In [57]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
import string
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import spacy
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder
from sklearn.impute import KNNImputer
from sklearn.impute import SimpleImputer
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from colorama import *
warnings.filterwarnings('ignore')
data=pd.read_csv('C:/Users/Vijayvardhan reddy/Desktop/Data_Science_Internship_data_set.csv')
data=data[(data['status'] == 'LOST') | (data['status'] == 'WON')]
data=data[["status","lost_reason","budget","lease","movein","room_type"]]
print(Fore.CYAN + Style.BRIGHT + "                                        ORIGINAL DATA GIVEN
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.GREEN + data)
print(Fore.YELLOW + "_____________________________________________________________________________________________")
a=data["status"] == 'WON'
data1=data[a]
print(Fore.CYAN + Style.BRIGHT + "                  DATA THAT HAS WON AS STATUS ATTRIBUTE                    ")
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.GREEN + data1)
print(Fore.YELLOW + "_____________________________________________________________________________________________")
sns.heatmap(data1.isnull())
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT +"              LABELLING OF THE MISSING VALUE in WON STATUS               ")
print(Fore.YELLOW + "_____________________________________________________________________________________________")
plt.title('Missing Values of WON Label Data')
plt.show()
le=LabelEncoder()
data1.loc[:, 'room_type'] = le.fit_transform(data1['room_type'])
knn=KNNImputer(n_neighbors=5)
num=[col for col in data1.columns if data1[col].dtypes !='O']
knn.fit(data1[num])
mode_val = data1['lease'].mode()[0]
data1['lease'].fillna(mode_val, inplace=True)
data1.loc[:, 'lease']=le.fit_transform(data1['lease'])
mode_budget = data1['budget'].mode()[0]
data1['budget'].fillna(mode_budget, inplace=True)
data1.loc[:, 'budget']=le.fit_transform(data1['budget'])
imputer = SimpleImputer(strategy='most_frequent')
imputed_data = imputer.fit_transform(data1[['movein']])
data1['movein'] = np.squeeze(imputed_data)
b=data["status"] == 'LOST'
data2=data[b]
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT + "                    DATA THAT HAS LOSS AS STATUS ATTRIBUTE                  ")
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.GREEN + data2)
print(Fore.YELLOW + "_____________________________________________________________________________________________")
sns.heatmap(data2.isnull())
print(Fore.YELLOW + "_____________________________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT +"        LABELLING OF THE MISSING VALUES OF LOSS STATUS ATTRIBUTE          ")
print(Fore.YELLOW + "_____________________________________________________________________________________________")
plt.title('Missing Values of LOSS Label Data')
plt.show()
imputed_data2 = imputer.fit_transform(data2[['budget']])
data2['budget'] = np.squeeze(imputed_data2)
imputed_lease = imputer.fit_transform(data2[['lease']])
data2['lease'] = np.squeeze(imputed_lease)
imputed_movein = imputer.fit_transform(data2[['movein']])
data2['movein'] = np.squeeze(imputed_movein)
data2.loc[:, 'room_type'] = le.fit_transform(data2['room_type'])
num=[col for col in data2.columns if data2[col].dtypes !='O']
knn.fit(data2[num])
df=pd.concat([data1,data2])
df['lost_reason'].fillna(0, inplace=True)
df['room_type'].fillna(0, inplace=True)
ax=sns.countplot(x='status',data=df)
for p in ax.patches:
    ax.annotate(format(p.get_height()), (p.get_x() + p.get_width() / 2., p.get_height()),
```

```
                ha = 'center', va = 'center', xytext = (0, 10), textcoords = 'offset points')
plt.title('Lead Status WON or LOST')
print(Fore.YELLOW + "_____________________________________________________________________________________")
plt.show()
print(Fore.YELLOW + "_____________________________________________________________________________________")
minority_class = df['status'].value_counts().idxmin()
majority_class = df[df['status'] != minority_class]
undersampled_majority = majority_class.sample(n=len(df[df['status'] == minority_class]), random_state=42)
undersampled_data = pd.concat([undersampled_majority, df[df['status'] == minority_class]])
undersampled_data = undersampled_data.sample(frac=1, random_state=42)
book = undersampled_data[['lost_reason']]
book['lost_reason'] = book['lost_reason'].astype(str).str.strip()
book = book[book['lost_reason'] != '']
book = [lost_reason.strip() for lost_reason in book.lost_reason]
book = [lost_reason for lost_reason in book if lost_reason]
text = ' '.join(book)
no_punc_text = text.translate(str.maketrans('', '', string.punctuation))
text_tokens = word_tokenize(no_punc_text)
my_stop_words = stopwords.words('english')
my_stop_words.append('the')
my_stop_words
no_stop_tokens = [word for word in text_tokens if not word in my_stop_words]
lower_words = [x.lower() for x in no_stop_tokens]
ps = PorterStemmer()
stemmed_tokens = [ps.stem(word) for word in lower_words]
nlp = spacy.load("en_core_web_lg")
doc = nlp(' '.join(no_stop_tokens))
lemmas = [token.lemma_ for token in doc]
vectorizer_n_gram_max_features = TfidfVectorizer(norm="l2",analyzer='word', ngram_range=(1,3), max_features = 1
tf_idf_matrix_n_gram_max_features =vectorizer_n_gram_max_features.fit_transform(book)
tfidf=tf_idf_matrix_n_gram_max_features.toarray()
text_data=pd.DataFrame(tfidf, columns=['availability', 'budget', 'interested', 'low', 'low availability', 'low
undersampled_data['status'] = undersampled_data['status'].replace({'WON': 1, 'LOST':0})
undersampled_data=undersampled_data.drop(['lost_reason'], axis=1)
undersampled_data['budget'] = undersampled_data['budget'].astype(str)
undersampled_data.loc[:, 'budget']=le.fit_transform(undersampled_data['budget'])
undersampled_data['lease']=undersampled_data['lease'].astype(str)
undersampled_data.loc[:, 'lease']=le.fit_transform(undersampled_data['lease'])
text_data = text_data.set_index(undersampled_data.index)
combined_df = pd.concat([undersampled_data, text_data], axis=1)
combined_df['movein'] = pd.to_datetime(combined_df['movein'])
combined_df['month'] = combined_df['movein'].dt.month
combined_df['year'] = combined_df['movein'].dt.year
combined_df['day'] = combined_df['movein'].dt.day
combined_df['weekday'] = combined_df['movein'].dt.day_name()
combined_df=combined_df.drop('movein', axis=1)
weekday_series = combined_df['weekday']
combined_df['weekday_label'] = LabelEncoder().fit_transform(weekday_series)
combined_df=combined_df.drop('weekday', axis=1)
sns.heatmap(combined_df.corr())

plt.title(Fore.CYAN + Style.BRIGHT  + '                  Correlation Heatmap of Lead data                    ')
print(Fore.YELLOW + "_____________________________________________________________________________________")
plt.show()
print(Fore.YELLOW + "_____________________________________________________________________________________")
combined_df.corr()
x=combined_df.drop('status', axis=1)
y=combined_df['status']
x_train, x_test, y_train, y_test = train_test_split(x,y , test_size=0.3)
model = LogisticRegression()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
accuracy_lr = accuracy_score(y_test, y_pred)
print(Fore.YELLOW + "_____________________________________________________________________________________")
print(Fore.RED + Style.BRIGHT + 'Logistic Regression model Accuracy is :', accuracy_lr*100,'%')
print(Fore.YELLOW + "_____________________________________________________________________________________")
proba_score=model.predict_proba(x_test)
proba_score=proba_score*100
combined_data = np.concatenate((x_test, proba_score), axis=1)
Lead_score_data=pd.DataFrame(combined_data, columns=['budget', 'lease', 'room_type', 'availability', 'budget',
       'low', 'low availability', 'low budget', 'not', 'not interested',
       'not responding', 'responding', 'month', 'year', 'day',
       'weekday_label', 'Lead_Score for 0(LOST) class','Lead_Score for 1(WON) class'])
print(Fore.YELLOW + "_____________________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT  + "             LEAD SCORES FOR THE GIVEN DATA                          ")
print(Fore.YELLOW + "_____________________________________________________________________________________")
print(Fore.GREEN +"")
print(Lead_score_data)
print(Fore.YELLOW + "_____________________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT  + "                            CONFIGURATION MATRIX
print(Fore.GREEN + "")
print(confusion_matrix(y_test,y_pred))
print(Fore.RED + "_____________________________________________________________________________")
print(Fore.CYAN + Style.BRIGHT  + "                   Classification Report for the prediction               ")
print(Fore.GREEN + "")
print(classification_report(y_test,y_pred))
print(Fore.RED + "_____________________________________________________________________________________")
```

ORIGINAL DATA GIVEN

```
       status           lost_reason                    budget  \
0      LOST    Not responding                          NaN
1      LOST        Low budget                          NaN
2      LOST    Not responding  £121 - £180 Per Week
3      LOST        Low budget                   0-0
4      LOST         Junk lead                          NaN
...          ...                    ...                      ...
46603  LOST  Low availability   £60 - £120 Per week
46604  LOST     Semester stay   £60 - £120 Per week
46605  LOST  Low availability  £241 - £300 Per week
46606  LOST  Low availability                  1108
46607  LOST  Low availability  £181 - £240 Per Week

                                            lease           movein  \
0                                             NaN              NaN
1                                             NaN              NaN
2            Full Year Course Stay 40 - 44 weeks  31-08-2022
3                                               0              NaN
4                                             NaN              NaN
...                                           ...              ...
46603  Complete Education Year Stay 50 - 52 weeks  01-09-2022
46604              Summer/Short Stay 8 - 12 weeks  29-09-2022
46605        Full Year Course Stay 40 - 44 weeks  20-09-2022
46606                                       294  30-08-2022
46607        Full Year Course Stay 40 - 44 weeks  01-09-2022

         room_type
0              NaN
1              NaN
2         Ensuite
3              NaN
4              NaN
...            ...
46603    Studio
46604    Studio
46605    Studio
46606          NaN
46607    Studio

[46317 rows x 6 columns]
```

DATA THAT HAS WON AS STATUS ATTRIBUTE

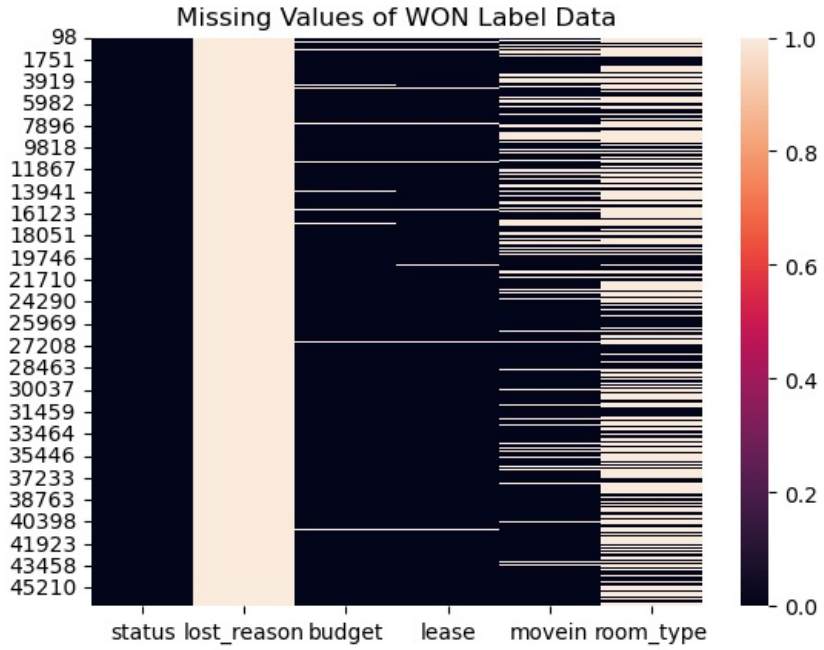
```
       status lost_reason                    budget  \
98     WON         NaN                    109
111    WON         NaN  £121 - £180 Per Week
139    WON         NaN  £181 - £240 Per Week
141    WON         NaN       > 300£ Per Week
152    WON         NaN  £121 - £180 Per Week
...        ...         ...                      ...
46559  WON         NaN                   139
46566  WON         NaN                   179
46570  WON         NaN                   180
46586  WON         NaN  £121 - £180 Per Week
46592  WON         NaN   £60 - £120 Per week

                                           lease           movein  \
98                                            51              NaN
111          Full Year Course Stay 40 - 44 weeks  09-09-2022
139          Full Year Course Stay 40 - 44 weeks  29-09-2022
141    Complete Education Year Stay 50 - 52 weeks  07-09-2022
152          Full Year Course Stay 40 - 44 weeks  31-08-2022
...                                          ...              ...
46559                                        42  01-10-2022
46566                                        51  10-09-2022
46570                                        44  16-09-2022
46586  Complete Education Year Stay 50 - 52 weeks  05-09-2022
46592        Full Year Course Stay 40 - 44 weeks  12-09-2022

             room_type
98                 NaN
111          Studio
139    Entire Place
141         Ensuite
152    Entire Place
...                ...
46559              NaN
46566              NaN
46570              NaN
46586        Studio
46592       Ensuite

[3073 rows x 6 columns]
```

LABELLING OF THE MISSING VALUE in WON STATUS

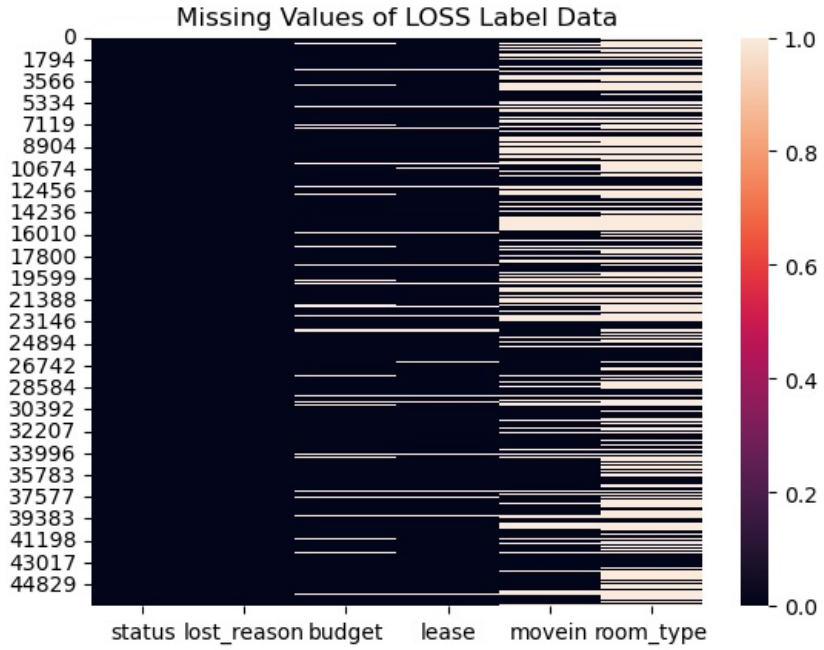


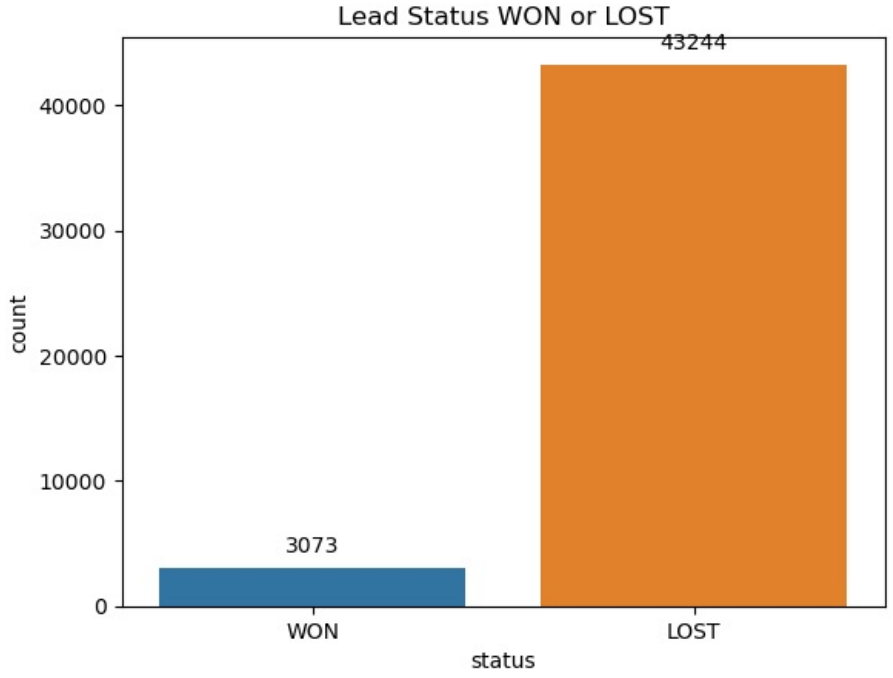
DATA THAT HAS LOSS AS STATUS ATTRIBUTE

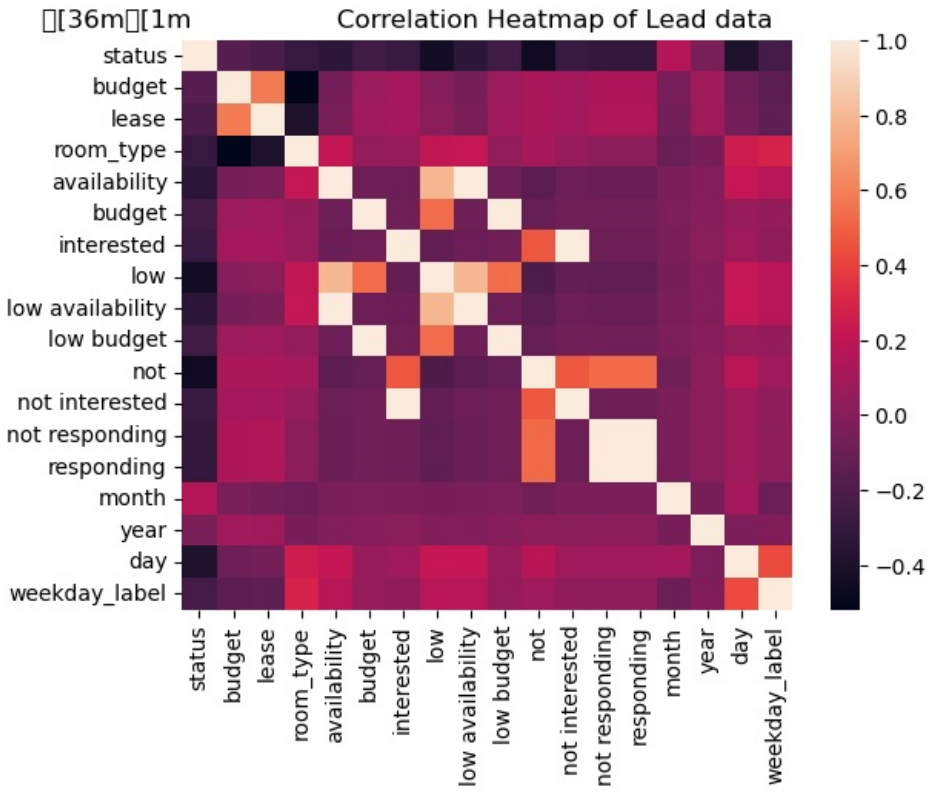
	status	lost_reason	budget	lease	movein	room_type
0	LOST	Not responding	NaN	NaN	NaN	NaN
1	LOST	Low budget	NaN	NaN	NaN	NaN
2	LOST	Not responding	£121 - £180 Per Week	Full Year Course Stay 40 - 44 weeks	31-08-2022	Ensuite
3	LOST	Low budget	0-0	0	NaN	NaN
4	LOST	Junk lead	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...
46603	LOST	Low availability	£60 - £120 Per week	Complete Education Year Stay 50 - 52 weeks	01-09-2022	Studio
46604	LOST	Semester stay	£60 - £120 Per week	Summer/Short Stay 8 - 12 weeks	29-09-2022	Studio
46605	LOST	Low availability	£241 - £300 Per week	Full Year Course Stay 40 - 44 weeks	20-09-2022	Studio
46606	LOST	Low availability	1108	294	30-08-2022	NaN
46607	LOST	Low availability	£181 - £240 Per Week	Full Year Course Stay 40 - 44 weeks	01-09-2022	Studio

[43244 rows x 6 columns]

LABELLING OF THE MISSING VALUES OF LOSS STATUS ATTRIBUTE


Missing Values of LOSS Label Data
0
1794
3566
5334
7119
8904
10674
12456
14236
16010
17800
19599
21388
23146
24894
26742
28584
30392
32207
33996
35783
37577
39383
41198
43017
44829
status
lost_reason
budget
lease
movein
room_type
1.0
0.8
0.6
0.4
0.2
0.0


Lead Status WON or LOST
43244
40000
30000
20000
10000
0
count
3073
WON
LOST
status


[36m[1m
Correlation Heatmap of Lead data
status
budget
lease
room_type
availability
budget
interested
low
low availability
low budget
not
not interested
not responding
responding
month
year
day
weekday_label
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

```
Logistic Regression model Accuracy is : 91.1062906724512 %
```

```
LEAD SCORES FOR THE GIVEN DATA
```

```
      budget  lease  room_type  availability    budget  interested       low  \
0      105.0  124.0        5.0      0.000000  0.000000         0.0  0.000000
1      508.0  115.0        5.0      0.605292  0.000000         0.0  0.516956
2      132.0   71.0        3.0      0.000000  0.000000         0.0  0.000000
3      210.0   73.0        5.0      0.000000  0.625953         0.0  0.465152
4       79.0  123.0        5.0      0.000000  0.000000         0.0  0.000000
...      ...    ...        ...           ...       ...         ...       ...
1839   477.0   42.0        3.0      0.000000  0.000000         0.0  0.000000
1840   407.0    0.0        5.0      0.000000  0.625953         0.0  0.465152
1841   394.0   71.0        3.0      0.000000  0.000000         0.0  0.000000
1842   153.0   55.0        3.0      0.000000  0.000000         0.0  0.000000
1843   362.0   92.0        0.0      0.000000  0.000000         0.0  0.000000

      low availability  low budget     not  not interested  not responding  \
0             0.000000    0.000000  0.4825             0.0        0.619352
1             0.605292    0.000000  0.0000             0.0        0.000000
2             0.000000    0.000000  0.0000             0.0        0.000000
3             0.000000    0.625953  0.0000             0.0        0.000000
4             0.000000    0.000000  0.0000             0.0        0.000000
...                ...         ...     ...             ...             ...
1839          0.000000    0.000000  0.0000             0.0        0.000000
1840          0.000000    0.625953  0.0000             0.0        0.000000
1841          0.000000    0.000000  0.0000             0.0        0.000000
1842          0.000000    0.000000  0.0000             0.0        0.000000
1843          0.000000    0.000000  0.0000             0.0        0.000000

      responding  month    year   day  weekday_label  \
0       0.619352    1.0  2022.0   9.0            3.0
1       0.000000    8.0  2022.0  31.0            6.0
2       0.000000   10.0  2022.0   9.0            3.0
3       0.000000    2.0  2023.0   1.0            6.0

4       0.000000    8.0  2022.0  31.0            6.0
...          ...    ...     ...   ...            ...
1839    0.000000    3.0  2022.0   9.0            6.0
1840    0.000000    7.0  2022.0   9.0            2.0
1841    0.000000    6.0  2022.0   8.0            6.0
1842    0.000000   10.0  2022.0   9.0            3.0
1843    0.000000    9.0  2022.0  19.0            1.0

      Lead_Score for 0(LOST) class  Lead_Score for 1(WON) class
0                        99.918192                     0.081808
1                        99.999366                     0.000634
2                         3.277446                    96.722554
3                        98.813113                     1.186887
4                        95.731731                     4.268269
...                            ...                          ...
1839                     37.102210                    62.897790
1840                     98.725623                     1.274377
1841                     27.725059                    72.274941
1842                      2.622441                    97.377559
1843                      3.559981                    96.440019

[1844 rows x 19 columns]
```

```
CONFIGURATION MATRIX
```

```
[[815 121]
 [ 43 865]]
```

```
Classification Report for the prediction
```

```
              precision    recall  f1-score   support

           0       0.95      0.87      0.91       936
           1       0.88      0.95      0.91       908

    accuracy                           0.91      1844
   macro avg       0.91      0.91      0.91      1844
weighted avg       0.91      0.91      0.91      1844
```

```
In [ ]:
```