Stock Price Prediction Pipeline Evaluation

Stock: TSLA
Fetch ID: fetch_20250617_093553
Model ID: model_tsla_20250807_220050
Variants: with_outliers, without_outliers
Date: August 07, 2025
Author: Diego Lozano

Date: August 07, 2025

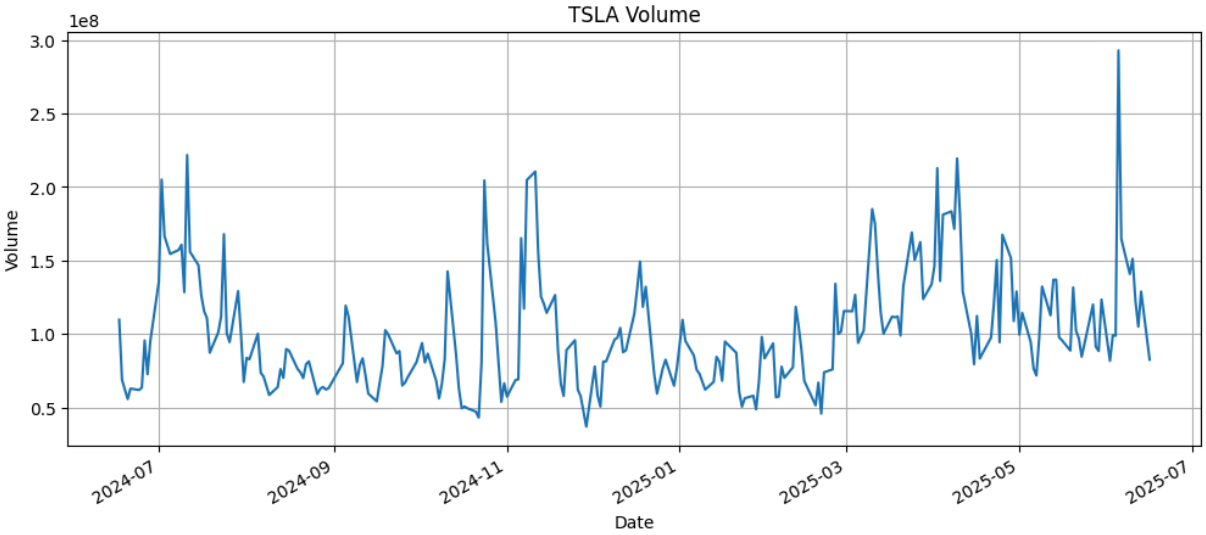
Stage: Inspect Raw Data

Objectives

- Verify the integrity of the raw stock data, including row count, date range, and data types.
- Identify any gaps in trading days and check for missing values or anomalies.
- Visualize key metrics such as 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.



Summary of findings

```
Data Summary:
Row Count: 250
Date Range: 2024-06-17 to 2025-06-16
Missing Values: 0
Anomalies: 1

Anomalies Details:
         date  column           value
1  2025-06-05  volume  292,818,655.00

Summary Statistics:
         open    high     low   close       volume
count     250     250     250     250          250
mean   290.45  297.72  282.84  290.46  100,814,958
std     72.62   74.00   70.21   71.99   39,521,561
min    177.92  183.95  177.00  181.57   37,167,621
25%    230.44  237.30  225.10  231.46   71,340,436
50%    263.05  274.11  257.10  265.41   93,858,026
75%    345.07  351.22  335.92  344.16  120,581,185
max    475.90  488.54  457.51  479.86  292,818,655
```

Conclusion: Comprehensive Analysis and Insights

- **Data Coverage**:
  - **Row Count**: `250` trading days from June 17, 2024, to June 16, 2025, aligning with typical annual trading days (`~252`) adjusted for holidays.
  - **Date Range**: Spans `364` days, including non-trading days, with `250` trading days reflecting U.S. market conventions (e.g., NASDAQ).

Trading Day Observations

- **Trading Days Declared Holiday**:
  - **January 9, 2025**: Declared a National Day of Mourning for former President Carter, resulting in the closure of equity markets.
- **Federal Holidays but still Trading Days**:
  - **October 14, 2024 (Columbus Day)**: Equity markets remained open.
  - **November 11, 2024 (Veterans Day)**: Equity markets remained open.
- This classification validates adherence to equity market calendars, ensuring data integrity by accurately reflecting trading activity on these dates.

- **Data Completeness and Types**:
  - **Missing Values**: None across all columns (`open`, `high`, `low`, `close`, `volume`, `mo5`).
  - **Data Types**: Numerical columns are `float64` for precision; index is `datetime64` for time-series analysis.
- **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 deviations above mean).
      - **IQR**: Exceeds upper bound (Q3 + 1.5 * IQR).
    - **Data Source Check**: Matches NASDAQ; Yahoo Finance shows `~1.8%` lower volume, likely due to adjustment.
    - **Event Context**: Triggered by President Trump's threat to cancel government contracts with Elon Musk's companies, causing a `14.2%` price drop (`$332.05` to `$284.70`).
    - **Action**: Flagged as `is_outlier = True` and retained for modeling.
- **Visualization Insights**:
  - **Closing Price**:
    - Rose from `$181.57` (mid-2024) to `$479.86` (early 2025), then declined to `~$300–$350` by June 2025.
    - Sharp `14.2%` drop on June 5, 2025, with partial recovery by June 13, suggesting event-driven volatility.
  - **Trading Volume**:
    - Right-skewed distribution (median: `~93.9M` shares, mean: `~100.8M` shares).
    - Spike to `~292.8M` shares on June 5, 2025, vs. baseline of `50–150M` shares.
    - Indicates high sensitivity to news, with June 5 as a notable outlier.

Conclusion:
The dataset is robust, covering `250` trading days with no missing values and correct market calendar alignment. A significant volume outlier on June 5, 2025 (`292,818,655` shares), validated by z-score (`4.858`) and IQR, is linked to a major event causing a `14.2%` price drop. Visualizations highlight volatility and recovery trends, with additional spikes suggesting event sensitivity. The outlier is retained for modeling.

Next Steps: Test model sensitivity to the outlier, explore lagged effects, and analyze other high-volume dates for better prediction accuracy.

Date: August 07, 2025

Stage: Model Performance Analysis

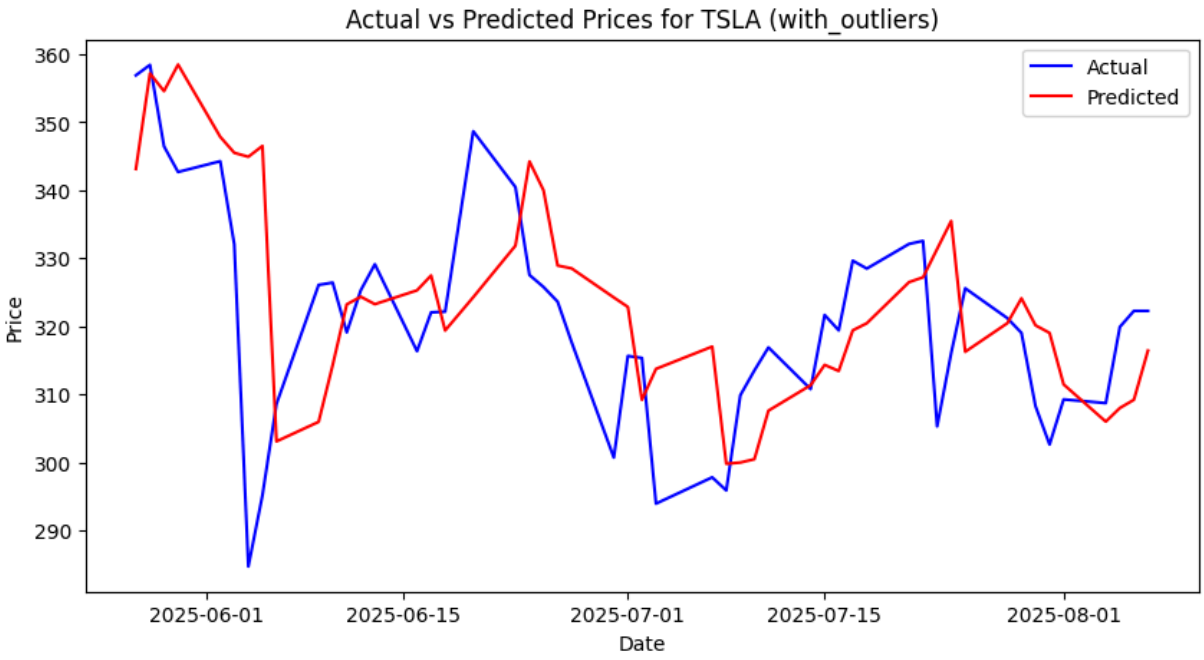
Objectives

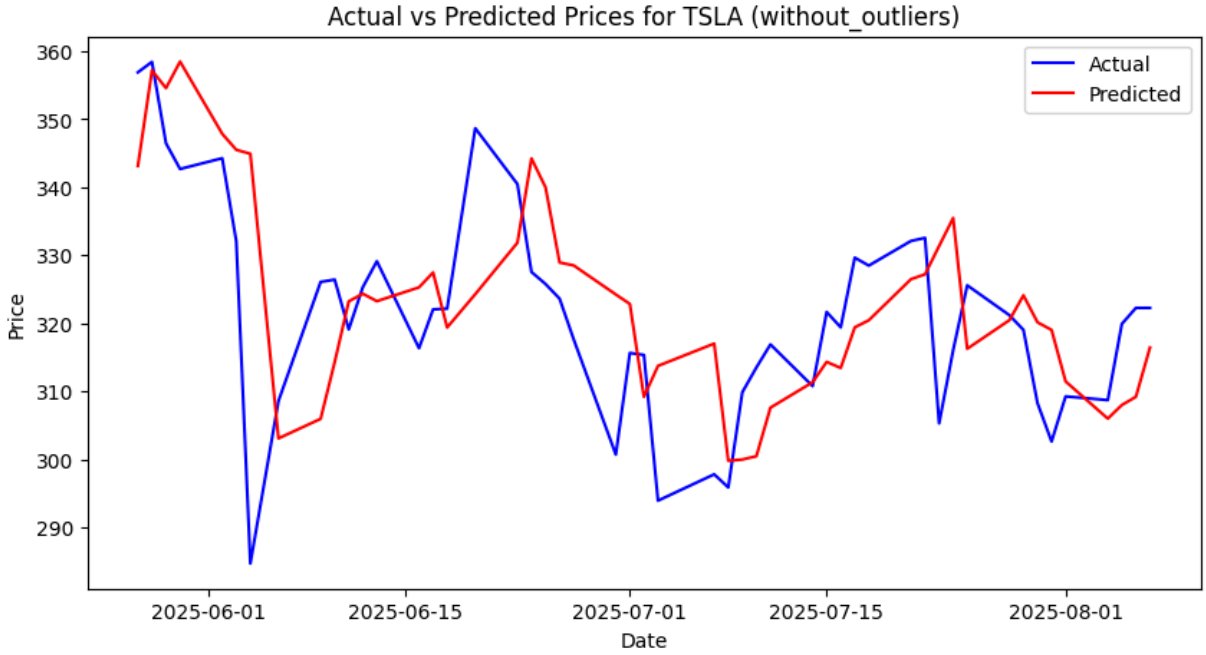
- Evaluate the performance of linear regression models (with and without outliers) using metrics such as RMSE, MAE, and R^2.
- Visualize predictions and residuals to assess model accuracy.
- Analyze the impact of outliers (e.g., June 5, 2025) on model performance.

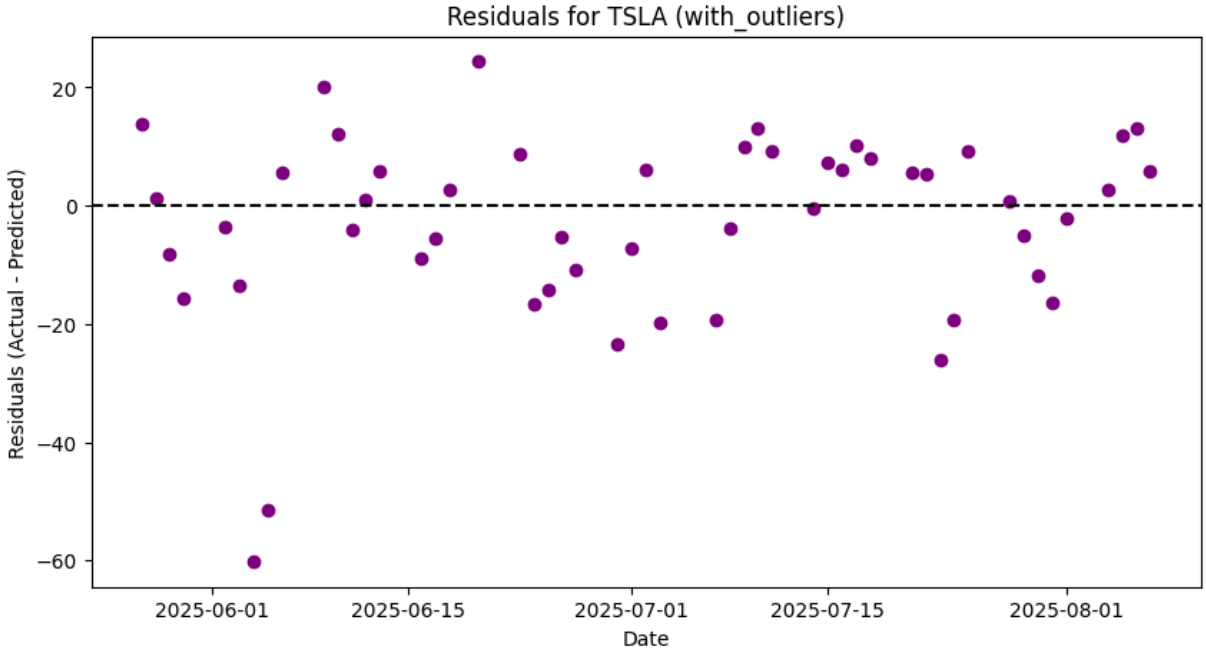
Model Evaluation

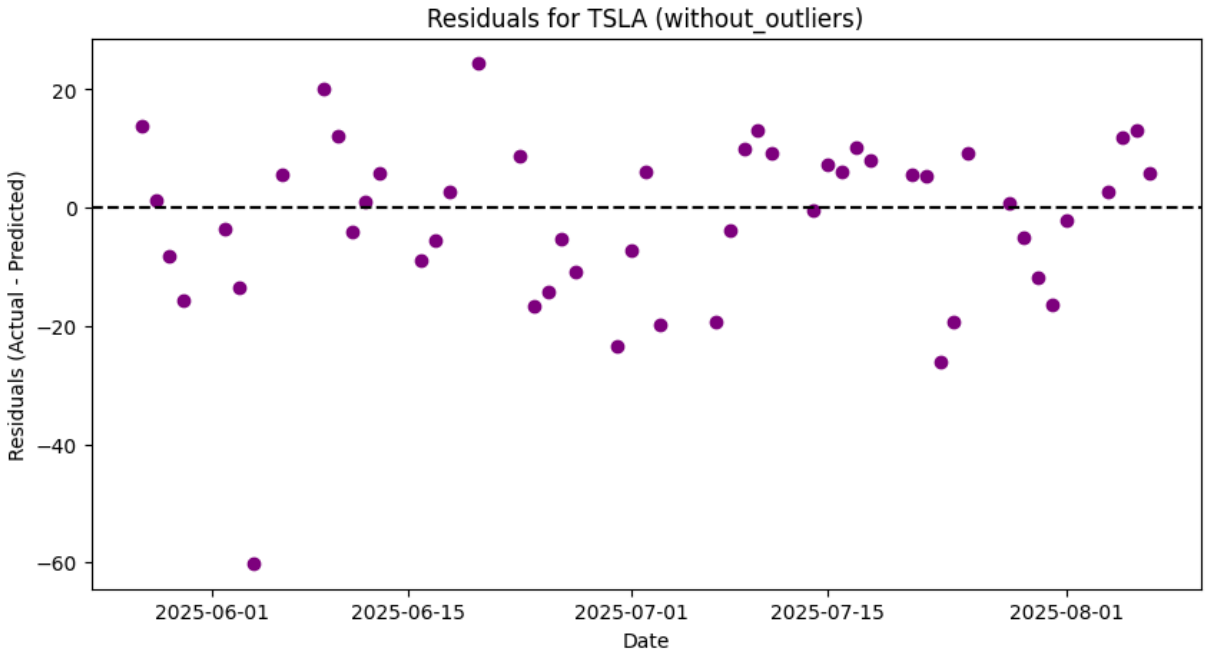
```
Model with outliers:
  RMSE: 15.99
  MAE: 11.63
  R2: -0.07
Model without outliers:
  RMSE: 14.43
  MAE: 10.84
  R2: 0.09
```

Visualization




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


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



Summary of Findings

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.

Recommendations and Next Steps

- **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.