Ride Hailing Supply and Demand Forecasting using Didi-Tech Dataset
Data Collection Report

Kimberly Williamson, Gopal Menon
Data Mining
Spring 2017, University of Utah

I. Background

Didi Chuxing is the leading ride hailing company in China and processes over 11 million trips, plans over 9 billion routes and collects over 50TB of data per day. They organized a worldwide algorithm challenge in the year 2016 [1] for forecasting ride supply and demand. Teams were asked to come up with a ride supply and demand forecasting algorithm based on a train and test dataset that was provided. Algorithms submitted by the teams were evaluated based on a published metric that measured the average error in the forecast.

II. Project Scope

The project team will use the 2016 Didi algorithm competition dataset to forecast taxi trip supply, demand, and expected fare for any given date, time, and location using the regression methods that will be covered in the Data Mining 2017 Spring semester at the School of Computing of the University of Utah. The accuracy of the forecast will be evaluated using same the average forecast error metric that was used in the competition.

III. How the Data was obtained

Training and Testing data was provided by Didi as part of the 2016 competition.

IV. Data Size

TABLE I
Train and Test number of rows

Dataset	Orders	Traffic	Weather	Cluster Map	POI
Train	8,540,614	193,553	4811	66	66
Test	557,985	8381	78	66	66

V. Data Format

Didi divides a city into n non-overlapping square districts $D = \{d_1, d_2, \ldots, d_n\}$ and divides one day uniformly into 144 time slots $t_1, t_2, \ldots, t_{144}$, each 10 minutes long. The training set contains 3 consecutive weeks of data for City M in 2016, and we need to forecast the supply-demand gap for a certain period in the 4^{th} and 5^{th} weeks of City M. Following are the tables in the dataset in tab separated format:

1) **Order Info** has the basic information of an order and contains order id, driver id, passenger id, start district hash, destination district hash, price and time
2) **District info** has shows the information about the districts to be evaluated and contains district hash and district id
3) **POI info** hasthe attributes of a district such as the number of different facilities, district hash and POI class
4) **Traffic Jam Info** the overall traffic status on the road in a district and contains district hash, number of road sections at different congestion levels and time
5) **Weather Info** has the weather for every 10 minutes and contains time, weather, temperature and pollution level

VI. Data Processing

Our plan is to create a single row for each order row by combining all the above tables and convert the data into an unnormalized format. Since the dataset contains numeric as well as categorical data counts, we will need to separate the counts for various categories and put them into their own columns. This will result in a table with the having more columns than were present in the original dataset.

VII. Data Simulation

The dataset has missing or incomplete data due to orders that were not fulfilled by drivers, many orders generated by a customer for the same ride, orders generated by third party applications and data that could not be collected due to glitches or technological limitations. This missing and incomplete data will need to be filled in before it is used for prediction of ride volumes and pricing. We plan to use matrix completion techniques for this purpose at this point of time. The Didi dataset is already separated into train and test parts and we plan to use k-fold cross validation in order to come up with a good model for prediction.

References

[1] "Algorithm Competition." *Algorithm Competition*. N.p., n.d. Web. 28 Jan. 2017.