Yellow Taxi Fare Prediction

Rahul Parmar[1], Murli Alva[2], Pathik Patel[3], and Vatsal Suthar[4]

[1]rahul.p1@ahduni.edu.in
[2]murli.a@ahduni.edu.in
[3]pathik.p@ahduni.edu.in
[4]vatsal.s3@ahduni.edu.in

Abstract

Taxis are preferred as one of the favorite ride in late 19th century in new york. Mostly preople used to prefer taxis over their personal vehicle. There are two types of taxis yellow and green. They are selected according to the location. fares vary according to location and time. In this project we are going to predict taxi fare and classify the areas where most taxis are booked.The yellow taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.in this project go for taxi price prediction and its accuracy towards future data.

Keywords :
Yellow taxi,New York,Machine learning,analysis, Regression,clustering,Sklearn library,python

1 Introduction

In NYC, taxicabs come in two varieties: yellow and green; they are widely recognizable symbols of the city. Taxis painted yellow (medallion taxis) are able to pick up passengers anywhere in the five boroughs. in Upper Manhattan, the Bronx, Brooklyn, Queens. The yellow taxi cab was first introduced in 1915 by a car salesman named John Hertz. Hertz decided to paint his taxis yellow because of a study by a Chicago university to establish what colour would grab the attention of passers-by more easily. The results proved that yellow with a touch of red was most noticeable. As a result, Hertz started to paint all his taxi cabs yellow and went on to start the Chicago based Yellow Cab Company in 1915.

Trips made by New York City's iconic yellow taxis have been recorded and provided to the TLC since 2009. Yellow taxis are traditionally hailed by signaling to a driver who is on duty and seeking a passenger (street hail), but now they may also be hailed using an e-hail app like Curb or Arro. Yellow taxis are the only vehicles permitted to respond to a street hail from a passenger in all five boroughs. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The records were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology service providers. The trip data was not created by the TLC, and TLC cannot guarantee their accuracy.

As technology continues to push the bar for information, one area in NYC that's lagging behind is its iconic Yellow Cab. With Uber, Lyft, Via, and other ride-hailing apps setting the pace, Yellow Cab's has teamed up with Google to become more data-centric. Yellow cab has asked Google to figure out how to predict the estimated fare amount by using a few features to determine the expected fare amount. The information that Google has obtained from yellow cab are past taxi rides that include:

Pickup Time ,Date Pickup Latitude,Longitude Dropoff Latitude,Longitude Passenger Count, Fare Amount. Yellow Cab, Google, and Kaggle have teamed up together in creating a playground type competition to allow people to view this dataset and create a machine learning algorithm that predicts the expected fare amount. Our goal is to analyze the dataset, manipulate, and create functions

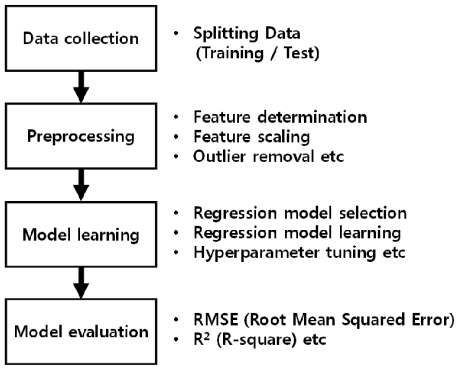

Figure 1: Flowchart The flow work of our method

to allow new data run through our code and machine learning algorithm to eventually offer an expected fare price.

2 Literature Survey

1. A case study: The New York City yellow cab System of Systems

Systems of systems engineering (SoSE) is "the process of planning, analyzing, organizing, and integrating the capabilities of a mix of existing and new systems into a system-of-systems capability that is greater than the sum of the capabilities of the constituent parts". The yellow cab case is a well documented example of a System of Systems, which consists of multiple integrated complex systems working together to achieve the objective of transporting passengers in New York City in a safe and efficient manner.

2. Exploring the Taxi and Uber Demand in New York City: An Empirical Analysis and Spatial Modeling

In this paper the increase in demand of taxis in New York city is explained in detail. It states that Uber, Yellow taxi, Green taxi demands have increased drastically from 2014 to 2015. The highest gainner being Uber(203

3 Implementation

In this project we are planning to predict fare using ML algorithm i.e., regression and to predict that from which area is most cabs are booked with the help of ML algorithm i.e., clustering on yellow taxi cab of New York city. During analysis we used different parameters to check which area has more cab booking, average number of passengers for each trip, average fare for the trip, and many more.... Etc.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Figure 2: Data The data and description of every attribute

3.1 Data Collection

yellow taxi cab of New York city.the dataset which has October 2020 month yellow taxi cab data `https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf`

3.2 Pre-processing

In the initial phase while analysing the data we have got some useful insight such as most cabs trips are with 1 passengers or 2 passengers. To sum up in the pandas library there are no duplicate entries in the dataset, performed the correlation of the columns in the dataset ,visualized the passenger's count. We merged all extra amount columns into the final amount. We have split the date and time into different columns. We have performed EDA on the dataset.

In the next step we compare our main attribute which is taxi price compares with its most correlated attribute trip distance , when seen there are lots of noisy value present and many outlier , so we processing mean-standard deviation method to removal outlier , after removing outlier we compare both attribute and its seems good like wise check for all and done for all

Last but not the least we standardize all the value to maintain value's variance and standard-deviation.

3.3 Model learning

We are doing feature engineering before train

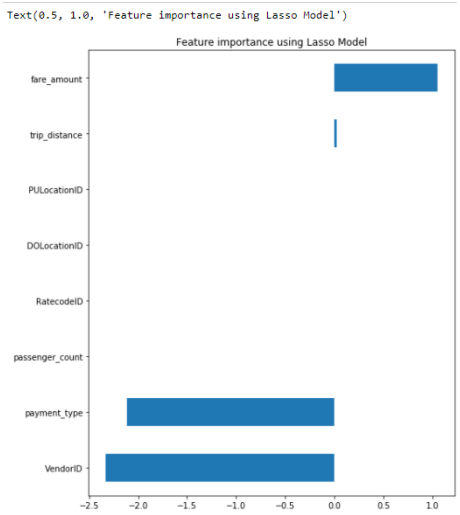

Figure 3: ALL Attribute of data evaluation

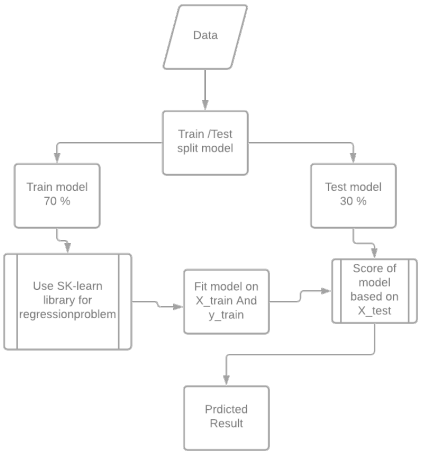

Figure 4: Flowchart The flow work of regression mode

our model, Extract information from date-time (day of week, month, hour, day). Taxi fares change during different hours of the day and on weekdays/weekends/holidays. Now that we have a clean dataset, we are ready to train a model for predicting taxi fare. Before we do that through, we will split the dataset into a train (80) percent and test (20) percent.

The model choose first is multivariate linear regression and according to that we doing feature selection for that we run testing our dataset to prove correlation between attribute like lasso and ridge regression and statistical analysis , PCA and than we find 8 best column for our model's independent variable and our final amount or fare price of taxi is targeted variable ,and then with the help of SK-learn library we choose linear regression and run on our x and y train and check accuracy for x and y test data so accuracy for linear regression is 83 percentage its good but it's still need some improvement.

The figure work flow of liner regression model The model choose next for analysis is that random forest Regressor or tree search Regressor and accuracy of the both is respectively 91 and 87 percentage its looking awesome for our data but need some more accurate accuracy in this case so we need find more model

The model I chose last is for this particular dataset was XGBoost. Another alternative would be LightGBM which many others have used as well in the project. The parameters are already preset in my model from hyperparameter tuning using grid search,and train the model and accuracy was awesome it's 93 present and getting needed accuracy.

3.4 Model Evaluation

In machine learning model first created linear regression on that we create model and fit it and check accuracy to check hoe model is ac curated or not. regression model lr-clf on LinearRegression and we get 83 parentage accuracy on our test data. Then use in model evaluation greed search cv which is train and test data with best parameter in various regression like linear regression , lasso , tree search regressor its accuracy is as following 87 , 66 , 88 percent respectively

rand-forest-regressor on RandomForest Regressorr and-forest-regressor is fit on X-train, y-train this code is use evaluate the train and test model based on randomforest regressor name as suggested algorithm works randomly in forest because its choose random data train model on it.and after use XGBoost to train model and to parameter training and getting 93 percent accuracy.

4 Results

Result of the project In the initial phase while analysing the data we have got some useful insight such as most cabs trips are with 1 passengers or 2 passengers, upper east side north of Manhattan has more bookings for the yellow taxi cab. To sum up in the pandas library there are no duplicate entries in the dataset, performed the correlation of the columns in the dataset ,visualized the passenger's count. We

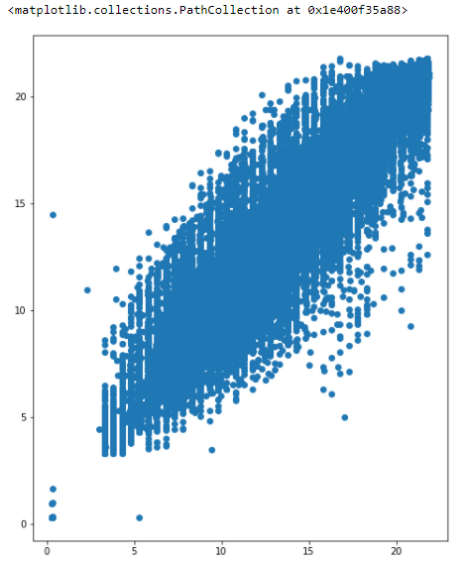

Figure 5: Scatter plot between Y-predicted and Y-target value

merged all extra amount columns into the final amount. We have split the date and time into different columns. We have performed EDA on the dataset.its seems amazing model train and after evaluating the data predict good its very accurate to predication, get value of the taxi fare price based on the trip distance , vendor id , pickup and drop off location id, payment type and fare amount.The result of prediction y' and y also get good result and its correlated and the graph of the y vs y' is looking good.

Conclusions

The New York City Yellow Cab SoS is a good example of a System of Systems. The objective of the TLC is to govern the SoS to ensure safe and efficient transportation of passengers in New York City. This helped to identify lessons learned based on which, best practices that could be extracted to other SoS have been established. This case also raises questions for discussion which can be applied to other case studies and be used as a benchmark for other cases. It is an initial condition for modeling and simulation of case studies overall. We show that the regression predictor does not always have better prediction accuracy than the random forest or XGboost. In the areas with low predictability, On the other hand, we find that every model make high predicative on the considering some more uncorrelated attribute.that, with high correlated attribute the prediction score is low not accurate very much.

Future work is in this project is necessary the data so many idea about location in the future doing work on geographical data make system which describe throw geographical data and predict fare price.

Acknowledgements

The authors thank: the New York City TLC for providing the data used in this paper;Special Thanks To,: Pro. Mehul Raval from ahmedabad University for his feedback on an early draft of this paper and project; and the anonymous reviewers for their insightful comments and suggestions.

References

[1] TLC NEW YORK, "New York City TLC," *IEEE Trans. DATA PAGE & Tech.,https://www1.nyc.gov/site/tlc/about/data-and-research.page*.

[2] Jimmy Gandhi California State University, Northridge , Alex Gorod University of Adelaide , "A case study: The New York City yellow cab System of Systems", *https://www1.nyc.gov/site/tlc/about/data-and-research.page* ,June 2011.

[3] Denis Khryashchev, Northridge,"A case study: Predicting taxi demand at high spatial resolution", *https://www.researchgate.net/publication/252023141_A_case_study_The_New_York_City_yellow_cab_System_of_Systems* ,Dec 2016.

[4] Coding nest, , "Project - New York Taxi Fare Prediction | Machine Learning", *https://www.youtube.com/watch?v=lJitd0ke2bs* ,Jul 2020.

[5] towards data science,"New York Taxi data set analysis Predicting taxi fare using Regression models", *https://towardsdatascience.com/new-york-taxi-data-set-\analysis-7f3a9ad84850* ,May 2019.

[6] towards data science, ,"Getting Started with XGBoost in scikit-learn",*https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost\-algorithm-long-she-may-rein-edd9f99be63d* ,Nov 2020.

.