ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

STOCK MARKET PRICE PREDICITON BY USING MACHINE LEARNING

B.Ramesh, T.Sai Krishna, R.Divya, M.Sowmya, T.Balu Prasad
Department of Computer Science and Engineering
Aditya Institute of technology and Management, Tekkali.

ABSTRACT
The aim of the project is to examine a number of price prediction to predict future stock returns based on past returns and numerical news indicators to construct a portfolio of multiple stocks in order to diversify the risk. We do this by applying supervised learning methods for stock price forecasting by interpreting the seemingly chaotic market data.

Nowadays, as the connections between worldwide economies are tightened by globalization, external perturbations to the financial markets are no longer domestic. With evolving capital markets, more and more data is being created daily.

The intrinsic value of a company's stock is the value determined by estimating the expected future cash flows of a stock and discounting them to the present, which is known as the book value. This is distinct from the market value of the stock, that is determined by the company's stock price. This market value of a stock can deviate from the intrinsic value due to reasons unrelated to the company's fundamental operations, such as market sentiment.

The fluctuation of stock market is violent and there are many complicated financial indicators. Only few people with extensive experience and knowledge can understand the meaning of the indicators and use them to make good prediction to get fortune. Most people have to rely solely on luck to earn money from stock trading. However, the advancement in technology, provides an opportunity to gain steady fortune from stock market and also can help experts to find out the most informative indicators to make better prediction. The prediction of the market value is of paramount importance to help in maximizing the profit of stock option purchase while keeping the risk low.

INTRODUCTION

The fluctuation of stock market is violent and there are many complicated financial indicators. However, the advancement in technology, provides an opportunity to gain steady fortune from stock market and also can help experts to find out the most informative indicators to make better prediction. The prediction of the market value is of paramount importance to help in maximizing the profit of stock option purchase while keeping the risk low.

Cite this article as: B.Ramesh, T.Sai Krishna, R.Divya, M.Sowmya & T.Balu Prasad., "Stock Market Price Prediciton By Using Machine Learning", International Journal & Magazine of Engineering, Technology, Management and Research (IJMETMR), ISSN 2348-4845, Volume 9 Issue 6, June 2022, Page 33-41.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The next section of the paper will be methodology where we will explain about each process in detail. After that we will have pictorial representations of the analysis that we have made and we will also reason about the results achieved. Finally, we will define the scope of the project. We will talk about how to extend the paper to achieve more better results.

This section will give you the detailed analysis of each process involved in the project. Each sub section is mapped to one of the stages in the project.

A. Data Pre-Processing

The pre-processing stage involves

• **Data discretization**: Part of data reduction but with particular importance, especially for numerical data

• **Data transformation**: Normalization.

• **Data Cleaning**: Fill in missing values.

• **Data Integration**: Integration of data files. After the data-set is transformed into clean data-set, the data-set is divided into training and testing sets so as to evaluate. Here, the training values are taken as the more recent values. Testing data is kept as 5-10 percent of the total dataset.

B. Feature Selection and Feature Generation

We created new features from the base features which provided better insights of the data like 50 day moving average, previous day difference, etc. To prune out less useful features, in Feature Selection, we select features according to the k highest scores, with the help of a linear model for testing the effect of a single regressor, sequentially for many regressors.

We used the **SelectKBest** Algorithm, with f regression as the scorer for evaluation. Furthermore, we added **Twitters Daily Sentiment Score**, as a feature for each company based upon the users tweets about that particular company and also the tweets on that company's page.

LITERATURE SURVEY

This section consists of various existing methodologies that are currently being implemented to segment, classify and count WBCs present in microscopic images which involves image processing techniques -**Holt-Winters** is the appropriate or correct mode when the time series has trend and seasonal factors. The series was divided into three components or parts that are trend, basis and seasonality. Holt-Winters find three trend, level, and seasonal smoothening parameters. It has two variants: Additive Holt Winters Smoothening model and Multiplicative Holt-Winters model. The former is used for prediction and the latter is preferred if there are no constant seasonal variations in the series. It is mainly popular for its accuracy and in the field of prediction it has outperformed many other models. In short—term forecasts of economic development trends, Holt-Winters exponential smoothing method with the trend and seasonal fluctuations is usually used. After removing the seasonal trends from the data, the following function is taken as an input and in return, the Holt-Winters makes the pre-calculations necessary for the purpose of forecasting. All parameters required for the forecasting purpose are automatically

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

initialized based on the function data. HWStock1_ng = HoltWinters(ds,gamma = FALSE) predHW = predict(HWStock1_ng,n.ahead = 9).

An artificial neural network (ANN) is a technique inspired from biological nervous system, such as the human brain [5, 10]. It has a great ability to predict from large databases [11]. On the basis of the back—propagation algorithm, ANN is generally used to forecast the stock market. In the back—propagation algorithm, a neural network of multilayer perceptron (MLP) is used. It consists of an input layer with a set of sensor nodes as input nodes, one or more hidden layers of computation nodes and computation nodes of the output layer. These networks often use raw data and data derived from the previously discussed technical and fundamental analysis.
A Multilayer Feed forward Neural Network is a neural network with an input layer, one or more hidden layers and an output layer. Inputs correspond to each training sample measured attributes. Inputs are passed to input layer simultaneously. The weighted outputs of these units are fed to the next layer of units that make up the hidden layer simultaneously. The weighted outputs of the hidden layers act as an input to another hidden layer, etc. The hidden layers number is an arbitrary design problem. The weighted output of the last the hidden layer acts as inputs to the output layer, which predicts the networks for certain samples. Important parameters of NN are learning rate, momentum and epoch,
Back propagation is a neural network learning algorithm [13]. The back propagation network learns by processing the sample set repeatedly and comparing the network prediction with the actual output. If the residual value exceeds the threshold value, the weight of the connections is modified to reduce the MSE between the forecast value and the original value.The weights are changed from the output layer to the first hidden layer in the opposite direction. Since the changes in the weights of the connections are made in the reverse direction, the name given to the algorithm is Back propagation [14]. Use the back propagation algorithm to perform the calculations and compare the predicted output and target output. The predicted value is not closer to the actual value and the weights are modified.

Recurrent neural networks (RNN) use back propagation to learn, but their nodes have a feedback mechanism. Because of this, RNN models can predict a stock price based on recent history and is recurrent [16]. model <- trainr(Y = trainy, X = trainx, learningrate = 0.05, hidden_dim = 5, numepochs = 2600) pred <- predictr(model,testx). Advantage Previous time points to input layer contains inputs. Disadvantage An RNN makes it possible to feed those words in through a much smaller set of input nodes.

Here I referred three different ways that can be use methodologies that can be used in the implementation of the code. Each methodology has their own advantage and disadvantage such as
Artificial Neural Network has Better performance compared to regression. Lower prediction error but Prediction gets worse with increased noise variation and the parameters used to predict the stock market price is Stock closing price.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Holt-Winters also know as Support vector machine for stock prediction does not lose much accuracy when applied to a sample from outside the training sample but exaggerate to minor fluctuations in the training data which decrease the predictive ability and the parameters used are Consumer investment, net revenue, net income, price per earnings ratio of stock, consumer spending.

RNN Recurrent Neural Network is Previous time points to input layer contains inputs but It possible to feed those words in through a much smaller set of input nodes and the parameters used are Input hidden and output layers.

So I choose to implement RNN in my code and using different AI concepts.

Experimental Investigations

Dataset:

The dataset used in this paper are taken from Kaggle[14] dataset. We have used data collect from past price of the stock in the data set .

There are approximately 2,500 images for each of 5 different cell types arranged into 5 different folders. Additional 410 pre-augmented images are also accompanied with the dataset. We used 70% of dataset as training data and 30% as testing data.

Segmentation:

Segmentation subdivides an image into constituent regions or objects which means partitioning the original image into multiple segments or different regions. The goal of segmentation is to extract more features and information from the image and analyze it in a more meaningful and easier manner. It takes the Pre-processed digital image as an input and gives the attributes of an image as output which is used for feature extraction and object recognition. This is the process of assigning a label to each pixel in an image such that the pixels with the same label share certain characteristics. There are many approaches for image segmentation, but we have used the Edge-based segmentation method which is more suitable for white blood cell detection. Image segmentation is the fundamental step to analyses images and extract data from them. It is an operation of partitioning an image into a collection of connected sets of pixels. The main purpose of interest in an image which helps in an image that helps in the annotation of the object scene. There are three main approaches of image segmentation which are region-based approach, boundary approach, and edge approach that depends on two properties discontinuity and similarity. Edgebased segmentation also known as edge detection is a process of locating the edges of an image. Edges usually correspond to points in the image where the grey value changes from one pixel to the other. From these edges, we can extract the features which give the information related to the image as it was discovered that the most important information lies in the edges of an image. Canny Edge Detection is used to get optimal results.

Feature Extraction:

In this paper, the Harris Corner Detection algorithm is used as the feature extraction technique. This technique is used to extract the important features from the image. It was first published in 1988 by Chris Harris and Mike Stephens. It is an operator for corner detection

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

which is widely used in machine learning algorithms to identify corners. This algorithm can detect the edges, flat area, and corners present in the image which helps in identifying whether the cell is White blood cell or not. A corner is a point that can be viewed as a two-edge intersection, where an edge reflects a sudden change in the brightness of the image. Corners are where a slight location shift will lead to a significant change in intensity both vertical and horizontal axes.

Packages:

We have used a lot of packages to implement this project

Keras:

This is an open source library in python which can be used to call neural network architectures. As it is an open source library many contributions are made from the community of python that helps Keras stay up-to date with latest highly functioning Deep Learning algorithms. It has TensorFlow working in the backend which is a huge open source data science and machine learning library.

A Keras model consists of multiple components:

- An architecture, or configuration, which specifyies what layers the model contain, and how they're connected.
- A set of weights values (the "state of the model").
- An optimizer (defined by compiling the model).
- A set of losses and metrics

Tensorflow:

TensorFlow is an open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications. It can build and train ML models easily using intuitive high-level APIs like Keras with eager execution, which makes for immediate model iteration and easy debugging. We use Tensorflow to easily train and deploy models in the cloud, on-prem, in the browser, or on-device no matter what language you use. It has a A simple and flexible architecture to take new ideas from concept to code, to state-of-the-art models, and to publication faster.

OpenCV:

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. Being a BSDlicensed product, OpenCV makes it easy for businesses to utilize and modify the code.

The library has more than 2500 optimized algorithms, which includes a comprehensive set of both classic and state-of-the-art computer vision and machine learning algorithms. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, extract 3D models of objects, produce 3D point clouds from stereo cameras, stitch images together to produce a high resolution image of an entire scene, find similar images from an image

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

database, remove red eyes from images taken using flash, follow eye movements, recognize scenery and establish markers to overlay it with augmented reality, etc.

Pandas:

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of thePythonprogramming language.Pandas is the most widely used tool for data munging. It contains high-level data structures and manipulation tools designed to make data analysis fast and easy. Pandas package is used to implement modifications and adjustments to our data frames. Pandas provide us wide range of tools to convert our data frame dimensions and dynamics as to our preferences.

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. This package of python is by far the best available for visual data analysis. It contains many exhaustive and complex plot options that help us to easily understand and analyse our data.

Sklearn:

Scikit-learn called in python as Sklearn is written mainly in python. It uses Numpy for high performance algebra and array functions and operations. We use sklearn in our program to perform normalization of the data to normalize the commodity values and then we use r2score module to analyse the performance of our algorithms. As we are using SVM (Support Vector Machine) which is also available in the Scikit-Learn package.

Anaconda:

Anaconda is an independent and open source distribution of Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), intended to make package management and deployment easier. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

We choose Anaconda because it brings many tools used in data science and machine learning with just one install, so it is great for short and simple setup. Like Verchenav, Anaconda also uses the concept of creating environments to separate individual libraries and versions. Anaconda also introduces its own package manager, called **conda**, from where you can set up a library.

Additionally, Anaconda still has a useful interaction with pip that allows you to install any additional libraries that are not available in Anaconda Package Manager.

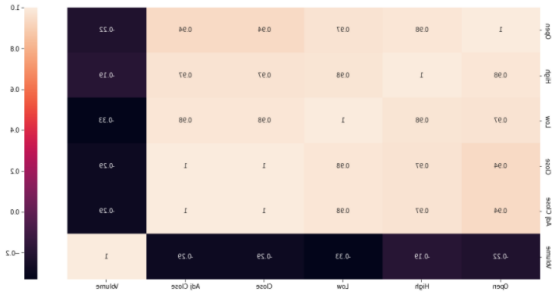
Conda:

Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux. Conda quickly installs, runs and updates packages and their dependencies. Conda easily builds, saves, loads, and switches between environments on its local computer. It was built for Python programs, but it can package and distribute software for any language.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

EXPERIMENTAL RESULTS

After the execution our model succeeded in classifying the white blood cells and counting the cells present in each category.



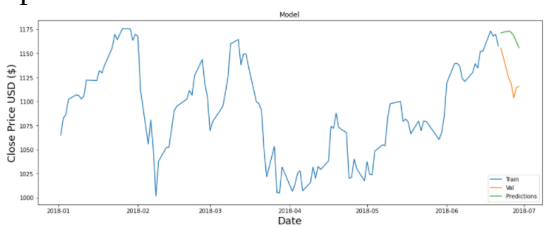
The proposed method is implemented on a system with 8 processors of Intel Core i7-8550U CPU with 2.0 GHz and 8 GB RAM. The RNN model was developed using "Keras" [15] and "Tensorflow" [16] libraries in python. We used "Anaconda Distribution" as platform to implement our model. The other libraries include Pandas, Matplotlib, Scikit Learn [17], and OpenCV [18].

. Accuracy is defined as the ratio of sample images that sample images.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

We achieved an accuracy of 97.89% through our proposed model by implementing Softmax function. We implemented different combinations of Activation functions, Optimizers, and Loss functions to compare the achieved accuracy with other models.

The activation map was displayed in the below output.



Resulted output:

Date	Close	Predictions
2018-06-22	1155.479980	1171.348511
2018-06-25	1124.810059	1173.380249
2018-06-26	1118.459961	1172.022827
2018-06-27	1103.979980	1168.291138
2018-06-28	1114.219971	1162.256104
2018-06-29	1115.650024	1156.105957

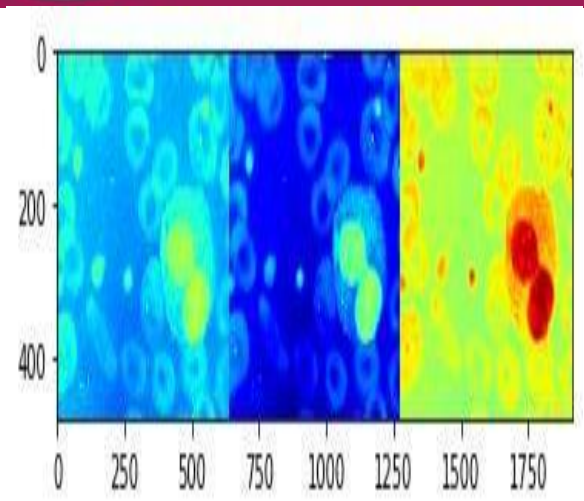
DISCUSSION OF RESULTS

We have done the classification on "Blood Cell Images" which was collected from Kaggle. All the models are trained with that dataset. The model trained is then utilized to test the data on other datasets.

Our proposed model has achieved an accuracy of 97.89%. In our model we used "Softmax" [19] function as Activation function, "RMSprop" [20] as Optimizer, and "Categorical_Crossentropy" [21] as Loss function. We implemented various combinations of Activation functions, Optimizers, and Loss functions to compare the proposed model with other models. The result is shown in the table below.

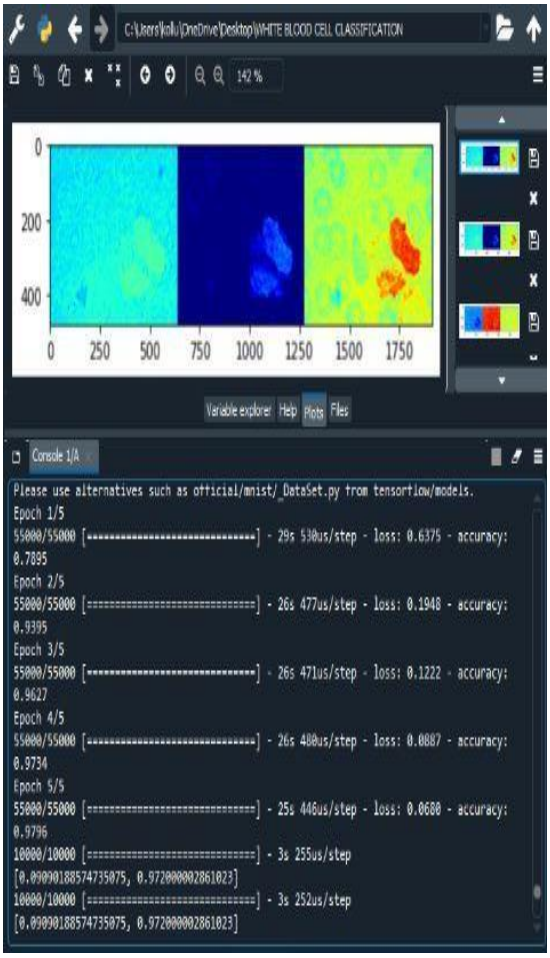
Among all the combinations Softmax, RMSprop, and Categorical_Entropy achieved highest accuracy. This is achieved because of the strength in the relationship of Softmax layers.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



The augmented images that are processed from the dataset is shown above.

The accuracy obtained from the proposed model is shown below.



CONCLUSION

In this paper we used Convolutional Neural Networks for the classification of white blood cells. Here we trained the models with the data from Kaggle. The models obtained were then used to classify the white blood cells using Convolutional Neural Networks. Our results showed that the projected technique will effectively segment white blood corpuscle cells with their nucleus. The outcomes of the model are very promising as it is proven to be able to classify better than other existing methods that are used. When we compare our model to other models, our model has a considerable gain in terms of accuracy. The model achieved an accuracy of 97 %.

Our method has a potential to be implemented to a large range of applications due to its adaptability and simplicity.

REFERENCES

[1] A.Kong , H.Zhu, “Predicting Trend Of High Frequency CSI 300 Index Using Adaptive Input Selection and Machine Learning Techniques,” Journal of System Science And Information ,Vol. 6, pp. 120_133, 2018.

[2] A. K. Jain, J. Mao, K. M. Mohiuddin, “Artificial Neural Networks: A Tutorial,” Computer, Vol 29, pp. 31_44, 2015

[3] S. m. Erfani, S. Rajasegarar, S. Karunasekera, et al, “High-dimensional and Large-Scale anomaly detection using a linear one-class SVM with Deep learning,” Pattern Recognition, pp.121-134,2016.

[4] Idrees, Sheikh Mohammad, M. Afshar Alam, and Parul Agarwal. "A Prediction Approach for Stock Market Volatility Based on Time Series Data." IEEE Access 7 (2019): 17287-17298.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[5] Picasso, Andrea, et al. "Technical Analysis and Sentiment Embeddings for Market Trend Prediction." Expert Systems with Applications (2019).

[6] Somarajan, Siddarth, et al. "Modelling and Analysis of Volatility in Time Series Data." Soft Computing and Signal Processing. Springer,Singapore,2019.609-618.

[7] Nyoni, Thabani. "Time series modeling and forecasting of the consumer price index in Belgium." (2019).

[8] N.Srinivasan, C.Lakshmi ” Stock Prediction and Analysis Using Intermittent training data with artificial neural networks”, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

[9] Mr. Pankaj K. Bharne, Dr. Sameer S. Prabhune “Survey on Combined Swarm Intelligence and ANN for Optimized Daily Stock Market Price”, 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp), Dec. 2017.

[10] Van Horne, J. C., & Parker, G. G. (2017). The random-walk theory: an empirical test. Financial Analysts Journal, 23(6), 87-92.

[11] Rasel, R. I., Sultana, N., & Hasan, N. (2016). Financial instability analysis using ANN and feature selection technique:Application to stock market price.

[12] aChung, H., & Shin, K. S. (2018). Genetic algorithm-optimized long short-term memory network for stock market prediction. Sustainability, 10(10), 3765.