Predicting Aircraft Delays Using Weather Data

Ceara Stewart, Cristobal Gavaldon, Kristan Rodriguez

Syracuse University

IST718: Big Data Analytics

Introduction

Weather has always influenced the ability for aircraft to be able to take off, fly and land. Low clouds, fog, rain, and snow can make it a challenge visually. High winds can determine if an aircraft will be able to take off and land. And thunderstorms and lightning can cause downdrafts, creating turbulence and making an aircraft challenging to fly. Weather, such as what was just listed, is a big reason that flights are delayed. This report aims to predict whether a flight will be delayed based on weather.

# Data Sets Used

Multiple datasets were used for this report. Five years of airport data were loaded into Python from the Bureau of Transportation Statistics. The data is from January 2019 to November of 2023, and each year was loaded into Python as its own dataset. A dataset consisting of the latitude and longitude of each airport within the US was also used. This dataset came from the OCHA Services HDX website. Weather data was pulled from the National Centers for Environmental Information, with monthly average precipitation and temperature values for each state being used. Throughout the development of the Python code, the datasets were joined together so that each airport was matched with its latitude and longitude, and each state listed its monthly average precipitation temperature.

## Data Cleaning and Data Manipulation

All seven of the data sets used included unnecessary and confusing information so the following cleaning/manipulation measures were implemented: removing columns, renaming columns, handling missing values, merging data sets, deriving a datetime columns. Removing and renaming columns are necessary to improve efficiency and clarity of data analysis and modeling. In addition, many of the functions/models used for this analysis cannot process missing values so they needed to be removed beforehand. Also, for all the dates between the data sets to match, datetime columns were derived from each to avoid inconsistencies when merging the data.

### Descriptive Analysis and General Modelling and Graphics

Descriptive statistics were run on each yearly airport data set. The descriptive statistics calculate the count, mean, standard deviation, and quartile range data for each column within the datasets.

The average weather delay for each airport was also calculated and compared across the five years. A map outlining the weather delay by the airport was then mapped for each year as well. During this process, it was found that Chicago based airports had the highest average weather delay value for 2019, in 2020 the Orlando based airports had the highest average weather delay value, for 2021 there were a multitude of airports with high average weather delays with the highest average being from the Colorado based airports, 2020 also saw a multitude of airports having high average weather delays with the highest average being from the Orlando based airports, and for 2023 Colorado based airports saw the highest average weather delay. This falls in line with national weather patterns as areas such as Colorado and Chicago see a large amount of precipitation, while Florida is plagued by hurricane season.

We also printed the airplane carriers and their count of weather delay occurrences for each year. Through this process we found that SkyWest Airlines Inc has the highest count of weather delays for each of the five years of data. When researching SkyWest, it was found that their main airport hubs revolve around Denver, CO, Chicago, IL, and Minneapolis, MN, which were found to be cities hit with heavy amounts of precipitation especially during the winter months.

Forecasting and Subsequent Analysis

Forecasting is the process of making predictions or estimates about future events or outcomes based on past and present data. It involves analyzing historical data patterns, trends, and relationships to make educated guesses or projections about what may happen in the future.

**Regression**

Regression analysis is a statistical method used to study the relationship between a dependent variable and one or more independent variables. It can be used to make predictions and understand the strength and nature of the relationship between variables. In a regression model, the dependent variable is the variable being predicated while the independent variable(s) are used to make the prediction.

To run our regression model, we first decided to create a variable that was the binary version of the weather\_delay variable. This binary variable assigned a 0 if a flight was not delayed by weather, and a 1 if it was delayed by weather. We chose to create this variable as we are trying to predict whether a flight is or is not delayed by weather, not the value of the weather delay itself. This binary variable was also used for the other forecasting models.

We used this binary weather delay variable as our dependent variable, and we selected year, month, latitude, longitude, elevation, precipitation and temperature as our independent variables. We first found the mean squared error of the model, or the average squared difference between the value observed and the predicted value for the model. The smaller the mean squared error value the better as this means the predicted value is close to the observed value. For the regression model ran, we obtained a MSE of 0.247 which is ideal as it is close to zero and tells us that predicted values produced by the model are close to any observed values.

We then added a constant term and fit the linear regression and printed the summary report produced. We found that year, month, latitude, longitude, elevation and precipitation were all statistically significant variables as their p-values were less than our alpha of 0.05. The elevation and longitude variables were not statistically significant and had a p-value greater than 0.05. This left us with the below equation:

*Weather\_Delay\_Binary = -30.2604 + 0.0153\*year – 0.0048\*month – 0.0040\*latitude – 0.0009\*precipitation – 0.0008\*temperature*

By inserting values into the beta variables within the equation, you can predict whether a flight delay due to weather occur. An example would be if we entered weather data from January 2024 for Chicago O’Hare International Airport. This would leave us with the below equation:

*Weather\_Delay\_Binary = -30.2604 + 0.0153\*2024 –0.0048\*01 – 0.0040 \*41.9786 – 0.0009 \*23.4 - 0.0008\*54.1*

This equation produces the value of 0.47 as the predicted weather delay for Chicago O’Hare International Airport and in terms of whether a flight would delay at O’Hare due to weather would be a yes.

Although we were able to predict if a delay due to weather would occur, we received a very small R-squared value for our model of 0.006. This suggests our model is not a good fit for the variables. But it is known that R-squared can be inflated when using complex models or when overfitting occurs, which is very possible given that our MSE was so low.

**Decision Trees**

Decision trees are a type of supervised machine learning algorithm that is used for both classification and regression tasks. It is a powerful model that learns simple decision rules from the data, represented in the form of a tree structure. A decision tree consists of nodes, branches, and leaves. Each node represents a decision based on an attribute of the data, each branch represents the outcome of the decision, and each leaf node represents the predicted outcome. At each node of the tree, the dataset is split into subsets based on the value of a feature. Once the tree is constructed, decision rules can be extracted from the tree structure and these rules provide insights onto how the model makes predictions.

To construct our decision tree, we selected the same dependent and independent variables used in the regression analysis with the binary weather delay variable as our dependent variable, and we selected year, month, latitude, longitude, elevation, precipitation and temperature as our independent variables. The data was then split into training and testing sets where the model was fitted to the training data for the dependent and independent variables. Then that model was used to predict the dependent test variable based on the test independent variables. This resulted in a mean squared error of 34.617, which is higher than what we got for the regression analysis leading to a conclusion that the predicted values are not as close to the observed values.

We then produced a decision tree from the fitted models where each node split subsequently leads to being able to predict if a flight will be delayed due to weather. An example would be if the month is less than or equal to August, you would move onto the next node where if the month is less than or equal to November you would move onto the node where year is less than or equal to 2021 and so on and so forth till if all of that criteria follows the nodes, the predicted flight would be delayed.

There are shortcomings to decision trees that could affect the predicted values. With more complex data, such as what we have, decision trees are prone to overfitting and can capture noise in the training data, leading to poor generalization performance on unseen data. Decision trees also have difficulty with capturing relationships between features thus making them less effective at capturing complex relationships between features. But despite these shortcomings, the node splits and selected features produced by the decision tree makes sense. Typically, winter months produce more precipitation that has a greater chance of delaying a flight, and with global warming, weather has become a lot more unpredictable in more recent years.

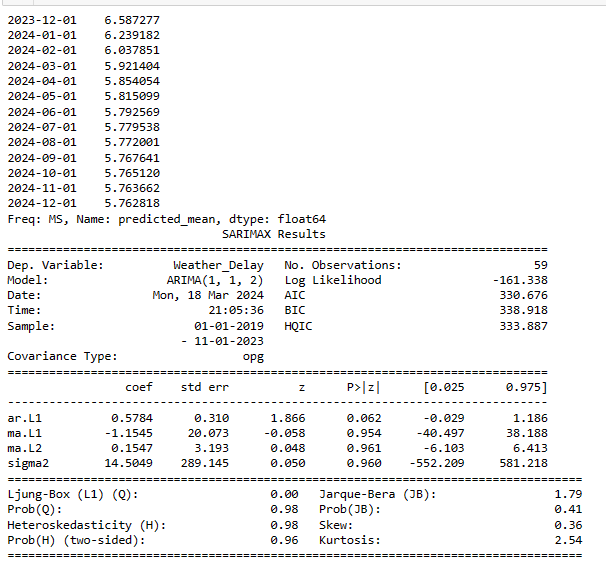
**ARIMA**

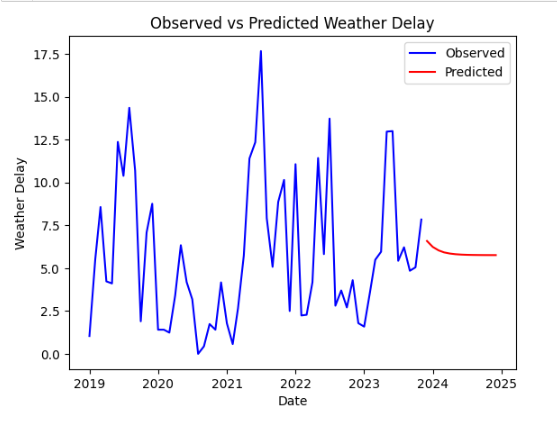
ARIMA stands for AutoRegressive Integrated Moving Average. It is a time series forecasting method that combines autoregression (AR), differencing (I), and moving average (MA) components into a single model. AutoRegressive (AR)represents the linear relationship between an observation and a certain number of lagged observations. It models the relationship between an observation and a linear combination of its previous values. Integrated (I) represents the differencing of the time series data to make it stationary. It involves subtracting the previous observation from the current observation to eliminate trends or non-stationarity in the data. Moving Average (MA) represents the linear relationship between an observation and a stochastic term based on the errors from a moving average model applied to lagged observations. It models the relationship between an observation and the residual errors of a moving average model. ARIMA models are particularly useful for capturing and modeling complex temporal patterns, trends, and seasonality in data.

To begin our ARIMA journey, we first conducted an ADF test for stationarity. We obtained a p-value of 0.0246 which is less than the alpha threshold of 0.05 meaning our data was stationary and did not need to be differenced. We then plotted autocorrelation and partial autocorrelation functions along with running an auto-ARIMA to determine our best model order. We determined from these tests that an order of (1,1,2) was most appropriate for our data.

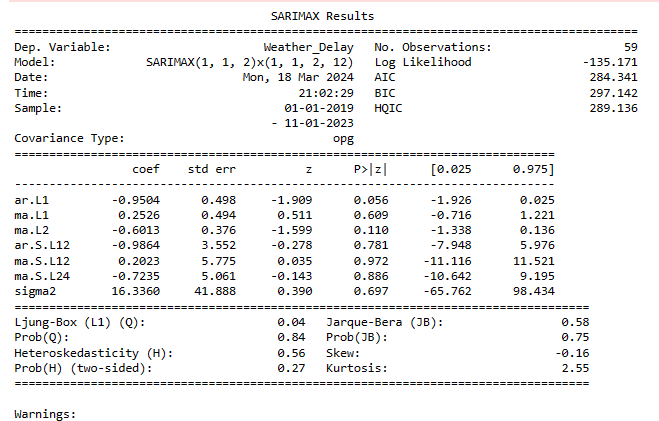
Since we had an expansive data set and wanted to ensure our dates were following a monthly frequency so that the best ARIMA model would be produced, we decided to only run our ARIMA models only on American Airlines flights at the Denver International Airport in Colorado. We ran a total of two ARIMA models, one without accounting for seasonality and then a seasonal ‘SARIMAX’ model.

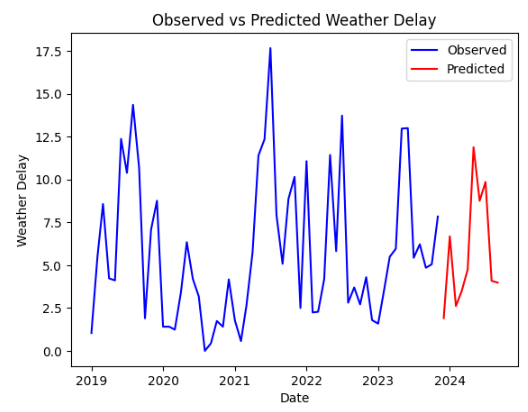
Our ARIMA model produced predictions for December 2023 to December 2024 data shown below along with the observed and predicted values for the weather delay variable plotted. We received a Ljung-Box statistic of 0.0, where anything below 0.05 indicates that the residuals are not independently distributed, suggesting that there is still information left in the residuals that the model did not capture.





When running the SARIMAX, which adds a seasonal component to the order, we obtained the summary and plot below. This model produced a Ljung-Box statistic of 0.04, which is still lower than 0.05 but is much closer than the first ARIMA model. This leads us to conclude that the SARIMAX model was a better fit for the data and better accounted for the residuals. Our plot also produces predicted values that much better align for what one would expect to see based on monthly weather changes and its effects on flight delays due to weather.





Conclusion

We set out to try and predict flight weather delays. After extensive data cleaning and manipulation, we were able to run three predictive models, a regression model, a decision tree model, and two ARIMA models. All models were able to predict flight weather delays but not all of them were good fits for the data. Our regression model did allow us to predict weather delays for flights, but the model had a low R-squared and our low MSE suggests that the data was overfitted. Our decision tree model produced a higher MSE than the regression model, but the node splits and selected features do make logical sense for predicting what values predict a weather delay. Our initial ARIMA produced an extremely low Ljung-Box statistic while our seasonal ARIMA did produce a higher Ljung-Box statistic but still not something above 0.05.

If we continued this project into the future, we would like to produce a regression model with not only a low MSE but a high R-squared. We would also like to tune the decision tree model to see if we can obtain a smaller MSE. And we would like to run the ARIMA model on all data and not just a subset as well as work to increase the Ljung-Box statistic to 0.05 or higher. But at the end of the day, we were able to predict if an airline flight would be delayed based on the weather.

Websites Used for Datasets

Airport Data:

<https://data.humdata.org/dataset/ourairports-usa>

Delay Data:

<https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp?qv52ynB=pun46&20=E>

Weather Data:

<https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series/7/tavg/12/0/2019-2024?base_prd=true&begbaseyear=2019&endbaseyear=2023>