Airline Company Satisfaction Logistic Regression

August 9, 2023

1 Airline Company Satisfaction Logistic Regression

1.1 Step 1: Imports

```
[1]: # Standard operational package imports.
     import pandas as pd
     import numpy as np

     # Important imports for preprocessing, modeling, and evaluation.
     from sklearn.preprocessing import OneHotEncoder
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     import sklearn.metrics as metrics

     # Visualization package imports.
     import matplotlib.pyplot as plt
     import seaborn as sns
```

1.1.1 Load the dataset

```
[2]: df_original = pd.read_csv("Invistico_Airline.csv")
```

1.1.2 Output the first 10 rows

```
[3]: df_original.head(10)
```

```
[3]:   satisfaction   Customer Type  Age   Type of Travel     Class  \
     0    satisfied  Loyal Customer   65  Personal Travel       Eco
     1    satisfied  Loyal Customer   47  Personal Travel  Business
     2    satisfied  Loyal Customer   15  Personal Travel       Eco
     3    satisfied  Loyal Customer   60  Personal Travel       Eco
     4    satisfied  Loyal Customer   70  Personal Travel       Eco
     5    satisfied  Loyal Customer   30  Personal Travel       Eco
     6    satisfied  Loyal Customer   66  Personal Travel       Eco
```

```
7    satisfied  Loyal Customer   10  Personal Travel       Eco
8    satisfied  Loyal Customer   56  Personal Travel  Business
9    satisfied  Loyal Customer   22  Personal Travel       Eco

   Flight Distance  Seat comfort  Departure/Arrival time convenient  \
0              265             0                                  0
1             2464             0                                  0
2             2138             0                                  0
3              623             0                                  0
4              354             0                                  0
5             1894             0                                  0
6              227             0                                  0
7             1812             0                                  0
8               73             0                                  0
9             1556             0                                  0

   Food and drink  Gate location  …  Online support  Ease of Online booking  \
0               0              2  …               2                       3
1               0              3  …               2                       3
2               0              3  …               2                       2
3               0              3  …               3                       1
4               0              3  …               4                       2
5               0              3  …               2                       2
6               0              3  …               5                       5
7               0              3  …               2                       2
8               0              3  …               5                       4
9               0              3  …               2                       2

   On-board service  Leg room service  Baggage handling  Checkin service  \
0                 3                 0                 3                5
1                 4                 4                 4                2
2                 3                 3                 4                4
3                 1                 0                 1                4
4                 2                 0                 2                4
5                 5                 4                 5                5
6                 5                 0                 5                5
7                 3                 3                 4                5
8                 4                 0                 1                5
9                 2                 4                 5                3

   Cleanliness  Online boarding  Departure Delay in Minutes  \
0            3                2                           0
1            3                2                         310
2            4                2                           0
3            1                3                           0
4            2                5                           0
5            4                2                           0
```

```
6            5                3                          17
7            4                2                           0
8            4                4                           0
9            4                2                          30

   Arrival Delay in Minutes
0                       0.0
1                     305.0
2                       0.0
3                       0.0
4                       0.0
5                       0.0
6                      15.0
7                       0.0
8                       0.0
9                      26.0

[10 rows x 22 columns]
```

1.2 Step 2: Data exploration, data cleaning, and model preparation

1.2.1 Explore the data

Check the data type of each column. Note that logistic regression models expect numeric data.

[4]:

```
df_original.dtypes
```

[4]:

```
satisfaction                         object
Customer Type                        object
Age                                   int64
Type of Travel                       object
Class                                object
Flight Distance                       int64
Seat comfort                          int64
Departure/Arrival time convenient     int64
Food and drink                        int64
Gate location                         int64
Inflight wifi service                 int64
Inflight entertainment                int64
Online support                        int64
Ease of Online booking                int64
On-board service                      int64
Leg room service                      int64
Baggage handling                      int64
Checkin service                       int64
Cleanliness                           int64
```

```
Online boarding                        int64
Departure Delay in Minutes             int64
Arrival Delay in Minutes             float64
dtype: object
```

1.2.2 Check the number of satisfied customers in the dataset

```
[5]: df_original["satisfaction"].value_counts()
```

```
[5]: satisfied       71087
     dissatisfied    58793
     Name: satisfaction, dtype: int64
```

1.2.3 Check for missing values

An assumption of logistic regression models is that there are no missing values. Check for missing values in the rows of the data.

```
[6]: df_original.isna().sum()
```

```
[6]: satisfaction                           0
     Customer Type                          0
     Age                                    0
     Type of Travel                         0
     Class                                  0
     Flight Distance                        0
     Seat comfort                           0
     Departure/Arrival time convenient      0
     Food and drink                         0
     Gate location                          0
     Inflight wifi service                  0
     Inflight entertainment                 0
     Online support                         0
     Ease of Online booking                 0
     On-board service                       0
     Leg room service                       0
     Baggage handling                       0
     Checkin service                        0
     Cleanliness                            0
     Online boarding                        0
     Departure Delay in Minutes             0
     Arrival Delay in Minutes             393
     dtype: int64
```

1.2.4 Drop the rows with missing values

```
[7]: df_subset=df_original.dropna(axis=0)
     df_subset.reset_index(drop=True)
     df_subset.isna().sum()
```

```
[7]: satisfaction                         0
     Customer Type                        0
     Age                                  0
     Type of Travel                       0
     Class                                0
     Flight Distance                      0
     Seat comfort                         0
     Departure/Arrival time convenient    0
     Food and drink                       0
     Gate location                        0
     Inflight wifi service                0
     Inflight entertainment               0
     Online support                       0
     Ease of Online booking               0
     On-board service                     0
     Leg room service                     0
     Baggage handling                     0
     Checkin service                      0
     Cleanliness                          0
     Online boarding                      0
     Departure Delay in Minutes           0
     Arrival Delay in Minutes             0
     dtype: int64
```

1.2.5 Prepare the data

For creating a plot (`sns.regplot`) of the model to visualize results, the independent variable `Inflight entertainment` cannot be "of type int" and the dependent variable `satisfaction` cannot be "of type object."

```
[8]: df_subset.astype({"Inflight entertainment":float})
     df_subset["Inflight entertainment"]
```

```
[8]: 0         4
     1         2
     2         0
     3         4
     4         3
              ..
     129875    5
```

```
129876    1
129877    2
129878    2
129879    3
Name: Inflight entertainment, Length: 129487, dtype: int64
```

1.2.6 Convert the categorical column satisfaction into numeric

[9]:
```
encoder=OneHotEncoder(drop="first")
encoded_data=encoder.fit_transform(df_subset[["satisfaction"]])
encoded_data.toarray()
df_subset[["satisfaction"]]=encoded_data.toarray()
df_subset
```

[9]:
```
        satisfaction      Customer Type  Age   Type of Travel     Class  \
0                1.0     Loyal Customer   65  Personal Travel       Eco
1                1.0     Loyal Customer   47  Personal Travel  Business
2                1.0     Loyal Customer   15  Personal Travel       Eco
3                1.0     Loyal Customer   60  Personal Travel       Eco
4                1.0     Loyal Customer   70  Personal Travel       Eco
…              …                …  …              …       …
129875           1.0  disloyal Customer   29  Personal Travel       Eco
129876           0.0  disloyal Customer   63  Personal Travel  Business
129877           0.0  disloyal Customer   69  Personal Travel       Eco
129878           0.0  disloyal Customer   66  Personal Travel       Eco
129879           0.0  disloyal Customer   38  Personal Travel       Eco

        Flight Distance  Seat comfort  Departure/Arrival time convenient  \
0                   265             0                                  0
1                  2464             0                                  0
2                  2138             0                                  0
3                   623             0                                  0
4                   354             0                                  0
…                  …           …                                …
129875             1731             5                                  5
129876             2087             2                                  3
129877             2320             3                                  0
129878             2450             3                                  2
129879             4307             3                                  4

        Food and drink  Gate location  …  Online support  \
0                    0              2  …               2
1                    0              3  …               2
2                    0              3  …               2
3                    0              3  …               3
4                    0              3  …               4
```

```
…                …            …  …             …
129875               5              3  …               2
129876               2              4  …               1
129877               3              3  …               2
129878               3              2  …               2
129879               3              3  …               3

        Ease of Online booking  On-board service  Leg room service  \
0                            3                 3                 0
1                            3                 4                 4
2                            2                 3                 3
3                            1                 1                 0
4                            2                 2                 0
…                        …               …               …
129875                       2                 3                 3
129876                       3                 2                 3
129877                       4                 4                 3
129878                       3                 3                 2
129879                       4                 5                 5

        Baggage handling  Checkin service  Cleanliness  Online boarding  \
0                      3                5            3                2
1                      4                2            3                2
2                      4                4            4                2
3                      1                4            1                3
4                      2                4            2                5
…                  …              …          …              …
129875                 4                4            4                2
129876                 3                1            2                1
129877                 4                2            3                2
129878                 3                2            1                2
129879                 5                3            3                3

        Departure Delay in Minutes  Arrival Delay in Minutes
0                                0                       0.0
1                              310                     305.0
2                                0                       0.0
3                                0                       0.0
4                                0                       0.0
…                            …                       …
129875                           0                       0.0
129876                         174                     172.0
129877                         155                     163.0
129878                         193                     205.0
129879                         185                     186.0

[129487 rows x 22 columns]
```

1.2.7 Create the training and testing data

[10]:
```
X=df_subset[["Inflight entertainment"]]
y=df_subset["satisfaction"]
X_train, X_test, y_train, y_test=train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)
```

1.3 Step 3: Model building

1.3.1 Fit a LogisticRegression model to the data

[11]:
```
clf=LogisticRegression().fit(X_train, y_train)
```

1.3.2 Obtain parameter estimates

[12]:
```
clf.coef_
```

[12]:
```
array([[0.99751462]])
```

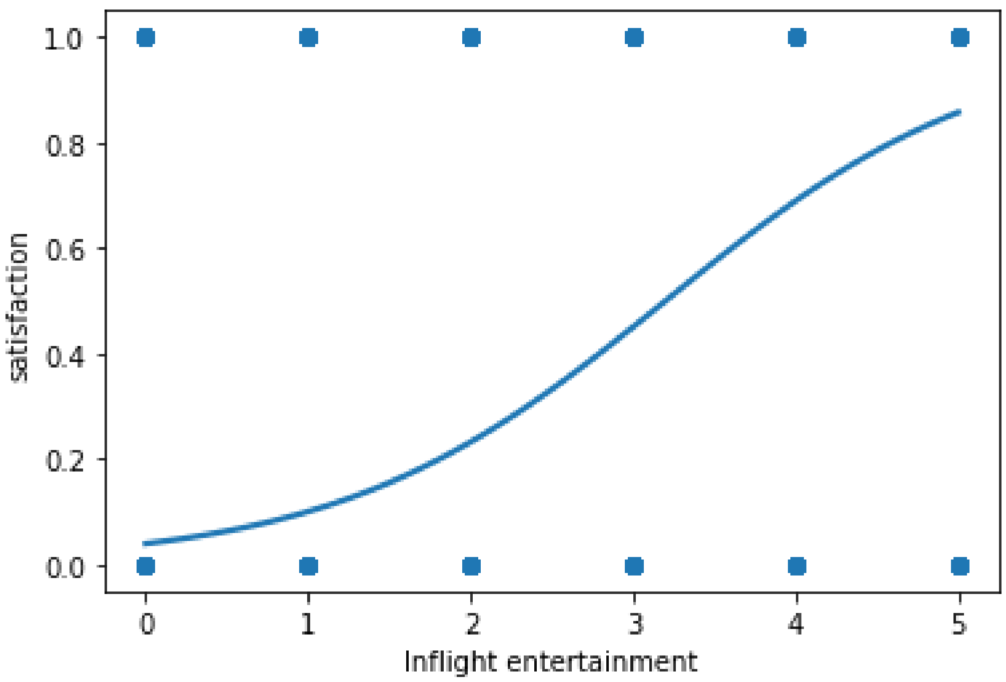
[13]:
```
clf.intercept_
```

[13]:
```
array([-3.19355406])
```

1.3.3 Create a plot of your model

[14]:
```
sns.regplot(x="Inflight entertainment", y="satisfaction", data=df_subset,␣
 ↪logistic=True, ci=None)
```

[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f68933954d0>
```



1.4 Step 4. Results and evaluation

1.4.1 Predict the outcome for the test dataset

```
[15]: y_pred=clf.predict(X_test)
      y_pred
```

```
[15]: array([1., 0., 0., …, 0., 0., 0.])
```

1.4.2 Use the `predict_proba` and `predict` functions on `X_test`

```
[16]: # Use predict_proba to output a probability.
      clf.predict_proba(X_test)
```

```
[16]: array([[0.14258068, 0.85741932],
             [0.55008402, 0.44991598],
             [0.89989329, 0.10010671],
             …,
             [0.89989329, 0.10010671],
             [0.76826225, 0.23173775],
             [0.55008402, 0.44991598]])
```

[17]:
```
# Use predict to output 0's and 1's.
clf.predict(X_test)
```

[17]:
```
array([1., 0., 0., …, 0., 0., 0.])
```

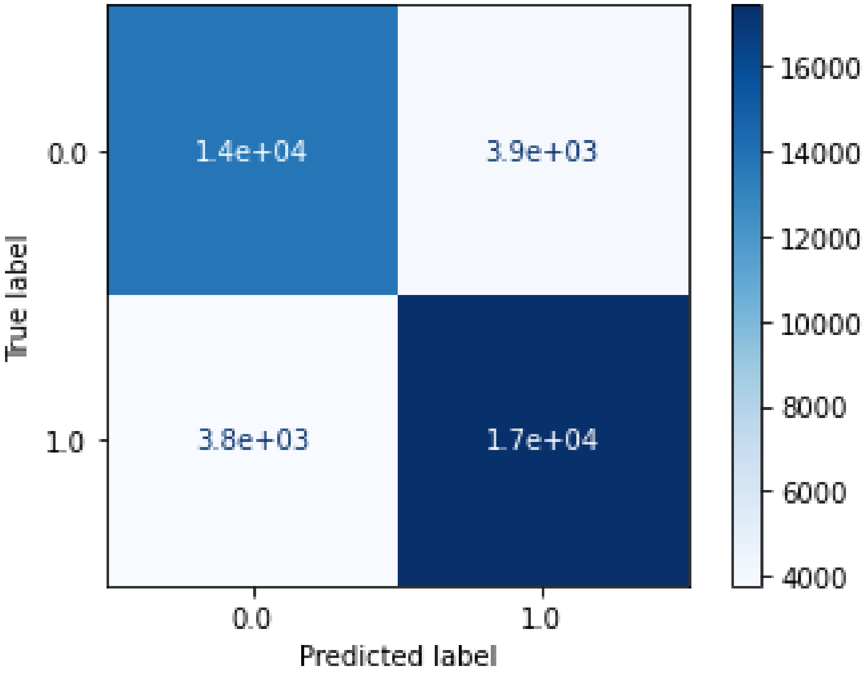
1.4.3 Analyze the results

[18]:
```
print("Accuracy score is " + "%.6f" % metrics.accuracy_score(y_test, y_pred))
print("Precision score is " "%.6f" % metrics.precision_score(y_test, y_pred))
print("F1 score is " "%.6f" % metrics.f1_score(y_test, y_pred))
```

```
Accuracy score is 0.801529
Precision score is 0.816142
F1 score is 0.818827
```

1.4.4 Produce a confusion matrix

[19]:
```
cm=metrics.confusion_matrix(y_test, y_pred, labels=clf.classes_)
disp=metrics.ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=clf.
 ↪classes_)
disp.plot(cmap=plt.cm.Blues)
plt.show()
```



1.5 Step 5. Conclusions

- Logistic regression accurately predicted satisfaction 80.2 percent of the time.
- The confusion matrix is useful, as it displays a similar amount of true positives and true negatives.
- Customers who rated in-flight entertainment highly were more likely to be satisfied. Improving in-flight entertainment should lead to better customer satisfaction.
- The model is 80.2 percent accurate. This is an improvement over the dataset's customer satisfaction rate of 54.7 percent.