|  |  |
| --- | --- |
|  | **Taxi Fare prediction** |
|  | Project done by :  R.V.S.R.S.Dheeraj  Shaik.Salma Sulthana  D.Manmadha Rao  D.Ramachandra Reddy  Asodi.Sri Lakshmi Vamsi Kalyan Reddy |

|  |
| --- |
| **[Final project report]** |

CONTENT PAGE

|  |  |  |
| --- | --- | --- |
| s.no | Chapters | Pg.no |
| 1.1 | Introduction | 2 |
| 1.2 | Objectives of research | 2-3 |
| 1.3 | Problem Statement | 3 |
| 1.4 | Industry Profile | 3 |
| 2 | Review of literature | 4 |
| 3 | Data Collection | 4-5 |
| 4 | Methodology | 5 |
| 4.1 | Exploratory Data Analysis | 5-6 |
| 4.1.1 | Figures and tables | 6-7 |
| 4.2 | Statistical Techniques and Data Visualization | 7-10 |
| 4.3 | Data Modeling using Supervised ML techniques | 10 |
| 5 | Findings and Suggestion | 10 |
| 6 | Conclusion | 10 |
| 7 | References | 11 |

**Taxi Fare Prediction**

**1.**  **Predict a rider's taxi fare**

**1.1 Introduction**

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.

Popular taxi services known as Uber and Ola provides their users with a prediction of taxi fare before the customer is mapped to a driver we try to provide a similar solution using a open data set that is a csv(comma separated file) file

In order to predict duration and fare,only data which would be available at the beginning of a ride was used.

Transportation has been proved as the most vital service large cities. Diverse modes of transportation are accessible. In large cities in the United States and cities around the world, taxi mode of conveyance plays a foremost role and used as the best substitute for the general public use of transportation to get their necessities

**1.2** **Objectives of research**

In this task, we are going to predict the fare amount for a taxi ride in New York City, given the pick-up, drop off locations and the date time of the pick-up. We will start from creating a simplest model after some basic data cleaning , this simple model is not Machine Learning, this we will move to more sophisticated models. Let’s get started.

The data set we have given is in form of longitudes and latitudes as the pickup and dropoff locations it is difficult to find distance in that format always so we convert the distance into normal form by using the Haversines formula .

Then we calculate the distance and according to that distance we predict the fare value

**1.3 Problem Statement**

By Considering the given data set, we have to predict the fare charges. depending on the distance between the pickup and dropoff locations. according the distance and the no.of passengers we should predict the fare value with a best accuracy rate.

* 1. **Industry Profile**

The industry today relies heavily on data analytics to make predictions. These predictions lead to successful business models that incentivise heavily from Machine learning. Popular taxi services known as Uber and Ola provides their users with a prediction of taxi fare before the customer is mapped to a driver we try to provide a similar solution using a open data set that is a csv(comma separated file) file named **train.csv**

Several advantages tend to make autos endearing to the public. It is preferable to a car on a number of grounds. It carries the same number of people on an average, takes one half of the driving space, one-third of the parking area and weight as that of a car. It results in lower wear and tear of roads. The vehicle itself wears out much less than a car and uses one third of national resources to produce it.

**2. Review of literature**

Taxis (and auto rickshaws) are public utility vehicles. More technically, they are called For Hire Vehicles (FHV). They are the transit transport of social agglomerates. They provide to-door service feeding mass and local transit, inter city transport such as to airports, rail heads and bus stations. From the passengers’ point of view, besides speed and availability, FHVs offers some novelty from routine driving, apart from offering certain conveniences resulting to them from the possibility of varying the routes at will. As carriers of fewer persons, they offer some privacy and freedom to the passengers to engage in a certain kinds of behavior that may be considered offensive or lacking in etiquette in bigger public carriers. According to Doolittle (1915), riding in open vehicles results in greater pleasure than in closed cars, for major portions in the year throughout the country.

Several advantages tend to make autos endearing to the public. It is preferable to a car on a number of grounds. It carries the same number of people on an average, takes one half of the driving space, one-third of the parking area and weight as that of a car. It results in lower wear and tear of roads. The vehicle itself wears out much less than a car and uses one third of national resources to produce it. All this reduces direct and indirect pollution. Due to their low weight they cannot produce fatal accidents among pedestrians and bicyclists easily. They offer protection against sun or rain than two wheelers.

**3.Data Collection**

The data that we have collected to predict this taxi fare is a open data source. It is one of the project that is given to Google and Kaggle for the for the solution and by the authority of that industry a data set is given that we are considering in this project. The data set given by the Industry will have more than 10 lakhs of lines it will be useful for analysing data more accurately.

Data fields:

**key**- Unique string identifying each row in both the training and test sets. Comprised of pickup\_datetime plus a unique integer, but this doesn't matter, it should just be used as a unique ID field.

Features:

* pickup\_datetime - timestamp value indicating when the taxi ride started.
* pickup\_longitude - float for longitude coordinate of where the taxi ride started.
* pickup\_latitude - float for latitude coordinate of where the taxi ride started.
* dropoff\_longitude - float for longitude coordinate of where the taxi ride ended.
* dropoff\_latitude - float for latitude coordinate of where the taxi ride ended.
* passenger\_count - integer indicating the number of passengers in the taxi ride.
* Observe the chapter 4.1.1 for the data explanation and data set

**Target:**

* fare\_amount - float dollar amount of the cost of the taxi ride. This value is only in the training set; this is what you are predicting in the test set.

**4. Methodology**

**4.1 Exploratory Data Analysis**

The next step to solve any analytics problems is to list down a set of hypothesis, which in our case are factors that will affect the cost of a taxi trip.

1. **Trip distance** : If the distance to be traveled is more, then fare should be higher.
2. **Time of Travel**: During peak traffic hours, the taxi fare may be higher.
3. **Day of Travel** : Fare amount may differ on weekday and weekends
4. **No of passengers**: depends on how many number of passengers have a ride at a time
5. **Use Haversine formula to calculate Distance:**

**Distance**

Calculate the distance based on longitude and latitude

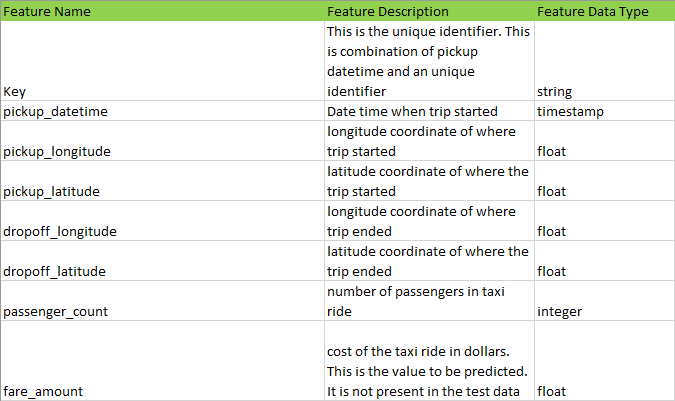
**Haversine formula:**

dlon = lon2 - lon1 dlat = lat2 - lat1 a = (sin(dlat/2))^2 + cos(lat1) cos(lat2) (sin(dlon/2))^2 c = 2 atan2( sqrt(a), sqrt(1-a) ) d = R c (where R is the radius of the Earth)

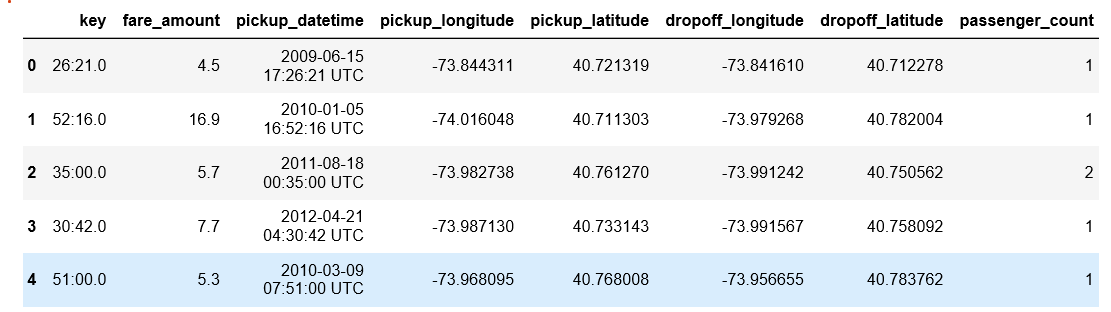
a = sin²(Δφ/2) + cos φ1 ⋅ cos φ2 ⋅ sin²(Δλ/2)

c = 2 ⋅ atan2( √a, √(1−a) ) d = R ⋅ c

**4.1.1 Figures and tables**



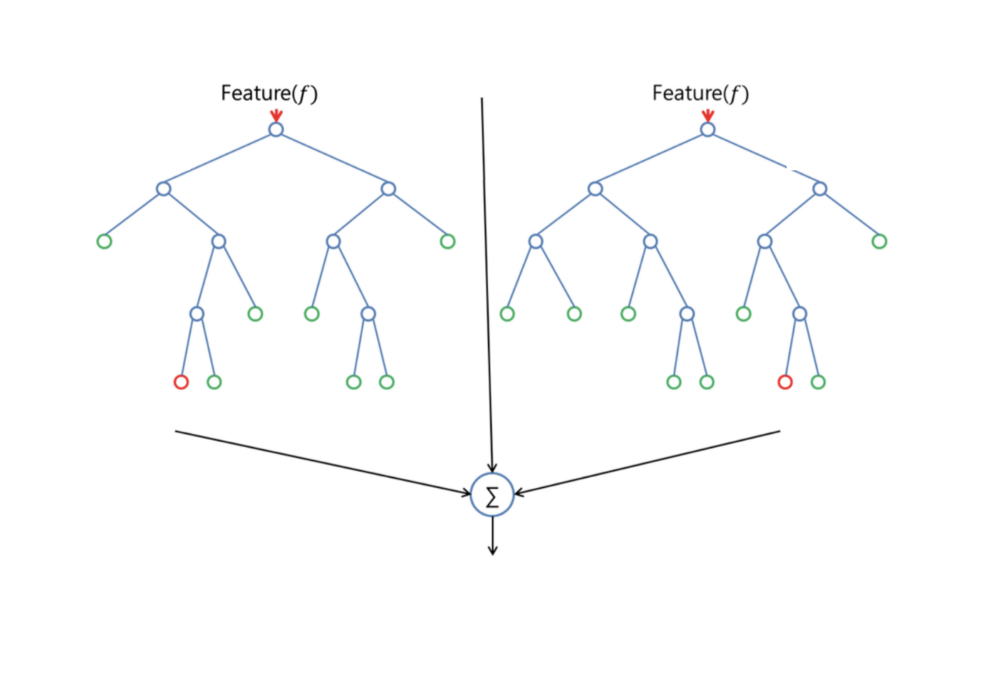
**🡪data.head(5)**

****

**4.2 Statistical Techniques and Data Visualization**

**Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it**

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, you don’t have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor.

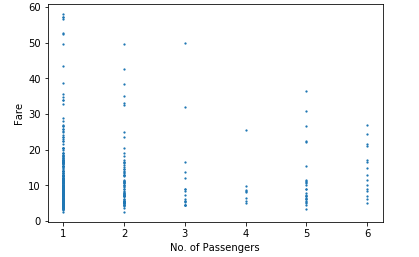
Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

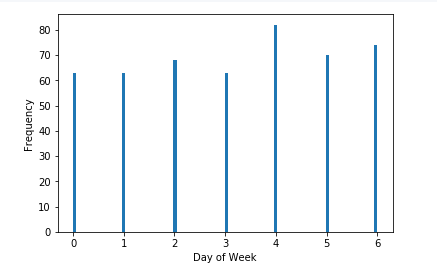
# ****1)Check the pickup date and time affect the fare or not****

# 

**2)number of passengers vs fare**



**3)Does the day of the week affect the fare?**



# 4.3 After applying the machine learning Algorithm to check accuracy:

# We used Random Forest Regression to check the accuracy here Random regressor is a one of the regression method it has advantages like it will be used in both the classification and regression and to our project Taxi fare prediction this Random Forest Regression is used because it is showing the accuracy of 84.3 % .

# 5. Findings and Suggestions:

# We have got the suggestions from our mentors and we also referred from the Github for the toughest parts in the program it was very helpful to our team to solve this project and we used the Node red in Watson studios to deploy this project.

# We people from our team suggest to provide a sample project before getting into the real project and make people to solve or predict any type of projects that may helpful in future.

# 6. Conclusion:

# Overall, our models for predicting taxi pickups in New York City performed well. The Random forest regression model performed best, likely due to its unique ability to capture complex feature dependencies. The decision tree regression model achieved a value of and 0.84 for R^2. Our results and error analysis for the most part supported our intuitions about the usefulness of our features, with the exception of the unexpected result that participation of rainfall  feature is not important for model performance. A model could be useful to city planners and taxi dispatchers in determining where to position taxicabs and studying patterns in ridership.

# 

# References:

# 1).<https://github.com/kuldeepnpatel/New-York-City-Taxi-Fare-Prediction/blob/master/New%20York%20City%20Taxi%20Fare%20Prediction.ipynb>

# 2.) <https://www.kaggle.com/c/new-york-city-taxi-fare-prediction>

# 