DEPARTMENT OF STATISTICS, ACTUARIAL AND DATA SCIENCE

CENTRAL MICHIGAN UNIVERSITY

MT. PLEASANT MI 48859

FALL 2021

Team Project for STA 591

Instructor: Dr. Carl Lee

Predicting Delays in Domestic Flights Operated by Large Air Carriers in the US.

Submitted by:

1. Adefolu, Aminat
2. Esivwenughwu, Emmanuel
3. Joslin, Priscilla
4. Ofori Aboah, Valerie

**ABSTRACT**

This study investigates the various factors that cause delays in flight arrivals and builds predictive models to determine if a flight would be delayed based on these factors. The major data mining and modelling software used in this project were SAS Enterprise Miner 14 and SAS base 9.4.

The target variable selected was Delay\_On\_Arrival which was created from ARRIVAL\_DELAY. We chose this as our target variable since the purpose of the study is to predict the delays in Domestic flights operated by Large Air Carriers in the US.

From the Chi-square and Cramer’s V plots, ARRIVAL\_TIME had the most impact on the target variable, Delay\_On\_Arrival, however in the Variable Worth’s plot, DEPARTURE\_DELAY had the most impact on the target variable, Delay\_On\_Arrival.

Our Mid-Term investigation so far has identified some important factors/input variables which include DEPARTURE\_DELAY, ARRIVAL\_TIME, WHEELS\_ON, WHEELS\_OFF, TAXI\_OUT, DEPARTURE\_TIME and SCHEDULED\_DEPARTURE. We considered these factors in building our predictive models as they had significant impacts on the target variable, Delay\_On\_Arrival, from the Chi-square, Cramer’s V, and Variable Worth plots we prepared.

**PURPOSE OF THE PROJECT**

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations. The purpose of this project includes the following:

1. To identify the various factors (inputs variables) associated with the target variable (Delay\_On\_Arrival).
2. To develop predictive models and select the best one that accurately predicts whether a particular flight would be delayed using the factors considered in 1.

**LITERATURE REVIEW**

“Generation and Prediction of Flight Delays in Air Transport” written by Qiang Li and Ranzhe Jing analyzes 2018 flight delay data from the Department of Transportation. The research goal was to discover the cause of flight delays and minimize their occurrence. Their primary method of modeling the data was regression with 93.92% of their predictive test errors within 15 min. Rather than keeping the given inputs, the researches divided them into two categories: Network and local. This replaced the existing factors with more novel factors. The picture of the map represents the network delays caused by larger airports earlier in the day. Li and Jing found that long term flight delays were caused by delays at larger airports earlier in the day whereas short term delays were more strongly associated with local delays.

**DATA SOURCES AND DESCRIPTION**

The data was selected from <https://www.kaggle.com/> which provides publicly available datasets ranging from various aspects of human endeavor and industries, medicine, travel, entertainment etc. The dataset we used is the 2015 Flight Delays and Cancellations created by the U.S. Department of Transportation. The data consisted of over 5 million observations and 31 variables which include AIRLINE, LATITUDE, LONGITUDE, ARRIVAL\_TIME, DEPARTURE\_TIME, WHEELS\_ON etc.

The dataset can be found at: <https://www.kaggle.com/usdot/flight-delays>.

**DATA CLEANING AND METHODS FOR VARIABLE SELECTION AND TRANSFORMATION**

For this study, we restricted the year to only flight delays in 2015.

The CSV file downloaded from the data source was read and converted to a SAS dataset using the PROC IMPORT and a total of 50,000 records was selected using simple random sampling.

Also, using the IF condition, a new binary variable (“Delay\_On\_Arrival”) was created based on the variable “ARRIVAL\_DELAY”, where a number 1 was assigned if “ARRIVAL\_DELAY” was greater than 15 and 0 otherwise. This made the total number of variables to be 32. In these 32 variables, we had 19 input variables, 12 rejected variables and 1 target variable. For level, we had 3 binary variables, 7 nominal variables, 2 unary and 20 interval variables.

This new variable (“Delay\_On\_Arrival”) was chosen as our target variable and since it gives the same information as “ARRIVAL\_DELAY”, “ARRIVAL\_DELAY” was rejected, using the default **“Advanced Advisor settings”**. Also, under the **“Advanced Advisor Settings”**, we made the Class Levels Count Threshold 20, and the Missing Percentage Threshold 50. This by default rejected variables that exceeded this threshold.

Table

Description automatically generated

Data Partition done was **Training: 50.0, Validation: 30.0, Test: 20.0**

Further Variable selection was later considered after we explored the data using StatExplot and Multiplot.

**DESCRIPTIVE SUMMARY OF TARGET AND SELECTED INPUTS**

The summary statistics of the Selected Input variables and Target Variable from the StatExplore Node is shown below:

For the Interval Input Variables

**Graphical user interface, application, table, Excel

Description automatically generated**

For the Class Input Variables

A screenshot of a computer

Description automatically generated with low confidence

The Variable summary statistics from the Output of the StatExplore is given below:

For the Class Variables:

Table

Description automatically generated

From the summary statistics for the class variables above, 2.29% of flights were cancelled, 31.22% of flights occurred on day\_of\_week 4 which is a Wednesday.

For the Interval Variables:

Table

Description automatically generated

From the summary statistics for the interval variables above, the average distance between airports is 850.1379 miles and the minimum distance between airport is 31 miles and maximum is 4983 miles.

Summary Statistics of the Class variable by Class Target from the Output of the StatExplore Node is shown below:

A picture containing text, receipt

Description automatically generated

Summary Statistics of the Interval variable by Class Target from the Output of the StatExplore Node is shown below:

Table, calendar

Description automatically generated

Table

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

**DESCRIPTIVE SUMMARY AND ASSOCIATION BETWEEN TARGETS AND SELECTED INPUTS**

This aspect includes results from the StatExplore and Multiplots nodes which involved exploratory analysis of the selected inputs and targets using graphs and the summary statistics from the output window.

The Chi-Square graph, Cramer’s V graph and Variable Worth graph and their corresponding table from the StatExplore node is shown below:

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Table

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

From the Chi-Square, Cramer V’s and Variable Worth’s plots of the StatExplore node, we see that the top seven input variables (which are consistent for the three plots) associated with the Target Variable (“Delay\_On\_time”) in no particular order are:

* DEPARTURE\_DELAY
* ARRIVAL\_TIME
* WHEELS\_ON
* WHEELS\_OFF
* TAXI\_OUT
* DEPARTURE\_TIME
* SCHEDULED\_DEPARTURE

For the graphs of the Multiplot node, we considered some of the mentioned input variables with that of the input variable:

Here, we see that there is truly a difference in the ARRIVAL\_TIME of the two levels of “Delay\_On\_Arrival”, which shows that this must be an important input that affects the target.

Chart, bar chart

Description automatically generatedAlso, the same is true for “DEPARTURE\_TIME” which is shown below:

Chart, bar chart

Description automatically generated

**METHOD OF ANALYSIS, RESULTS AND DISCUSSIONS**

Since our target is a class target which is binary, we considered two types of models: Decision tree and Regression models and based on the options we selected, we had a total of four (4) models.

Below is a diagram to show the various nodes connecting the input with various models we selected:

Diagram

Description automatically generated

Results of the various nodes and models will be discussed next.

**MODEL 1: DECISION TREE**

For this model, we considered the various inputs without any transformation (that is no transformation variable node was introduced prior to this model).

The default splitting (growing the tree based on training data) rule was **ProbChiq** because we had a class target which is binary.

Method selected for the optimal tree size (pruning the tree based on validation data) was **Assessment** and the assessment measure selected was **Misclassification** which selects the tree which minimizes the misclassification rate.

Results from the Decision tree node are displayed below:

Timeline

Description automatically generated with medium confidence

**Tree Diagram of model 1**

Table

Description automatically generated

**Variable Importance of model 1**

The above variable importance shows the extra contribution to the target variable each input variable has when other input variables are also considered.

**Graphical user interface, text, application

Description automatically generated**

**Fit Statistics table of model 1**

**Chart, bar chart

Description automatically generated**

**Leaf Statistics plot of model 1**

**Note: Full tree diagram is available in the xml sent via email for this project.**

**MODEL 2: DECISION TREE (2)**

For this model, we considered a transformation before the model. For the transformation node, the default methods selected are:

Interval inputs – Maximum Normal

Interval targets – None

Class inputs – Maximum Normal

Class targets - None

The default splitting (growing the tree based on training data) rule was **ProbChiq** because we had a class target which is binary.

Method selected for the optimal tree size (pruning the tree based on validation data) was **Assessment** and the assessment measure selected was **Average Square Error** which selects the tree which minimizes the average square error.

Results from the Decision tree (2) node are displayed below:

A picture containing diagram

Description automatically generated

**Tree Diagram of model 2**

**Table

Description automatically generated with medium confidence**

**Variable Importance of model 2**

The above variable importance shows the extra contribution to the target variable each input variable has when other input variables are also considered.

**Graphical user interface, text, application

Description automatically generated**

**Fit Statistics table of model 2**

**Chart, bar chart, histogram

Description automatically generated**

**Leaf Statistics plot of model 2**

**Note: Full tree diagram is available in the xml sent via email for this project.**

**IMPUTATION NODE OPTIONS FOR THE REGRESSION MODELS**

The various options were selected in the imputation node which was connected prior to every of the two regression models.

CLASS VARIABLE

DEFAULT INPUT METHOD– Tree

DEFAULT TARGET METHOD– None

NORMALIZE VALUES - Yes

INTERVAL VARIABLE

DEFAULT INPUT METHOD– Tree

DEFAULT TARGET METHOD– None

**MODEL 3: REGRESSION**

For the first regression model (model 3: regression), we considered the various inputs without any transformation (that is no transformation variable node was introduced prior to this model). The various options selected for this model are:

REGRESSION TYPE – Logistic Regression

LINK FUNCTION – Logit

SELECTION MODEL – Stepwise

SELECTION CRITERION – Akaike Information Criterion (AIC).

Results from the Regression node are displayed below:

Table

Description automatically generated with medium confidence

**Likelihood Ratio Test for Testing Global Hypothesis of model 3**

**Table

Description automatically generated**

**Type 3 Analysis of Effect of model 3**

**Table

Description automatically generated**

**Analysis of Maximum Likelihood Estimates of model 3**

**Text

Description automatically generated**

**Odds Ratio Estimates of model 3**

From the above, for input variable: **CANCELLED** level 0 is 999 times better at predicting the values of the target variable: **Delay\_On\_Arrival** than level 1.

For a unit increase in the value of the target variable Delay\_On\_Arrival, there is on average a 18.6% increase of **IMP\_DEPATURE\_DELAY**.

For a unit increase in the value of the target variable Delay\_On\_Arrival, there is on average a 14.4% increase of **IMP\_TAXI\_IN**

For a unit increase in the value of the target variable **Delay\_On\_Arrival**, there is on average a 14.6% increase of **IMP\_TAXI\_OUT**

**Graphical user interface, application

Description automatically generated**

**Fit Statistics of model 3**

**MODEL 4: REGRESSION (2)**

For the second regression model (model 4: regression (2)), we considered a transformation before the model. For the transformation node, the default methods selected are:

Interval inputs – Maximum Normal

Interval targets – None

Class inputs – Maximum Normal

Class targets – None

The various options selected for this model are:

REGRESSION TYPE – Logistic Regression

LINK FUNCTION – Logit

SELECTION MODEL – Stepwise

SELECTION CRITERION – Validation Misclassification.

Results from the Regression (2) node are displayed below:

A picture containing application

Description automatically generated

**Likelihood Ratio Test for Testing Global Hypothesis of model 4**

Table

Description automatically generated

**Type 3 Analysis of Effect of model 4**

**Table

Description automatically generated**

**Analysis of Maximum Likelihood Estimates of model 4**

**Text

Description automatically generated**

**Odds Ratio Estimates of model 4**

From the above, for input variable: **CANCELLED** level 0 is 999 times better at predicting the values of the target variable: **Delay\_On\_Arrival** than level 1

**Graphical user interface, application

Description automatically generated**

**Fit Statistics of model 4**

**MODEL COMPARISON:**

This node was connected to compare the four models above and select the best (optimal) of these models. The model selection options chosen are:

SELECTION STATISTICS – ROC

SELECTION TABLE – Test

Results from the model comparison node which includes the ROC chart and the table showing the selected model are shown below:

Graphical user interface, chart, line chart

Description automatically generated

**ROC chart of the model comparison node**

**Table

Description automatically generated**

**Partial table showing the Fit statistics table of the model comparison node and best model selected based on ROC index of the test table**

From the above the selected model is model 3: the first regression model because it has the highest ROC index based on the test table compared to the other models.

**SUMMARY AND CONCLUSION**

From all models we considered, the results and findings all point to some key points about the dataset.

1. Information about flights delay on arrival (either Yes = 1 or No = 0) would not be possible if the flights was cancelled (that is value for the cancelled variable is 1), these was shown in the odd estimates table of the two regression models. This is intuitive, in the sense that if a flight never departs, it is pointless (makes no sense) to talk about its delay on arrival at its destination.
2. A flights departure delay is highly associated with delay on arrival, this was shown by all models where the variable importance of the input variable DEPARTURE\_DELAY was high which signifies this input contributes more to the target Delay\_On\_Arrival when all other inputs are also considered. This makes perfect sense because a flight which departs late is likely to arrive late.
3. Taxi in and Taxi out also played major roles in affecting results of the target variable. Though they mean different things depending on the context. In general, **taxi out** is seen as the period during which a flight is ready to take off and the various checks done before this while **taxi in** is a similar situation during a flight’s arrival. In both cases, if the airport is a busy one (more flights operating at that time), each of these flight’s arrival / departure time will be affected.

**LIMITS AND FUTURE WORK**

Our research successfully identified models used for predicting whether a flight delay will occur or not. In the future, our goal would be to use these models to limit the length of flight delay and their occurrence at all. This would require a more in-depth analysis of our models and then extrapolate the knowledge gained to have a practical application. Additionally, our team could use the knowledge gained from this project and apply it to international flight data or to researching a link between lost luggage and flight delays.

Due to time and resources, our team was unable to track the effect of switching airlines, or which airlines more often resulted in flight delays. It could be helpful to track flight delays for each individual airport or airline and see how the model differs from our current national flight delay model.

**REFERENCES**

Below is a list of references for this project:

1. <https://www.kaggle.com/usdot/flight-delays>
2. STA 591 course materials from Central Michigan University’s blackboard site: <https://blackboard.cmich.edu/webapps/blackboard/content/listContent.jsp?course_id=_199249_1&content_id=_8179079_1&mode=reset>
3. Idea on useful prediction that can be done with the dataset: <https://www.kaggle.com/usdot/flight-delays/code>
4. Literature Review: Li Q, Jing R. Generation and prediction of flight delays in air transport. IET Intelligent Transport System (Wiley-Blackwell). 2021;15(6):740-753. doi:10.1049/itr2.12057

**APPENDIX**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name:** | **Description:** | **Role:** | **Level:** |
| DEPARTURE\_TIME | WHEEL\_OFF - TAXI\_OUT | Input | Interval |
| DISTANCE | Distance between two airports | Input | Interval |
| DEPARTURE\_DELAY | Total Delay on Departure | Input | Interval |
| ELAPSED\_TIME | AIR\_TIME+TAXI\_IN+TAXI\_OUT | Input | Interval |
| FLIGHT\_NUMBER | Flight Identifier | Input | Interval |
| DIVERTED | Aircraft landed on airport that out of schedule | Input | Binary |
| AIRLINE | Airline Identifier | Input | Nominal |
| ARRIVAL\_TIME | WHEELS\_ON+TAXI\_IN | Input | Interval |
| AIR\_TIME | The time duration between wheels\_off and wheels\_on time | Input | Interval |
| DAY | Day of the Flight Trip | Input | Nominal |
| DAY\_OF\_WEEK | Day of week of the Flight Trip | Input | Nominal |
| CANCELLED | Flight Cancelled (1 = cancelled) | Input | Binary |
| SCHEDULED\_TIME | Planned time amount needed for the flight trip | Input | Interval |
| WHEELS\_OFF | The time point that the aircraft's wheels leave the ground | Input | Interval |
| SCHEDULED\_DEPARTURE | Planned Departure Time | Input | Interval |
| TAXI\_OUT | The time duration elapsed between departure from the origin airport gate and wheels off | Input | Interval |
| TAXI\_IN | The time duration elapsed between wheels-on and gate arrival at the destination airport | Input | Interval |
| WHEELS\_ON | The time point that the aircraft's wheels touch on the ground | Input | Interval |
| SCHEDULED\_ARRIVAL | Planned arrival time | Input | Interval |
| WEATHER\_DELAY | Delay caused by weather | Rejected | Interval |
| YEAR | Year of the Flight Trip | Rejected | Unary |
| AIRLINE\_DELAY | Delay caused by the airline | Rejected | Interval |
| ARRIVAL\_DELAY | ARRIVAL\_TIME-SCHEDULED\_ARRIVAL | Rejected | Interval |
| CANCELLATION\_REASON | Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security | Rejected | Nominal |
| ORIGIN\_AIRPORT | Starting Airport | Rejected | Nominal |
| DESTINATION\_AIRPORT | Destination Airport | Rejected | Nominal |
| MONTH | Month of the Flight Trip | Rejected | Unary |
| LATE\_AIRCRAFT\_DELAY | Delay caused by aircraft | Rejected | Interval |
| TAIL\_NUMBER | Aircraft Identifier | Rejected | Nominal |
| AIR\_SYSTEM\_DELAY | Delay caused by air system | Rejected | Interval |
| SECURITY\_DELAY | Delay caused by security | Rejected | Interval |
| Delay\_On\_Arrival | Flight Delay greater than 15 minutes = 1. | Target | Binary |