1	0	0	
4	33	0	1	0	0	5	may	198	1	-1	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
45206	51	0	825	0	0	17	nov	977	3	-1	0	1	0	0	0	0	
45207	71	0	1729	0	0	17	nov	456	2	-1	0	1	0	0	0	0	
45208	72	0	5715	0	0	17	nov	1127	5	184	3	1	0	0	0	0	
45209	57	0	668	0	0	17	nov	508	4	-1	0	0	0	1	0	0	
45210	37	0	2971	0	0	17	nov	361	2	188	11	0	0	0	1	0	

45211 rows × 38 columns

In [13]:

```
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import LabelEncoder
for column in bank.columns:
        if bank[column].dtype == np.number:
            continue
        bank[column] = LabelEncoder().fit_transform(bank[column])
bank.head()
```

Out[13]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	40	4	1	2	0	3036	1	0	2	4	8	261	0	0	0	3	0
1	26	9	2	1	0	945	1	0	2	4	8	151	0	0	0	3	0
2	15	2	1	1	0	918	1	1	2	4	8	76	0	0	0	3	0
3	29	1	1	3	0	2420	1	0	2	4	8	92	0	0	0	3	0
4	15	11	2	3	0	917	0	0	2	4	8	198	0	0	0	3	0

In [14]:

```
# Find and Replace Encoding for month catergorical variable
data1['month'].value_counts()
```

Out[14]:

```
may    13766
jul     6895
aug     6247
jun     5341
nov     3970
apr     2932
feb     2649
jan     1403
oct      738
sep      579
mar      477
dec      214
Name: month, dtype: int64
```

In [15]:

```
order={'month':{'jan':1,'feb':2,'mar':3,'apr':4,'may':5,'jun':6,'jul':7,'aug':8,'sep':9,'oct':10,'nov':11,'dec':12}}
```

In [16]:

```
data1=data1.replace(order)
```

In [17]:

```
data1
```

Out[17]:

	age	default	balance	housing	loan	day	month	duration	campaign	pdays	previous	y	job_admin.	job_blue-collar	job_entrepreneur	job_housemaid	jo
0	58	0	2143	1	0	5	5	261	1	-1	0	0	0	0	0	0	
1	44	0	29	1	0	5	5	151	1	-1	0	0	0	0	0	0	
2	33	0	2	1	1	5	5	76	1	-1	0	0	0	0	1	0	
3	47	0	1506	1	0	5	5	92	1	-1	0	0	0	1	0	0	
4	33	0	1	0	0	5	5	198	1	-1	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
45206	51	0	825	0	0	17	11	977	3	-1	0	1	0	0	0	0	
45207	71	0	1729	0	0	17	11	456	2	-1	0	1	0	0	0	0	
45208	72	0	5715	0	0	17	11	1127	5	184	3	1	0	0	0	0	
45209	57	0	668	0	0	17	11	508	4	-1	0	0	0	1	0	0	
45210	37	0	2971	0	0	17	11	361	2	188	11	0	0	0	1	0	

45211 rows × 38 columns

In [18]:
```
data1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 38 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  45211 non-null  int64
 1   default              45211 non-null  int32
 2   balance              45211 non-null  int64
 3   housing              45211 non-null  int32
 4   loan                 45211 non-null  int32
 5   day                  45211 non-null  int64
 6   month                45211 non-null  int64
 7   duration             45211 non-null  int64
 8   campaign             45211 non-null  int64
 9   pdays                45211 non-null  int64
 10  previous             45211 non-null  int64
 11  y                    45211 non-null  int32
 12  job_admin.           45211 non-null  uint8
 13  job_blue-collar      45211 non-null  uint8
 14  job_entrepreneur     45211 non-null  uint8
 15  job_housemaid        45211 non-null  uint8
 16  job_management       45211 non-null  uint8
 17  job_retired          45211 non-null  uint8
 18  job_self-employed    45211 non-null  uint8
 19  job_services         45211 non-null  uint8
 20  job_student          45211 non-null  uint8
 21  job_technician       45211 non-null  uint8
 22  job_unemployed       45211 non-null  uint8
 23  job_unknown          45211 non-null  uint8
 24  marital_divorced     45211 non-null  uint8
 25  marital_married      45211 non-null  uint8
 26  marital_single       45211 non-null  uint8
 27  education_primary    45211 non-null  uint8
 28  education_secondary  45211 non-null  uint8
 29  education_tertiary   45211 non-null  uint8
 30  education_unknown    45211 non-null  uint8
 31  contact_cellular     45211 non-null  uint8
 32  contact_telephone    45211 non-null  uint8
 33  contact_unknown      45211 non-null  uint8
 34  poutcome_failure     45211 non-null  uint8
 35  poutcome_other       45211 non-null  uint8
 36  poutcome_success     45211 non-null  uint8
 37  poutcome_unknown     45211 non-null  uint8
dtypes: int32(4), int64(8), uint8(26)
memory usage: 4.6 MB
```

Model Building

Dividing our data into input and output variables

In [19]:
```
x=pd.concat([data1.iloc[:,0:11],data1.iloc[:,12:]],axis=1)
y=data1.iloc[:,11]
```

In [20]:
```
# Logistic regression model
classifier=LogisticRegression()
classifier.fit(x,y)
```

Out[20]:
```
LogisticRegression()
```

Model Predictions

In [21]:
```
# Predict for x dataset
y_pred=classifier.predict(x)
y_pred
```

Out[21]:
```
array([0, 0, 0, ..., 1, 0, 0])
```

In [22]:

```
y_pred_df=pd.DataFrame({'actual_y':y,'y_pred_prob':y_pred})
y_pred_df
```

Out[22]:

	actual_y	y_pred_prob
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
...	...	...
45206	1	1
45207	1	0
45208	1	1
45209	0	0
45210	0	0

45211 rows × 2 columns

Testing Model Accuracy

In [23]:

```
# confusion matrix for the model Accuracy
confusion_matrix=confusion_matrix(y,y_pred)
confusion_matrix
```

Out[23]:

```
array([[39146,   776],
       [ 4046,  1243]], dtype=int64)
```

In [24]:

```
# The mode accuracy is calculated by (a+d)/(a+b+c+d)
(39156+1162)/(39156+766+4127+1162)
```

Out[24]:

```
0.8917741257658535
```

In [25]:

```
# As Accuracy = 0.8933, which is greater than 0.5; Thus [;,1] Threshold value>0.5=1 esle [:,0] Threshold value<0.5=0
classifier.predict_proba(x)[:,1]
```

Out[25]:

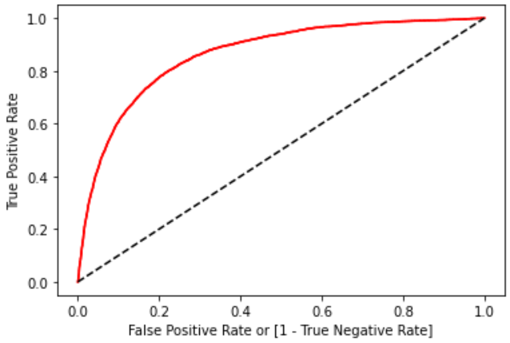
```
array([0.04476732, 0.02155565, 0.01339782, ..., 0.79033147, 0.07604012,
       0.12629756])
```

In [26]:

```
# ROC Curve plotting and finding AUC value
fpr,tpr,thresholds=roc_curve(y,classifier.predict_proba(x)[:,1])
plt.plot(fpr,tpr,color='red')
auc=roc_auc_score(y,y_pred)

plt.plot(fpr,tpr,color='red',label='logit model(area  = %0.2f)'%auc)
plt.plot([0,1],[0,1],'k--')
plt.xlabel('False Positive Rate or [1 - True Negative Rate]')
plt.ylabel('True Positive Rate')
plt.show()

print('auc accuracy:',auc)
```



```
auc accuracy: 0.6077890835891568
```

In [ ]: