



# **iJRASET**

International Journal For Research in  
Applied Science and Engineering Technology



---

# **INTERNATIONAL JOURNAL FOR RESEARCH**

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume: 9      Issue: XII      Month of publication: December 2021**

**DOI: <https://doi.org/10.22214/ijraset.2021.39599>**

**[www.ijraset.com](http://www.ijraset.com)**

**Call:  08813907089**

**E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)**

# Analysis of Trends in Stock Market

Abhinav N Patil<sup>1</sup>, Ananya P Mitra<sup>2</sup>, Karan A Paranjpe<sup>3</sup>, Sanjesh Pawale<sup>4</sup>

<sup>1, 2, 3, 4</sup>B.Tech, Dept of S&T, Vishwakarma University

**Abstract:** *In the period of AI applications for each and every fragment of examination, breaking down financial exchange costs and patterns has become more famous than previously. We have gathered information of securities exchange utilizing python back end and have proposed a far reaching customization of element designing and profound learning based model for anticipating the pattern of market, The proposed arrangement incorporates pre-handling of the securities exchange dataset, usage of various element designing strategies, joined with a redid profound learning based framework for pattern forecast, Evaluations have been directed on the models and reason that our proposed arrangement beats because of element designing that we constructed. Framework accomplishes decent exactness without the distant element of information overfitting and because of definite assessment of term lengths and strategies, this adds to the stock examination research local area both in monetary and specialized areas.*

**Keywords:** *Machine-learning, Deep-learning, Python, API, Metrics, Feature Engineering, Stock Market.*

## I. INTRODUCTION

Stock market is one of the significant fields that financial backers are especially drawn to and hence market value expectation is consistently a hotly debated issue for analysts from both monetary and specialized areas. In this exploration, our goal is to assemble a condition-of-craftsmanship expectation model for value pattern which centers around transient value pattern forecast.

Stock market expectation has been a functioning space of exploration for quite a while. The Efficient Market Hypothesis (EMH) states that securities exchange costs are to a great extent driven by new data and follow an arbitrary walk design. However this theory is broadly acknowledged by the examination local area as a focal worldview administering the business sectors as a rule, a few group have endeavored to separate examples in the manner securities exchanges act and react to outer upgrades.

In this paper, we test a theory dependent on the reason of conduct financial matters, that the feelings and opinions of people influence their dynamic cycle, hence prompting an immediate relationship between's "public opinion" and "market opinion". We perform feeling investigation on freely accessible Twitter information to track down the public state of mind and the level of participation into 2 classes – Positive, Negative. We utilize these opinions and past United Airlines stock qualities to foresee future stock developments and afterward utilize the anticipated qualities in our portfolio the executives methodology.

The remainder of the paper is coordinated as follows. Second some portion of presentation is the writing review we performed on recently acknowledged papers. The second and third area momentarily examines our overall methodology towards taking care of the issue and the accompanying segments talk about the singular parts more meticulously. In Section 4, we momentarily talk about the dataset that we have utilized for this paper and information preprocessing measures took on. Area 5 fundamentally has the execution part and examines the opinion investigation method created by us with the end goal of this paper. It likewise incorporates, exhaustively, the means we followed to foresee United Airlines stock qualities utilizing our opinion investigation results and presents our discoveries. In Section 6, we examine the future work that can be performed on this venture and casing an end on our exploration.

## II. LITERATURE SURVEY

The latest work [reference 1] proposes a similar hybrid neural network architecture, integrating a CNN with a bidirectional long short-term memory to predict the stock market index, while researchers frequently proposed different neural network solution architectures, it bought further discussions about the topic if the high cost of training such models is worth the result or not.

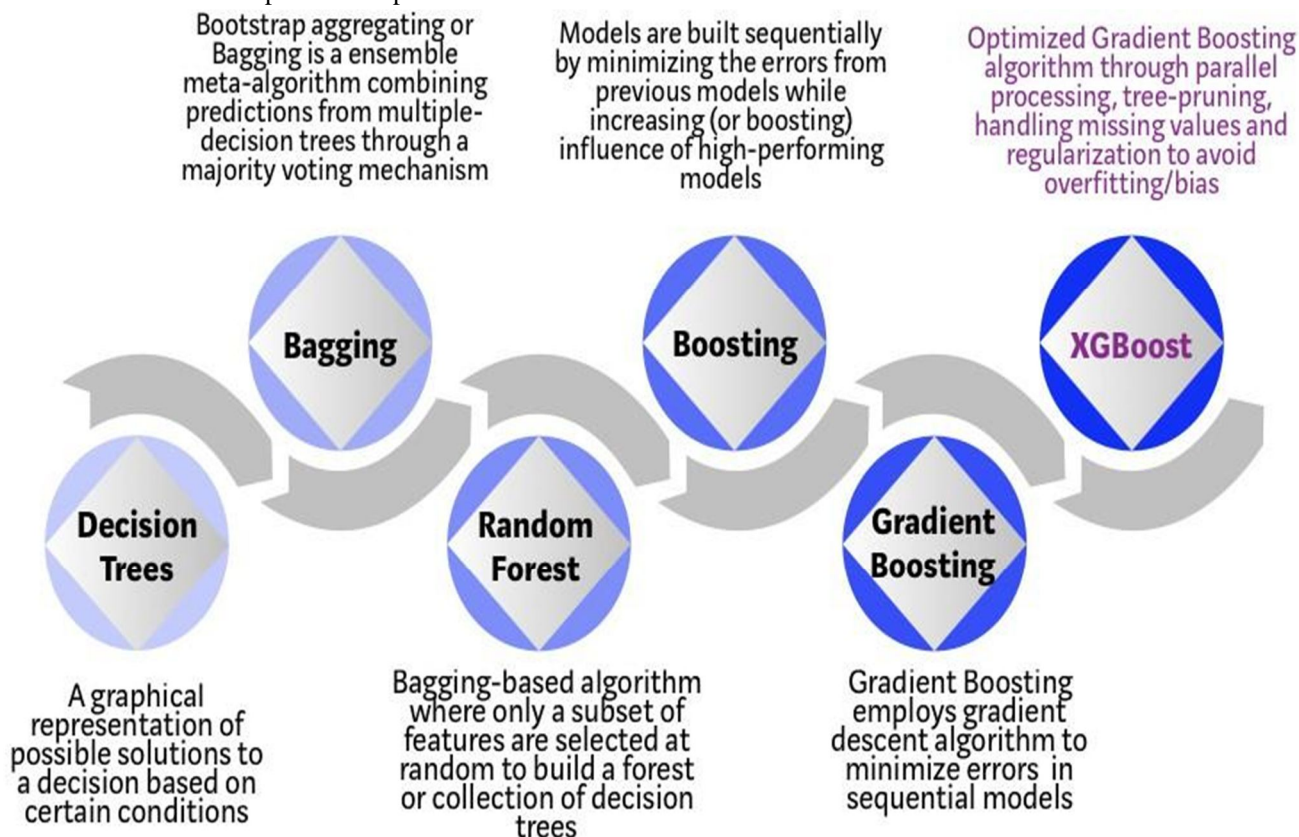
There are three key contributions of our work: (1) a new dataset extracted and cleansed (2) a comprehensive feature engineering and (3) customized machine learning models.

As per the productive market speculation, in the monetary market, stock costs quickly change in accordance with new data once it becomes public, making the expectation of financial exchange's developments unthinkable [2]. This judgment is right for conventional investigations utilizing numerous direct relapse examination. Be that as it may, stock costs will become unsurprising assuming the dynamic and nonlinear connections in financial exchanges are shown. This is the inspiration for applying neural organizations to monetary time series investigation. Up to this point, use of neural organizations to anticipate the developments of monetary business sectors has stood out for some researchers.

White [13] proposed neural organization demonstrating and learning strategies to look for and unravel nonlinear normalities of resource value developments and anticipated IBM day by day stock costs, Chenoweth [4] concentrated on the exchanging framework dependent on the future upsides of every day S&P 500 record. Zhang [15] completed different examinations in four sorts of Stock including Lujiazui, Fuhua Industrial, Changchun Hualian, Shanghai Petrochemical utilizing BP network with unique info information. Wang [12] presented the sliding window method and RBF network into non-straight time series, the outcome was great. A portion of these explores only involved the previous upsides of the stock list as the contribution of neural organizations in order to acquire gauges, while others utilized extra monetary elements as information sources

### III. XGBOOST

XGBoost is a choice tree-based Machine Learning calculation that utilizes a slope supporting structure. In forecast issues including unstructured information (pictures, text, and so forth) counterfeit neural organizations will generally beat any remaining calculations or systems. Notwithstanding, with regards to little to-medium organized/even information, choice tree based calculations are viewed as top tier at the present time.



### IV. REGRESSION, RANDOM FOREST

Random forest calculations anticipate the result esteems dependent on input highlights from the information took care of in the framework. The go-to philosophy is the calculation assembles a model on the highlights of preparing information and involving the model to foresee the incentive for new information.

Random Forest is a Supervised Machine Learning Algorithm that is utilized broadly in Classification and Regression issues. It constructs choice trees on various examples and takes their larger part vote in favor of order and normal in the event of relapse.

One of the main elements of the Random Forest Algorithm is that it can deal with the informational collection containing consistent factors as on account of relapse and clear cut factors as on account of characterization. It performs better for characterization issues.



### V. DATASET

We are using the yahoo finance datasets for model training. An user interface has been created so that the user can input the company's stock name of his/her choice and get the trends and final prediction price.

Date	Tweets
2021-12-19	RT JasonPester Ive been thinking about that 27...
2021-12-19	RT JasonPester In case you already forgot and ...
2021-12-19	RT JasonPester If you got your holiday shoppin...
2021-12-19	RT JasonPester Geez Apple AAPL demanded a TaxH...
2021-12-19	RT spcryptocalls ADA 40 so far in 10 HoursHODL...
...	...
2021-12-10	saylor elerianm Yall had AAPL stock though
2021-12-10	WNBD TSLA OTCMarkets AAPL GUYS TIME TO LOOK A...
2021-12-10	ctm MEMES amp MOVERS TRADING TSLA AAPL AMC ...
2021-12-10	We are consolidating at yet another high of 17...
2021-12-10	Apple heading for 3T valuation AAPL

869 rows × 2 columns

Date	Tweets
12/19/2021	RT
12/19/2021	RT
12/19/2021	RT
12/19/2021	RT @JasonPester: Geez, #Apple #AAPL demanded a #TaxHoliday to return funds to its home country USA... but Tim Cook just throws \$275 Billion&#x20ac;
12/19/2021	RT
12/19/2021	Market traders use the weekend and head to the #TRICKTRADES #Youtube channel, actual professional market education you can apply to your trading. #SPY #QQQ #AAPL #FB #TSLA #AMD #TWTR #MU #AMZN #MSFT #GOOGL #
12/19/2021	1743 funds increasing their holding in \$AAPL vs last quarter!! #aapl #stocks #investing Find out more: https://t.co/mzNQioP8S8
12/19/2021	The Big Five #FAANG Stock and Sector Service is now posted for Sunday #AMZN #AAPL #FB #NFLX #TSLA #GOOGL and Sector Funds - #IBB #IYR #XLE #XLF #XLK #XRT (link: https://t.co/bihMz399my) #trading #investing #Fi
12/19/2021	A beautiful Sunday morning. What&#x201c;s your plan for the crypto market today? #AirdropKing #BitcoinCrash #Binance #BTC #Algorand #Cryptos #1inch #AAVE #100xgem #ANKR #Airdropspector #AdidasNFT #AAPL #Algorithms
12/19/2021	102 more funds holding \$AAPL vs last quarter!! #aapl #stocks #investing Find out more: https://t.co/mzNQioP8S8
12/19/2021	2279 funds holding \$AAPL as one of their top ten holdings. This is up 34 vs last quarter!! #aapl #stocks #investing Find out more: https://t.co/mzNQioP8S8
12/19/2021	RT @zmtts134: DWAC STOCK CHART UPDATE!! QUICK TECHNICAL ANALYSIS ON DIGITAL WORLD ACQU... https://t.co/kAAUr2omzF via @YouTube #dwac #DWAC&#x20ac;
12/19/2021	DWAC STOCK CHART UPDATE!! QUICK TECHNICAL ANALYSIS ON DIGITAL WORLD ACQU... https://t.co/kAAUr2omzF via @YouTube #dwac #DWAC #truthSocial #Truth #Trump #Trump2024 #TrumpWasRight #TRUMP2024ToSaveAm
12/19/2021	Subscrib
12/19/2021	2279 funds holding \$AAPL as one of their top ten holdings. This is up 34 vs last quarter!! #aapl #stocks #investing Find out more: https://t.co/mzNQioP8S8
12/19/2021	i always believe in my \$TSLA, \$NIO, \$AAPL! Friends, do you always trust the stocks you hold? #TSLA /#NIO /#AAPL! https://t.co/CN8Dv5YEab
12/19/2021	\$SPY \$TSLA \$QQQ \$AMZN #Aliens #ufotwitter #AAPL #BTC #CelebrationBowl! #DoctorStrange #ETH #FIFArabCup #GVSvFB #honkwtselfieday #PL2022 #johandjohannes #kpoppredictions #LennyaleThata watch this video, has!
12/19/2021	RT @Wen_Moon_Token: Buy buy buy. #bsc #PancakeSwap #CryptocurrencyNews #Crypto #ToTheMoon #Airdrop #BNB #BTC #jeet #ApeSwap #altcoins #wenm&#x20ac;
12/19/2021	BTC Leveraged Trade -- Looking at this. Your thoughts? #AAVE #AAPL #ADA #1inch https://t.co/fRrF25j9l
12/19/2021	Buy buy buy. #bsc #PancakeSwap #CryptocurrencyNews #Crypto #ToTheMoon #Airdrop #BNB #BTC #jeet #ApeSwap #altcoins #wenmoon #AAPL #ETH #NFTCommunity #Metaverse @elonmusk @1goonrich @BillGates https://t.
12/19/2021	RT
12/19/2021	RT
12/19/2021	#ADA
12/19/2021	\$AAPL not too sure if it has made up it's mind which way its going to go. I think you may see this drop and test the resistance line or even the \$165.70 level before moving back to the upside. #Apple #AAPL #stocks #stockmarket

### VI. IMPLEMENTATION AND RESULTS

#### A. Data Preprocessing

The dataset obtained from Kaggle initially had a pickle file type which was converted to a csv file for ease of data preprocessing and data training.

The columns necessary for further training and analysis were extracted such as:

- 1) Closing Price
- 2) Adjusted closing price
- 3) Tweets

During the conversion of the pickle file to a csv file, the encoding format used is UTF-8 for ease of translation of any Unicode character to the matching binary string and vice-versa.

```
In [2]: # df = df.asfreq('H')

stocks_dataf = pd.read_pickle('C:/Users/Lenovo/Downloads/Twitter_Dataset.pkl')
stocks_dataf.columns=['closing_price', 'adj_close_price', 'Tweets']

stocks_dataf.to_csv('C:/Users/Lenovo/Downloads/Twitter_Dataset.csv', encoding='utf8', index=False)
stocks_dataf=pd.read_csv('C:/Users/Lenovo/Downloads/Twitter_Dataset.csv')
stocks_dataf.head()
```

### B. Sentiment Analysis

Sentiment analysis was a significant piece of our answer since the result of this examination was used as a learning factor in our prescient model. There has been a great deal of examination going on in grouping a piece of text as one or the other positive, negative or unbiased, and we will be taking the advantage of this in our venture. Utilizing feeling investigation, we will break down the tweets as being good or negative and taking their compound outcomes forward.

### C. Filtering Punctuations

We shall filter out unnecessary punctuations from each tweet as they do not contribute to the analysis and might also deteriorate the resultant sentiment.

```
In [3]: stocks_dataf['adj_close_price'] = stocks_dataf['adj_close_price'].apply(np.int64)
stocks_dataf = stocks_dataf[['adj_close_price', 'Tweets']]
stocks_dataf['Tweets'] = stocks_dataf['Tweets'].map(lambda x: x.lstrip('.-'))
```

### D. Sentiment Analyzer

We shall now apply the SentimentIntensityAnalyzer for the computation of the positivity and negativity of the tweets.

The SentimentIntensityAnalyzer class from VADER, gives a sentiment intensity score to sentences.

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a vocabulary and rule-based sentiment investigation device that is explicitly sensitive to feelings communicated in web-based media. VADER utilizes a mix of feeling dictionary for example a rundown of lexical elements (e.g., words) which are for the most part named by their semantic direction as one or the other positive or negative. VADER not just tells about the Positivity and Negativity score yet additionally enlightens us regarding how positive or negative a feeling is.

If the polarity score is:

- 1) Below 0, it is termed as a Negative Tweet
- 2) Above 0 and approximately greater than 0.5, it is termed as a Positive Tweet
- 3) Otherwise it becomes a Neutral Tweet.

Finally, a compound, i.e. an overall score is calculated to determine the ultimate sentiment of the tweet to be either Positive or Negative.

```
In [7]: from nltk.sentiment.vader import SentimentIntensityAnalyzer
import unicodedata
sentiment_i_a = SentimentIntensityAnalyzer()
for indexx, row in stocks_dataf.T.iteritems():
    try:
        sentence_i = unicodedata.normalize('NFKD', stocks_dataf.loc[indexx, 'Tweets'])
        sentence_sentiment = sentiment_i_a.polarity_scores(sentence_i)
        dataframe.at[indexx, 'Comp'] = sentence_sentiment['compound']
        dataframe.at[indexx, 'Negative'] = sentence_sentiment['neg']
        dataframe.at[indexx, 'Neutral'] = sentence_sentiment['neu']
        dataframe.at[indexx, 'Positive'] = sentence_sentiment['pos']
    except TypeError:
        print (stocks_dataf.loc[indexx, 'Tweets'])
        print (indexx)
```

```
In [8]: dataframe.head()
```

Out[8]:

	adj_close_price	Comp	Negative	Neutral	Positive
0	12469	-0.9814	0.159	0.749	0.093
1	12472	-0.8521	0.116	0.785	0.099
2	12474	-0.9993	0.198	0.737	0.065
3	12480	-0.9982	0.131	0.806	0.062
4	12398	-0.9901	0.124	0.794	0.082

To visualize the sentiment analysis score, the following snippet plots a pie chart displaying the percentage of positive and negative scores.

```
In [9]: posi=0
nega=0
for i in range (0,len(dataframe)):
    get_val=dataframe.Comp[i]
    if(float(get_val)<(-0.99)):
        nega=nega+1
    if(float(get_val)>(-0.99)):
        posi=posi+1
posper=(posi/(len(dataframe)))*100
negper=(nega/(len(dataframe)))*100
print("% of positive tweets= ",posper)
print("% of negative tweets= ",negper)
arr=np.asarray([posper,negper], dtype=int)
mlpt.pie(arr,labels=['positive','negative'])
mlpt.plot()
```

Random forest improves on bagging because it decorrelates the trees with the introduction of splitting on a random subset of features. This means that at each split of the tree, the model considers only a small subset of features rather than all of the features of the model. That is, from the set of available features  $n$ , a subset of  $m$  features ( $m = \text{square root of } n$ ) are selected at random.

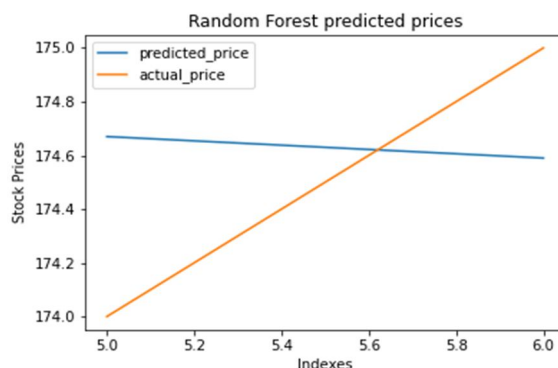
```
from treeinterpreter import treeinterpreter as ti
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
import sys
rf = RandomForestRegressor()
rf.fit(numpy_df_train, y_train)
prediction, bias, contributions = ti.predict(rf, numpy_df_test)
```

```
idx=np.arange(index[0],(index[1]+1))
```

```
predictions_df_ = pd.DataFrame(data=prediction[:, index = idx, columns=['Prices'])
predictions_df_
```

	Prices
5	174.67
6	174.59

```
ax = predictions_df_.rename(columns={"Prices": "predicted_price"}).plot(title='Random Forest predicted prices')#predicted value
ax.set_xlabel("Indexes")
ax.set_ylabel("Stock Prices")
fig = y_test.rename(columns={"Prices": "actual_price"}).plot(ax = ax).get_figure()#actual value
fig.savefig("random forest2.png")
```



Think of XGBoost as gradient boosting on ‘steroids’ (well it is called ‘Extreme Gradient Boosting’ for a reason!). It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time.

```

from xgboost import XGBRegressor
%matplotlib inline
from sklearn import metrics

model = XGBRegressor(objective='reg:squarederror', n_estimators=1000)
model.fit(numpy_df_train, y_train)

XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, enable_categorical=False,
              gamma=0, gpu_id=-1, importance_type=None,
              interaction_constraints='', learning_rate=0.300000012,
              max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
              monotone_constraints=()), n_estimators=1000, n_jobs=12,
              num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
              reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact',
              validate_parameters=1, verbosity=None)

predicted = model.predict(numpy_df_test)
predicted

array([171.00093, 178.99806], dtype=float32)

print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predicted)))
print('R-squared :', metrics.r2_score(y_test, predicted))

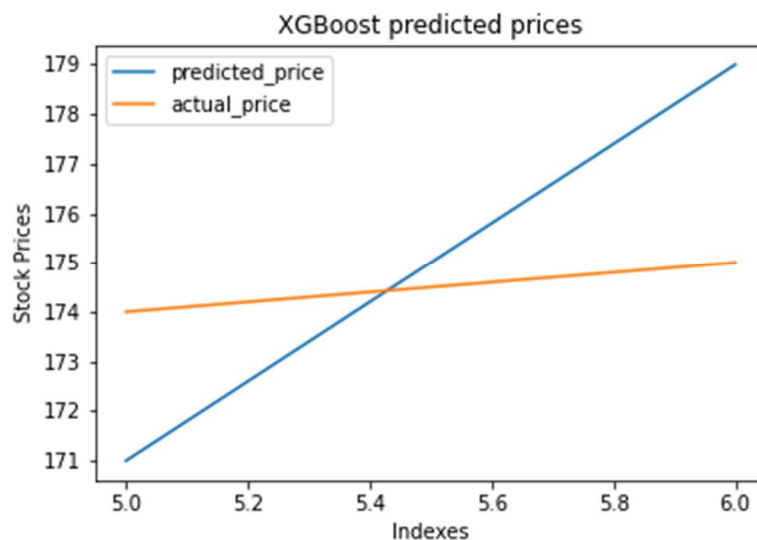
Root Mean Squared Error: 3.534042796515338
R-squared : -48.9578339504078

predictions_df_2 = pd.DataFrame(data=predicted[:,], index = idx, columns=['Prices'])

ax = predictions_df_2.rename(columns={"Prices": "predicted_price"}).plot(title='XGBoost predicted prices')#predicted value
ax.set_xlabel("Indexes")
ax.set_ylabel("Stock Prices")
fig = y_test.rename(columns={"Prices": "actual_price"}).plot(ax = ax).get_figure()#actual value
fig.savefig("XGB.png")

```





## VII. CONCLUSION

We conclude our research paper with the facts we reticulated against the comparison of both algorithms we select and on the project as whole. Random forest is expected to always give better result when bagging technique is involved on large scale with in accordance with bootstrapping technique, however, when it comes to minimizing error coefficient for exponential data and gradient boosting, XGBoost is the way to go since it essentially provides more than just gradient tree and using just one algorithm like XGBoost, user can essentially catch out-of-core computing as well as parallel multiple tree nodding together. Trend study in stock has benefitted using these two approaches as the former gives option to bag and bootstrap tree on advance level and the later provides excellent collection of tools to prune and have multiple approaches on how we model the data on every node level. This paper concludes with a starting base for further research on multi-level tree grading algorithms used to analyze stock trends and predict the movement on a bigger and complex scale.

## VIII. ACKNOWLEDGMENTS

We would like to express our heartfelt appreciation to Prof. Sanjesh Pawale who helped us with the project report. Due to our lack of experience and competence, we ran into a few roadblocks during the project, but he helped us overcome them and integrate our vision in the end. We would like to thank him for his guidance and comments, which helped our entire team understand the project report's finer details. Finally, we would like to express our gratitude to Vishwakarma University's administration for enabling us to participate in these activities.

## REFERENCES

- [1] McNally S, Roche J, Caton S. Predicting the price of bitcoin using machine learning. In: Proceedings—26th Euromicro international conference on parallel, distributed, and network-based processing, PDP 2018. pp. 339–43. <https://doi.org/10.1109/PDP2018.2018.00060>.
- [2] Nagar A, Hahsler M. News sentiment analysis using R to predict stock market trends. 2012. <http://past.rinfinance.com/agenda/2012/talk/Nagar+Hahsler.pdf>. Accessed 20 July 2019.
- [3] Ni LP, Ni ZW, Gao YZ. Stock trend prediction based on fractal feature selection and support vector machine. Expert Syst Appl. 2011;38(5):5569–76. <https://doi.org/10.1016/j.eswa.2010.10.079>.
- [4] Pang X, Zhou Y, Wang P, Lin W, Chang V. An innovative neural network approach for stock market prediction. J Supercomput. 2018. <https://doi.org/10.1007/s11227-017-2228-y>.
- [5] Piramuthu S. Evaluating feature selection methods for learning in data mining applications. Eur J Oper Res. 2004;156(2):483–94. [https://doi.org/10.1016/S0377-2217\(02\)00911-6](https://doi.org/10.1016/S0377-2217(02)00911-6).
- [6] Shen J, Shafiq MO. Learning mobile application usage—a deep learning approach. ICMLA. 2019;2019:287–92.
- [7] Sirignano J, Cont R. Universal features of price formation in financial markets: perspectives from deep learning. Ssrn. 2018. <https://doi.org/10.2139/ssrn.3141294>.
- [8] Thakur M, Kumar D. A hybrid financial trading support system using multi-category classifiers and random forest. Appl Soft Comput J. 2018;67:337–49. <https://doi.org/10.1016/j.asoc.2018.03.006>.
- [9] Tsai CF, Hsiao YC. Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis Support Syst. 2010;50(1):258–69. <https://doi.org/10.1016/j.dss.2010.08.028>.





- [10] Wang X, Lin W. Stock market prediction using neural networks: does trading volume help in short-term prediction?. n.d.
- [11] Zubair M, Fazal A, Fazal R, Kundi M. Development of stock market trend prediction system using multiple regression. Computational and mathematical organization theory. Berlin: Springer US; 2019. <https://doi.org/10.1007/s10588-019-09292-7>.
- [12] Zhang S. Architectural complexity measures of recurrent neural networks, (NIPS). 2016. pp. 1–9.
- [13] Tushare API. 2018. <https://github.com/waditu/tushare>. Accessed 1 July 2019.
- [14] Wang X, Lin W. Stock market prediction using neural networks: does trading volume help in short-term prediction?. n.d.
- [15] Tsai CF, Hsiao YC. Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis Support Syst. 2010;50(1):258–69. <https://doi.org/10.1016/j.dss.2010.08.028>.
- [16] Shih D. A study of early warning system in volume burst risk assessment of stock with Big Data platform. In: 2019 IEEE 4th international conference on cloud computing and big data analysis (ICCCBDA). 2019. pp. 244–8.
- [17] Qiu M, Song Y. Predicting the direction of stock market index movement using an optimized artificial neural network model. PLoS ONE. 2016;11(5):e0155133.
- [18] Malkiel BG, Fama EF. Efficient capital markets: a review of theory and empirical work. J Finance. 1970;25(2):383–417.
- [19] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. Knowl Based Syst. 2018;164:163–73. <https://doi.org/10.1016/j.knosys.2018.10.034>.
- [20] Lee MC. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Syst Appl. 2009;36(8):10896–904. <https://doi.org/10.1016/j.eswa.2009.02.038>.



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)