



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

1. Leading ride hailing company in China
2. Processes over 11 million trips
3. Plans over 9 billion routes
4. Collects over 50TB of data per day
5. Organized algorithm challenge in 2016 [1]
 - for forecasting ride supply and demand
 - we used the Didi algorithm dataset
 - to forecast taxi trip supply & demand

Below are the regression methods used:

1. Stochastic Gradient Descent
2. Gaussian Kernel Ridge
3. Gaussian Kernel
4. Hierarchical Clustering
5. Using top 10 eigen vectors
 - (a) Polynomial degrees 2, 3 and 4
 - (b) Gradient Boosted
 - (c) Ridge
 - (d) Lasso

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

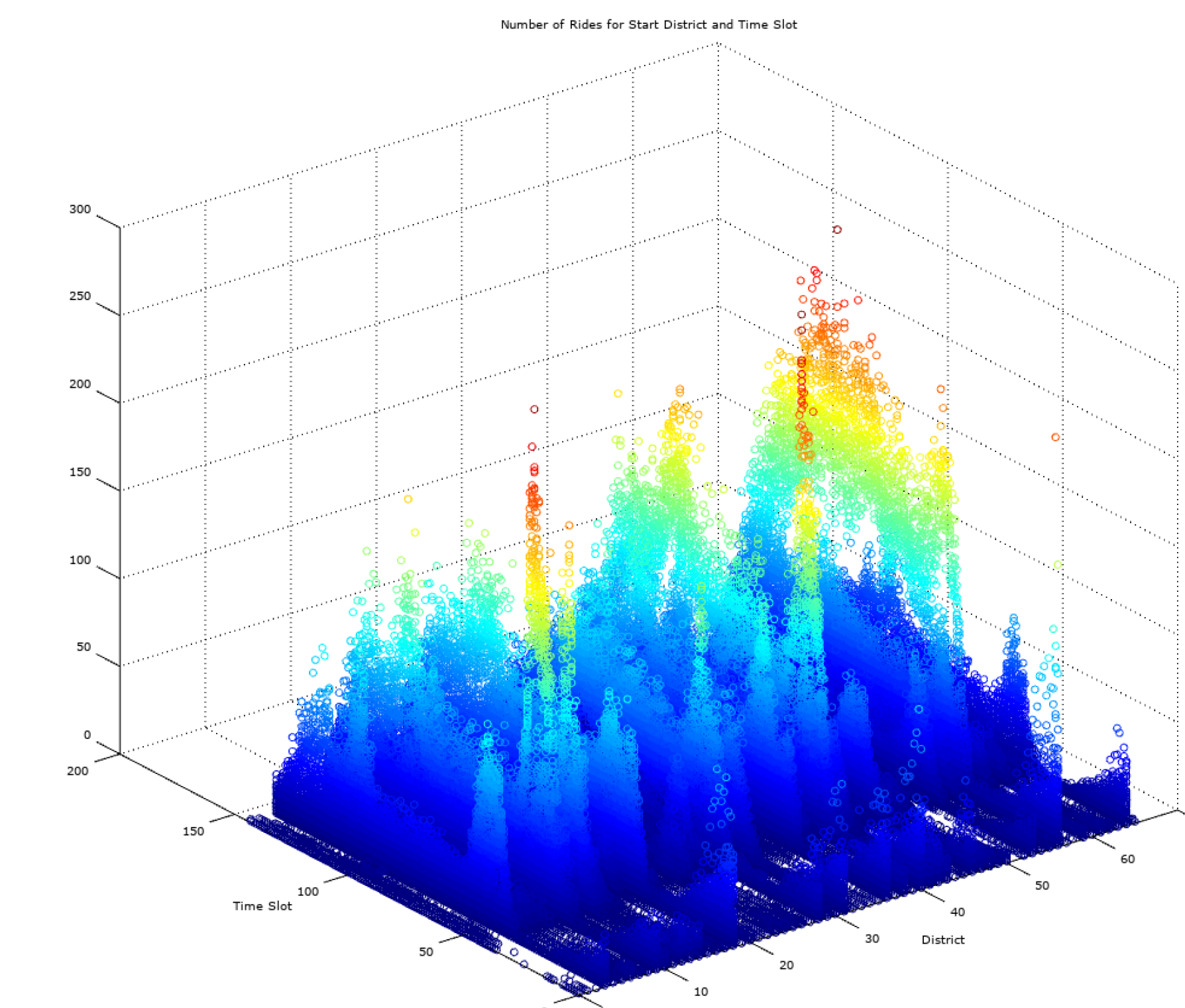


Figure 1: Number of rides scatter plot for start district and time slot.

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

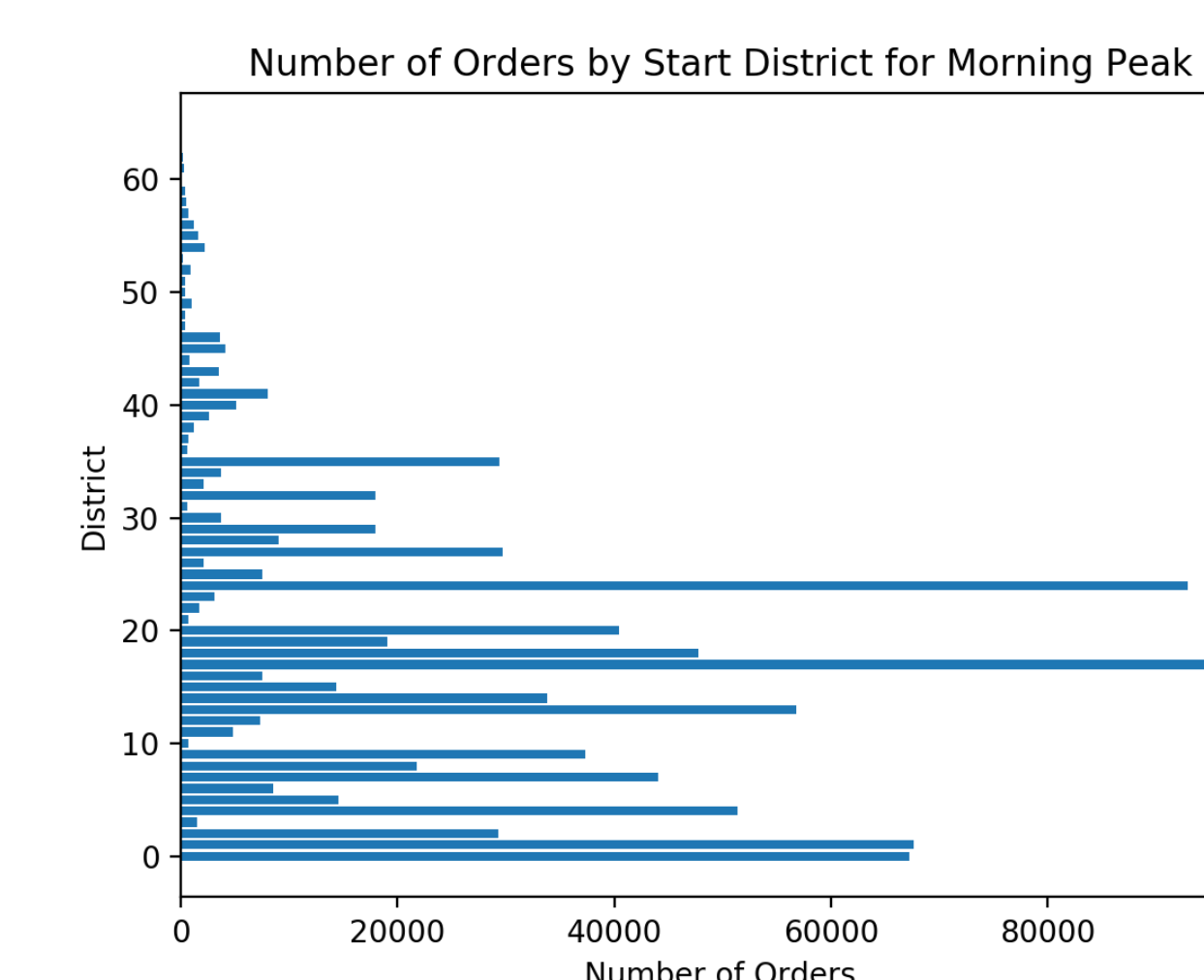


Figure 2: Morning peak orders

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.

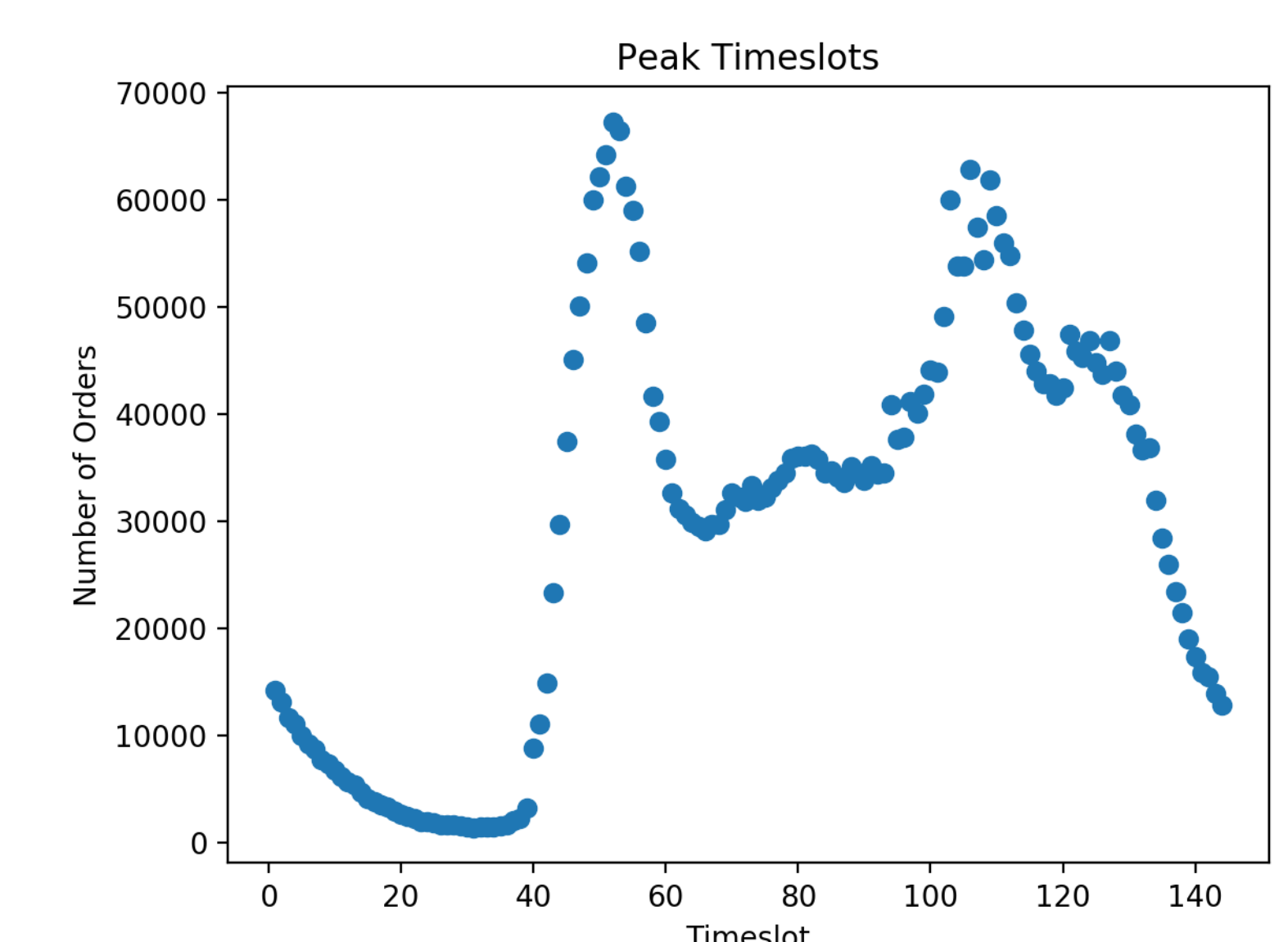


Figure 3: Number of rides scatter plot for a timeslot (Monday-Friday).

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