Predicting taxi fare prices in New York City using various regression models and evaluating their performance.

Gerli Poopuu GERL0016@STUD.KEA.DK

Jakub Kriz JAKU0526@STUD.KEA.DK

Vaidaras Pranskaitis VAID004@STUD.KEA.DK

Grzegorz Goraj GRZE0203@STUD.KEA.DK

Abstract

This paper includes an analysis of a New York City taxi dataset coupled with fare price predictions obtained from the following regression models: Linear Regression, Extra-Trees Regression, Bagging Regression, Ada-Boost Regression, Decision-Tree Regression, Random-Forest Regression, Gradient-Boosting Regression and Multi-layer Perceptron Regression. The features from the dataset are described, analyzed and then transformed into new features that are used to construct the aforementioned models. The predictive power of these models is evaluated according to the following performance metrics: Root Mean Square Error (RMSE) and Coefficient of Determination (R2). The results show that Multi-layer Perceptron predictions are marginally better than the other mentioned models, having a RMSE of 3.395 and R2 of 0.849.

1. Introduction

In this paper we attempt to estimate taxi fare prices in New York City using various regression models trained with features that were derived from a dataset of New York City taxi rides (Kag, 2018). The performance of each model is assessed by comparing the RMSE (Fürnkranz et al., 2010) and R2 (Renaud and Victoria-Feser, 2010) of each model.

2. Methods

2.1 Statistical and Machine Learning Methods

Sci-kit learn, a Python library was our main source of algorithms. The geographic location of each data-point was clustered using the K-Means algorithm (Jain, 2010). To predict fare prices the following regression models were used: Linear Regression (Sedgwick, 2013), Extra-Trees Regression (Galelli and Castelletti, 2013), Bagging Regression (Sutton, 2004), Ada-Boost Regression (Collins et al., 2002), Decision-Tree Regression (Podgorelec and Zorman, 2015), Random-Forest Regression (Cutler et al., 2012), Gradient-Boosting Regression (Zemel and Pitassi, 2001) and Multi-layer Perceptron Regression (Murtagh, 1991).

2.2 Feature Transformation

The pick-up and drop-off latitude/longitude data-points were transformed into metric dis-

tances. This was done according to latest World Geodetic System (WGS84) (Agency et al., 2005) using Geod from the Python library, pyproj.

2.3 Visualization

Matplotlib, a Python library, was used to generate all visualizations. The dataset is explored through the use of histograms, bar-charts, scatter-plots, line-plots and colored geographic plots that highlight important features. Tables and scatter-plots are used to visualize the performance metrics of the models.

3. Dataset

The dataset used for this paper is taken from an online platform called Kaggle (Kag, 2018). It is an online community of data scientists and machine learners owned by Google (Lardinois et al., 2017). The original dataset consist of 55 million data-points, each containing 8 features (Table 1). Since our main goal is to find the best model for predicting taxi prices in New York city, *Fare Amount* is chosen to be our dependant variable.

Variable	Data type
Key	String
Fare Amount	Ratio
Pick-up Date-time	Interval
Pick-up Longitude	Interval
Pick-up Latitude	Interval
Drop-off Longitude	Interval
Drop-off Latitude	Interval
Passenger Count	Ratio

Table 1: Variables in the dataset.

4. Analysis

4.1 Dependent Variable

The dependent variable is *fare amount*, expressed in United States Dollars. As can be seen in Figure 1, the vast majority of fares are under 20 dollars with a small edge peak in the 40-60 dollar range.

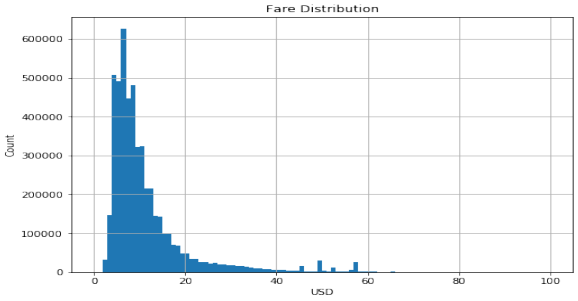

Figure 1: Fare amount distribution

4.2 Passenger Count

Figure 2 shows that the majority of rides have 1 passenger and rarely exceed 6 passengers. It is also interesting to note that there are rides with 0 passengers, these could potentially be corrupt data-points or a taxi providing a courier service. The correlation analysis value between *passenger count* and *fare amount* is low at 0.0162078, indicating that there is little to no relationship between these variables.

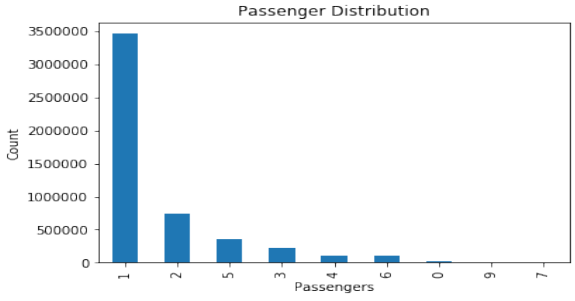

Figure 2: Passenger count distribution

4.3 Pick-up Date-time

The dates in the dataset range from 2009 to 2015 with there being less data-points from 2015, this is shown in Figure 3.

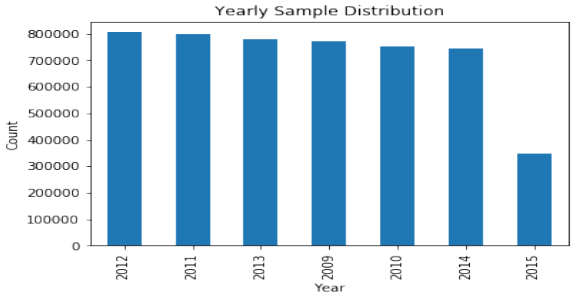

Figure 3: Passenger count distribution by year

The relationship between the *fare amount* and the *pick-up date-time* is shown in Figures 4 and 5. Figure 4 shows that average fares seem to rise over the years and that average fare amounts rise at around 5am and 3pm. Figure 5 shows that average fares drop on Saturdays.

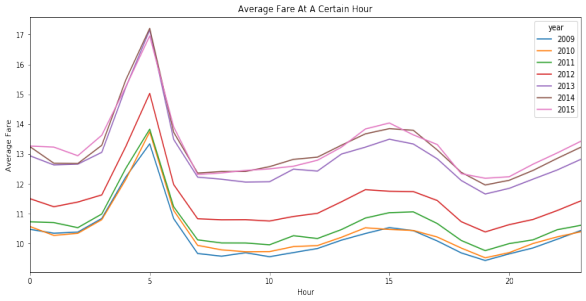

Figure 4: Average fare vs time of day

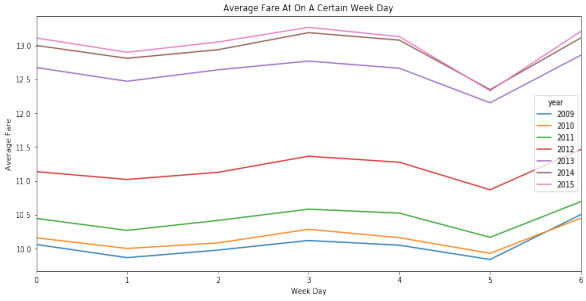

Figure 5: Average fare vs weekday

4.4 Pick-up and Drop-off Co-ordinates

Pick-up and *drop-off* co-ordinates for each ride are plotted in Figure 6 and give an overview of the geographic location of each data-point. The most dense *pick-up location* is in central New York followed by two airports.

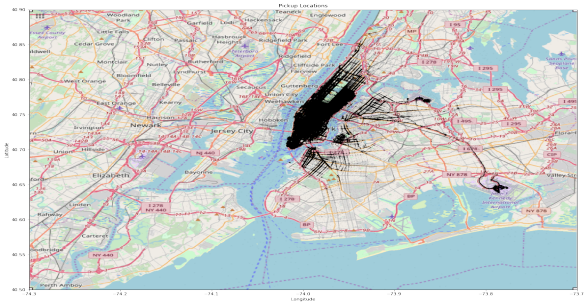

Figure 6: Taxi pick-up locations

The relationship between *pickup location* and *fare amount* can be seen in Figure 7, it shows that fares are lowest in the center of the city and rise in the outskirts, peaking at the airports.

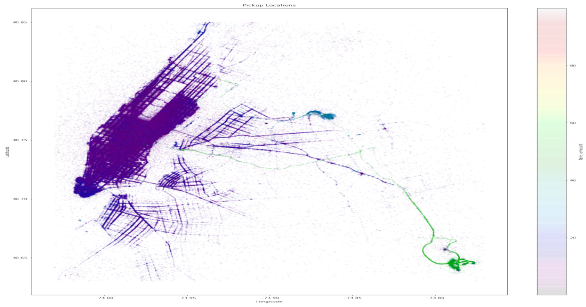

Figure 7: Taxi pick-up locations with fare heat-map

5. Fare Price Prediction

5.1 Data Preprocessing

5.1.1 Train test split

In order to measure the generalization ability of our models we split the data into training, test and validation sets. The training set contained 1,000,000 samples and the test-set coupled with the validation-set contained 100,000 samples.

5.1.2 Removing anomalous data-points

The dataset contained fare amounts that were negative and some that were extremely higher than the mean. It also contained data-points with *latitude* and *longitude co-ordinates* that

pointed outside the locality of New York. All of these data-points were removed. It should also be noted that we excluded the *passenger count* column since it has a low correlation with *fare amount*.

5.1.3 Additional features

As the original dataset did not seem to provide features in a format that would be beneficial to our models we decided to add new features derived from the *pick-up/drop-off co-ordinates* and *date* variable.

Distance. We derived the distance travelled in kilometers between the *pick-up/drop-off co-ordinates*. The relationship between *fare amount* and *distance* is show in Figure 8. The figure indicates that there is a correlation between the *distance* and *fare amount* and this is confirmed by a correlation analysis output value of 0.737013.

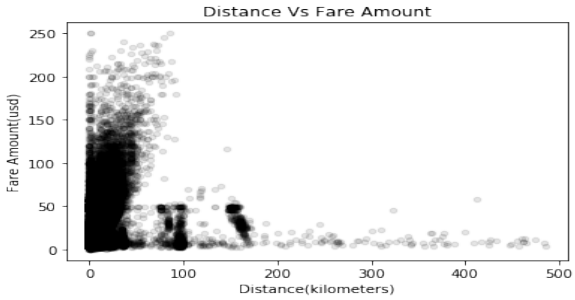

Figure 8: Fare vs distance

Pick-up and Drop-off Clusters. Using K-Means we were able to categorize each co-ordinate into localized clusters. This was done in order to give context to each co-ordinate value that our models could hopefully take advantage of. The generated clusters are shown in Figure 9 and their relationship with the fare amount in Figure 10. Note that Figure 10 is similar to Figure 7.

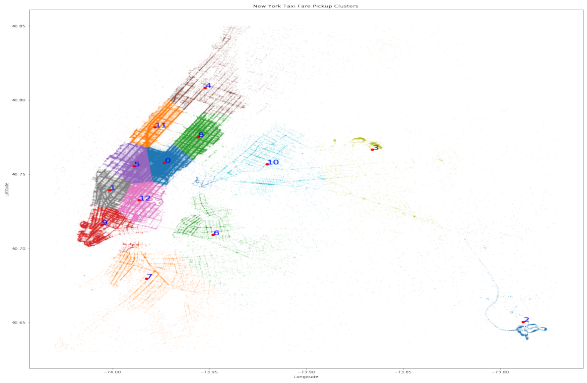

Figure 9: Cluster zones

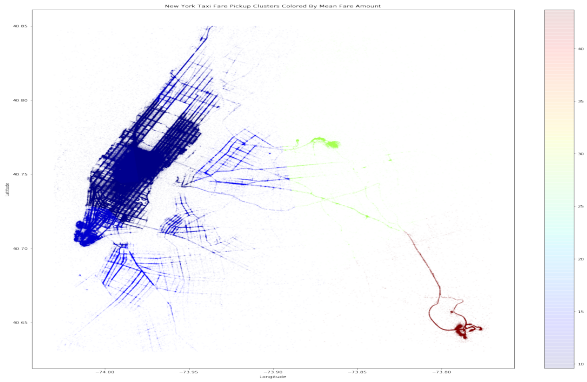

Figure 10: Mean fare amount per cluster

Hour, Weekday and Year. In order to make use of the date timestamp we decided to derive the *hour, weekday* and *year* for each respective data-point. This was done in hopes to provide features that would potentially enable our model to contextualize the relationship between *fare amount* and the *pick-up time*. This relationship is shown in the previous Figures 4 and 5.

5.1.4 Describing Our New Features

Distance is in kilometers and is expressed as ratio data-type, *pick-up* and *drop-off* clusters are integer values ranging from 0-12 as shown in Figure 9. *Hour*, *weekday* and *year* are also integer values as shown in Figures 4 and 5.

5.1.5 Scaling

As neural networks tend to be more sensitive to non-scaled values we decided to use input features scaled from 0 to 1 for our Multi-layer Perceptron (MLP) regression model. The rest of our models used non-scaled input values.

5.2 Regression Models

The regression models used are mentioned in section 2.1. The performance metric results RMSE and R2 can be seen in table 2. In the table, MLP1 and MLP2 are Multi-layer Perceptron networks. MLP1 consisting of hidden layers with 20 and 50 perceptrons respectively and MLP2 with 3 hidden layers of 100 perceptrons. The results show that MLP2 performs the best in terms of RMSE and R2 and that Ada-Boost performs the worst. In order to improve the result, we also decided to average the predictions of the Extra-Trees, Bagging, Random-Forest, Gradient-Boosting and MLP2 models into a new prediction. This resulted in better performance in terms of a slight decrease in RMSE and increase in R2. Figure 11 shows the ground truth versus each models prediction.

Model	RMSE	R2
Linear	4.264	0.761
Extra-Trees	3.576	0.832
Bagging	3.565	0.833
Ada-Boost	4.789	0.699
Decision-Tree	4.562	0.727
Random-Forest	3.400	0.848
Gradient-Boosting	3.550	0.835
MLP1	3.608	0.829
MLP2	3.395	0.849
Average	3.287	0.858

Table 2: Regression models' performance results.

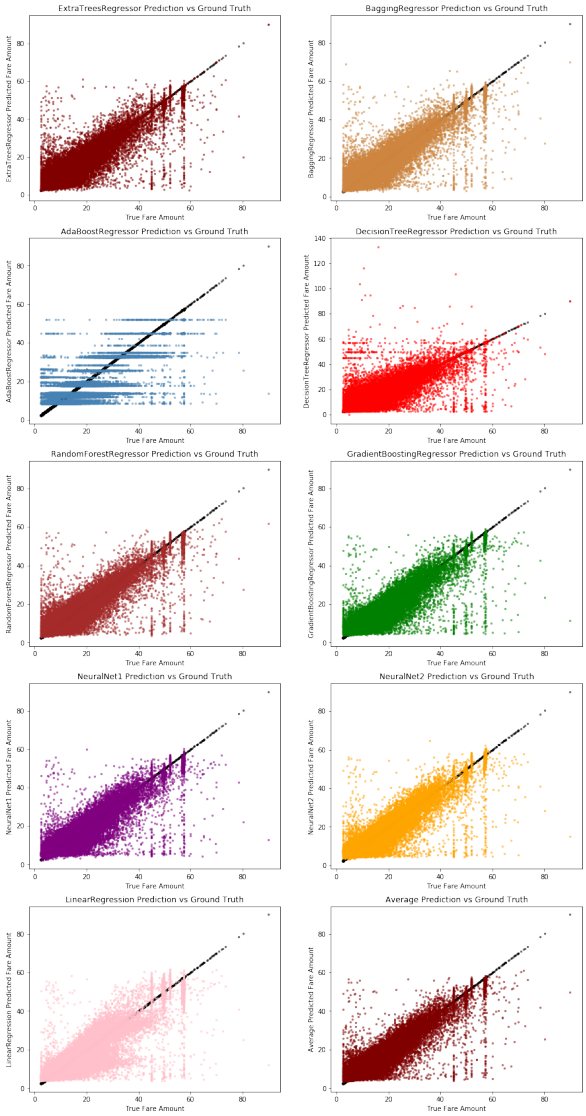

Figure 11: Models vs ground truth

6. Findings

The main goal of our research was to determine how effectively different regression models could predict the *taxi fare* in New York based of *pick-up* and *drop-off locations* coupled with the *time of pickup*. We introduced new features derived from these variables. These features were used as inputs to our models. The predictive power of each model was evaluated in terms of RMSE and R2 where a lower RMSE and a

higher R2 were desirable. Our results showed that a Multi-layer Perceptron network of 3 hidden layers with 100 perceptrons performed the best with an RMSE of 3.395 and R2 of 0.849. More desirable results were also achieved by averaging the predictions of the best models leading to a slight decrease of the RMSE value to 3.287 and increase of the R2 value to 0.858.

Acknowledgments

We wish to acknowledge our teacher Henrik Strøm for his guidance and support.

References

New York City Taxi Fare Prediction | Kaggle, 2018. URL `https://www.kaggle.com/c/new-york-city-taxi-fare-prediction`.

National Geospatial-intelligence Agency, Abbreviated Frame, Associated Trs, Global Type, Dimensional Last Version, Reference Epoch, Brief Description, Frame Earth, Iers Reference Pole, B I H Conventional, Terrestrial Pole, Iers Reference Meridian, The Irm, and B I H Zero Meridian. World Geodetic System 1984. Technical report, 2005.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 2002. ISSN 08856125. doi: 10.1023/A:1013912006537.

Adele Cutler, D. Richard Cutler, and John R. Stevens. Random forests. In *Ensemble Machine Learning: Methods and Applications*. 2012. ISBN 9781441993267. doi: 10.1007/9781441993267_5.

Johannes Fürnkranz, Philip K. Chan, Susan Craw, Claude Sammut, William Uther, Adwait Ratnaparkhi, Xin Jin, Jiawei Han, Ying Yang, Katharina Morik, Marco Dorigo, Mauro Birattari, Thomas Stützle, Pavel Brazdil, Ricardo Vilalta, Christophe Giraud-Carrier, Carlos Soares, Jorma Rissanen, Rohan A. Baxter, Ivan Bruha, Rohan A. Baxter, Geoffrey I. Webb, Luís Torgo, Arindam Banerjee, Hanhuai Shan, Soumya Ray, Prasad Tadepalli, Yoav Shoham, Rob Powers, Yoav Shoham, Rob Powers, Geoffrey I. Webb, Soumya Ray, Stephen Scott, Hendrik Blockeel, and Luc De Raedt. Mean Squared Error. In *Encyclopedia of Machine Learning*. 2010. doi: 10.1007/978-0-387-30164-8_528.

S. Galelli and A. Castelletti. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology and Earth System Sciences*, 2013. ISSN 10275606. doi: 10.5194/hess-17-2669-2013.

Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010. ISSN 01678655. doi: 10.1016/j.patrec.2009.09.011.

Frederic Lardinois, Matthew Lynley, and John Mannes. Google is acquiring data science community Kaggle | TechCrunch, 2017. URL `https://techcrunch.com/2017/03/07/google-is-acquiring-data-science-community-kaggle/`.

Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 1991. ISSN 09252312. doi: 10.1016/0925-2312(91)90023-5.

Vili Podgorelec and Milan Zorman. Decision Tree Learning. In *Encyclopedia of Complexity and Systems Science*. 2015. doi: 10.1007/978-3-642-27737-5_117-2.

Olivier Renaud and Maria Pia Victoria-Feser. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 2010. ISSN 03783758. doi: 10.1016/j.jspi.2010.01.008.

Philip Sedgwick. Simple linear regression, 2013. ISSN 17561833.

Clifton D. Sutton. Classification and Regression Trees, Bagging, and Boosting, 2004. ISSN 01697161.

Richard S Zemel and Toniann Pitassi. A Gradient-Based Boosting Algorithm for Regression Problems. *NIPS '01*, 2001. ISSN 1049-5258.