New York City Yellow Taxi Trips Dataset

100518243[1]

[1] University of Derby, Kedleston Road DE22 1GB, UK
100518243@unimail.derby.ac.uk

Abstract. Exploring and exploiting the transportation data of New York City's famous yellow cab is a window into the factors that are reason for everything from policy shifts to service upgrades. This document seeks to implement a diverse set of algorithms, both supervised and unsupervised in various contexts to derive insight from the classification of the taxi vendors who supply the services, cluster the location points to their respective boroughs, predict the trip distance for the purpose of better financial planning but also predicting the total fare that could give both passengers and the service providers better outlook for their spending and budgeting purposes.

Keywords: Linear Regression, Naïve Bayes, KMeans, Regression Trees

1 Introduction

The iconic yellow cab taxi is one of, if not the most popular sights in the world. Their daily contribution to moving the residents of New York across the various boroughs moving the money makers to Wall Street to the experience of Chinatown. Given this vast transport network, vast amounts of data are generated across the many trips that are taken by the residents of New York and as such makes for a fascinating investigation into what insights can be gleamed from modern transportation data together with the economics of ride hailing networks.

This report will leverage data from the NYC Taxi and Limousine Commission[17] that provides data record information into yellow taxi trips across New York City from 2009 until present. **Table 1** shows an overview of the dataset and its fields prior to any pre-processing and analysis being done in relation to this report.

The aim of this report is to investigate the economics of ride hailing against various factors. The objectives of this report include:

- To perform classification on the vendorID.
- Perform cluster analysis on the trip pickup and drop-off locations.
- To predict the total amount of a trip based on the predictors available.
- To predict the trip distance based on the predictors available.

The algorithms selected for this report that will allow the meeting of the above-mentioned objectives will include Decision Tree Regressor, Linear Regression, Naïve Bayes Classifier and K-means Clustering.

2 Algorithm literature review

2.1 Regression Tree

The regression tree forms a part of the tree-based methods, a set of supervised learning methods that operate on the basis of segmenting the predictor space into a set of simplified regions.[4] They are able to do so by applying binary conditional splits where appropriate and in doing so are capable to provide a clear distinct explanation given their tree-like shape of what conditions caused the split. This method is termed as recursive binary splitting and it grows the tree at it partition only ceasing growth when it terminates at a condition where each terminal node has fewer than some minimum number of observations.[4]

Given that they are meant to be approached for the prediction of a continuous variable as opposed to their classification instance, their supervised method makes for quick training and an even easier prediction and application.[12]

In the article [8], they follow the implementation with a boosted version of the regression tree, a gradient boosted decision tree algorithm, in the bid to predict a closer to accurate driving range for electric cars.

2.2 K-Means Clustering

The K-means clustering algorithm is widely acknowledged as one of the most popular unsupervised methods. It enables researchers and scientists to observe and discover the structure within the data as opposed to actually predicting the variable. It is prevalent in its role involving market segmentation, image segmentation[2], outlier detection[9] as well as geographical based cluster analysis in [7]. This application of K-Means within this document's analysis follows along the hard-clustering principle[13] that establishes that each data point or object belongs completely to one cluster or it doesn't. This enables us to properly establish and assign the pickup and drop off location points to the right RateCodeID.

The resulting Kmeans function after fitting is usually explainable through a few significant values that are able to define not just how its centroids are located but also how the separation is made between the clusters. The *tot.withinss* is defined as the total within cluster sum of squares which is the sum of cohesiveness within the items of each cluster, also referred to as the intercluster similarity. The *betweenss* is defined as the between sum of squares which is the sum of cohesiveness between each of the individual clusters, also referred to as intracluster similarity. The key golden rule to any clustering is to always have a lower tot.withinss but a higher betweenss.[13]

2.3 Naïve Bayes Classification

The naïve bayes follows the Bayes theorem that considers the independence of each predictor variable in the model without going on the possible correlations and relationships that consider many predictor variables not independent of each other.

The Bayes theorem operates on conditional probability, expounded as the probability of an event occurring on the based on information about events in the past.[11] It may be denoted as follows;

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

In which P(A|B) is the conditional probability of event A occurring given that event B, P(A) is the probability of event A occurring, P(B) is the probability of event B occurring and P(B|A) is the conditional probability of event B occurring given event A.[11]

These classifiers are popular for their simplicity as well as ability to handle large data for multi-class prediction problems with minor modification. However, they are held back by their insistence to assume all predictors are independent which in turn limits the algorithms practical application to real-world scenarios.[16] They have a strong industrial application in text classification[5], spam filtering and sentiment analysis[10] among others.

Given that this algorithm is a classifier implementation, aside from accuracy, the key measures of a well implemented classifier are described by the *sensitivity* and *specificity*. Sensitivity (true positive rate) is the probability of attaining positive classification given the true outcome is similarly positive. The latter, (true negative) is the probability of getting a negative classification when the outcome is also negative.[6]

The naïve bayes classification was approached within this study as a binary classifier purposefully to categorize the trips with their respective vendors. The benefit here being that they would be capable of identifying trips even with a possibility of some missing values.

2.4 Linear Regression

The use of linear regression is admittedly common within the data science industry as they excel strongly where the response variable is quantitative or continuous in nature and are able to accept both qualitative and quantitative predictors to influence the outcome of the linear model. The generalized linear model can be denoted as shown below;

$$Y = \beta 1 + \beta 2X + \epsilon \quad (2)$$

where $\beta 1$ is the intercept and $\beta 2$ is the slope, under a collective name of regression coefficients, ϵ is known as the error term. Together, they are the predictors that describe Y.[14]

Their merits stem not just from having their outputs explained as probabilistic interpretations but as well as having them regularize their data to avoid any overfitting. Despite such strong advantages, they have a weakness in their inflexibility to capture more complex relationships.[15]

Measures of importance are the *Adjusted R-squared* and *Multiple R-squared* metrics that make it possible explain to the proportion of variation in the response variable that has been explained by the model and in doing so the higher their value the better the model was at explaining the response.[14]

In the context of this document, the linear regression was useful in the prediction of the total amount a trip would cost given several other predictors. Similar implementation as to the use of linear regression is reported in this paper,[3] that investigates modeling taxi demand such that with the right information, planners are capable of reducing driver downtime and improving transportation systems.

3 Algorithm application

3.1 Regression Tree

The regression tree was made use of in this dataset under the case of predicting the trip distance against several other predictors. This use case makes for a supervised learning problem where we already have data that has been tagged i.e. labelled and draw on that to make predictions on unforeseen data.

It is important to discern the trip distance as it relates to other factors as it can enable the taxi operators to be knowledgeable into how far trips are going to better give drivers forewarning of the times to work that can allow them to earn more money.

Using the *library(tree),* we have access to the Decision Tree Regressor used for this analysis. After splitting the dataset into train and test sets, they are fitted into the tree model. Due to the nature of the regression tree, we are able to discern that of all the predictors, only "total_amount", "fare_amount" & "ride_duration" are shown to have played a key role in the construction of the tree. This is illustrated with **Fig. 1**. Predicting the trip distance using this tree model provided an accuracy at just over **90.12%.**

3.2 K-Means Clustering

K-Means clustering is an unsupervised learning method meaning that has the ability to group together objects that reflect a strength of similarity to object with a closer shared likeness. The dataset provided variables denoted as RateCodeID which was the final code in effect at the end of the trip, PULocationID where the taximeter was engaged and DOLocationID where the taximeter was disengaged. All these variables relate to the geographical information of the trip that was taken and as such made them the best candidates to cluster such that one is able to understand where trips commencing and concluding. This could prove beneficial to the taxi companies as a tool for market segmentation to better provide specific services or improvements in locations that are served accordingly.

The kmeans function comes in-built into R therefore not requiring any installation. The data used for the clustering was a selected subset of the whole dataset to only include the above mentioned variables. The kmeans algorithm was applied with the subset data as the input data source and the clusters being set to **5**, to match the distinct rate codes available in the dataset.
The investigation into the structure of the resulting kmeans object reveals that a *tot.withinss* of **1.09e+08** and a *betweenss* of **3.97e+08**. The importance of these two values is what ultimately describe a good clustering where a lower value of *withinss* and higher value of *betweenss* are crucial. The **Fig. 2** is able to illustrate the effective manner in which the clustering was applied and makes it possible to discern at the cross points of each point, a trip with its respective pickup and drop-off location ID as well as the respective RateCodeID that was applied at the end of the trip.

3.3 Naïve Bayes Classification

The naïve bayes classification is a technique that works on the foundational basis of Bayes' theorem where an assumption on the independence of each predictor is not dependent on the presence of any other features.

The use of naïve bayes was brought on by the installation of the *library(e1071)*. The key value metric for this classification was to be capable of assigning a taxi vendor to a certain trip. This can prove beneficial in the circumstance where the taximeter might be offline for whatever reason or a data error in recording the trip. This classifier makes it able to group trips using the predictors into a class and it was done as a binary classification as there are only two distinct yellow taximeter vendors and as such only two distinct classes.

Similar to the previous supervised learning method, decision tree regressor, the data for this analysis was split into train and test data for training and later prediction to measure the accuracy of the classification model.

Following the model fitting, prediction was carried out on the test dataset and a confusion matrix from the *caret* package was used to study the efficiency of the model. The naïve bayes model was able to outcome an accuracy at just over **91.25%** with a sensitivity at **0.9** and specificity of **0.9199**.

3.4 Linear Regression

The linear regression approach is useful for its ability to identify the relationship between predictor variables and the response as well as providing an equation to mathematically predict response variable from the predictors.[6]

The linear regression model is particularly useful when predicting continuous values as is the case for its implementation in this analysis. The total amount of a trip is one of the most valuable data variables in the dataset as it is the one most likely to be influenced by most of the other predictors of the dataset ranging from the fare_amount, trip_distance, pickup_hour and pickup_day_of_the_week. It is to that end this regression makes it possible to study the relationship between these predictors but also to generate a linear equation that makes it possible for the taxi service providers to make

use of the predictions for a range of other services such as providing estimated cost to customers inside a taxi app as evidenced by the project[1].

The linear model approach followed similar supervised learning methods with splitting data into train and test data and then fitting the data to a linear model.

The summary generated from the linear model fitting returns a number of metrics that enable the understanding into the model. The most important of which are the Multiple R-squared at **0.9996** & Adjusted R-squared at **0.9995**.

The test data was applied in the prediction and the model was able to correctly predict the total amount of a trip given the majority of predictors at **99.97%.** This accuracy is quite exceptional for a simple linear model. In order to fully understand the impact of the accuracy, the RMSE (Root Mean Squared Error) and Mean Absolute Error (MAE) were extracted using the *caret* package and reported back at **0.279** and **0.155**, respectively. The understanding behind this is that when predicting the total amount a passenger is likely to pay for a trip, the cost will differ between **0.15** and **0.28** cents. The mark of a good linear model is a low RMSE and MAE. The prediction from the linear model is illustrated by **Fig. 3**.

4 Conclusion

Through the extensive exploration and evaluation taken to understand and analyse the New York City Yellow Taxi Dataset, there are several key takeaways that can be garnered from this study but also improvements that could be incorporated into the analysis with different choice of data as well as choice of algorithm to tackle different tasks.

The linear regression model given its great accuracy and power could be need review to investigate any overfitting as well as normalizing any data used in its prediction to cast any doubts of its results. The KMeans cluster was a revelation into the very correlative relationship between the pickup and drop off location IDs and with information such as the longitude and latitude data it would be further enhanced to display a richer visualization. The classification of the VendorIDs aided by the Naïve Bayes classifier provides the ability for the VendorIDs data to borrow the strengths from the market-basket analysis that the Bayes classifier is so popular for. Finally, regression tree and its application with the trip distance was able to raise the fundamental areas, used in its fitting, that can be investigated when improving the taxi services.

5 References

1. Combessie, G.N.& A.: Taxi Cab Fare Prediction Machine Learning in Real Time, https://blog.dataiku.com/predicting-taxi-fares-in-new-york-using-machine-learning-in-real-time, last accessed 2021/01/13.
2. Dhanachandra, N. et al.: Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. Procedia Comput. Sci. 54, 764–771 (2015). https://doi.org/10.1016/j.procs.2015.06.090.

3. Faghih, S. et al.: Taxi and Mobility: Modeling Taxi Demand Using ARMA and Linear Regression. Procedia Comput. Sci. 177, 186–195 (2020). https://doi.org/10.1016/j.procs.2020.10.027.
4. Gareth James et al.: An Introduction to Statistical Learning with Applications in R. Springer International Publishing (2015).
5. Kolluri, J., Razia, S.: Text classification using Naïve Bayes classifier. Mater. Today Proc. (2020). https://doi.org/10.1016/j.matpr.2020.10.058.
6. Robert I. Kabacoff: R in Action: Data Analysis and Graphics with R. Manning Publications (2015).
7. Sanjanamohapatra 99: Clustering Taxi Geolocation Data To Predict Location of Taxi Service Stations- Answering Important…, https://medium.com/analytics-vidhya/clustering-taxi-geolocation-data-to-predict-location-of-taxi-service-stations-answering-important-82535ed9bf57, last accessed 2021/01/13.
8. Sun, S. et al.: A Machine Learning Method for Predicting Driving Range of Battery Electric Vehicles, https://www.hindawi.com/journals/jat/2019/4109148/, last accessed 2021/01/13. https://doi.org/10.1155/2019/4109148.
9. Wei, Y. et al.: MSD-Kmeans: A Novel Algorithm for Efficient Detection of Global and Local Outliers. ArXiv191006588 Cs Stat. (2019).
10. Xu, F. et al.: E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. Inf. Process. Manag. 57, 5, 102221 (2020). https://doi.org/10.1016/j.ipm.2020.102221.
11. A Step By Step Guide To Implement Naive Bayes In R, https://www.edureka.co/blog/naive-bayes-in-r/, last accessed 2021/01/13.
12. Decision Trees in R, https://www.datacamp.com/community/tutorials/decision-trees-R, last accessed 2021/01/13.
13. K-Means Clustering in R Tutorial, https://www.datacamp.com/community/tutorials/k-means-clustering-r, last accessed 2021/01/13.
14. Linear Regression With R, https://r-statistics.co/Linear-Regression.html, last accessed 2021/01/13.
15. ML - Advantages and Disadvantages of Linear Regression, https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/, last accessed 2021/01/13.
16. Naive Bayes Classification using Scikit-learn, https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn, last accessed 2021/01/13.
17. TLC Trip Record Data - TLC, https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, last accessed 2020/12/09.

6 Appendix

Name	Data Type	Values (sample)
doLocationId	string	161, 236
Extra	double	0.5,1.0

Fare_Amount	double	6.5, 4.5
improvement_surcharge	string	0.3, 0
MTA_tax	double	0.5, -0.5
Passenger_count	int	1, 2
Payment_type	string	CSH, CRD
puLocationId	string	237, 161
rateCodeId	int	1, 2
storeAnd FwdFlag	string	N, Y
tipAmount	double	1.0, 2.0
tollsAmount	double	5.33, 4.8
totalAmount	double	7.0, 7.8
tpepDropoffDateTime	timestamp	2013-11-0301:25:00 2013-11-03 01:26:00
tpepPickupDateTime	timestamp	2013-11-0301:15:00 2013-11-03 01:14:00
tripDistance	double	1.0, 0.9
vendorID	string	1, 2

Table 1. A data dictionary showcasing the fields and datatypes of the dataset prior to preprocessing.

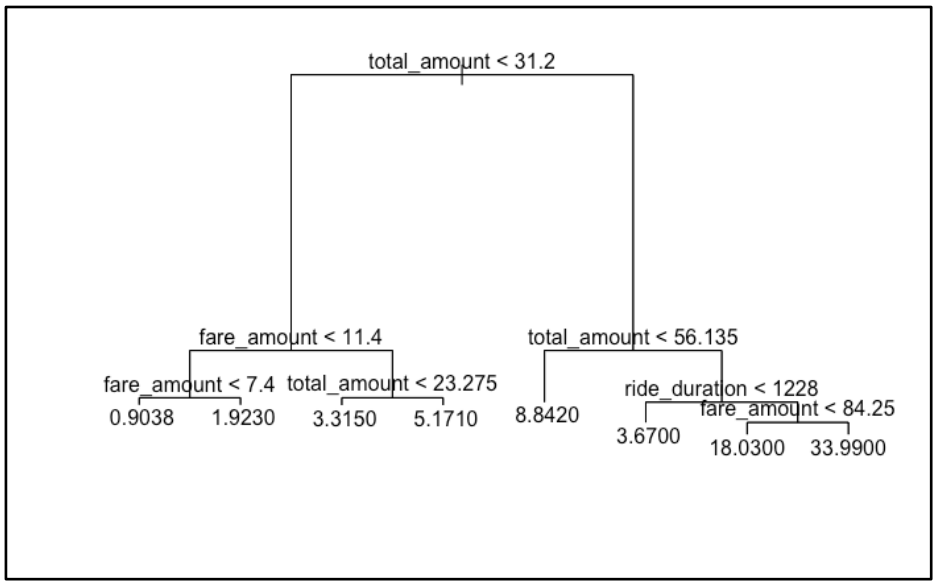

Fig. 1. Decision Tree Regressor plotted.

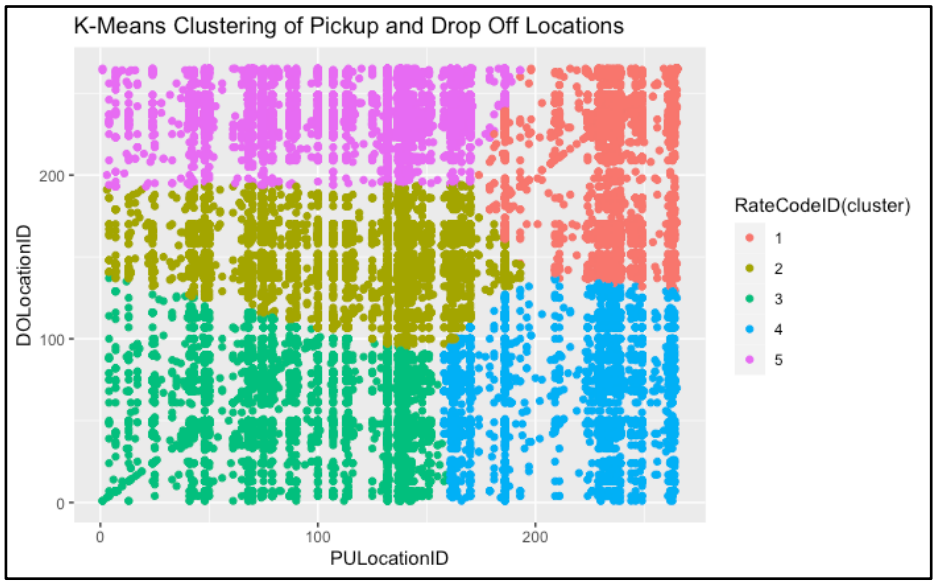

Fig. 2. KMeans Cluster showing the pickup and drop-off locations clustered into their respective RateCodeID clusters.

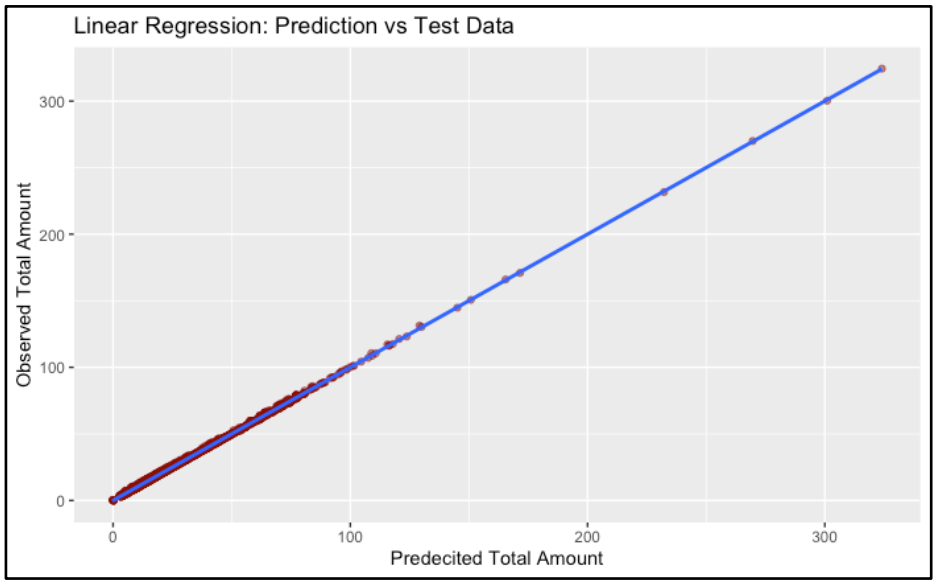

Fig. 3. A plotted linear regression model showing the actual total amount against the predicted total amount.