MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL (M.P.)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MINOR PROJECT

ON

STOCK TREND PREDICTION USING MACHINE LEARNING TECHNIQUES

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF TECHNOLOGY

SUBMITTED BY:

KRISHNA CHANDAN AYYAGARI(141112003)

SNEHIT CHAKKA(141112100)

GOUTHAM KALLEMPUDI(141112050)

SAI KALYAN GOGINENI(141112080)

UNDER THE GUIDANCE

OF:

DR. NAMITA TIWARI

SESSION 2016-2017

MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL(M.P)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **Snehit Chakka**, **Goutham kallempudi**, **Sai Kalyan Gogineni** and **Krishna Chandan Ayyagari** students of B.Tech 3rd Year (Computer Science & Engineering), have successfully completed their project "**STOCK TREND PREDICTION USING MACHINE LEARNING TECHNIQUES**" in partial fulfillment of their minor project in Computer Science & Engineering.

DR.NAMITA TIWARI

(Project Guide)

PROF.SANYAM SHUKLA

(Project Coordinator)

MAULANA AZAD NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL(M.P)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, hereby, declare that the following report which is being presented in the Minor Project Documentation entitled “STOCK TREND PREDICTION USING MACHINE LEARNING TECHNIQUES” is the partial fulfillment of the requirements of the third year (sixth semester) Minor Project in the field of Computer Science And Engineering. It is an authentic documentation of our own original work carried out under the able guidance of Dr. Namita Tiwari and the dedicated co-ordination of Prof.Sanyam Shukla. The work has been carried out entirely at Maulana Azad National Institute of Technology, Bhopal. The following project and its report, in part or whole, has not been presented or submitted by us for any purpose in any other institute or organization.

We, hereby, declare that the facts mentioned above are true to the best of our knowledge. In case of any unlikely discrepancy that may possibly occur, we will be the ones to take responsibility

SNEHIT CHAKKA (141112100)

GOUTHAM KALLEMPUDI (141112050)

SAI KALYAN GOGINENI (141112080)

KRISHNA CHANDAN AYYAGARI (141112003)

ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide Dr. Namita Tiwari, for her valuable help and guidance. We are thankful for the encouragement that she has given us in completing this project successfully. Her rigorous evaluation and constructive criticism were of great assistance.

It is imperative for us to mention the fact that this minor project could not have been accomplished without the periodic suggestions and advice of our project co-ordinator Prof. Sanyam Shukla.

We are also grateful to our respected director Dr. N. S Choudhary for permitting us to utilize all the necessary facilities of the college.

Needless to mention is the additional help and support extended by our respected HOD, Dr. R. K. Pateriya, in allowing us to use the departmental laboratories and other services.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind co-operation and help.
Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing the much-needed support and encouragement.

CONTENTS

ABSTRACT

This project focuses on predicting the stock price trends of a company in the near future using machine learning techniques. The main objective of this project is to predict the stock trend of companies of Newyork stock exchange (NYSE) for the required forthcoming days using various machine learning techniques. Unlike some other approaches which are concerned with company fundamental analysis (e.g. Financial reports, market performance, sentimental analysis etc.), the feature set is derived from the time series of the stock itself and is concerned with the potential movement of the past price. The machine learning model takes different banking attributes as inputs such as on balance volume, rsindex, ROCR, Williams Accumulation etc., and predicts the stock's trend as positive or negative. Here, a subset of stock technical indicators is critical parameters for predicting the stock trend. It explores different ways of validation and shows that overfitting tend to occur due to fundamentally noisy nature of a single stock price. The machine learning techniques include logistic regression and kernel based Support Vector Machine (SVM) and the different kernels used in the SVM being Linear kernel and Gaussian kernel (RBF). Experimental results suggest that we are able to achieve more than 70% accuracy on predicting a 3-10 day average price trend with RBF kernelled SVM algorithm. These techniques are compared and resulting in SVM being the most suited technique.

INTRODUCTION

Stock market investment is sought to be one of the most popular and fetching methodologies for the semi-short term and long term capital investments in today's day and age. The investment in Stock markets is a practice however subjected to various risks for its unpredictability due to a large number of influencing factors; most of which are fairly predictable today due to the development of various new age computational techniques using complex algorithms and extensive data analysis.Today, there are an increasing number of researchers that are being initiated each day to maximize the accuracy of these predictions by developing various algorithms with newer sets of attributes taking into account a variety of other factors to help the buyers, stock brokers, and the companies to formulate a reliable opinion of the futuristic performance of the companies.The various computable methodologies used today are often satisfiable accurate to considerably reduce the risks in this investment.

Machine learning is a branch of computer science that relies on the implementation of artificial intelligence that enables the computers to learn and perform functionalities without explicitly programming them. It is broadly classified into 3parts; Supervised learning, Unsupervised learning, and Reinforcement.

SUPERVISED LEARNING:

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples., In supervised learning, each example is a pair consisting of an input object (typically a vector) and the desired output value (also called the supervisory signal).

In supervised learning, both the input and the desired output is provided and this data is labeled for classification to provide a learning basis for future data processing.

UNSUPERVISED LEARNING:

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.In unsupervised learning, no datasets are provided, instead, the data is clustered into different classes.

REINFORCING:

Reinforcement Learning is another type of *Machine Learning* where it allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal.

In this project, we make use of the Supervised Learning by implementing Logistic Regression and SVM.

Supervised learning problems are again categorized into Regression and Classification.

Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
Example: Logistic Regression, SVM.

Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Example: Linear Regression.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

RELATED WORK

LOGISTIC REGRESSION:

logistic regression is a type of probabilistic statistical classification model. Logistic regression is a classification machine learning problem where the output of the machine is a discrete value. If the output is more than two discrete values it is a multivariate logistic regression.Logistic Regression uses Gradient Descent algorithm for convergence.

Regression fits a good line but sometimes a curve too for the given labeled examples. The general equation of the line is **y=a*x +c** where m=slope and c=constant, and in this equation, there is only one value of x (feature), but if there are more features we define the term **a*x+c** as **a*x 1+b*x2+c** or **($\boldsymbol{theta^T}$ * x)** where x is a vector of all features and theta is a vector containing constants and initially the values of theta are zeros. Unlike linear regression where the output values are continuos and h(x) the output of the machine it is, a whole number. Logistic regression also uses the same cost function, but h(x) should not be continous rather it must have a value between -1 and 1. To make h(x) value between – 1 and 1 apply sigmoid function to ($\boldsymbol{theta^T}$ $*$ **x**) .

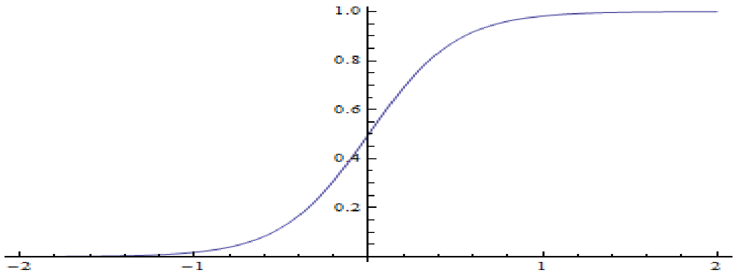
Sigmoid Function:

A sigmoid function is a bounded differentiable real function that is defined for all real input values and has a positive derivative at each point. The logistic function has this further, important property, that its derivative can be expressed by the function itself (i.e if s(z) is sigmoid function then its derivative **s'(z) = s(z) * [s(z) – 1]**). Sigmoid function is defined as

s(z) =1/(1 + e^{-z}). Here **z = $theta^T$ * X** where X is feature set

$$h(x) = 1/(1+e^{-(theta^T * x)})$$

Sigmoid Function



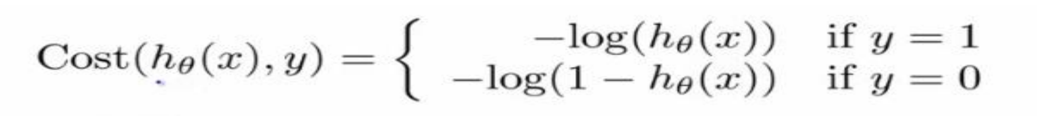
GRADIENT DESCENT:

To make the value of h(x) nearly equal to y, we have to find the optimal value for theta. For finding optimal theta the cost function **J(theta) =** $\sum[ylog(h(x))$ – $(1-y)\log(1-h(x))]^2$ should continuously differentiated until convergence. α is learning rate, Lower the value of α more is the guarantee that the global minima value for theta is obtained generally the value of α is in the range **[0.001 , 0.1]** . Once theta is obtained the value

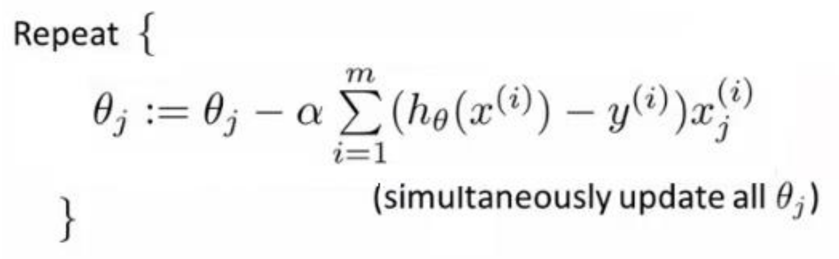
($theta^T$ * x) is calculated , if the value is greater than zero then the given example is positive example or else example is negative.

($\boldsymbol{theta^T}$ * x) >= 0 => Positive Example

($\boldsymbol{theta^T}$ * x) <0 => Negative Example



$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

SUPPORT VECTOR MACHINE (SVM):

In machine learning, **support vector machines** are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called **support vector clustering** and is often used in industrial applications either when data are not labeled or when only some data are labeled as a pre-processing for a classification pass.

IMPLEMENTATION OF SUPPORT VECTOR MACHINE:

SVM is basically used for classification purpose. It draws a decision boundary between positive and negative examples. Unlike logistic regression, here an example is considered as positive if **Z>1** and negative if **Z<-1.** So there are gutters between +1 and -1. SVM tries to draw a decision boundary that maximizes the width of the road that separates positive and negative examples. The main idea is to find the widest separating street.

We can write the constraint:

$$\boldsymbol{y_i(w.x_i + b)} \mathbf{>=1}$$

y_i =+1 for positive examples and y_i =-1 for negative examples.

We classify unknown 'u' as follows:

$$\textbf{f(u)=}\ \boldsymbol{w.u+b}$$

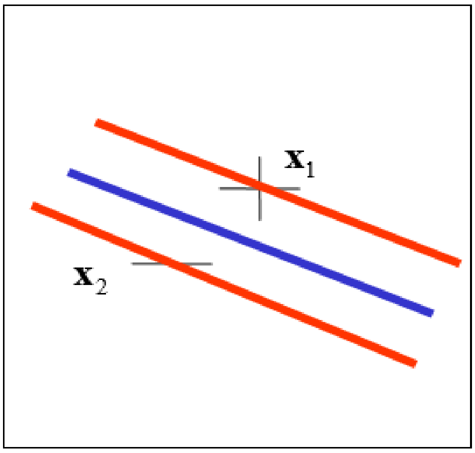
For all positive examples:

$$\textbf{f(x}_{+}\textbf{)=}\ \boldsymbol{y_i(w.x_{+}+b)}\ \textbf{>=1}$$

For all negative examples:

$$\textbf{f(x}_{-}\textbf{)=}\ \boldsymbol{y_i(w.x_{-}+b)}\ \textbf{<-1}$$

Distance between gutters is shown in figure.1:



$\boldsymbol{w.x_1+b}$ **=1** ------equation (1)

$\boldsymbol{w.x_2+b}$ **=-1** --------equation (2)

Subtracting equation (1) from equation (2)

$\boldsymbol{w(x_1-x_2)}$**=2**

Dividing the above equation with unit vector

$\frac{w}{||w||}.\ (x_1-x_2) = \frac{2}{||w||}$

Now to maximize the distance the factor ||w|| should be minimized.

For mathematical convenience, it can be transformed as

$$\frac{1}{2}||\boldsymbol{w}||^2$$

Minimizing the above equation can give the maximum distance between the gutters.

One approach to finding the minimum value is by using Lagrange's method. The resulting Lagrange multiplier that to optimize is:

$$L = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_i \boldsymbol{\alpha}_i(\boldsymbol{y}_i\,(\boldsymbol{w}.\boldsymbol{x}_i + \boldsymbol{b}) - \mathbf{1})$$

By partial derivatives, we can get the values of alpha, w, b etc.

1. The partial derivative of L with respect to b:

   $$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}} = \mathbf{0}$$

   We obtain $\sum_i \boldsymbol{\alpha}_i \boldsymbol{y}_i = \mathbf{0}$

2. The partial derivative of L with respect to w:

   $$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w}} = \mathbf{0}$$

   We obtain $\boldsymbol{w} = \sum_i \boldsymbol{\alpha}_i \boldsymbol{y}_i \boldsymbol{x}_i$

The decision boundary is drawn by the equation:

$$h(x) = \sum_i \boldsymbol{\alpha}_i \boldsymbol{y}_i \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + \boldsymbol{b} \geq 0$$

To classify an unknown *x* we should compute a kernel function $K(x_i, x)$ against each vector x_i

For positive gutter :

h(x)= $\sum_i \boldsymbol{\alpha_i y_i K(x_i, x)} + \boldsymbol{b}$ **=1**

For negative gutter:

h(x)= $\sum_i \boldsymbol{\alpha_i y_i K(x_i, x)} + \boldsymbol{b}$ **=-1**

SUPPORT VECTOR MACHINE KERNELS:

There are different types of kernels. Some of them are as follows

1.Linear Kernel

2.Polynomial Kernel

3.Radial Basis Function(RBF) or Gaussian Kernel.

In this project, linear and Gaussian kernels are used.

Linear Kernel:

$$\boldsymbol{K(\vec{u}, \vec{v}) = \vec{u}.\vec{v}}$$

If the examples are not separable they are transformed into another space.

$$\boldsymbol{K(\vec{u}, \vec{v})} = \emptyset(\boldsymbol{\vec{u}}).\emptyset(\boldsymbol{\vec{v}})$$

$\emptyset(\vec{u})$ that transforms input vectors into another dimension where they can be linearly separable.

Gaussian Kernel:

$$\boldsymbol{K}(\vec{\boldsymbol{u}},\vec{\boldsymbol{v}}) = \boldsymbol{exp}(-\frac{\|\vec{\boldsymbol{u}}-\vec{\boldsymbol{v}}\|^2}{2\boldsymbol{\sigma}^2})$$

Here σ^2 is the standard deviation of input vectors. It defines the steepness of the rise around the landmark. In 2D decision boundaries will be circles around positive and negative examples. As the value of σ^2 increases the feature falls to zero rapidly.

PROPOSED WORK

- TAKE STOCK DATA SET
- APPLY TECHNICAL FUNCTIONS TO THE DATASET TO OBTAIN FEATURE SET
- NORMALIZE THE FEATURE SET
- MAKE LABELING TO THE FEATURE SET
- APPLY LOGISTIC REGRESSION ALGORITHM
- CALCULATE THE ACCURACY OF DATA USING LOGISTIC REGRESSION
- APPLY SUPPORT VECTOR MACHINE
    - SVM WITH LINEAR KERNEL
    - SVM WITH GAUSSIAN KERNEL
- CALCULATE THE ACCURACY OF DATA USING SVM
- COMPARE THE ACCURACY

SOFTWARE AND HARDWARE REQUIREMENTS

SOFTWARE:

The following soft-wares were used for this project:

- Operating System – Microsoft® Windows® 8, Ubuntu or above.
- Octave Software – To run octave codes of stock trend prediction using machine learning techniques.
- Microsoft Excel- To take dataset as input and for preprocessing of the dataset.

HARDWARE:

The following hardware configuration is required to run the various soft-wares for this project:

- Processor: Intel® Core™ i3 CPU
- Memory: 4GB RAM
- Storage required: Maximum of 100MB
- Graphics card: not needed

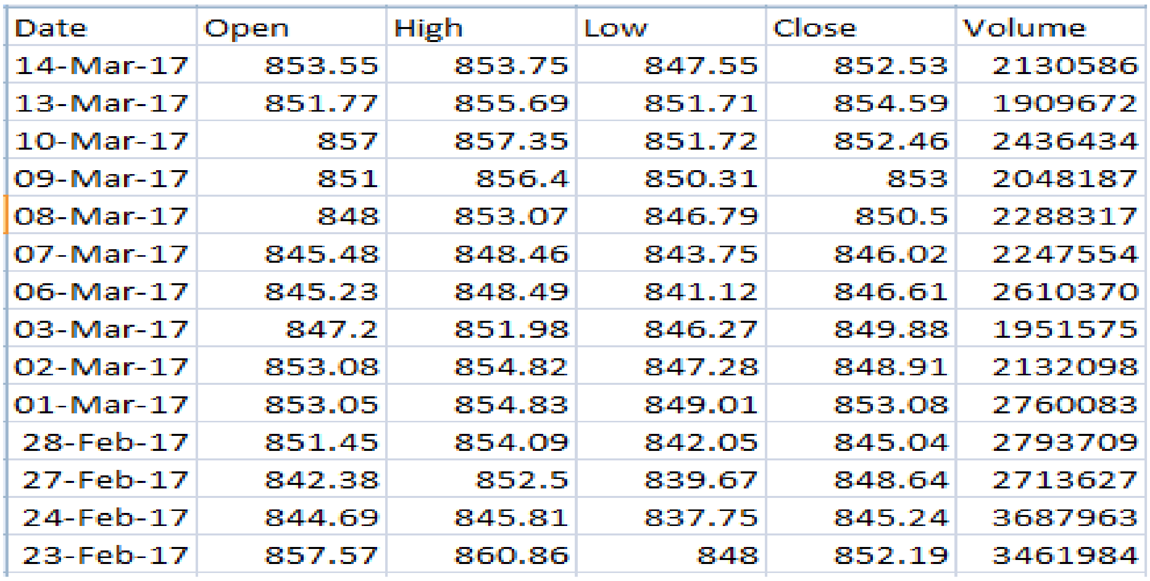
IMPLEMENTATION

DATASET EXTRACTION:

To perform certain operations and programs we need data. Data means information, it is of any form. A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set. Data sets that are so large that traditional data processing applications are inadequate to deal with them are known as big data.

Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis.

For this project, we took a traditional dataset of Amazon to perform a machine learning algorithm to predict the stock trend of that company.Every data set has some attributes as columns and tuples as rows. In this data set, there are five attributes. they are open value, closing value, high value, low value, date and volume which means a number of products. There are 1761 tuples which are the daily trade values of the company for five years. For some idea of the data set, we are showing the screenshot of small part of the dataset.



Date	Open	High	Low	Close	Volume
14-Mar-17	853.55	853.75	847.55	852.53	2130586
13-Mar-17	851.77	855.69	851.71	854.59	1909672
10-Mar-17	857	857.35	851.72	852.46	2436434
09-Mar-17	851	856.4	850.31	853	2048187
08-Mar-17	848	853.07	846.79	850.5	2288317
07-Mar-17	845.48	848.46	843.75	846.02	2247554
06-Mar-17	845.23	848.49	841.12	846.61	2610370
03-Mar-17	847.2	851.98	846.27	849.88	1951575
02-Mar-17	853.08	854.82	847.28	848.91	2132098
01-Mar-17	853.05	854.83	849.01	853.08	2760083
28-Feb-17	851.45	854.09	842.05	845.04	2793709
27-Feb-17	842.38	852.5	839.67	848.64	2713627
24-Feb-17	844.69	845.81	837.75	845.24	3687963
23-Feb-17	857.57	860.86	848	852.19	3461984

DATA CLEANSING:

Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The data set we took for our project also contains an attribute date, which is irrelevant.we removed it by data cleansing.

FEATURE SET:

To predict some value or to identify a certain trend we need some attributes on which that value will depend.These values are known as features and set of certain features is known as "Feature Set".These are the technical indicators which will decide the trend on basis of the values present in our data set.A stock technical indicator is a series of data points that are derived by applying a function to the price data at time t and study period n. Below is a table of indicators that I compute from time series and transform to features: Efficiency of the prediction depends on the size of the feature set.As the size of the feature set increases,the number of technical indicators will increase and it results in more efficient prediction.

For the feature set to predict the stock trend of a company, we have to carefully select the technical indicators that will be calculated on basis of function of time by using the values in the dataset. We chose 34 technical indicators as our feature set that will affect the stock trend of a company.

Feature_set=[

‘rsindex’,’volroc’,’negvolidx’,’posvolidx’,’adline’,’hhigh’,’llow’,
’medprice’,’onbalvol’,’prcroc’,‘pvtrend’,’typprice’,’wclose’,
’willad’,’adosc’,’chaikosc’,’stochosc’,’tsaccel’,’tsmom’,
’chaikvolat’,’willpctr’]

Some other technical indicators are below in detail

Description	Indicators	Name	Formulae
It reflects the level of the close relative to the highest high for the look-back period.	WILLR	Williams %R	(highest-lowest)/(highest-closed)*100
Measures the change in price of n periods.	MOM	Momentum	Price(t)-price(t-n)
It is used to identify overbought or oversold conditions of a stock.	RSI	Relative Strength Index	RSI = 100 - 100 / (1 + RS) Where RS = Average gain of up periods during the specified time frame / Average loss of down periods during the specified time frame Default time frame=14 days
Compute rate of change relative to previous trading intervals.	ROCR	Rate of Change	Price(t)/price(t-n)*100

NORMALIZATION:

It performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute A. Min-Max transformation maps a value v of A to v^l in the range **[new_min_A,new_max_A]** by computing

$$\mathbf{v^l} = \frac{v - min_A}{max_A - min_A} \mathbf{(new\_max_A – new\_min_A) + new\_min_A}$$

DATA LABELLING:

This is the column of values we have to predict from the training set and from the feature set. We have to calculate or predict the stock price trend at the interval of 3 days as “label3",5 days as "label5" and 10 days as "label10".This means that if we compare the changes in closing prices of last y days and we predict the trend then it is "label y".

Procedure to calculate label:

First we have to take the training set and calculate the difference between the closing prices of this day and next y days and if it is a positive value then we will assign "+1" to label y column ,else we have to assign "0" to label y coloumn.we have to establish a relationship between feature space and this label y column through some machine learning algorithms and by using this relationship in testing set we will predict the trend by values in "label y" column. If it is a positive value it will increase or else it will decrease.

The function **Y (t) = close (t)-close (t-x)** gives the label value for a given feature record

For 10 days: **Y (t) = close (t) – close (t-10)**

For 3 days : **Y (t) = close (t) – close (t-3)**

From the above function stock is negative (0) if Y (t) is less than zero or else positive (1)

Y (t) > =0 -> label =1

Y (t) < 0 -> label =0

Now apply Logistic regression on dataset and compute accuracy.

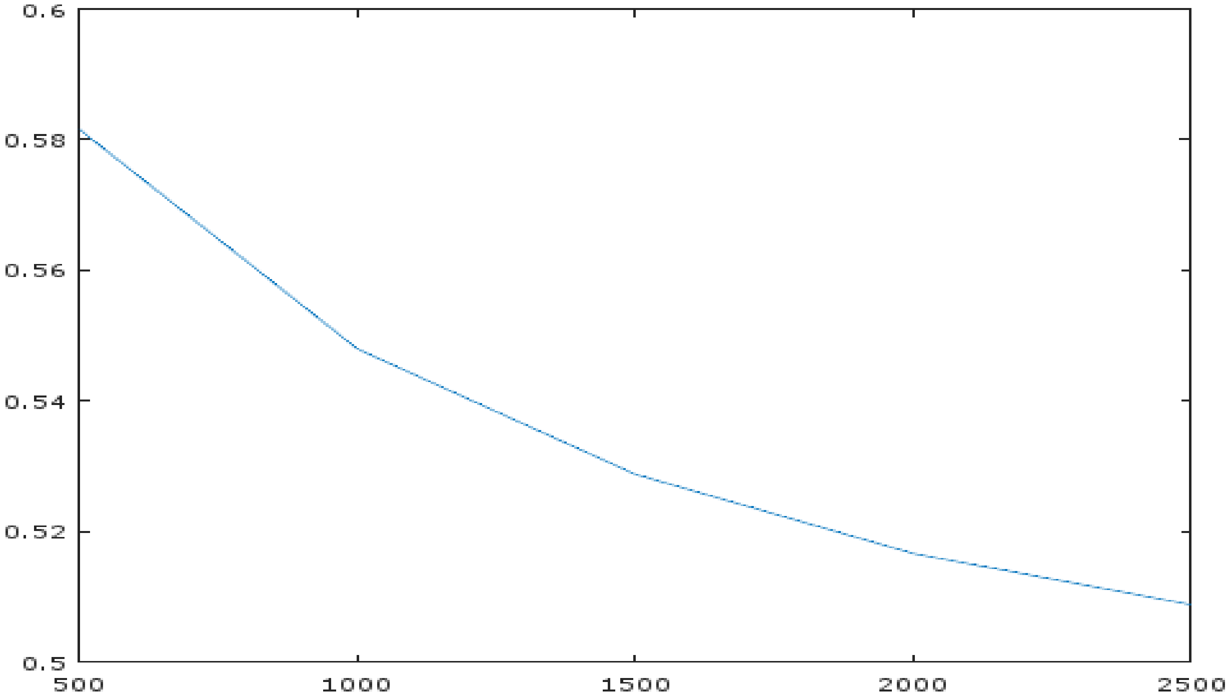
Similarly,applying SVM on dataset and compute accuracies of different kernels.

RESULT AND ANALYSIS

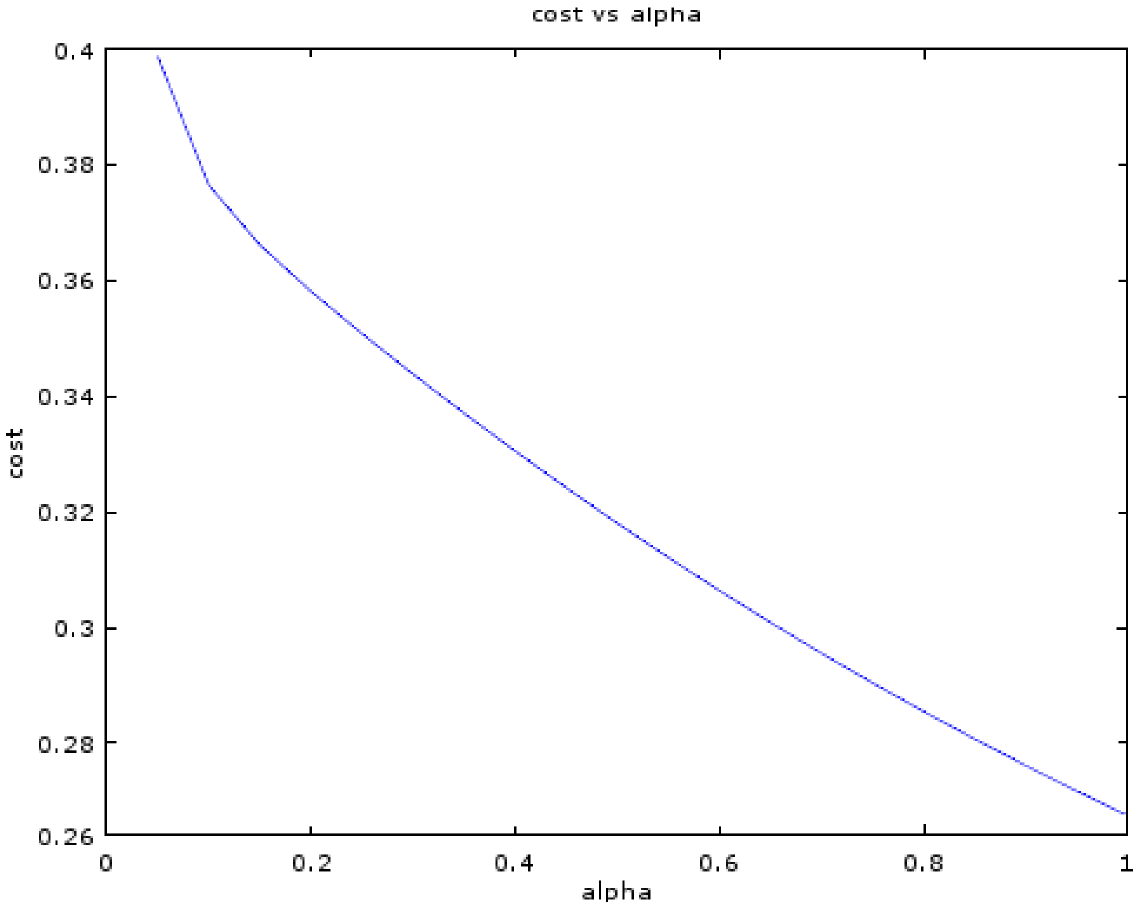
EFFICIENCY:

In logistic Regression, the parameter alpha is known as learning rate.Learning rate is the main thing to get a good efficient algorithm, generally, the value of alpha ranges from **0.001 to 10**. Besides learning rate, the number of iterations in gradient descent also affects the algorithm, So how to choose a good value of alpha and iterations? The possible way for this is with an increase in the value of alpha and learning rate the cost obtained by alpha and iterations should decrease this ensures that the algorithm is performing better but how much? Gradient descent always chooses a value of alpha and iterations for which the cost is minimum. Hence while iterating for all the values of alpha and iterations there will be one combination for which the cost is minimum. Choose those values of alpha and iterations.Below is the graph between cost and iterations (a) and graph between cost and alpha values (b).

(a)



(b)



ACCURACY:

Efficiency in general terms is defined as the number of correctly predicted examples from the total examples. There can be four possibilities for any machine learning problem Correctly Identified **(True positive),** Correctly Rejected **(True Negative),** Incorrectly Identified **(False positive),** Incorrectly Rejected **(False Negative**).

$$\textbf{Accuracy =} \quad \frac{(\textbf{True positive + True negative})}{\textbf{True positive+Truenegative + False positive + False negative}}$$

PRECISION AND RECALL:

Two new metrics precision and recall both give a value between 0 and 1

Precision

How often does our algorithm false alarm?

Precision = true positives / (true positive + false positive)

High precision is good

Recall

How sensitive is our algorithm?

Recall = true positive / (true positive + false negative)

High recall is good

By computing, precision and recall get a better sense of how an algorithm work

The accuracies of logistic regression and SVM of Amazon stock are shown in Table I.

Days	Logistic Regression	SVM(Linear Kernel)	SVM(Gaussian Kernel)
Next 3-day	73.97%	80.23%	77.29%
Next 5-day	79.84%	83.95%	84.54%
Next 10-day	79.45%	86.10%	83.75%

TABLE I

The accuracies of logistic regression and SVM of Microsoft stock are shown in Table II.

Days	Logistic Regression	SVM(Linear Kernel)	SVM(Gaussian Kernel)
Next 3-day	71.56%	79.78%	74.36%
Next 5-day	74.34%	82.12%	82.89%
Next 10-day	78.46%	84.14%	81.75%

TABLE II

The result shows that for Amazon the prediction is robust and above 80%. For Amazon, it suggests that longer-period prediction is significantly better than the short-term. For Microsoft, the split is between 5-day, shorter than that it tends to have very high noise while 7-day and 10-day prediction is extremely good.Microsoft stock has more generalized indicators weighted heavier than others so that it has low variance. Amazon’s top features are all 1-3 days volatility so that the model tend to overfit. Actually, the experiment shows the testing accuracy for Amazon is 80% which is significantly higher than Microsoft stock.

Conclusion And Future Work

The result is very helpful in real-world investment for people having no idea on investment. By learning from past data we are able to get above 70% accurate prediction on the next couple day’s trend. For future work, it worth adding sentiment data as features to augment the technical features. The challenge is how to eliminate as much as noise in sentiment data and quantify them.The future work also includes dynamically extracting stock data from NYSE and predicting for stock trend.

Applications:

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit

Many trading applications and websites rely on stock market prediction model to predict the stock price trend of various companies. This helps the traders to invest their money in buying the shares of a company with good stock price value.

A good prediction model with more efficiency will give good results and decides the flow of millions of money

REFERENCES:

- *Machine learning tutorials by Andrew NG* *https://www.coursera.org/learn/machine-learning*
- *MACHINE LEARNING AN ALGORITHMIC PERSPECTIVE by Stephen Marsland.*
- *OCTAVE implementation tool is used for implementing algorithms.* *https://www.gnu.org/software/octave/doc/interpreter/*
- *Data sets are available in Google finance.com* *https://www.google.com/finance/historical?q=NASDAQ%3AAMZN&ei=bhTuWKiwMcveuQTHpoaADw*
- *An SVM-based Approach for Stock Market Trend Prediction* *http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:introduction_to_technical_indicators_and_oscillators*
- *http://cs229.stanford.edu/proj2014/Xinjie%20Di,%20Stock%20Trend%20Prediction%20with%20Technical%20Indicators%20using%20SVM.pdf*