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

doi:10.3233/APC220050

Analyzing Ola Data for Precise Price Prediction Using XGBoost Technique Comparing with LASSO Regression

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

Abstract: XGBoost algorithm and Lasso regression and compare r-square, Mean Square Error (MSE), Root MSE, and RMSLE values. 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 of the groups considered. It was iterated 20 times for efficient and accurate prediction of cab price prediction with G power in 80% and threshold 0.05%, CI 95% mean and standard deviation. The sample size calculation was done with clincle. The pretest analysis was kept at 80%. The sample size calculation was done using clincalc. The statistical analysis shows that the significance value for calculating r-squared and MSE was 0.63 and 0.581(p>0.05), respectively. The XGBoost algorithm gives a slightly better accuracy rate with a mean r-squared percentage of 72.62%, and the Lasso regression algorithm has a mean r-square of 70.47%. Through this, the prediction is made for the online booking of cabs or taxis, and the Xgboost algorithm gives a slightly better r-squared value and MSE values than the Lasso regression algorithm.

Keywords. XGBoost regression, LASSO regression, Fare prediction, Novel exploratory data analysis, Machine Learning.

1. Introduction

The objective of this study is to use a machine learning method called XGBoost Algorithm to predict the fare amount for online cab services before the trip starts by comparing the r-squared and MSE values with the Lasso regression algorithm [1]. The central importance of this study is predicting the prices of online cab services. The price of the trip which will be started can be shown before the trip starts. This process is shown as the price prediction. The price prediction shows the trip's fare by calculating the given values of the attributes. The attributes are the central values to be calculated to show the prediction. The attributes include location, date-time, passenger count, and fare amount. The existing fare amount should be changed or updated through the program, and the fare amount is updated through the weather conditions,

[1] P.Sriramya, Dept. of AI&DS, Saveetha School of Engineering, SIMATS, *Chennai, India.* E-mail: sriramyap@saveetha.com

day or night, etc.; these conditions affect the fare amount and update it [2]. The research applications show the uses of the proposed algorithm XGBoost and where it is used. Mainly it is used in cache awareness and out-of-core computing, Parallelized tree building [3], efficient handling of missing data, and regularization of avoiding overfitting.

According to the literature the survey made, there are more than 20,000 related articles published in Google Scholar and Scopus indexed journals of Machine Learning. Many existing articles show the prediction of crude oil prices using different Machine learning approaches. The XGBoost algorithm is considered to be used for predicting crude prices rise among them. In this work, Crude oil price forecasting is done using XGBoost, and it shows the accurate forecasting of crude oil prices. Forecasting crude oil prices gives profit to the country's economy. It is an excellent need for estimation where the quantity of crude oil is deficient [4]. In the research, CEEMDAN and XGBoost based approaches to forecasting crude oil prices show the novel approach which integrates to complete empirical ensemble decomposition. Here they used CEEMDEN to decompose nonstationary and nonlinear sequences. Results show that the proposed algorithm outperforms the other with high accuracy of 0.8241(82%) [5]. We can also find the Bitcoin price prediction using machine learning to approach sample dimension engineering [6]. This article classifies the bitcoin prices by daily price and [7] achieves a better performance with high accuracies. It shows that the pilot study of the importance of sample dimension in machine learning accuracy is (71%) [8]. In this article [9], the insurance claim amount of past claim data is usually substantial.

Furthermore, many of the data's features have missing values. As a result, they use machine learning models that can deal with both types of data. XGBoost is suitable for both data characteristics. XGBoost gives better accuracy in terms of other methods. Accuracy prediction achieved using XGBoost is 75%.

The drawback of the existing algorithm is that it gives an inaccurate prediction of prices with less execution performance and cannot make group selection if there is a group of variables. This article aims to create the most efficient and accurate prediction of prices without errors. By giving the dataset, we compare the attributes and get the accuracy prediction.

2. Materials and Methods

The research was performed in the Data Analytics Lab of 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 the fare amount prediction, and machine learning classification algorithms are used. Group 1 is the Lasso regression algorithm, and the XGBoost 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 lasso regression, and the proposed work is XGBoost regression for price prediction. The total number of samples evaluated on the proposed methodology is 75 in each of the two groups. Attributes are the crucial part of showing the prediction price of the fare. The required samples for this analysis are done using G power calculation. The minimum power of the analysis is fixed as 0.8, and the maximum accepted error is fixed as 0.5 [10, 11].

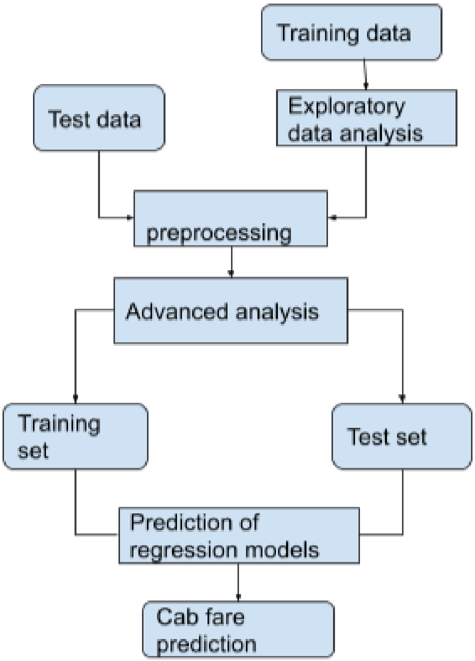

Figure 1. Architecture diagram of the proposed system were training and testing data goes through advanced analysis and in the last step prediction of the regression analysis is done using XGBoost

2.1. *LASSO Regression Algorithm*

Lasso regression is the existing algorithm in this work. A Novel exploratory data analysis is applied to analyze the input data and summarize their main characteristics. The training dataset goes through the novel exploratory data analysis to extract the main feature for data extraction. Lasso regression is a linear regression algorithm that uses shrinkage; shrinkage is taken where data values are shrunk towards a central point similar to the mean. It mainly considers simple and sparse models. This regression is mainly suitable for models which show high levels of multicollinearity. It mainly performs L1 regularization, which gives a penalty equal to the correct value of magnitude coefficients. This method puts a constraint on the addition of the absolute values of the model then, the sum has to be less than the fixed value to apply a shrinkage where it penalizes the coefficients of regression variables, shrinking them to zero [12].

2.2. *Extreme Gradient Boost Algorithm*

XGBoost is the proposed algorithm in this article. Figure 1 shows the architecture diagram of the process done in the proposed algorithm. The testing procedure includes training the dataset before continuing and after testing it, training and evaluating the algorithms. The testing procedure includes training the dataset before continuing and after testing it, training and evaluating the algorithms.

Here the XGBoost also runs as the tree sets where the data is divided into three sets. In XGBoost, even when the trees get inaccurate predictions, the algorithm pushes to execute them and re-execute them until the accurate values occur. These weak learners are regression trees, each tree maps an input data to one of its leaves which contains the continuous score. XGBoost minimizes regularised object function that combines a convex loss function. It adds the new trees that predict residuals combined with previous trees to make the final prediction. [13] The detailed process of the XGboost algorithm, where the dataset is divided into trees and executed [14].

The software and hardware requirements are included in the testing setup for the implementation. The necessary tools are Jupyter notebook or Google Colab and the programming language utilized in Python programming. The minimum hardware requirements are a Windows 7 or 8, or 10 operating system, a 1GHz processor, and 1 GB of RAM. Github repository was used to gather the input data set. Both dependent and independent attributes are present in the input datasets. **Table 1** shows an example of the input dataset obtained from the Kaggle website.

Table 1. Sample Input Dataset

fare_a mount	Pickup _datetime	Pickup _longitude	Pickup _latitude	Dropoff _longitude	Dropoff _latitude	Passenge r _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

SPSS version 21 was used to compare parameters like r squared and MSE statistically. Dependent attributes are fare_amount, which will be present only in the training dataset. Independent attributes are pickup_datetime, pickup_logitude, pickup_latitude, dropoff_longitude, dropoff_latitude, and passenger_count, which will exist in both the data sets. Analysis was done for r-squared and MSE. An independent sample T-test was done to calculate the r-squared value and Mean Square Error.

3. Results

In this study, we observed that XGBoost Algorithms have a slightly better r-squared value than Lasso regression Algorithm ($p<0.001$, Independent sample t-test). When the algorithms are compared, XGBoost has a higher r-square value of 72.62% compared to Lasso regression with 70.47%. Furthermore, the mean square error of XGboost (53.21%) is lesser than Lasso regression (54.39%). As there is a marginal difference in accuracy, XGBoost is statistically better when compared to Lasso 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 value is almost the same as in both of the algorithms in the case of training data. According to the results achieved in training data, there is a better improvement in the MSE value of XGBoost (5.270) compared to Lasso Regression (5.735). Figure 2b gives the comparative analysis of Test data for the performance evaluation parameters r square, MSE, RMSE, and RMSLE.

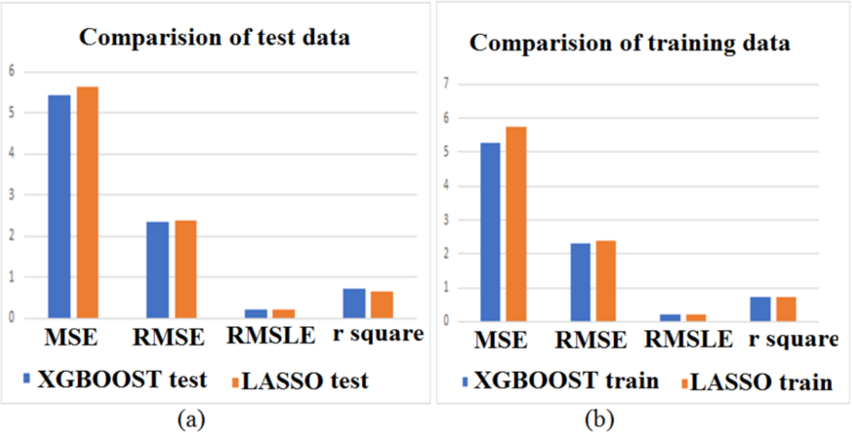

Figure 2a and 2b. Comparative analysis of Test and training data for the performance evaluation parameters r square, MSE, RMSE and RMSLE

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

	Training Data		Testing Data	
	XGBoost	**Lasso regression**	**XGBoost**	**Lasso regression**
MSE	5.270	5.735	5.437	5.620
RMSE	2.295	2.394	2.331	2.370
RMSLE	0.217	0.228	0.223	0.224
R square	0.735	0.711	0.719	0.643

From **Table 3**, it is observed that there is a slight significant increase in r-square values in both the algorithms in the case of testing data. According to the results achieved in testing data, there is a slight improvement in the MSE value of XGBoost (5.437) compared to Lasso Regression(5.62). Since testing data is considered for the results, we can prove that XGBoost can accurately predict the price.

Table 3. 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	XGBOOST	10	72.62	.640	.202
	LASSO	10	70.47	.637	.201
MSE	XGBOOST	10	53.21	1.957	.619
	LASSO	10	54.39	2.164	.684

A brief descriptive statistical analysis was performed to obtain Mean, Std. Deviation and Std. Error Mean for r-squared, and MSE values of XGBoost Algorithm and Lasso regression Algorithm are presented in Table 3. An independent sample t-test was performed with a fixed confidence level to obtain the t-test Equality of Means presented in **Table 4**.

Figure 3 gives the Bar chart representing the comparison of XGBoost and Lasso regression in terms of r-squared and MSE. The mean accuracy for XGBoost is lesser than Lasso regression and the standard deviation of XGBoost with E-LSB is better than Lasso.

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]

		Levene's Test for Equality of variances		t-test for Equality of Means					95% Confidence interval of the Difference	
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
r-square	Equal variances assumed	.241	.630	7.508	18	.011	2.144	.280	1.54403	2.743
	Equal variances not assumed			7.508	18.00	.011	2.144	.280	1.54403	2.743
MSE	Equal variances assumed	0.316	.581	-1.274	18	.219	-1.176	.934	-3.11482	.76421
	Equal variances not assumed			-1.274	17.823	.219	-1.176	.934	-3.11482	.76421

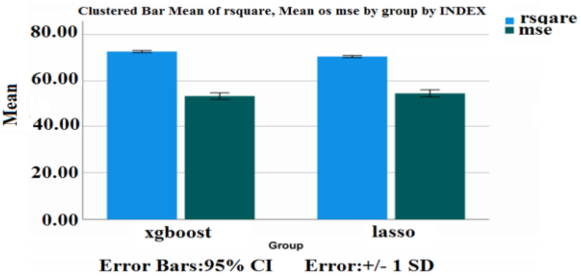

Figure 3. Bar chart representing the comparison XGBoost and Lasso regression in terms of r-square and MSE

4. Discussion

The XGBoost runs as the tree sets where the data is divided into three sets and are executed. After the execution, the weak learners who show inaccurate values are again executed until they show the accurate values because of this particular property [15]; XGBoost is slightly better than the Lasso algorithm, and the significance value is less than 0.630. The average r-square value of XGBoost is 72.62%. XGBoost is used for mining applications. The research uses ten years of historical data on stock market indices. It investigates how XGBoost differs from the remaining techniques [16]. It discusses the regularization techniques that the methods offer and their effect on the techniques. It also shows why the XGBoost is superior to the remaining techniques [3]. Similar findings are that the XGBoost provides more accurate values or predictions, outperforms other state-of-the-art predictions, and is capable of capturing a decent amount of variations [17]. There are no opposite findings observed in this work.

The limitation of this work is that when there is an over fitted model, it performs worse on the testing dataset. Although the study results are slightly better in both experimental and statistical analysis, work limitations exist. The Lasso regression fails to make the grouped selection; it takes only one variable from a group and ignores the remaining values. As a result, future work could include improving the algorithm to compute dynamic ride-sharing during peak traffic hours. To deal with this complexity, deep neural networks can be used.

5. Conclusion

These results obtained showed a slightly better accuracy standard for producing a near accurate estimation result. Based on the significance value (0.630) achieved through SPSS. XGBoost mean accuracy is 72.622% and Lasso regression mean accuracy is 70.478%. The Mean Accuracy error of XGBoost was also lower when compared to the Lasso regression Algorithm. Thus, the XGBoost algorithm has slightly better accuracy when compared to the Lasso regression algorithm.

References

[1] Swamynathan M. Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. Apress, 2019.

[2] United States. Civil Aeronautics Board, Burnstein MH. The Effect of Discout Fares and Charter Operations on Yields and Operating Revenues in Transatlantic Operations. 1973.

[3] Brownlee J. XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. Machine Learning Mastery, 2016.

[4] Gumus M, Kiran MS. Crude oil price forecasting using XGBoost. 2017 International Conference on Computer Science and Engineering (UBMK). Epub ahead of print 2017. DOI: 10.1109/ubmk.2017.8093500.

[5] Zhou Y, Li T, Shi J, et al. A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices. Complexity 2019; 2019: 1–15.

[6] Sriramya P, Karthika RA. Solving Grid Scheduling Problems Using Selective Breeding Algorithm. First International Conference on Sustainable Technologies for Computational Intelligence 2020; 581–591.

[7] Nielsen JP, Asimit A, Kyriakou I. Machine Learning in Insurance. MDPI, 2020.

[8] Chen Z, Li C, Sun W. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics 2020; 365: 112395.

[9] Khusna W, Murfi H. An analysis of the proportion of feature subsampling on XGBoost - A case study of claim prediction in car insurance. INTERNATIONAL CONFERENCE ON SCIENCE AND APPLIED SCIENCE (ICSAS2020). Epub ahead of print 2020. DOI: 10.1063/5.0031366.

[10] Dieckmeyer M, Rayudu NM, Yeung LY, et al. Prediction of incident vertebral fractures in routine MDCT: Comparison of global texture features, 3D finite element parameters and volumetric BMD. Eur J Radiol 2021; 141: 109827.

[11] Baudier E. Minimax Behaviour and Price Prediction. Risk and Uncertainty 1968; 283–310.

[12] Chan-Lau MJA. Lasso Regressions and Forecasting Models in Applied Stress Testing. International Monetary Fund, 2017.

[13] Ghosh A. Prediction of gain-of-function and loss-of-function mutations using Combined Annotation Dependent Depletion (CADD) .DOI: 10.26226/morressier.595a9c56d462b80296c9fb08.

[14] Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. DOI: 10.7287/peerj.preprints.2911.

[15] Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters 2020; 136: 190–197.

[16] Zhang Y, Wang Y, Xu J, et al. Comparison of Prediction Models for Acute Kidney Injury Among Patients with Hepatobiliary Malignancies Based on XGBoost and LASSO-Logistic Algorithms. Int J Gen Med 2021; 14: 1325–1335.

[17] Malhotra S. Python machine learning based detection of Parkinson's Disease using XGBoost. DOI: 10.14293/s2199-1006.1.sor-.ppwx1hp.v1.