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.

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 an 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 multiple non-overlapping square districts and divides one day uniformly into 144 time slots, 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 the city. Following are the tables in the dataset in tab separated format [1], where fields that are underlined will not be used for regression, *categorical* fields are shown in italics and numeric fields are shown in normal font:

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 the information about the districts to be evaluated and contains *district hash* and *district id*.
3) **POI info** has *district hash* and the associated attributes for the places of interest in the district such as the number of different facilities and *POI class*.
4) **Traffic Jam Info** has 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

The data will be processed in order to convert into a format that can be used for regression. Wrapper classes will be created for POI, Traffic and Weather data that will have getter methods for returning data. The order data will be joined with the POI data using the start and destination district hashes, with traffic jam data using start and destination district hashes and the time of the order, and with weather data using the time of the order. The data for joining will be retrieved using the getter methods described above. Wherever time is used for joining, the closest traffic jam and weather times will be used. Categorical fields will be converted to lists of the form $C = \{c_1, c_2, \ldots, c_p, \ldots, c_k\}$, where k is the total number of categories present and all entries are 0, except the one entry that is 1, corresponding to the category that is present. Time fields will be converted in a similar manner where the entry with a 1 will be at a position corresponding to the time slot number. Additionally there will be binary elements in the list corresponding to weekend day and holiday.

VII. Data Simulation

The dataset has missing or incomplete data due to orders that were not fulfilled by drivers, multiple 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. Unfullfilled orders will be removed and not used as inputs into the regression models. We plan to use matrix completion techniques to fill in missing data when data is required. 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.