STOCK PRICE PREDICTION

RISHABH GUPTA(2014086)
KUSHAGRA MAHAJAN(2014055)

INTRODUCTION

In this project, machine learning algorithms were used to predict the closing price of stocks of companies across various sectors (IT, Pharmaceutical ,Banking Industry) and find out if same model gives best performance across stocks in different sectors.

Stock price prediction has always been one of the most challenging problems affecting the lives of millions of people around the world. Recently the problem has gained great popularity due to its unpredictable behaviour. This work is our attempt to explore the performance of some of the common techniques on stock datasets.

DATASETS

We explore three datasets taken from Quandl.com :- **HDFC Bank**, **TCS**, **CIPLA Pharmaceuticals Pvt. Ltd**.

The datasets provided stock values for the past 8 years(3000 samples).

- 80%(2400 samples) of the data was chosen for training.
- 10% of the training data(240 samples) were used for validation.
- 20% of the total data was used for testing.

FEATURE EXTRACTION

The original data had the following features: Opening, Highest, Lowest, Last, Closing prices, Total trade Quantity and Turnover.
The features used for Learning Model are given in Table-3 .

EVALUATION METRICS

Mean Squared Error was used as an Evaluation Metrics for our Models.

PRE-PROCESSING

Technique	Applied	Remarks
Normalization	YES	The Dataset was scaled between 0 and 1.
PCA	NO	The number of features were not large.
Missing Values Correction	YES	Taking average of surrounding values

TECHNIQUES APPLIED

We have explored

- Simple and Regularized Linear Regression(LASSO and RIDGE).
  - Stochastic Gradient Descent
  - Batch Gradient Descent
- Polynomial Kernel Regression
- Gaussian Kernel Regression
- Support Vector Regression
  - Linear kernel
  - Polynomial kernel
  - Gaussian kernel

METHODOLOGY

- We have used **Grid Search** along with **Cross Validation** Technique in each model(wherever possible) to hyper-tune the parameters and come up with the ones which gives minimum MSE.
- Using the best parameters various techniques were applied and MSE was computed.
- Nifty feature was removed and all techniques were again applied against best parameters and the results were compared.

CHALLENGES

Finding out the appropriate features was a big challenge. Lot of papers were consulted and research was done to come up with relevant features in stock market prediction. Some of the relevant works explored were:

- Feature Investigation for Stock market Prediction by Hui Lin[1].
- FEATURE SELECTION FOR PRICE CHANGE PREDICTION by Zehra Çataltepe1 , Savaş Özer2 , Vahide Unutmaz Barın2[2].

OBSERVATIONS

Note: All the values are normalized between 0 and 1.So MSE values are according to normalization.

- **Linear and Regularized Regression**

After hyper-tuning the parameters (alpha and delta) Linear ,LASSO and Ridge Regression was applied along with Stochastic and Batch Gradient Descent algorithm. The results corresponding to best parameters are as follows:

Table-1 : MSE corresponding to best alpha and delta for Regularized Linear Regression.

Stock	Best Model	Best Parameters	MSE
CIPLA	Simple Linear Regression	Alpha=0.05 Delta=0	0.001425
TCS	LASSO Linear Regression	Alpha=0.1 Delta=0.1	0.001067
HDFC BANK	LASSO Linear Regression	Alpha=0.00005 Delta=0.01	0.002279

- **Polynomial Kernel Regression**

Extracting the polynomial features and then applying Linear Regression across different degree gives the following results against each stock:

Table-2 : MSE corresponding to best degree in Polynomial Kernel Regression.

Stock	Best Degree	MSE
CIPLA	2	0.00006993
TCS	3	0.000884
HDFC	3	0.000149

- **Gaussian Kernel Regression**

Applying Gaussian Kernel Regression against best value of sigma across each stock gives the following results.

Table-3 : MSE corresponding to best sigma in Gaussian Kernel Regression.

Stock	Best Sigma	MSE
CIPLA	0.0002	0.0210
TCS	0.2599*	0.000127
HDFC	0.02	0.000553

* Standard deviation of Closing Prices of training set

- **Support Vector Regression**

Applying Support Vector Regression using linear, polynomial and Gaussian Kernel across each stock gives the following results.

Table-4 : MSE corresponding to best alpha and sigma in Support Vector Regression.

Stock	Kernel	MSE
CIPLA	Linear	0.00036
	Polynomial	0.015
	Gaussian	0.00246
TCS	Linear	0.000155
	Polynomial	0.00448
	Gaussian	0.0003196
HDFC	Linear	0.00000355
	Polynomial	0.010
	Gaussian	0.000144

To explore and prove the importance of good feature selection we eliminated an important feature(NIFTY closing price) and recalculated the MSE.

The Best Accuracy is compared in both cases across all Models and stocks.

Table-5 : Comparison of MSE when NIFTY feature was removed.

Stock	Technique	MSE (With Nifty)	MSE (Without Nifty)
CIPLA	**Regularized**	0.002279	0.00248
	Polynomial	**0.00006993**	0.0000703
	Gaussian	0.00045	0.020
	Support Vector	0.000359	0.000341
TCS	**Regularized**	0.001067	0.001198
	Polynomial	0.000884	0.001966
	Gaussian	**0.000127**	0.0003548
	Support Vector	0.000155	0.000155
HDFC	**Regularized**	0.001425	0.001465
	Polynomial	0.000149	0.0009838
	Gaussian	0.00045	0.0005535
	Support Vector	**0.00000355**	0.00000379

RESULTS

The Best Model corresponding to each stock is as follows:

Table-6 : Best Model of each Stock.

STOCK	BEST MODEL	MSE
CIPLA	Polynomial Kernel Regression	0.00006993
TCS	Gaussian Kernel Regression	0.000127
HDFC BANK	Support Vector Regression (Linear Kernel)	0.00000355

- We observe that the same model does not give best results across different stocks. Hence stocks across different sectors (IT, Banking, Pharmaceutical) move differently with features and historical data.
- The complexity of the best models of different stocks(Polynomial, Gaussian, Support Vector Regression) shows that the closing price of stock are not merely linearly related to the features but are rather more complex.
- The closing Price of HDFC BANK is most predictable among the 3 stocks as it gives the least MSE.
- The MSE values were found to be mostly higher when NIFTY feature was removed. This shows that the choice of good features is an important determinant in good performance from the machine learning models.

REFERENCES

- https://www.quandl.com/data/NSE?keyword=
- http://web.itu.edu.tr/~cataltepe/pdf/2011_ICMFECataltepe.pdf
- http://www.ntu.edu.sg/home/elpwang/PDF_web/13_ICONIP.pdf
- http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Source Code

- https://github.com/git-rishabh/Stock-Price-Prediction

Additional Tables and Graph

Table-7 : Features used in our Learning Model.

Previous Day Closing Price
Same Day Opening Price
Previous Day Nifty Closing Price
10 Days Moving averages
15 Days Moving averages
20 Days Moving averages
40 Days Moving averages
10 Days Momentum
40 Days Momentum
Average Difference between Opening Price and Closing Price (5 Days)
Average Difference between Opening Price and Closing Price (10 Days)
Average Difference between Opening Price and Closing Price (40 Days)
Average Difference between Highest Price and Lowest Price (5 Days)
Average Difference between Highest Price and Lowest Price (10 Days)
Average Difference between Highest Price and Lowest Price (40 Days)
Volatility
Average Turnover for 10 Days

Table-8 : Best Value of Alpha and Delta by applying cross validation and grid search

Stock		**Alpha**	**Delta**	**MSE**
CIPLA	**LASSO**	0.05	0	0.033688
	RIDGE	0.05	0	0.3688
TCS	**LASSO**	0.1	0.1	0.00139
	RIDGE	0.10	0	0.00144
HDFC	**LASSO**	0.00005	0.01	0.00146
	RIDGE	0.00001	0	0.00188

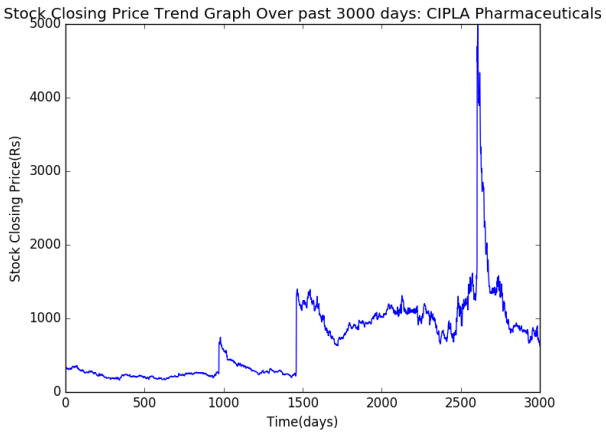
Table-9 : MSE corresponding to best alpha and delta for Regularized Linear Regression.

Stock	**Parameters**	**Gradient Descent**	**MSE**
CIPLA	Alpha=0.05	Stochastic	0.002474
	Delta=0	Batch	**0.002279**
TCS	Alpha=0.1	Stochastic	**0.001067**
	Delta=0.1	Batch	0.001116
HDFC	Alpha=0.00005	Stochastic	**0.001425**
	Delta=0.01	Batch	0.001754

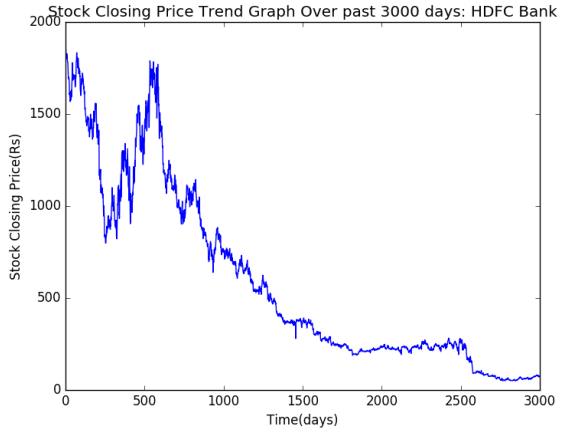
Plots of Closing Price for 3000 days.

Note: These plots does not tell anything about the best model in each stock and any model should not be evaluated against these plots.

CIPLA DATASET



HDFC DATASET



TCS DATASET