Complex Data Munging & Statistical Modeling on S&P 500 Stock Dataset

Objective

- Prepare and clean messy financial data
- Engineer predictive features
- Fit logistic-regression to predict next-day stock price movement

Dataset

- 5 years daily stock prices of S&P 500 companies
- Source : Kaggle “S&P 500 stock data”

Data Challenges

- Missing values in price and volume data
- Outliers caused by extreme price movements
- Mixed data types requiring careful conversion
- Complex multi-index time series data for many tickers
- Need to align features with target variables for supervised learning

Data Cleaning & Normalization

- Imputed missing values by forward and backward filling within each stock group
- Detected outliers using z-score and capped extreme values to reduce skew
- Converted ticker symbols to categorical dtypes and volumes to integers
- Created MultiIndex of (Ticker, Date) and pivot tables for normalization
- Result: Clean, consistent, and well-structured dataset ready for feature engineering

Feature Engineering

- Created polynomial features (e.g., squared close price) to capture nonlinearities
- Generated interaction term: close price x volume to model combined effects
- Calculated rolling 5-day mean of closing prices for trend smoothing
- Constructed lagged target variable for next-day price movement classification
- Applied advanced pandas techniques: groupby + transform, custom apply, Categorical dtypes

Statistical Modeling Results

- Model: Logistic regression predicting binary target (price up/down next day)
- Parameter estimates and statistical significance reported with confidence intervals
- Key variables (polynomials and interaction terms) found significant via Wald tests
- Odds ratios showed interpretables impact of features on likelihood of price increases
- Visualizations highlighted model-predicted probabilities across close prices

Insights & Conclusion

- Multi-step cleaning vital for preparing noisy finance data
- Feature engineering improved model’s ability to capture price dynamics
- Logistic regression provided interpretable probabilistic outputs and hypothesis testing
- Framework applicable to other financial datasets and modeling problems

Acknowledgement & References

- Data source: Kaggle S&P 500 Stock Dataset (all_stock_5yr.csv)
- Libraries: pandas, numpy, statsmodels, matplotlib
- Project developed using Python Jupyter notebooks for reproducibility