International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

FINANCIAL ANALYSIS USING DATA MINING

Akshaya Rajagopalan[1], Vidhya. S[2], P.R Bhavya[3]

[1]*Akshaya Rajagopalan, Dept. of Computer Application, M.O.P Vaishnav College for women, TamilNadu, India*
[2]*Vidhya.S, Dept. of Computer Application, M.O.P Vaishnav College for women, TamilNadu, India*
[3]*P.R Bhavya, Dept. of Computer Application, M.O.P Vaishnav College for women, TamilNadu, India*

Abstract - *Data mining refers to extracting or mining knowledge from large amounts of data. Data mining is used in prediction of various diseases, stock market trends and weather conditions. The paper uses data mining algorithms to predict whether the NASDAQ share market will increase every month. The prediction is done using KNN classification, Rule based classification and Deep Learning technique. The three algorithms are compared with each other for determining the one with higher accuracy.*

Key Words: (Size 10 & Bold) Key word1, Key word2, Key word3, etc (Minimum 5 to 8 key words)…

1. INTRODUCTION

Data mining refers to extracting information from huge sets of data. In general, we can say that data mining is the process of mining knowledge from data. The information or knowledge extracted can be used for various applications such as market analysis, fraud detection, customer retention and many more.

Stock market trends are very unstable. This leads to high risk in stock returns. The stock returns for an investor is important as they invest their money on stocks to gain profit. In order to gain high stock returns, knowledge on stock trends is necessary. There are market experts who predict when an investor can buy, sell or hold shares. The prediction however is difficult due to complex and unstable nature of stock market. On the other hand, there is huge amount of data generated by the stock market which can be used for prediction. There are various ways to predict stock market trends.

There are various algorithms for classification and clustering data. The paper uses the data set of NASDAQ share market of 10 years for every month from Jan-2009 to Jan-2019 from it official website. The dataset was given label as increasing which has either yes or no as its value based on previous month and current month price of stock.

1.1 Literature Review

[1]Examining the stability and persistence of the long-short performance of fundamental signals over time. This analysis is important because previous studies (e.g., Sullivan, Timmermann, and White 2001) argue that the analysis of sub period stability is a remedy against data mining.

[2]Tokenization: Each news article or financial report document is split into meaningful words called tokens. They are: Data standardization, Stop-word-removal, Stemming, Abbreviation processing, Filter tokens.

[3]To develop the new models, two methods were used: DA and decision trees.
Using these methods, we identified the most significant variables from the basic sets in terms of prediction bankruptcy/non bankruptcy of companies within the Slovak business environment.

[4]There are often inconsistencies in how entity names and addresses are entered, in addition to outright errors and typos. There is implicit semantic knowledge included in a name, e.g., a name may contain National association or State Bank of in its name. This complicates matching based on a similarity score that is obtained using some edit distance metric.

[5]Mentioning that the final purpose behind text classification processes is to facilitate text analyzing in order to recognize hidden textual patterns with reduced manual efforts. To do this, we focus on utilization of classified financial footnotes in the last phase of the research, which consists of representation of classified output in each category.

[6]If for denser structure trees all effective features in first prediction are selected by the proposed hybrid model, results in better accuracy as ‘‘BF Tree’’, ‘‘LAD Tree’’, and ‘‘FT Tree’’. Otherwise, it is possible that accuracy drops, like ‘‘CART and Rep’’ TREEs. The selected features have different effect on the accuracy of forecasting. Some trees with large structure, such as J48 Graph and J48 Tree.

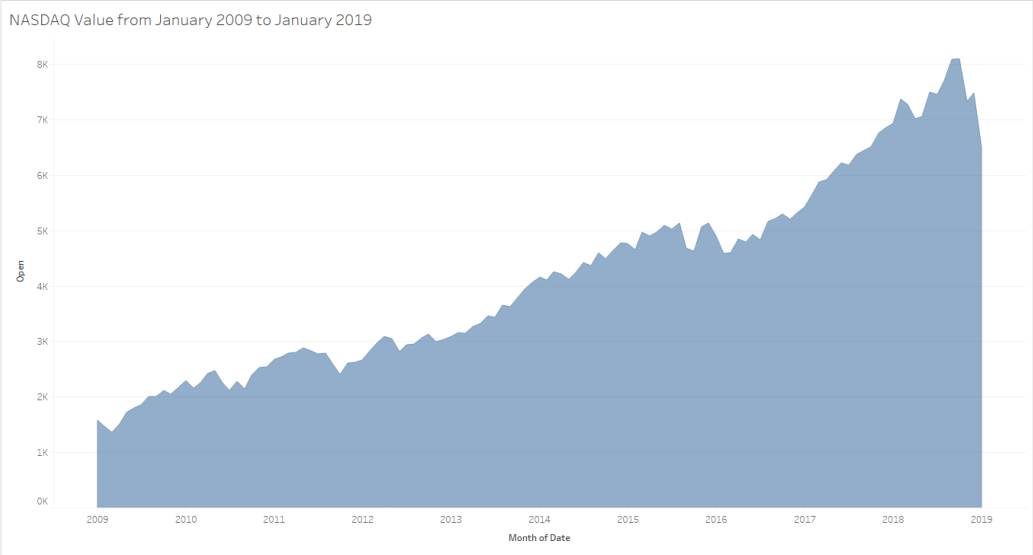
[7]The LDA results from the previous section show that Comcast has serious customer service issues. We want to further explore these issues with sentiment analysis, which is a supervised learning method and can also be used to visualize and summarize big data. Sentiment analysis can be carried out in various models, such as maximum entropy.

[8]As cryptocurrencies continue to develop they have merited the attention of policy-makers and regulators for a host of differing reasons. However, there have been three specific situations that have been quite unique to cryptocurrency markets.

[9] Always aiming to increase credit volume while reducing default risks. Therefore, credit scoring analyses are crucial to aid faster decision making, reduce the costs of loan analysis, monitor existing accounts more closely, predict default risks, and ensure that institutions can detect possible risks while developing their competitiveness

2. ANALYSIS

The analysis is done using Data mining tool named RapidMiner Studio. It is about the usage of three main algorithms: K Nearest Neighbour, Rule Based Classification and Deep Learning for predicting whether the NASDAQ market will increase every month or not and with higher accuracy. The label for prediction is increasing with values yes or no. This data was then imported to RapidMiner and many algorithms were applied. In the end only three algorithms had significant accuracy. These three are compared and the one with higher accuracy is best suited for this prediction. The working, results and performance of each algorithm is given below.



K NEAREST NEIGHBOURS

K Nearest Neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g. distance functions). It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. In KNN-classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k- nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 121 examples with 7 dimensions of the following classes:
  no
  yes
```

accuracy: 71.28% +/- 13.89% (micro average: 71.07%)

	true no	true yes	class precision
pred. no	86	31	73.50%
pred. yes	4	0	0.00%
class recall	95.56%	0.00%	

The table above shows the how accurate the result is. The first parameter is with respect to prediction no which is true no: 86 and true yes: 31. The second parameter is prediction yes which is true no: 4.Therefore, the total accuracy is 71.28% with error of +/- 13.89%.

RULE BASED CLASSIFICATION

The term rule-based classification can be used to refer to any classification scheme that makes use of IF-THEN rules for class prediction. The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent. The antecedent part the condition consists of one or more attribute tests and these tests are logically ANDED. The consequent part consists of class prediction.

RuleModel

```
else no  (90 / 31)

correct: 90 out of 121 training examples.
```

Table View Plot View

accuracy: 74.49% +/- 11.64% (micro average: 74.38%)

	true no	true yes	class precision
pred. no	90	31	74.38%
pred. yes	0	0	0.00%
class recall	100.00%	0.00%	

The table above shows the how accurate the result is. The first parameter is with respect to prediction no which is true no: 90and true yes: 31. The second parameter is prediction yes which is true no: 0 and true yes: 0.Therefore, the total accuracy is 74.49% with error of +/- 11.64%.

DEEP LEARNING

Deep Learning is a set of Machine Learning algorithms which have one or more hidden layers in Neural Networks. Now Deep Learning systems are used for almost any tasks other Machine Learning algorithms are used for. These can be Classification, Dimensionality Reduction, Object Recognition, Clustering etc.

Model Metrics Type: Binomial
Description: Metrics reported on full training frame
model id: rm-h2o-model-deep_learning-847701
frame id: rm-h2o-frame-deep_learning-797658
MSE: 0.01230316
R^2: 0.9354371
AUC: 1.0
logloss: 0.089967564
CM: Confusion Matrix (vertical: actual; across: predicted):

	no	yes	Error	Rate
no	90	0	0.0000	= 0 / 90

yes 0 31 0.0000 = 0 / 31
Totals 90 31 0.0000 = 0 / 121
Gains/Lift Table (Avg response rate: 25.62 %):
Group Cumulative Data Fraction Lower Threshold Lift Cumulative Lift Response Rate Cumulative Response Rate Capture Rate Cumulative Capture Rate Gain Cumulative Gain

1	0.01652893	0.949065	3.903226	3.903226	1.000000	1.000000	0.064516	0.064516	290.322581	290.322581
2	0.02479339	0.938766	3.903226	3.903226	1.000000	1.000000	0.032258	0.096774	290.322581	290.322581
3	0.03305785	0.936683	3.903226	3.903226	1.000000	1.000000	0.032258	0.129032	290.322581	290.322581
4	0.04132231	0.934789	3.903226	3.903226	1.000000	1.000000	0.032258	0.161290	290.322581	290.322581
5	0.05785124	0.903776	3.903226	3.903226	1.000000	1.000000	0.064516	0.225806	290.322581	290.322581
6	0.10743802	0.876076	3.903226	3.903226	1.000000	1.000000	0.193548	0.419355	290.322581	290.322581
7	0.15702479	0.839690	3.903226	3.903226	1.000000	1.000000	0.193548	0.612903	290.322581	290.322581
8	0.20661157	0.801775	3.903226	3.903226	1.000000	1.000000	0.193548	0.806452	290.322581	290.322581
9	0.30578512	0.151490	1.951613	3.270270	0.500000	0.837838	0.193548	1.000000	95.161290	227.027027
10	0.40495868	0.087053	0.000000	2.469388	0.000000	0.632653	0.000000	1.000000	-100.000000	146.938776
11	0.50413223	0.052242	0.000000	1.983607	0.000000	0.508197	0.000000	1.000000	-100.000000	98.360656
12	0.60330579	0.044419	0.000000	1.657534	0.000000	0.424658	0.000000	1.000000	-100.000000	65.753425
13	0.70247934	0.035038	0.000000	1.423529	0.000000	0.364706	0.000000	1.000000	-100.000000	42.352941
14	0.80165289	0.029978	0.000000	1.247423	0.000000	0.319588	0.000000	1.000000	-100.000000	24.742268
15	0.90082645	0.021018	0.000000	1.110092	0.000000	0.284404	0.000000	1.000000	-100.000000	11.009174
16	1.00000000	0.005519	0.000000	1.000000	0.000000	0.256198	0.000000	1.000000	-100.000000	0.000000

Status of Neuron Layers (predicting increasing, 2-class classification, bernoulli distribution, CrossEntropy loss, 44,752 weights/biases, 529.2 KB, 1,210 training samples, mini-batch size 1):
Layer Units Type Dropout L1 L2 Mean Rate Rate RMS Momentum Mean Weight Weight RMS Mean Bias Bias RMS

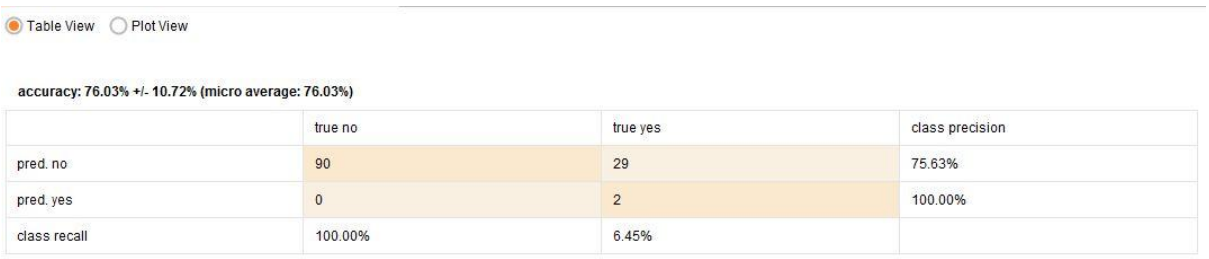
1 841 Input 0.00 %
2 50 Rectifier 0.00 % 0.000010 0.000000 0.026497 0.097324 0.000000 0.000063 0.048879 0.485324 0.071063
3 50 Rectifier 0.00 % 0.000010 0.000000 0.013653 0.026169 0.000000 -0.000503 0.143624 0.997198 0.013364
4 2 Softmax 0.000010 0.000000 0.002090 0.001478 0.000000 -0.079601 0.434169 -0.000000 0.007881

Scoring History:

Timestamp Duration Training Speed Epochs Iterations Samples Training MSE Training R^2 Training LogLoss Training AUC Training Lift Training Classification Error

2019-09-22 15:33:36 0.000 sec 0.00000 0 0.000000 NaN NaN NaN NaN NaN NaN
2019-09-22 15:33:36 0.453 sec 703 rows/sec 1.00000 1 121.000000 0.18056 0.05247 0.53647 0.76416 1.95161 0.19835
2019-09-22 15:33:37 1.078 sec 1549 rows/sec 10.00000 10 1210.000000 0.01230 0.93544 0.08997 1.00000 3.90323 0.00000

H2O version: 3.8.2.6-rm9.0.0



The above table shows the how accurate the result is. The first parameter is with respect to prediction no which is true no: 90 and true yes: 29. The second parameter is prediction yes which is true yes: 2.Therefore, the total accuracy is 76.03% with error of +/- 10.72%.

3. CONCLUSION

In this paper, dataset of NASDAQ share market has been analysed to predict whether the price will increase or not in upcoming month. The three algorithms- KNN, Rule model and Deep Learning are applied and their performances are compared. On comparing these three , Deep Learning has highest accuracy of 76.03% +/-10.72%. Rule based classifier comes next with 74.49% +/-11.64%. The least among the three is KNN with 71.28 +/-13.89%. Therefore we can conclude that Deep Learning is best suited for predicting whether the NASDAQ share market will increase each month or not.

REFERENCES

[1.]Fundamental Analysis and the Cross-Section of Stock: A Data-Mining Approach,Lingling Zheng,2019
[2].Predicting Stock Market Behavior using Data Mining Technique, Ayman Elsayed Khedr ,2017
[3]. Analysis of Impact of Using the Trend Variables on Bankruptcy Prediction Models Performance, Beáta GAVUROVÁ,2017
[4]. Research Challenges in Financial Data Modeling and Analysis,Lewis Alexander ,Sanjiv R. Das ,2017
[5].Financial Footnote Analysis: Developing a Text Mining Approach,Maryam Heidari, Carsten Felden ,2018
[6].Developing an approach to evaluate stocks by forecasting effective features with data mining ,Sasan Barak,2018.
[7.] Analyzing the impact of user-generated content on B2B Firms' stockperformance: Big data analysis with machine learning methods,Xia Liu,2019
[8.]Cryptocurrencies as a Financial Asset: A systematic analysis,Shaen Corbet,2019
[9]. -Comparison of Data Mining Classification Algorithms Determining the Default Risk, Begüm Çığşar and Deniz Ünal,2017

BIOGRAPHIES

Akshaya Rajagopalan a student of M.O.P Vaishnav College currently in her final semester pursuing BCA.

S.Vidhya a student of M.O.P Vaishnav College currently in her final semester pursuing BCA.

Bhavya.P.R a student of M.O.P Vaishnav College currently in her final semester pursuing BCA.