Prediction of Telsa stock using machine learning.

A Capstone Phase-II project report submitted in partial

fulfilment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

By

B. Nagaraju Kumar	(2203A51072)
S. Prathibha	(2203A51060)
M. Akhila	(2203A51119)

Under the guidance of

Dr. Arpita Baronia

Associate Professor, CS & AI.

Ananthasagar, Warangal.

CERTIFICATE

This is to certify that this project entitled “ Prediction of Telsa stock using machine learning." is the bonafide work carried out by B. Nagaraju Kumar, S. Prathibha, M. Akhila, as a Capstone Phase-II project for the partial fulfilment to award the degree BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING during the academic year 2023-2024 under our guidance and Supervision.

Dr. Arpita Baronia
Assoc. Prof. CS & AI
S R Engineering College,
Ananthasagar, Warangal.

Dr.M.Sheshikala
Assoc. Prof. & HOD(CSE),
S R Engineering College,
Ananthasagar,Warangal.

External Examiner

ABSTRACT:

Technological advancements have revolutionized stock market forecasting, with machine learning methods proving more accurate than traditional statistical approaches. Comparing various models, it was found that integrated algorithms like Random Forest exhibit superior performance in predicting Tesla's stock closing prices. These methods mitigate risks associated with investing while maximizing dividends, crucial for effective resource allocation and macroeconomic expansion in a country's financial market.

I. INTRODUCTION

Stock markets, consisting of stockbrokers and dealers, facilitate the trading of stock shares, increasing liquidity and investor appeal. However, stock market investments are inherently risky due to the potential for rapid fluctuations in stock prices, making forecasting stock prices challenging. Traditional statistical methods, being linear, often fail to predict sudden spikes or drops in stock prices accurately, proving insufficient for the volatile and unpredictable nature of stock data.

To address this, various approaches have been developed, including Linear Regression, Support Vector Machines (SVM), Random Forest, Long Short-Term Memory (LSTM), and Autoregressive Integrated Moving Average (ARIMA) models, to predict stock prices. These models are applied to Tesla stock data from June 29, 2010, to July 12, 2022, obtained from Kaggle. The data is processed to remove null values and normalize it for model compatibility.

The experiment aims to compare the performance of these models using metrics like Mean Absolute Error (MAE) or Root Mean Square Error (RMSE). The dataset is divided into training (June 29, 2010, to September 7, 2018) and testing (July 10, 2018, to July 12, 2022) sets, with 2121 data points for training and 909 for testing. The experiment seeks to assess model accuracy and provide updated recommendations for future research and project expansion in stock price prediction.

It is important to note that stock markets, such as the London Stock Exchange and New York Stock Exchange, cater to publicly listed and privately held companies' shares. Investments in mixed-ownership share trade involve publicly exchangeable ordinary shares on rare occasions.

II. LITERATURE SURVEY

The research discussed focuses on using various machine learning algorithms and statistical techniques to predict Tesla stock price. They emphasize the importance of legitimate business forecasts for investors and financial analysts. In these studies, linear regression, polynomial regression, XGBoost, ARIMA, Prophet, LSTM, etc. are used to analyze historical data and make predictions about future market prices. methods are used. Metrics used include mean square error (MSE), root mean square error (RMSE), mean error (MAE), and prediction accuracy. Some studies also include sentiment analysis on social media such as Twitter to understand public sentiment and its impact on stock prices. Overall, these studies contribute to the growing body of research on the use of machine learning and statistical methods to improve the accuracy of cost estimates.

. Problem Statement:

Tesla Inc. It is a well-known energy company and predicting its stock price accurately is beneficial for investors and traders. However, it is very difficult to estimate the price of the product due to the lack of lines and weakness. The aim of the project is to accurately predict Tesla's future price data and to develop and evaluate advanced learning models that include interactive learning such as RNN and LSTM, as well as traditional methods such as regression, tree-based and hybrid methods. Relevant data is based on historical data. Goals include generating preliminary data, model design, performance evaluation using metrics such as MSE, RMSE, MAPE, and R-squared, comparing different methods, and exploring improvements and future research directions. Performance can help determine the potential value of Tesla shares.

III. EXPERIMENTAL RESULTS

In this section, we explain the details of experimental results. In this section initially, we explain the details of datasets, then the results of our experiments.

1. Experimental setup

For this project, I have obtained my dataset from Kaggle. This dataset contains 2814 rows of data and 7 columns (features) that we could focus on to build our prediction model i.e., I have used 7 attributes to predict the rise in stock of Tesla.

Date	Open	High	Low	Close	Adj Close	Volume
########	3.8	5	3.508	4.778	4.778	93831500
########	5.158	6.084	4.66	4.766	4.766	85935500
########	5	5.184	4.054	4.392	4.392	41094000
########	4.6	4.62	3.742	3.84	3.84	25699000
########	4	4	3.166	3.222	3.222	34334500
########	3.28	3.326	2.996	3.16	3.16	34608500
########	3.228	3.504	3.114	3.492	3.492	38557000
########	3.516	3.58	3.31	3.48	3.48	20253000
########	3.59	3.614	3.4	3.41	3.41	11012500
########	3.478	3.728	3.38	3.628	3.628	13400500
########	3.588	4.03	3.552	3.968	3.968	20976000
########	3.988	4.3	3.8	3.978	3.978	18699000
########	4.14	4.26	4.01	4.128	4.128	13106500
########	4.274	4.45	4.184	4.382	4.382	12432500
########	4.37	4.37	4.01	4.06	4.06	9126500
########	4.132	4.18	3.9	4.044	4.044	6262500
########	4.1	4.25	4.074	4.2	4.2	4789000
########	4.238	4.312	4.212	4.258	4.258	3268000
########	4.3	4.3	4.06	4.19	4.19	4611000

Table 1: Data set

1. **Date**: This column represents the date for each data point. It indicates the specific day when the stock market opened and these measurements were recorded.

2. **Open**: The "Open" price is the opening price of TSLA stock on a given trading day. It's the price at which the stock started trading at the beginning of the trading session.

3. **High**: The "High" price is the highest price that TSLA stock reached during the trading session on that particular day.

4. **Low**: The "Low" price is the lowest price to which TSLA stock dropped during the trading session.

5. **Close**: The "Close" price is the closing price of TSLA stock at the end of the trading day. It's the last price at which a trade occurred before the market closed.

6. **Adj Close**: The "Adjusted Close" price is similar to the closing price but adjusted for factors such as dividends, stock splits, and other corporate actions that may affect the stock's historical prices. It's often used for more accurate historical performance analysis.

7. **Volume**: The "Volume" column represents the trading volume for TSLA stock on a given day. It indicates the total number of shares traded during the trading session.

These parameters are essential for conducting technical analysis, building predictive models, and making investment decisions. For stock prediction, analysts and data scientists often use historical data like this to develop models that attempt to forecast future stock prices or to identify patterns and trends in the market.

Data Cleaning:

Data Cleanin meanz da process of findings the incorrects, incomplete, inaccurate, irrelevant, or missing part of da data and then modifications, replacins , or deletin dem accordin' to da necessity! Data cleaning is consider'd a foundational elemints of basic data science. Data is da most valuable ting for Analytics and Machine learnin'. In computin' or Business, data is needed everywhere. When it comes to da present world data, it is not improbable dat data may contain incomplete, inconsistent, or missin' values. If da data is corrupted then it may hinder da process or provide inaccurate results. In my dataset, there are no strings, so there is no need for data cleaning.

DATA VISUALIZATION:

Data visualization be the graphical representation of information for it data into a pictorial or graphical format. Data visualization tools provide an accessible way for see and understanding trends, patterns in data, and outliers. Data visualization tools and technologies be essences to analyzing massive amounts of information and make data-driven decisions. The concept of using pictures to understand data be use for centuries. General types of data visualization be charts, tables, graphs, maps, dashboards, etc... Below be the graphs of our data set which be plotted between each feature and price.

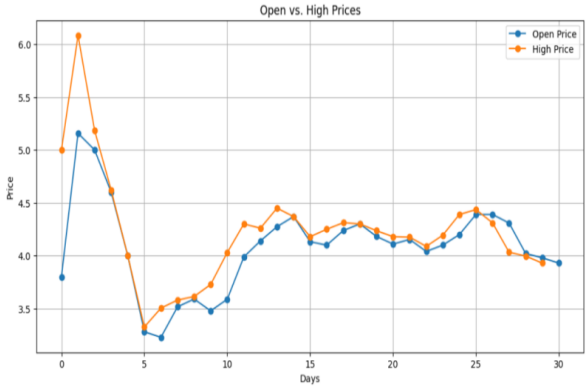

Figure1.Graph: Open price vs High price

(LINEAR REGRESSION)

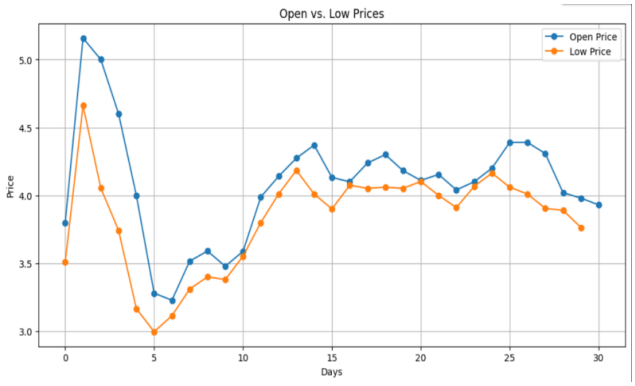

Figure 2.Graph: Open price vs low price

(RIDGE REGRESSION)

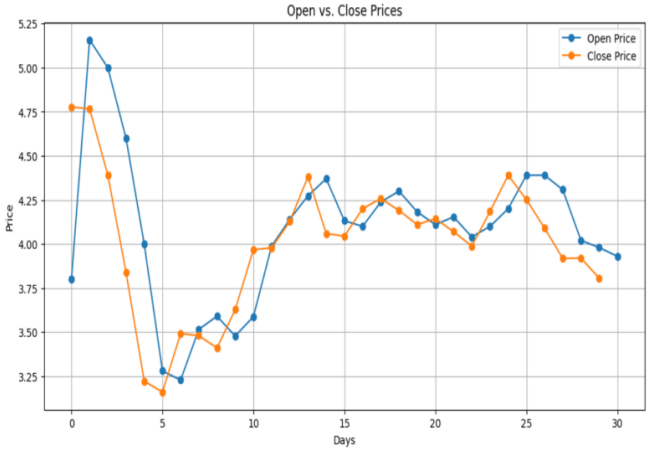

Figure3.Graph: Open price vs Close price

(SVM)

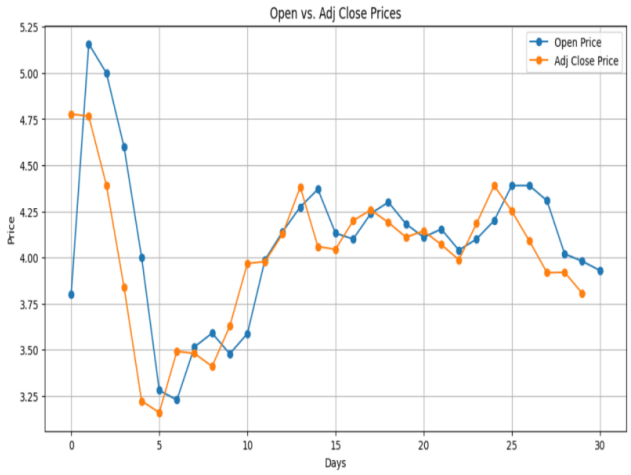

Figure 4. Graph: Open price vs Adj Close price

(KNN)

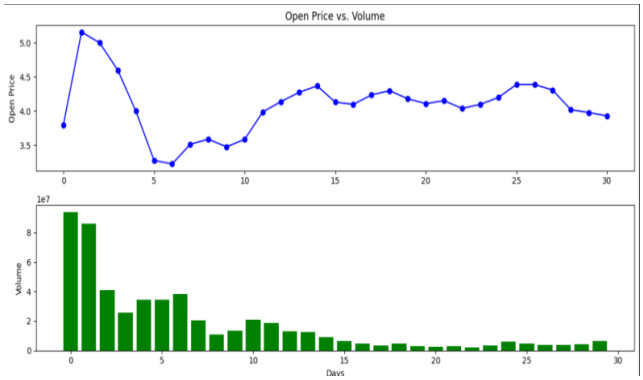

Figure 5.Graph: Open price vs volume
(LASSO REGRESSION)

METHODOLOGY:

Enough methods are performed on the data to evaluate the data set and gather knowledge about the data. Let's perform some Machine Learning models and Experimentation to create a model that helps us to achieve the goal I stated in the problem definition. In this, we talk about the various machine-learning algorithms used for the project. They are Linear regression, Logistic regression, Svm, Knn, Ridge regression, and Lasso regression.

1. LINEAR REGRESSION:

Linear regression is a supervised machine learning algorithm used to establish a linear relationship between a dependent variable and one or more independent features. When there's a single independent feature, it's termed Univariate Linear regression; with multiple features, it's Multivariate Linear regression. The algorithm aims to find the best linear equation to predict the dependent variable based on independent variables, presenting a straight line indicating their relationship, with the slope indicating the change in the dependent variable for a unit change in the independent variable. In regression, Y is the dependent variable to be predicted from the independent variable(s) X, with various functions or models utilized for regression tasks.

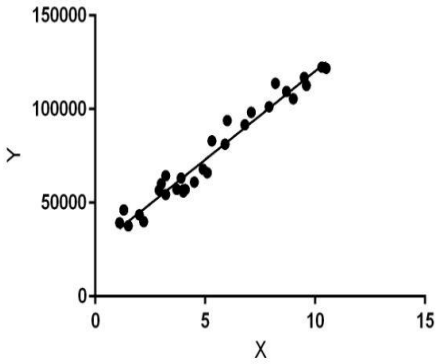

Figure 6: Linear Regression

Formula and Examples of Multiple Linear Regression:

$yi=\beta0+\beta1x1+\beta2x2+...+\beta xi+\epsilon$

where for i=n observations:

yi = dependent variable

xi = feature(independent) variables

β0 = y-intercept (constant term)

βp = slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

In the real estate market, linear regression can be applied to predict housing prices. Independent variables might include square footage, number of bedrooms, location, and other property features, while the dependent variable is the housing price.

In the business world, linear regression can be used to predict customer churn based on various factors like customer tenure, usage patterns, and customer service interactions. Churn (yes/no) is the dependent variable, and customer-related factors are the independent variables.

Utility companies can employ linear regression to forecast energy consumption based on historical data, seasonal trends, and weather conditions. Energy consumption is the dependent variable, and time, weather, and other relevant factors serve as independent variables.

Retailers can use linear regression to predict future sales based on factors such as advertising spending, time of year, and past sales data. Sales are the dependent variable, and marketing budgets and time-related factors are independent variables.

2. RIDGE REGRESSION:

Ridge regression, a regularization technique for linear regression, adds a regularized term to the cost function to prevent overfitting by constraining the weights. This term, controlled by the parameter 'alpha', utilizes the L2 norm to minimize the weights' magnitude. A higher 'alpha' strengthens regularization, reducing variance in estimates. Scaling inputs with Standard Scaler from sklearn is crucial due to the model's sensitivity to input scaling. Ridge regression enhances linear regression by mitigating overfitting, making it a valuable tool in machine-learning projects.

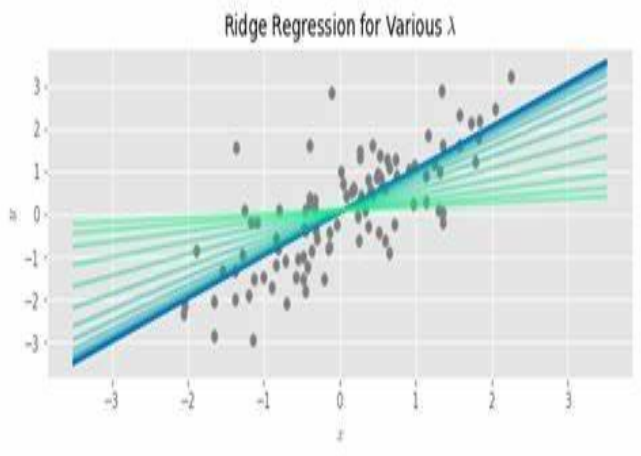

Figure 7: Ridge regression

3. LASSO REGRESSION:

In linear regression, the relationship between input variables and the target variable forms a line or hyperplane. Model coefficients are determined by minimizing the sum squared error between predictions and actual values. However, large coefficients in linear regression can lead to instability, especially in datasets with fewer observations than predictors. To address this, penalized linear regression introduces additional costs for large coefficients, known as the L1 penalty. This penalty minimizes coefficients' sizes and enables some to reach zero, effectively removing irrelevant features from the model.

FORMULA FOR LASSO REGRESSION:

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients)

Where,

λ denotes the amount of shrinkage.

λ = 0 implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model

λ = ∞ implies no feature is considered

The bias increases with an increase in λ

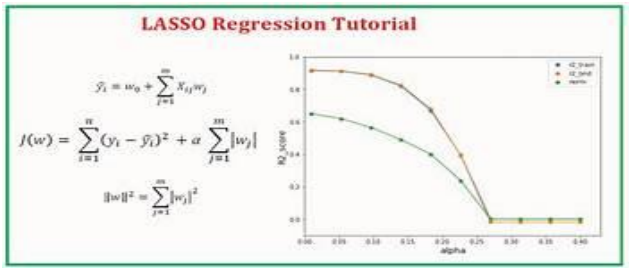

Fig 9: Lasso Regression

Example of Lasso Regression:

In this section, depicted how ilk application the Lasso Regression algorithm. Firstly, allow presentin' a typical regression dataset. Takin' a housing dataset. A housing dataset be an ordinary machine learnin' dataset containin' 506 rows of data with 13 numerical input variables and a numerical aim variable. By applyin' a test harness of repeatin' stratified 10-fold cross-validation with three repeaters, a naive model could fetch a mean absolute error (MAE) near 6.6. A top-performin' form can fetch an MEA on this test like near 1.9. These yield boundaries for the performance expectation on this data put.

SUPPORT VECTOR MACHINE:

Support Vector Machines (SVMs) is versatile supervised learning algorithm use for classifocation and regreshion tasks, preferring datas with many features or claration separation margins. They aims to finds a hyperplane with the large margin between classees in training data, facilitating accurately classifcation. In SVM, each data point is plotted in an N-dimensional space (were N is the number of features), and an optimial hyperplane is determine to separate classess. While inherent binary, SVMs can handles multi-class problems by creations a binary classifier for each classs in the dataset. This approache enables SVMs to effectively classifying datas acroses multiple classes!!!

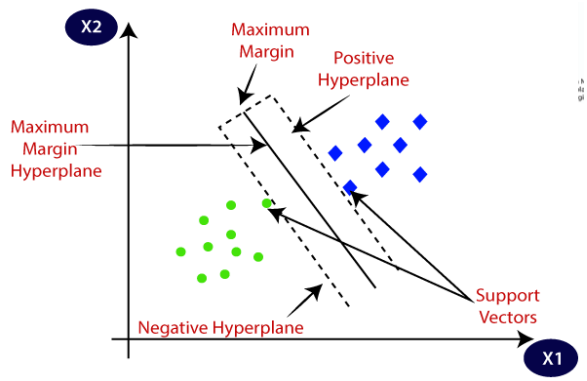

Fig 10. Support Vector Machine

FORMULA:

$f(x) = \text{sign}(\beta 0 + \beta 1*X1 + \beta 2*X2 + ... + \beta n*Xn)$

f(x) is the decision function that predicts the class label.

X1, X2, ..., Xn are the feature values.

β0, β1, β2, ..., βn are the coefficients to be learned during the training process.

sign is the sign function, which returns +1 if the expression is positive and -1 if it's negative.

EXAMPLES:

Support Vector Regression (SVR) have ability to foretell numeric quantities, such as stock costs, by discover a hyperplane that finest fits the data while decrease deviations. Through SVR, instances consist of prophesying dwelling costs founded on features like square footages and total of sleeping rooms or foreseeing a corporation's earnings founded on previous data and financial signals!!!

KNN REGRESSION:

◆K-Narest Nawbor iz a simpl Machin Larning algoritm bassed on Supervised Leaning tecnics!

◆The KNN algorithm assume the similarity among the new case/data and available cases, placing the new case in a categories most similar to an available category!!!

◆KNN algorithm store all the data available and classifying a new data point based on similarities! Meaning, new data appears so it much easier classified into suitable category by using K-NN algorithm.

◆The KNN algorithm can be used for Regression as well as for Classification but it is mostly used for Classification problems.

◆KNN be an no-param fraudulent algorithm, that have not done any thinks on base data.

◆It be likewise known as an idle scholar algorithm coz it do not learn from the teaching group at once rather it saves the data package and at the point of categorization, it executes an action on the data package.

◆The KNN algorithm's training phase simply store the dataset, and when new data arrive, it classifying data into category that is much similar to new data!

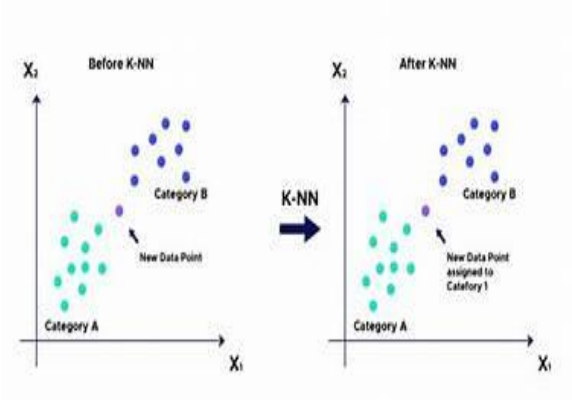

Fig 11. KNN REGRESSION

Formula

KNN algorithm during its training phase just keeping the data set then if it receives newer data, classifying that information into a category that quite resembles the recent data. K-Nearest Neighbors (KNN) be an supervised machine learning algorithm for classification and regression. It forecasting based on the majority class of the k-nearest data points in the feature space. The formulation for KNN might be articulated as:

For Classification:

For a new data point x, KNN predicts class c by selecting the mode (most common class) among the k-nearest neighbors' classes in the training dataset.

For Regression:

For a new data point x, KNN predicts the target value by taking the mean (average) of the k-nearest neighbors' target values in the training dataset.

Examples:

1. KNN was trained on dataset measuring flowers, if there's a new flower like attributes K nearest neighbors, it predicts class based on common class among neighbors!

2. Housing' estimate price predicament KNN, a' it can estimating' a house price by takin' the average of k nearby neighboring' houses with alike characteristics, such as square footage 'n bedrooms number.

3. KNN be used detect anomalies in credit card transactions through calculations distances between transactions and flagging those significantly different from their neighbors!

IV. CONCLUSION

The use of various machine learning algorithms has significantly advanced the field of disease prediction, benefiting conditions such as cardiac problems, renal disorders, breast cancer, and neurological diseases. To further enhance prediction accuracy and efficiency, there is a need to develop more complex machine learning models in the future. Continuous calibration of these models after initial training is crucial for superior performance and more credible predictions. Additionally, expanding the scope of datasets to include a wider representation of demographic factors is essential to prevent overfitting and improve model precision. Utilizing more sophisticated feature selection techniques can further enhance model performance by focusing on the most relevant features, leading to more accurate predictions and simplifying the model. Overall, future disease prediction models will be complex, regularly calibrated, utilize broader and representative datasets, and implement advanced feature selection methods to improve accuracy, effectiveness, and applicability in healthcare and disease diagnosis.

REFERENCES:

[1] Hendri Mahmud Nawawi,1, a) Muhammad Iqbal,2, b) Yudhistira Yudhistira,2, c) Imam Nawawi,3, title is Deep Learning for Tesla's Stock Prices Prediction.

[2] Ogulcan E. Örsel . title is comparative study of machine learning modals for stock price prediction.

[3] Dr. Chaitanya Kishore Reddy.M. title is Analysing Tesla Stock Prices Using Machine Learning Algorithm

[4] Yalong Kong* Faculty of Science, The Chinese University of Hong Kong, Hong Kong, China

[5] Desai Mitesh Madhusudan U.G. Student, Department of Information Technology, B. K. Birla college of Arts, Science and Commerce (Autonomous), Kalyan 421 306 , Maharashtra, India