Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220086

Analyzing Ola Data for Predicting Price Based Trip Distance Using Random Forest and Linear Regression Analysis

G. Venkat Sai Tarun [a] and P.Sriramya[b,1]
[a]*Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b]*Professor, Dept. of AI&DS, Saveetha School of Engineering,*
SIMATS, Chennai, Tamilnadu, India

Abstract. The paper aims to create a most efficient and accurate cab fare prediction system using machine learning algorithms and comparing them. The machine learning algorithms are Random forest algorithm and Linear regression and comparing the r-square, mean square error (MSE), Root MSE and Root Mean Squared Logarithmic Error (RMSLE) values. We implement the Random forest and linear regression algorithms to predict the prices of the system and to get the best accuracy when comparing both the algorithms. The algorithms should be efficient to predict the prices of the trips before the starting of the trip. The sample size considered for this work is N=10 for each of the groups considered. Totally it was iterated 20 times for efficient and accurate analysis on prediction of price with G-power in 80% and threshold 0.05%, CI 95% mean and standard deviation. The sample size calculation was done with clincle. Based on the statistical analysis the significance value for calculating the r-square was found to be 0.034. The Random forest algorithm gives a slightly better accuracy rate with a mean r-square percentage of 71.67% and the linear regression has mean r-square value of 70.57%. By this process, the prediction is done for the price prediction of the online cab rental system and the Random forest algorithm gives a better r-square value compared to the Linear regression algorithm.

Keywords. Linear regression, Random Forest Regression, Fare prediction, Novel Exploratory data analysis, Machine Learning

1. Introduction

This paper aims to predict the prices for the online cab rental system using machine learning algorithms with Random forest algorithm and Linear a regression algorithm. The price of the trip which is going to be started is shown before the trip starts. This process is shown as the price prediction [1]. It is mainly important for the accurate prediction of the prices. The price of the trip which is going to be started is calculated using the given values of the attributes. The attributes like Location, date-time, passenger count. The existing fare amount is also given as an attribute, the existing fare amount should be updated or changed through the program and the fare amount is

[1]P.Sriramya, Dept. of AI&DS, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Chennai, India. E-mail: sriramyap@saveetha.com

updated through the weather conditions, day or night, etc., these conditions affect the fare amount and update it [2]. The applications of the research show the usage of the proposed algorithm which is the Random Forest algorithm and where it is mainly used. It is mainly used in decision-making based on the different circumstances, predicting the values and is so useful in medical industries and natural language processing tasks [3].

The existing papers show many algorithms for the prediction of values, one of the best algorithms for prediction analysis is random forest algorithm. In papers like evaluating multiple classifiers for stock price prediction, it shows the accurate prediction of stock prices and their updating of the prices. The purpose of this research work is to benchmark ensemble methods and against single classifier models. Results show the Random forest algorithm is the top algorithm followed by the support vector machines then the paper using machine learning algorithms for house price prediction; it shows the analysis of 5359 townhouses of metropolitan information systems to improve the accuracy. This article demonstrates that the random forest algorithm based on accuracy outperformed the other model [4]. In the paper artificial neural networks artificial network stock price prediction, it presents an artificial neural network approach to predict the stock market indices. It focuses on the importance of choosing the correct input features along with preprocessing [5]. In the paper using the bitcoin transaction graph to predict the price of bitcoin, we investigate the predictive power of the block chain network, based on future price bitcoin. It obtains a bitcoin price movement classification accuracy of 75%. Limited research has been performed to analyze network influence on overall bitcoin prices [6].

The drawbacks of the existing research show the inaccurate predictions of prices with less performance of the execution and the overfitting model shows the drawbacks. This research work aims to create the most efficient and accurate prediction of prices without errors [7]. By giving the dataset we compare the attributes and get the accuracy prediction

2. Materials and Methods

The investigation was performed in the Department of Data Science, Saveetha School of Engineering (SSE), SIMATS. The sample size taken for conducting the experiment was 1SIMATIC groups are considered as classifiers algorithms in order to classify the prediction of the fare amount, machine learning algorithms are used. The Linear regression algorithm is in group 1 and the Random forest algorithm is in group 2, and they are evaluated for further r-square, MSE, RMSE and RMSLE values to determine which approach is the best. The existing work is linear regression and the proposed work is Random forest algorithm for price prediction. The total number of samples that are evaluated on the proposed methodology is 75 in each of the 2 groups. Attributes are the crucial part to show the predicted on price of the fare. The required samples for this analysis are done us G the power calculation the minimum power of the analysis is fixed as 0.8 and maximum accepted error is fixed as 0.5.

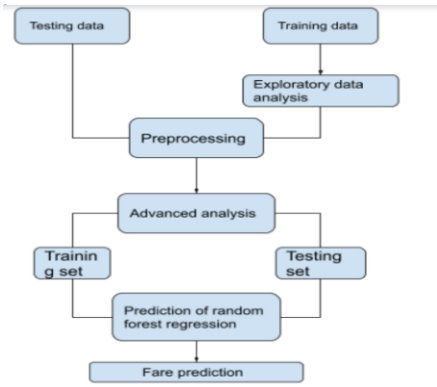

Figure 1. Multiple linear the egression architecture diagram.

a. Linear Regression Algorithm

The existing algorithm in this diagrams the linear regression algorithm. To examine the incoming data and highlight their key properties, a novel exploratory data analysis is used. To extract the major characteristic for data extraction, the training dataset is subjected to a unique exploratory data analysis. Linear regression is a machine learning approach for mapping numeric inputs to numeric outputs by fitting a line through the data points. It's the process of selecting a line that best matches the data points on the plot so that we can use it to forecast output values for inputs that aren't in the data set we have, with the assumption that such outputs will fall on the line. Performance depends on various factors including how clean and consistent the data is [8].

b. Multiple Linear Regression Algorithm

Random forest regression is the proposed algorithm in this paper. **Fig. 1** shows the architecture diagram of the process done in the proposed algorithm. Random forest algorithm is a tree-based algorithm that uses quality. A randomness of multiple decision trees. Random forest is a supervised algorithm where the training dataset is given as input and predicted values are found. The decision tree algorithms have disadvantages like low accuracy in executing and inaccurate predictions. These disadvantages can be solved using the Random forest algorithm. Here in this algorithm, the data is divided into tree-sets, and the program is executed in different types of ways tree sets accuracy is found [1]. The testing procedure includes training the dataset before continuing the testing process and after testing it, training and evaluating the algorithms. The testing procedure includes training the dataset before continuing the testing process and after testing it, training and evaluating the algorithms. In Random forest algorithm, where the dataset is divided into trees tend to executed.

The testing setup of the implementation includes the software requirements and hardware requirements. The software requirements are the tools which are Jupyter notebook or Google Colab and the software used is python. Hardware requirements are Windows version with 7 or 8 or 10, minimum processor of 1 ghz and memory with minimum 1 GB. The input data set was collected from the Github repository. The datasets consist of both dependent and independent attributes. **Table 1** gives the sample of the input dataset taken.

Table 1. Sample Input Dataset

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.5	2009-06-15 17:26:21 UTC	-73.844311	40.721319	-73.84161	40.712278	1
16.9	2010-01-05 16:52:16 UTC	-74.016048	40.711303	-73.979268	40.782004	1
5.7	2011-08-18 00:35:00 UTC	-73.982738	40.76127	-73.991242	40.750562	2
7.7	2012-04-21 04:30:42 UTC	-73.98713	40.733143	-73.991567	40.758092	1
5.3	2010-03-09 07:51:00 UTC	-73.968095	40.768008	-73.956655	40.783762	1

For statistical comparisons of metrics like r squared and MSE, SPSS version 21 was employed. The dependent attributes are fare amount and fare amount Pickup datetime, pickup_longitude, pickup_lattitude, dropoff_longitude and latitude, and passenger count are independent pr, operties that will appear in both data sets. The r-squared and MSE were calculated. The r-squared value and Mean Square Error were calculated using an independent sample T test.

3. Results

In this study we have observed that the Random forest algorithm, has slightly better r-squared value than linear regression algorithm squared, independent sample t-test). When the algorithms are compared with each other, Random forest algorithm has a higher r-squared value of 71.67%, when compared to linear regression with value of 70.57%. And the mean square error the (MSE) of the Random forest algorithm (53.31%) is lesser than linear regression (54.22%). As there is a marginal difference in accuracy, random forest is statistically better when compared to linear regression. **Figure 2a** gives the comparative analysis of Training data for the performance evaluation parameters r square, MSE, RMSE and RMSLE**.** From **Table 2** it is observed, that the r-square and MSE values are almost the same as in both the algorithms in the case of training data.

Table 2. Comparison of the performance evalu T-test metrics for training and testing data values achieved.

	Training Data		Testing Data	
	Random forest	**Linear regression**	**Random forest**	**Linear regression**
MSE	4.017	5.795	5.601	5.627
RMSE	2.004	2.407	2.306	2.72
RMSLE	0.191	0.230	0.220	0.225
R square	0.796	0.706	0.714	0.712

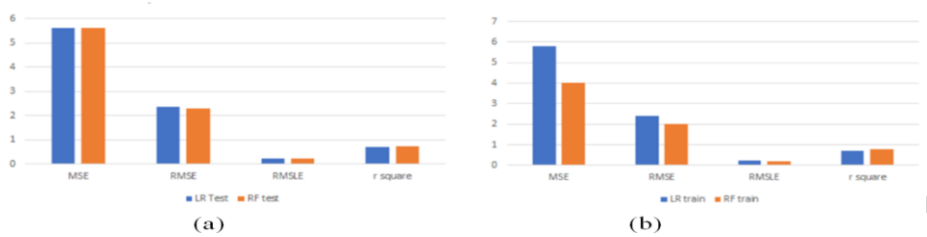

Figure 2a and 2b. Comparative analysis of Test and Figures data for the performance evaluation parameters r square, MSE, RMSE and RMSLE.

In Figure 2b gives the comparative analysis of Test data for the performance evaluation parameters r square, MSE, RMSE and RMSLE**.** From **Table 3** it is observed that there is a slight significant increase in r-square and MSE values in both the algorithms in the case of testing data. Since testing data is considered for the results we can prove that Random forest is able to predict the price in a scanner.

Table 3. Group Statistics: Comparison of Random Forest and Multiple linear algorithms by varying rsquare parameters. Algorithms linear has square value of 71.69 for and the Random Forest results in a m value of 71.29 for r-square.

	Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
r-square	RF	10	71.67	.803	.254
	LINEAR	10	70.57	.782	.247
MSE	RF	10	53.31	1.921	.607
	LINEAR	10	63.05	2.701	.854

Table 4. Independent Sample T test for the two groups has been carried out and it is observed that there is a slight difference in r-squared and MSE between Multiple linear and Random Forest algorithms. [significance is 0.945 (r-square) and 0.266 (MSE), p>0.05]

		F	**Sig.**	**T**	**Df**	**Sig. (2-tailed)**	**Mean Difference**	**Std.Error Difference**	**Lower**	**Upper**
r-square	Equal variances assumed	.001	0.034	3.122	18	.011	0.006	1.107	.362	1.852
	Equal variances not assumed			3.122	17.988	.011	0.006	1.107	.362	1.852
MSE	Equal variances assumed	0.473	0.05	0.081	18	.348	0.936	.085	1.048	2.287
	Equal variances not assumed			0.081	16.248	.349	0.936	.085	1.048	2.287

A brief descriptive statistical analysis was performed to obtain Mean, Std. Deviation and Std. Error Mean for r-squared and MSE values of Random forest algorithm and linear regression algorithm which is presented in **Table 4**. An independent sample t-test was performed with a fixed confidence level to obtain t-test Equality of means which is presented the in Table 4. **Figure 3** gives the bar chart representing the comparison of Random forest and linear regression in terms of r-square and MSE. The mean accuracy for random forest is lesser than linear regression and the standard deviation of Random forest with E-LSB is better than linear regression.

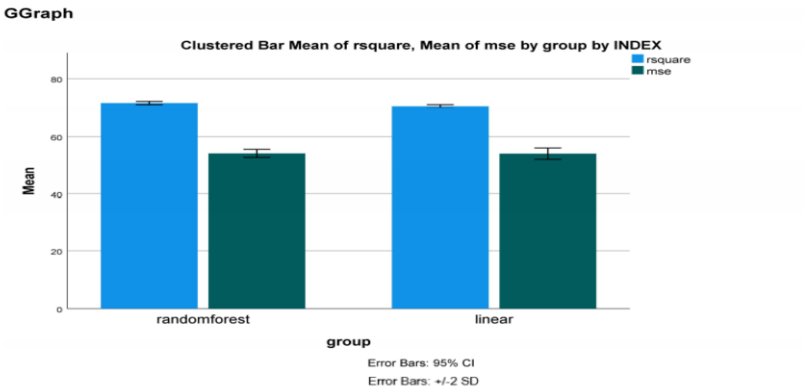

Figure 3. Bar chart representing the comparison of Random Forest (RF) and Linear regression in terms of r-square and MSE.

4. Discussion

The Random forest algorithm is a tree-based algorithm that uses quality features of multiple decision trees. Here in this algorithm, the data is divided into tree sets, and the program is executed in different types of ways where the accuracy is found, the values are accurate due to this property. Random forest is slightly better than linear regression and with the reference to the significance value to be less than 0.034. The average r-square value of random forest is 71.67%.

In the work, Machine learning using novel exploratory data analysis to predict taxi fare, it investigates how random forest is used to solve regression and classification problems. It shows random forest is the best model which has the lowest RMSE value and shows why the Random forest is the best algorithm in machine learning techniques [9]. The similar findings are that the Random forest provides more accurate values or predictions and is capable of capturing a decent amount of variations [10].

The limitation of this work is that when there is an over fitted model it performs worse on the testing dataset. Although the results of the study are slightly better in both experimental and statistical analysis there are limitations in the work. Linear regression cannot obtain accurate values when the points on the graph do not fit in the line. It considers only normalized regression. This study also excludes the impacts of urban use properties on taxi demand, which is the future work. Obtaining certain environmental qualities can be difficult. As a result, we will offer a strategy for forecasting environmental variables in order to better predict taxi demand in the future.

5. Conclusion

In this research, the aim was to predict the fare amount with accurate values. Here we proposed a method for cab fare prediction using machine learning techniques, these results showed a slightly better accuracy standard for producing a near accurate estimation result. Based on the significance value (0.034) achieved through SPSS. The

random forest mean accuracy is 71.67% and the Linear regression mean accuracy is 70.57%. The mean square error of random forest was also lower when compared to linear regression algorithm. Thus, the Random forest algorithm has slightly better accuracy when compared to the Linear regression algorithm.

References

[1] Mantas CJ, Castellano JG, Moral-García S, Abellán J. A comparison of random forest based algorithms: random credal random forest versus oblique random forest [Internet]. Vol. 23, Soft Computing. 2019. p. 10739–54. Available from: http://dx.doi.org/10.1007/s00500-018-3628-5

[2] Roh KH. A Study on the Method of Taxi Fare Calculation [Internet]. Vol. null, Management & Information Systems Review. 2007. p. 201–31. Available from: http://dx.doi.org/10.29214/damis.2007..23.009

[3] Koning M, Smith C. Decision Trees and Random Forests: A Visual Introduction for Beginners. Independently Published; 2017. 168 p.

[4] Kumar GK, Kiran Kumar G, Malathi Rani D, Koppula N, Ashraf S. Prediction of House Price Using Machine Learning Algorithms [Internet]. 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). 2021. Available from: http://dx.doi.org/10.1109/icoei51242.2021.9452820

[5] Liang J, Song W, Wang M. Stock Price Prediction Based on Procedural Neural Networks [Internet]. Vol. 2011, Advances in Artificial Neural Systems. 2011. p. 1–11. Available from: http://dx.doi.org/10.1155/2011/814769

[6] Loh EC. Emerging Trend of Transaction and Investment: Bitcoin Price Prediction using Machine Learning [Internet]. Vol. 9, International Journal of Advanced Trends in Computer Science and Engineering. 2020. p. 100–4. Available from: http://dx.doi.org/10.30534/ijatcse/2020/1591.42020

[7] Simple Linear Regression [Internet]. Applied Linear Regression. 2005. p. 19–46. Available from: http://dx.doi.org/10.1002/0471704091.ch2

[8] Späth H. Linear Lp Regression with Linear Constraints [Internet]. Mathematical Algorithms for Linear Regression. 1992. p. 217–48. Available from: http://dx.doi.org/10.1016/b978-0-12-656460-0.50011-2

[9] Panda G. Machine Learning using Exploratory Analysis to Predict Taxi Fare [Internet]. Vol. 7, International Journal for Research in Applied Science and Engineering Technology. 2019. p. 516–22. Available from: http://dx.doi.org/10.22214/ijraset.2019.8073

[10] Hansch R. Stacked Random Forests: More Accurate and Better Calibrated [Internet]. IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium. 2020. Available from: http://dx.doi.org/10.1109/igarss39084.2020.9324475