**Is your flight going to be delayed?**

**Proposal**

Flight delay is one of the greatest inconvenience in everyone’s life. In the U.S., there are on average 3,500 flights delayed every day. Over 20% flights are not on time or canceled. It disrupts our family travel plans, scheduled business meetings and causes people stuck at airport for a very long time. The motivation of this project is to deliver an application for users to find out the probability of their flight’s status before they book the flight, so that they can make a choice of their flight accordingly.

I build my prediction model using 30 years of historical airline traffic data in the U.S., which includes flights, airlines, and destinations. The final dataset contains more than 100 million entries (Due to the time limitation, I only used the lastest 5 years data). The database is scrapped from <http://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time>

I first ran sanity check on the data. For example, I removed entries with missing values and flights which were cancelled or diverted, and impute the delay cause data as 0 if there is no delay. Obvious mistakes of the data are also checked and deleted. Second, I performed exploratory analysis. In particular, I look at the information of flight delay length, destination, airline information, and departure time. In this step, several features are constructed based on the original features provided by the database. Through the exploratory analysis, I obtained the following interesting findings.

1. Low cost airlines have less flight delay frequency than the legacy carrier.
2. Certain airlines are for sure will be delayed at certain city.
3. Smaller hubs have higher delay frequency.
4. It is better to avoid flights in the morning.

To continue, I plan to include more features to model and predict flight performance. Some available and interesting features in this dataset are flight time (day of the month, day of the week, hour), carrier, origins and destinations, delayed length in minutes, duration, distance, wheels time, taxi time, and causes category of the delay. In addition, weather is a very common factor that causes flight delay. For example, weather leads to more than half of flight delays in national airport during the past year (see <http://www.transtats.bts.gov/OT_Delay/ot_delaycause1.asp?type=3&pn=1>).  Therefore, I plan to acquire the daily/hourly weather data (of the same 30-year span) by scraping API provided by <https://developer.weathersource.com/documentation/>.

Then, I plan to perform feature reduction before model fitting. Linear regression or principle component analysis can be considered and compared during feature reduction. Next, I will implement machine learning on fitting the model with cross validation. Since this is a classification problem, I plan to compare logistic regression, support vector machine, random forest, and gradient boosting machines. Moreover, I also plan to model the subgroup, which might share more similar features and improve model prediction. For clustering subgroups, I plan to use unsupervised learning methods, such as K-means clustering.

To visualize the result, I plan to build an App which provides the prediction of flight status for users. Users can input the flight number and time, the App will provide the probability of this flight delayed. In the meanwhile, the App will also recommend flights which meet users’ need and will less likely be delayed.

Let’s not wait in the airport any longer!