Tesla Share Price Prediction Since It Got Listed: A Comparitive Study And Multiple Classifiers

Palak Shahu
Electronics and Computer Science, RBU
shahupr_2@rknec.edu

Krish Sharma
Electronics and Computer Science, RBU
sharmak3@rknec.edu

***Abstract*- This project aims to predict stock price movement using both Support Vector Machines (SVM) and Linear Regression models. Our motivation lies in the need to develop reliable, automated trading strategies leveraging historical stock data. We processed datasets containing key financial indicators and built predictive models using SVM for classification and Linear Regression for understanding directional trends. Results demonstrate the models' effectiveness, supported by metrics like accuracy, precision, and cumulative return plots.**

2. Introduction

The stock market's volatility, driven by factors like economic shifts and investor sentiment, makes accurate price prediction challenging yet crucial for informed trading, risk management, and optimizing returns. Traditional methods often rely on subjective judgment, limiting their effectiveness.

Predicting stock market trends is critical for optimizing trading strategies and managing financial risk. This report investigates two machine learning models—Support Vector Machines (SVM) and Linear Regression—to predict stock price movements using historical data from Tesla and Apple.

Key Inputs:

- **Daily High, Low, Open, Close Prices**
- **Engineered Features**:
  - **Open-Close**: Difference between opening and closing prices
  - **High-Low**: Difference between high and low prices

Model Outputs:

- **SVM**: Binary prediction (price goes up or down)
- **Linear Regression**: Continuous stock price, converted to directional prediction

Motivation: We aim to improve financial decisions by using machine learning to find patterns in stock data, enabling systematic trading strategies that minimize subjective judgment. This offers both financial gains and a technical challenge.

Impact: Accurate predictions could transform investment strategies by offering data-driven insights, improving financial decision-making efficiency.

3. Related work

1. **Technical Analysis**: Silva et al. (2017) used indicators like moving averages and RSI with logistic regression to predict price movement. ***Pros***: Simple, interpretable. ***Cons***: Relies solely on historical data, ignoring market sentiment.
2. **Ensemble Learning**: Alwazani et al. (2021) applied Random Forest and Gradient Boosting to enhance prediction accuracy. ***Pros***: Reduces overfitting, captures non-linear patterns. ***Cons***: Complex, less interpretable.
3. **Deep Learning**: Choudhury et al. (2019) used LSTM networks to capture time-dependent stock trends. ***Pros***: Suited for time series, captures long-term dependencies. ***Cons***: Computationally heavy, prone to overfitting.
4. **Hybrid Models**: Abdallah (2020) combined machine learning with sentiment analysis from news. ***Pros***: Considers external factors, comprehensive. ***Cons***: Increased complexity, data-heavy.
5. **ARIMA**: Kim (2019) used ARIMA for time series stock forecasting. ***Pros***: Statistically robust for linear trends. ***Cons***: Limited by linearity and stationarity assumptions.

Summary: State-of-the-art approaches like deep learning and hybrids excel in accuracy but are complex. Simpler models, often used in practice, balance interpretability with predictive power.

4. Dataset and Features

Tesla Dataset

- **Source**: Custom CSV file containing historical stock prices of Tesla sourced from

Kaggle.
- The dataset contains 2,967 entries.
- **Features:** columns for **Date**, **Open**, **High**, **Low**, **Close**, **Adjusted Close**, and **Volume**.

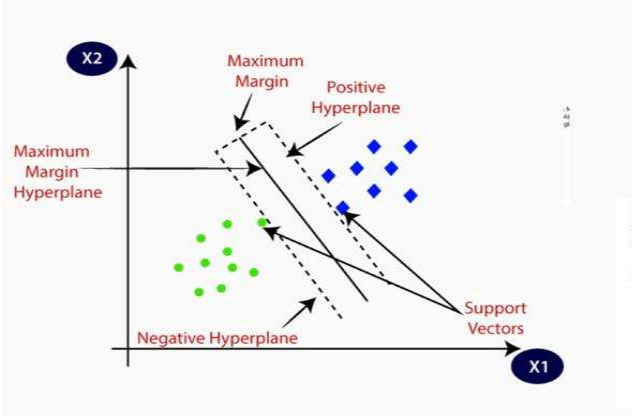
- **Feature Engineering:**
  - Open-Close: Difference between opening and closing prices.
  - High-Low: Difference between the highest and lowest prices in a day.
- **Target Variable**: Binary indicator (1 if the next day's closing price is higher, 0 otherwise).
- **Preprocessing**: Dropped date columns, set the index to DateTime format, and created new features.
- **Time-Series Discretization**: The dataset is structured by trading days, with each row representing one day of Tesla's trading data. This is already discretized by daily intervals.
- **Citation : SJ. *Tesla Share Price Since It Listed*. Kaggle, CC0: Public Domain, Updated 3 years ago,** https://www.kaggle.com/datasets/surajjha101/tesla-share-price-for-last-5-years .**Accessed [2/11/2024].**
- **Example Data Points** Here are some example records from the dataset:

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	29-06-2010	3.8	5	3.508	4.778	4.778	93831500
3	30-06-2010	5.158	6.084	4.66	4.766	4.766	85935500
4	01-07-2010	5	5.184	4.054	4.392	4.392	41094000

5. Methods

1. Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification tasks. It aims to find the optimal hyperplane that separates data points of different classes in a high-dimensional space. The optimization problem minimizes:



$$W.X_i + b \geq 1, \forall X_i, \text{with } y_i = 1$$

$$W.X_i + b \leq -1, \forall X_i, \text{with } y_i = -1$$

$$\text{as, } y_i\,(W.X_i + b) \geq 1, \forall X_i$$

So the problem becomes:-

$$\text{Maximize } \frac{2}{||W||}, \text{ s.t. } y_i\,(W.X_i + b) \geq 1, \forall X_i$$

$$\text{or, Minimize } \frac{1}{2}||W||^2, \text{ s.t. } y_i\,(W.X_i + b) \geq 1, \forall X_i$$

- So, now the problem is to find W, b that solves

$$\text{Minimize } \frac{1}{2}||W||^2, \text{ s.t. } y_i\,(W.X_i + b) \geq 1, \forall X_i$$

Hyperparameters: We used the default settings of sklearn's SVC model and evaluated the results on our test set.

Classification and Regression: Effective for both tasks. **Margin Maximization**: Separates classes by maximizing margin. **Support Vectors**: Uses closest points to define boundary. **Kernels**: Supports linear and non-linear (e.g., RBF) kernels. **High-Dimensional Performance**: Handles many features well. **Regularization (C)**: Balances margin size and error minimization. **Binary and Multi-Class**: Adapts for multiple classes. **Overfitting Resistant**: Robust with clear class separation.

2. Linear Regression

It models the relationship between a dependent variable Y and one or more independent variables X by fitting a linear equation to the observed data.

$$Y = a + bX$$

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \qquad a = \frac{\sum Y - b\sum X}{N}$$

- **Sigmoid Function**: Uses sigmoid to output values between 0 and 1, representing probabilities.
- **Linear Decision Boundary**: Creates a linear boundary between classes.
- **Interpretable Coefficients**: Coefficients show feature impact on the prediction.
- **Odds and Log-Odds**: Models the log-odds of the probability of class membership.
- **Regularization**: Can use L1 or L2 regularization to prevent overfitting.
- **Multi-Class Extension**: Adaptable to multi-class (softmax or one-vs-rest).

- **Fast and Efficient**: Low computational cost, good for large datasets.

6. Experiments/Results/Discussion

The experiments in this study aimed to evaluate the performance of different machine learning models—Linear Regression,Random Forest, and Support Vector Machine—Binary prediction (price goes up or down) and Continuous stock price, converted to directional prediction.

Support Vector Machine

Hyperparameter -Kernel: Experiment with nonlinear kernels like RBF or polynomial for complex patterns.**Regularization:** Fine-tune the C parameter using grid search or randomized search.**Cross-Validation:** Implemented k-fold cross-validation for robust model 5 folds evaluation. **Accuracy**: Calculated on the test set. We observed a low level of moderate accuracy 0.48 performance.

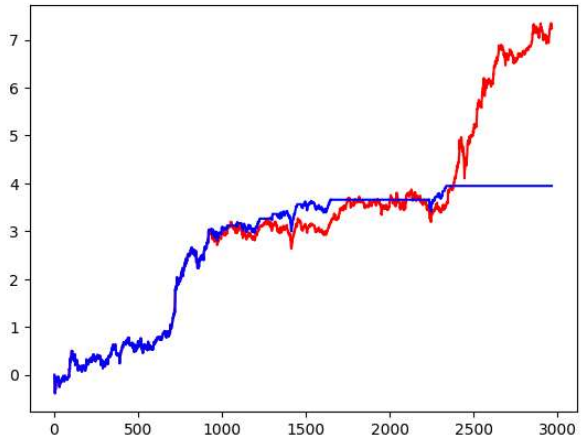


Fig .Cumulative returns and Cumulative Strategy

Random Forest Classifier

Hyperparameter Tuning: The Random Forest model was optimized with the following hyperparameters:

- **Maximum Depth:** 10
- **Minimum Samples Split:** 5
- **Number of Estimators:** 200

Overall Performance:

The model achieved an overall accuracy of **45.33%**. While this accuracy is moderate, it's essential to consider the class imbalance and the nature of the classification problem.

Class-wise Performance:

- **Class 0:**
  - Precision: 45% (Of the predicted positive cases, 45% were actually positive)
  - Recall: 26% (Of the actual positive cases, 26% were correctly identified)
  - F1-Score: 33% (Harmonic mean of precision and recall)
- **Class 1:**
  - Precision: 45%
  - Recall: 67%
  - F1-Score: 54%

ROC-AUC Score:

The ROC-AUC score of **0.4898** indicates the model's ability to distinguish between positive and negative classes. A higher ROC-AUC score would suggest better discriminatory power.

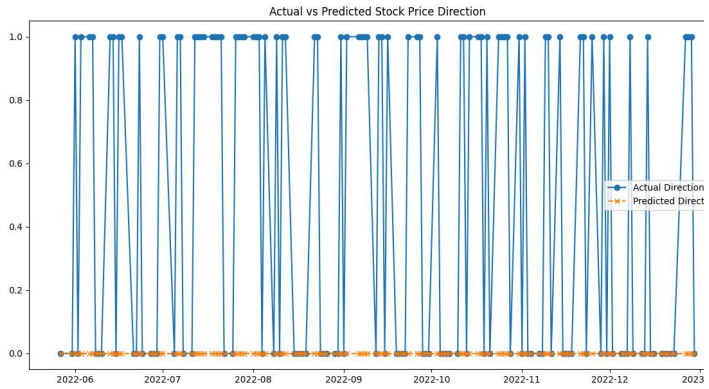


Fig.Actual vs Predicted prices direction

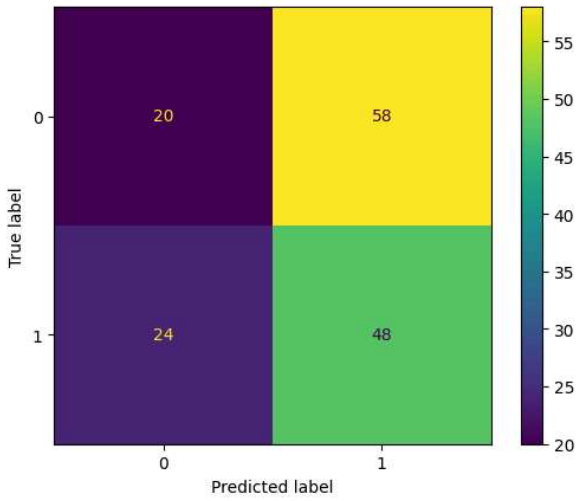


Fig . Confusion Matrix of Random Forest

Linear Regression

The linear regression model, trained on the specified features ('Previous_Close' and 'MA_5'), demonstrated a directional prediction accuracy of 52%. This indicates that the model correctly predicted the direction of the market movement (upward or downward) 52% of the time.

Previous_Close : 0.0019

'MA_5' : -0.0019

'MA_10' : -0.0002

However, the model's precision is 1.00 , recall is 0 , and F1-score is 0.6842. struggles with identifying positive instances, as evidenced by the low precision, recall, and F1-score. This suggests that the model is overly conservative, potentially due to a high classification threshold or an imbalanced dataset. To improve performance, consider adjusting the threshold, balancing the dataset, exploring additional features, tuning hyperparameters, or trying different models.

Evaluation Metrics:

- **Directional Prediction Accuracy:** 0.52
- **Precision:** 1.00
- **Recall:** 0.00
- **F1-Score:** 0.6842

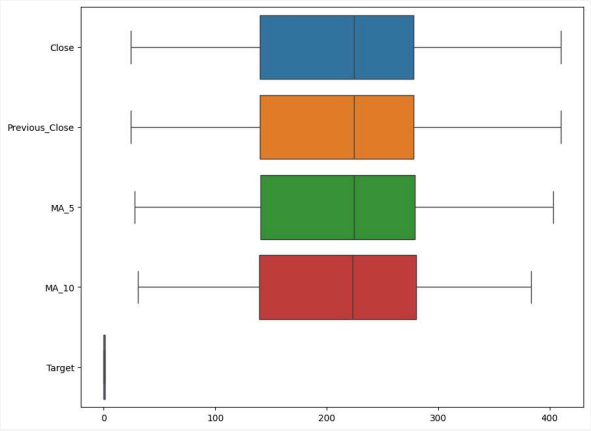


Fig : Boxplot for all features

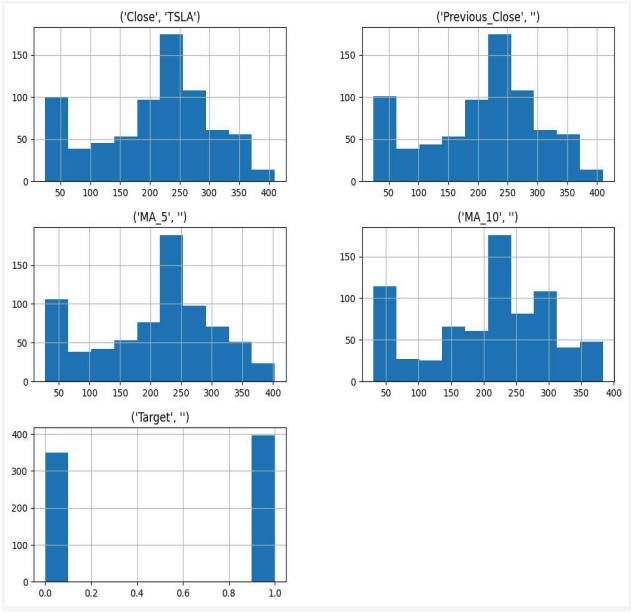


Fig : Histogram for each numerical column

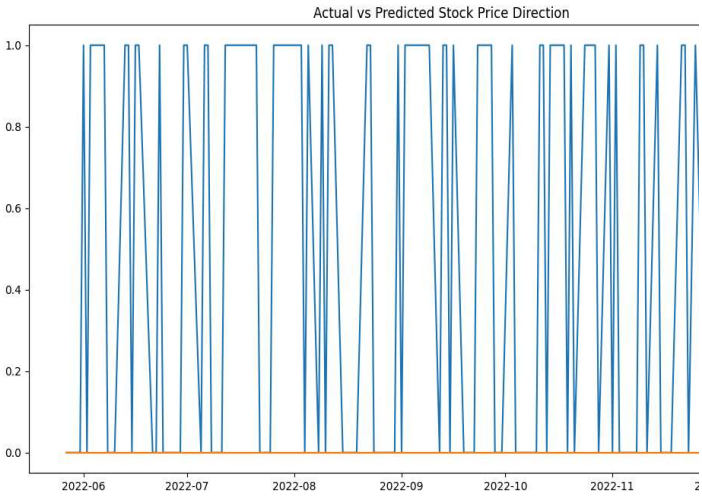


Fig. Actual vs Predicted Stock prices direction

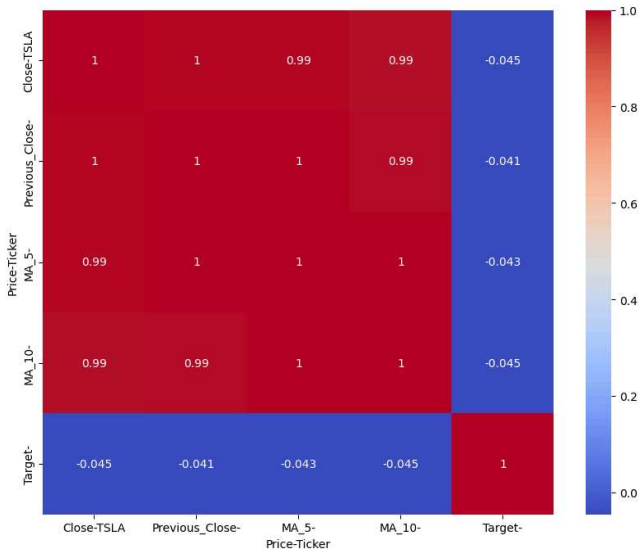


fig: correlation matrix plotted as heat map

7. Conclusion and Future Work

This study explored the application of machine learning techniques, specifically Support Vector Machines (SVM) , Linear Regression and Random Forest Classifier to predict stock price movements. While all models demonstrated moderate predictive power, they also revealed certain limitations.

The SVM model, though capable of handling complex patterns, showed room for improvement in terms of accuracy, potentially through more advanced kernel functions and hyperparameter tuning. The Linear Regression model, while achieving a decent directional prediction accuracy, struggled with identifying positive instances, highlighting the need for careful feature engineering and model selection.

Future Work: Data Enhancement and Feature Engineering:Expand Dataset: Incorporate data from a wider range of stocks, including those from different sectors and regions.Enrich Features: Explore additional features such as volatility indices, momentum indicators, and sentiment analysis from news and social media.Time Series Analysis: Utilize time series analysis techniques to capture seasonal and cyclical patterns in the data

8. Appendices

Appendix A : Data Preprocessing Steps
Handling Missing Values: Missing values were filled using the median or mode of each feature column to ensure the dataset remained complete for model training1) Standardization and Normalization:: Features were scaled to have a mean of zero and standard deviation of one,improving the convergence of gradient-based algorithms.
2) Resampling:: Techniques like undersampling or over sampling could be applied in future iterations to address class imbalance and improve recall on.

Appendix B: Evaluation Metrics
1) Precision:: Indicates the accuracy of positive predictions,
calculated as TP / (TP + FP).
2) Recall:: Measures the ability to identify all relevant instances, calculated as TP / (TP + FN).
3) F1-Score:: The harmonic mean of precision and recall,
balancing the trade-off between them.
4) Accuracy:: The proportion of correct predictions out of
all predictions, useful for overall performance but limited in
imbalanced datasets.

9. Contributions

A. Palak Shahu:
Led data preprocessing, including handling missing values,
Exploratory data analysis , and feature selection. Conducted the implementation of the Random Forest model and linear regression model. Contributed to writing the report sections for Dataset and Features, Methods, and Results.

B.Krish Sharma:
Focused on the SVM model implementation and hyperparameter tuning. Conducted extensive performance evaluations and prepared the classification reports and confusion matrices.Created pictoral representations.Authored the section of introduction and co- authored the sections on Related Work, Experiments/ Results/Discussion, and Conclusion/Future Work.

10. References

[1] Linear regression analysis study
January 2018Journal of the Practice of Cardiovascular Sciences 4(1):33
DOI:10.4103/jpcs.jpcs_8_18
License CC BY-NC-SA 4.0

[2]Estimating Classification Accuracy for Unlabeled Datasets Based on Block Scaling September 2023
International Journal of Engineering and Technology Innovation 13(4):313-327
DOI:10.46604/ijeti.2023.11975
LicenseCC BY-NC 4.0

[3]Support Vector Machines: Theory and ApplicationsSeptember 200
DOI:10.1007/3-540-44673-7_12
SourceDBLP
Conference: Machine Learning and Its Applications, Advanced Lectures