International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2062108

Predictive Analysis of Taxi Fare using Machine Learning

Pallab Banerjee[1], Biresh Kumar[2], Amarnath Singh[3], Priyeta Ranjan[4], Kunal Soni[5]

[1,2,3]Assistant Professor, Department of Computer Science and Engineering, Amity University Ranchi, Jharkhand, India

[4,5]B.Tech Scholar, Department of Computer Science and Engineering, Amity University Ranchi, Jharkhand, India

ABSTRACT

This research aims to study the predictive analysis, which is a method of analysis in Machine Learning. Many companies like Ola, Uber etc uses Artificial Intelligence and machine learning technologies to find the solution of accurate fare prediction problem. We are proposing this paper after comparative analysis of algorithms like regression and classification, which are useful for prediction modeling to get the most accurate value. This research will be helpful to those, who are involved in fare forecasting. In previous era, the fare was only dependent on distance, but with the enhancement in technologies the cab's fare is dependent on a lot of factors like time, location, number of passengers, traffic, number of hours, base fare etc. The study is based on Supervised learning whose one application is prediction, in machine learning.

Keywords : Machine Learning, Fare Prediction, Predictive Analysis, Supervised Learning, Feature Selection.

I. INTRODUCTION

Artificial Intelligence(AI) is the superset of Machine Learning(ML), and machine learning is superset of Deep Learning. ML is useful in model building as data is being feed to the machine, using algorithms further training and testing performed on those huge data so that the machine becomes capable of performing operations on its own on the new data given to it. It is divided into 3 types:

- Supervised learning: A supervision is required during the learning phase of machine. Both the input and desired outputs are available in it, model is prepared to predict the desired output. Eg. Regression and Classification.
- Unsupervised learning: It doesnot require any supervision, the model learns itself by finding the pattern within the dataset. Only input is given, model trains itself and output comes. Eg. Clustering and Association.
- Reinforcement learning: In this learning, the model is prepared using hit and trial method. It is dependent in nature. Its input are output of preceded process. Eg. Puzzle, chess etc.

In this research we have used supervised learning approach, because it suits the best as per the requirement of predictive analysis. Prediction is performed as data is collected from past, the model is trained to handle new data and predict the desired output.

Steps involved in this approach

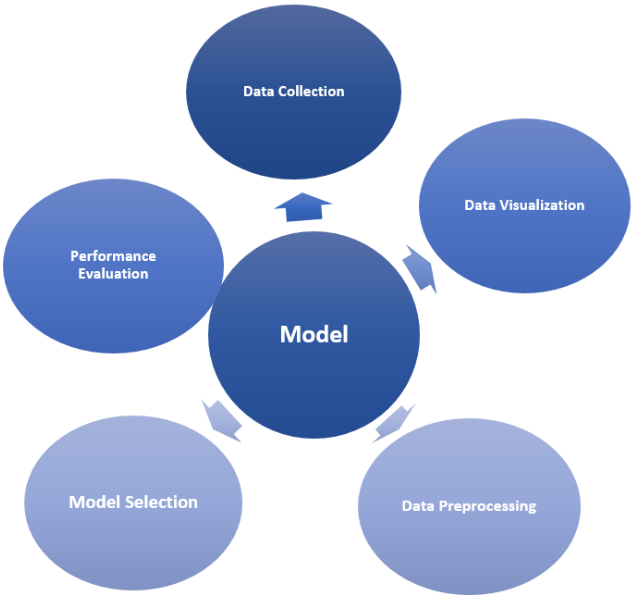
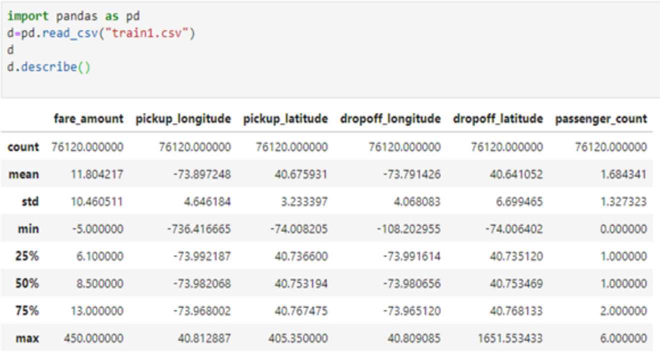


Fig 1.(a) 5 steps in model building

II. METHODS AND MATERIAL

We were looking for a dataset online which had information about the date and time, pickup and dropoff latitudes and longitudes, and the fare amount charged for that journey.

Data collection was done from the website of Kaggle[1], which consists of approximately 8 columns namely key, fare amount, pickup date and time, pickup longitude, pickup latitude, dropoff longitude, dropoff latitude and lastly, passenger count. It also consisted of around 5 million rows of data. Out of that data we have used approximately 80 thousand rows from year 2009 to 2016. Here is the following data view:



```
import pandas as pd
d=pd.read_csv("train1.csv")
d
d.describe()
```

	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	76120.000000	76120.000000	76120.000000	76120.000000	76120.000000	76120.000000
mean	11.804217	-73.897248	40.675931	-73.791426	40.641052	1.684341
std	10.460511	4.646184	3.233397	4.068083	6.699465	1.327323
min	-5.000000	-736.416665	-74.008205	-108.202955	-74.006402	0.000000
25%	6.100000	-73.992187	40.736600	-73.991614	40.735120	1.000000
50%	8.500000	-73.982068	40.753194	-73.980656	40.753469	1.000000
75%	13.000000	-73.968002	40.767475	-73.965120	40.768133	2.000000
max	450.000000	40.812887	405.350000	40.809085	1651.553433	6.000000

Fig.1(b) Data Collection

It includes the techniques, method we are have applied they are as follows:

1. DATA VISUALIZATION:

Data visualization helps to visualize the data easily using different graphs, charts, plots etc. So here it is represented using scatter plot, bar graph and histogram, so that data can be analysed properly. All these analysis is done in JupyterLab [2].

Longtitude, latitude of pickup and dropoff location plotting: Maximum travels had been done in the range of -73 to 40, but the rests are outliers which are present in the range of -700 and 400. This visualization is important for further processes.

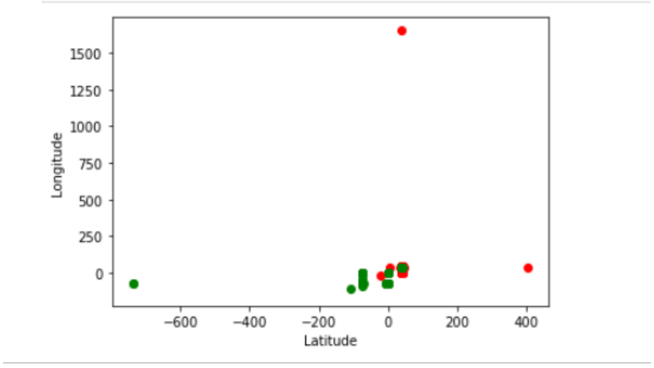


Fig.2.(a) Scatter Plot

Representation of data of passenger counts in the form of Bar graph:

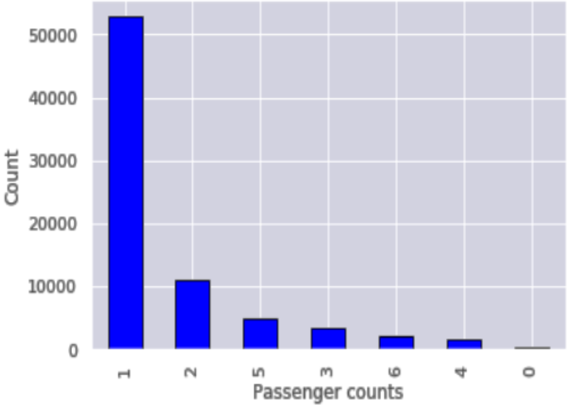
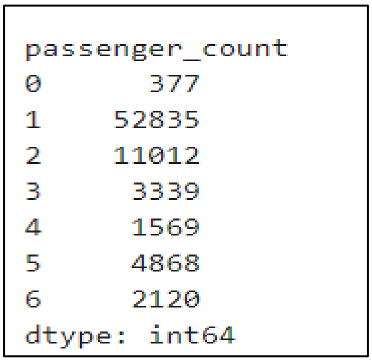


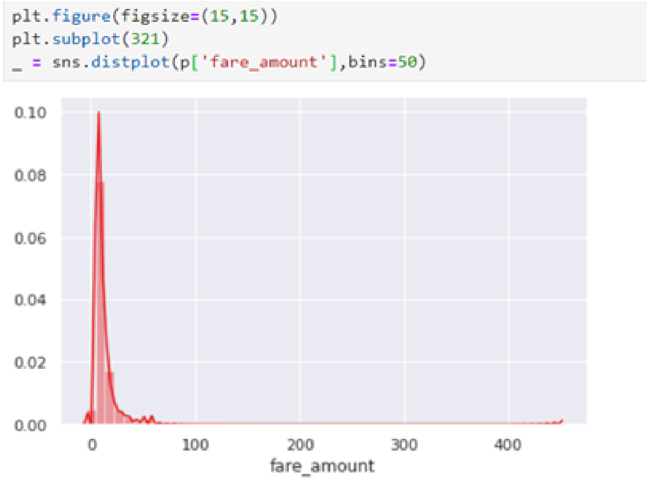
Fig.2.(b) Bar Graph

No of passengers:



```
passenger_count
0         377
1       52835
2       11012
3        3339
4        1569
5        4868
6        2120
dtype: int64
```

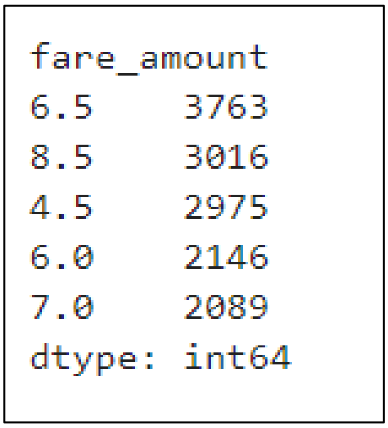
Dependent Variable in the dataset is Fare amount, its visualization is adequate for the data analysis part, so it is shown below:



```
plt.figure(figsize=(15,15))
plt.subplot(321)
_ = sns.distplot(p['fare_amount'],bins=50)
```

Fig.2.(c) Histogram

Maximum fare amount is in the range of 6.5 and 8.5, as the top 5 most paid amounts are shown in the below picture:



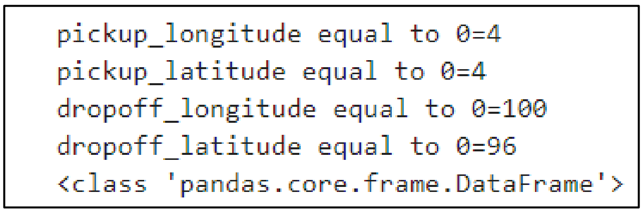
```
fare_amount
6.5     3763
8.5     3016
4.5     2975
6.0     2146
7.0     2089
dtype: int64
```

2. DATA PREPROCESSING

In machine learning, data preprocessing is a major process of cleaning the data. Basically it is a process to convert raw data into clean data. Here cleaning of data will be done in 2 ways: Missing data, Data out of the defined range called Noisy data

- **Missing Data**

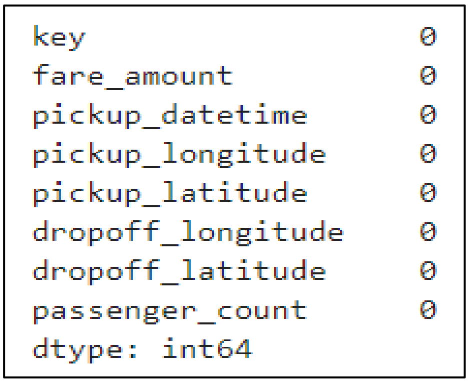
Total missing values are dropped during the cleaning process, before cleaning there were lots of missing values which were filled with the method of median. Other methods are also there like mean and mode, but mode sometimes gives biased value. So, mean and median are more preferable.



```
pickup_longitude equal to 0=4
pickup_latitude equal to 0=4
dropoff_longitude equal to 0=100
dropoff_latitude equal to 0=96
<class 'pandas.core.frame.DataFrame'>
```

Fig.2.(d). Before Cleaning missing values

After dealing with missing values, all data are left with no missing values. as it is seen below:

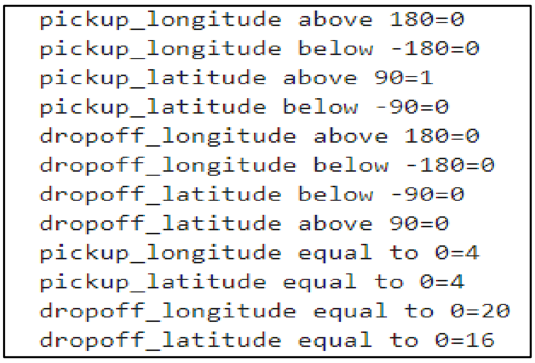


```
key                  0
fare_amount          0
pickup_datetime      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    0
dropoff_latitude     0
passenger_count      0
dtype: int64
```

Fig.2.(e) After Cleaning missing values

- **Noisy Data**

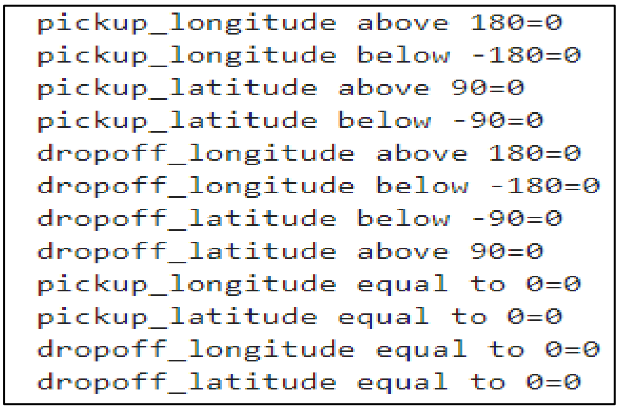
Those values which are out of the range of longitude and latitude are eliminated.



```
pickup_longitude above 180=0
pickup_longitude below -180=0
pickup_latitude above 90=1
pickup_latitude below -90=0
dropoff_longitude above 180=0
dropoff_longitude below -180=0
dropoff_latitude below -90=0
dropoff_latitude above 90=0
pickup_longitude equal to 0=4
pickup_latitude equal to 0=4
dropoff_longitude equal to 0=20
dropoff_latitude equal to 0=16
```

Fig.2.(f) Before Cleaning of noisy data

There are values which are equal to 0 we have also removed them. After cleaning noisy data, no noisy data is present as it follows:



```
pickup_longitude above 180=0
pickup_longitude below -180=0
pickup_latitude above 90=0
pickup_latitude below -90=0
dropoff_longitude above 180=0
dropoff_longitude below -180=0
dropoff_latitude below -90=0
dropoff_latitude above 90=0
pickup_longitude equal to 0=0
pickup_latitude equal to 0=0
dropoff_longitude equal to 0=0
dropoff_latitude equal to 0=0
```

Fig.2.(g)After cleaning noisy data

3. FEATURE SELECTION

In Machine learning, it is also known as attribute or variable selection. Basically, this is the process of selecting those attributes which contribute most to the target variable. Here fare_amount is the target variable or prediction variable, while others are independent variables.

Correaltion matrix helps to show the high correlation among the variables. According to the matrix, all the independent variables are important for the prediction variable as they all contribute to it. Other analyses are also present, but we have used correlation analysis for feature selection.

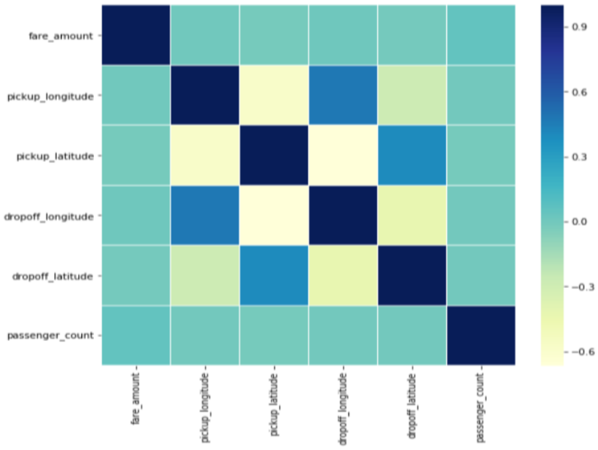


Fig.2.(f) Correlation Matrix[3]

This is done using python[4] language:
corrmat = train.corr()f, ax = plt.subplots(figsize =(9, 8)) sns.heatmap(corrmat, ax = ax, cmap ="YlGnBu", linewidths = 0.1)

4. MODELING

After data preprocessing an important step comes and that is modelling also known as model selection. Model selection is the process of selecting a model among many models for a predictive problem. Our problem is to predict the fare_amount. This is a Regression problem. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical. So, we are going to deal with regression models on training data and predict it on test data. In this research, we are using Random forest is a tree-based algorithm which can be used to solve regression, classification problems and Linear regression model, which is also helpful in regression models. We splitted the whole data into 2 parts: train data (75%) and test data(25%). After that different models are approached.

Random Forest

This model is based on supervised learning which helps in regression as well as classification. Random forest is better than a single decision tree, because it is made up of various decision trees which is collectively helpful in predicting the target value. A collection of trees gives more accurate value than a single tree.

Linear Regression

It is an algorithm based on **supervised learning**. It gives target prediction value based on independent variables. It is mostly used for finding out the relationship between dependent and independent variables.

Linear regression performs the task to predict a dependent variable value based on a given independent variable. So, this technique finds out a linear relationship between both type of variables. The model is based on the equation as shown below:

$$y=a+bx$$

where x is independent variable and y is dependent variable, but its aim is to find the appropriate values of a and b.

II. RESULTS AND DISCUSSION

We will evaluate performance of validation dataset. For evaluation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are widely adopted in many recommendation systems to measure the difference between the predicted scores and users' actual ratings[5].

We will deal with specific regression error metrics like –

- R square : The higher the value, the better the model. The values are taken as percentages between 0 to 1.
- MSE (Mean square Error): It is the average error rate which is the difference between the original value and predicted value.
- RMSE (Root Mean Square Error): It is the error rate by the square root of MSE. RMSE of 0.6 is small but it is not that small anymore. However, although the smaller the RMSE, the better.

The model which has the highest value of R square and the lowest value of RMSE is considered to the best model and the most accurate one also. According to our calculations and research, we found that Random Forest is the most suitable model for this regression problem.

Model	R square	MSE	RMSE
Random Forest	0.5	2.163	1.470
Linear Regression	0.4	2.642	1.625

III. CONCLUSION

After training and testing the results shown are fairly accurate. Random forest is useful in regression as well as classification whereas linear regression helps to find the linear relation among the variables. Hence we reached to the conclusion that Random forest is the best because it gives more accurate value as compared to linear regression model. That is why Random forest algorithm is the best fit for the model selection as it has the lowest RMSE value and the highest R square value. More further future scope is there if will apply more different types of approaches like XgBoost Regression technique or Ridge Regression technique.

III. REFERENCES

[1]. Machine learning in medicine: a practical introduction by Jenni A. M. Sidey-Gibbons & Chris J. Sidey-Gibbons

[2]. Model Buildingonline]: www.towardsdatascience.com

[3]. Datasets available online: www.kaggle.com/datasets

[4]. Scikit-learn: M L in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5]. JupiyterLabonline]:https://mybinder.org/v2/gh/jupyterlab/jupyterlab-demo/try.jupyter.org?urlpath=lab

[6]. Website online]: www.geeksforgeeks.com

[7]. Research:publication/323661651_Feature_selection_in_machine_learning_A_new_perspective

[8]. Journal of Machine Learning Research 12 (2011) 2825-2830 Submitted 3/11

[9]. Revised 8/11; Published 10/11. Scikit-learn: Machine Learning in Python

[10]. Exploring Correlation in Python - GeeksforGeeks

[11]. Feature Engineering for Predictive Modeling Using ... – AAAI by U Khurana

[12]. Sebastian Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, OnlineAvailable https://arxiv.org/abs/1811.12808,2016

[13]. James, G, D Witten, T Hastie, and R Tibshirani. An introduction to statistical learning. Vol. 6. , New York, Springer., 2013.

[14]. Analysis of a Random Forests Model - Journal of Machine by GÃŠ Biau

[15]. A full linear regression analysis –Research paper by Seber

[16]. Research work by Weijie Wang 1 and Yanmin Lu , Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model.

[17]. R square, error metrics Online]: www.researchgate.com

Cite this article as :

Pallab Banerjee, Biresh Kumar, Amarnath Singh, Priyeta Ranjan, Kunal Soni, "Predictive Analysis of Taxi Fare using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 2, pp. 373-378, March-April 2020. Available at doi : https://doi.org/10.32628/CSEIT2062108
Journal URL : http://ijsrcseit.com/CSEIT2062108