**1. Develop a logistic regression model for the Flight Delays case.**

**a. Develop in Python the predictor variables (14 variables) and outcome variable (‘FL\_STATUS’) and partition the data set (80% for training and 20% for validation partitions). Train a logistic regression model using LogisticRegression() with the training 2 data set and display in Python the model’s parameters (intercept and regression coefficients). Provide these parameters in your report and also present the mathematical equation of the trained logistic regression model**.

To develop the predictor variables (14 variables) and outcome variable (‘FL\_STATUS’) and partition the data set:

**Text

Description automatically generated with low confidence**

To train the regression model using LogisticRegression() with the training 2 data set and display the model’s parameters (intercept and regression coefficients):

**Graphical user interface, text

Description automatically generated**

Outcome:

**Text, letter

Description automatically generated**

The mathematical equation of the trained logistic regression model:

**Logit =** 0.097 + 0.027\*SCH\_TIME - 0.028\*DEP\_TIME + 0.01\*DISTANCE + 0.0\*FL\_NUM - 0.53\*WEATHER + 0.048\*WK\_DAY - 0.022\*MTH\_DAY + 0.061\*CARRIER\_DH + 0.708\*CARRIER\_DL - 0.83\*CARRIER\_MQ + 0.375\*CARRIER\_OH - 0.012\*CARRIER\_RU + 0.074\*CARRIER\_UA - 0.042\*CARRIER\_US

**b. In Python, make predictions and identify probabilities p(0) and p(1) for the validation data set. For the first 50 records in the validation data set, display a table that contains the actual and predicted flight arrivalstatus, and probabilities p(0) and p(1). Present this table in your report, and comment on the predicted vs. actual flight arrival status.**

To make predictions and identify probabilities p(0) and p(1) for the validation data and to display a table that contains the actual and predicted flight arrival status, and probabilities p(0) and p(1) for first 50 records:

Text

Description automatically generated

Outcome:

Table

Description automatically generated

Table

Description automatically generated

From the table, it seems that the model is performing relatively well, with the vast majority of flights being correctly classified as either delayed or on time. However, there are some instances where the predicted value does not match the actual classification, such as with flights 1026, 1846, and 1957. These records are misclassified as their actual flight status was 0 (‘delayed’), but the logistic regression model classified them as 1 (‘ontime’).

**c. Identify and display in Python confusion matrices for the training and validation partitions. Present them in your report and comment on accuracy (misclassification) rate for both partitions and explain if there is a possibility of overfitting.**

To identify and display in Python confusion matrices for the training and validation partitions:

Text

Description automatically generated

Outcome:

A picture containing shape

Description automatically generated

According to the confusion matrices of the training and validation partitions, the logistic regression model has a high accuracy, which indicates it fits well for the validation date set and is suitability for classifying flight arrival status.

The misclassification rate:

Training Partition: 1 – Accuracy = 1 – 0.8994 = 10.06%

Validation Partition: 1 – Accuracy = 1 – 0.8934 = 10.66%

The accuracy of the model for the validation records is slightly higher than the training records, which suggests that overfitting is not a concern in this case. Therefore, the trained logistic regression model can be used with confidence for classification purposes.

**d. Create and display in Python the Lift chart only for ‘delayed’ flight status. For that, use p(0) for .sort\_values() and p(0) in liftChart(). Also use ncols=1 in plt.subplots() for a single plot, and remove ax=axes[1] from liftChart(). Present this Lift chart in your report and briefly explain what the chart demonstrates and what conclusion(s) can be made**

To create and display in Python the Lift chart only for ‘delayed’ flight status:

Text

Description automatically generated

Outcome:

Chart, bar chart

Description automatically generated

The lift chart for the ‘delayed’ flight status displays the comparison between the proportion of 0 (i.e., 'delayed') classifications generated by the model and the proportion of 'delayed' flight status selected randomly for different percentiles in the validation partition. In the top 10% of the data that is most likely to be 0 based on the logistic model, the proportion of 0's is 3.9 times higher than the proportion of 0's selected randomly.

**2. Compare results of logistic regression model vs. classification tree model for the same data set.**

**a. Present and compare in your report the validation confusion matrix for the logistic regression model in 2c of this case versus the validation confusion matrix using the GridSearchCV() algorithm for the classification tree in the previous case study. Using the accuracy value (misclassification rate), which model would you recommend applying for classification (prediction) of flight arrival status? Briefly explain your answer.**

The validation confusion matrix for the logistic regression model:

A picture containing shape

Description automatically generated

The validation confusion matrix using the GridSearchCV() algorithm for the classification tree:

A picture containing background pattern

Description automatically generated

The misclassification rate for the logistic regression model:

Training Partition: 1 – Accuracy = 1 – 0.8994 = 10.06%

Validation Partition: 1 – Accuracy = 1 – 0.8934 = 10.66%

The misclassification rate for the classification tree:

Training Partition: 1 – Accuracy = 1 – 0.8960 = 10.40%

Validation Partition: 1 – Accuracy = 1 – 0.8685 = 13.15%

As shown above, the misclassification rate for the validation partition in the classification tree model (13.15%) is higher than the one in the logistic regression model (10.66%). Thus, considering the two given confusion matrices, **the logistic regression model** achieved the highest accuracy of 0.8934 or 89.34% in the validation partition. Therefore, we recommend to use this model for the classification of the flight arrival status.