Taxi fare prediction

Raghav K
Student, B.Tech, Computer Science and Engineering
Vellore Institute of Technology,
Chennai, Tamil Nadu, India
raghav.k2020@vitstudent.ac.in

Keerthana V
Student, B.Tech, Computer Science and Engineering
Vellore Institute of Technology,
Chennai, Tamil Nadu, India
keerthana.v2020a@vitstudent.ac.in

Abstract—More and more people are using taxi services nowadays. Services employ dynamic pricing to balance supply and demand in an effort to improve service quality, benefiting both drivers and consumers. However, dynamic pricing sometimes causes issues for travelers since "unpredictable" pricing frequently makes it difficult for them to make quick selections. It is crucial to provide passengers with additional information in order to solve this issue, and projecting dynamic costs is a workable fix. In this study, we concentrate on the estimate of dynamic pricing, projecting the price for each unique passenger order. Passengers' anxieties will be allayed by price prediction, which will enable them to determine if they may find a cheaper price at close-by locales or within a short amount of time. The prediction is made by figuring out the link between dynamic pricing and attributes taken from the dataset. We train a single linear model as a representative and evaluate its results using actual service data from multiple angles. Additionally, we analyze the contribution of model-based characteristics at various levels to determine which factors influence dynamic pricing the most. Finally, based on the findings of the assessment, we forecast dynamic pricing using an effective decision tree, random forest and principal component analysis model. We anticipate that the research will contribute to happy travelers by providing an accurate prognosis.

Keywords—component, formatting, style, styling, insert (key words**)**

I. Introduction

Taxi ridesharing provides a solution to one of the main issues with travel and transportation. A recent study found that the majority of customers preferred taxi ridesharing. The two most popular methods for consumers to utilise ridesharing are to either arrange a trip using an app or hire a cab to use the services offline. The suggested system searches for routes that fulfil certain travel requirements. The term "stability of route" here refers to how often a taxi driver has to reroute or transfer. It may also be thought of as the likelihood that an unforeseen road condition would arise and effect how much time and how far the driver and the passenger need to go.

The suggested approach takes into account both origin and destination. Additionally, the existing systems return a genuine cab as soon as one is located rather than looking for the best one. Such incomplete information-based systems miss both the possible best matches with the lowest cost and fail to filter out the first irrelevant taxis.

By incorporating RSP, the existing system addresses HDFS (Hadoop Distributed File System) issues (Random Sample Partitioning). The HDFS findings are more susceptible to mistake. Because HDFS does not preserve the data's attributes, this further creates another flaw in the system. By permitting pickup location id and drop off location id ridesharing, taking into account passengers on the same route, dynamically updating requests, and indexing a taxi that supports the best route, the suggested system resolves the technique of the present ridesharing services.

II. PROBLEM STATEMENT

You are a newly established taxi firm. After a successful pilot project, you want to expand your taxi service across the nation. After gathering historical data for your pilot project, you are now required to use analytics to estimate fares. You must create a system that can estimate the cost of a city taxi journey.

Dataset: https://www.kaggle.com/code/neilclack/taxi-trip-cleaning-analysis/data

Attributes:
trip_distance- Total distance took from source to destination
pickup_location_id- id number of the pickup location
dropoff_location_id – id number of the drop location
year, month, day, day_of_week,, hour_of_day – feature extracted variables to predict the fare amount accurately depending on various situations.
trip_duration- total duration of the trip
passenger_count- number of passengers travelling in that particular cab

Target:
Fare amount: Cost of the taxi ride. This value is only in the training set; this is what we are predicting in the test set.

III. RELATED WORK

In present days, there is a practise of joining another passenger's journey after abandoning one's own while still considering the vehicle's capacity.This is called as Slugging. Even though this could seem to be quite close to our suggested approach, only one pickup or delivery site is taken into account. An improved pickup and delivery site for passengers is planned for our system. Here, it would seem more practical to choose a spot that is close to both passengers.

The driver may provide additional limits such as his or her own start and departure times and the quantity of available ride-sharing seats in addition to only the pickup and destination constraints. Therefore, instead of merely sharing the normal limits when the suggested system locates a driver, his or her information (driver's schedule) would also be communicated. In this manner, the system will not choose the specified driver for the trip if the requested trip exceeds the driver's established timetable.

This paper presents a perspective where decentralisation is the main objective. Instead than relying on a central system to link all the nodes in the system, each node in the system solely executes local activities. For a ride, each node is paired with other nodes that have a comparable path. The passengers' journey time will be shortened as a result. Although this technique seems to operate with seamless efficiency, there is a substantial likelihood of endless node grouping and ungrouping. Therefore, if a node is rejected once, our algorithm takes that into account and does not group that specific node again.

Another issue that is addressed is choosing a decent route. For the calculation of the route, it takes into account a road network that has previously undergone testing. The priority in this case is a shorter journey while taking into account travel time, danger, and deadline. However, dynamic monitoring of road conditions is not practised in the actual world. Dynamic road conditions are also evaluated in our suggested approach (traffic, sudden road damage).

A huge data set may be represented using the RSP distributed data model as a collection of disconnected data pieces. This enables the system to analyse using that specific RSP block. Although this procedure is rapid and just needs a modest amount of storage, the outcomes are not particularly encouraging. Since the RSP block only represents a tiny portion of the whole dataset, there is a significant risk of overlooking crucial data items. Our model adopts a similar strategy. Instead of RSP, we use Key Feature Extraction. Therefore, it is possible to depict the newly created characteristics as a mixture of the initial set of features.

The prediction of fares over certain time periods may also be done using large amounts of data. A technique to estimate the waiting time for a passenger at a certain time and location using previous taxicab trajectory data, for instance, was reported by Qi et al. [2]. A gradient-boosted tree was utilised by Zhang and Haghani [3] to forecast and enhance the journey duration. Wavelet-based neural networks are utilised by Yingjun et al. [5] to estimate the number of taxicabs, while Moreira et al. [4] employed taxi data to anticipate demand using streaming data. The demand for bus passengers is finally predicted by Zhou et al. [6] using mobile device use.

The advantage of big data may be used in addition to such investigations. For instance, [7] developed an algorithm to determine taxi fares based on massive amounts of data. The major goal is to ascertain the effect of applying a varied fee within various distance ranges on driver profitability. Egan and Jakob [9] produced a marketing strategy for on-demand transportation services that takes into account market processes, while Bai and Wang [8] built a taxicab routing and fee rate estimate by integrating the calculating algorithm into a mobile application.

Fortunately, in recent years, the area of transportation research has seen a significant increase in the number of academics interested in enhancing taxi services. A queuing network solution was suggested by Zhang et al. [10] to reduce the waiting and searching times for customers. This model also takes traffic congestion on urban road connections into account. The approach provided by Jayasooriya and Bandara [1] for calculating the economic consequences of traffic congestion. The productivity of the labour for both private and public motorised transportation is decreased by road traffic congestion, according to the authors. Unfortunately, bus and car/van transit is the root cause of heavy traffic congestion.

Additionally, Wong et al. [11] suggested a strategy to enhance taxi services. To start, a client satisfaction survey on a taxi service was carried out. In order to determine the areas that need to be improved the most in the service quality of urban taxis, an updated linear regression model was created. The present level of a taxi service was then analysed using a six level of service (LOS) instrument. In addition, Ulk et al. [12] provided a cost study to evaluate the costs associated with taxi services on a digital platform (e.g., Uber, Bolt, and Taxify). Thankfully, their research is comparable to this one in that a cost analysis was carried out.

The crucial distinction, however, is that [12] did not take into account the traffic situation for the model study. The model seems to be static and not data-driven. The method is based on a paper-based survey for [11]. Their results are remarkable. But today's data-driven methodologies are more practical than a traditional poll at the analysis stage. Sadly, adopting simply a paper-based survey lengthens the process of gathering and analysing data. To fill in the gaps in the research by [11, 12], as well as the influence of road traffic congestion on the economic cost noted previously, is what inspired us.

The cost measurement techniques described in [1] are also used to determine the cost of a taxi trajectory's journey.

IV. Data Analytics Models

The Random Forest and Decision Tree techniques used will effectively help us to analyze the cut-off fare asked by the taxi aggregators.This allows each driver and passenger to reduce the time it takes to locate one another. Drivers don't have adequate information on where passengers and different cabs should go. As a result, a cab center will coordinate the taxicab fleet and efficiently provide uniform requests to the entire municipality.

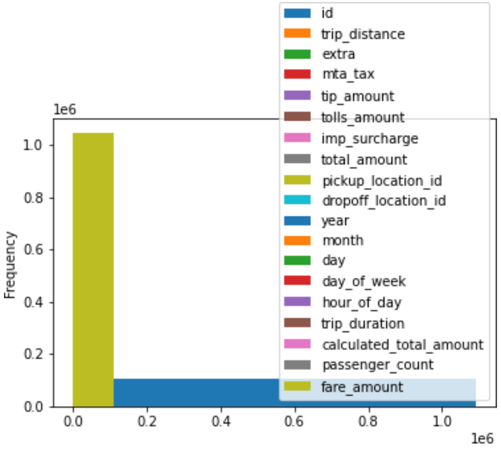
The technique makes use of the scheduled cab's latitude and longitude as well as the day of trip and the month. With this dataset, an unsupervised learning model is developed, and the model is used to predict the pickup of the cab on the cluster. The proposed project method is broken down into six parts namely System Architecture, Raw Data (Dataset), Data Importing, Data Visualization, Testing Data, Predicted Scheduling of cab using algorithm. The latitude will be grouped and categorized from the acquired dataset depending on the frequency of journeys taken by the cab during the day.

When these requirements are met, data preprocessing will be performed on these datasets. Data visualization is described as evaluating the performance of a model using graphs and metrics that compute performance. Data visualization is mostly used to categorize data into new levels so that the technique may be applied to an observation of each output variable produced from an observed input variable.

The cab's scheduling can be predicted based on the user's location, and the proposed method finds the nearest hotspot, which is defined as a cluster of points analyzed using k-means clustering and informs the cab about the hotspot that is closest to the user's location and is booked to pick up the user. The technique utilized is based on the Radom Forest approach, which is associated with unsupervised learning. It builds decision trees on different samples and takes the majority voting for the process of classification K-Means divides things into clusters that share commonalities and are distinct to objects in another cluster The clustering algorithm is categorized into three steps, where 'n' random records are taken from dataset of 'k' records , decision trees is constructed for each of the sample, each of the decision trees generates an output. Thus, the final output will be based on majority voting (classification) or on the average (regression)

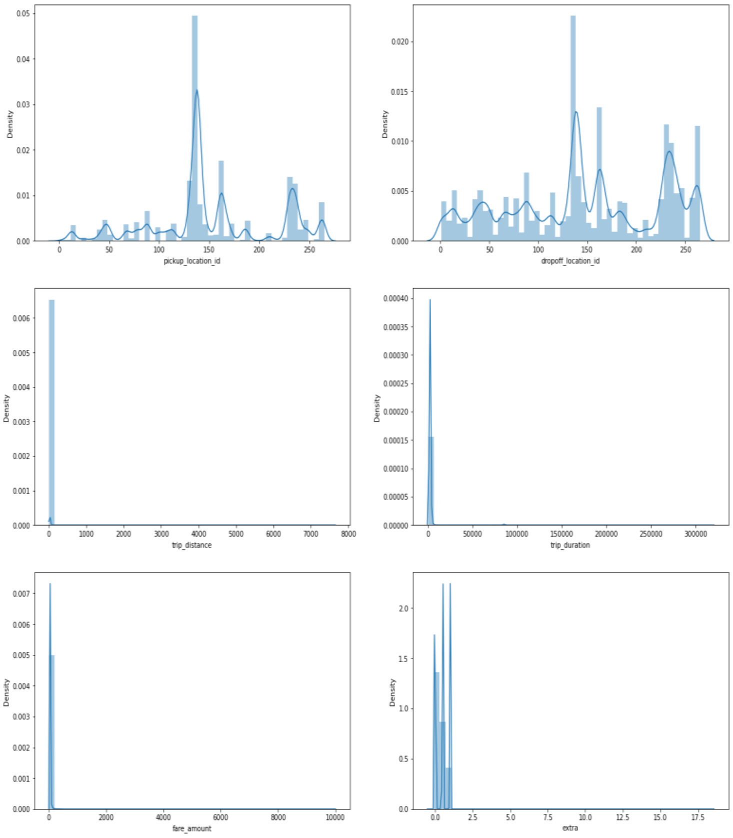
A. Results and Discussions: (Along with codes, figures and graphs)

"data.plot.hist()" is used to draw the histogram corresponding to data frame's columns

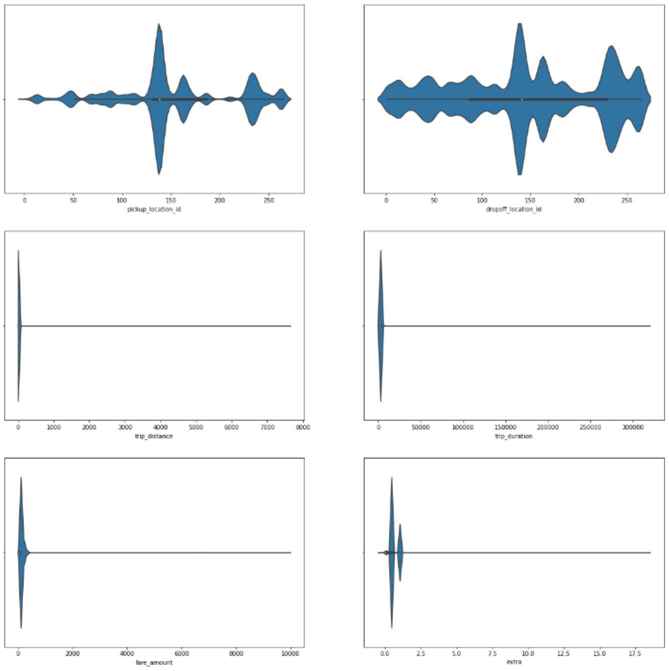


"subplot" helps in plotting multiple plots on a single figure using the following command

plt.figure(figsize = (20, 20))
plt.subplot(321)_ = sns.distplot(data['pickup_location_id'])



"violinplot" is used to observe the numeric distribution of th e dataset as shown in the command below
plt.figure(figsize = (20, 20))
plt.subplot(331)_ = sns.violinplot(data['mta_tax'])

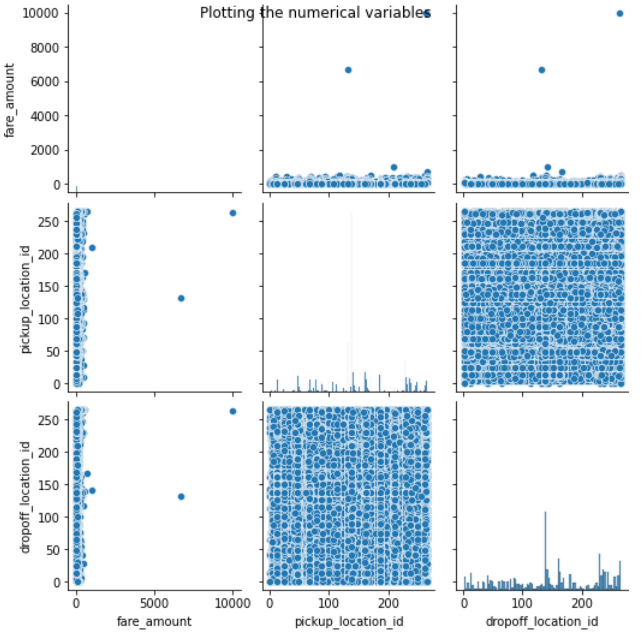


"pairplot" is used to plot multiple bivariate distribution in a given dataset as shown below
variables = ['fare_amount','pickup_location_id','dropoff_loc ation_id']

a = sns.pairplot(data = data[variables], kind = 'scatter', dropna = True)

a.fig.suptitle('Plotting the numerical variables')

plt.show()



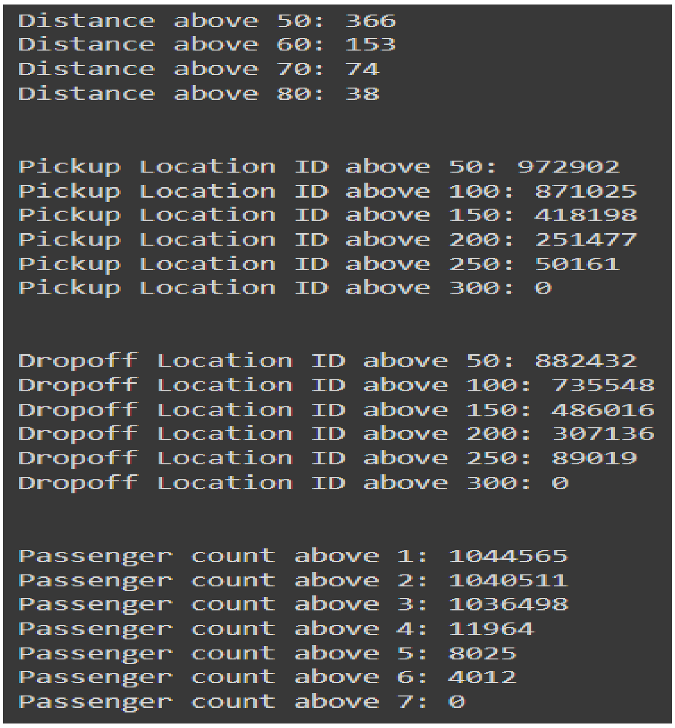
We also check the different statistics of the data namely the trips where distance travelled is greater than a particular value, if passenger count is not greater than 6 and more than 1 as shown in the commands below

print('Distance above 50: {}'.format(sum(data['trip_distance']>50)))

print('Pickup Location ID above 100: {}'.format(sum(data['pickup_location_id']>100)))

print('Dropoff Location ID above 150: {}'.format(sum(data['dropoff_location_id']>150)))

print('Passenger count above 4: {}'.format(sum(data['passenger_count']>4)))

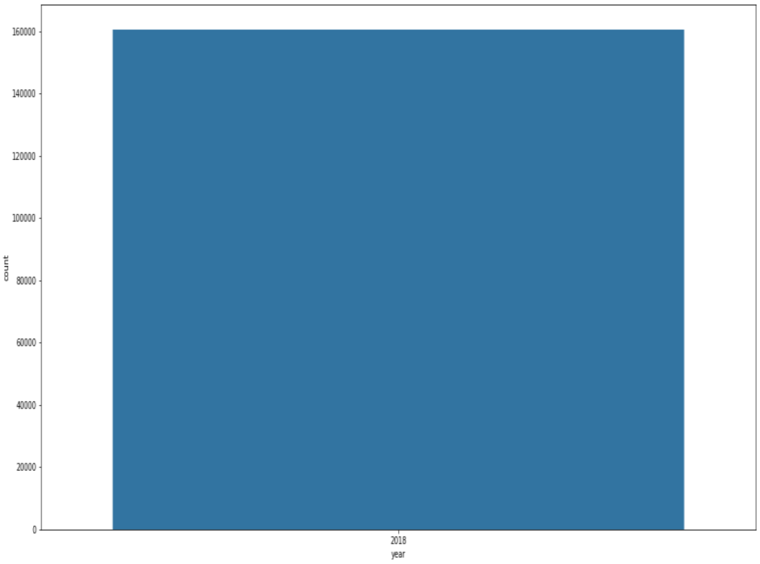


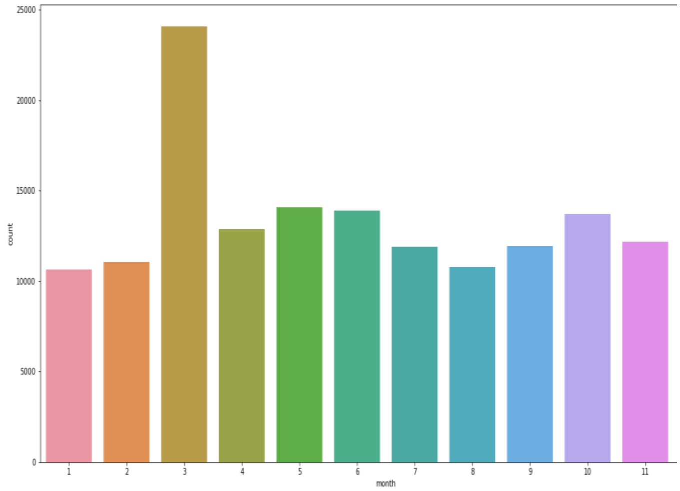
```
Distance above 50: 366
Distance above 60: 153
Distance above 70: 74
Distance above 80: 38


Pickup Location ID above 50: 972902
Pickup Location ID above 100: 871025
Pickup Location ID above 150: 418198
Pickup Location ID above 200: 251477
Pickup Location ID above 250: 50161
Pickup Location ID above 300: 0


Dropoff Location ID above 50: 882432
Dropoff Location ID above 100: 735548
Dropoff Location ID above 150: 486016
Dropoff Location ID above 200: 307136
Dropoff Location ID above 250: 89019
Dropoff Location ID above 300: 0


Passenger count above 1: 1044565
Passenger count above 2: 1040511
Passenger count above 3: 1036498
Passenger count above 4: 11964
Passenger count above 5: 8025
Passenger count above 6: 4012
Passenger count above 7: 0
```

We also work on the feature engineering part for the timestamp variable as shown

plt.figure(figsize = (20, 10))

sns.countplot(data['year'])

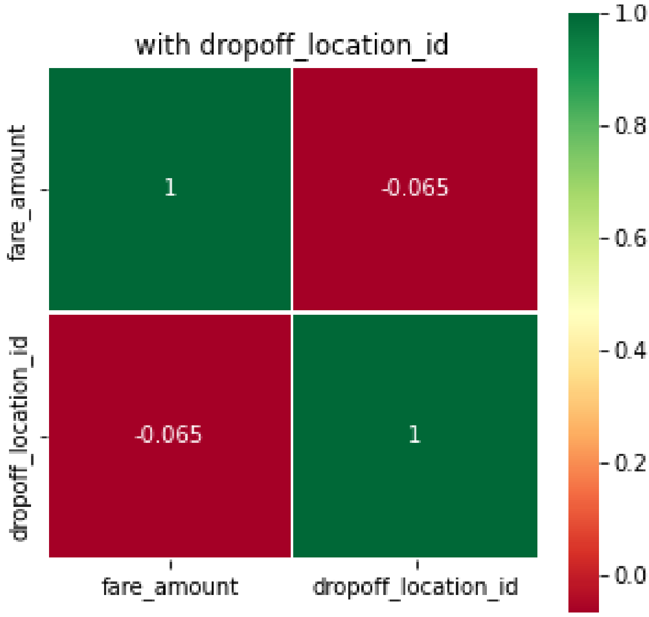


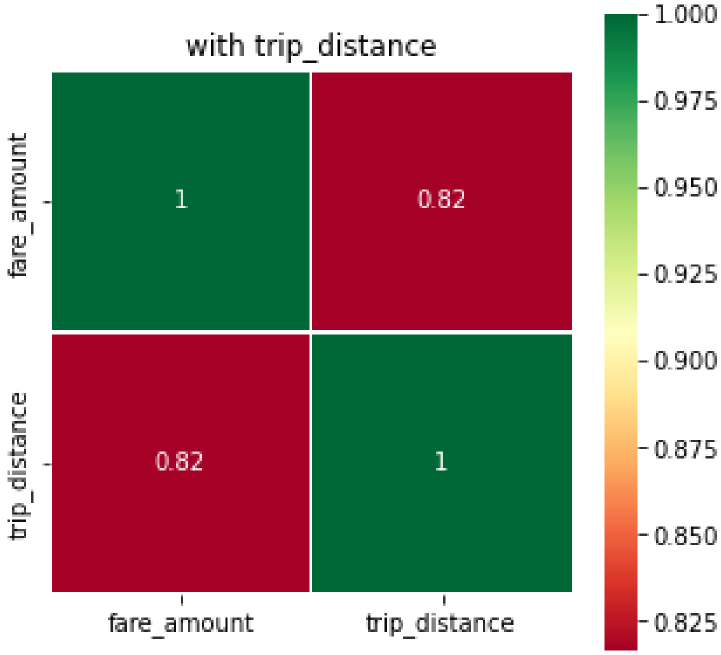


B. Feature Selection

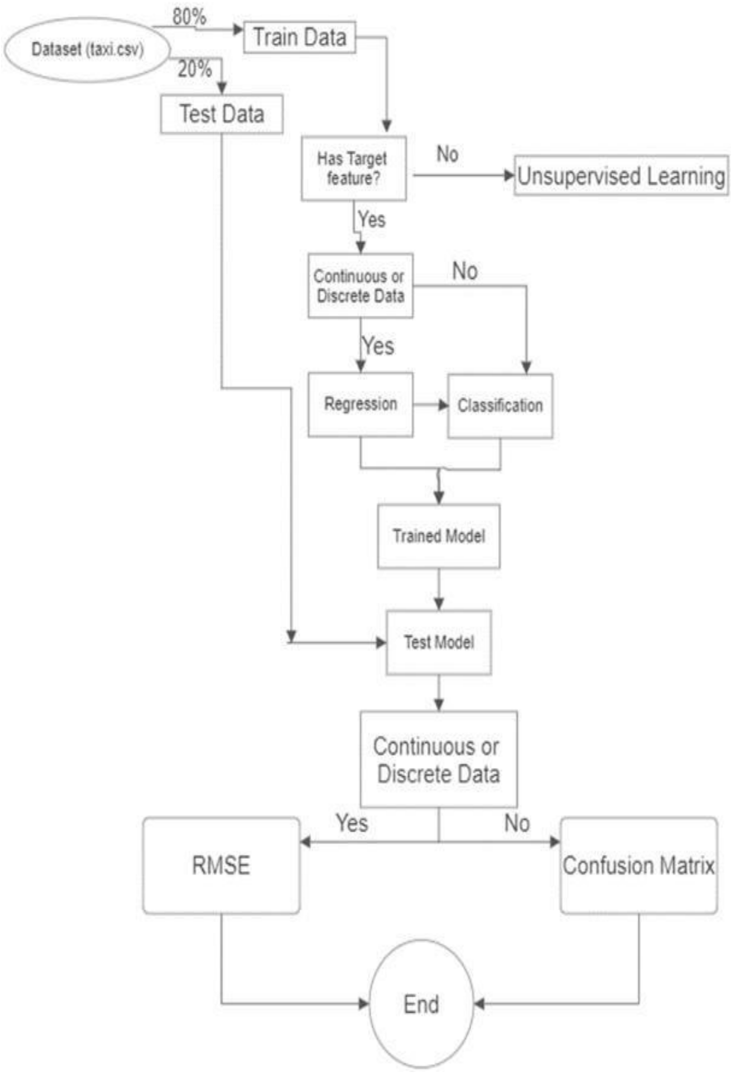
1. Correlation Analysis

Statistically correlated features move together directionally. Linear models assume feature independence.And if features are correlated that could introduce bias into our models





C. Block Diagram



ACKNOWLEDGMENT

We would like to thank our faculty of the course, Machine Learning, CSE4020, Dr. Jothi R who had been a guide and helped us in gaining insights into this project right from the first review of helping us to implement the project of Taxi fare prediction

CONCLUSION

Thus, from the data analysis, we found out that the Random Forest model has a higher accuracy of 91.12% while the Decision Tree model has a accuracy of 90.641%. While, we tried out with other models, we inferred that that the parameters can be more tuned and thus the accuracy score can be enhanced further considering the data standardization, data plotting and thus checking for any null values in the dataset. Thus, we have shown that our data trains and test well on Random Forest algorithm

REFERENCES

1. Jayasooriya SACS, Bandara YMMS. Measuring the economic costs of traffic congestion. In: 3rd international Moratuwa engineering research conference, MERCon 2017; 2017. p. 141–6. https://doi.org/10.1109/MERCon.2017.7980471.
2. Qi G, Pan G, Li S, Wu Z, Zhang D, Sun L, Yang LT. How long a passenger waits for a vacant taxi? Large-scale taxi trace mining for smart cities. In: Proceedings of the 2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing, GreenCom-iThings-CPSCom 2013, vol. 4; 2013. p.1029–36. https://doi.org/10.1109/GreenCom- iThings-CPSCom.2013.175.
3. Zhang Y, Haghani A. A gradient boosting methodto improve travel time prediction. Transp Res C Emerg Technol. 2015;58:308–24. https://doi.org/10.1016/j.trc.2015.02.019.
4. Moreira-Matias L, Gama J, Ferreira M, Mendes-Moreira J, Damas L. Predicting taxi—passenger demand using streaming data. IEEE Trans Intell Transp Syst. 2013;14(3):1393–402. https://doi.org/10.1109/TITS.2013.2262376.
5. Yingjun Y, Cui H, Shaoyang Z, Yingjun Y. A prediction model of the number of taxicabs basedon wavelet neural network. Procedia Environ Sci. 2012;12:1010–6. https://doi.org/10.1016/j.proenv.2012.01.380.
6. Zhou C, Dai P, Wang F, Zhang Z. Predicting the passenger demand on bus services for mobile users.Pervasive Mob Comput. 2016;25(2013):48–66. https://doi.org/10.1016/j.pmcj.2015.10.003.
7. Phiboonbanakit T, Horanont T. How does taxi driver behavior impact their profit? Discerning thereal driving from large scale GPS traces. In: UbiComp 2016 adjunct—proceedings of the 2016ACM international joint conference on pervasive and ubiquitous computing. 2016. https://doi.org/10.1145/2968219.2968417.
8. Bai YW, Wang EW. Design of taxi routing and fare estimation program with re-prediction methodsfor a smart phone. In: 2012 IEEE I2MTC— proceedings of the international instrumentation and measurement technology conference; 2012. p. 716–21. https://doi.org/10.1109/I2MTC.2012.6229165.
9. Egan M, Jakob M. Market mechanism design for profitable on-demand transport services. Transp Res B Methodol. 2016;89:178–95. https://doi.org/10.1016/j.trb.2016.04.020.
10. Zhang W, Honnappa H, Ukkusuri SV. Modeling urban taxi services with e-hailings: a queueing network approach. Transp Res Procedia. 2018;38:751–71. https://doi.org/10.1016/j.trpro.2019.05.039.
11. Wong RCP, Szeto WY. An alternative methodology for evaluating the service quality of urban taxis. Transp Policy. 2018;69:132–40. https://doi.org/10.1016/j.tranpol.2018.05.016.
12. Čulík K, Kalašová A, Otahálová Z. Alternative taxi services and their cost analysis. Transp Res Procedia. 2020;44:240–7. https://doi.org/10.1016/j.trpro.2020.02.047.