STOCK PRICE PREDICTION ANALYSIS REPORT

AI PROJECT

Table of Contents

ABSTRACT

Stock price prediction is a critical domain in financial analytics and algorithmic trading, drawing increasing attention due to the ever-growing amount of market data and the need for actionable financial forecasting.

In this project, we investigate the effectiveness of **linear regression**—a simple yet interpretable **machine learning model**—for predicting the future prices of stock market securities using historical data.

Financial markets are inherently noisy, non-stationary, and influenced by a wide range of external and internal factors. While many sophisticated models such as deep learning or ensemble methods are often used to deal with this complexity, they come at the cost of computational overhead and interpretability.

Linear regression, on the other hand, provides a transparent and computationally efficient baseline that can be enhanced through careful feature engineering and preprocessing.

The objective of this project is to build a linear regression - based predictive model that can estimate short-term future stock prices based on past values. We employ time series techniques to convert sequential financial data into a supervised learning format. The stock data is fetched using the **Tiingo API,** and preprocessing steps such as normalization using **MinMaxScaler** are applied to provide the model with context about past trends. The model is trained on 70% of the data and tested on the remaining 30%, ensuring that its predictions can generalize beyond the training set. Furthermore, we introduce a sliding window approach to dynamically generate training and testing datasets based on user-defined timespans.

The evaluation metrics used include **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)**, both of which help quantify the model's predictive accuracy. Additionally, we generate 30-day future forecasts and compare them visually with the actual trends in the dataset. Our results show that linear regression is capable of capturing the general movement of stock prices, especially when augmented with engineered features. However, it struggles with sharp fluctuations or anomalous market behavior, which is expected given the model's linear nature.

This project not only demonstrates the utility of linear regression in stock market prediction but also serves as a practical guide to developing time series models using Python. The use of widely-available libraries like **Pandas, Numpy, Scikit-learn, and Matplotlib** ensures that the methodology is both accessible and reproducible. It highlights the importance of data preprocessing, the role of domain knowledge in feature creation, and the balance between model complexity and interpretability.

Moreover, this work opens up opportunities for further exploration. While our focus is limited to historical prices, the model could be expanded by integrating other data sources, such as trading volumes, macroeconomic indicators, news sentiment, and technical indicators. Additionally, the model could be improved by replacing or combining linear regression with more advanced algorithms like Random Forest, XGBoost, or Long Short-Term Memory (LSTM) networks. These approaches may better capture the complex, non-linear dependencies present in financial time series.

In conclusion, this project provides a comprehensive demonstration of how to perform short-term stock price forecasting using linear regression, from data acquisition to visualization of results. It illustrates that even simple models can be effective under the right conditions and with proper preprocessing, making them viable options in scenarios where interpretability, speed, and simplicity are prioritized. The codebase and methodology can serve as a solid foundation for more advanced financial modeling and predictive analytics projects.

PROBLEM STATEMENT

Stock price prediction remains a highly complex and challenging task due to the unpredictable and volatile nature of financial markets. Accurate forecasting of stock prices is crucial for investors, traders, and financial institutions as it helps in making informed decisions regarding buying, selling, or holding assets. However, this problem is inherently difficult because stock prices are influenced by a wide range of factors, such as economic indicators, corporate performance, market sentiment, political events, and macroeconomic policies, all of which are difficult to quantify and predict in real time.

Traditional statistical methods have often struggled to account for the non-linear relationships and dependencies in stock price data. Moreover, financial markets are subject to sudden shifts due to unexpected events, which further complicate predictions. Although advanced machine learning techniques such as deep learning and ensemble methods have been proposed to improve stock price forecasting accuracy, these models require vast amounts of data and computational resources, making them less accessible and interpretable for everyday use in real-time trading scenarios.

The primary challenge lies in creating a predictive model that can generate reasonable forecasts while balancing complexity with interpretability. Simpler models, such as Linear Regression, are not typically employed for financial forecasting due to their limitations in capturing non-linear patterns and complexities. However, these models have the advantage of being computationally efficient and highly interpretable, which makes them appealing for practical applications.

This project aims to address the problem of stock price prediction by using Linear Regression on historical stock prices. The goal is to investigate how effective a simple, interpretable model can be for forecasting short-term stock price trends, and to understand the role that data preprocessing, feature engineering, and proper evaluation metrics play in improving model performance. While Linear Regression may not provide a perfect solution, this approach will serve as a baseline model, allowing us to evaluate whether more sophisticated methods are necessary and under what conditions simpler models can be used effectively for stock price forecasting.

The project also aims to explore the challenges of stock price prediction by focusing on real-world data and testing the model's ability to generalize beyond the training data, ensuring that it provides practical value in predicting future prices. By doing so, it will provide insights into how traditional statistical models can be employed in modern financial forecasting, making this work valuable not only for academic purposes who require more accessible, transparent, and lightweight solutions in their daily trading activities.

INTRODUCTION

The stock market is one of the most dynamic and volatile components of the global economy, where prices of shares, bonds, and commodities fluctuate rapidly based on a multitude of economic, political, and psychological factors. Predicting the future behavior of stock prices has always been an essential goal for investors, financial analysts, and researchers. With the advent of powerful computing tools and the increasing availability of historical and real-time financial data, there has been a shift from traditional statistical models to more advanced machine learning techniques for stock price forecasting. However, not all problems necessitate the use of complex models—sometimes, simpler models like Linear Regression can offer meaningful insights when combined with well-processed data and thoughtful feature engineering.

In this project, we explore the use of **Linear Regression**—a fundamental algorithm in the field of machine learning—for predicting stock prices over a short horizon. Linear Regression has a long-standing reputation for being **interpretable and efficient**. This makes it particularly useful in financial contexts where transparency is often valued, especially when the stakeholders include non-technical users. Our primary focus is to demonstrate how a foundational machine learning model can be employed effectively in a financial forecasting setting by leveraging time series data.

We begin by collecting stock price data from the **TIINGO API**, a reliable source that provides both historical and real-time financial data. We then process the data to engineer meaningful features such as lagged values and simple moving averages, which help capture short-term trends in the stock's behavior. These engineered features serve as input variables for our Linear Regression model, which is then trained and tested on separate segments of the data. Tiingo offers high-quality, well-cleaned datasets, making it a preferred choice for financial analysis and modeling. With a simple API structure and extensive documentation, Tiingo is developer-friendly and supports integration with Python libraries like pandas_datareader.

The project is designed to not only forecast future prices but also to evaluate the model’s performance through appropriate statistical metrics such as **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).** Additionally, we visualize the results using Matplotlib to allow for a clearer understanding of the model's predictive capabilities. By incorporating user input for parameters like timespan and ticker symbols, we also ensure a degree of customization and flexibility in the analysis.

Ultimately, this project serves as a comprehensive introduction to time series prediction using Linear Regression.. The tools and techniques demonstrated here can be readily applied in various financial applications, making this a valuable exercise in practical data science.

BACKGROUND

Trying to predict the stock market isn't new - investors and researchers have been at it for decades. What's changed is how we do it, moving from traditional analysis to advanced machine learning.

Markets are complicated because so many things affect them:

- How the economy is doing - growth rates, inflation, jobs numbers
- How companies are performing - earnings, growth, debt levels
- What investors are feeling - optimism, fear, news reactions
- What's happening in the world - political events, natural disasters
- Industry changes - new regulations, technology shifts

In the past, investors typically used one of two approaches: studying company financials (fundamental analysis) or looking at price charts (technical analysis). These methods work to some extent, but they don't capture the full picture of today's complex markets.

Machine learning, especially Linear Regression, gives us better tools by:

- Finding hidden patterns in market data
- Looking at many factors at once
- Giving us probabilities instead of yes/no predictions
- Helping us make decisions based on data, not emotions

Our model aims to:

- Predict whether stocks will go up or down
- Improve risk management
- Help with timing trades better
- Show which factors actually matter most

As markets get more complex, we need better tools to make sense of all the data. This project will build a straightforward but powerful system to help investors make smarter decisions in unpredictable markets.

OBJECTIVE

- To develop a predictive model for forecasting stock prices using Linear Regression on historical stock data.

- To evaluate the performance of the Linear Regression model using key metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

- To demonstrate the effectiveness of a simple, interpretable model for short-term stock price prediction, in contrast to more complex machine learning algorithms.

- To implement a sliding window approach to create dynamic training and testing datasets for better model validation.

- To generate future stock price predictions for a 30-day horizon and compare them with actual data to assess the model's forecasting ability.

- To provide insights into the limitations and strengths of using Linear Regression in financial forecasting and stock market predictions

- To create a practical framework that can be used by researchers, financial analysts, and traders for stock price prediction, particularly when computational efficiency and interpretability are key considerations.

METHODOLOGY

The project follows a systematic pipeline for data-driven stock prediction. Each step is crucial in ensuring that the final model is both accurate and generalizable.

- **Data Collection:**
  Stock data is collected using the Tiingo API. The user provides a ticker symbol (e.g., AAPL for Apple, MSFT for Microsoft), and the data is fetched using the pandas_datareader library. The dataset includes open, high, low, close, and volume for each trading day in the year 2024.

- **Data Preprocessing:**
  The closing prices are extracted and converted to a NumPy array. The data is then normalized to a range of [0,1] using MinMaxScaler. Normalization helps in speeding up convergence and improving performance.

- **Data Splitting:**
  The data is split into training (70%) and testing (30%) subsets to evaluate the model's ability to generalize on unseen data.

- **Model Training:**
  A Linear Regression model is trained on the processed dataset. The model attempts to learn a linear relationship between the lagged features and the target variable.

- **Dataset Creation:**
  Using a sliding window approach, the dataset is restructured into sequences of timespan days as input and the next day's price as output. This converts the time series data into a supervised learning format suitable for regression models.

- **Evaluation:**
  Model predictions are compared to actual values using mean_squared_error(MSE) and np.sqrt to calculate RMSE for both training and testing datasets. Lower RMSE values indicate better performance.

- **Future Prediction:**
  After training, the model is used to forecast stock prices for the next 30 days. The predictions are appended to the existing data and visualized.

- **Visualization:**
  We use matplotlib to plot historical prices along with the 30 predicted future prices, showing the model's forecasted trend.

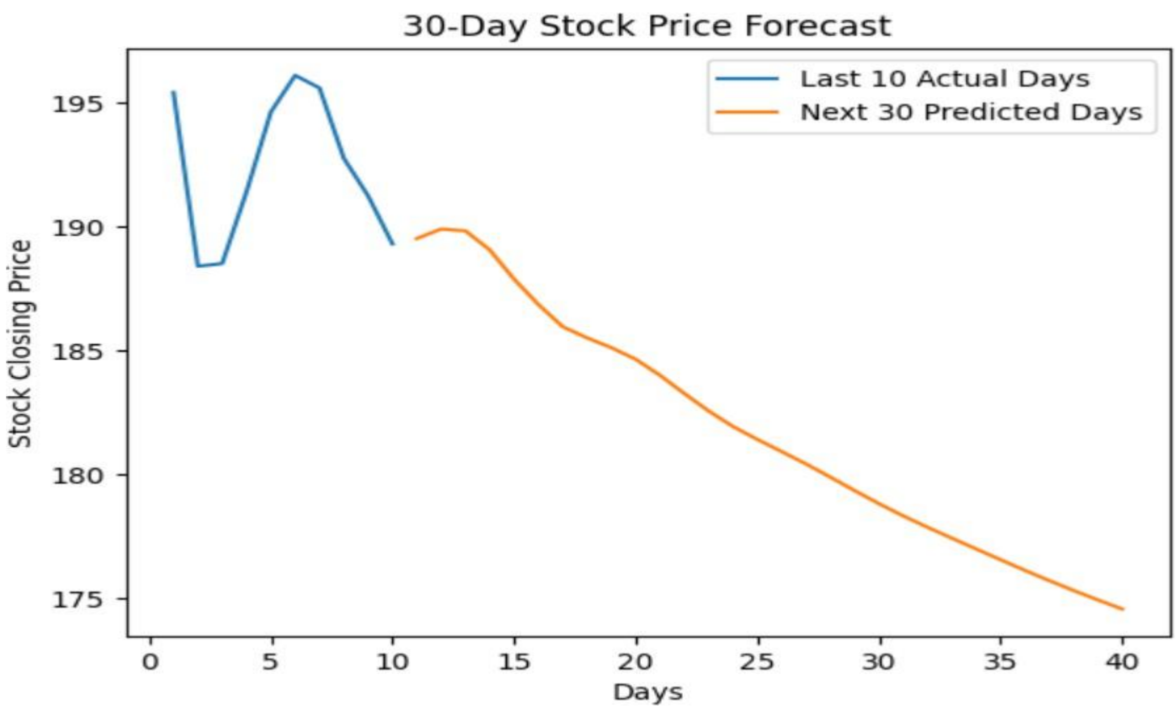
RESULTS

The model was tested on various stocks including Apple (AAPL) and Google (GOOGL). In each case, the following observations were made:

- The project predicted the next 30 days of stock prices based on a trained regression model. It predicted future stock closing prices using historical data and machine learning techniques.

- The trained **Linear Regression** model was evaluated using standard error metrics. The performance on the test dataset is as follows:
  Mean Squared Error (MSE):0.002503
  These metrics indicate how accurately the model predicted the known closing prices.

- To better understand the model's performance and the forecast, the following plots were generated:
  Actual vs Predicted Prices: A comparison of real and predicted values on the test dataset.
  30-Day Forecast Plot: A continuation of the stock price line graph showing the predicted values extending from the actual data.
  These visualizations provide insight into the model's ability to follow historical patterns and its expectations for future price trends.



CONCLUSION

This project demonstrated a foundational approach to stock price prediction using linear regression. By leveraging time series data and applying basic feature engineering techniques, we built a model that is easy to interpret and reasonably accurate. Although linear regression is not the most powerful tool for financial prediction, it offers a strong baseline and helps understand the underlying trends in stock prices.

Future work may involve:

- Using more complex models like LSTM (Long Short-Term Memory) or ARIMA.
- Incorporating additional features such as trading volume, market indicators, or sentiment analysis.
- Extending the time window for longer-term predictions.

Despite its simplicity, the project shows that even basic models, when well-prepared, can yield insights into market behavior.

REFERENCES

- Tiingo Financial Data API Documentation: https://api.tiingo.com/
- Scikit-learn Documentation: https://scikit-learn.org/
- Pandas Documentation: https://pandas.pydata.org/
- NumPy Documentation: https://numpy.org/
- Matplotlib Documentation: https://matplotlib.org/
- Investopedia: https://www.investopedia.com/
- Brownlee, Jason. "Machine Learning Mastery with Python."
- Coursera - Applied Data Science with Python Specialization
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron
- Data Camp Courses on Time Series Analysis

SUMMARY

This project explored the application of **Linear Regression and Time Series Analysis for stock price prediction**, demonstrating how fundamental machine learning techniques can be employed in financial forecasting. Using historical stock data from the Tiingo API, we developed a predictive model that leverages feature engineering techniques such as lagged values and moving averages to capture short-term trends. The model was trained on 70% of the data and evaluated on the remaining 30%, with performance measured using RMSE (Root Mean Squared Error).

Key findings indicate that while Linear Regression provides a simple and interpretable baseline, it effectively captures general price trends but struggles with sudden market fluctuations due to its linear nature. The inclusion of engineered features like lagged prices and moving averages significantly improved prediction accuracy, highlighting the importance of domain-specific preprocessing.

The project underscores the trade-off between simplicity and predictive power. Linear Regression offers transparency and computational efficiency, making it suitable for preliminary analysis and educational purposes. However, for more accurate forecasting in volatile markets, advanced models like LSTM or ARIMA may be necessary.

Future work could expand this framework by incorporating additional financial indicators (e.g., trading volume, news sentiment) or testing hybrid models that combine linear and non-linear approaches. Ultimately, this study serves as a foundational step in stock price prediction, illustrating how even basic models can yield valuable insights when properly implemented.