2nd National Conference on Engineering Applications of Emerging Technology
in association with International Journal of Scientific Research in Science,
Engineering and Technology | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
[https://ijsrset.com | doi : https://doi.org/10.32628/IJSRSET]

Survey of Regression-Driven Stock- Market-Price-Predictors

Jeevan UM[1], N. Dhanvina[1], N. Dharshan[1], Sushmitha S[2]
[1]Student, [2]Assistant Professor
Department of Computer Science and Engineering Global Academy of Technology, Bangalore, Karnataka, India

ABSTRACT

The alluring profits in the Stock market in the investment loci have invoked a ubiquitous interest, besides as a recreational hobby by many also as a significant career both in consultancy and investment. The Stock market price prediction is a sophisticated task intrinsically involving the company's earnings, the market competition, demand, and stability apart from the extrinsic parameters- exchange rate, political stability, the government policy decisions. Thus, multiple algorithms have been written to predict the non -linear and fluctuating Stock price sensitized by the market emotions. Different algorithms have different principles and varying degrees of accuracy; It is attempted to survey their accuracy based on a Machine -Learning-Technology-Based-System to analyze the share statistics.

We use Regression, a widespread family of algorithms that despite its modest ideology produces good results on large datasets, is quite intuitive, and easily represented mathematically. The selection of the dataset has no small role and with that in mind we use the data set collected from Kaggle, recording the parameters open, close, high, low, date, time, volume, and the bid prices for the stock at separate instants.

Keywords—regression, stock, machine learning, survey

I. INTRODUCTION

Machine learning algorithms are broadly classified primarily into Supervised and Unsupervised ed Algorithms. An Algorithm is said to be Supervised when the right solution exists; the function of the input X against the output the training and test sets are used approximate the function to obtain the value for any new value of x. Regression and Classification are the major types of problems in Supervised algorithms. When the result is obtained for the new value of x, the obtained result y is compared to the values already available and is corrected. The two tasks Regression and Classification involve obtaining a numerical value in the former and a categorical attribute in the latter.

Regression – A regression problem is used when the output variable is a real or continuous like the value of money, etc. It has various types of models and involves fitting the data in the best hyper-plane. When there is single variant, it is simple and multiple when the behaviour of multiple variables is studied. Python, a language alluring my data analysts and scientists and Scikit_learn, widely used module for Machine learning has been used. The following variants of Regression are studied-

1. Linear Regression
2. Robust Regression
3. Ridge Regression
4. Lasso Regression

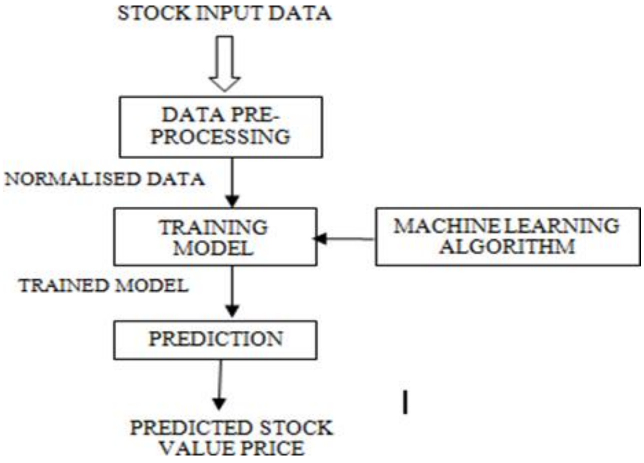

Fig1: System Architecture

II. MATHEMATICAL ANALYSIS AND STUDY OF ALGORITHMS

A. Simple Linear Regression

The response is obtained using a single feature when two variables are linearly related. A line in which data is best fit is called regression line and is represented by the equations

B. Robust Regression

Robust regression methods are designed to overcome certain limitations of traditional parametric and non- parametric methods and to be resistant to the violations of the underlying data-generating process assumptions and the robustness of the large data sets is tested

C. Ridge regression

Ridge regression acts as a possible solution to the imprecision of least square estimators[6] when linear regression models have certain multicollinear (highly correlated) independent variables—by creating a ridge regression estimator thus reducing the standard errors. The formula for ridge regression is

D. Lasso regression

Lasso regression is a regression analysis method that performs both variable selection and regularization overcoming the prediction error when only a few covariates have a strong relationship with the outcome and unlike ridge selection also performs covariate selection as well by excluding certain coefficients from impacting the prediction by forcing it to a zero-value through the sum of the absolute value of the regression coefficients to be less than a fixed value, besides this it also uses soft thresholding. It is Mathematically represented as

III. FEATURES OF THE DATASET

Ticker- Change in the price

Date

Time

Open- Open price of day

High- Highest price in day

Low- Lowest price in day

Close- Close price of day

IV. MEASURE OF PERFORMANCE AND MATHEMATICAL SIGNIFICANCE

MAE The Mean absolute error measuring the average of the residuals in the dataset represents the average of the absolute difference between the actual and predicted values in the dataset.

MSE Mean Squared Error measuring the variance of the residuals represents the average of the squared difference between the original and predicted values in the data set.

RME Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals. Since it is differentiable, it is used as a default metric for computing the loss by data analysts.

R^2 The coefficient of determination or R-squared is a scale-free score (i.e. irrespective of the values being small or large, the value of R square will be less than one) representing the proportion of the variance in the dependent variable explained by the linear regression model.

Adjusted R squared, a modified version of R square is adjusted for the number of independent variables in the model, and would always be less than or equal to R^2.In adjoined expression the number of observations in the data and the number of the independent variables in the data are denoted by n and k respectively.

OLS regression serves as a good supervised algorithm that has a training procedure and a deployment procedure though it over-fits in some circumstances. the only shortcoming being that here, there's no means to stop the training when it over fits.

The least squares method finds application in a wide variety of fields, including finance and investing. Financial analysts, use the method to quantify the relationship between two or more variables—such as a stock's share price and its earnings per share (EPS). By performing this type of analysis the future behavior of stock prices or other factors are predicted. To illustrate, consider the case of an investor considering whether to invest in a gold mining company. for instance an investor intending to know how sensitive the company-stock-price is to changes in the market price of gold, would use the least squares to trace the relationship between the two variables over time onto a scatter plot to help the investor predict the degree to which the stock's price would likely rise or fall for any fluctuations in the price of gold.

the Data set is first cleaned by removal of null and duplicates by the use of the Exploratory Data Analytic tools of the pandas library, further the attributes the data holds is analysed and following which the data is obtained in suitable array dimensions and pipelining is performed for preprocessing

the data set is now split into the training and the testing components in the ratio 80:20 by the train_test_split method of the sklearn library. RANSACRegressor, Ridge, Lasso, Sequential modules of the sklearn and keras are used to fit the data against the test component to obtain the scatterplot of the open price against the close price

V. SALIENT ASPECTS OF THE CODE

```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression(normalize=True)
lin_reg.fit(X_train,y_train)
from sklearn.linear_model import
RANSACRegressor
model =
RANSACRegressor(base_estimator=LinearReg
ression(), max_trials=100) model.fit(X_train,
y_train)
from sklearn.linear_model import Ridge model =
Ridge(alpha=100, solver='cholesky', tol=0.0001,
random_state=42)
model.fit(X_train, y_train) pred =
model.predict(X_test)
from sklearn.linear_model import Lasso model =
Lasso(alpha=0.1,
precompute=True, positive=True,
selection='random', random_state=42)
model.fit(X_train, y_train)
```

```
test_pred = model.predict(X_test) train_pred =
model.predict(X_train)
print('Test set evaluation:\n_ ')
print_evaluate(y_test, test_pred)
print('================================
===')
print('Train set evaluation:\n_ ')
print_evaluate(y_train, train_pred)
results_df_2 =
pd.DataFrame(data=[[use_model_here,
*evaluate(y_test, test_pred) , cross_val(Lasso())]],
columns=['Model', 'MAE', 'MSE', 'RMSE', 'R2
Square', "Cross Validation"]) results_df =
results_df.append(results_df_2, ignore_index=True)
a = df2['Open'] b = df2['Close']
X2 = sm.add_constant(a) est = sm.OLS(b, X2) est2 =
est.fit() print(est2.summary())
```

VI. RESULTS

Table 1: Results of various algorithms

Model	Linear	Robust	Ridge	Lasso
MAE	3.845779	3.845779	6.633434	3.850596
MSE	30.11601	30.11601	64.82109	30.12317
RMSE	5.487806	5.487806	8.051155	5.488458
R2 Square	0.997035	0.997035	0.993619	0.997035
Cross Validation	0.925135	0.925135	0.925135	0.92513

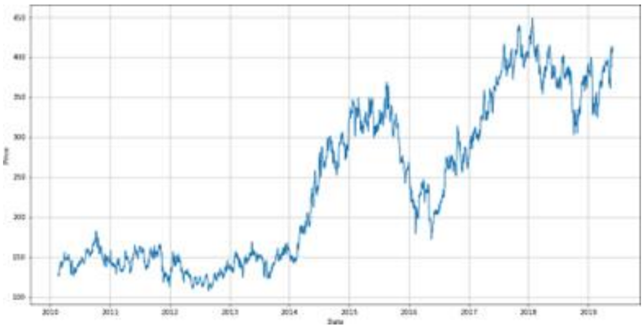

Fig 2:Data Representation_1

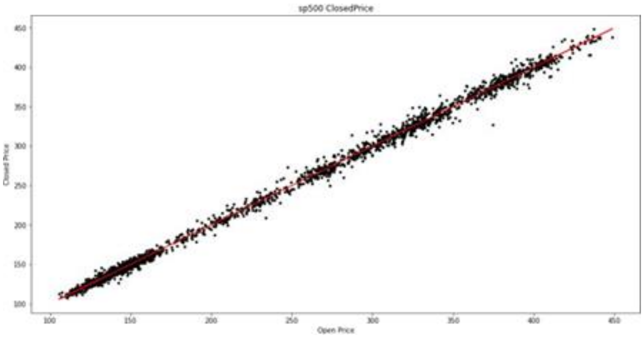

Fig 3: Data Representation_2

VII. REFERENCES

[1]. Stock Market Predication Using A Linear Regression- Dinesh Bhuriya ,Girish Kaushal, Upendra Singh

[2]. Application of a Multi-factor Linear Regression Model for Stock Portfolio Optimization-Zhihao PENG, Xucheng LI

[3]. Developing a Prediction Model for Stock Analysis-R. Yamini Nivetha, Dr. C. Dhaya

[4]. A Comparative Analysis on Linear Regression and Support Vector Regression- Kavitha S, Varuna S, Ramya R

[5]. Performan ce measures -MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-o f-determination-adjusted-r-squared-which-metric- is-better-cd0326a5697e

[6]. Hilt, Donald E.; Seegrist, Donald W. (1977). Ridge, a computer program for cal culating ridge regression estimates