CAPP 30123 Project Proposal

Group Name: Hide on Bush

Group Member:

- *Dongcheng Yang*
- *Siyuan Peng*
- *Yaling Xu*
- *Zeyu Xu*

Data:

- Size: 9.6 GB (300-700MB per year, altogether 20 years)
- Source: Flight Delays and Cancellation Dataset from the Bureau of Transportation Statistics
- Time Period: 1989 – 2008
- Record Description: Approximately 726,919,200 lines of records, 30 variables in total
- Link: http://stat-computing.org/dataexpo/2009/the-data.html

Hypotheses:

The goal of this project is providing rational travel suggestions based on the exhaustive information of each flight in Flight Delays and Cancellation Dataset. We believe that the expected delay time is highly related to the departure month, departure time, distance, airline companies, latitude and longitude of origin and destination. Specifically, the following research questions will be studied:

(1) We will analyze the impact of different departure month and time on the expected delay time.

(2) We plan to combine the flight data with geographic data to get the latitude and longitude of origin and destination airports for each flight, then explore the relationship between geographic characters and expected delay time.

(3) We will examine more in details the distribution of delays for every airline. For example, the proportion of flights that delay, the trend of delay time, delays at arrival and at departure, etc. Thus it is possible to establish a ranking of the companies.

(4) For advanced analysis, we conduct analysis of whether there is any relationship between flight delays and the financial situation of an airline is provided. We will add cross terms of airlines

and time or geographic characters to put forward further suggestions. Moreover, our goal is to create a model that predicts delays.

Algorithm:

- Heatmap: Ideally, we will build a dynamic heatmap reflecting the delays covering most of the airports in the U.S. The heatmap will reflect both the number of delays and the mean of the delay time. Thus, we can deduce from these observations in average delays, both between the different airports but also between the different airlines. It will be necessary to adopt statistical inference that is specific to the company and the airport.
- Cross-validation: A robust method called cross-validation will be used to build a prediction model. This method consists of performing a first separation of the data in *training* and *test* sets. As always, learning is done on the training set, but to avoid over-learning, it is split into several pieces that are used alternately for training and testing.
- Logistic regression: Logistic regression will be used to identify the association between the probability of delay and financial factors of an airline. Specifically, the revenue growth, current ratio reduce, leverage, airline size, and operating revenue per employee increase will influence the likelihood of a flight being delayed.