I.J. Intelligent Systems and Applications, **2023, 5, 12-27**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2023.05.02

Predicting Automobile Stock Prices Index in the Tehran Stock Exchange Using Machine Learning Models

Arash Salehpour*
Department of Computer Engineering, Islamic Azad University, Rasht Branch Iran
E-mail: arash.salehpour4@gmail.com
ORCID iD: https://orcid.org/ 0000-0003-2770-1619
*Corresponding Author

Received: 10 April 2023; Revised: 11 July 2023; Accepted: 24 August 2023; Published: 08 October 2023

Abstract: This paper analyses the performance of machine learning models in forecasting the Tehran Stock Exchange's automobile index. Historical daily data from 2018-2022 was pre-processed and used to train Linear Regression (LR), Support Vector Regression (SVR), and Random Forest (RF) models. The models were evaluated on mean absolute error, mean squared error, root mean squared error and R2 score metrics. The results indicate that LR and SVR outperformed RF in predicting automobile stock prices, with LR achieving the lowest error scores. This demonstrates the capability of machine learning techniques to model complex, nonlinear relationships in financial time series data. This pioneering study on a previously unexplored dataset provides empirical evidence that LR and SVR can reliably forecast automobile stock market prices, holding promise for investing applications.

Index Terms: Machine Learning, Stock Price Prediction, Linear Regression, Support Vector Regression, Random Forest, Tehran Stock Exchange.

1. Introduction

Predicting stock market prices has long been an essential and challenging task in financial forecasting. With the advancement of machine learning techniques in recent years, researchers have increasingly applied machine learning models to stock market prediction. In this paper, we investigate the application of three machine learning techniques - Linear Regression, Support Vector Regression (SVR), and Random Forest - for predicting the stock prices of the Car Index in the Tehran Stock Exchange. The Tehran Stock Exchange [1] is the largest in Iran, with over 330 companies listed across 37 industries. The Car Index consists of stocks of major automotive companies in Iran, including Iran Khodro and SAIPA, which account for over 90% of domestic car production. Predicting the Car Index is essential for investors and policymakers to understand the performance of this vital sector of the Iranian economy. Previous works have applied machine learning models to predict stock market indices[2]. For instance, we used neural networks to predict the Tehran Stock Exchange index while utilizing SVR for the Istanbul Stock Exchange index. However, the application of machine learning specifically for the Tehran Stock Exchange Car Index has been limited. Through this study, we contribute to the literature on stock prediction by evaluating different machine-learning techniques for this nascent index[3]. Accurately forecasting stock prices has long been challenging in financial economics research. While numerous models have been proposed, they have yet to emerge as robust and reliable market volatility predictors over the short and long term [4-7]. Fundamental and technical analysis remains the predominant approaches, but each has limitations [8]. Recent studies have increasingly applied machine learning techniques to stock prediction, but their performance tradeoffs still need to be clarified, especially on emerging market datasets [9, 10]. This paper aims to evaluate different machine learning models for forecasting the automobile stock index of the Tehran Stock Exchange. As a primary emerging market with substantial volatility, modeling the Tehran exchange poses unique challenges. To date, no extensive studies have tested predictive capabilities on its automobile index specifically. We compare three popular methods - Linear Regression (LR), Support Vector Regression (SVR), and Random Forest (RF) - on this untapped dataset to assess their forecasting accuracy. The automobile sector's complexity and nonlinearity will rigorously examine each technique's robustness. The Tehran Stock Exchange (TSE) is Iran's largest stock exchange, established in 1967. It has grown to become one of the largest stock exchanges in the Middle East, with over 300 listed companies and a market capitalization of around $200 billion as of 2018. The TSE's main index is called TEDPIX, which tracks the performance of the top 100 companies on the exchange. The TSE and TEDPIX play an important role in Iran's economy by providing capital for companies and investment opportunities for domestic and foreign investors. Several factors can influence stock returns and volatility in the TSE, such as macroeconomic conditions, political events, oil price shocks, and investor behavior. Oil price volatility in particular has been found to significantly impact returns in the TSE, given Iran's dependence on oil exports. Global financial crises can also transmit shocks to the TSE, evidenced by the impact of the 2008 crisis.

Our core objectives are

- To establish benchmark machine learning predictive performance on the automobile index,
- To identify the strengths and limitations of each modeling approach, and
- To determine the best-suited technique for financial time series forecasting on emerging markets.
- This study will provide novel empirical insights into automobile stock prediction while advancing the literature on applying machine learning to noisy, real-world financial datasets.

The TSE has a combination of experienced institutional investors and less knowledgeable individual investors participating. This can result in the possibility of price bubbles and volatility. The TSE displays characteristics of an emerging market with potential risks from economic and political uncertainty. Despite this, it also offers growth potential as Iran's financial markets continue to develop.

The rest of the paper is organized as follows. Section 2 Literature Review includes Overview of stock forecasting techniques, Machine learning for financial forecasting and Studies on the Tehran Stock Exchange Section 3 describes Data description, Data preprocessing, Feature engineering and Models (LR, SVR, RF) Section 4 explains Train/test split, Model implementation details, Performance metrics and Experimental process flow section 5 presents the results. Model predictions, Quantitative performance metrics and Comparison and analysis, Section 6 Discussion, Interpretation of results, Insights gained and Comparative strengths and weaknesses of model’s section 7 Conclusion describes Summary of findings, Contributions and Future work.

2. Literature Review

Financial market forecasting has been of significant interest to researchers and practitioners for decades. This review aims to explore different forecasting techniques, focusing on applying machine learning algorithms. The effectiveness of these techniques in predicting stock market fluctuations will be discussed.

2.1. Early Methods of Financial Forecasting

Abu-Mostafa and Atiya (1996) [11] gave an overview of financial forecasting, stressing the significance of comprehending the fundamental drivers that influence market behavior. They also discussed the difficulties in predicting market trends and stressed the importance of using reliable forecasting models to enhance investment strategies.

2.2. Chaos Theory and Financial Markets

The application of chaos theory in financial markets is used to detect the existence of nonlinear dynamics and their potential influence on market behavior. Blank (1991) [12] Investigated the presence of chaos in futures markets and found evidence of nonlinear dynamics Similarly, Decoster et al. (1992) provided evidence of chaos in commodity futures prices These studies contribute to understanding financial market complexity and the need for advanced forecasting methods to capture such nonlinear behavior.

2.3. Machine Learning in Financial Forecasting

Applying machine learning algorithms in financial forecasting has become increasingly popular. Huang et al. (2005) [13] They suggested using support vector machines (SVMs) to predict the direction of the stock market movement. Their findings showed that SVMs efficiently identify intricate patterns in financial data. Moody and Saffell (2001) [14] utilized direct reinforcement learning to predict the stock market movement and discovered that my approach yielded better results than other forecasting methods. Hsu et al. (2016) [15] comparing machine learning algorithms and financial economics models, it was discovered that machine learning methods typically display better forecasting performance than traditional models. These studies have demonstrated the potential of machine learning algorithms to enhance the accuracy of financial forecasting.

2.4. Deep Learning Methods

Financial forecasting tasks have been incorporating deep learning techniques, a type of machine learning that has gained popularity. Long et al. (2019) [16] They suggested utilizing deep learning-based feature engineering to forecast stock price movement. They showed that their strategy effectively captures intricate connections in financial data. Hochreiter and Schmidhuber (1997) [17] Long Short-Term Memory (LSTM) networks were introduced as a type of recurrent neural network. They have become famous for analyzing time series data, such as financial forecasting tasks.

2.5. Literature Review on Stock Return Modeling and Forecasting in the Tehran Stock Exchange

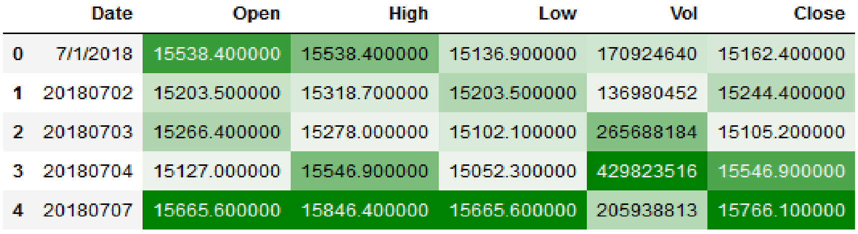
Several studies have examined stock return predictability and volatility modeling in the TSE. Shahrestani and Rafei [18] used Markov switching vector autoregressive models to analyze the impact of oil price shocks on TSE returns. They found asymmetric effects, with oil price increases having a more significant impact than decreases. Ebrahimi and Hajizadeh [19] proposed a novel data envelopment analysis model to assess the performance measurement of TSE firms using flexible inputs and outputs. Ramezanian et al. [20] developed an integrated machine learning model combining genetic network programming and neural networks to forecast daily TSE returns. Their hybrid approach outperformed benchmarks and individual models. Chizari et al. [21] studied the impact of pharmaceutical firms' intellectual capital in explaining TSE market valuation and performance. Human capital was found to be significant in the pharmaceutical industry. Several studies have modeled volatility dynamics in the TSE using GARCH models. Nejad et al. [22] applied Markov regime-switching GARCH models to account for structural breaks when estimating oil price volatility spillovers. Abounoori et al. [23] also used Markov switching GARCH to capture volatility clustering and asymmetry in the TSE effectively. Regarding investor behavior, Jalilvand et al. [24] provided evidence of informed institutional investors and uninformed individuals influencing price formation in the TSE. Mothlagh et al. [25] examined the relationship between abnormal TSE returns and changes in market value added as a measure of intellectual capital. They found a significant positive association between the two. Regarding forecasting, Ebrahimpour et al. [26] proposed a mixture of MLP-experts’ models that combined neural network experts optimized for different TSE time series data segments. Their ensemble approach improved short-term trend forecasting over single models. Lastly, Jahan-Parvar and Mohammadi [27] evaluated the risk-return relationship in the TSE using different asset pricing models and found evidence of higher returns being associated with higher systematic risk.

In this review section, we have looked at the evolution of financial forecasting techniques, from their early beginnings to the latest machine learning and deep learning algorithms. As chaos theory shows, the unpredictability and complexity of financial markets emphasize the importance of advanced forecasting methods that can identify and track these underlying patterns. By utilizing machine learning and deep learning techniques, we can enhance the accuracy of financial market predictions. However, more research is necessary to understand the potential of these methods in different market conditions and for various asset classes.

3. Data Preparation

3.1. What is the Car index in TEDPIX?

All companies operating in the stock market belong to a specific industry; Even if we don't want to limit our vision to the stock market, this is still true of all businesses. Boutiques are part of the clothing industry, and restaurants are part of the food industry. Various industries are represented on the stock exchange. In fact, we must say that the companies whose shares are offered in this market are from various industries. From the information and communication industry to the cement, iron, and plaster industries All the companies that operate in a certain field are placed together in an industrial group - which is related to their field of activity. Well, as expected, automobile manufacturing, as one of the country's largest industries, has various companies in the stock market. When talking about the car market, it means the automotive industry and parts manufacturing; that is, all the companies that operate in these two fields - and have exposed their shares for sale - are included in this collection. Note: You should not confuse the car exchange with the sale of cars in the commodity exchange. These two categories are completely separate from each other. Automotive index is a list of all the companies that make cars. These companies are part of the group that makes cars and car parts. All the companies that manufacture cars and their parts, and have exposed their shares for sale, are present in the stock market in the form of the car industry. Approximately more than 30 companies in the field of automobile manufacturing are present at the Tehran Stock Exchange and have offered their shares. The automotive industry in the stock market consists of the motor vehicle production subgroup   and the Subgroup of motor vehicle spare parts and accessories.



	Date	Open	High	Low	Vol	Close
0	7/1/2018	15538.400000	15538.400000	15136.900000	170924640	15162.400000
1	20180702	15203.500000	15318.700000	15203.500000	136980452	15244.400000
2	20180703	15266.400000	15278.000000	15102.100000	265688184	15105.200000
3	20180704	15127.000000	15546.900000	15052.300000	429823516	15546.900000
4	20180707	15665.600000	15846.400000	15665.600000	205938813	15766.100000

Fig.1. Figure 1 data head pandas’ data frame

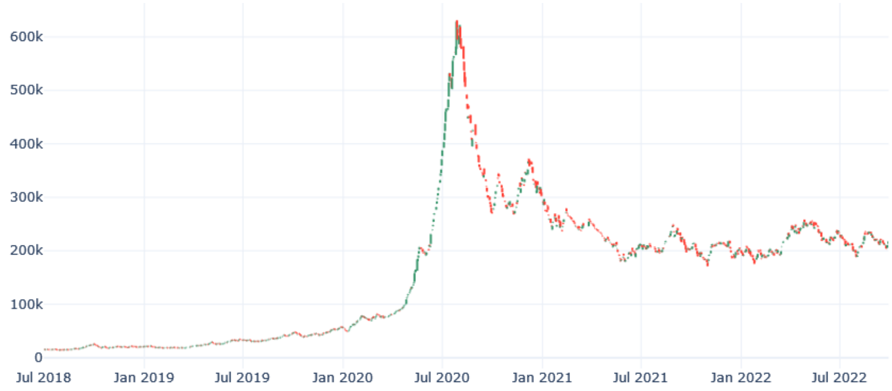


Fig.2. Car index close price base on OHLC data

3.2. *Data Description*

We employ historical daily data from 2018-07-01 to 2022-09-28, sourced from the online and freely available en.tsetmc.com, Before using this data for training and testing, it has been cleaned up. The data set is set up so that: "Date", "Open", "High", "Low", "Close", "Vol", "9-Day EMA", "5-Day SMA", "10-Day SMA", "15-Day SMA", "30-Day SMA", "50-Day SMA", "RSI", and "MACD_Signal" as OHLC and technical indicators.

3.3. *Seasonal-trend Decomposition Using LOESS (STL)*

A good way to break up time series that is often used in economic and environmental studies, by fitting regression models to the data locally, the STL method separates a time series into its trend, seasonal, and rest parts When you use the seasonal decompose function, you get a result object in return. Each of the four data points from the decomposition is made available to the user in the form of an array by the result object. This bit of code shows how to separate a series into its trend, seasonal, and residual parts, assuming that an additive model is being used[6].

$$y_i = s_i + t_i + r_i$$

- y_i = The value of the time series.
- s_i = The value of the seasonal component.
- t_i = The value of the trend component at.
- r_i = The value of the remainder component at.

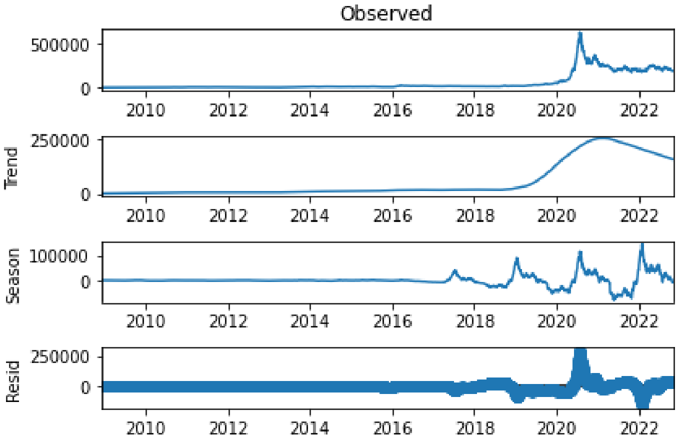


Fig.3. Seasonal-trend decomposition

3.4. *Technical Indicators*

Technical indicators, which are utilized by traders who follow technical analysis, are heuristic or pattern-based indications generated by the price, volume, and/or open interest of an asset or contract. Technical analysts utilize indicators to forecast future price changes by examining previous data. The Relative Strength Index (RSI), stochastics, and moving average convergence and divergence are a few examples of typical technical indicators (MACD)[28, 29].

Table 1. Technical indicators we use

Moving Averages Simple	5,10,15,30,50
Moving Averages Exponential	9
Relative Strength index	N=14
MACD	span=12, min_periods=12
MACD Signal	span=9, min_periods=9

A. *Moving Average*

A moving average (MA) is a stock indicator used often in technical analysis in the world of finance. The purpose of generating a stock's moving average is to create a continuously updated average price in order to smooth out the price data. The effects of random, short-term changes on the price of a stock over a certain time period are reduced by using the moving average calculation. A typical stock indicator in technical analysis is the moving average (MA). By generating a continuously updated average price, the moving average aids in leveling the price data over a given time. A simple moving average (SMA) is a formula that averages a series of prices over a predetermined period of time, often a certain number of days. An exponential moving average (EMA) is a weighted average that provides more weight to a stock's price in recent days, making it a more sensitive indicator to fresh data [28].

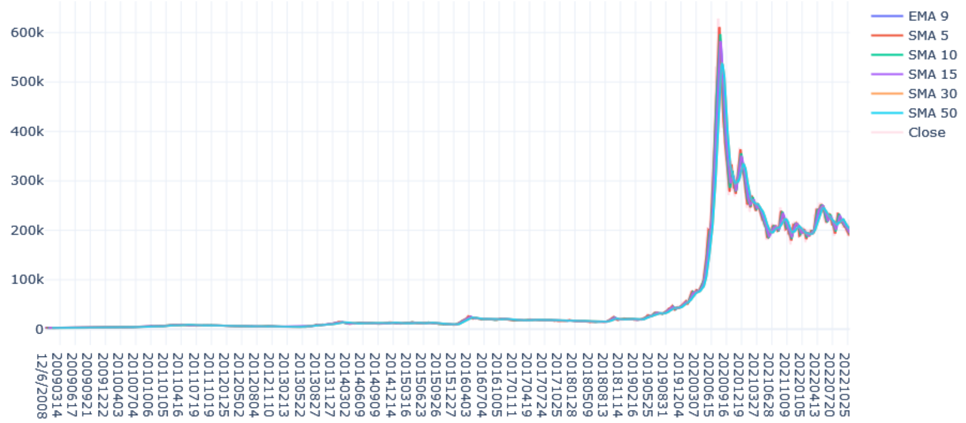


Fig.4. Moving averages

B. *Relative Strength Index*

Technical analysis uses the relative strength index (RSI). To assess whether a security's price is overpriced or undervalued, RSI evaluates the speed and amplitude of recent price fluctuations. An oscillator (a line graph) representing the RSI is shown, with a scale from 0 to 100. The indication was created by J. New Concepts in Technical Trading Systems, written by Welles Wilder Jr. and published in 1978, presented these ideas. Beyond identifying overbought and oversold assets, the RSI has other capabilities. It may also signal assets that are poised for a price correction or trend reversal. It may serve as a buying and selling cue. An overbought scenario is often indicated by an RSI value of 70 or above. An oversold state is indicated by a value of 30 or below [28].

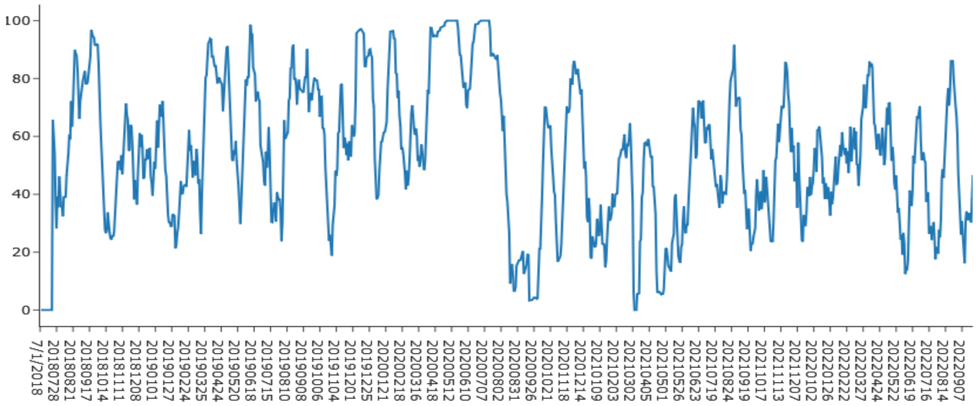


Fig.5. Relative strength index

C. Moving Average Divergence and Convergence

The connection between two exponential moving averages (EMAs) of the price of a security is shown by the trend-following as moving average convergence/divergence (MACD, or MAC-D). The 26-period EMA is subtracted from the 12-period EMA to generate the MACD line. The result of the math is the MACD line. The signal line is a nine-day moving average of the MACD line. It is drawn on top of the MACD line and can be used as a buy or sell signal. When the MACD line crosses above the signal line, traders can buy the asset. When the MACD line crosses below the signal line, traders can sell (or "short") the security. There are various ways to interpret MACD indicators, but the most popular ones include crosses, divergences, and rapid increases and falls[29, 30].

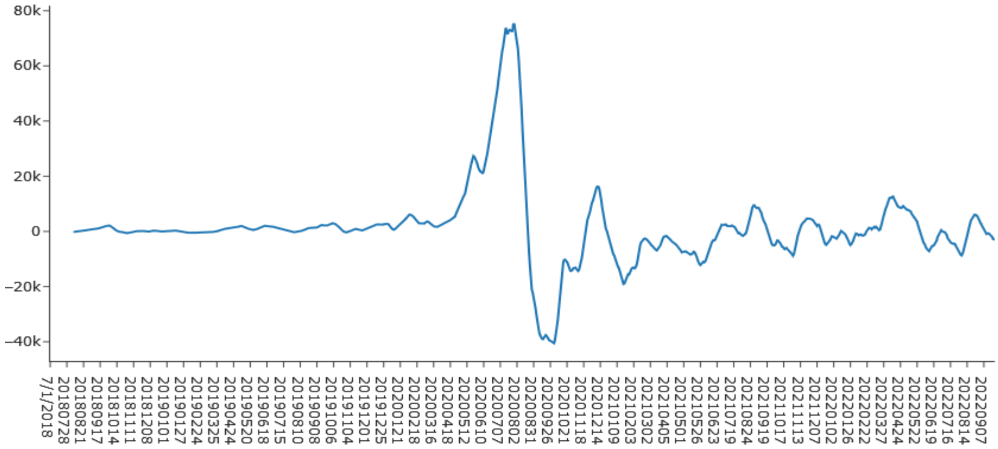
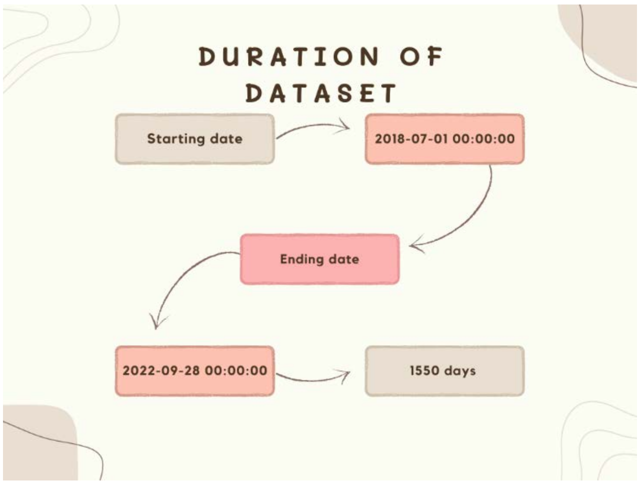


Fig.6. MACD

3.5. EDA - Exploratory Data Analysis

Exploratory data analysis is a crucial process that involves investigating data informally to look for patterns, identify outliers, test hypotheses, and validate presumptions using summary statistics and graphical representations. Exploratory data analysis is the term used to describe this approach.

A. Get the Duration of Dataset


DURATION OF DATASET

Starting date

2018-07-01 00:00:00

Ending date

2022-09-28 00:00:00

1550 days

Fig.7. Duration of data set

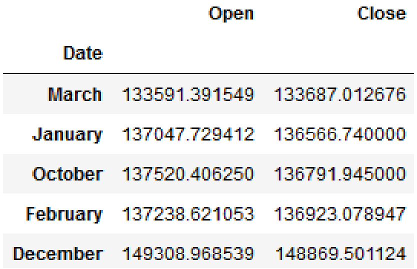
B. Month Wise Comparison between Stock Actual, Open and Close Price

Fig. 8, 9 and 10 depict a month wise comparison of the car index stock price and have shown the open and close price of each month.

3.6. Heatmap for Visualizing Data Correlations

Fig11. Visualize how well features correlate with each other with a heatmap the degree to which two variables are linked is shown by a statistical measure called correlation. There are two main kinds of correlations: positive and negative. There is a positive correlation when two variables move in the same direction, like when one goes up as the other goes up, Likewise, the other. a Heat map allows us to easily identify potential outliers within a dataset. Values tending toward darkness are negatively correlated, and those tending towards light are positively correlated. The darker

the color, the closer the value is to 0. Values tending towards dark red are negatively correlated, and those tending towards light are positively correlated. The darker the colour, the closer the value is to 0.



	Open	Close
Date		
March	133591.391549	133687.012676
January	137047.729412	136566.740000
October	137520.406250	136791.945000
February	137238.621053	136923.078947
December	149308.968539	148869.501124

Fig.8. Month wise comparison

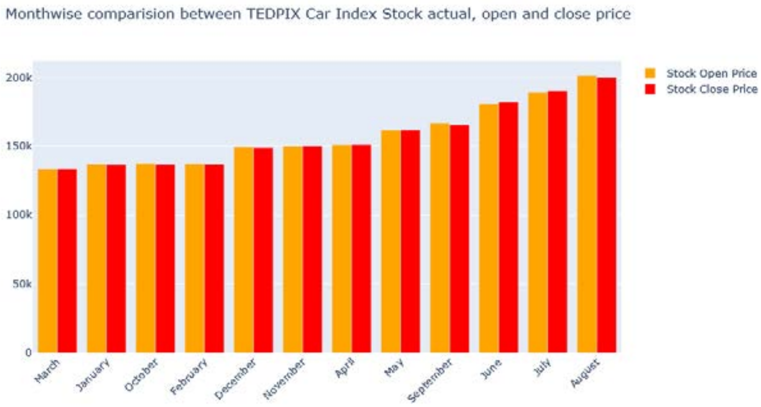


Fig.9. Month wise comparison bar chart

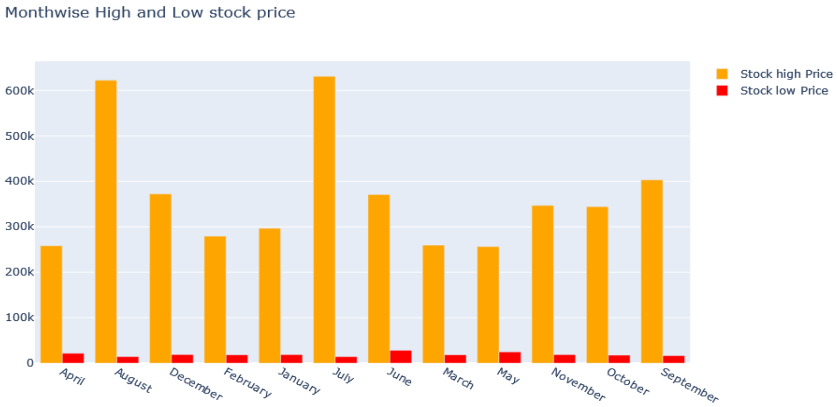


Fig.10. Month wise high and low bar chart

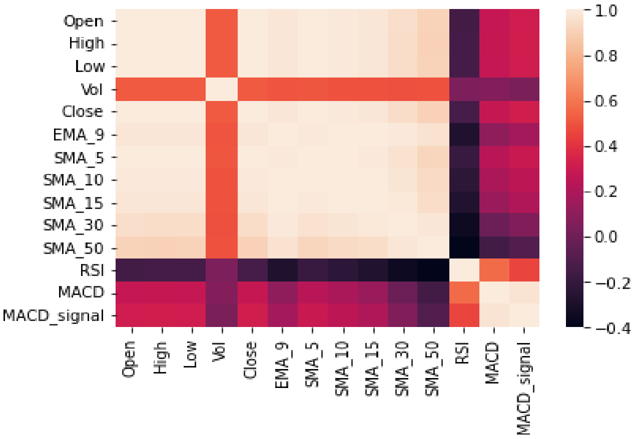


Fig.11. Heatmap

4. Methods

In this section, we'll provide a quick overview of the techniques used to generate the experiments' predictions. As previously noted, we apply the independently learned tree learning methodologies Linear Regression, Support Vector Regression, and Random Forest in this article. These techniques were chosen because they have a solid track record of success in the real world. Because of this, they are often used in literature to solve regression problems and make predictions about the stock market[31]. In this paper, we have used python in Jupyter Notebook and Numpy, Pandas Mathplotlib, Ploty and Sckit learn for implementation. Fig.12 depicts the supervised learning algorithms we use in this work and fig.13 depict s comparison ep by step process.

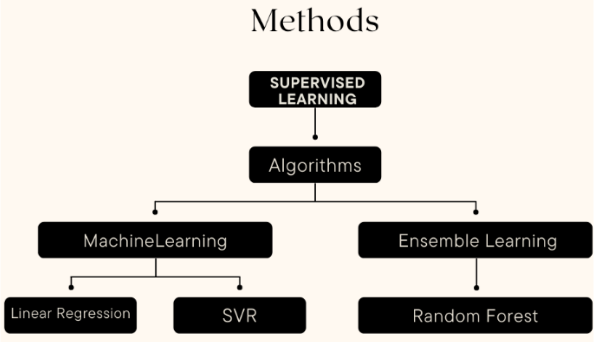


Fig.12. Methods

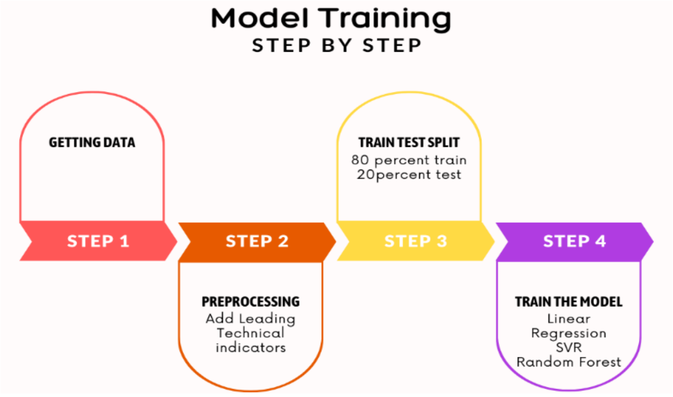


Fig.13. Model training step by step

4.1. Linear Regression

A statistical method known as linear regression analysis is used to forecast time series data as well as to model the connection between two variables. It predicts the value of one variable based on the value of another. This method's fundamental premise is to use a linear equation, $Y = a + b X$, to determine the connection between two variables. This equation represents the relationship between the independent variable (X) and the dependent variable (Y), or the variable that has to be forecasted[4, 32] In linear regression, linear predictor functions are used to model relationships, with the model's unknown parameters being estimated from the data. They are known as linear models. Linear regression is a supervised algorithm, supervised algorithms learn to predict a specific value based on historical data Model fitting is a crucial factor to take into account while choosing the model for the study. The explained variance of a linear regression model (usually represented as R2) will always rise when independent variables are added to the model. However, overfitting may happen when the model is given too many variables, which limits its potential to be generalized[33].

Table 2. Metrics

Test set evaluation:	Train set evaluation:		Actual	Predicted
MAE: 1659.9098334007872 **MSE:** 4072235.431345109 **RMSE:** 2017.9780552189136 **R2 Square** 0.9894212712805945	**MAE**: 1264.8380984724013 **MSE**: 5376583.286553996 **RMSE**: 2318.746059091852 **R2 Square** 0.999712710784184	**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

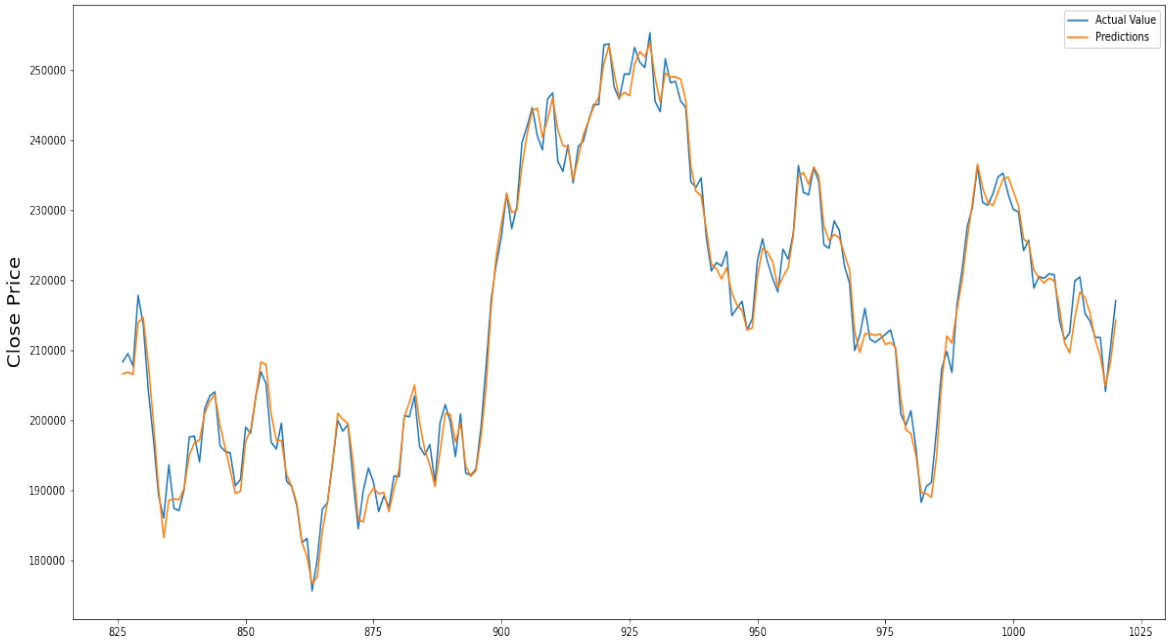


Fig.14. Predicted by LR

4.2. SVR

The SVM regression method is called "Support Vector Regression," or "SVR." Support vector regression is a type of supervised learning that is used to predict discrete values. The theory behind SVMs and support vector regression is the same. The most important part of SVR is finding the best fit line. In SVR, the best-fitting line is the one that has the most points. SVR is a type of support vector machine (SVM) that is used for regressions. There are some small changes from SVM. SVR is made up of sparse solutions and the use of kernels[32]. The SVR is different from other regression models because it doesn't try to minimize the difference between the actual and predicted values. Instead, it tries to find the best line that fits within a certain value. The difference between the hyperplane and the boundary line is the threshold value[34].

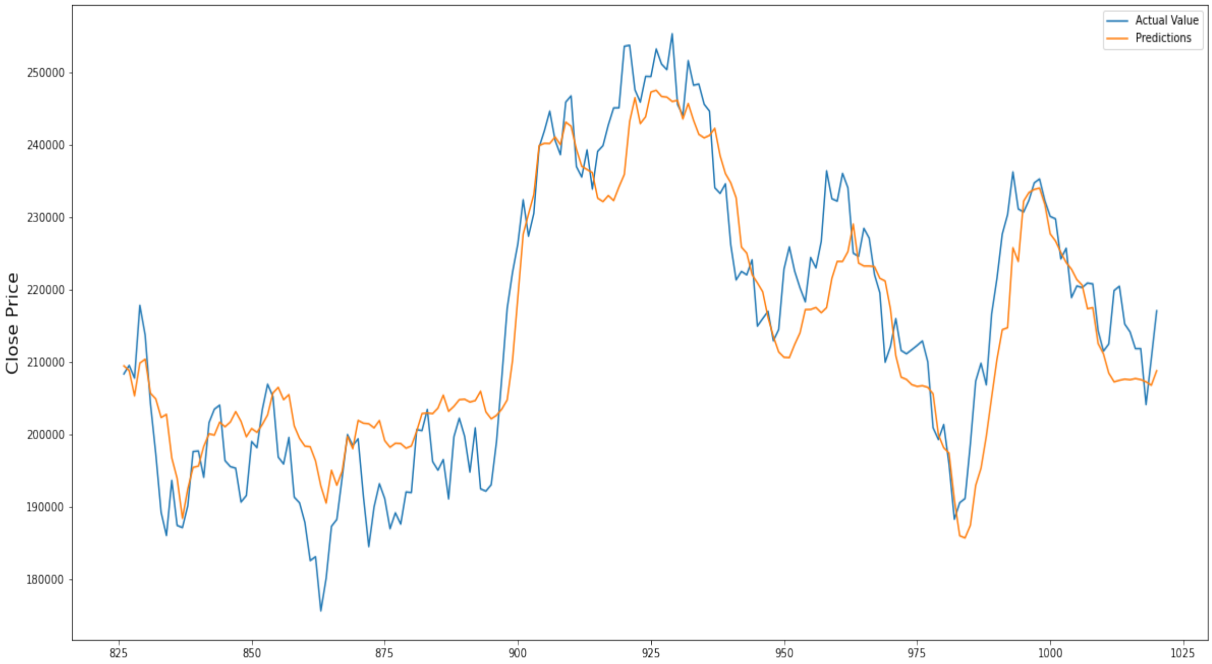


Fig.15. Predicted by SVR

Table 3. SVR Metrics

Test set evaluation:	Train set evaluation:		**Actual**	**Predicted**
MAE**: 1798.754797073944** MSE**: 5288867.02899446** RMSE**: 2299.7536887663555** R2 Square **0.9853249494021616**	**MAE**: 880.423621923374 **MSE**: 3378500.2059487947 **RMSE**: 1838.069695617877 **R2 Square** 0.999817975654345	**826**	208425.0	188807.609286
		827	209596.0	188734.997432
		828	207842.0	189700.017268
		829	217901.0	195014.196277
		830	213772.0	194849.089003

4.3. Random Forest

An ensemble technique called Random Forests (RF) can be applied to both classification and regression problems. The final prediction is created by adding the predictions of each decision tree in the RF model for making predictions. For example, the final prediction for regression tasks is the average of all the predictions made by the trees. By selecting a random selection of features from the training data using a bootstrap sample, the trees are trained independently [35]. came up with the idea of RF, but [36] made it better by combining the random subset method with its bagging method.

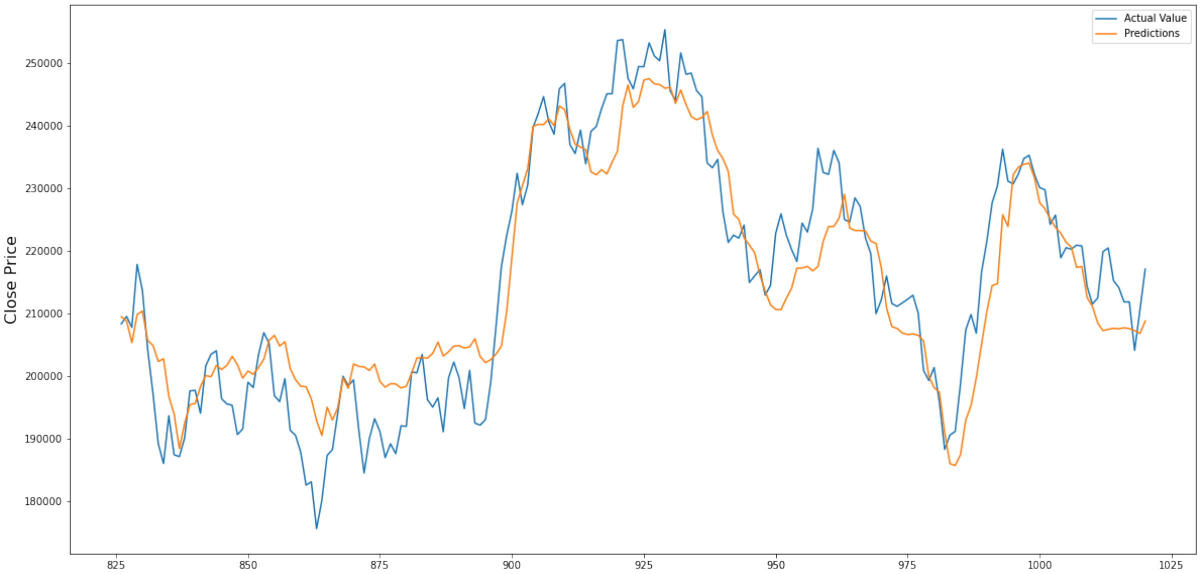


Fig.16. Predicted by random forest

Table 4. Random forest metrics

Test set evaluation:	Train set evaluation:		Actual	**Predicted**
MAE: 5897.585733821735 **MSE:** 54096382.1913612 **RMSE:** 7355.024282173458 **R2 Square** 0.8594700720152961	**MAE**: 1196.6510924153147 **MSE**: 4630266.984437989 **RMSE**: 2151.8055173360785 **R2 Square** 0.9997525890142344	**826**	**208425.0**	**209536.427**
		827	**209596.0**	**208870.294**
		828	**207842.0**	**205393.136**
		829	**217901.0**	**209913.923**
		830	**213772.0**	**210457.397**

5. Performance Measure

5.1. Mean Absolute Error (MAE)

The degree of mistake in your measurements is known as absolute error. That which separates the measured value from the "actual" value is the difference. The average of all absolute errors is known as the mean absolute error (MAE). The equation is:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|}$$

n = the number of errors,
Σ = summation symbol (which means “add them all up”),
|xi – x| = the absolute errors.

5.2. Mean Squared Error (MSE)

An estimator calculates the average squared error, or the difference in value between the estimated and actual values. describes the proximity of a regression line to a set of points. This is accomplished by squaring the distances between the points and the regression line (also known as the "errors"). The squaring is required to eliminate any unfavorable indications. Additionally, it emphasizes bigger discrepancies.

$$\text{MSE} = \frac{1}{n}\sum_{i-1}^{n}(Y_i - \hat{Y})^2$$

MSE = Mean Squared Error
n = number of data point
Y_i = observed values
$\hat{Y}_i$ = predicted values

5.3. Root Mean Squared Error

One of the methods most often used to assess the accuracy of forecasts is root means square error, also known as root mean square deviation. It illustrates the Euclidean distance between measured true values and forecasts. The residuals' standard deviation is (RMSE) (prediction errors). Data points' distance from the regression line is measured by residuals, which are The RMSE represents the degree of dispersion of these residuals. Otherwise put.

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(Predicted_i - Actual_i)^2}{N}}$$

5.4. R2 Square

One of the most important metrics for evaluating a continuous target regression model is the r2 score, varies between 0 and 100%. It is closely related to the MSE, the proportion of the variance in the dependent variable that is predictable from the independent variables. Another definition is “(total variance explained by model) / total variance.” So, if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases. In scientific studies, the R-squared may need to be above 0.95 for a regression model to be considered reliable.

$$\text{R2} = 1 - \frac{\text{Unexplained Variation}}{\text{total Variation}}$$

5.5. Experimental Design and Process

The overall experimental process followed these steps:

Data Splitting: The historical automobile stock price dataset was split into 80% train set and 20% test set. No separate validation set was used. The train and test sets were shuffled randomly to avoid ordering effects.

Model Implementation: The Linear Regression, Support Vector Regression, and Random Forest models were implemented in Python using the scikit-learn library. Default parameters were used for the models without any specific tuning or optimization.

Hardware: The experiments were run locally on a personal laptop with Intel i5 CPU and 16GB RAM without leveraging GPUs.

Evaluation: Each model was trained on the train set and iteratively evaluated on the test set. Performance metrics like MAE, MSE, RMSE and R2 Score were recorded. No statistical significance testing was done to compare the models.

Process Flow: The overall process flow was:

- Import and pre-process data
- Split into train/test sets
- Train each model on train set
- Evaluate on test set
- Record performance metrics
- Compare model results

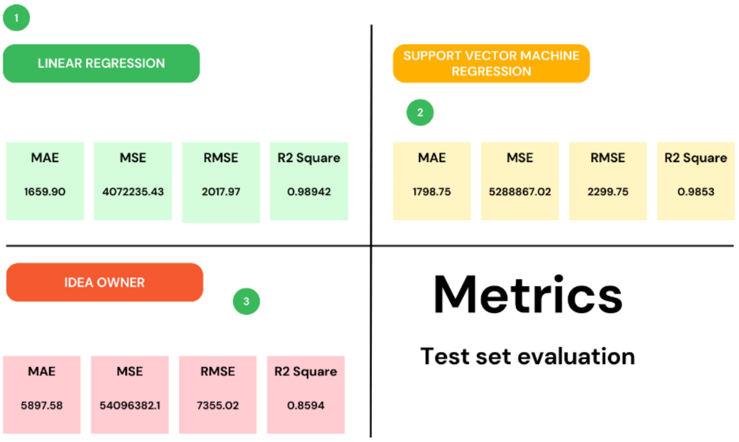
6. Empirical Results

In this experiment, the car index of the Tehran stock exchange is used as the input data. In order to increase the accuracy of the prediction[10]. the data also uses leading technical indicators as complementary features, this experiment was carried out under the Windows system of Jupyter Notebook by using the Python language As can be seen from Fig. 15 it is evident that the LR and SVR are better than other algorithms, so the SVR and LR are thought to be effective. MAE calculates the observations' absolute distance. (The entries of the dataset) from the predictions on a regression, taking the average over all observations. MAE tells us how on average, how large of a forecast error can we anticipate Both the MAE and RMSE can change. from zero to. They have scores that lean negatively, with lower values being preferable. The best MAE score is for Linear Regression with a score of 1659, A regression line's proximity to a set of points is indicated by MSE. To accomplish this, it squares the distances between the points and the regression line (these distances are the "errors"). To get rid of any warnings, there must be a square. For MSE, there is no ideal value. In other words, the lower the value, the better, and 0 denotes a perfect model. Since there is no right or wrong response, the MSE's primary benefit is in helping us choose one prediction model over another. There is no optimal MSE value. In other words, the lower the score, the better, and a value of 0 indicates that the model is flawless. Since there is no right or incorrect response, the MSE is most helpful for determining which prediction model to apply since there is no correct answer.

Further examination and discussion of the results

The results show that the Linear Regression (LR) model achieved the lowest error scores overall, with a test set MAE of 1659, MSE of 4072235, RMSE of 2017, and highest R2 of 0.9894. The LR model could most accurately capture the patterns in the data to make stock price predictions. The simplicity of a linear model seems well-suited to this problem and dataset. The Support Vector Regression (SVR) model attained the second-best performance with slightly higher errors than LR, which is still competitive. The test set errors were MAE of 1798, MSE of 5288867, RMSE of 2299, and R2 of 0.9853. The ability of SVR to handle nonlinear relationships using kernels appears beneficial for modeling stock prices.

Comparatively, the Random Forest (RF) model performed significantly worse than LR and SVR. The test errors for RF were relatively high, with MAE of 5897, MSE of 54096382, RMSE of 7355, and low R2 of 0.8594. The RF model averaging multiple decision trees could have provided better predictive accuracy. The randomness and nonlinearity captured by the trees may not fit the patterns in this data well. The models' relative performance aligns with expectations based on their underlying algorithms. The linear assumptions of LR make it best suited for modeling stock price trends. At the same time, SVR's kernels add flexibility for nonlinear relationships. However, RF's decision trees are likely to overfit this problem. The metrics indicate that LR and SVR were able to learn the input-target mappings for stock price forecasting reliably. At the same time, RF had overfitting issues leading to poor generalization. Given their accuracy on this emerging market dataset, the results are promising for using LR and SVR for financial forecasting problems. However, tuning and ensembles of models could improve RF predictions in future work. The results demonstrate superior predictive performance for LR and SVR compared to RF on this novel automobile stock price dataset. The analysis provides insights into the strengths and weaknesses of the different modeling approaches for this problem context. Further hyperparameter tuning and investigation of additional input features may further enhance the model accuracies in future experiments.



LINEAR REGRESSION

MAE	MSE	RMSE	R2 Square
1659.90	4072235.43	2017.97	0.98942

SUPPORT VECTOR MACHINE REGRESSION

MAE	MSE	RMSE	R2 Square
1798.75	5288867.02	2299.75	0.9853

IDEA OWNER

MAE	MSE	RMSE	R2 Square
5897.58	54096382.1	7355.02	0.8594

Metrics

Test set evaluation

Fig.17. Test set metrics

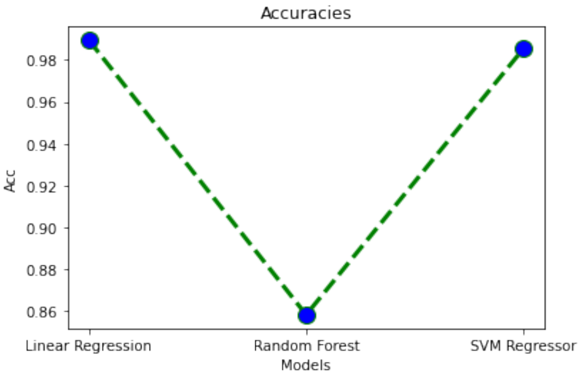


Fig.18. Model accuracies

7. Conclusions and Future Direction

This study evaluated the performance of three popular machine learning models - Linear Regression, Support Vector Regression, and Random Forest - in predicting daily stock prices of the automobile index on the Tehran Stock Exchange. The historical daily data from 2018-2022 was pre-processed and split into train and test sets. The models were trained on the train set and evaluated on the test set using MAE, MSE, RMSE, and R2 metrics.

The results demonstrate that overall, the linear regression and support vector regression models achieved lower error scores than the random forest model in predicting automobile stock prices. Linear regression obtained the lowest MAE, MSE, and RMSE on the train and test sets. Support vector regression also attained competitive performance, with the second-lowest error scores after linear regression. In contrast, random forest displayed significantly higher errors.

These findings suggest that for this dataset, the linear regression and support vector regression techniques were better able to capture the patterns and relationships between the input features and stock prices. The simplicity and linear assumptions of linear regression proved effective for the stock price modelling task. Support vector regression's capability to handle nonlinear relationships using kernels was also helpful. On the other hand, random forest averaging of multiple decision trees provided a different level of predictive accuracy.

The promising linear and support vector regression performance demonstrates that machine learning techniques can reliably forecast stock market prices, even for new datasets from emerging markets like the Tehran Stock Exchange. This study provides empirical evidence and quantitative benchmarks on an untapped dataset in the automobile industry.

Additional machine learning and deep learning architectures like neural networks and extended short-term memory networks can be evaluated for future work. More input features derived from candlestick charts, sentiment analysis, and macroeconomic data may improve predictions further. Testing combinations of models in ensemble approaches could also be beneficial. Overall, this pioneering research is a foundation for further analysis of machine learning for financial forecasting in Iran's stock markets.

To establish benchmark machine learning predictive performance on the automobile index:

- The study implements three standard machine learning algorithms - Linear Regression, Support Vector Regression, and Random Forest.
- These models are trained and tested on the historical daily stock price data of the Tehran Stock Exchange's automobile index.
- Performance metrics like MAE, MSE, RMSE, and R2 are calculated on both train and test sets.
- This provides a quantitative benchmark of predictive accuracy using different machine learning techniques on the new automobile index dataset.

To identify the strengths and limitations of each modelling approach:

- The results showed linear regression and support vector regression achieved lower errors compared to random forest.
- This highlights the effectiveness of linear regression's simplicity and SVR's ability to capture nonlinear patterns.
- Random forest's error rates were higher, indicating its averaging of decision trees was less suited.
- The metrics and analysis identify the comparative strengths and weaknesses of the models.
- To determine the best-suited technique for financial time series forecasting on emerging markets:
- The Tehran Stock Exchange represents an emerging market with unique properties.
- The automobile index captures a vital industry sector but has not been extensively modeled before.
- The empirical results suggest linear regression and SVR are better suited than random forests for this dataset.
- This provides data-driven guidance on which techniques generalize better to new financial time series data from emerging economies.

By implementing a set of standard machines learning models, systematically evaluating their performance, and comparing their strengths and limitations, the study fulfils its goals of benchmarking predictive accuracy, assessing modelling approaches, and identifying effective techniques for the problem context. The methods are aligned with and help achieve the stated research objectives.

This study makes several key contributions to advance the state of knowledge in applying machine learning for stock market forecasting, particularly in emerging economies like Iran.

Firstly, it establishes the first benchmarks for predictive accuracy on the previously unmodeled automobile index dataset from the Tehran Stock Exchange. There have been limited applications of machine learning models on Iranian stock data before. By training and rigorously evaluating standard algorithms like Linear Regression, SVR, and Random Forest, this work provides a quantification of model performance as a baseline for future experiments.

Secondly, the comparative analysis of the strengths and weaknesses of the models gives data-driven insights into which techniques are better suited for financial time series forecasting in this problem context. The results suggest that simpler linear models and nonlinear kernels are effective on this data, while ensemble tree-based methods face limitations. This guides appropriate model selection for related datasets.

Thirdly, the promising accuracy of the linear and kernel models demonstrates the feasibility of using machine learning to reliably predict stock prices on emerging market data. Despite different properties compared to developed markets, the algorithms are able to capture relevant patterns in the Iranian index. This justifies further research into machine learning for stock forecasting in such economies.

In addition, this work identifies several directions to explore to further improve predictive performance, including evaluating neural networks, adding more input features, and combining models. As one of the first studies on the Tehran automobile index, it lays the foundation for expanding machine learning applications in Iranian stock markets.

By benchmarking model performance on a new dataset, providing data-driven comparisons, demonstrating promising accuracy, and outlining future work, this study makes significant contributions to advancing the application of machine learning techniques for stock price forecasting, especially on emerging market data. It provides both quantitative results and scientific insights to progress the field.

Author Contributions

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

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

publicly available datasets were analyzed in this study these data can be found here: http://en.tsetmc.com/Site.aspx, Tehran Securities Exchange Technology Management Co. – TSETMC

Conflicts of Interest

The authors declare no conflict of interest.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

Acknowledgements

The author would like to thank the anonymous reviewers and respected editors for taking valuable time to go through the manuscript.

References

[1] Tehran Stock Exchange. (2020). Tehran Stock Exchange Statistics.; Available from: https://www.tse.ir/en/home.html.

[2] Moghaddam, A.H., M.H. Moghaddam, and M. Esfandyari, Stock market index prediction using artificial neural network. Journal of Economics, Finance and Administrative Science, 2016. 21(41): p. 89-93.

[3] Ozdemir, O., A. Aslanargun, and S. Asma, ANN forecasting models for ISE national-100 index. Journal of Modern Applied Statistical Methods, 2010. 9: p. 579-583.

[4] Ayala, J., et al., technical analysis strategy optimization using a machine learning approach in stock market indices. Knowledge-Based Systems, 2021. 225: p. 107119.

[5] Pai, P.-F. and C.-S. Lin, A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 2005. 33(6): p. 497-505.

[6] Wei, L.-Y., A hybrid model based on ANFIS and adaptive expectation genetic algorithm to forecast TAIEX. Economic Modelling, 2013. 33: p. 893-899.

[7] Arash Salehpour, E.S., A Regression Analysis on the Car Index in the Tehran Stock Exchange. Journal of Soft Computing Paradigm, 2022. 4(4): p. 238-251.

[8] 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.

[9] Kanwal, A., et al., BiCuDNNLSTM-1dCNN â€” A hybrid deep learning-based predictive model for stock price prediction. Expert Systems with Applications, 2022. 202: p. 117123.

[10] Wu, J.M.-T., et al., A Tool based on ML-driven Graphical Model for Stock Price Prediction by Leading Indicators. IFAC-PapersOnLine, 2020. 53(5): p. 692-697.

[11] Abu-Mostafa, Y.S. and A.F. Atiya, Introduction to financial forecasting. Applied Intelligence, 1996. 6(3): p. 205-213.

[12] Blank, S.C., " Chaos" in futures markets? A nonlinear dynamical analysis. The Journal of Futures Markets (1986-1998), 1991. 11(6): p. 711.

[13] Huang, W., Y. Nakamori, and S.-Y. Wang, Forecasting stock market movement direction with support vector machine. Computers & operations research, 2005. 32(10): p. 2513-2522.

[14] Moody, J. and M. Saffell, learning to trade via direct reinforcement. IEEE transactions on neural Networks, 2001. 12(4): p. 875-889.

[15] Hsu, M.-W., et al., Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert systems with Applications, 2016. 61: p. 215-234.

[16] Long, W., Z. Lu, and L. Cui, Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 2019. 164: p. 163-173.

[17] Hochreiter, S. and J.r. Schmidhuber, long short-term memory. Neural computation, 1997. 9(8): p. 1735-1780.

[18] Shahrestani, P. and M. Rafei, the impact of oil price shocks on Tehran Stock Exchange returns: Application of the Markov switching vector autoregressive models. Resources Policy, 2020. 65: p. 101579.

[19] Ebrahimi, B. and E. Hajizadeh, A novel DEA model for solving performance measurement problems with flexible measures: An application to Tehran Stock Exchange. Measurement, 2021. 179: p. 109444.

[20] Ramezanian, R., A. Peymanfar, and S.B. Ebrahimi, An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market. Applied Soft Computing, 2019. 82: p. 105551.

[21] Chizari, M.h., et al., The impact of Intellectual Capitals of Pharmaceutical Companies Listed in Tehran Stock Exchange on their Market Performance. Procedia Economics and Finance, 2016. 36: p. 291-300.

[22] Nejad, M.K., F. Jahantigh, and H. Rahbari, The Long Run Relationship between Oil Price Risk and Tehran Stock Exchange Returns in Presence of Structural Breaks. Procedia Economics and Finance, 2016. 36: p. 201-209.

[23] Abounoori, E., Z. Elmi, and Y. Nademi, Forecasting Tehran stock exchange volatility; Markov switching GARCH approach. Physica A: Statistical Mechanics and its Applications, 2016. 445: p. 264-282.

[24] Jalilvand, A., M.R. Noroozabad, and J. Switzer, Informed and uninformed investors in Iran: Evidence from the Tehran Stock Exchange. Journal of Economics and Business, 2018. 95: p. 47-58.

[25] Mothlagh, S.S., F. Samadi, and Z. Hajiha, The Relationship of the Content of the Market Value in the Explanation of Abnormal Stock Returns of Listed Companies in Tehran Stock Exchange. Procedia Economics and Finance, 2016. 36: p. 113-122.

[26] Ebrahimpour, R., et al., Mixture of MLP-experts for trend forecasting of time series: A case study of the Tehran stock exchange. International Journal of Forecasting, 2011. 27(3): p. 804-816.

[27] Jahan-Parvar, M.R. and H. Mohammadi, Risk and return in the Tehran stock exchange. The Quarterly Review of Economics and Finance, 2013. 53(3): p. 238-256.

[28] Fernando, J. Moving Average (MA): Purpose, Uses, Formula, and Examples. 2022; Available from: https://www.investopedia.com/terms/m/movingaverage.asp.

[29] Pattewar, T. and D. Jain. Impact of Covid'19 on Indian Stock Prediction by Technical Indicators. in 2022 International Conference for Advancement in Technology (ICONAT). 2022.

[30] Gamboa Valero, H., Big Data finance: trading strategy creation using Deep Reinforcement Learning in University of Manchester 2021 University of Manchester University of Manchester

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

[32] Neter, J., et al., Applied linear statistical models. 1996.

[33] Seal, H.L., Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. Biometrika, 1967. 54(1-2): p. 1-24.

[34] Smola, A.J. and B. SchÃ¶lkopf, A tutorial on support vector regression. Statistics and Computing, 2004. 14(3): p. 199-222.

[35] Ho, T.K. Random decision forests. in Proceedings of 3rd international conference on document analysis and recognition. 1995. IEEE.

[36] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

Authors' Profiles

Arash Salehpour holds a Master's in Computer Science Software Engineering from Islamic Azad University, Rasht Branch, Iran. He has experience in computer science, programming, and deep learning. He has research with an emphasis on finance and economics. His primary research directions are bibliometric analysis, portfolio construction, valuation, investor behaviour, stock price prediction, and big data analytics in economy and finance.

How to cite this paper: Arash Salehpour, "Predicting Automobile Stock Prices Index in the Tehran Stock Exchange Using Machine Learning Models", International Journal of Intelligent Systems and Applications(IJISA), Vol.15, No.5, pp.12-27, 2023. DOI:10.5815/ijisa.2023.05.02