Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

A Regression Analysis on the Car Index in the Tehran Stock Exchange

Arash Salehpour[1], Elaheh Salehpour[2]

[1]Department of Computer Engineering, Islamic Azad University, Rasht Branch Iran
[2]Department of Mechanical Engineering, University College of Nabi Akram, Tabriz Iran

E-mail: [1]arash.salehpour@yahoo.com, [2]elaheh.salehpour@yahoo.com

Abstract

One of the best ways to make money on the capital market is to buy shares on the stock exchange. The stock market has a nonlinear and chaotic system that is influenced by political, economic, and psychological conditions, and systems such as regression can be used to predict stock prices. In this research, different regression models are used, each of which measures information in a different way and tests the ability to predict the behaviour of index prices with this information. This paper examines linear regression, robust regression, ridge regression, polynomial regression, and elastic net on the historical daily data from 2018-07-01 to 2022-09-28 in the Car index of the Tehran Stock Exchange. Based on the empirical results, it is found that the best R2 score has been attained by the robust regression model. MSE, RMSE, MAE, and R2 for all models have been compared.

Keywords: Stock market prediction, Regression, Stock market, Machine learning, Tehran stock exchange

1. Introduction

One of the available options for cash investment is the stock market and securities. According to the nonlinear relationships between the variables affecting the stock price, artificial intelligence is one of the most suitable approaches to predict the stock price. The issuance of bonds and stock market shares is one method of providing capital for investment. People in this market expect to make a profit. The first and most important factor that is faced by the investor when investing in the stock market is the stock price factor. Stock price forecasting is an important part of the financial markets that has gotten a lot of attention from academics and experts over the past few decades. The importance of this issue comes from the fact that stock price prediction in financial markets is one of the important variables in the

DOI: https://doi.org/10.36548/jscp.2022.4.003
Received: 19.09.2022, received in revised form: 27.10.2022, accepted: 12.11.2022, published: 21.11.2022

field of investment decisions, securities pricing, derivatives, and risk management. Because stock market investors are always interested in knowing the next trend of prices, the actors in this market are trying to find and apply methods so that they can increase the profit of their capital by predicting the future stock price. Therefore, it seems necessary that appropriate, correct, and scientific methods are used to determine the future price of stocks for investors [1]. In financial time series forecasting, predicting how the price of a stock will change is seen as a difficult task. A correct prediction of stock price changes can bring a lot of profit to investors. Due to the complexity of stock market data, it is very difficult to develop efficient models for forecasting. Since the stock price is one of the most important factors in investment, decisions and its forecasting can play an important role in this field.

In this research, an attempt has been made to present a model so that, based on that, the movement of the target index stock price can be predicted with high accuracy. Investors constantly review past pricing history and use it to influence their future investment decisions. In the stock market, investors are warned not to follow the market trend. It is assumed that the best bet on market movements is that they will continue in the same direction. This concept is rooted in behavioural finance. With so many stocks to choose from, why do investors keep their money in falling stocks rather than rising ones? Studies have shown that mutual fund inflows are positively correlated with market returns. In fact, when more people invest, the market grows, which encourages even more people to buy. This is positive feedback reversion to the mean. Experienced investors who have seen many ups and downs in the market often believe that the market will even out over time. Historically, high market prices often discourage these investors from investing, while historically low prices may represent an opportunity. The tendency of stock prices to converge on an average value over time is called mean reversion.

The future of the stock market is algorithmic trading. Large banks, hedge funds, and institutional investors routinely use computerized trading algorithms. Algorithmic trading has revolutionized the stock market and the industry around it. A significant percentage of the transactions that are currently carried out around the world are done through robots. Is it possible for the machines to predict the stock values is a query. Scientists, analysts, and researchers around the world have long been trying to find a way to answer these questions. In the past few years, through artificial intelligence technology, decisions about stocks have been made about what to invest in and when. Stock price forecasting with artificial intelligence and machine learning is the process of predicting the future value of stocks

traded on the stock exchange for profit. With so many factors involved in stock price forecasting, forecasting stock prices with high accuracy is challenging, and this is where artificial intelligence and machine learning play a vital role.

In fact, the most important challenge in the stock market today is to predict its price. Stock price data represent financial time series that are more difficult to predict due to their dynamic nature. By treating stock data as a time series, past stock prices and other parameters can be used with artificial intelligence for predicting stock prices for the next day or week. Machine learning models like Recurrent Neural Networks (RNN) or Long Short -Term Memory (LSTM) are often used to predict time series data like weather forecasts, election results, and of course the stock prices. The idea is to weigh the importance of recent and older data and determine which parameters have the most influence on "current" or "future" day prices. The machine learning model assigns weights to each market characteristic and determines how much history the model should look at to predict future stock prices. In this research, some most important models of regression are used to predict the stock price.

1.1 Car index in Tehran Stock Exchange

All the companies that operate in the stock market belong to a certain industry and are placed in one of the categories of different industries in the capital market. All the companies that operate in the same field are placed next to each other and form their own industry. The car industry is one of the most important and well-known industries in the capital market. Talking about the car industry in the stock market, it includes the industry of automobile manufacturers and the manufacturing of its related parts, as well as all listed and over the counter companies that are active in these two fields. Care to distinguish between the automobile industry and the sale of automobiles on the commodity exchange, is needed. These two branches are completely separate and different from each other. This industry has many symbols for large or small markets.

2. Data Description

A historical daily data from 2018-07-01 to 2022-09-28, sourced from the online and freely available en.tsetmc.com, has been employed and is depicted in Fig.1. Before using this data for training and testing, it has been cleaned up. The dataset is set up with columns of "Date", "Open", "High", "Low", "Close", and "Vol".

Starting date: 2018-07-01 00:00:00

Ending date: 2022-09-28 00:00:00

Duration: 1550 days 00:00:00

	Date	Open	High	Low	Vol	Close
0	2018-07-01 00:00:00	15538.400000	15538.400000	15136.900000	170924640	15162.400000
1	2018-07-02 00:00:00	15203.500000	15318.700000	15203.500000	136980452	15244.400000
2	2018-07-03 00:00:00	15266.400000	15278.000000	15102.100000	265688184	15105.200000
3	2018-07-04 00:00:00	15127.000000	15546.900000	15052.300000	429823516	15546.900000
4	2018-07-07 00:00:00	15665.600000	15846.400000	15665.600000	205938813	15766.100000

Figure 1. Data Head -Pandas Data Frame

2.1 Data Info

Fig.2 shows the dataset information derived from pandas data frame.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1021 entries, 0 to 1020
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    1021 non-null   datetime64[ns]
 1   Open    1021 non-null   float64
 2   High    1021 non-null   float64
 3   Low     1021 non-null   float64
 4   Vol     1021 non-null   int64
 5   Close   1021 non-null   float64
dtypes: datetime64[ns](1), float64(4), int64(1)
memory usage: 48.0 KB
```

Figure 2. Dataset information

2.2 OHLC Data visualization

The chart of data visualization based on OHLC (Open-high-low-close chart) is depicted in Figure 3, and it is visualised using the open high, low, and close prices from the dataset in Jupyter Notebook in Python using the Plotly library.

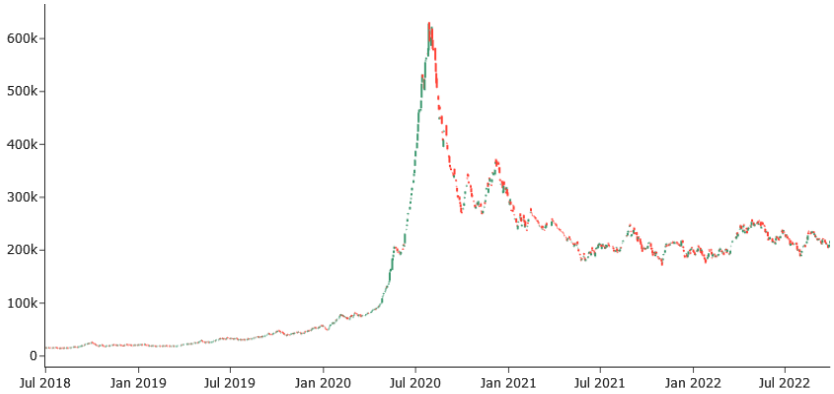

Figure 3. Data visualization based on OHLC

2.3 Heat Map

Heatmaps are usually used to plot the correlation between numerical columns in the form of a matrix. Looking at the output, it can be seen that drawing a box for each combination of rows and columns is essentially, which is what the heat map does. The colour of the box depends on the size of the house. For example, if there is a high correlation between two features, the corresponding house or box is white. On the other hand, if there is no correlation, the corresponding house remains black. This colour spectrum changes from the lowest value to the highest value of all houses of the matrix, which can be seen on the right side of the table shown in fig.4.

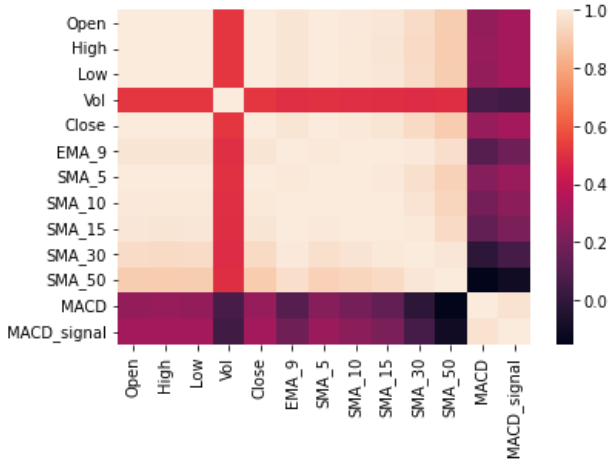

Figure 4. Heat Map

2.4 MACD

Fig.5 depicts the Moving Average Convergence Divergence (MACD) oscillator, which indicates the strength, direction, and acceleration of a stock trend. The data used to reach the appropriate output of this oscillator are the moving averages 26, 12, and 9.

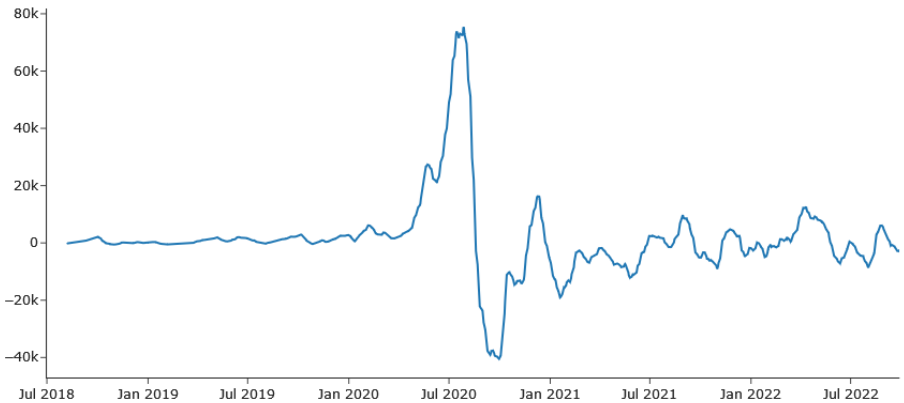

Figure 5. MACD

2.5 Moving Averages

Fig.6 depicts the moving average, which is one of the most common technical analysis indicators that forms the basis of many trading tools. In statistics, a "moving average" is a calculation that is performed to analyse data points by averaging different subsets of the entire dataset.

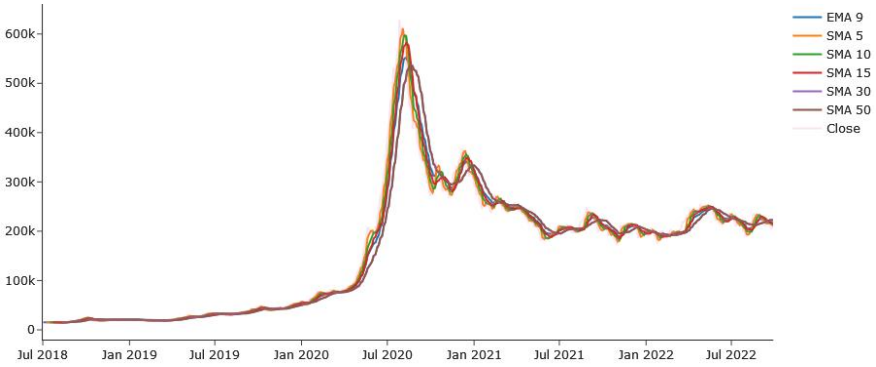

Figure 6. Moving Averages

2.6 Relative strength index

RSI is a technical indicator and is used in financial market analysis. There are many indicators, but the RSI is one of the most popular indicators. It was invented by Welles Wilder in 1978, and it shows the current and past strength and weakness of the chart based on the trades made in a recent period. Fig.7 shows the Relative Strength Index (RSI) indicator.

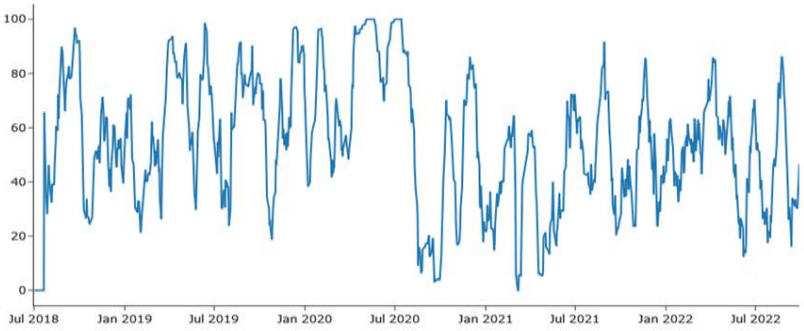

Figure 7. Relative strength index

3. Methods

Using data to discover the relationship between them is the basis of data analysis. One of the tools for relationship measurement and modelling is the use of statistical regression tools. Today, in order to analyze and discover the model of Big Data, various regression methods have been developed. The use of simple linear regression analysis is widely used in

various data analyzing sciences, especially in the subject of machine learning, chemistry, and biological sciences. Regression analysis is a process to estimate relationships between variables. This method includes many techniques for modelling and analyzing specific and unique variables, focusing on the relationship between the dependent variable and one or more independent variables.

In artificial intelligence, the mathematical method of finding the relationship between two or more variables is known as regression. This concept is widely used in machine learning to predict the behavior of one variable depending on the value of another variable. Many factors work together to create or cause an event, or situation. All factors are imagined together, and these items are predictive variables. In this paper, Python in Jupyter Notebook has been used. NumPy, Pandas Mathplotlib, Ploty, and Sckit learn are available for Windows implementation. Fig. 8 depicts the features used in X_train, obtained from pandas in X_train dataset, for training the model and Fig.9 depicts the model process step by step.

	Vol	EMA_9	SMA_5	SMA_10	SMA_15	SMA_30	SMA_50	RSI	MACD	MACD_signal
50	430029465	17056.386921	18155.84	17562.04	17204.080000	16054.640000	15776.718	78.234583	824.169913	622.950041
51	562050399	17234.213057	18456.56	17705.97	17432.493333	16207.933333	15849.998	78.174879	896.202822	677.733055
52	371010029	17455.877101	18835.02	17926.93	17663.420000	16362.163333	15933.942	78.691197	1005.075216	743.328367

Figure 8. Features in X-train head

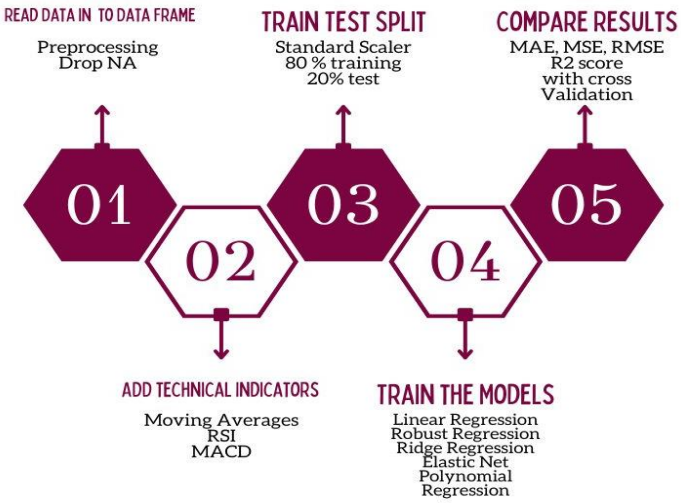

Figure 9. Step by Step Algorithm

3.1 Linear Regression

Linear Regression (LR) can be considered as the simplest type of method. Linear regression is divided into two categories: Simple linear and Multiple linear. These two types

include a complete and extensive structure and meet the requirements of a wide range of analyses. Fig.10 shows the actual vs predicted Linear Regression chart.

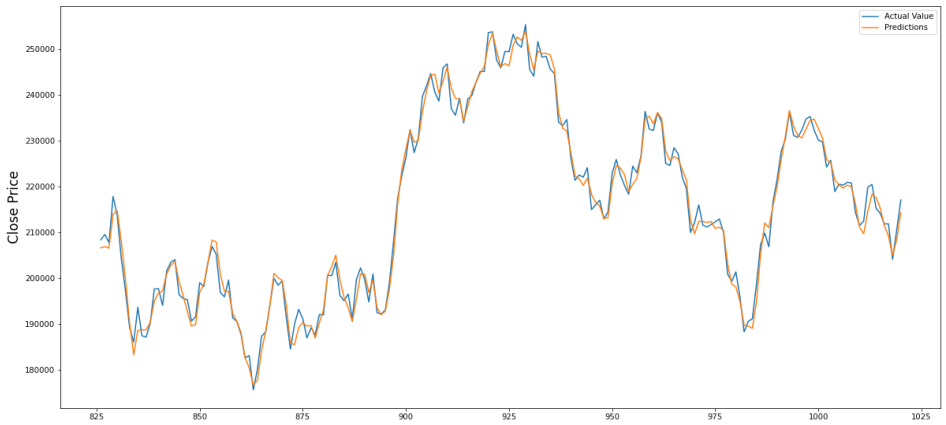

Figure 10. Linear Regression chart

Table 1. Empirical Result -LR

Test set evaluation:	**Train set evaluation:**	
MAE: 1660.316266166592 **MSE:** 4069485.4508623504 **RMSE:** 2017.296569883157 **R2 Square** 0.9894284150958285	**MAE**: 1263.1094374179263 **MSE**: 5378333.954227737 **RMSE**: 2319.1235314721243 **R2 Square** 0.9997126172400286	

	Actual	Predicted
826	208425.0	206705.972492
827	209596.0	206943.125830
828	207842.0	206606.391357
829	217901.0	214025.356532
830	213772.0	214760.088788

3.2 Robust Regression

It is often believed that regression methods are stable and resistant to outliers [2]. Popular regression methods such as least squares perform admirably when the assumptions upon which they are based turn out to be accurate; however, these methods can struggle when presented with data that is not ideal. For instance, the approach that uses the fewest squares particularly has extreme data points. The fact that the data are non-variable also contributes to the complexity of the situation [3]. In order to achieve stable regression, a variety of parametric and non-parametric methods have been proposed. It is a more stable alternative to the least squares method, and in it, the absolute value of the error is utilized as opposed to the second power of the regression error. Another name for this method is the minimum absolute value method [4, 5]. Several parametric and nonparametric methods for stable regression

have been described. In contrast to least squares, which takes the square root of the regression error into account, the absolute value of the error is employed in the minimum absolute value technique, which is considered more accurate [2]. Fig .11 shows the actual vs predicted Robust Regression chart.

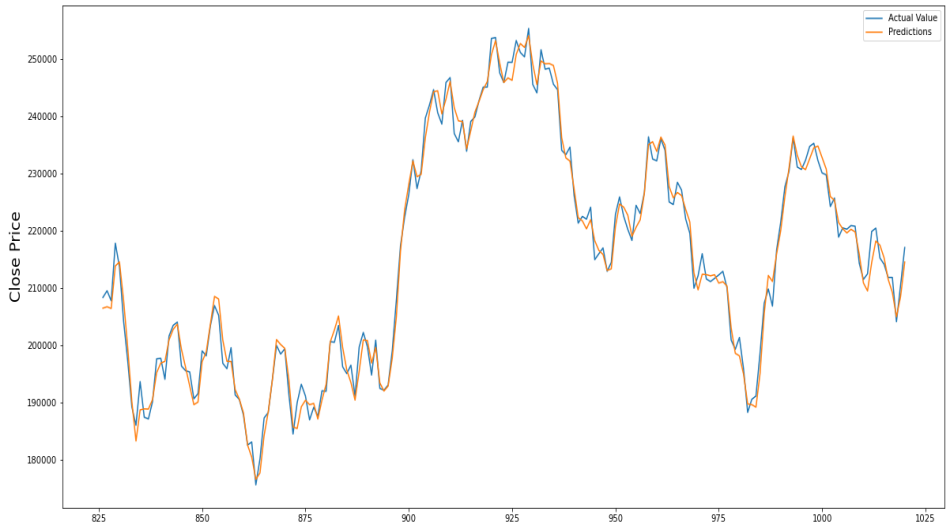

Figure 11. Robust Regression chart

Table 2. Empirical Result -Robust Regression

Test set evaluation:	**Train set evaluation:**	
MAE: 1656.655472572617 **MSE:** 4049794.076793364 **RMSE:** 2012.4100170674374 **R2 Square** 0.9894795687454394	**MAE**: 1272.2912728432873 **MSE**: 5407956.055290078 **RMSE**: 2325.501248180718 **R2 Square** 0.9997110344299553	

	Actual	Predicted
826	208425.0	206705.972492
827	209596.0	206943.125830
828	207842.0	206606.391357
829	217901.0	214025.356532
830	213772.0	214760.088788

3.3 Ridge Regression

In multiple regression discussions, the number of independent variables to be used in a model becomes an issue. By increasing the numbers, overfitting occurs, and by reducing them, it is possible to deal with underfitting issues. If the regression model suffers from overfitting, its error will depend more on the new values. With the presence of less than necessary variables in the model, underfitting increases the variance of the model. Therefore, by increasing the number of variables, the problem of collinearity and overfitting appears,

and by decreasing them, the variance of the model will increase. One of the ways to overcome these problems in multiple regression is to use the Ridge Regression model. Since there are many model variables or multiple collinearities, the variance of the estimators is inflated and becomes a peak collinearity. Therefore, this regression method overcomes this problem [6]. Fig .12 shows the actual vs predicted Ridge Regression chart.

Figure 12. Ridge Regression chart

Table 3. Empirical Result -Ridge Regression

Test set evaluation:	**Train set evaluation:**	
MAE: 5816.609559764159 **MSE:** 54644931.57607733 **RMSE:** 7392.221017804955 **R2 Square** 0.8580450671922837	**MAE**: 8246.774864341962 **MSE**: 161228343.10075405 **RMSE**: 12697.57233099123 **R2 Square** 0.9913850187399574	

	Actual	Predicted
827	209596.0	205701.761069
828	207842.0	204365.009837
829	217901.0	210997.467201
830	213772.0	214447.982298
831	204278.0	210268.343901

3.4 Elastic Net

Although the OLS least squares estimator has desirable features [7], especially unbiasedness, it can suffer from the large variance problem in some situations. If the number of regressors explanatory variables is large in comparison to the sample size, or if there are many correlated regressors, least square estimates are very sensitive to random errors, with a

high variance and poor performance. One of the solutions for this problem is the use of pure elastic regression patterns [8]. Fig.13 shows the actual vs predicted Elastic Net chart.

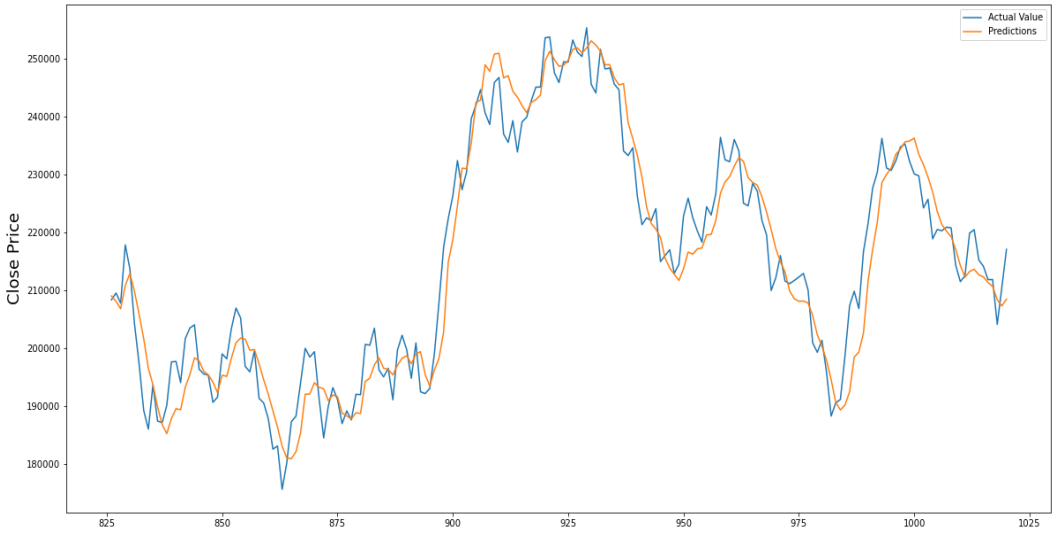

Figure 13. Elastic Net

Table 4. Empirical Result -Elastic Net

Test set evaluation:	**Train set evaluation:**	
MAE: 4182.558838954719 **MSE:** 28750191.91281336 **RMSE:** 5361.920543314061 **R2 Square** 0.9253136303133543	**MAE**: 4098.977498555391 **MSE**: 48259618.19986959 **RMSE**: 6946.914293401754 **R2 Square** 0.9974213237051697	

	Actual	Predicted
826	208425.0	209073.364138
827	209596.0	208175.727464
828	207842.0	206865.584783
829	217901.0	210903.726165
830	213772.0	212972.010260

3.5 Polynomial Regression

This type of average value is considered the most effective and widely used method of multivariate analysis. If several independent variables can predict a dependent variable and form a linear relationship, then it can be said that this is a multiple regression method. The mentioned item can be researched and compared across multiple models. All these models ultimately create a work that is very useful and profitable. In regression, all the relationships between independent and dependent variables can be found and, the sales plan of a company in the future can be predicted. In regression analysis, how to predict the changes in the

dependent variable by using the independent variables is understood. The purpose of this model is to help understand how the sum of squares is calculated. How to calculate the mentioned phenomenon can be referred from related sites [9] and [10]. Fig.14 shows the actual vs predicted Polynomial Regression chart.

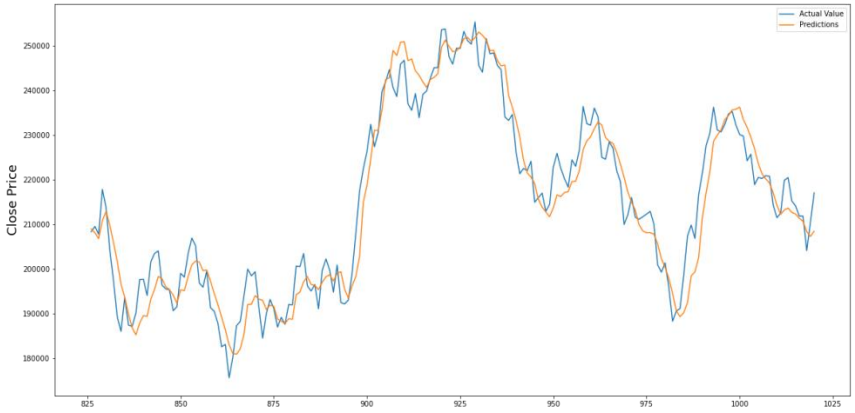

Figure 14. Polynomial Regression chart

Table 5. Empirical Result -Polynomial Regression

Test set evaluation:	**Train set evaluation:**	
MAE: 2064.7156730507118 **MSE:** 6301803.047009912 **RMSE:** 2510.339229468781 **R2 Square** 0.9836293686842609	**MAE**: 1085.5124721698949 **MSE**: 3049824.434869593 **RMSE**: 1746.3746547833293 **R2 Square** 0.9998370374597448	

	Actual	Predicted
826	208425.0	209073.364138
827	209596.0	208175.727464
828	207842.0	206865.584783
829	217901.0	210903.726165
830	213772.0	212972.010260

4. Empirical Results

Table 6 and Fig. 15 show the empirical results based on the experiments.

Table 6. Empirical Results of all models

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	1660.316266	4.069485e+06	2017.296570	0.989428	0.989274
1	Robust Regression	1656.655473	4.049794e+06	2012.410017	0.989480	0.989454
2	Ridge Regression	5816.609560	5.464493e+07	7392.221018	0.858045	0.989274
3	Elastic Net Regression	4182.558839	2.875019e+07	5361.920543	0.925314	0.966240
4	Polynomail Regression	2064.715673	6.301803e+06	2510.339229	0.983629	0.000000

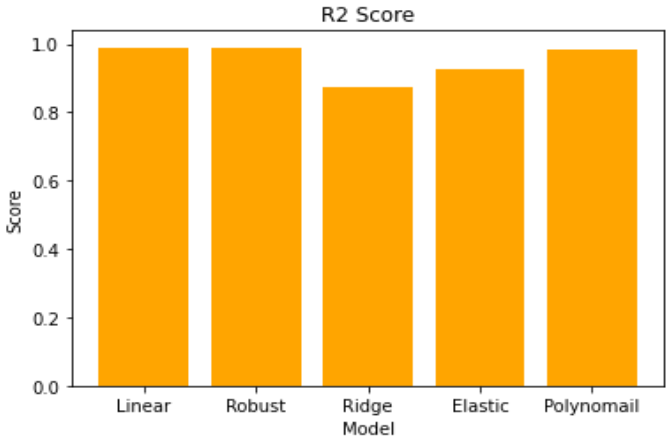

Figure 15. R2 Score Bar chart

5. Conclusion

Because the world economy is growing quickly and there are more and more capital markets, investors need access to methods that are both reliable and strong for predicting the prices of stocks in the future. Before computers and other electronic forecasting tools were widely used in the stock market, investors had to rely on a variety of manual forecasting techniques in order to maximize profits while minimizing losses. In this study, stock value changes has been predicted using different regression models. For this purpose, the price data of the car index has been used. A number of technical indicators were first calculated. By using different regression models, an optimal model for stock price prediction can be provided. The results show that, using robust regression provides considerable accuracy when compared to other forecasting methods, and it provides better results than the existing methods. Due to the fact that this dataset has not been analysed before, it could not be compared with other methods using machine learning models. In the future, this database would be researched with other machine learning and deep learning algorithms and the results would be compared.

Author contributions: All authors listed have made a substantial, direct, and intellectual contribution to the work and have approved it for publication.

Data Availability Statement: publicly available datasets were analysed in this study. These data can be found in: http://en.tsetmc.com/Site.aspx, Tehran Securities Exchange Technology Management Co. – TSETMC

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Bazrkar, M.J. and S. Hosseini, Predict Stock Prices Using Supervised Learning Algorithms and Particle Swarm Optimization Algorithm. Computational Economics, 2022.

[2] Andersen, R., Modern Methods for Robust Regression Quantitative Applications in the Social Sciences. September 2007, University of Toronto, Canada: SAGE.

[3] Masoumi, M., et al., Economic and non-economic determinants of Iranian pharmaceutical companiesâ€™ financial performance: an empirical study. BMC Health Services Research, 2019. 19(1): p. 1011.

[4] Patel, J., et al., Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 2015. 42(4): p. 2162-2172.

[5] Strutz, T., Data Fitting and Uncertainty: A practical introduction to weighted least squares and beyond, ed. Springer.

[6] Ahmadi, E., et al., New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic. Expert Systems with Applications, 2018. 94: p. 21-31.

[7] Arash Salehpour, Bibliometric Review of Applications of Deep Learning in Marketing: Advances in Resources and Top Trend Analysis. Journal of Artificial Intelligence and Capsule Networks, 2022. 4 (4): p. 230-244.

[8] Moodie, M., et al., Cost-Effectiveness of Fiscal Policies to Prevent Obesity. Current obesity reports, 2013. 2: p. 211-224.

[9] Kumar, D., P.K. Sarangi, and R. Verma, A systematic review of stock market prediction using machine learning and statistical techniques. Materials Today: Proceedings, 2022. 49: p. 3187-3191.

[10] Ghobaei-Arani, M., et al., An efficient task scheduling approach using moth-flame optimization algorithm for cyber-physical system applications in fog computing. Transactions on Emerging Telecommunications Technologies, 2020. 31(2): p. e3770.