Junwei Zhuge (juz25@pitt.edu)

Ayush Mishra

February 23, 2016

**Final Project - Proposal**

* Topic: Pittsburgh Bike Share
* The general approach you plan to take:

1. Clean the data, deleting the rows that have missing values
2. Separate StartTime and StopTime into 6 columns, including day, month, year, hour, minute, and second, merge “AM” or “PM” into hour, using 24-hour type to represent those two values.
3. The data set is using “second” to represent TripDuration. We found that the result calculated from “StopTime - StartTime” is not equal to TripDuration. So we need to decide which value we should use to represent the time that the user spent. For convenience, we decided to choose TripDuration to represent the time the user spent.
4. We may do some aggregation on the attribute “FromStationName” based on the region stations belong to. Using a relatively bigger region instead of only an individual station to show a more general situation of bike sharing.
5. For the task - to predict the total count of bikes rented during each hour, we will use the new attribute “hour” as “y”, and other attributes as “x”. And use popular algorithms to build classification models, including logistic regression, knn, Naive Bayes, Decision Tree, SVM, ada, and ridge regression. Using ROC and accuracy to do the evaluation. We may also use importance matrix to select the most important variables. And use those variables to build a more accurate and robust model.
6. For the “rebalancing problem”, we may build a rebalancing system with all the stations in pittsburgh and 2 vehicles to transfer those bicycles with respect of the amount of bicycles “out” of the station and the amount of “into” the station as well as time-related considerations. Finally, we may draw a route of the repositioning operation by the result of a good model and solution.