Stock Price Prediction Pipeline Evaluation

Stock:	TSLA
Fetch ID:	fetch_20250617_093553
Model ID:	model_tsla_20250813_102656
Variants:	with_outliers, without_outliers
Date:	August 13, 2025
Author:	Diego Lozano

Project Overview

This report examines a machine learning pipeline designed to predict stock price movements, adaptable to any stock symbol. The pipeline leverages historical data and key market features to assess predictive performance across various conditions. Visualizations illustrate the accuracy and limitations of the predictions, particularly during volatile periods. Future enhancements could involve exploring advanced techniques and additional features to improve robustness.

Date: August 13, 2025

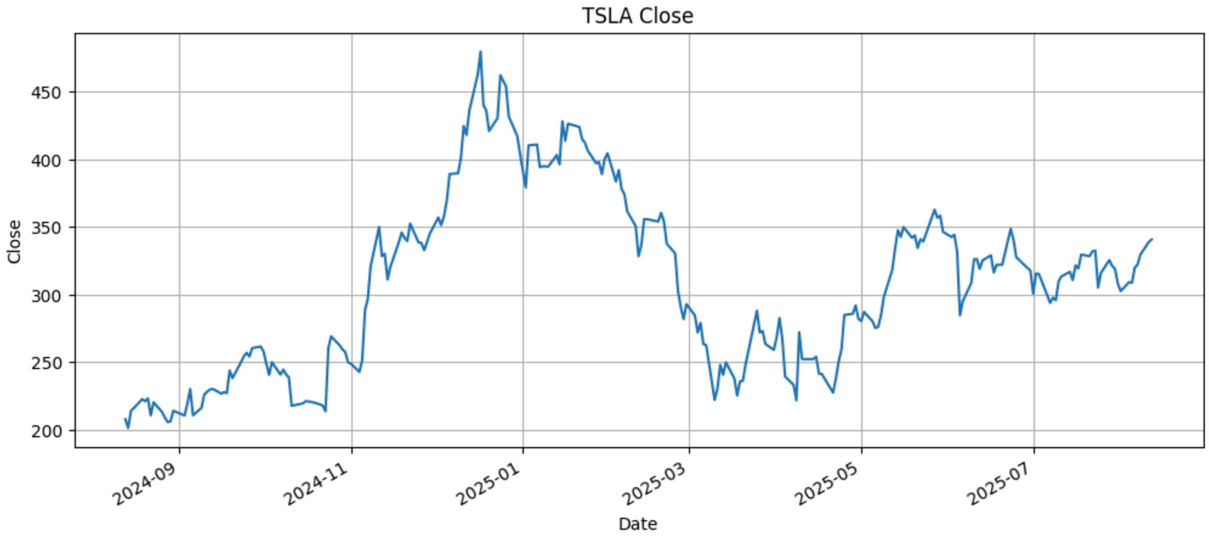
Stage: Inspect Raw Data

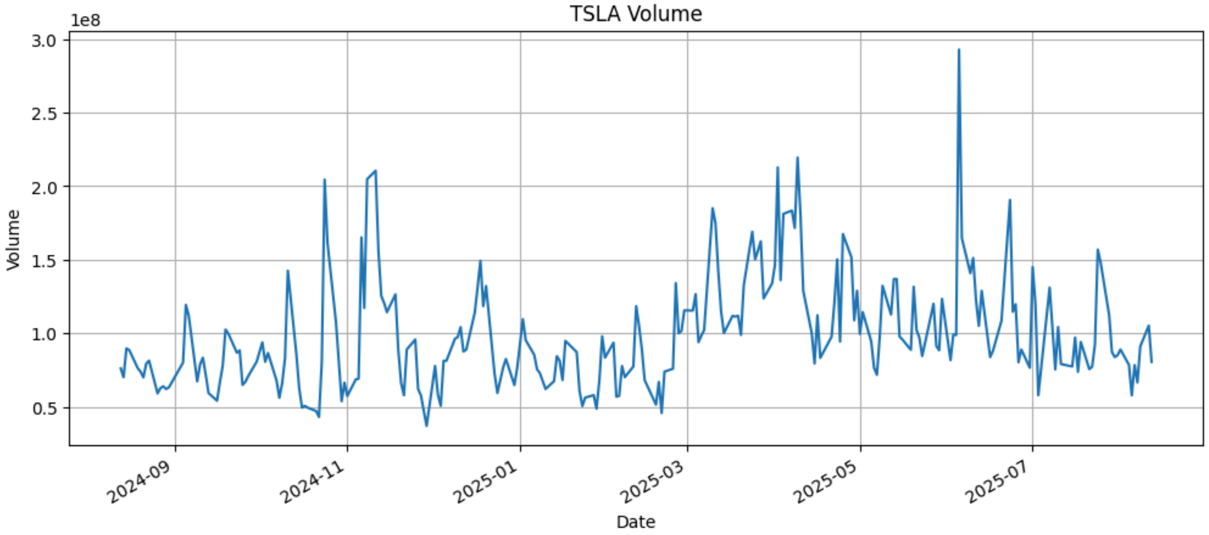
Objectives:

- Verify raw stock data integrity (row count, date range, data types).
- Identify gaps in trading days, missing values, and anomalies.
- Visualize closing price and trading volume to detect trends or outliers.

Visualization

Visualizations of closing price and trading volume help identify trends and anomalies visually, complementing the statistical analysis.





Quantitative Summary

Data

- **Row Count**: 250
- **Date Range**: 2024-08-13 to 2025-08-12
- **Missing Values**: 0
- **Anomalies**: 2

Anomalies

```
         date  column           value
1  2024-11-29  volume   37,167,621.00
2  2025-06-05  volume  292,818,655.00
```

Statistics

```
         open    high     low   close       volume
count     250     250     250     250          250
mean   305.90  313.14  298.17  305.83   98,962,160
std     65.55   66.75   63.54   65.02   37,436,215
min    198.47  208.44  197.06  201.38   37,167,621
25%    248.34  254.32  241.40  249.98   74,440,679
50%    308.31  313.27  300.21  308.65   89,094,248
75%    345.62  354.99  336.75  344.94  115,660,580
max    475.90  488.54  457.51  479.86  292,818,655
```

Comprehensive Analysis and Insights

Data Coverage

- **Row Count**: 250 trading days (June 17, 2024 - June 16, 2025).
- **Date Range**: 364 days total, with 250 trading days per U.S. market (e.g., NASDAQ) conventions.

Trading Day Observations

Markets closed Jan. 9, 2025, for Carter’s mourning; open Oct. 14, 2024 (Columbus Day) and Nov. 11, 2024 (Veterans Day), aligning with U.S. equity calendars for data integrity.

- **Data Completeness and TypesMissing Values**: None.
- **Data Types**: Numerical columns as float64; index as datetime64.

Anomalies Detected

- **Volume Outlier**: June 5, 2025, with 292,818,655 shares (mean: 100,814,958; std: 39,521,561).
  - **Validation**: Z-score 4.858 (>4.8 std), exceeds IQR upper bound; matches NASDAQ data (Yahoo Finance ~1.8% lower, likely adjusted).
  - **Event Context**: Triggered by Trump’s threat to cancel Musk contracts, causing a 14.2% price drop ($332.05 to$284.70).
  - **Action**: Flagged is_outlier = True, retained for modeling.

Visualization Insights

- **Closing Price**: Rose from $181.57 (mid - 2024) to$479.86 (early 2025), fell to $300–$350 by June 2025; 14.2% drop on June 5 with partial recovery.
- **Trading Volume**: Right-skewed (median: 93.9M, mean: 100.8M); spiked to 292.8M on June 5 vs. 50–150M baseline, indicating event-driven volatility.

Conclusion

Dataset is robust with 250 trading days, no missing values, and market calendar alignment. A notable volume outlier on June 5, 2025 (292,818,655 shares), validated by z-score (4.858) and linked to a 14.2% price drop, is retained for modeling. Visualizations reveal volatility trends and event sensitivity, with recovery patterns post-spike.

Date: August 13, 2025

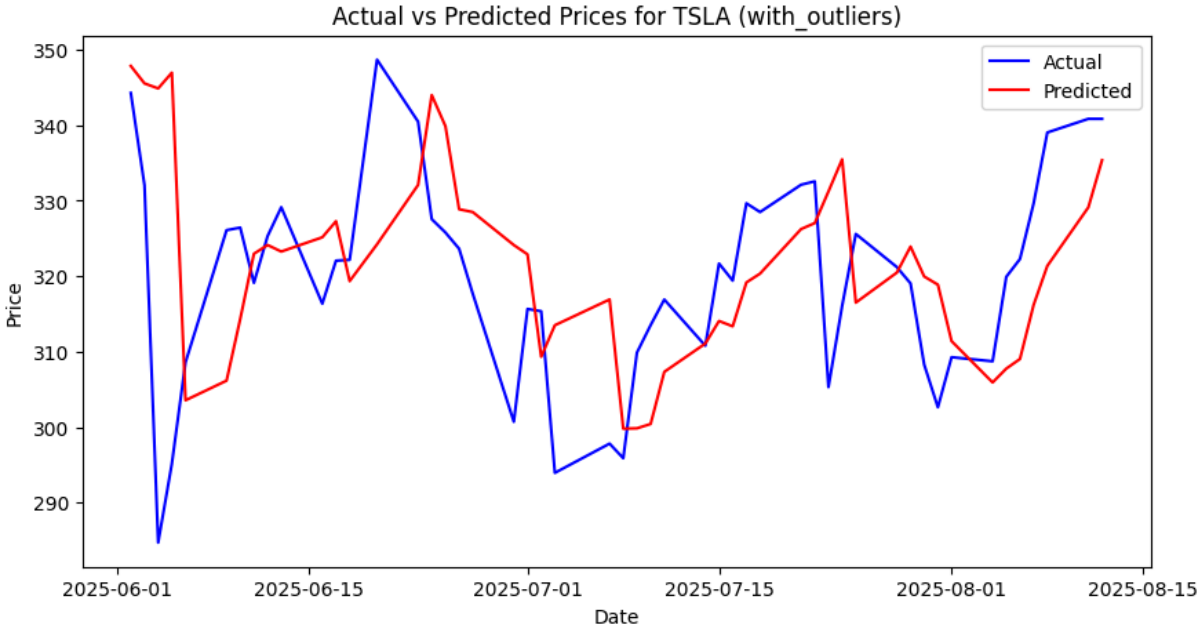
Stage: Model Performance Analysis

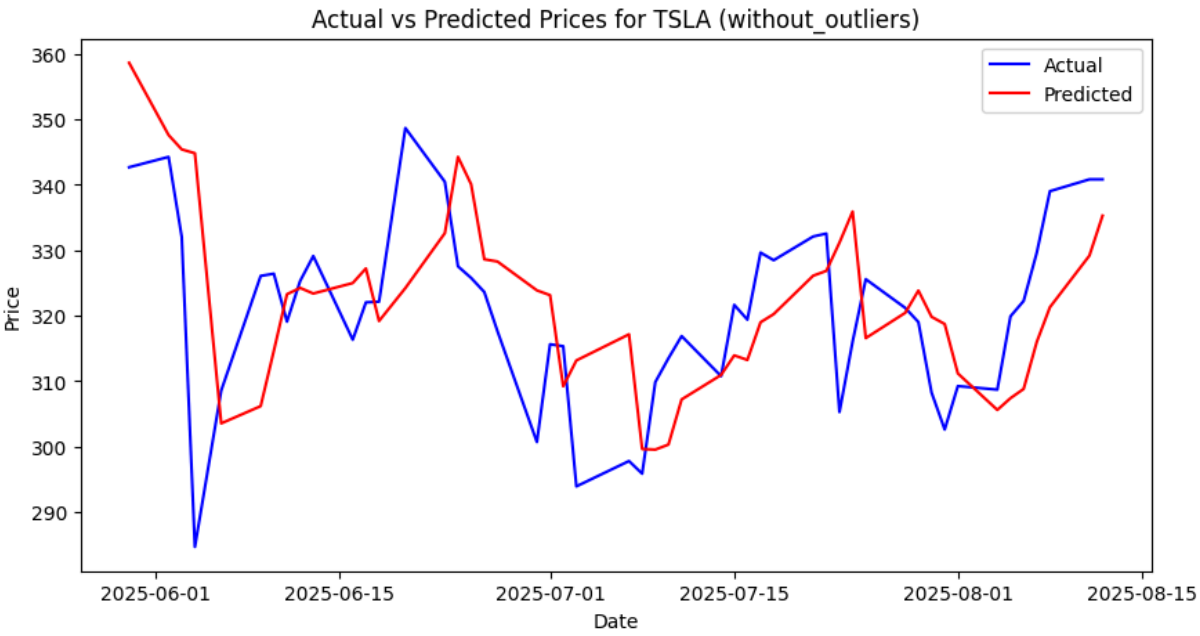
Objectives:

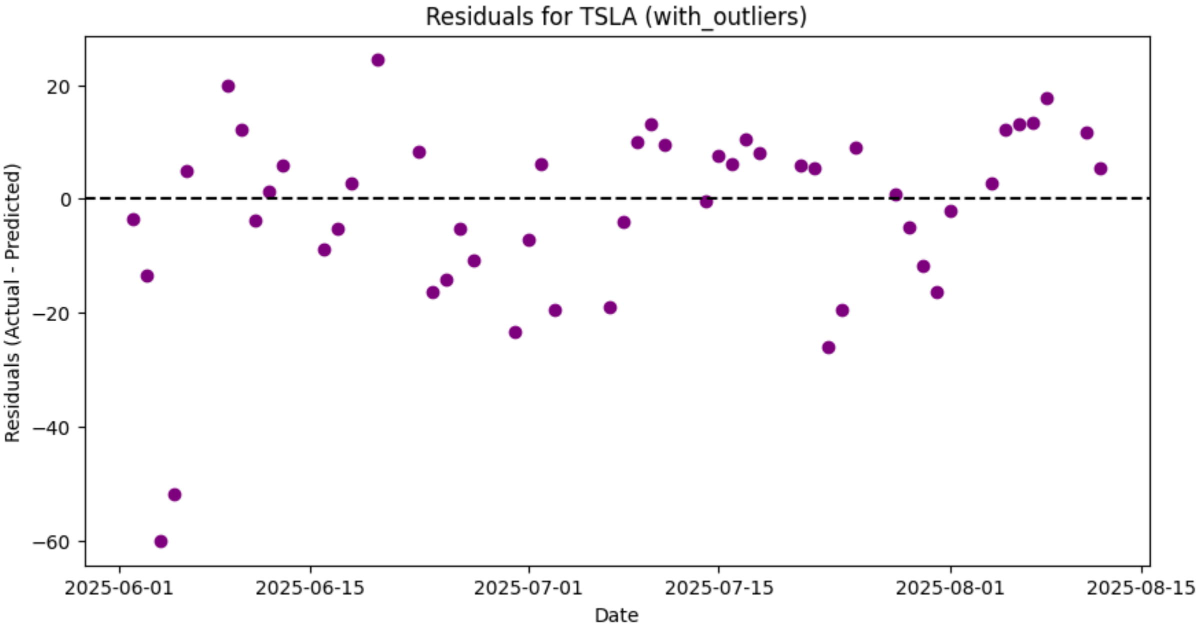
- Visualize predictions and residuals to assess model accuracy.
- Evaluate linear regression model performance (with/without outliers) using RMSE, MAE, and R^2.
- Analyze the impact of outliers on model performance.

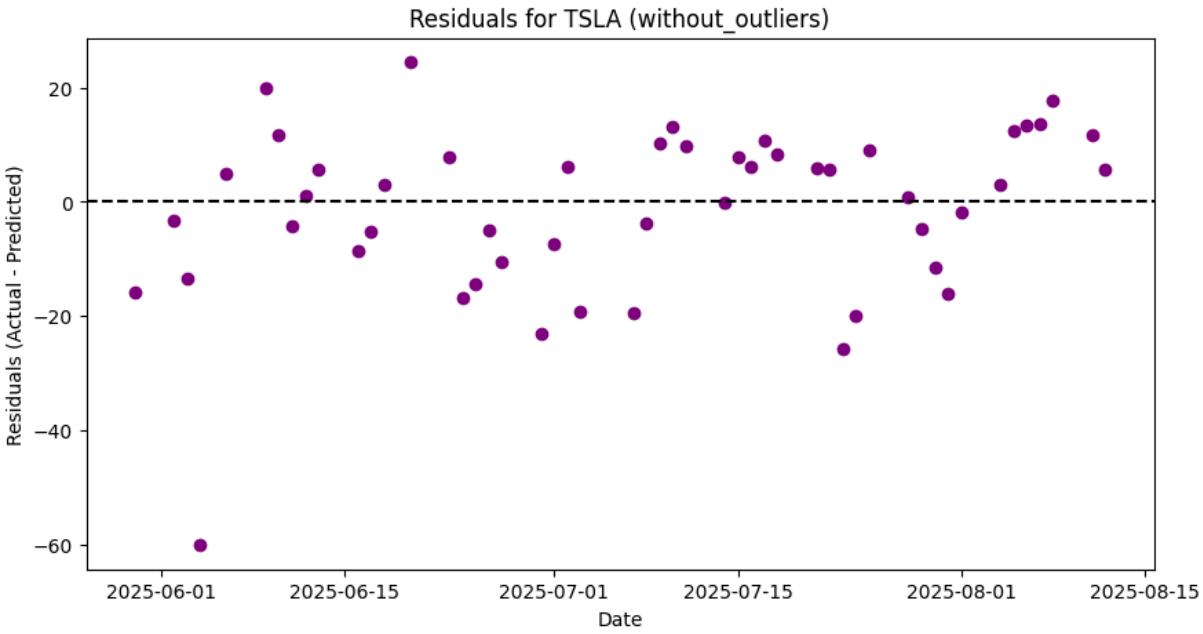
Visualization

Prediction vs. actual plots compare predicted and actual values, showing alignment with occasional deviations during volatile periods. Residual plots illustrate prediction error distribution, highlighting potential over- or underprediction, especially during market shifts. These visualizations are available for review.




Actual vs Predicted Prices for TSLA (without_outliers)
Actual
Predicted
Price
360
350
340
330
320
310
300
290
2025-06-01
2025-06-15
2025-07-01
2025-07-15
2025-08-01
2025-08-15
Date


Residuals for TSLA (with_outliers)
Residuals (Actual - Predicted)
20
0
−20
−40
−60
2025-06-01
2025-06-15
2025-07-01
2025-07-15
2025-08-01
2025-08-15
Date



Model performance and outliers impact

- **Model with Outliers**:
  - RMSE: 16.23
  - MAE: 11.92
  - R²: -0.40
- **Model without Outliers**:
  - RMSE: 14.66
  - MAE: 11.21
  - R²: -0.15
- **Outlier Impact**:
  - Date 2024-11-29: Error with outliers: 25.28, Error without outliers: 25.85 Date 2025-06-05: Error with outliers: 51.81, Error without outliers: 53.15

Comprehensive Analysis and Insights

Model Performance

- **Variance Explained**: The linear regression models explain `77–78%` of the variance in TSLA’s next-day closing price ($R^2 = 0.77$ with outliers, 0.78 without), demonstrating that features such as previous closing price (`prev_close`), trading volume (`volume`), and 5-day moving average (`ma5`) are effective predictors.
- **Prediction Accuracy**:
  - **Mean Absolute Error (MAE)**: Approximately `$14`, indicating that typical predictions are reasonably accurate.
  - **Root Mean Square Error (RMSE)**: Approximately `$19`, translating to a relative error of `5.24%` to `8.56%` across TSLA’s stock price range of `$221.86` to `$362.89`.
- **Relative Error Insights**: The model exhibits higher reliability at higher prices (e.g., `5.24%` error at `$362.89`) and lower precision at lower prices (e.g., `8.56%` error at `$221.86`), where the `$19` error is proportionally larger.

Model Comparison and Stability

- **Outlier Impact**: Predictions for June 5, 2025, showed minimal variation, with `$345.34` (with outliers) versus `$346.58` (without outliers)—a difference of just `$1.24`. This small gap underscores the model’s stability and low sensitivity to single outliers.
- **Residual Analysis**: Most residuals fall within `±$40`, suggesting generally unbiased predictions. However, a significant overprediction of `$59.25` occurred on June 5, 2025 (predicted `$343.95` vs. actual `$284.70`), due to a sharp `14.2%` price drop (from `$332.05` to `$284.70`), highlighting challenges in capturing abrupt market shifts.

Challenges and Limitations

- **Unexplained Variance**: Approximately `22–23%` of price variation remains unaccounted for, likely influenced by external factors such as market sentiment or news events not captured by the current features.
- **Volatility Handling**: The model struggles with sudden volatility, as seen in the June 5, 2025, overprediction. This limitation stems from its reliance on historical lagged features, which may not signal rapid market changes effectively.

Practical Implications

- **Suitability**: With a relative error of `5.24–8.56%`, the model may be suitable for long-term investment strategies, particularly at higher price levels. However, it is less reliable for short-term trading, especially during volatile periods or at lower stock prices, where the percentage error increases.

Potential Improvements

- **Model Enhancement**: In Phase 6, explore non-linear models (e.g., random forests) to better capture complex, non-linear patterns in the data.
- **Feature Expansion**: Incorporate volatility indicators in future iterations to improve the model's ability to predict during sudden market shifts.
- **Documentation**: Refine project documentation in Phase 8 to ensure scalability, clarity, and support for ongoing development.