Analysis of airport On-Time Performance (SAS EM Project)

Mohamed El Faiz[1]

[1] melfaiz@student.uliege.be (s198260)

I. INTRODUCTION

Flight delay is one of the most common and unpleasant experience that people dread to have. Every year, a lot of flights get delayed which involves some cost both for the airline and the passenger in different ways. The passenger's time and money get affected and at the same time, the airline's reputation is at stake. Delay is treated as one of the most remembered performance indicators of the airline.

Therefore, statistics of the flight delays becomes crucial factor in understanding the flight's performance.

This study presents the analysis driven from flight delay data for the Logan International Airport from the United States for the year 2018. This study analyses the variety of factors responsible for and associated with flight delays for different airlines.

Logan International Airport is an international airport that is located mostly in East Boston and partially in Massachusetts, United States. It opened in 1923, has six runways and four passenger terminals, and employs an estimated 16,000 people. It is the largest airport in both the Commonwealth of Massachusetts and the New England region in terms of passenger volume and cargo handling, as well as the 16th-busiest airport in the United States. The airport saw 42,522,411 passengers in 2019, the most in its history.

FIG. 1. Aerial view of Logan International Airport

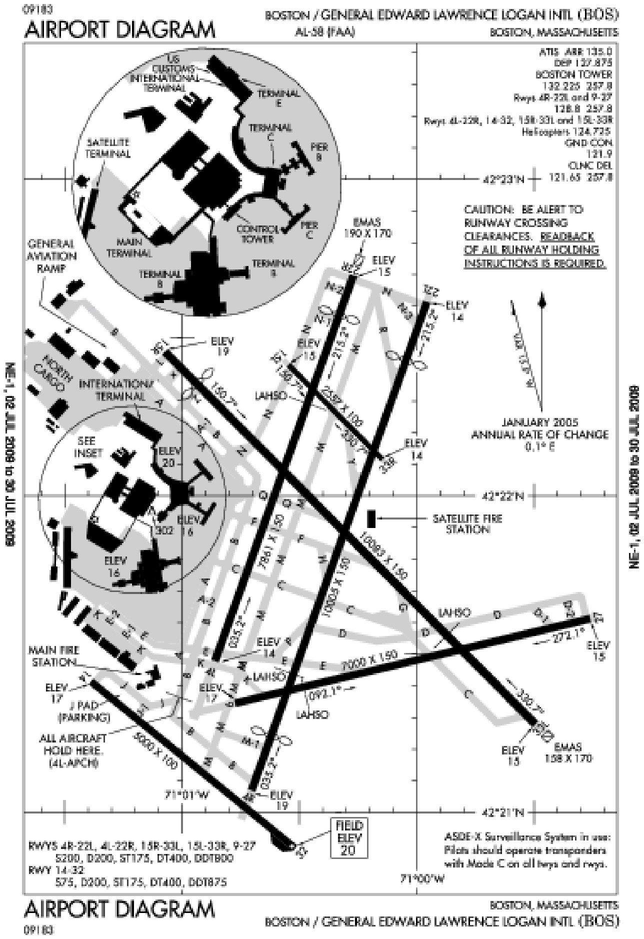
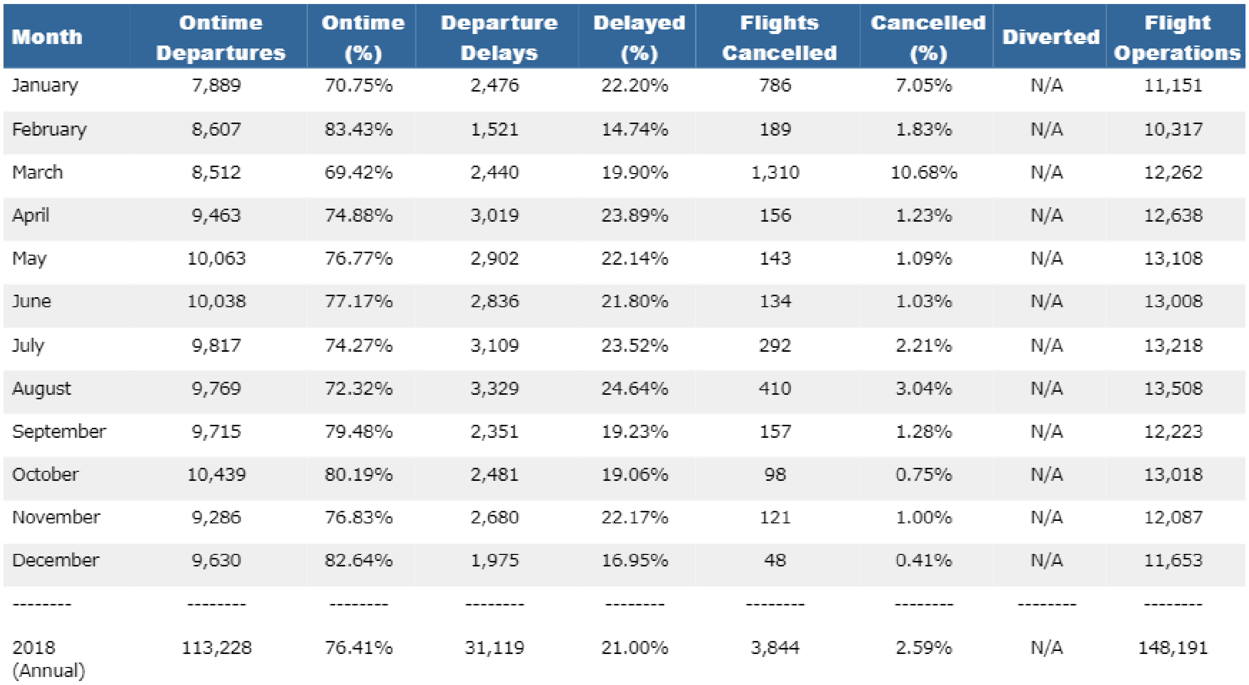


FIG. 2. Logan International Airport Diagram

II. DATA COLLECTION AND PREPARATION

The data was obtained from the Bureau of Transportation Statistics at the United States Department of Transportation.

The first database is the Airline Ontime Performance Data [2]. It contains scheduled and actual departure and arrival times reported by certified U.S. air carriers that account for at least one percent of domestic scheduled passenger revenues. The database contains on-time arrival data for non-stop domestic flights by major air carriers, and provides such additional items as day of month, day of week, original airport ID, destination airport ID, departure time, departure delay, departure delay 15 minutes, departure time slots. It also provides such addi-

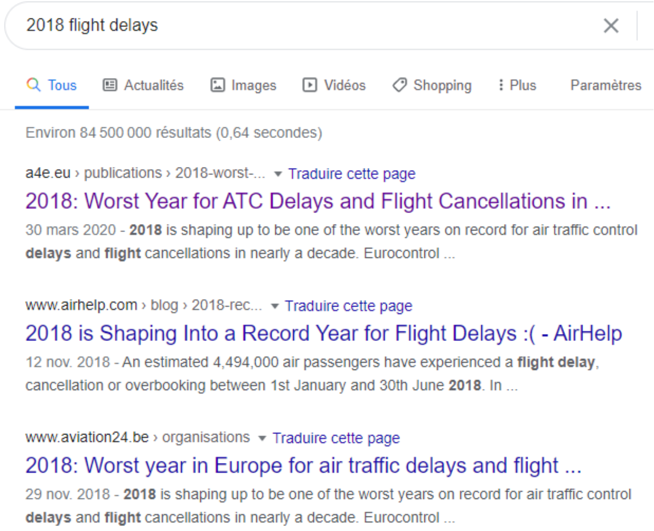


Month	Ontime Departures	Ontime (%)	Departure Delays	Delayed (%)	Flights Cancelled	Cancelled (%)	Diverted	Flight Operations
January	7,889	70.75%	2,476	22.20%	786	7.05%	N/A	11,151
February	8,607	83.43%	1,521	14.74%	189	1.83%	N/A	10,317
March	8,512	69.42%	2,440	19.90%	1,310	10.68%	N/A	12,262
April	9,463	74.88%	3,019	23.89%	156	1.23%	N/A	12,638
May	10,063	76.77%	2,902	22.14%	143	1.09%	N/A	13,108
June	10,038	77.17%	2,836	21.80%	134	1.03%	N/A	13,008
July	9,817	74.27%	3,109	23.52%	292	2.21%	N/A	13,218
August	9,769	72.32%	3,329	24.64%	410	3.04%	N/A	13,508
September	9,715	79.48%	2,351	19.23%	157	1.28%	N/A	12,223
October	10,439	80.19%	2,481	19.06%	98	0.75%	N/A	13,018
November	9,286	76.83%	2,680	22.17%	121	1.00%	N/A	12,087
December	9,630	82.64%	1,975	16.95%	48	0.41%	N/A	11,653
--------	--------	--------	--------	--------	--------	--------	--------	--------
2018 (Annual)	113,228	76.41%	31,119	21.00%	3,844	2.59%	N/A	148,191

FIG. 3. Report for delayed and cancelled flights in Logan International in 2018 [1]

tional items as flight numbers, taxi-out and taxi-in times, air time, and non-stop distance.

Both Carrier Informations [3] and Aircraft Informations [4] databases were also joined to the first database to get additional features.



2018 flight delays

Environ 84 500 000 résultats (0,64 secondes)

a4e.eu › publications › 2018-worst-... ▾ Traduire cette page

2018: Worst Year for ATC Delays and Flight Cancellations in ...

30 mars 2020 - **2018** is shaping up to be one of the worst years on record for air traffic control **delays** and **flight** cancellations in nearly a decade. Eurocontrol ...

www.airhelp.com › blog › 2018-rec... ▾ Traduire cette page

2018 is Shaping Into a Record Year for Flight Delays :(- AirHelp

12 nov. 2018 - An estimated 4,494,000 air passengers have experienced a **flight delay**, cancellation or overbooking between 1st January and 30th June **2018**. In ...

www.aviation24.be › organisations ▾ Traduire cette page

2018: Worst year in Europe for air traffic delays and flight ...

29 nov. 2018 - **2018** is shaping up to be one of the worst years on record for air traffic control **delays** and **flight** cancellations in nearly a decade. Eurocontrol ...

FIG. 4. Top Google results after searching for ”2018 flight delays”

The time period chosen for this project was the year 2018 mainly because we don’t have Aircraft Information [4] after 2018 but also because it was considered a very busy year that had the most flight delays ofer the decade which gives us more delays data to work on. We will only work on non-cancelled and direct flights. Moreover, for purposes of this study, a flight is considered delayed if it arrived at (or departed) the gate 15 minutes or more after the scheduled arrival (departure) time as noted in the Ontime performance dataset.

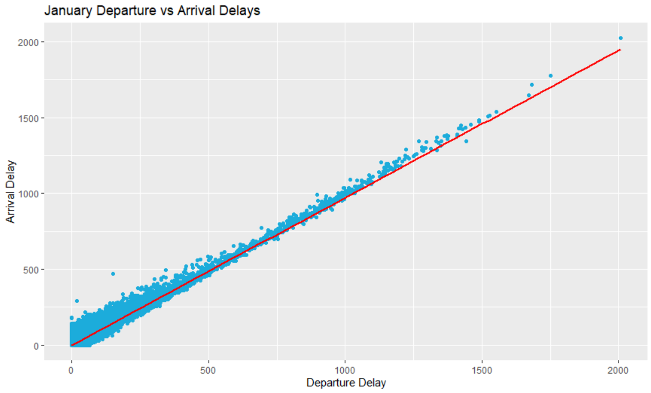


FIG. 5. Arrival vs Departure delays scatter plot for january 2018

To select just the meaningful columns, multiple plots were done. The scatter plot in the Figure 5 shows us there is a linear relationship between the departure delay and arrival delay. This means that we could consider working only with departure delays as it’s more likely to have an arrival delay if we have a departure delay.

One other point of interest was to see if the days of the week had an effect on flights delays. The box plot

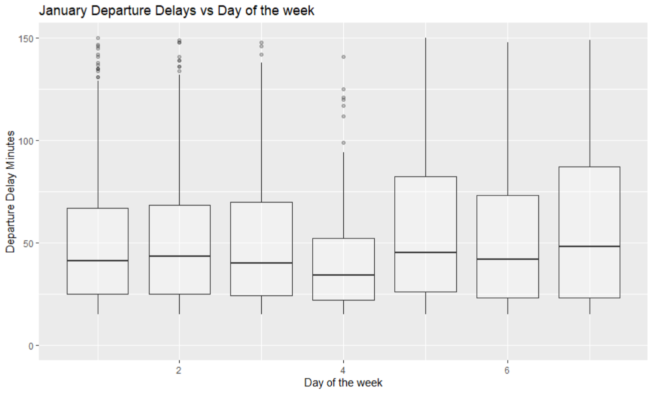


FIG. 6. Mean and variance of Departure delays over days of the week

in Figure 7 shows that although the median delay values can be comparable over the week, there is a higher range of delays in Fridays and the week-end in comparison to the rest of the week.

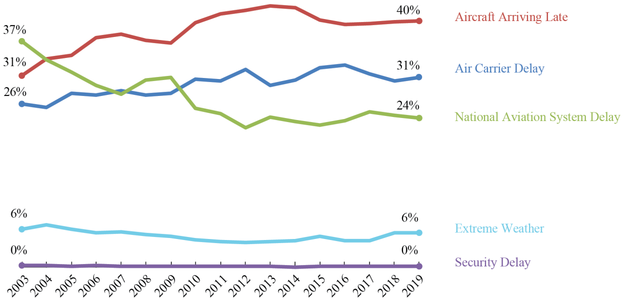


FIG. 7. Delay Cause by Year, as a Percent of Total Delay Minutes

The On-time Performance database also contains reported causes of delay by the airlines in broad categories that were created by the Air Carrier On-Time Reporting Advisory Committee. The categories are explained below:

Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

Weather delay: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.

National Aviation System (NAS) delay: Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

Late-arriving aircraft delay: A previous flight with same aircraft arrived late, causing the present flight to depart late.

Security delay: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

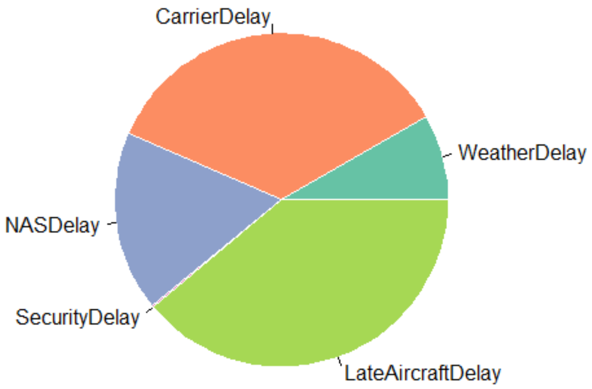
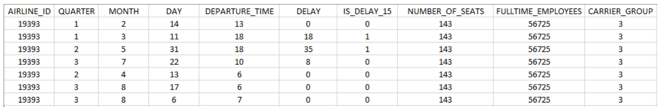


FIG. 8. Delays causes for over 15 minutes delays in 2018

But while the pie plot in Figure 8 shows that bad weather makes up less than 10% of the total delay minutes in 2018. It is worth knowing that there is another category of weather within the NAS category [5]. This type of weather slows the operations of the system but does not prevent flying. Delays or cancellations coded "NAS" are the type of weather delays that could be reduced with corrective action by the airports or the Federal Aviation Administration.

III. PREDICTIVE MODELING WITH SAS EM

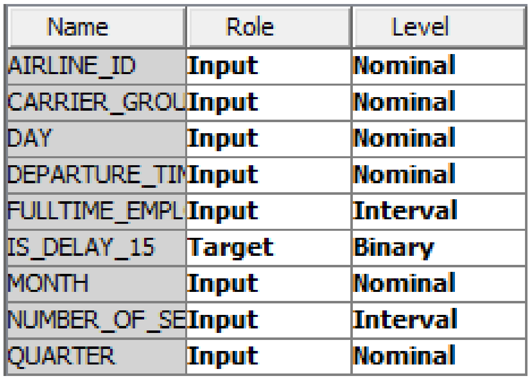
After filtering cancelled and diverted flights and joining the datasets using R language. The data from different month was grouped in one table with 140000 row. The columns kept for the predictive modeling are in the figure 9.



AIRLINE_ID	QUARTER	MONTH	DAY	DEPARTURE_TIME	DELAY	IS_DELAY_15	NUMBER_OF_SEATS	FULLTIME_EMPLOYEES	CARRIER_GROUP
19393	1	2	14	13	0	0	143	56725	3
19393	1	3	11	18	18	1	143	56725	3
19393	2	5	31	18	35	1	143	56725	3
19393	3	7	22	10	8	0	143	56725	3
19393	2	4	13	6	0	0	143	56725	3
19393	3	8	17	6	0	0	143	56725	3
19393	3	8	6	7	0	0	143	56725	3

FIG. 9. Table first 7 rows

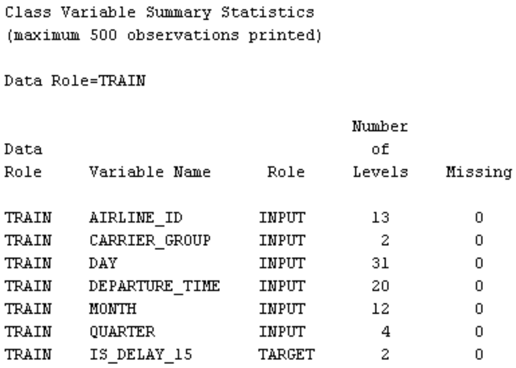
In the predictive model, IS_DELAY_15 was selected as target variable, it's a binary variable equal to 1 when the delay is over 15 minutes and to 0 otherwise. AIRLINE_ID, QUARTER, MONTH, DAY and DEPARTURE_TIME are input variables. NUMBER_OF_SEATS, FULLTIME_EMPLOYEES and CARRIER_GROUP were also added as input variables as they looked relevant to the problem.



Name	Role	Level
AIRLINE_ID	**Input**	**Nominal**
CARRIER_GROU	**Input**	**Nominal**
DAY	**Input**	**Nominal**
DEPARTURE_TI	**Input**	**Nominal**
FULLTIME_EMPL	**Input**	**Interval**
IS_DELAY_15	**Target**	**Binary**
MONTH	**Input**	**Nominal**
NUMBER_OF_SE	**Input**	**Interval**
QUARTER	**Input**	**Nominal**

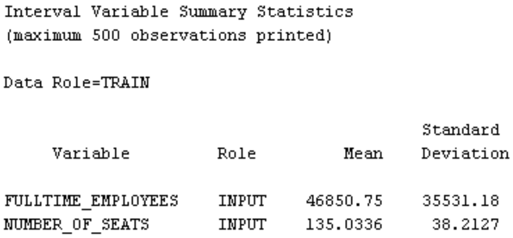
FIG. 10. Variables in file import node

A StatExplore node was used after the FileImport node to explore the inputs and check for missing values. The data preparation part was successful as we have no missing values.



```
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                      Number
Data                                    of
Role       Variable Name     Role     Levels    Missing

TRAIN      AIRLINE_ID        INPUT      13         0
TRAIN      CARRIER_GROUP     INPUT       2         0
TRAIN      DAY               INPUT      31         0
TRAIN      DEPARTURE_TIME    INPUT      20         0
TRAIN      MONTH             INPUT      12         0
TRAIN      QUARTER           INPUT       4         0
TRAIN      IS_DELAY_15       TARGET      2         0
```

FIG. 11. Nominal variables in StatExplore node

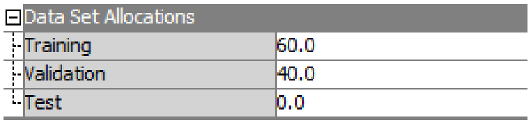


```
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                               Standard
     Variable          Role        Mean       Deviation

FULLTIME_EMPLOYEES    INPUT    46850.75     35531.18
NUMBER_OF_SEATS       INPUT    135.0336      38.2127
```

FIG. 12. Interval variables in StatExplore node

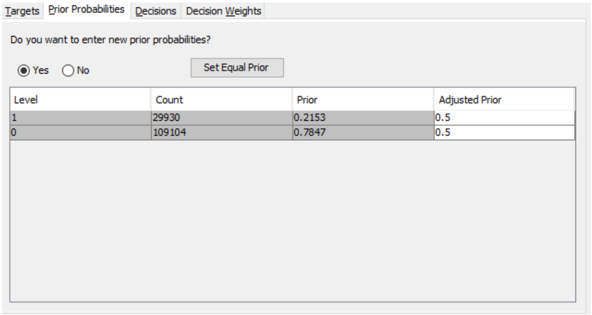
A data partition node was added after the file import and the rate of the training data was 60% and the validation was 40%.

We should node that the imported file was saved as a data source to use the prior probability windows and set the prior probabilities equal for the classes 0 and 1 of the target IS_DELAY_15.



Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0

FIG. 13. Data partition node



Do you want to enter new prior probabilities?

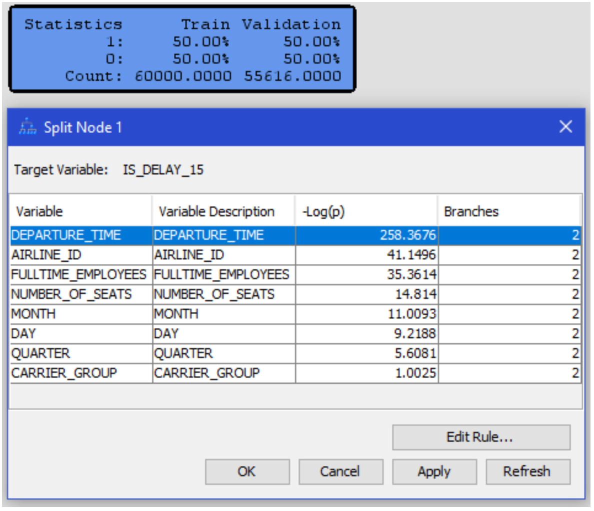
● Yes ○ No

Level	Count	Prior	Adjusted Prior
1	29930	0.2153	0.5
0	109104	0.7847	0.5

FIG. 14. Ajusting prior probabilities to equal

A. Interactive Tree

The first model in the prediction part is a Decision tree. We can already have an insight about the importance of each column in splitting the data by looking at each relative importance value or logworth while creating the first node.



```
Statistics       Train  Validation
        1:      50.00%      50.00%
        0:      50.00%      50.00%
    Count: 60000.0000  55616.0000
```

Target Variable: IS_DELAY_15

Variable	Variable Description	-Log(p)	Branches
DEPARTURE_TIME	DEPARTURE_TIME	258.3676	2
AIRLINE_ID	AIRLINE_ID	41.1496	2
FULLTIME_EMPLOYEES	FULLTIME_EMPLOYEES	35.3614	2
NUMBER_OF_SEATS	NUMBER_OF_SEATS	14.814	2
MONTH	MONTH	11.0093	2
DAY	DAY	9.2188	2
QUARTER	QUARTER	5.6081	2
CARRIER_GROUP	CARRIER_GROUP	1.0025	2

FIG. 15. Logworth of each input at the first split

The first split in Figure 17 was made over the DEPARTURE_TIME variable, where 63.88% of the flights with departure time between 2PM and 10PM were delayed at least by 15 minutes. The next split were done automati-

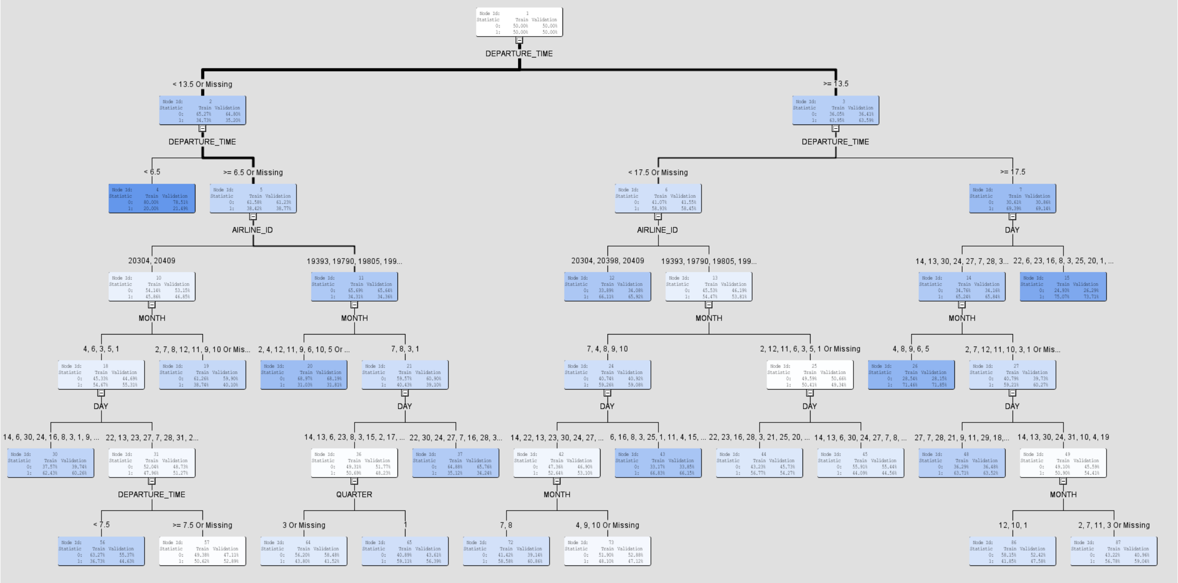


FIG. 16. Autonomous decision tree view

cally by SAS EM using the split node option on our two first leaves.

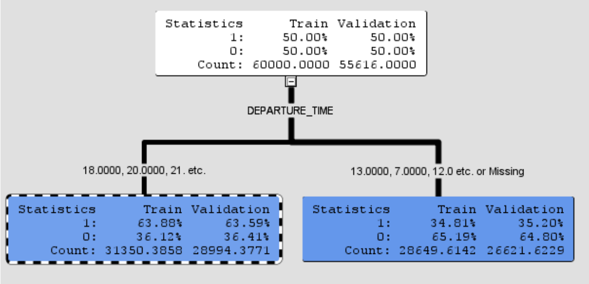

Statistics Train Validation
1: 50.00% 50.00%
0: 50.00% 50.00%
Count: 60000.0000 55616.0000
DEPARTURE_TIME
18.0000, 20.0000, 21. etc.
13.0000, 7.0000, 12.0 etc. or Missing
Statistics Train Validation
1: 63.88% 63.59%
0: 36.12% 36.41%
Count: 31350.3858 28994.3771
Statistics Train Validation
1: 34.81% 35.20%
0: 65.19% 64.80%
Count: 28649.6142 26621.6229

FIG. 17. First node in the decision tree

The Subtree Assessment Plot below shows the Average Square Error that corresponds to each tree in the sequence as the data is sequentially split. The performance of the intercative tree on the training sample becomes monotonically better as the tree becomes more complex. However, the training had to be stopped at 15 leaves before the model overfits as we can see in the figure 18.

B. Autonomous Tree

The figure 16 shows the view of the Autonoumous tree which showed a lower average square error. The Average Square Error plot of the Autonomous tree is in the Figure 19.

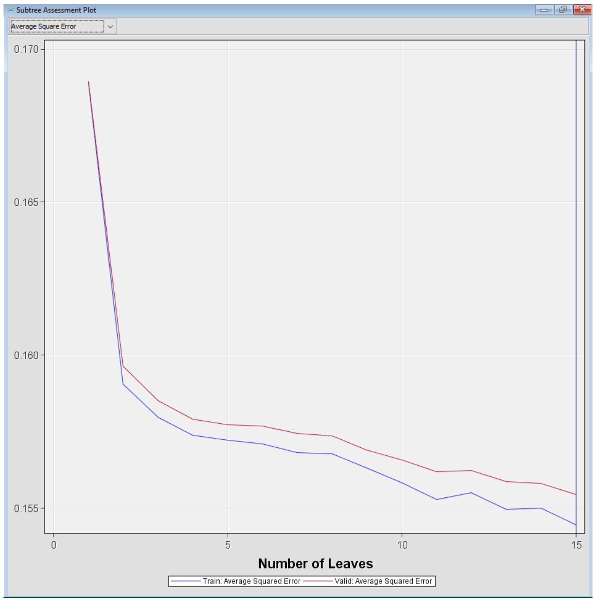


FIG. 18. Subtree Assessment Plot for the Interactive Tree

C. Regression

In the Type 3 Analysis of Effects in the regression output, we can see the number of parameters that each input contributes to the model, as well as each input's statistical significance. A value near 0 in the Pr > ChiSq column indicates a significant input. We can see that the selected inputs were all significant to the regression.

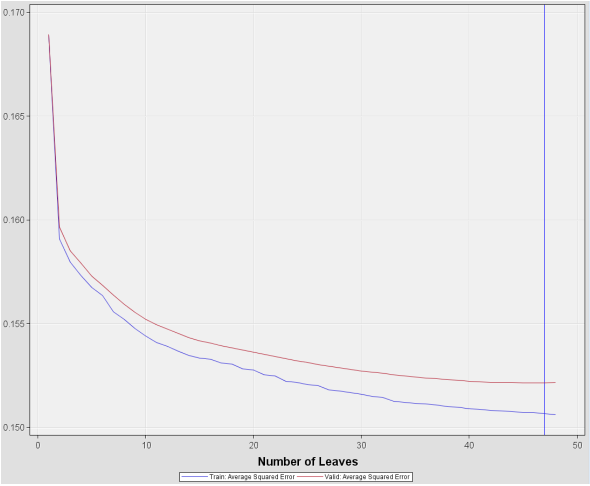
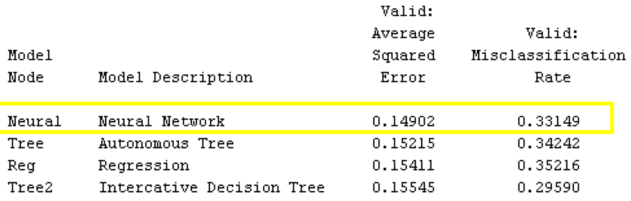


FIG. 19. Average Square Error Plot for the Autonomous Tree

D. Conclusions

So can we can you predict if a flight will be late? How confident are we in our predictions and why? To answer those questions, a model comparison node was added. After linking our four models (Interactive tree, autonomous tree, regression model and neural network) to the node and running it, the Output window lists values for a variety of statistics of fit for each model. Statistics are listed for both the training and validation data partitions. The Model Comparison tool selects the model with the smallest validation misclassification rate as the best model. The final diagram is shown in the Figure 21 .



Model Node	Model Description	Valid: Average Squared Error	Valid: Misclassification Rate
Neural	Neural Network	0.14902	0.33149
Tree	Autonomous Tree	0.15215	0.34242
Reg	Regression	0.15411	0.35216
Tree2	Intercative Decision Tree	0.15545	0.29590

FIG. 20. Models comparison

SAS Enterprise Miner has selected the **Neural Network** model as the best model. The model is 76% confident about the classifications with the lowest Misclassification Rate value 0.33 and an AVERAGE SQUARE ERROR of 14.9% in the validation data .

[1] BTS. On-time performance - flight delays at a glance , https://www.transtats.bts.gov/homedrillchart_month.asp.

[2] Bureau of Transportation Statistics. Airline on time performance data http://www.transtats.bts.gov/dl_selectfields.asp?table _id=236db_short_name=on-time.

[3] Bureau of Transportation Statistics. Carrier information https://www.transtats.bts.gov/dl_selectfields.asp?table _id=312db_short_name=air%20carrier%20financial.

[4] Bureau of Transportation Statistics. Aircraft information https://www.transtats.bts.gov/dl_selectfields.asp?table _id=314db_short_name=air%20carrier%20financial.

[5] Understanding the reporting of causes of flight delays and cancellations, https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations.

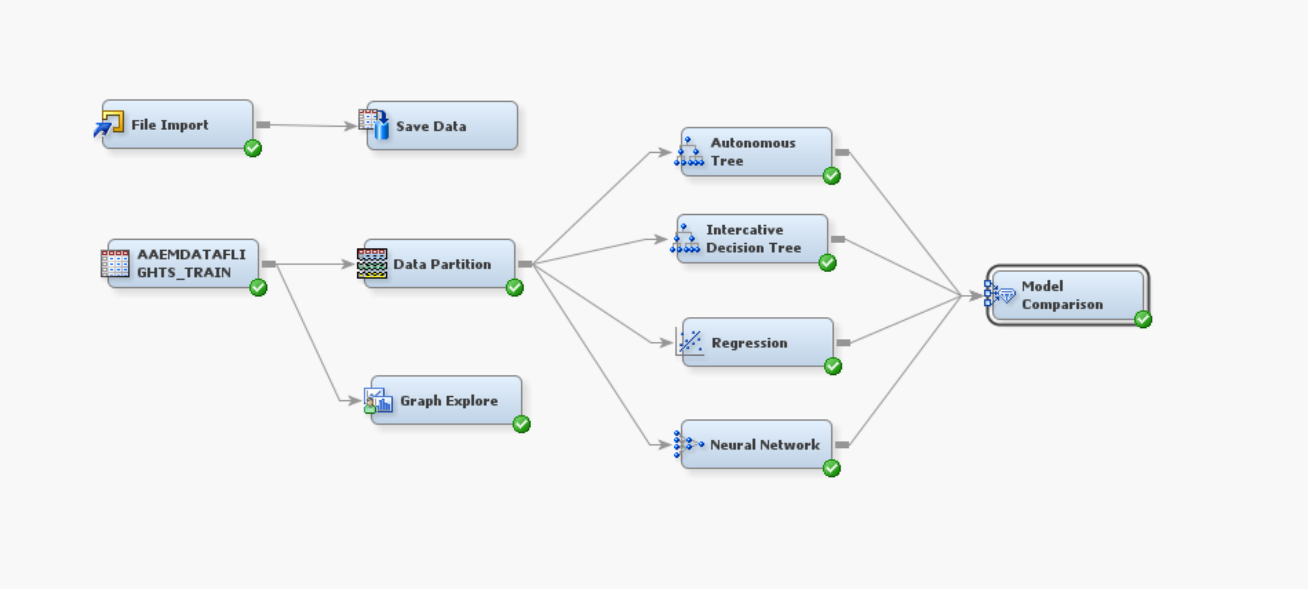


FIG. 21. Predective modeling diagram in SAS EM