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

doi:10.3233/APC220071

Ola Data Analysis for Dynamic Price Prediction Using Multiple Linear Regression and Random Forest Regression

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. This research aims to create the most efficient and accurate cab fare prediction system using two machine learning algorithms, the Multiple linear Regression algorithm and the random forest algorithm, and compare parameters r-square, Mean Square Error (MSE), Root MSE, and RMSLE values to evaluate the efficiency of two machine learning algorithm. Considering Multiple linear Regression as group 1 and random forest algorithms as implemented, the 2 group process was to predict prices and get the best accuracy to compare algorithms. The algorithm should be efficient enough to produce the exact fare amount of the trip before the trip starts. The sample size for implementing this work was N=10 for each group considered. The sample size calculation was done with clincle. The pretest analysis was kept at 80%. The sample size is estimated using G-power. Based on the statistical analysis significance value for calculating r-squared, MSE was 0.945 and 0.266($p>0.05$), respectively. The Multiple linear algorithms give a slightly better accuracy rate with a mean r-squared percentage of 71.69%, and the Random forest algorithm has a mean r-square of 71.29%. Through this, prediction is made for online booking of cabs or taxis, and the Multiple linear algorithms give a slightly better r-squared value than the Random forest algorithm.

Keywords. Multiple Linear regression, Random Forest regression, Fare prediction, Novel exploratory data analysis, Machine Learning.

1. Introduction

This research aims to predict fare amount for online cab services with a machine learning algorithm using a Multiple linear regression algorithm to evaluate the r squared value with a Random forest algorithm [1]. The primary importance of this study is predicting the prices of online cab services. The trip cost which will start can be shown before the trip begins. This process is shown as price prediction. The price prediction offers the trip's fare by calculating the given values of the attributes.

The attributes are the central values to be calculated to show prediction—the attributes like location, date-time, passenger count, and existing fare amount [2]. The existing fare amount should be changed or updated through the program, and the fare amount is updated through weather conditions, day or night, etc. These conditions

[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.

affect the fare amount and update it. The research applications show the uses of the proposed algorithm, Multiple linear regression, and where it is used [3]. It mainly correlates two or more independent variables and obtains statistically significant relations. Under certain conditions, multiple linear regression allows us to derive projected values for specific variables. [4].

According to the literature survey, more than 20,000 related articles were published in Google Scholar, and Scopus indexed journals of Machine Learning. Many existing reports show predictions of cab fare with different approaches. The multiple linear regression algorithm is one of the best algorithms for prediction analysis. This study focuses on the mobile application for the stock prediction that uses improved multiple linear regression to provide accurate stock price predictions. IMLR is a hybrid technique that combines the average moving approach with multiple linear regression in multiple linear regression. It analyzes the daily stock history and predicts stock prices [5].

In this research work, multiple linear regression algorithms are used to forecast and explain the creation of Spanish day-ahead electricity prices. The accuracy of multiple linear regression models is improved at the expense of complexity. [6] show a quantile regression model based on gradient boosted regression trees. The price trend of maize is forecasted using several linear regression analysis models under big data, although the estimated price will diverge from the actual model. The forecast value is derived from the study of independent factors. Then it was utilized to predict corn [2] better. In the investigation of Multiple linear regression predictor analysis for electricity price forecasting, error rates are considerably reduced using multiple linear regression.

2. Materials and Methods

The research was performed in the Data Analytics Lab in the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences. The sample size taken for experimenting was 10. Two groups are considered classifiers algorithms to classify prediction of fare amount, and machine learning classification algorithms are used. Group 1 is the Random forest algorithm, and the Multiple linear regression algorithm is group 2, and they are compared for more r-square, MSE, RMSE, and RMSLE values for choosing the best algorithm. The current work is random forest regression, and the proposed work is multiple linear regression for price prediction. The total number of samples evaluated on the proposed methodology is 75 in each of the two groups. Attributes are a crucial part of showing the prediction price of the fare. The g power calculation calculates the required samples for this assay. [7]. The minimum analytical power is set at 0.8, while the maximum allowed error is set to 0.5. Figure 1 depicts the work's intended architecture

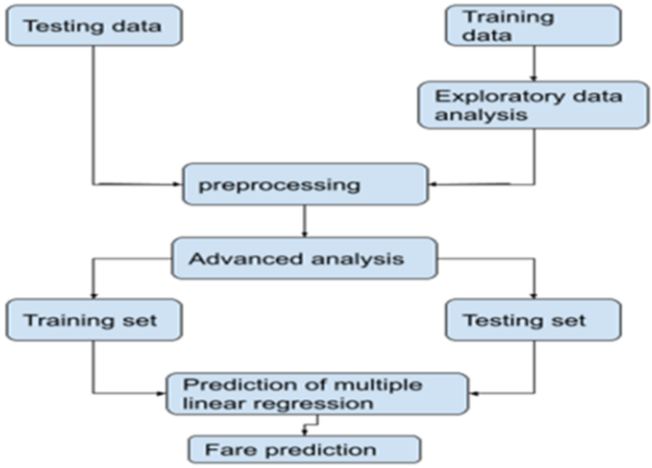

Figure 1. Multiple linear regression architecture diagram.

2.1. Random Forest Algorithm

Random forest regression is the existing algorithm in this work. A Novel exploratory data analysis is applied to analyze input data and summarize their main characteristics. The training dataset goes through novel exploratory data analysis to extract the main feature for data extraction. The random forest algorithm is a tree-based technique that employs numerous decision trees' quality features. Random forest is a supervised method that takes as input a training dataset and finds predicted values. The decision tree algorithms have disadvantages like low execution accuracy and inaccurate predictions. These disadvantages can be solved using the Random forest algorithm. Here in this algorithm, the data is divided into three sets, and the program is executed in different ways where accuracy is found [8].

2.2. Multiple Linear Regression Algorithm

The Multiple linear regression algorithm is the proposed algorithm in this article. The testing procedure includes training the dataset before continuing and after testing it, training and evaluating these algorithms. The testing procedure includes training the dataset before continuing the testing process and, after trying it, training and evaluating

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

2.3. Statistical Analysis

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 latitude, dropoff longitude and latitude, and passenger count are independent attributes that will appear in both data sets. The r-squared and MSE were calculated. The r-squared value and Mean Square Error were computed using an independent sample T-test. Algorithm ($p>0.05$, Independent sample t-test). When these algorithms are compared, Multiple linear regression has a higher r-square value of 71.69% when compared to random forest regression of 71.29%. And mean square error of multiple linear regression (54.54%) is lesser than Random Forest (57.40%). As there is a marginal difference in accuracy, Multiple linear regression is statistically better when compared to random forest regression.

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

Table 2. Comparison of the performance evaluation metrics for training data values achieved.

	Multiple linear	Random Forest
MSE	4.017	5.724
RMSE	2.004	2.392
RMSLE	0.191	0.227
R square	0.796	0.712

3. Results

In this study, we observed that multiple linear regression algorithms have a slightly better r-squared value than Random Forest observed that r-squared and MSE values are almost the same as in both algorithms in case of training data. According to the results achieved in training data there is better improvement in MSE value of Multiple Linear Regression (4.017) when compared to Random Forest (5.724). Figure 2 gives comparative analysis of Training data for performance evaluation parameters r square, MSE, RMSE and RMSLE**.** From Table 2.

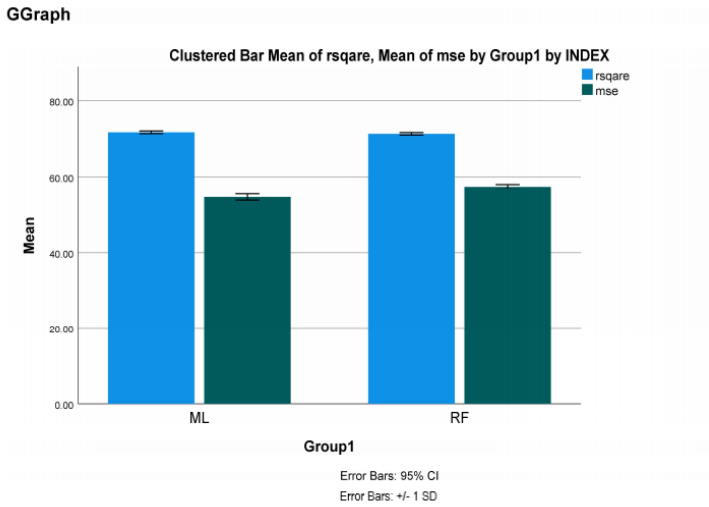

Figure 3. Comparative analysis of Training data for the performance evaluation parameters r square, MSE, RMSE and RMSLE**.**

Figure 3 gives comparative analysis of Test data for 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 values in both algorithms in case of testing data. Accordingto the results achieved in testing data there is slight improvement in MSE value of Multiple Linear Regression (5.601) when compared to Random Forest (5.651). Since testing data is considered for results, we can prove that Multiple linear regression is able to predict price in an accurate manner.

Table 3. Comparison of the performance evaluation metrics for testing data values achieved.

	Multiple linear	**Random Forest**
MSE	5.601	5.651
RMSE	2.366	2.377
RMSLE	0.220	0.225
R square	0.734	0.708

A brief descriptive statistical analysis was performed to obtain Mean, Std. Deviation and Std. Error Mean for r-squared and MSE values of Multiple linear regression algorithm and Random Forest 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 in Table 5.

Table 4. Group Statistics: Comparison of Random Forest and Multiple linear algorithm by varying rsquare parameters. Multiple linear has a mean value of 71.69 for and the Random Forest results in a mean value of 71.29 for r-square.

	Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
r-square	Multiple linear	10	71.69	.486	.153
	RF	10	71.29	.467	.147
MSE	Multiple linear	10	54.54	1.172	.370
	RF	*10*	*57.40*	*.843*	*.266*

Table 5. Independent Sample T test for the two groups 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
r-square	Equal variances assumed	.005	.945	1.865	18	.039	0. .398	.280
	Equal variances not assumed			1.865	17.973	.039	0 .398	.280
MSE	Equal variances assumed	1.317	.266	-6.248	18	.001	-2.85	.934
	Equal variances not assumed			-6.248	16.348	.001	-2.85	.934

4. Discussion

Multiple linear regression is a machine learning approach that lets us transfer numeric inputs to numeric outputs by fitting a line through the data points. It's the process of identifying a line that best fits the data points on a plot so that we can use it to forecast output values for inputs that aren't present in the data set we have, with the hope that those outputs will fall on the line. Multiple linear regression is slightly better than random forest algorithm in terms of determining model variation and relative contribution of each independent variable in total variance. This property shows that multiple linear regression is slightly better than random forest algorithm, with a significance value of less than 0.945. The average r square value of multiple linear regression is 71.69%.

In the article Multiple linear regression model for predicting bidding price, it shows accuracy estimated from statistics of validation data r square, MAPE is used as estimators of model [9]. Multiple linear regression is one of the best machine learning algorithms for finding predictions of values and shows why it is the best algorithm to use [2]. The similar findings are that multiple linear regression provides more accurate values or predictions and is capable of capturing a decent amount of variations[10]. There are no opposite findings observed for this work.

The limitation of this work is that, when there is an overfitted model it performs worse on the testing dataset. Despite the fact that the study's outcomes are marginally superior in both experimental and statistical analysis, the work has limitations. Multiple

linear regression cannot obtain accurate values when the points on the graph do not fit in line. It considers the effect of more than one explanatory variable on some outcomes. As a result, future work could include improving the algorithm so that it can compute dynamic ride-sharing during peak traffic hours. To deal with this complexity, deep neural networks can be used.

5. Conclusion

The research work, 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 (0.945) achieved through SPSS. Multiple linear regression mean accuracy is 71.69% and random forest mean accuracy is 71.29%. The mean square error of multiple linear regression is lower when compared to random forest algorithms. Thus, multiple linear regression has slightly better accuracy when compared to random forest algorithms.

References

[1] Politis DN. Model-Based Prediction in Regression. Model-Free Prediction and Regression 2015; 33–56.
[2] Ulgen T, Poyrazoglu G. Predictor Analysis for Electricity Price Forecasting by Multiple Linear Regression. 2020 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM). Epub ahead of print 2020. DOI: 10.1109/speedam48782.2020.9161866.
[3] Smith PF, Ganesh S, Liu P. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. Journal of Neuroscience Methods 2013; 220: 85–91.
[4] Roback P, Legler J. Review of Multiple Linear Regression. Beyond Multiple Linear Regression 2021; 1–38.
[5] Izzah A, Sari YA, Widyastuti R, et al. Mobile app for stock prediction using Improved Multiple Linear Regression. 2017 International Conference on Sustainable Information Engineering and Technology (SIET). Epub ahead of print 2017. DOI: 10.1109/siet.2017.8304126.
[6] Sharma N. XGBoost. The Extreme Gradient Boosting for Mining Applications. GRIN Verlag, 2018.
[7] Yuen KC, Zhu L, Zhang D. Lifetime Data Analysis. 2002; 8: 401–412.
[8] Xu R. Improvements to random forest methodology. DOI: 10.31274/etd-180810-3436.
[9] Noi P, Degener J, Kappas M. Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data. Remote Sensing 2017; 9: 398.
[10] Kaushal A, Shankar A. House Price Prediction Using Multiple Linear Regression. SSRN Electronic Journal. DOI: 10.2139/ssrn.3833734.