Ride Hailing Supply and Demand Forecasting using Didi-Tech Dataset

{ Gopal Menon and Kimberly Williamson } University of Utah Data Mining Spring 2017

Introduction

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. We used the 2016 Didi algorithm competition dataset to try and forecast taxi trip supply and demand for any given date, time, and location using regression models covered in the Data Mining 2017 Spring semester at the School of Computing of the University of Utah. Below is a sample of some methods used:

1. Linear Regression using Stochastic Gradient Descent
2. Polynomial Regression
3. Gaussian Kernel Regression
4. Hierarchial Clustering

Ride Hailing Data

- Taxi Cab Data for January 2016
- 1.3 Million Rows of Training Data Reduced to 921,000 Rows
- 21 Days of Training Data and 5 Days of Testing Data

Some of the Data Elements Included:

- Hashed Start and Destination District
- Time of Pickup
- Points of Interest
- Weather
- Traffic

Categorical values were converted into a regression friendly format.

Regression Results

Regression Type	Mean Squared Error
Linear using Stochastic Gradient Descent	245.50
Gradient Boosting using features based on top 10 eigen vectors	285.97
Ridge Regression using features based on top 10 eigen vectors	293.83
Lasso Regression using features based on top 10 eigen vectors	293.83
Polynomial degree 2 regression using features based on top 10 eigen vectors	2.34×10^{53}
Polynomial degree 3 regression using features based on top 10 eigen vectors	3.53×10^{73}
Polynomial degree 4 regression using features based on top 10 eigen vectors	9.31×10^{93}
Gaussian Kernel Regression	376.90

Table 1: Regression Results

References

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

Patterns In Ride Hailing Data

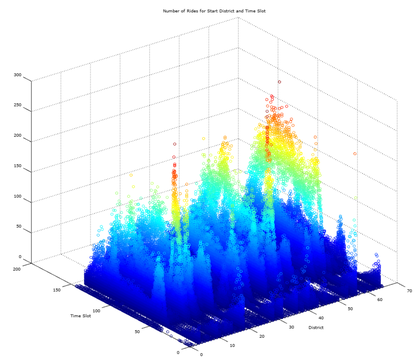

Figure 1: Number of rides scatter plot for start district and time slot.

Commercial vs Residential

- Commercial - Districts where Didi Chuxing Customers Work
- Residential - Districts where Didi Chuxing Customers Live
- Peak Timeslots and Districts were Visible
- Pattern Not Strong Enough to Reduce Complexity

Outliers

- Boxplot
  - Computed with Inter-Quartile Distance Between the First and Third Quartiles
  - Included Elements that are up to 1.5 Times the Inter-Quartile distance
  - Outliers Removed had Little Effect
- Clustering
  - Agglomerative Clustering
  - Sparsed Data did not Create Well-Formed Clusters
  - Unable to Remove Outliers

However it was doubtful that the reported outliers were accurate since it is actual data and the reported number was large.

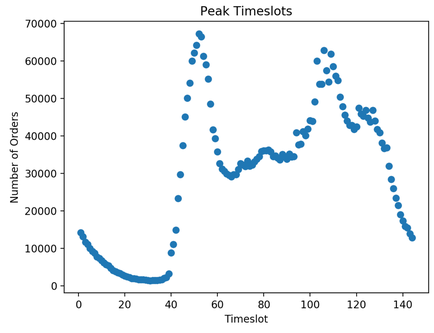

Figure 2: Number of rides scatter plot for a timeslot (Monday-Friday).

Conclusion

- Traditional Regression Methods Do Not Work for Complicated Situations with Human Behavior
- Deep Learning Techniques Would be Better Fit to Handle this Project
- Missing Geo-Location Restricted Ability to Deliver Meaningful Results