**CASE STUDY 3**

1. **a**. Logistic regression extends the ideas of linear regression to the situation where the outcome variable, Y, is categorical. The Logistic regression model is used for classifications.

In flight\_df data frame, the outcome variable, FL\_STATUS is a categorical variable with two classes (‘ontime’ and ‘delayed’). Hence the logistic regression model may be used in this case to predict a new observation into an existing class, referred to as classification. The multiple linear regression model is used for doing predictions of the numerical outcome variable. Here the outcome variable is not numerical but categorical which is to be classified into an existing class. Hence the multiple linear regression model is not applied in this case.

In multiple linear regression the aim is to predict the value of the continuous Y for a new record, in logistic regression the goal is to predict which class a new record will belong to, or simply to classify the record into one of the classes.

**b.** The flight\_df Data frame dimensions:

Number of rows in the data frame: 2201

Number of columns in the data frame: 11

Shape of the data frame (rows, columns): (2201, 11)

After removing the ‘DEST’ and ‘ORIGIN’ columns from the flight\_df data frame, the data types of the remaining columns in flight\_df data frame are shown below.

SCH\_TIME int64

CARRIER object

DEP\_TIME int64

DISTANCE int64

FL\_NUM int64

WEATHER int64

WK\_DAY int64

MTH\_DAY int64

FL\_STATUS object

The data types of the columns are integer data type for all the columns except for ‘CARRIER’ and ‘FL\_STATUS’, which have object data types. The ‘CARRIER’ and ‘FL\_STATUS’ are converted to binary variables

Modified list of 15 column variables (including dummy variables):

Index(['SCH\_TIME', 'DEP\_TIME', 'DISTANCE', 'FL\_NUM', 'WEATHER', 'WK\_DAY', 'MTH\_DAY', 'CARRIER\_DH', 'CARRIER\_DL', 'CARRIER\_MQ', 'CARRIER\_OH', 'CARRIER\_RU', 'CARRIER\_UA', 'CARRIER\_US', 'FL\_STATUS'], dtype='object')

Modified data types of columns in data frame:

SCH\_TIME int64

DEP\_TIME int64

DISTANCE int64

FL\_NUM int64

WEATHER int64

WK\_DAY int64

MTH\_DAY int64

CARRIER\_DH uint8

CARRIER\_DL uint8

CARRIER\_MQ uint8

CARRIER\_OH uint8

CARRIER\_RU uint8

CARRIER\_UA uint8

CARRIER\_US uint8

FL\_STATUS\_ontime uint8

**c.** We choose to convert an outcome variable with two classes into a binary outcome variable for purposes of simplification, reflecting the fact that decision-making is binary (1 for ‘ontime’ and 0 for ‘delayed’). For the logistic regression the left hand side of the equation is not Y(numerical variable) to be predicted, but, logit (log of odds), which maps to probabilities, in turns help to classify the outcome categorical variable into an existing class.

The outcome variable that we need to classify is flight arriving status i.e., “FL\_STATUS\_ontime” using the 14 predictor variables such as 'SCH\_TIME', 'DEP\_TIME', 'DISTANCE', 'FL\_NUM', 'WEATHER', 'WK\_DAY', 'MTH\_DAY', 'CARRIER\_DH', 'CARRIER\_DL', 'CARRIER\_MQ', 'CARRIER\_OH', 'CARRIER\_RU', 'CARRIER\_UA', 'CARRIER\_US'.

The goal is to build a logistic regression model that identifies flights which are most likely to be on time in future schedules.

We can also identify flights which are most likely to be delayed too in future schedules using logistic regression model.

1. **a**. Parameters of Logistic Regression Model with Multiple Predictors

Intercept: 0.051

Coefficients for Predictors

SCH\_TIME DEP\_TIME DISTANCE FL\_NUM WEATHER WK\_DAY MTH\_DAY \

Coeff: 0.033 -0.034 0.009 0.0 -0.247 0.097 -0.021

CARRIER\_DH CARRIER\_DL CARRIER\_MQ CARRIER\_OH CARRIER\_RU \

Coeff: 0.35 0.782 -1.054 0.236 -0.05

CARRIER\_UA CARRIER\_US

Coeff: 0.053 -0.043

The mathematical equation of the trained logistic regression model:



Logit = 0.051 + 0.033\*SCH\_TIME - 0.034\*DEP\_TIME + 0.009\*DISTANCE + 0.0\*FL\_NUM - 0.247\*WEATHER + 0.097\*WK\_DAY - 0.021\*MTH\_DAY + 0.035\*CARRIER\_DH + 0.782\*CARRIER\_DL - 1.054\*CARRIER\_MQ + 0.236\*CARRIER\_OH - 0.05\*CARRIER\_RU + 0.053\*CARRIER\_UA - 0.043\*CARRIER\_US

Where logit = log(Odds)

**b.** Classification for Validation Partition

Actual Classification p(0) p(1)

1276 1 1 0.1392 0.8608

1446 1 1 0.0795 0.9205

335 1 1 0.0916 0.9084

1458 1 1 0.1136 0.8864

2038 1 1 0.0738 0.9262

1314 1 1 0.0672 0.9328

389 1 1 0.1599 0.8401

1639 1 1 0.1263 0.8737

2004 1 1 0.0951 0.9049

403 1 1 0.3002 0.6998

979 1 1 0.0488 0.9512

65 1 1 0.0691 0.9309

2105 1 1 0.1434 0.8566

1162 1 1 0.1037 0.8963

572 1 1 0.3036 0.6964

1026 0 1 0.0649 0.9351

1044 1 0 0.5242 0.4758

1846 0 1 0.4611 0.5389

1005 1 1 0.1545 0.8455

1677 1 1 0.0503 0.9497

It is observed that, in the first twenty records of the validation partition, only three records are classified incorrectly (1026, 1044, 1846). The actual values are 0, 1 and 0 where as the classifications are 1,0 and 1 respectively for the mentioned records.

**c.** Confusion matrices for Original logistic model:

Training Partition

Confusion Matrix (Accuracy 0.9015)

Prediction

Actual 0 1

0 141 120

1 10 1049

Validation Partition

Confusion Matrix (Accuracy 0.8990)

Prediction

Actual 0 1

0 90 77

1 12 702

**Training data set:**

Accuracy = 141+1049/1320 = 0.9015 = 90.15 %

Misclassification = 1- 0.9015 = 0.0985 = 9.85 %

**Validation data set:**

Accuracy = 90+702/881 = 0.8990 = 89.90%

Misclassification = 1- 0.8990 = 0.101 = 10.1 %

There is not much difference in the accuracy scores of training and validation sets. We can say that there is no possibility of overfitting for this model. Hence this model can be used for making classifications.

d. Chart

Description automatically generated

In this Lift chart, taking the 10% of the records that are ranked by the model as 'most probable 0's' yields 4.1 times as many 0's as would simply selecting 10% of the records at random.

In other words, there are only 19.45 percent of total records i.e. 428 out of 2201 records with flight status as ‘delayed’. If we pick randomly 10%(220) of the records, 43(19.45% of 220) records will have flight status as ‘delayed’. But the logistic regression model yields 4.1 times(4.1\*43 = 176 records) as many 0’s (delayed flights) as would simply selecting 10 % of the records at random.

1. a. Using the GridSearchCV() algorithm from case study 2 with control parameters as following: (a) maximum depth (number of splits) in the range from 2 to 30; (b) minimum impurity decrease per split of 0, 0.0005, and 0.001; and (c) minimum number of node records (samples) to split in the range from 5 to 30.

Each time we run the GridSearchCV() and Classification tree, the results are varying. The most frequently displayed result is shown here.

Improved score:0.8659

Improved parameters: {'max\_depth': 11, 'min\_impurity\_decrease': 0.001, 'min\_samples\_split': 11}

**Confusion matrices for grid search classification tree:**

**Training Partition**

Confusion Matrix (Accuracy 0.9250)

Prediction

Actual 0 1

0 169 92

1 7 1052

**Validation Partition**

Confusion Matrix (Accuracy 0.8785)

Prediction

Actual 0 1

0 90 77

1 30 684

The difference in the accuracy scores of training and validation data sets is not significant enough to claim the possibility of overfitting. We can say that there is no possibility of overfitting for this model. Hence this model can be used for making classifications.

b) **Original logistic model:**

**Training data set:**

Accuracy = 141+1049/1320 = 0.9015 = 90.15 %

Misclassification = 1- 0.9015 = 0.0985 = 9.85 %

**Validation data set:**

Accuracy = 90+702/881 = 0.8990 = 89.90%

Misclassification = 1- 0.8990 = 0.101 = 10.1 %

**Grid search classification tree**

**Training data set:**

Accuracy = 169+1052/1320 = 0.9250 = 92.50 %

Misclassification = 1- 0.9250 = 0.075 = 7.5 %

**Validation data set:**

Accuracy = 90+684/881 = 0.8785 = 87.85 %

Misclassification = 1- 0.8785 = 0.1215 = 12.15 %

Although there is no possibility of overfitting in both the logistic regression model and the Grid Search Classification tree, the accuracy rates for the validation data sets of the **Logistic regression model** are more compared to the accuracy rates for the validation data of the Grid search classification tree. The **Logistic regression model** is recommended for making classifications in this case of flight arrival status.

1. **a.** Actual 14 predictor Variables: SCH\_TIME, DEP\_TIME, DISTANCE, FL\_NUM, WEATHER, WK\_DAY, MTH\_DAY, CARRIER\_DH, CARRIER\_DL, CARRIER\_MQ, CARRIER\_OH, CARRIER\_RU, CARRIER\_UA, CARRIER\_US

Best Variables from Backward Elimination Algorithm (11 predictor variables):

['SCH\_TIME', 'DEP\_TIME', 'DISTANCE', 'WEATHER', 'MTH\_DAY', 'CARRIER\_DH', 'CARRIER\_DL', 'CARRIER\_MQ', 'CARRIER\_OH', 'CARRIER\_UA', 'CARRIER\_US']

The variables such as ‘FL\_NUM’, ‘WK\_DAY’, ‘CARRIER\_RU’ are removed from this model in Backward Elimination Algorithm.

**Logistic Regression Model for Training Set Using Backward Elimination**

Intercept [0.1]

Predictor Coefficient

0 SCH\_TIME 0.03

1 DEP\_TIME -0.03

2 DISTANCE 0.01

3 WEATHER -0.36

4 MTH\_DAY -0.02

5 CARRIER\_DH 0.31

6 CARRIER\_DL 0.87

7 CARRIER\_MQ -1.17

8 CARRIER\_OH 0.32

9 CARRIER\_UA 0.09

10 CARRIER\_US -0.05

**b. Confusion matrices for logistic regression model using Backward Elimination:**

Training Partition

Confusion Matrix (Accuracy 0.9030)

Prediction

Actual 0 1

0 146 115

1 13 1046

Validation Partition

Confusion Matrix (Accuracy 0.8978)

Prediction

Actual 0 1

0 87 80

1 10 704

The difference in the accuracy scores of training and validation data sets is not significant enough to claim the possibility of overfitting. We can say that there is no possibility of overfitting for this model. Hence this model can be used for making classifications.

Confusion matrices for original logistic regression model:

Training Partition

Confusion Matrix (Accuracy 0.9015)

Prediction

Actual 0 1

0 141 120

1 10 1049

Validation Partition

Confusion Matrix (Accuracy 0.8990)

Prediction

Actual 0 1

0 90 77

1 12 702

There is not much difference in the accuracy scores of both the training and validation data and no possibility of overfitting in both the original logistic regression model and the logistic regression model based on Backward elimination algorithm, the accuracy rates for the validation data sets of the original Logistic regression model (with 14 predictors) and logistic regression model based on Backward elimination algorithm (with 11 predictors) are almost same. Both methods are performing well in terms of classification of the validation data.

The accuracy rates for the validation data sets of the original Logistic regression model (with 14 predictors) slightly more (0.8990) compared to the accuracy rates for the validation data of the logistic regression model based on Backward elimination algorithm (0.8978). If we want the classifications to be more accurate, the **original** **Logistic regression model** is recommended for making predictions in this case of flight arrival status.

However, the accuracy scores of validation data are not much different for two models. The parsimonious model which is the logistic regression model based on the **Backward Elimination** with 11 number of predictors can be recommended for making classifications. The simplest model that performs better results.