Amazon EC2 Price Prediction Using Machine Learning

Samindar Vibhute[*],Prof. S. S. Redekar[**]

[*]Computer Science & Engineering,Ashokrao Mane Group of Institutes, Wathar (India)
[**]Computer Science & Engineering, Ashokrao Mane Group of Institutes, Wathar (India)

Abstract-:
Amazon Ec2 is Amazon elastic cloud compute. Ec2 is cloud service provided by amazon.com. Users can bid for spot instances and aquire the cloud for that much time. By using the LSTM algorithm we can predict the spot prices effectively with compare to other previous works which were in regression random forest, logistic regression, SVC RBF, SVC Poly. Users can plot the graph of simple moving average, rolling mean and standard deviation, time series and range slider which gives you information about the valatality of that instance. And last we cxan predict the instance price one day ahead or one weeek ahead.
Index Terms-Amazon Price Prediction, Random Forest, Spot Price

I.INTRODUCTION:

The on incorporate scalability feature of cloud computing force cloud carriers to overestimate their assets to meet top load need of its clients. Which occurs at unique time durations and may not to be overlap. thanks to over-estimation, a significant range of cloud assets are idle at some point of off top hours. Cloud carriers additionally face the disturbance of allocating assets, preserving seeable user's unique process necessities and statistics middle ability. differing types of users, a pair of styles of necessities, similarly alleviate the useful resource allocation hassle. Also, concern cloud assets differ thanks to today's utilization based totally totally pricing plans and market condition. so as to manage those require fluctuations, greater bendy pricing plans are required to market assets in step with actual time marketplace need. Spot pricing became delivered with the help of using Amazon EC2 in December 2009 to limit operational cost.

Amazon gives three pricing modules , all requiring a price from some cents to three dollars, keep with hour, to keep with strolling example. The module offer unique assurances concerning while times is also released and terminated. Paying an once a year price customers purchase the capacity to release one reserved example anytime they want. Clients also can additionally as a substitute pick to forgo the once a year price and take a look at to shop for an on imply example after they need it, at a more robust hourly price and to not be employing a assure that launching is accessible at any given time. Both reserved and on-demand for times still be lively till terminated with the help of using the client. Most inexpensive pricing version is spot example, which offers no assure concerning both release and termination time. While setting missive of invitation for a distinct segment example, customers bid the foremost hourly charge they're inclined to purchase obtaining it . The request is accepted if the bid is healthier than the spot charge, otherwise it waits. Amazon publishes a fresh spot charge and launches all ready example requests with a most charge exceeding this value, the days will run till customers terminate them or the spot charge will increase above their most charge.

II. Literature Review:

Several works focus on use of machine learning and ensemble methods for solving prediction problems of various applications.
In L. Zhang et al. brought a hybrid approach, via way of means of incorporating multi output guide vector regression and particle swarm optimization, for c programming language forecasting of the carbon futures costs. Specifically, we look into the feasibility of forecasting the 2 bounds (maximum and lowest value) of carbon futures costs collection concurrently via way of means of MSVR-PSO with a few ability predictors that have robust effect on carbon futures costs. The proposed MSVRPSO approach and 5 decided on competition are evolved over the duration from August, 2010, to June, 2013, and their out-of-pattern prediction performances are established over the duration from June 2013, to November, 2014. According to the experimental results, conclusions may be drawn: (1) the proposed MSVR-PSO approach has the better forecasting overall performance relative to 5 competitions, indicating that it's far a promising opportunity for c programming language forecasting of carbon futures costs; (2) introducing a few ability predictors that have robust have an impact on carbon futures costs.

In "Improved short-term load forecasting using bagged neural networks" paper, a hybrid approach inclusive of Prophet Version and LSTM version and BPNN version is provided to forecast short-time period electric load. The proposed technique takes the gain of every version to forecast through making use of now no longer most effective linear algorithm of the electric load however additionally makes use of the non-linear algorithm found in electric load to enhance the forecasting accuracy. The overall performance of the proposed version is proven through forecasting actual time electric load data. Simulation outcomes corroborate that the proposed hybrid version has the bottom cost of RMSE, MAE, and MAPE in contrast to standalone fashions like LSTM and Prophet in addition to hybrid fashions including hybrid ARIMA SVM. The proposed hybrid version in day beforehand forecast time horizon on common forecasts 81.36 (RMSE), 0.91% (MAPE), and 80.11 (MAE).

In "Price prediction and insurance for online auctions", we use PDAs because they can be described and compared using "hard" features/specifications (memory size, speed, screens type, operating system). In contrast, "soft" products such as clothing items don't have the same kinds of attributes that can be used to compare different kinds of items. Features such as size, material and

colour do exist but they are not the kind of attributes that "define" the style of the product. To apply the algorithms in that context we can use ideas described in some earlier work to first of all extract product attributes from free-text descriptions of products available online , and then use these attributes as part of the learning process. This would extend the applicability of our approach to "soft" products such as apparel, fashion items, antiques, and collectibles.

The approach to statistic relies on past market prices. To compute an unlimited amount of information, simulation approaches are often quite expensive. Machine Learning is amongst the foremost widely used techniques for forecasting time series-based prices of the spot case. Machine learning may be a great improvement over other techniques used for forecasting. Without actually being programmed explicitly, software applications tend to become more accurate, it's through the categories of algorithms which basically is machine learning. Algorithms built to receive computer file and predicting output supported statistical analysis also update outputs as new data becomes available, this basically is machine learning.

III. Proposed System:

This study aims to compare existing systems and compare with LSTM Algorithm. It can predict the trend of amazon ec2 and predict the spot instances for one day ahead and one week ahead. We apply LSTM algorithm and plotting graphs for simple moving average, volume traded, valatality and comparison in actual and predicted data. By observing the results we use 55-60% data for greater accuracy in prediction.

MACHINE LEARNING APPROACHES FOR PREDICTION:

Simple Linear Regression:

Linear regression is a machine learning algorithm which is under supervised learning. The regression task is performed by this algorithm. Based on independent variables, regression models predict an Actual value. To determine the relationship between variables and forecasts, it is mainly used.

Simple Linear Regression:-

A Simple Linear Regression algorithm uses one independent variable to predict the value of a numerical dependent variable.

$$y=\beta 0+\beta 1x+€$$

Y is the predicted independent variable value.
B0 is the intercept of the predicted value y when x is o
B1 is the regression coefficient variable that is how anyone except y value as x value goes high.
X is the stand-alone variable that is independent.

Multiple linear regression formula:

$$y=\beta 0+\beta 1X1+\ldots\ldots+\beta nXn$$

y= it is a predicted value of one of the dependent variable
β0 = y-intercept (value of y when all other parameters are set to 0)
β1X1=the regression coefficient of the first independent variable ()
…= Same for how many independent values you are testing
βnXn= the regression coefficient of the last independent variable
€= model error – how much variation in our estimation.

LSTM Architecture:

It works by using special gates to allow each LSTM layer to take information from both previous layers and the current layer.The main advantage of this is that it allows each LSTM cell to remember patterns for a certain amount of time.The thing to be noted is that LSTM can remember important information and at the same time forget irrelevant information. LSTM is a modified version of recurrent neural networks. This is like vanishing gradient problem of the RNN algorithm.LSTM is well suited to classifying, processing, and predicting time series with unknown lags. A back-propagation algorithm is used to train the model. RNNs eliminate gradient data.

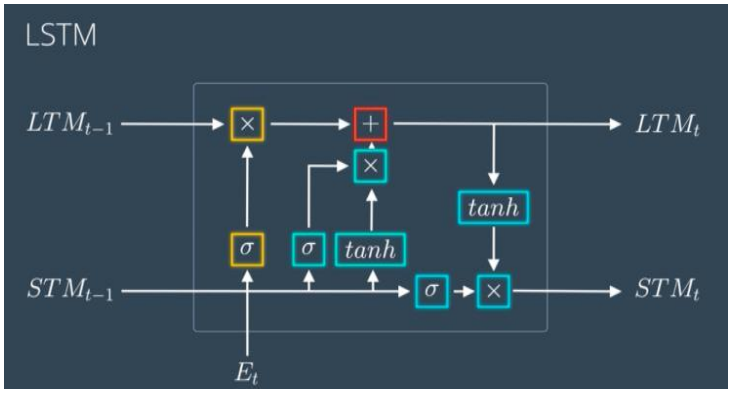


Fig1. LSTM Architecture

[Forget gate – Forget information which is not useful
Learn gate – Current information stored as next gate
Short term memory – Load previous data
Long term memory – STM and learn gate combination]

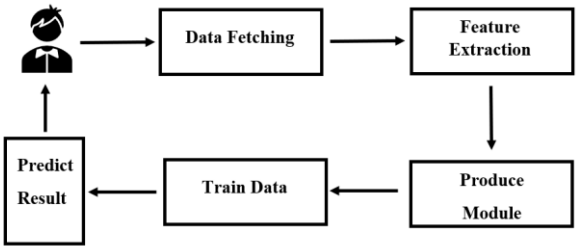


Fig2. Proposed Work Flow

Flow Chart of the proposed system

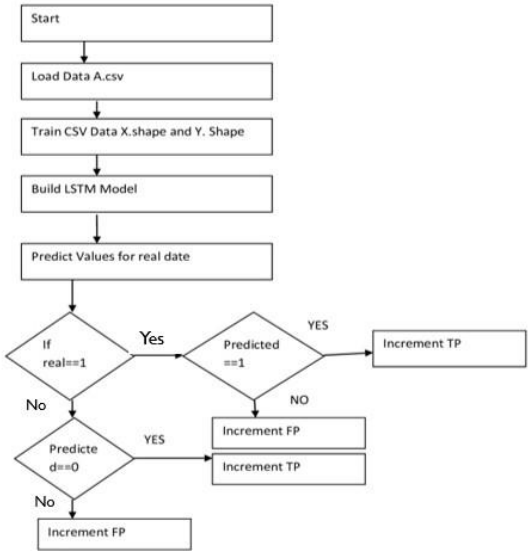


Fig3. Flow Chart of Proposed System

Fig3. Shows a Flowchart of the Proposed System.
First load the data in the CSV file then train this CSV data like X shape and Y shape. After that, build LSTM model. For real data, we have to predict some values like if the real value is equal to 1 and the prediction is equal to 1 then increment true positive TP. If R is equal to 1 and prediction is not b equal to one, then increment false positive. If real is equal to 0 and predict is zero, then increment true positive TP. If real is equals to zero and predict is equal to one, then increment false positive.

The Confusion Matrix shows the difference between Actual and Predicted values. A table-like structure measures the performance of our Machine Learning classification model.A True Positive is a value which was both actual and predicted to be positive. The value was actually negative, but was misinterpreted as positive. Also known as a Type I error.

Objective of Work:
1. To review the current literature to understand the existing methods
2. To analyze the existing methods for finding amazon prediction with the Current Date.
3.To collect and train the data.
4. To develop LSTM algorithms for price prediction and find the best result.
5. To test the accuracy of the developed algorithm using confusion Matrix.
6. To validate the results

Methods of Implementation:

Pre processing Data:
Analysis of spot price instance history data. AWS is composed of several regions and availability zones

Region	Name	Launch Date	Zones
us-east-1	US East (N. Virginia)	2006	5
us-west-2	US West (Oregon)	2011	3
us-west-1	US West (N. California)	2009	3
eu-west-1	EU (Ireland)	2007	3
eu-central-1	EU (Frankfurt)	2014	2
ap-southeast-	Asia Pacific (Singapore)	2010	2
ap-northeast-1	Asia Pacific (Tokyo)	2011	3
ap-southeast-	Asia Pacific (Sydney)	2012	2
sa-east-1	South America (Sao Paulo)	2011	3
ap-northeast-2	Asia Pacific (Seoul)	2016	2

Table1.Regionwise Dataset

Amazon Web Service is composed of several regions and availability zones as shown in Table to cut back the consequences of outages and provide facility to launch instances in regions that are closer to user thereby reducing latency. Each region has multiple redundant availability zones that provide fault tolerance to cloud resources.

Fitting features in Dataset:
Spot price history data contains:
(1) region/availability zone
(2) instance type
(3) platform/operating system
(4) price
(5)Time

Training Dataset Model:
Small variation in OOB Error is observed when training dataset size is between 7 to 14 daysIncreasing the RRFs training data size beyond 14 days leads to high OOB Error.This clearly shows that patterns in spot prices change considerably within a short period of time and only alittle window size of past history traces is required to accurately predict spot prices. We take optimal amount of past spot price history adequate 7 days for training dataset where prediction error are minimal. The dataset size incorporates the effect of all important feature variables in spot pricing-namely hour, day and weekday.Small variation in OOB Error is observed when training dataset size is varied between 7 and 14 days

IV. Mathematical Formulation:
Precision:- Out of all positive classes predicted correctly by the model, how many were true? The number of correct outputs provided by the model. The formula below calculates it.

$$\text{Precision} = \frac{\text{TP}}{TP + FP}$$

Recall: - This measures our model's ability to correctly predict out of total positive classes. Recall levels must be as high as possible.

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

F1 score is statistical measure to rate performance. It is defined as harmonic mean between precision and recall. This showed F1 score is between 0-50 . (0 Shows the lowest performance and Highest shows the Better performance.)

$$\textbf{F1Score} = \frac{2*Recall*Precision}{Recall+Precision}$$

V. RESULTS:

1. LSTM Network

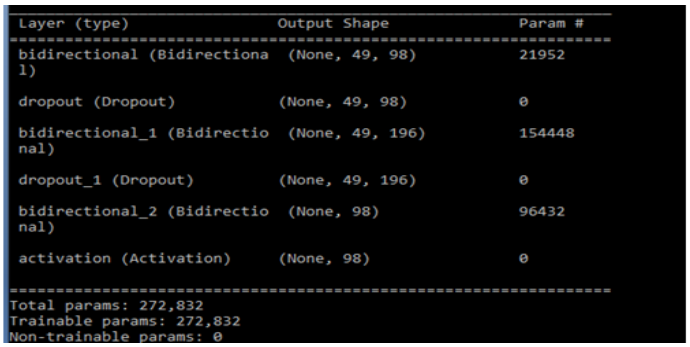


Fig5.LSTM Network

The Fig5. The LSM network is shown. Neural network architecture is developed in both directions by long-short-term memory (bi-lstm). An LSM only allows input to flow forward, while a bidirectional neural network allows input to flow both forward and backward. Flowing intake in both directions preserves past and future information in bi-directional systems. A bidirectional LSTM effect can be seen in the above result. Additionally, it displays layer outputs and parameters per layer, as well as total parameters, trainable parameters, and no trainable parameters in the evaluation section.

2. Raw Data Loading:

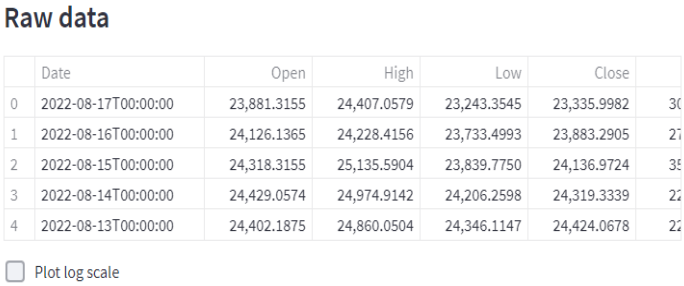


Fig6.Raw Data Loading

Fig6. Shows Raw Data extraction that data from Dataset is loading with the till date and generate the raw data graph. Every data has values of Open, High, Low, and Close. In March 2022, we got the data as shown in the above figure. The range slider shown above shows the graph of year vs. volume. The range slider is developed by using the data collected from yahoo data finance. The tick data were collected through a Web scraper that pulled data from the APIs of the Finance Amazon exchange from March 13, 2022 to January 17, 2023, resulting in approximately 50,000 unique trading records, including Price, Trading Volume, Open, Close, High, and Low points.

3. Forecast Plot for 365 Days

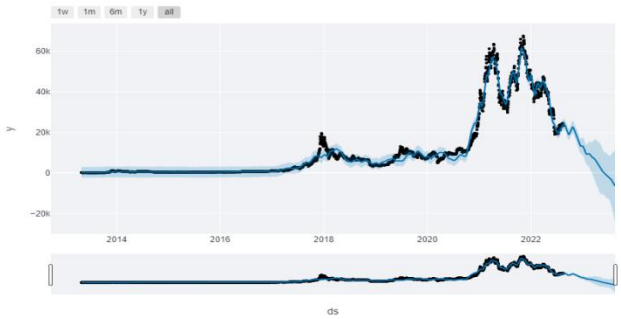


Fig7. Forecast Plot for 365 Days

The Fig7. Shows the forecast plot for the next 365 days. In July 2021, we got the data as shown in figure 7. The data collected from yahoo fiancé is a trend, yhat_lower, yhat_upper, trend_upper and lower. The range slider shown above in Fig 7 is the graph of day's vs. volume. The range slider is developed by using the data collected from yahoo data finance shows the forecasting plot with the next 365 days' values means it will predict the values for every month values will increase or decrease.

4. Forecast Component

Fig8. Forecast Component

Fig 8. This chart shows the price distribution yearly, weekly, and monthly. In this paper, we observed moderate growth between January 2013 and January 2023, followed by a rapid rise to a peak at the beginning of 2022. The first 75% of the price datasets were used for training, and the remaining 25% for testing.

5. Closing Price Trend

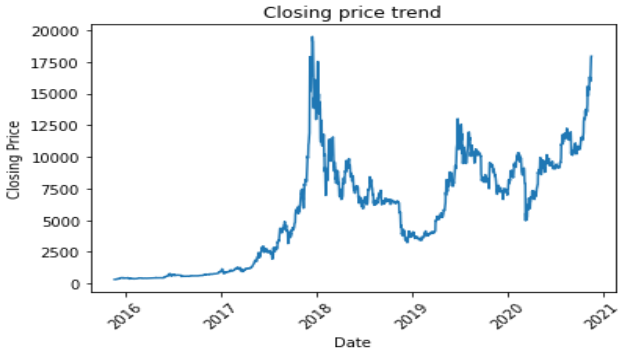


Fig 9. Closing Price Trend

The Fig.9 Shows Closing Price Trend .The damage is noted the worth of a stock at the top of the commerce hours. It's typically noted by traders as a benchmark value to match the performance with historical costs

6. Time Series with Range Slider

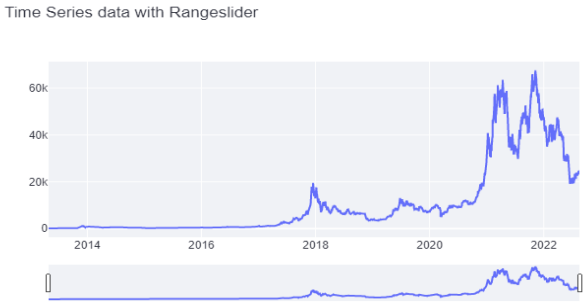


Fig 10. Time Series with Range Slider

The Fig 10 Shows Time Series with Range Slider. The range slider allows how the stock market values is increases or decreases with time ranges.

7. Confusion Matrix

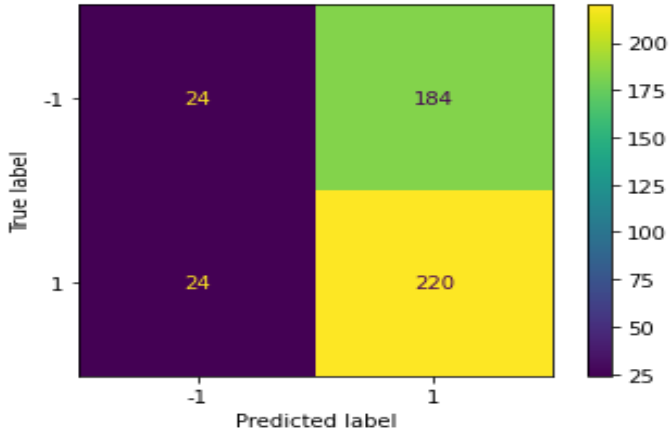
Confusion matrixes are visual representations of actual versus predicted values. A table-like structure measures the performance of our Machine Learning classification model.A confusion matrix for a binary classification problem looks like this

Precision:- Out of all positive classes predicted correctly by the model, how many were true? The number of correct outputs provided by the model. The formula below calculates it.

$$\text{Precision} = \frac{\text{TP}}{TP + FP}$$

Recall: - This measures our model's ability to correctly predict out of total positive classes. Recall levels must be as high as possible.

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$



	precision	recall	f1-score	support
-1	0.51	0.12	0.19	208
1	0.55	0.91	0.68	244
accuracy			0.54	452
macro avg	0.53	0.51	0.43	452
weighted avg	0.53	0.54	0.45	452

Fig11. Confusion Matrix

8. Simple Moving Average

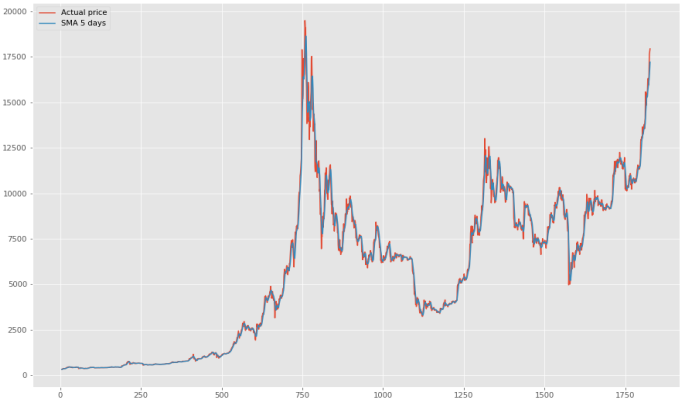


Fig.12 Simple Moving Average

The Fig12. Shows the Simple moving average.Moving averages area unit one amongst the core indicators in technical analysis, and there are a unit a range of various versions. SMA is that the best moving average to construct. it's merely the typical value over the required amount. the typical is named "moving" as a result of its premeditated on the chart bar by bar, forming a line that moves on the chart because the average worth changes.

9. Rolling Mean & Standard deviation

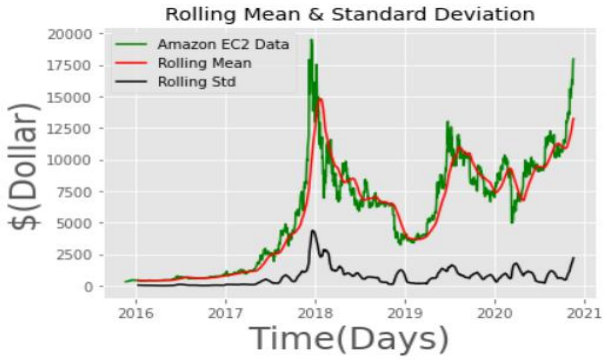


Fig 13. Rolling Mean Vs. Standard deviation

The Fig13. Shows the Rolling Mean Vs. Standard Deviation. Standard deviation is commonly wont to live the volatility of returns from assets or methods as a result of it will facilitate live volatility. Higher volatility is mostly related to the next risk of losses, thus investors need to examine higher returns from funds that generate higher volatility. For instance, a indicator fund ought to have comparatively low variance compared with a growth fund.

10. Expected VS Predicted Forecasting

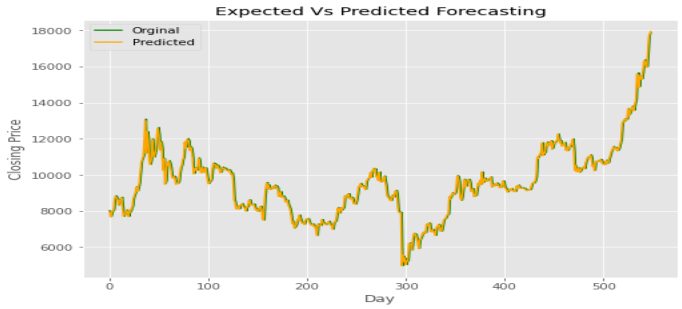


Fig14. Expected VS Predicted Forecasting

The Fig 14. Shows Expected Vs Predicted Forecasting. The green line indicates the original close price. The orange line indicates the predicted price.

11. F1 Score Comparison

The Fig15. Shows F1 score comparison of the used algorithm. This paper uses the random forest & regression algorithms.

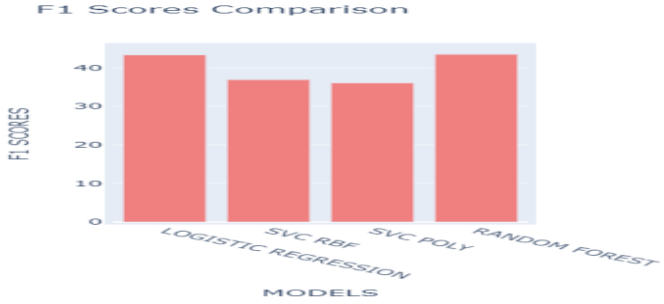


Fig 15 F1 Score Comparison

ACKNOWLEDGMET:

My sincere gratitude goes out to Professor S.S. Redekar and Prof. P.S.Powar, my research supervisors, for their patient guidance, enthusiasm.I owe a debt of deepest gratitude to our esteemed Prof. S.S.RedekarHead of Department of Computer Science and Engineering, Ashokrao Mane Group of Institution for his guidance, support, motivation and encouragement during the course of this training work. His readiness for consultation at all times, his educative comments, his concern during this task has been invaluable.I take opportunity to thank Prof.P.S.Powarof Computer Science and Engineering Department for their cooperation. It is our pleasure to thank all those who have rendered their help during the period of my research work.

CONCLUSION:

By this study and research we can definitely conclude that by applying LSTM Approach, we can definitely predict better results than other algorithms like regression random forest, SVC POLY, SVC RBF and logistic regression. We can also conclude that by training 55-60% of raw data we can predict better results. And hence we definitely sure to say LSTM algorithm is new version of prediction using machine learning approach. And by this prediction users can bid for cheaper rates for the instances. We also can predict the trend of that instances which will go higher or lower in future.

REFERENCES:

1. Lu Zhang, Junbiao Zhang, Tao Xiong,1 Chiao Su "Interval Forecasting of Carbon Futures Prices Using a NovelHybrid Approach with Exogenous Variables" Received 7 April 2017; Accepted 5 July 2017; Published 9 August 2017.

2. S. Khwaja, M. Naeem, A. Anpalagan, A. Venetsanopoulos, and B. Venkatesh, "Improved short-term load forecasting using bagged neural networks," Electric Power Systems Research, vol. 125, pp. 109–115, Aug. 2015.

3 R. Ghani, "Price prediction and insurance for online auctions," Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD "05, 2005.

4 M. Mattess, C. Vecchiola, and R. Buyya. Managing peak loads by leasing cloud infrastructure services from a spot market. In 2010 IEEE 12th International Conference on High Performance Computing and Communications (HPCC), pages 180– 188, Sept 2010. doi: 10.1109/HPCC.2010.77.

5Sangho Yi, Junyoung Heo, Yookun Cho, and Jiman Hong. Taking point decision mechanism for page-level incremental checkpointing based on cost analysis of process execution time. J. Inf. Sci. Eng., 23(5):1325– 1337, 2007. URLhttp://www.iis.sinica.edu.tw/page/jise/ 2007 /200709_01.html.

6.Dennys C. A. Mallqui, R. Fernandes Predicting the direction, maximum, minimum and closing prices of daily exchange rate using machine learning techniques Received date : 24 January 2018 Revised date : 29 August 2018 Accepted date : 20 November 2018.

7.Samiksha Marne et. Al Predicting Price of Cryptocurrency - A Deep Learning Approach International Journal of Engineering Research & Technology (IJERT) NTASU - 2020 Conference Proceedings

AUTHORS:

FirstAuthor–Samindar Vibhute, M. Tech Student, Computer Science & Engineering, Ashokrao Mane Group of Institutes, Wathar(India), andsamindarvibhute.007@gmail.com.

SecondAuthor–Prof. S. S. Redekar, Assistant Professor,Computer Science & Engineering, Ashokrao Mane Group of Institutes(India),and

ssr@amgoi.edu.in