Prateek Mechineni

Flight Data Final Project

**Introduction:**

For this project, I decided to analyze the Flight data from the past year from Kaggle (<https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay>). The dataset is a CSV file that contains 539383 instances and 8 different features:

Text

Description automatically generated with low confidence

Metadata:

* ID: Just list of items
* Airline: Type of Arline
* Flight: Flight Number
* AirportFrom: 3 letter airport code
* AirportTo: 3 letter airport code
* DayofWeek: Number to represent day of the week
* Time: Time of flight (Add number of min in column to Midnight of that day to get time. Ex. 15 means 12:00am + 15min = 12:15am flight time.
* Length: Time of flight in Min
* Delay: 1 for delay, 0 for no Delay.

**Business Questions:**

For this project, my first objective was to perform some exploratory analysis on the data and see what I can gather and determine. I wanted to then use my R skills of using machine learning to see if I can create a ML model that can predict if a flight can be delayed or not.

Key Questions:

* What airline has the most delays?
* Which airport has the most delays?
* Which day has the most delays?
* Which airports are best to fly in and out of?
* Does flight distance affect delays?
* Does flight time affect delays?
* Can I create an ML model to help predict if flights are going to be delayed or not?

**Data Cleansing and Preparation:**

For this project, I first performed some general analysis on the dataset and decided to split my dataset into two data frames with one that contained all the numerical columns and one that contained all the object columns. This allowed me to create graphs with Seaborn and Matplot and answer some of the key questions more easily.

To create the ML model, I had to split the data into a training set and a test set since ML models need a training set to help accurately predict models. I then used Python’s Scikit-Learn library to import ML models and tools to create and test the various types of models. I used Oridinal Encoding to process the data. I did a test run, a Decision Tree model, a Gaussian model, and a KNeighbor model. I chose these since they are well incorporated into the SciKit-learn library and each one is a slightly different model.

**Data Analysis:**

To answer my key questions, I transformed the data into graphs to give a visual analysis.

1. What airline has the most delays?

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

1. Which airport has the most delays?

Chart, histogram

Description automatically generated

1. Which day has the most delays?

Chart, histogram

Description automatically generated

1. Which airports are best to fly in and out of?

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. Does flight distance affect delays?

Chart, histogram

Description automatically generated

1. Does flight time affect delays?

Chart, histogram

Description automatically generated

ML Analysis:

Test Run with Dummy Classifier:

Text

Description automatically generated

Decision Tree:

Text

Description automatically generated

GaussianNB:

Text

Description automatically generated

KNeighbors:

Text

Description automatically generated

**Conclusions:**

Overall, I considered this a very successful project. Looking at the business questions I wanted to answer, I was able to see that WN (Southwest) and DL (Delta) were the airlines with the most delays. Atlanta airport showed to be the airport with the most delayed flights. For days of the week, Wednesday and Thursday had the greatest number of delays but the differences from the other days of the week were not drastic. Once again Atlanta airport is shown to be the worst flight to fly in and out of.

Interestingly, I found that flight length doesn’t seem to be influencing delays but flight time is a major indicator of flight delays. As time goes on later in the day, the more delays there are. I am assuming that during the start of the day that there are fewer issues in the airport to cause delays. Once a delay occurs, I think it creates a snowball effect that pushes other flights into delays as well.

For the ML analysis, the dummy model came out with about 50.5% accuracy which is what we are expecting for our baseline models. In the end, the KNeighbors Classifier was able to give the best accuracy but it was only around 61%. Overall, I think I needed more data columns such as weather or the reason why the flight was canceled and then incorporated that into my ML models to get better accuracy and precision scores.