In [1]:

```
import pandas as pd
import numpy as np
from scipy import stats
from matplotlib import pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score

import warnings
warnings.filterwarnings('ignore')
```

In [3]:

```
dd = pd.read_csv('bank-full.csv',sep=';')
dd
```

Out[3]:

| | age | job | marital | education | default | balance | housing | loan | contact | day |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 58 | management | married | tertiary | no | 2143 | yes | no | unknown | 5 |
| **1** | 44 | technician | single | secondary | no | 29 | yes | no | unknown | 5 |
| **2** | 33 | entrepreneur | married | secondary | no | 2 | yes | yes | unknown | 5 |
| **3** | 47 | blue-collar | married | unknown | no | 1506 | yes | no | unknown | 5 |
| **4** | 33 | unknown | single | unknown | no | 1 | no | no | unknown | 5 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **45206** | 51 | technician | married | tertiary | no | 825 | no | no | cellular | 17 |
| **45207** | 71 | retired | divorced | primary | no | 1729 | no | no | cellular | 17 |
| **45208** | 72 | retired | married | secondary | no | 5715 | no | no | cellular | 17 |
| **45209** | 57 | blue-collar | married | secondary | no | 668 | no | no | telephone | 17 |
| **45210** | 37 | entrepreneur | married | secondary | no | 2971 | no | no | cellular | 17 |

45211 rows × 17 columns

In [4]:

```
# EDA
```

In [5]:

```
dd.shape
```

Out[5]:

```
(45211, 17)
```

In [6]:

```
dd2 = pd.get_dummies(dd,columns=['job','marital','education','contact','poutcome','month'])
dd2
```

Out[6]:

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

45211 rows × 49 columns

In [7]:

```
dd2.head(30)
```

Out[7]:

| | age | default | balance | housing | loan | day | duration | campaign | pdays | previous | ... | montl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 58 | no | 2143 | yes | no | 5 | 261 | 1 | -1 | 0 | ... | |
| **1** | 44 | no | 29 | yes | no | 5 | 151 | 1 | -1 | 0 | ... | |
| **2** | 33 | no | 2 | yes | yes | 5 | 76 | 1 | -1 | 0 | ... | |
| **3** | 47 | no | 1506 | yes | no | 5 | 92 | 1 | -1 | 0 | ... | |
| **4** | 33 | no | 1 | no | no | 5 | 198 | 1 | -1 | 0 | ... | |
| **5** | 35 | no | 231 | yes | no | 5 | 139 | 1 | -1 | 0 | ... | |
| **6** | 28 | no | 447 | yes | yes | 5 | 217 | 1 | -1 | 0 | ... | |
| **7** | 42 | yes | 2 | yes | no | 5 | 380 | 1 | -1 | 0 | ... | |
| **8** | 58 | no | 121 | yes | no | 5 | 50 | 1 | -1 | 0 | ... | |
| **9** | 43 | no | 593 | yes | no | 5 | 55 | 1 | -1 | 0 | ... | |
| **10** | 41 | no | 270 | yes | no | 5 | 222 | 1 | -1 | 0 | ... | |
| **11** | 29 | no | 390 | yes | no | 5 | 137 | 1 | -1 | 0 | ... | |
| **12** | 53 | no | 6 | yes | no | 5 | 517 | 1 | -1 | 0 | ... | |
| **13** | 58 | no | 71 | yes | no | 5 | 71 | 1 | -1 | 0 | ... | |
| **14** | 57 | no | 162 | yes | no | 5 | 174 | 1 | -1 | 0 | ... | |
| **15** | 51 | no | 229 | yes | no | 5 | 353 | 1 | -1 | 0 | ... | |
| **16** | 45 | no | 13 | yes | no | 5 | 98 | 1 | -1 | 0 | ... | |
| **17** | 57 | no | 52 | yes | no | 5 | 38 | 1 | -1 | 0 | ... | |
| **18** | 60 | no | 60 | yes | no | 5 | 219 | 1 | -1 | 0 | ... | |
| **19** | 33 | no | 0 | yes | no | 5 | 54 | 1 | -1 | 0 | ... | |
| **20** | 28 | no | 723 | yes | yes | 5 | 262 | 1 | -1 | 0 | ... | |
| **21** | 56 | no | 779 | yes | no | 5 | 164 | 1 | -1 | 0 | ... | |
| **22** | 32 | no | 23 | yes | yes | 5 | 160 | 1 | -1 | 0 | ... | |
| **23** | 25 | no | 50 | yes | no | 5 | 342 | 1 | -1 | 0 | ... | |
| **24** | 40 | no | 0 | yes | yes | 5 | 181 | 1 | -1 | 0 | ... | |
| **25** | 44 | no | -372 | yes | no | 5 | 172 | 1 | -1 | 0 | ... | |
| **26** | 39 | no | 255 | yes | no | 5 | 296 | 1 | -1 | 0 | ... | |
| **27** | 52 | no | 113 | yes | yes | 5 | 127 | 1 | -1 | 0 | ... | |
| **28** | 46 | no | -246 | yes | no | 5 | 255 | 2 | -1 | 0 | ... | |
| **29** | 36 | no | 265 | yes | yes | 5 | 348 | 1 | -1 | 0 | ... | |

30 rows × 49 columns

In [8]:

```
# To see all columns
```

In [9]:

```
pd.set_option('display.max.columns',None)
dd2
```

Out[9]:

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

45211 rows × 49 columns

In [10]:

```
dd2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 49 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   duration             45211 non-null  int64
 7   campaign             45211 non-null  int64
 8   pdays                45211 non-null  int64
 9   previous             45211 non-null  int64
 10  y                    45211 non-null  object
 11  job_admin.           45211 non-null  uint8
 12  job_blue-collar      45211 non-null  uint8
 13  job_entrepreneur     45211 non-null  uint8
 14  job_housemaid        45211 non-null  uint8
 15  job_management       45211 non-null  uint8
 16  job_retired          45211 non-null  uint8
 17  job_self-employed    45211 non-null  uint8
 18  job_services         45211 non-null  uint8
 19  job_student          45211 non-null  uint8
 20  job_technician       45211 non-null  uint8
 21  job_unemployed       45211 non-null  uint8
 22  job_unknown          45211 non-null  uint8
 23  marital_divorced     45211 non-null  uint8
 24  marital_married      45211 non-null  uint8
 25  marital_single       45211 non-null  uint8
 26  education_primary    45211 non-null  uint8
 27  education_secondary  45211 non-null  uint8
 28  education_tertiary   45211 non-null  uint8
 29  education_unknown    45211 non-null  uint8
 30  contact_cellular     45211 non-null  uint8
 31  contact_telephone    45211 non-null  uint8
 32  contact_unknown      45211 non-null  uint8
 33  poutcome_failure     45211 non-null  uint8
 34  poutcome_other       45211 non-null  uint8
 35  poutcome_success     45211 non-null  uint8
 36  poutcome_unknown     45211 non-null  uint8
 37  month_apr            45211 non-null  uint8
 38  month_aug            45211 non-null  uint8
 39  month_dec            45211 non-null  uint8
 40  month_feb            45211 non-null  uint8
 41  month_jan            45211 non-null  uint8
 42  month_jul            45211 non-null  uint8
 43  month_jun            45211 non-null  uint8
 44  month_mar            45211 non-null  uint8
 45  month_may            45211 non-null  uint8
 46  month_nov            45211 non-null  uint8
 47  month_oct            45211 non-null  uint8
 48  month_sep            45211 non-null  uint8
dtypes: int64(7), object(4), uint8(38)
memory usage: 5.4+ MB
```

In [11]:

```
dd2.isna().sum()
```

Out[11]:

```
age                  0
default              0
balance              0
housing              0
loan                 0
day                  0
duration             0
campaign             0
pdays                0
previous             0
y                    0
job_admin.           0
job_blue-collar      0
job_entrepreneur     0
job_housemaid        0
job_management       0
job_retired          0
job_self-employed    0
job_services         0
job_student          0
job_technician       0
job_unemployed       0
job_unknown          0
marital_divorced     0
marital_married      0
marital_single       0
education_primary    0
education_secondary  0
education_tertiary   0
education_unknown    0
contact_cellular     0
contact_telephone    0
contact_unknown      0
poutcome_failure     0
poutcome_other       0
poutcome_success     0
poutcome_unknown     0
month_apr            0
month_aug            0
month_dec            0
month_feb            0
month_jan            0
month_jul            0
month_jun            0
month_mar            0
month_may            0
month_nov            0
month_oct            0
month_sep            0
dtype: int64
```

In [12]:

```
dd2.describe()
```

Out[12]:

| | age | balance | day | duration | campaign | pdays | 45 |
|---|---|---|---|---|---|---|---|
| **count** | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | |
| **mean** | 40.936210 | 1362.272058 | 15.806419 | 258.163080 | 2.763841 | 40.197828 | |
| **std** | 10.618762 | 3044.765829 | 8.322476 | 257.527812 | 3.098021 | 100.128746 | |
| **min** | 18.000000 | -8019.000000 | 1.000000 | 0.000000 | 1.000000 | -1.000000 | |
| **25%** | 33.000000 | 72.000000 | 8.000000 | 103.000000 | 1.000000 | -1.000000 | |
| **50%** | 39.000000 | 448.000000 | 16.000000 | 180.000000 | 2.000000 | -1.000000 | |
| **75%** | 48.000000 | 1428.000000 | 21.000000 | 319.000000 | 3.000000 | -1.000000 | |
| **max** | 95.000000 | 102127.000000 | 31.000000 | 4918.000000 | 63.000000 | 871.000000 | |

In [13]:

```
# Custom Binary Encoding of Binary o/p variables
```

In [17]:

```
dd2['default'] = np.where(dd2['default'].str.contains("yes"), 1, 0)
dd2['housing'] = np.where(dd2['housing'].str.contains("yes"), 1, 0)
dd2['loan'] = np.where(dd2['loan'].str.contains("yes"), 1, 0)
dd2['y'] = np.where(dd2['y'].str.contains("yes"), 1, 0)
dd2
```

Out[17]:

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

45211 rows × 49 columns

In [18]:

```
dd2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 49 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   duration             45211 non-null  int64
 7   campaign             45211 non-null  int64
 8   pdays                45211 non-null  int64
 9   previous             45211 non-null  int64
 10  y                    45211 non-null  int32
 11  job_admin.           45211 non-null  uint8
 12  job_blue-collar      45211 non-null  uint8
 13  job_entrepreneur     45211 non-null  uint8
 14  job_housemaid        45211 non-null  uint8
 15  job_management       45211 non-null  uint8
 16  job_retired          45211 non-null  uint8
 17  job_self-employed    45211 non-null  uint8
 18  job_services         45211 non-null  uint8
 19  job_student          45211 non-null  uint8
 20  job_technician       45211 non-null  uint8
 21  job_unemployed       45211 non-null  uint8
 22  job_unknown          45211 non-null  uint8
 23  marital_divorced     45211 non-null  uint8
 24  marital_married      45211 non-null  uint8
 25  marital_single       45211 non-null  uint8
 26  education_primary    45211 non-null  uint8
 27  education_secondary  45211 non-null  uint8
 28  education_tertiary   45211 non-null  uint8
 29  education_unknown    45211 non-null  uint8
 30  contact_cellular     45211 non-null  uint8
 31  contact_telephone    45211 non-null  uint8
 32  contact_unknown      45211 non-null  uint8
 33  poutcome_failure     45211 non-null  uint8
 34  poutcome_other       45211 non-null  uint8
 35  poutcome_success     45211 non-null  uint8
 36  poutcome_unknown     45211 non-null  uint8
 37  month_apr            45211 non-null  uint8
 38  month_aug            45211 non-null  uint8
 39  month_dec            45211 non-null  uint8
 40  month_feb            45211 non-null  uint8
 41  month_jan            45211 non-null  uint8
 42  month_jul            45211 non-null  uint8
 43  month_jun            45211 non-null  uint8
 44  month_mar            45211 non-null  uint8
 45  month_may            45211 non-null  uint8
 46  month_nov            45211 non-null  uint8
 47  month_oct            45211 non-null  uint8
 48  month_sep            45211 non-null  uint8
dtypes: int32(4), int64(7), uint8(38)
memory usage: 4.7 MB
```

In [19]:

```
# Model Building
```

In [22]:

```
x = pd.concat([dd2.iloc[:,0:10],dd2.iloc[:,11:]],axis=1)
y = dd2.iloc[:,10]
```

In [23]:

```
# Logistic Regression model
classifier = LogisticRegression ()
classifier.fit(x,y)
```

Out[23]:

```
LogisticRegression()
```

In [24]:

```
# Model Prediction
```

In [25]:

```
y_prediction = classifier.predict(x)
y_prediction
```

Out[25]:

```
array([0, 0, 0, ..., 1, 0, 0])
```

In [26]:

```
y_prediction_df = pd.DataFrame({'actual_y':y,'y_preb_prob':y_prediction})
y_prediction_df
```

Out[26]:

| | actual_y | y_preb_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

In [27]:

```
# Testing Model Accuracy
```

In [28]:

```
confusion_matrix = confusion_matrix(y,y_prediction)
confusion_matrix
```

Out[28]:

```
array([[39155,   767],
       [ 4127,  1162]], dtype=int64)
```

In [29]:

```
# THE MODEL ACCURACY IS CALCULATED BY (a+d)/(a+b+c+d)
(39153+1162)/(39153+769+4127+1162)
```

Out[29]:

```
0.8917077702329079
```

In [30]:

```
print(classification_report(y,y_prediction))
```

```
              precision    recall  f1-score   support

           0       0.90      0.98      0.94     39922
           1       0.60      0.22      0.32      5289

    accuracy                           0.89     45211
   macro avg       0.75      0.60      0.63     45211
weighted avg       0.87      0.89      0.87     45211
```

In [32]:

```
fpr,tpr,thresholds = roc_curve(y,classifier.predict_proba (x)[:,1])

auc = roc_auc_score(y,y_prediction)
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()
```