Air Flight Data: Delayed New Year Flights

Andrew Tran, Chris Alipio, Fiona Tang, Jeffrey Lam, Sean Pattarateranon
Department of Computer Information Systems
California State University
Los Angeles
E-mail: spattar@caltstatela.edu

1. Abstract

Airline delays have become under increased scrutiny lately among travelers and popular press. Flight delays are caused by several attributes such as airport congestion, airspace congestion, weather and customer disturbance. We will be using Airline Delay and Cancellation Data, 2016 - 2018 Flight info. of US domestic flights. Our focus is to analyze and predict future cancellations of flights and delays during the New Year's holiday. The years we base our datasets will be between 2016-2018 with additional information implementation of machine learning. By doing this, we will demonstrate regression lines for cancellation flights and top 15 destination flights.

2. Introduction

The airline industry faces real challenges on the daily based on changes on competition, technical and customer behavior. Flight delays have always been an issue among travelers throughout the year with even more increasing delays throughout peak times during the year. Especially during the holiday seasons when millions of travelers are seeking flights to visit their family and friends which contributes to flight delays at all-time high in recent years. According to U.S Department of Transportation from data in 2017 an average flight is expected to be delayed by 15 minutes or more after its scheduled arrival. As a result of having one flight delayed will prolong other flights behind it thus causing airport or airspace congestion. Thus, airlines need to develop, identify and implement better business strategies. For our project our intention is to analyze what variables factor into flight delays. By doing so, we might be able to find out what some of the most common factors for flight delays are.

3. Related Work

In a research article by Shaowu Cheng, flight delays and their causal factors were studied using spatial analysis. It ranked these factors based on an analytical hierarchical process and found that technical failures and weather conditions were the most influential factors (Cheng et al., 2019). The study also analyzes previous research that relied on machine learning algorithms and found that they tend to show an aggregation pattern in temporal dimensions. This means that high delays were generally clustered while low delays were surrounded by low delays. As such, the study contends that the correlation between two delay values depends on spatial attributes such as location and distance. In short, the study used spatial analysis and found that weather conditions and technical failures at previous airports were the largest factor on departure delay. Our work differs in that we will be relying on machine learning algorithms.

A similar study was conducted on the investigation of flight delays. The study used a fuzzy evaluation method and found that these factors had the highest correlation with flight delays: flow of air traffic, airline issues, and weather conditions according to the degree of correlation (Ma et al.,2017). We will utilize a similar method by running a linear regression to predict values in Azure Machine Learning. We will then use a permutation feature importance module to determine which factors have the largest role in the cause of flight delays.

4. Background / Existing Work

The Airline Delay and Cancellation data is collected through www. kaggle.com which originally retrieved from Bureau of Transportation Statistics, which stores flights on-time performance. We use dataset from year 2016 to 2018 to demonstrate based on our focus and goals (Flight Delays and Cancellations). The tools we used for our analysis purpose are Microsoft Azure Machine Learning Studio, to build and deploy predictive analytics, and Kibana to visualize the data.

5. Result

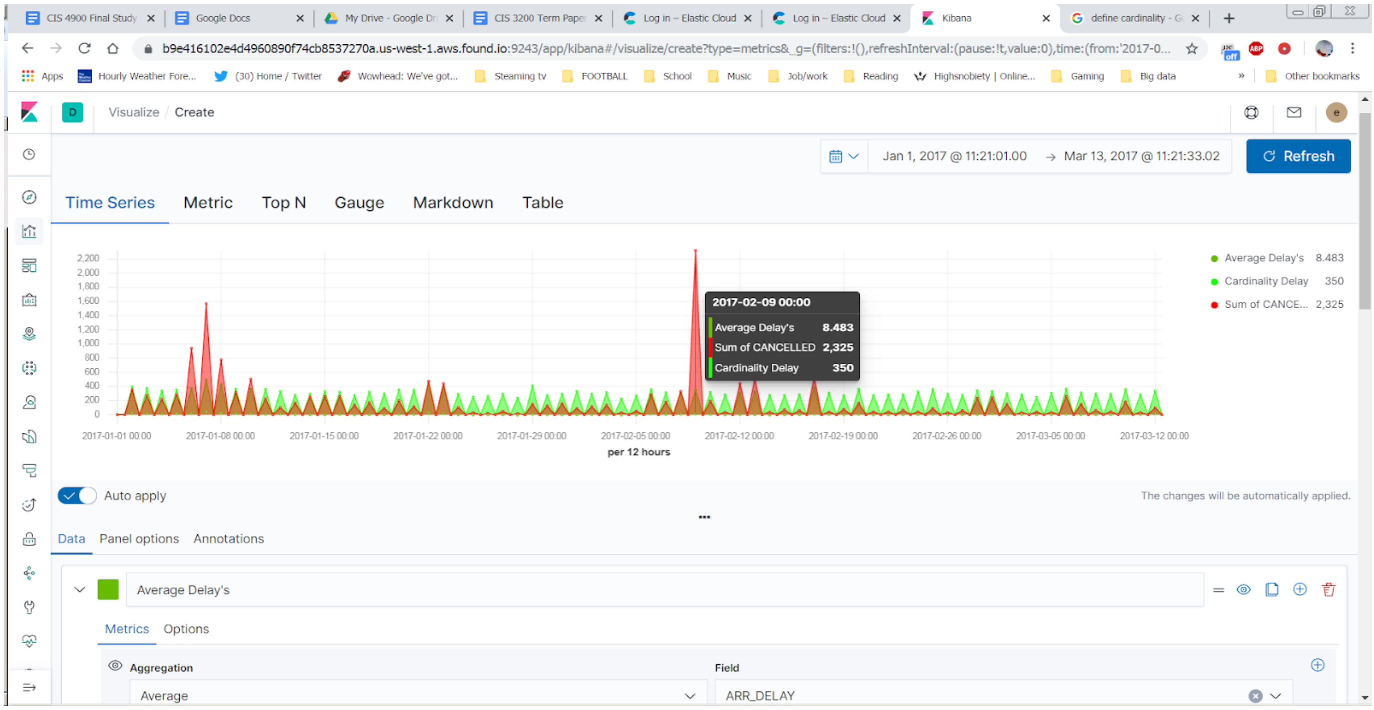
Figure 1. Data Visualization using Kibana (2017 Time Series, Cancellations - Sum of Cancellations.

Figure 2. Data Visualization using Kibana (2017 Delays - Average Delays, Departure Delays and Cardinality Delays).

Figure 3. Microsoft Azure Machine Learning (Linear Regression, 2017 Flight Delay Azure ML Experiment Diagram).

Figure 4. Microsoft Azure Machine Learning (Linear Regression, 2017 Flight Delay Permutation Feature Importance).

Figure 5. Microsoft Azure Machine Learning (Linear Regression, 2017 Flight Delay Evaluation Model results).



b9e416102e4d4960890f74cb8537270a.us-west-1.aws.found.io:9243/app/kibana#/visualize/create?type=metrics&_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:'2017-0...

Jan 1, 2017 @ 11:21:01.00 → Mar 13, 2017 @ 11:21:33.02

● Average Delay's	8.483
● Cardinality Delay	350
● Sum of CANCE...	2,325

2017-02-09 00:00

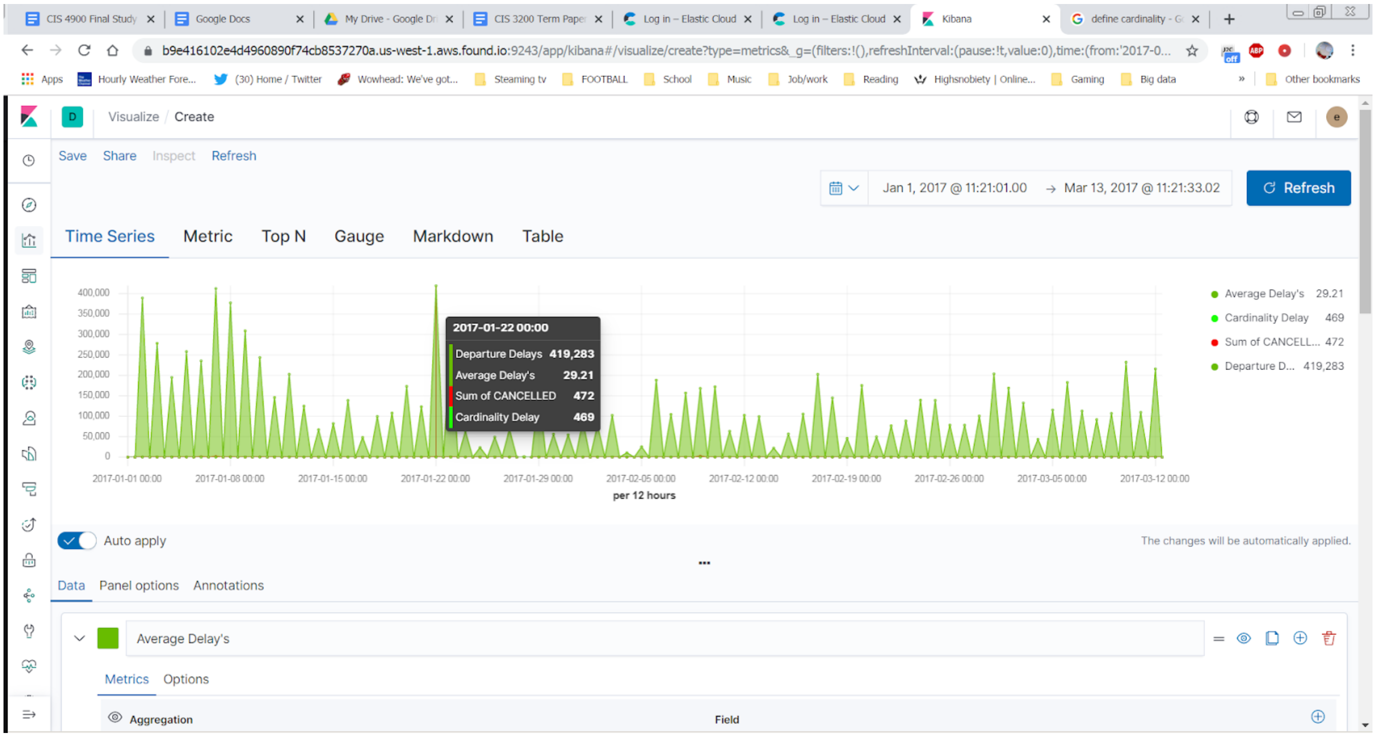
Average Delay's	8.483
Sum of CANCELLED	2,325
Cardinality Delay	350

Average Delay's

Aggregation: Average

Field: ARR_DELAY

“Figure 1. Data Visualization using Kibana (2017 Time Series, Cancellations - Sum of Cancellations).”



b9e416102e4d4960890f74cb8537270a.us-west-1.aws.found.io:9243/app/kibana#/visualize/create?type=metrics&_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:'2017-0...

Jan 1, 2017 @ 11:21:01.00 → Mar 13, 2017 @ 11:21:33.02

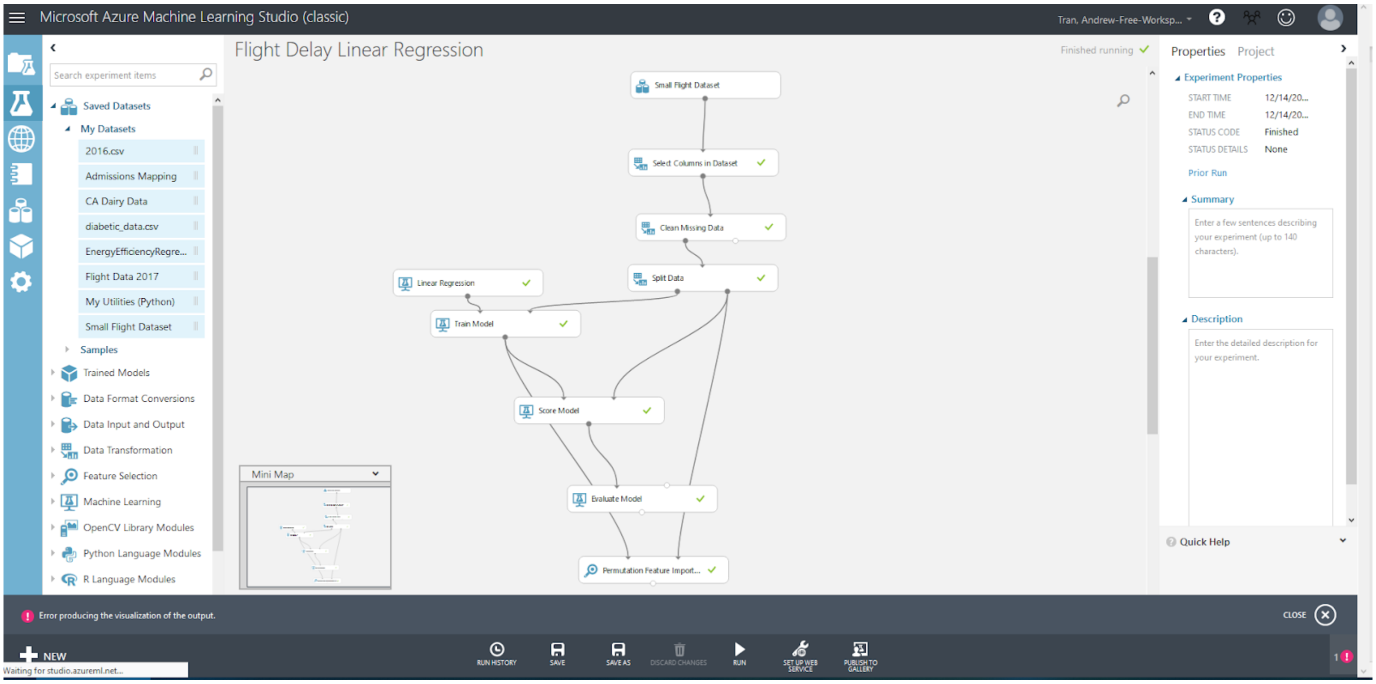
● Average Delay's	29.21
● Cardinality Delay	469
● Sum of CANCELL...	472
● Departure D...	419,283

2017-01-22 00:00

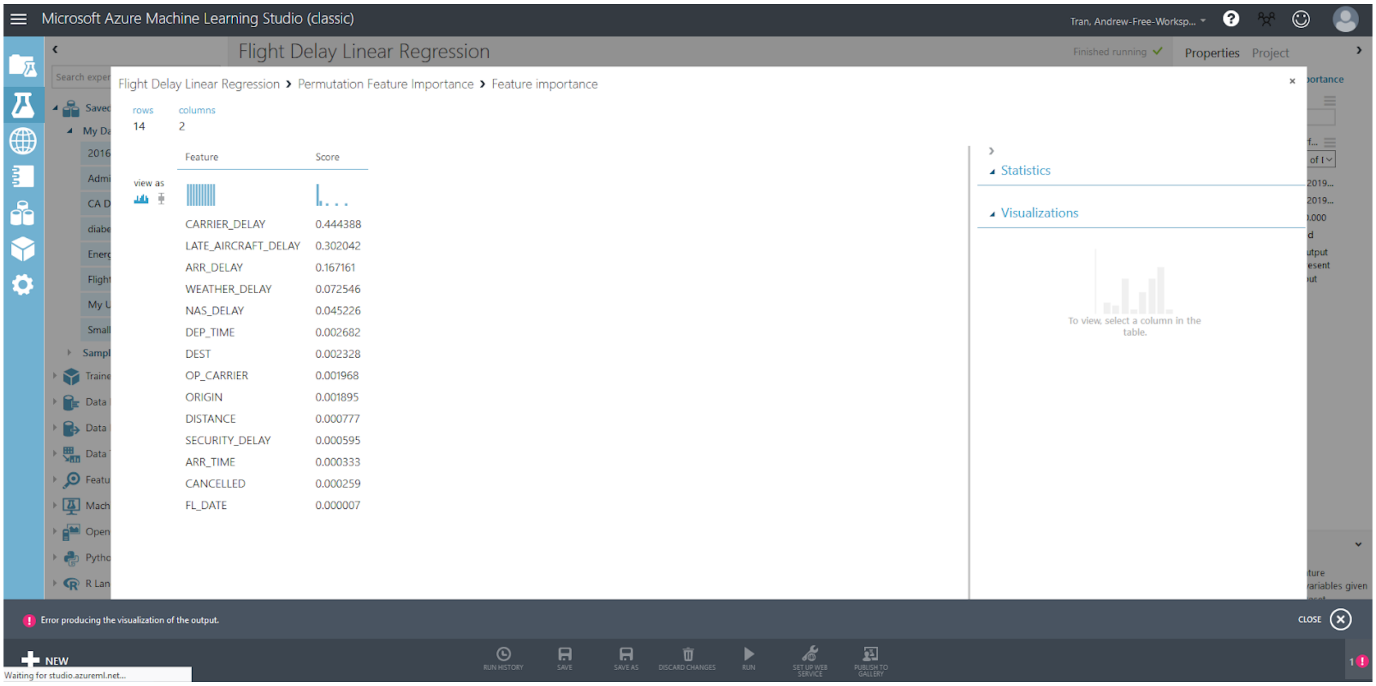
Departure Delays	419,283
Average Delay's	29.21
Sum of CANCELLED	472
Cardinality Delay	469

Average Delay's

“Figure 2. Data Visualization using Kibana (2017 Delays - Average Delays, Departure Delays and Cardinality Delays).”



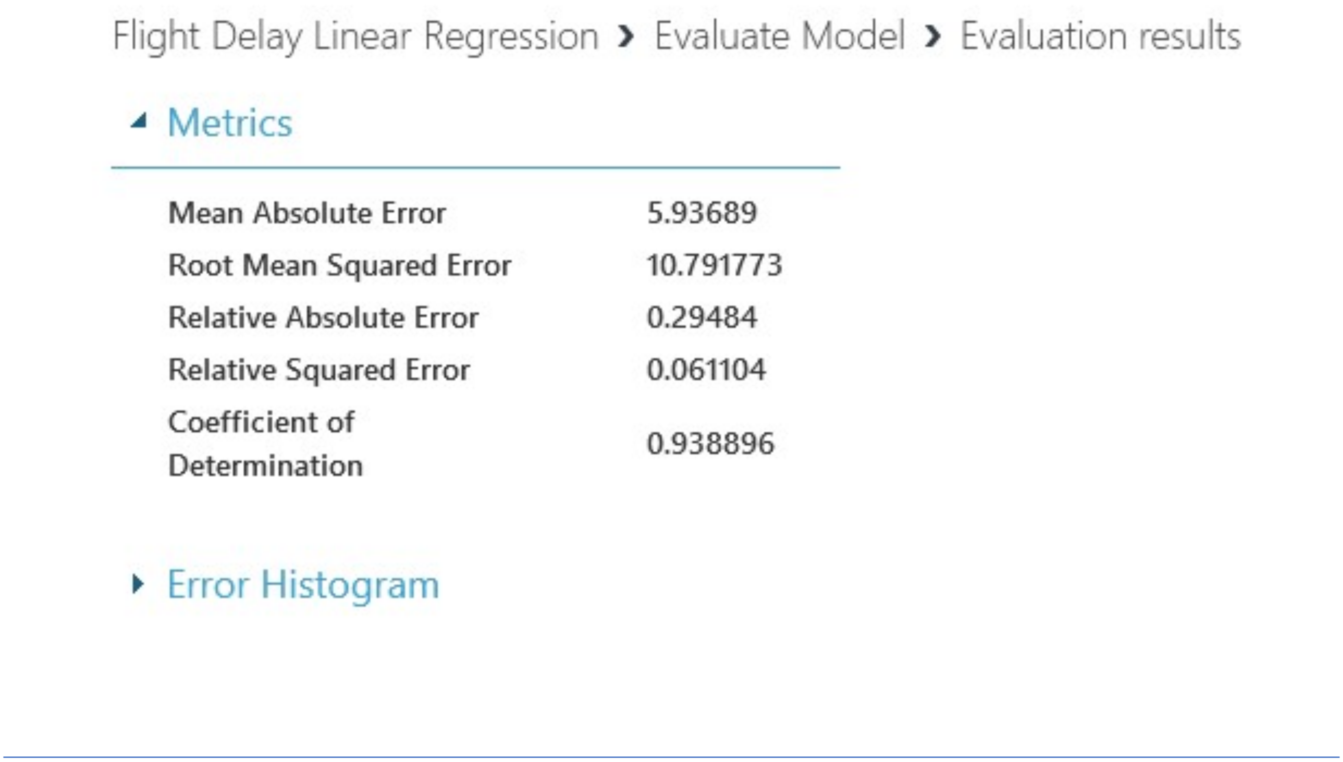
“Figure 3. Microsoft Azure Machine Learning (Linear Regression Model, 2017 Flight Delay Azure ML Experiment Diagram).”


Microsoft Azure Machine Learning Studio (classic)
Tran, Andrew-Free-Worksp...
Flight Delay Linear Regression
Finished running ✓
Properties
Project
Flight Delay Linear Regression › Permutation Feature Importance › Feature importance
view as
Statistics
Visualizations
To view, select a column in the table.
Error producing the visualization of the output.
CLOSE
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY
Waiting for studio.azureml.net...

rows 14 columns 2

Feature	Score
CARRIER_DELAY	0.444388
LATE_AIRCRAFT_DELAY	0.302042
ARR_DELAY	0.167161
WEATHER_DELAY	0.072546
NAS_DELAY	0.045226
DEP_TIME	0.002682
DEST	0.002328
OP_CARRIER	0.001968
ORIGIN	0.001895
DISTANCE	0.000777
SECURITY_DELAY	0.000595
ARR_TIME	0.000333
CANCELLED	0.000259
FL_DATE	0.000007

“Figure 4. Microsoft Azure Machine Learning (Linear Regression Model, 2017 Flight Delay Permutation Feature Importance).”



Flight Delay Linear Regression › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	5.93689
Root Mean Squared Error	10.791773
Relative Absolute Error	0.29484
Relative Squared Error	0.061104
Coefficient of Determination	0.938896

Error Histogram

“Figure 5. The result of Coefficient of Determination from Flight Delay Linear Regression model is 0.938.”

6. Conclusion

The main concentration is done on analytic platform and data visualization. We have built Linear Regression model of machine learning algorithms using Microsoft Azure Machine Learning Studio, and we have completed a process of modeling, training, testing, and evaluation of the experiment. The result data we gathered from our Flight Delay Linear Regression experiment is the most accurate.

The result of Coefficient of Determination from Linear Regression model is 0.938, which interpret that the end result of analytic from dataset is 94% predictable. As a result, Linear Regression model yield the best result at 94%, which is accurate.

This project has helped our team a better understanding of Microsoft Azure Machine Learning and the analytic process which is completely a new topic for our team.

7. References

7.1 Research articles

[1] Cheng, S., Zhang, Y., Hao, S., Liu, R., Luo, X., & Luo, Q. (2019). Study of Flight Departure Delay and Causal Factor Using Spatial Analysis. *Journal of Advanced Transportation*, *2019*, 1–11. doi: 10.1155/2019/3525912

[2] Ma, S., Wang, X., & Hu, H. (2017). Research on Flight Delay Based on Fuzzy Evaluation Algorithm. *Journal of Applied Mathematics and Physics*, *05*(10), 1923–1937. doi: 10.4236/jamp.2017.510163

7.2 Dataset link

[1] https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

7.3 GitHub link

[1] https://github.com/spatta983/CIS3200-Group4

7.4 Microsoft Azure Machine Learning link

[1] https://gallery.cortanaintelligence.com/Experiment/Flight-Delay-Linear-Regression