Group Member: Hengchuan Zhang, Robin Sah, Li Guan

Project Proposal: Big Data Analytics for Predicting Airline Delays

Delays are common on all types of transport vehicles. However, airline delays differ from other delays due to differences in the geographical distance traveled.

In this project, our team will be focusing on predicting airline delays with the use of big data. We train our model based on past airline data and use airline API or web scraping technique to get the real time data to finish our project. The final outcome should be deployed on a website and tell users which airline is likely to experience delay in the near future.

The data set that we use to train our model has 28 columns. We will be the date, airline company name, origin airport, destination airport, planned departure time, actual departure time, total delay before departure, the time that airline actual leave the ground, the time that airline actual touch the ground, planned arrival time, actual arrival time, total delay on arrival, canceled, airline travel time, distance, carrier delay, weather delay, air system delay and security delay to finalize a machine learning model. Ideally, the final model could give a percentage of delay with those information above.

There are two main challenges for this project. One is the data collection. The real time data mainly has two sources to get. First one is the airline API. However, the free tier API only has a limited number of requests in total, which makes it hard for users to get enough data. Second way to get real time data is web scraping. However, web scraping is not good at getting streaming data. We may need to collect those data in batch and update those data in a period of time. Second challenge is the machine learning part. None of us in our group has experience in machine learning. Researching and learning how to use this technique will be stressful to us.