Ride Hailing Supply and Demand Forecasting using Didi-Tech Dataset Intermediate 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 [**?**] 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 regression models explicated 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 the average forecast error metric used in the competition.

III. Current Progress

In order to run the regression models, the categorical values in the Didi algorithm competition dataset needed to be converted into a regression friendly format. Depending on the type of categorical value, the new values are lists that consist of 0 values when the category was not present in an order and 1 or a respective count when the category is present in the order. The project team has coded the input values for the regression models. The input values consist of:

1) An order key, comprised of the values for location, weather, traffic, and points of interests.
2) The cost, which is the median price of orders for a specific date and time.
3) And the number of orders for a specific date and time.

With the input values created, the project team can now experiment with multiple regression models.

IV. Stochastic Gradient Descent

The project team has experimented with the Stochastic Gradient Descent (SGD) model of regression. The results of the initial regression can be found below. Approximately 13.76% of the values were determined to have any significance, meaning the coefficient for the value was greater than 1 or the value was less than -1, in predicting the number of orders. The majority, $\approx 92.12\%$, of the significant values used to predict the number of orders were part of the time categories. Meaning the time of the day had the most significant impact on the number of orders.

Approximately 23.90% of the values were determined to have any significance, meaning the coefficient for the value was greater than 1 or the value was less than -1, in predicting the order price. The majority, $\approx 77.55\%$, of the significant values used to predict the number of orders were part of the district categories. Meaning the districts, both start and destination, had the most significant impact on the order price.

TABLE I
Mean Squared Error

Predictive Variable	Mean Squared Error
Number of Orders	182.777
Order Price	518.663

V. Future Work

The project team plans to experiment with other models of regression including Ridge, Lasso, and Ordinary Least Squares. During review of the regression model output the project team may adjust the input values to reduce the Order Keys value by eliminating variables that are shown to have minimal impact.

References

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