Automotive Stock Price Prediction

Steven Sullivan, Robert Lignowski, Uditi Shah, Ahmad Javed

Introduction

- Objective: Forecast future stock prices for Ford, Tesla, Toyota, and Volkswagen

- Motivation:
  - Understand the impact of market trends, policy changes, and international relationships on automotive stocks
  - Learn how to create and implement an accurate time series forecasting model

- Tools: Predictive Models and Algorithms
  - ARIMA,
  - Prophet
  - XGBoost
  - LSTM

- Data Sources: Historical stock prices (Investing.com)

Predictive Modeling Problem Definition

- Goal: Use time series and machine learning models to predict future stock price
- Train models on historical data (2020–2024)
- Forecast prices for early 2025
- Evaluation:
  - RMSE (Root Mean Squared Error)
  - MAPE (Mean Absolute Percentage Error)

Data Pre-Processing and Acquisition

Data Collection: Historical stock price data for Volkswagen, Toyota, Tesla, and Ford were gathered from reliable financial data sources(www.investing.com).

Data Cleaning: The datasets were examined for missing values, inconsistencies, and anomalies. Appropriate methods, such as imputation and removal, were applied to address these issues.

Feature Engineering: Additional relevant features were created to enhance the predictive power of the models. This included generating technical indicators and other derived metrics.

Data Transformation: The data were transformed as necessary, including normalization or scaling, to ensure compatibility with the modeling techniques employed.

Data Splitting: The processed data were divided into training and testing sets, typically using an 80/20 split, to facilitate model evaluation and validation.

Final Dataset Preparation: The cleaned and transformed datasets were finalized for input into the forecasting models, ensuring readiness for the subsequent modeling phase.

Exploratory Data Analysis(EDA) and Visualisation

Cleaned and analyzed stock data for Tesla, Ford, Toyota, and Volkswagen.

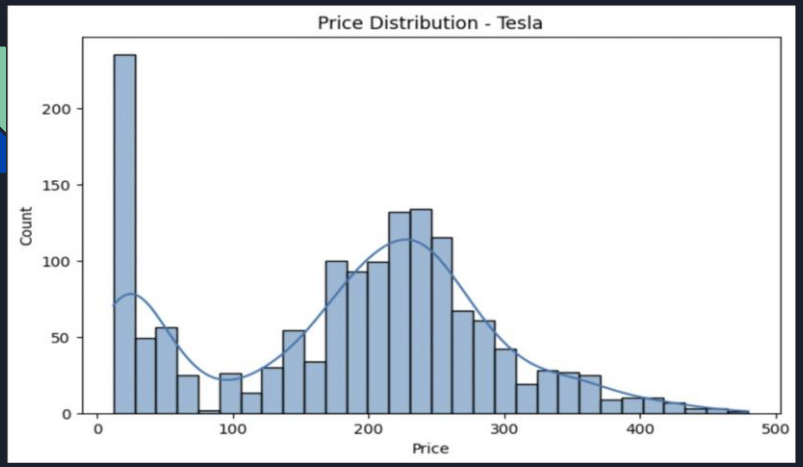
Histograms and trend lines showed Tesla had the most volatility; Toyota was the most stable.

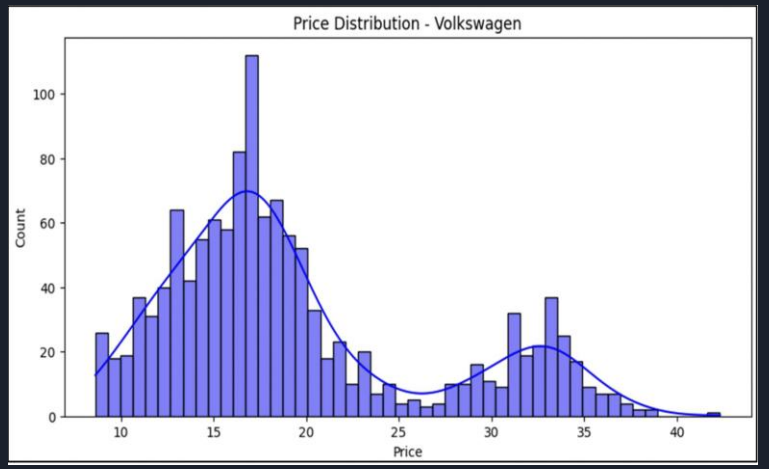
Volume spikes aligned with major events, especially in Tesla and Ford.

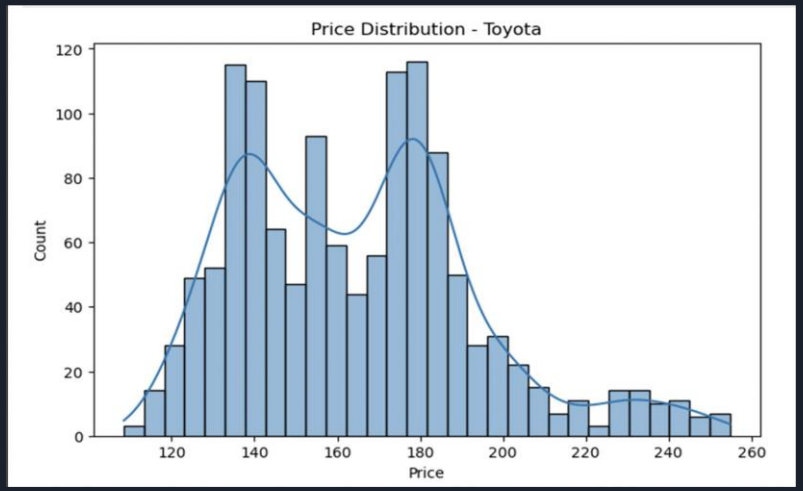
Box plots revealed Tesla had significant outliers; Ford and Toyota were steadier.

Correlation matrices showed strong relationships among price variables; volume and change % offered added insights.

No missing values were found in any dataset.


Price Distribution - Tesla
Count
Price
0
50
100
150
200
0
100
200
300
400
500


Price Distribution - Volkswagen
Count
Price
0
20
40
60
80
100
10
15
20
25
30
35
40


Price Distribution - Toyota
Count
Price
0
20
40
60
80
100
120
120
140
160
180
200
220
240
260


Distribution of Closing Prices - Ford
Frequency
Closing Price
0
50
100
150
200
250
300
5
10
15
20
25

First Predictive Model - ARIMA

- Captures linear trends and seasonality in data
- Uses past values to predict future values
- Best for stable, consistent patterns with minimal noise
- Traditionally used for financial or economic data

Other Feature and Models

- Expanded beyond ARIMA to include Prophet, XGBoost, and LSTM
- Forecasts evaluated using RMSE and MAPE
- XGBoost achieved best performance overall on most stocks
- Hybrid model (LSTM + SARIMA) developed for Volkswagen
- Hybrid model achieved the lowest RMSE (0.5388) and MAPE (3.60%)
- Highlights potential of combining statistical and deep learning approaches

Volkswagen Dataset

Model	RMSE	MAPE
ARIMA	1.72	17%
Prophet	2.25	13.3%
XGBoost	1.50	9.99%
LSTM	0.55	3.64%
Hybrid (LSTM + SARIMA)	**0.54**	**3.60%**

Results

- Across companies, there was little to no seasonality or patterns over a five-year period
- All companies showed a spike in stock price in late 2021/late 2022
  - But there were little overall trends
- Predictions
  - XGBoost was the most accurate, with the lowest RMSE and MAPE scores.
    - Also handled noise and short-term fluctuations
  - LSTM models also worked well - especially for Tesla
  - Prophet and ARIMA were less effective
    - Prophet underperformed
    - A hybrid between SARIMA and LSTM was effective for Volkswagen
  - This occurred because the model had low seasonality
- Overall economic and global factors can result in a change of stock prices
  - best to choose models that fit that data present to get the most accurate predictions.

Lessons Learned

- Importance of data preprocessing
  - Handling missing values
  - Syncing timelines
  - Changing data types (i.e. string to float)

- Time series forecasting is highly sensitive to seasonality and trends

- Learning models like XGBoost AND LSTM perform better with non-seasonal/ irregular data compared to more traditional models (ARIMA and LSTM)

- External influences like policy and economic factors can’t always be captured purely by numerical data