**Proposal for Capstone Project**

**Title:** Predicting the likelihood of a US domestic flight cancellation

**Problem:** Imagine you have a trip coming in next few days and someone tells you that “*your flight has high chance of being canceled, so be aware of that and rethink about your travel and hotel bookings etc.*”. That would be helpful for you and all other passengers traveling out there. Even though the flight cancellation rate is not high (about 2-3% in US domestic market), that one rare event causes a lot of troubles to passengers in terms of rescheduling their travel plans. There are many factors such as flight date and time, origin and destination airport location, airline type, weather etc. which might affect the cancellation rate. We propose to use data from various sources containing these factors and other factors too, and build a machine learning model for predicting the likelihood of flight cancellation for U.S. domestic flights operating at selected airports.

**Who might care?** Travel planner and booking companies such as booking.com, expedia.com, kayak.com, priceline.com, etc. can use such a model to predict the likelihood of the cancellation of a flight. They can then inform their customers well in advance, even before the airlines' management informs the passengers, about the probability of the cancellation of their upcoming flight. From the travelers point of view, it would be very convenient for them. On the other hand, such a predictive model would enhance the product base of travel planner companies. Moreover, there is a possibility of developing an app which travelers can use to know about their flight cancellation likelihood in advance.

**Data:** The flight data will be acquired from the Bureau of Transportation Statistics (https://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID=236&DB\_Short\_Name=On-Time). This data contains information about each flight and their on-time performance, including date of flight, carrier type, origin and destination airports, flight distance, delay, cancellation, diversion etc. In this project, we will mainly focus on the data collected during 2015-2016, and on some selected airports only. Often the flight cancellations are caused due to bad weather conditions. Therefore, it is also important to have the hourly weather data which will be accessed through wunderground.com API (https://www.wunderground.com/weather/api), for example. We will acquire the weather data for those selected airport locations and only for years 2015 and 2016.

**Modeling approach:** Since we want to predict the likelihood of a flight getting cancelled, a supervised classification algorithm is a perfect choice to build the predictive model. The classification algorithm not only classifies classes (in our case, two classes: “cancelled” and “not cancelled”) but also predict the probability of each class. This data set contains imbalanced classes (only about 2-3% of the data will have "canceled" flag in them), which makes the project challenging. We plan to use different approaches to tackle the imbalanced class issue and choose the best one. Also, we will try various classification algorithms and pick the one which performs best. To evaluate the model’s performance, we will split the data in two parts: (i) the first part will contain only 2015 data, this will be the training data set, and (ii) the second part will contain only 2016 data, this will be used to evaluate how well the trained model generalizes to unseen data.

**Possible limitations:** The prediction of the model will depend on how good the prediction of the weather is, say after 3 days. In other words. if we want to predict the likelihood of the flight cancellation for a flight which is scheduled after 3 days, we would need to know the weather after 3 days (which we do not know "accurately" today). This means that the model will predict better if the weather prediction is better or if we are trying to predict the flights not far ago than the scheduled departure.

**Deliverables:**

1. Codes (notebooks) for:
   1. data acquisition
   2. data cleaning
   3. data exploration analysis
   4. machine learning model development
2. Report on the capstone project
3. Presentation on the capstone project
4. Cleaned data sets