STA 208 Project Report - NYC Taxi data pickup prediction

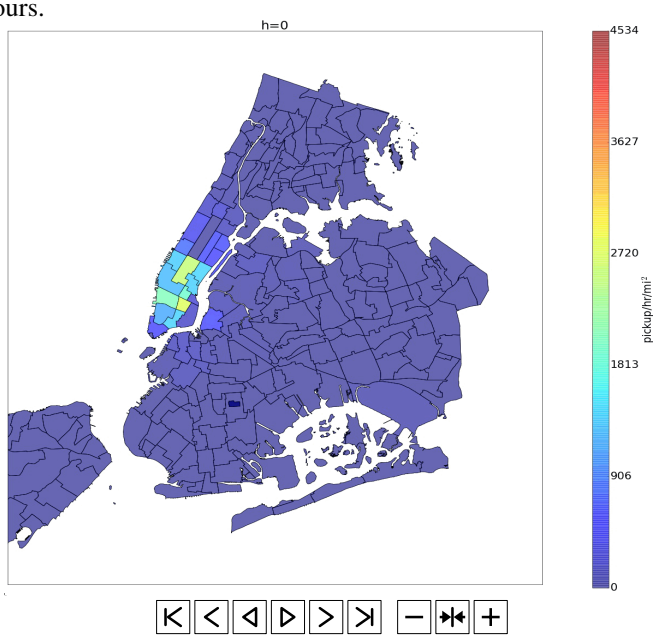
Wenhao Wu, Xingtai Li, Guicheng Wu

***Abstract*— From 2012 to 2015, there were about 170 million taxi trips in New York City each year, including both green cab and yellow cab. Accurate prediction of taxi pickup numbers can help people monitor the traffic efficiently. With this purpose in mind, we used regression models to predict the number of pickups in certain region in New York City, given trips in previous hour and weather information. We split 195 areas in NYC and applied five different models to train our features, including Linear regression, Lasso, Support Vector Regression, Bayesian ridge regression and Decision Tree. Then we found out the optimal estimator for each model by tuning the parameters. Based on our validation, we found that Decision Tree Regressor has best R^2 score of 0.966 and lowest MSE, which is 180.**

I. INTRODUCTION

An accurate prediction of the pickup number of taxi in certain blocks of New York City can help taxi drivers to position themselves smartly, guide customers to call the cab at right time, and help transportation officers forecast the potential traffic jams in the future. In our project, we are going to predict the number of pickups in certain region of New York City based on the current time, last-hour pickup numbers and weather information. We implemented different regression models to predict the expected pickups and compared their performance based on R^2 and mean squared error metrics.

Because the size of data(about 240 GB) is extremely large, we stored our data in the PostgreSQL database to enhance the performance of querying data. Clicking the button under the picture to see the change of taxi pickup numbers in 24 hours.



II. METHODOLOGY

A. Data set and processing tools

The primary dataset studied in this project is the New York City (NYC) Taxi & Limousine Commission (TLC) Trip Record Data [1] during 2014 including both the yellow and green taxi trips, which in csv formats amount to a total of 25GB. Among various fields, we are mainly interested in the time and venue of the pickup location. In order to enable efficient manipulations of the data, the trip data are firstly written into a PostgreSQL database with PostGIS extension using package `sqlalchemy` with `geoalchemy2`. After data cleansing, a total of 177312293 records are written into the database.

Due to the sheer size of the dataset, it would be extremely difficult to study the temporal and spatial distribution of the trips in the continuous domain. Moreover, instead of pinpointing the density of pickup at a certain coordinate and time, a more practical and informative result would be how many taxi pickups per square mile is expected during the next hour in each of the (administrative) neighborhoods. Therefore we decide to map the pickup longitude and latitude for each trip are mapped to the 195 Neighborhood Tabulation Areas (NTAs) [2] and the 2166 Census Tracts (CTs) [3] in 2010 as in [4] and count the total number of pickups in the spacial resolution of 1 NTA/CT and the temporal resolution of 1 hour.

As we believe that the distribution of taxi trips are also related to daily events such as weather conditions and holidays, we also incorporate the daily average temperature and precipitation data recorded at New York Central Park Belvedere Tower observing station available from NOAA [5] as well as the holiday calendar for 2014.

The models and evaluation metrics we applied come from the elegant `sklearn` python package for machine learning. The visualization of the regression results are implemented with Matplotlib with the Basemap toolkit.

B. Feature selections

Our goal of project is to predict the number of pickups in certain area and time. Inspired by the MIT 2013-2014 Big Data Challenge [6] and the related course project [7], as well as the recent algorithm competition by DiDi Research [8], we selected the following seven features to build the regression model as shown in Table III. In order to justify this feature selection, in Fig. 1 we plot the total number of pick-ups in NYC for each hour in 2014 and its FFT, which clearly indicates that data is subject to the multiplication of a weekly periodic pattern and a daily periodic pattern and in

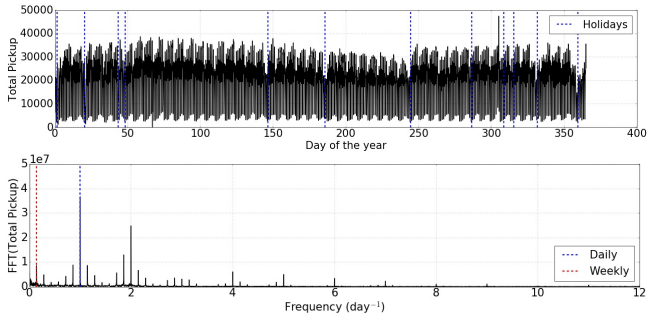


Fig. 1. Temporal and frequency pattern of the total pickup count.

turn justifies the usage of both DOW and hour as features. Moreover, it is clear that holidays always lead to significant decrease in taxi pickup. On the other hand, the weather and short-term historical pickup count are also believed to be useful features as shown [7], [8].

Among these parameters, there are several categorical data, which are not continuous values. In order to use regression algorithm to build the model, we coded these categorical data in binary integer representation. To be specific, we used the `LabelEnoder` function to encode labels between 0 and *n-1*. We list all features in table 1.

TABLE I
FEATURE SELECTION.

Feature name	Data Type	Meaning
pickup gid	categorical	The id of pickup location.
pickup dow	categorical	Day of week (0..6)
pickup hour	categorical	Hour of pickup (0..23)
temperature	numerical	Daily Temperature
precipitation	numerical	Daily Precipitation
holiday	categorical	True/False
count 1	numerical	pickups at previous hour

We extracted most features from the SQL database that we loaded before. As for the temperature and precipitation, we got the data from the National Weather Service Forest Website [9]. Since we separated the area of city in the database, we mapped the information of weather back to the database and assigned each trip their corresponding weather values.

For each categorical data type, we encoded its binary integer representation. Moreover, we used the *StandardScalar* to remove the mean and scale to unit variance. In total there are 35 features for each trip in given area (7 days+24 hours+temperature+precipitation+holiday+count).

C. Models

We applied five different models to train our feature data, then used the models to predict the pickup count number given a particular grid. These models included linear regression, Lasso, support vector regression, Bayesian ridge regression and decision tree. Also we found out the optimal parameters for each model by using $GridSearchCV$ method to tune the hyper-parameters. The results showed that the decision tree worked best. The optimal parameters of decision tree are as follows: $max\ depth = 100$ and $min\ sample\ leaf = 10$.

1) Linear regression: Linear regression is the most basic regression method, however, it has demonstrated wide availability and scientific acceptance in solving prediction and forecasting problems. Our project is to use NYC taxi data and weather data to predict the pickup count number in a given grid. Obviously it is a prediction problem which suit the linear regression method well.

2) Lasso: Lasso is considering both the parameter weights and linear regression in order to enhance the prediction accuracy. We can think lasso as the extension of least square model.

3) Support vector regression: Support vector regression use the same principle as support vector machine. The only difference is that it's output the prediction value based on the given explanatory variables.

4) Bayesian ridge regression: Bayesian linear regression is also based on linear regression. In Bayesian linear regression, the statistical analysis is undertaken within the context of Bayesian inference.

5) Decision tree: Decision tree has wide applications in business areas to choose strategies. The goal of decision tree is to create a model that predicts the value of a target variable based on several input variables. Intuitively, it can be applied to predict our taxi pickup count number in a specific grid.

D. Performance Metrics

To quantify the performance of different regression models, we used the train-test split function in *sklearn* package. For each location, we split the trips in this area and make prediction of pickup numbers based on the trained model.

There are several metric functions to accessing the prediction error of regression model, including mean squared error, mean absolute error and R^2 coefficient of determination function.

We used both R^2 and mean squared error as metrics to evaluate our model. Coefficient of determination is useful in regression because it gives the proportion of the variance of one variable that is predictable from the independent variable [10]. This statistic gives us the first impression about the goodness of the fit of a model. The equation of R^2 score is as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y_i})}{\sum_{i=0}^{n}(y_i - \overline{y})} \quad (1)$$

Where $\overline{y} = \frac{1}{n}\sum_{i=0}^{n} y_i$ and $\hat{y_i}$ is our predict value for a particular area. The largest possible R^2 score is 1. The larger the R^2 score, the better the accuracy of prediction.

Moreover, if regressors have similar R^2 scores, we want to use mean squared error(MSE) to calculate the variance of predictions compared to true value. We want to minimize the variance because large change of prediction may cause traffic jam at certain time. The mean squared error(MSE) is

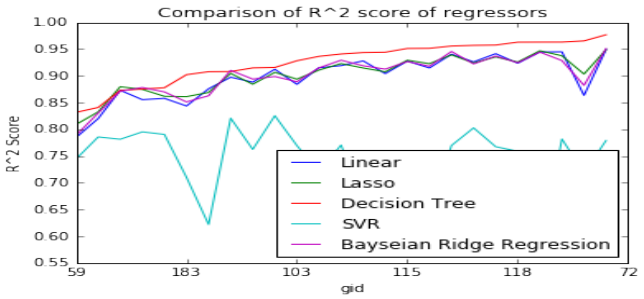


Fig. 2. Comparison of r2 scores between different Regressors

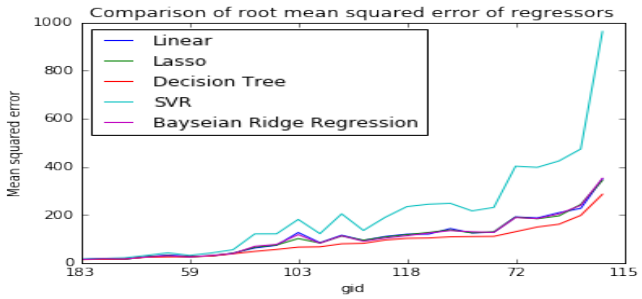


Fig. 3. Comparison of mean squared error

calculated as follows

$$MSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

III. RESULTS

A. Comparison of Models

We used 1.7 million trips (195gids*365days*24hours) in 2014 to train the regressors model. For each area, we implemented split method to train and test all five regressors. To be specific, we used 80% of the trips as training set and 20% as testing set. Then for each regressor, we calculated the R^2 and root mean squared error.

Based on our observation, areas has pickup density less than 100 trips per squared mile per hour is highly unpredictable. Therefore, we have selected 26 areas with pickup density higher than 100 trips per squared mile per hour as our targets. In Table 2 we have shown the performance of each regressor at area 27, where the density of pickup is 213.5 pickups per squared mile per hour.

TABLE II
AVERAGE PREDICTION RESULT FOR EACH REGRESSOR AT AREA 27

Regressor	Parameter	R^2	RMSE
Linear		0.8987	128.5
Lasso	$\alpha = 0.01$	0.9005	130.51
SVR	rbf kernel, $C = 1e5$	0.6710	241.61
Bayesian Ridge	$\alpha = 1e-05$,$\lambda = 1e-06$	0.9009	131.06
Decision Tree	max length =100,min leaf = 10	0.9241	104.8

To verify that the performance of regressors are similar in all locations, we trained and tested the model in 26 areas of New York City. Then we sorted the result of R^2 and mean squared error according to their performance and presented in fig 2 and fig 3. Note that the x-axis in both figures are the location ids that have been sorted according to their scores. From the graph we found that the Decision Tree Regressor has better R^2 score and lower rooted mean squared error compared to other model. A higher R^2 suggests that the prediction is close to the true pickup number while the low rooted mean square error gives us a lower variance.

Moreover, we found that linear regression, Lasso, and Bayesian Ridge regression has similar results. This is reasonable because all these three models implement linear regression approach to make prediction based on finding the relationship between trips. On the other hand, the Supporter Vector Regressor behaves poorly in our scenario because of the non-linear relationship between each feature in our case. Even though we can use kernel trick to project the feature into higher dimension, categorical features are still not linear separable.

B. Decision Tree Regression

After the comparison between different model selections, we focused on Decision Tree Regression method because of its better performance. To show that the performance of Decision Tree Regressor converges as the train data size increases, we have plotted the learning curve for the regressor at area 27 in fig 4. Note that we split 80% training set and 20% testing set for each train size.

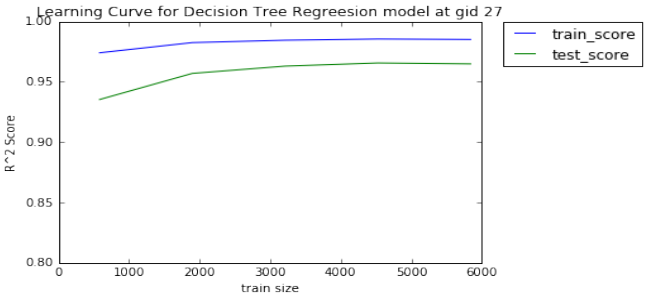


Fig. 4. Learning curve for decision tree regressor at gid 27

Both train set and test set converge to their optimal score. As the train size increases, the prediction score of Decision Tree converges to nearly 0.98, indicating that there is little overfitting problem with our model.

TABLE III
WEIGHTS OF EACH FEATURE IN DECISION TREE

pickup dow	0-6	0.0167
pickup hours	0-23	0.1076
temperature	numerical	0.01098
precipitation	numerical	0.00256
holiday	True/False	9.91e-05
previous count	Numerical	0.86196

Then we construct the tree diagram for the Decision Tree Regressor at area 27. The subsection of tree diagram shows the relative weights of features. In table III, we have shown the relative weights of each features in our trained Decision Tree Regressor. Moreover, based on the Decision Tree model, we have constructed the diagram of decision process at area 27 [Fig 5.].

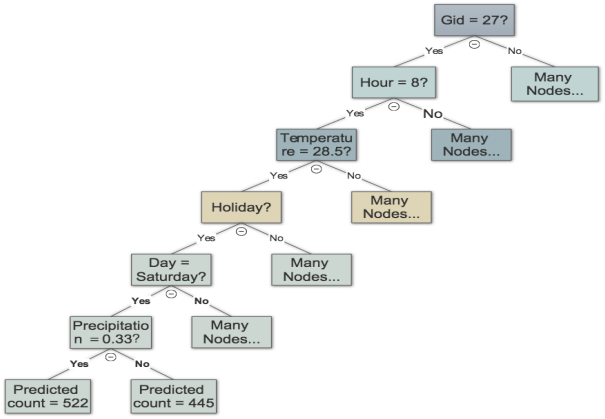


Fig. 5. Final trained decision tree

Based on the subsection of decision tree we conclude that the prediction of pickup count heavily depends on the count of previous hour. This is reasonable because the our current pickup number is based on the number of previous hour. Among other features, pickup hours determines the decision of subsection in the first place because it has relative higher weights than other features. Therefore, we can determine the informative weight of each feature by using Decision Tree regression successfully. Then we use built Decision Tree regression model to make prediction of pickup number at area 27 within 7 days. We have randomly selected one week in June, 2015 and compared the predicted value with the true value. The result is show in Fig.6 below.

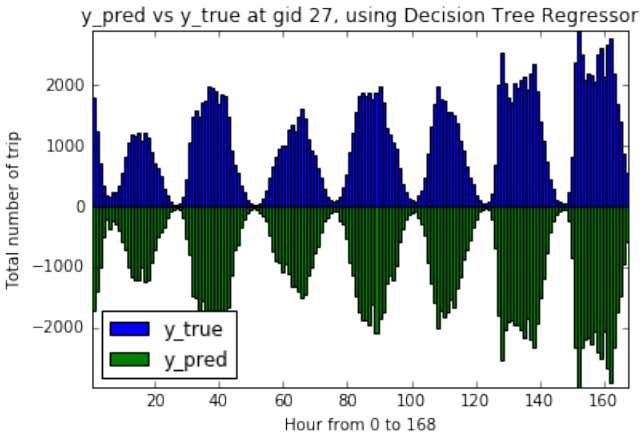
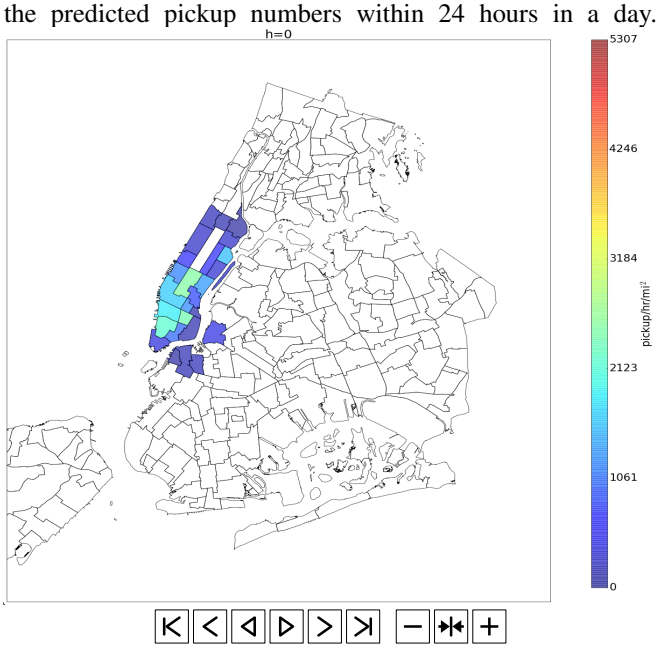


Fig. 6. y_pred vs y_true at gid 27, using Decision Tree Regressor

Based on the comparison of predicted value and true value we found that our prediction using Decision Tree regression model has successfully captured the variance of pickup number at different time period. At every daytime there is peak of calling cabs with in the area, while the number has decreased significantly at late night. Moreover, the number of pickups on weekends is larger than the number on weekdays.

The picture below is our predicted NYC taxi pickup numbers for the chosen 26 areas within 24 hours in a day. You can click the buttons to see the differences the predicted pickup numbers within 24 hours in a day.



IV. Conclusions

In order to predict the pickup number of taxi in New York City at given time and area, we selected several numerical and categorical features and implemented five regression models. Based on our result, the performance of Decision Tree regression model is the best compared to other linear regressors and supporter vector regressors. By capturing the relative weights of each feature, Decision Tree model achieved a R^2 of 0.9241 and rooted mean squared error of 104.8 on average (given populated areas). Moreover, the increasing scores of learning curve as the train size increases eliminates our concerns about overfitting.

A. Future recommendation

There may be relationships between areas that are near to each other. We can extract features such as pickup numbers in previous hour from areas that is close to current region ID. However, we still have to decide the relative weights of these features to prevent overfitting problem.

Besides feature selection, we can add more prediction model besides traditional regression methods, such as Neural Network regression and unsupervised learning like k-means clustering.

REFERENCES

[1] N. T. . L. Commission. (2016) Nyc tlc trip data. [Online]. Available: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[2] N. Department of City Planning. (2014) Neighborhood Tabulation Areas. [Online]. Available: https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq

[3] ——. (2014) Census Tracts. [Online]. Available: https:// data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku

[4] T. W. Schneider. (2015) Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. [Online]. Available: http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

[5] National Oceanic And Atmospheric Administration (NOAA). Data Tools: Find a Station. [Online]. Available: http://www.ncdc.noaa.gov/cdo-web/datatools/findstation

[6] MIT Computer Science and Artificial Intelligence Lab (CSAIL). MIT Big Data Challenge. [Online]. Available: http://bigdata.csail.mit.edu/challenge

[7] J. Grinberg, A. Jain, and V. Choksi, "Predicting taxi pickups in new york city," *Final Paper for CS221 Artificial Intelligence, Computer Science Department, Stanford University*, 2014.

[8] Didi Research Institute. Di-Tech Challenge. [Online]. Available: http://research.xiaojukeji.com/competition/

[9] N. W. Service. (2016) NOAA Online Weather Data- New York. [Online]. Available: http://w2.weather.gov/climate/xmacis.php?wfo=okx

[10] N. J. D. Nagelkerke, "A Note on a General Definition of the Coefficient of Determination," *Biometrika, Vol. 78, No. 3. (Sep., 1991), pp. 691-692.*, 2008.