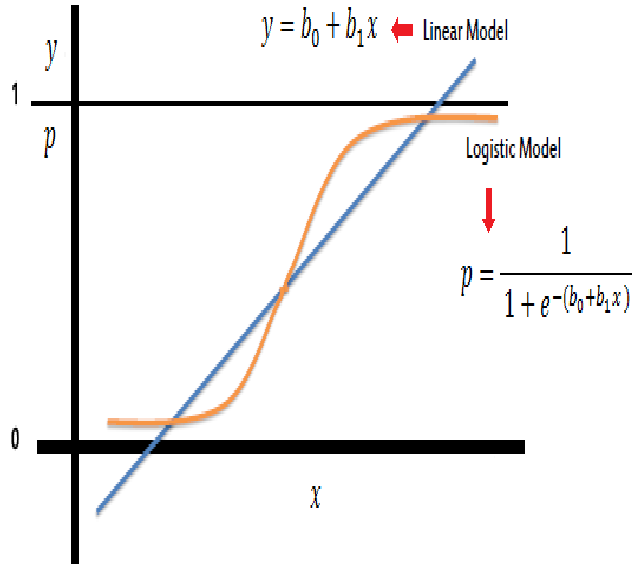
Case Study #3

Logistic Regression for Flight Status



Adriel Naranjo
ed2149

Data Mining - BAN 620
Instructor: Dr. Zinovy Radovilsky

1. Upload, explore, clean, and preprocess data.
    a. A logistic regression model may be used in this study since the outcome that we want to predict can be represented as a binary variable where FL_STATUS = 0 would be a flight that was delayed and FL_STATUS = 1 would be a flight that was on time. Logistic regression is specifically designed for binary classification tasks of Nominal categorical variables and estimates the probability of the binary outcome using the logit function. Conversely, multiple linear regression is not appropriate for this case study. It predicts continuous numerical values that can fall outside the [0,1] range, making it unsuitable for binary classification tasks like predicting flight status.
    b.

```python
try:
    flight_df = pd.read_csv('FlightDelays.csv')
except:
    print("FlightDelays.csv is not in the present working directory")
```

```python
# Remove 'DEST' and 'ORIGIN' variables from the flight_df data frame.
flight_df = flight_df.drop(['DEST', 'ORIGIN'], axis=1)
```

```python
# Convert 'CARRIER' and 'FL_STATUS' into category.
flight_df.CARRIER = flight_df.CARRIER.astype('category')
flight_df.FL_STATUS = flight_df.FL_STATUS.astype('category')
```

```python
# Convert 'CARRIER' and 'FL_STATUS' into binary variables.
flight_df = pd.get_dummies(flight_df, columns=['CARRIER', 'FL_STATUS'], prefix_sep='_', drop_first=True)
```

```python
flight_df.head()
```

	SCH_TIME	DEP_TIME	DISTANCE	FL_NUM	WEATHER	WK_DAY	MTH_DAY	CARRIER_DH	CARRIER_DL	CARRIER_MQ	CARRIER_OH	CARRIER_RU	CARI
0	1455	1455	184	5935	0	4	1	0	0	0	1	0	
1	1640	1640	213	6155	0	4	1	1	0	0	0	0	
2	1245	1245	229	7208	0	4	1	1	0	0	0	0	
3	1715	1709	229	7215	0	4	1	1	0	0	0	0	
4	1039	1035	229	7792	0	4	1	1	0	0	0	0	

    c. In logistic regression, the output variable 'FL_STATUS' needs to be converted into a binary variable to align with the model's design for binary classification tasks and does not understand what to do with variables represented as strings. Logistic regression estimates the probability that an instance belongs to one of two classes. By representing 'FL_STATUS' as binary where FL_STATUS=0 indicates a delayed flight and FL_STATUS=1 indicates an on-time flight, the model can effectively predict the probability of a flight status.
2. Develop a logistic regression model for the Flight Delays case.
    a.

logit(p)= 0.115 + 0.025(SCH_TIME) − 0.026(DEP_TIME) + 0.009(DISTANCE) + 0(FL_NUM) − 0.753(WEATHER) + 0.069(WK_DAY) − 0.022(MTH_DAY) + 0.059(CARRIER_DH) + 0.9(CARRIER_DL) − 1.004(CARRIER_MQ) + 0.37(CARRIER_OH) + 0.031(CARRIER_RU) + 0.054(CARRIER_UA) + 0.154(CARRIER_US)

```
Parameters of Multiple Predictors (14) Logistic Regression Model
Intercept: 0.115
              SCH_TIME  DEP_TIME  DISTANCE  FL_NUM  WEATHER  WK_DAY  MTH_DAY  \
Coefficient:     0.025    -0.026     0.009     0.0   -0.753   0.069   -0.022

              CARRIER_DH  CARRIER_DL  CARRIER_MQ  CARRIER_OH  CARRIER_RU  \
Coefficient:       0.059         0.9      -1.004        0.37       0.031

              CARRIER_UA  CARRIER_US
Coefficient:       0.054       0.154
```

b. Overall, the Logistic Regression model does well at predicting records with FL_STATUS = 1 (on time) but produces two false positives for FL_STATUS = 0 (delayed) by wrongly labeling the records as on time. It seems that the positive class (1) is more dominant than the negative class (0) in this validation partition. This imbalance is likely affecting the model's performance.

Classification for Validation Partition

	Actual	Classification	p(0)	p(1)
1276	1	1	0.1506	0.8494
1446	1	1	0.0730	0.9270
335	1	1	0.1008	0.8992
1458	1	1	0.1206	0.8794
2038	1	1	0.0986	0.9014
1314	1	1	0.0811	0.9189
389	1	1	0.1300	0.8700
1639	1	1	0.1623	0.8377
2004	1	1	0.0967	0.9033
403	1	1	0.2379	0.7621
979	1	1	0.0615	0.9385
65	1	1	0.0743	0.9257
2105	1	1	0.1841	0.8159
1162	1	1	0.1365	0.8635
572	1	1	0.2444	0.7556
1026	0	1	0.0620	0.9380
1044	1	1	0.4702	0.5298
1846	0	1	0.4088	0.5912
1005	1	1	0.1422	0.8578
1677	1	1	0.0814	0.9186

c. The Logistic regression model with 14 predictors does well at obtaining an accuracy of 0.8968 on the training partition and 0.8971 on the validation partition. Being that the margin between these two accuracies is minimal, we can conclude that there are no significant signs of overfitting on the training partition.

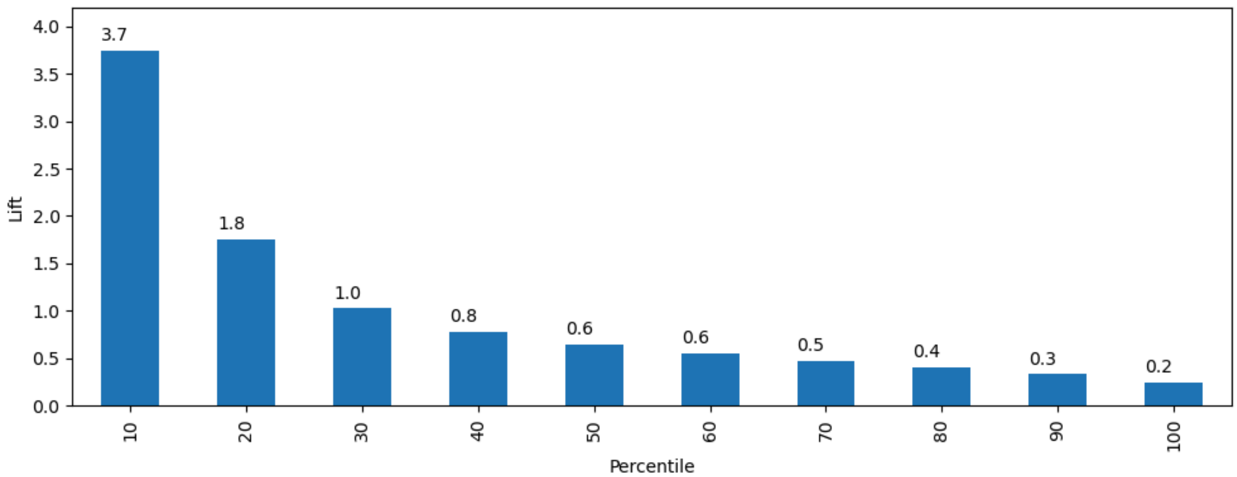
Training Partition
Confusion Matrix (Accuracy 0.8968)

Actual	Prediction 0	Prediction 1
0	151	153
1	6	1230

Validation Partition
Confusion Matrix (Accuracy 0.8971)

Actual	Prediction 0	Prediction 1
0	58	66
1	2	535

d. A decile lift chart shows how much better a logistic model is compared to random assignments. In the decile lift chart for delayed flight status we can see that the lift chart suggests that the model performs exceptionally well in the top 10% of the flights, with a lift value of 3.7, meaning it is 3.7 times better at identifying delayed flights compared to random selection. However, the model's performance deteriorates as we move down the deciles, becoming worse than random selection from the 7th decile onward. Performing worse than random selection past the 7th decile suggests that the predictive model needs further refinement and optimization to improve its effectiveness in identifying delayed flights.



3. Compare results of logistic regression model vs. classification tree model for the same data set.
   a. Since the accuracy between the logistic regression model and the optimize classification model on the validation partition produces a small difference, we can say that either of these models would be good for predicting flight status. There is one trade-off to compare in this case and that is interpretability. If we want to have a model that can be explained through coefficients/parameters then using the logistic regression model would be preferred.

Logistic Regression Model

```
Validation Partition
Confusion Matrix (Accuracy 0.8971)

       Prediction
Actual   0    1
     0  58   66
     1   2  535
```

Optimized Classification Tree Model

```
Confusion Matrix (Accuracy 0.8941)

       Prediction
Actual   0   1
     0  69  55
     1  15 522
```

4. Extra Credit
   a. Logistic regression model based on the best predictor variables from Backward Elimination. Predictors were removed from this model: ‘FL_NUM’

```
Parameters of Backwards Elimination Logistic Regression Model
Intercept: 0.547
              SCH_TIME  DEP_TIME  DISTANCE  WEATHER  WK_DAY  MTH_DAY  \
Coefficient:     0.025    -0.025     0.008    -2.88   0.057    -0.02

              CARRIER_DH  CARRIER_DL  CARRIER_MQ  CARRIER_OH  CARRIER_RU  \
Coefficient:        0.32       0.909      -0.619       1.462        0.11

              CARRIER_UA  CARRIER_US
Coefficient:       0.288       0.246
```

b. Based on the results below, either of these models can be chosen to predict FL_STATUS. This is because the difference in accuracy between the original model and the model created based off of features from backwards elimination is very similar. When looking at the confusion matrices for both, we can also see their ability to predict FL_STATUS = 0 is also similar. If I had to choose between these models then I would say that the model that uses the predictors from the backwards elimination is better since its accuracy is still higher. In reality, an exhaustive search or Ordinary Least Squares should be further applied to see which variables relate to the target variable.

```
Training Partition
Confusion Matrix (Accuracy 0.8968)

       Prediction
Actual    0    1
     0  151  153
     1    6 1230

Validation Partition
Confusion Matrix (Accuracy 0.8971)

       Prediction
Actual   0   1
     0  58  66
     1   2 535
```

```
Training Partition of Backwards Elimination Model
Confusion Matrix (Accuracy 0.8994)

       Prediction
Actual    0    1
     0  150  154
     1    1 1235

Validation Partition of Backwards Elimination Model
Confusion Matrix (Accuracy 0.8986)

       Prediction
Actual   0   1
     0  58  66
     1   1 536
```