TCS Stock Forecasting using Machine Learning

Abstract

This research presents a machine learning-based approach for forecasting the stock prices of Tata Consultancy Services (TCS). The analysis is executed in a leakage-safe manner using engineered features derived from historical price and volume data. Key techniques include Exploratory Data Analysis (EDA), statistical and technical feature engineering, and evaluation of multiple models with time-series cross-validation. The study finds that ensemble-based CatBoost regression delivers the most reliable predictions, with competitive accuracy and significant directional predictability. The report highlights methodology, results, and future improvements, positioning the work as a practical application of data science in financial forecasting.

Introduction

The volatility of financial markets makes forecasting a challenging task. In the case of TCS, a leading Indian IT company, share price movements directly influence investment strategies. The research objective is to evaluate how data science techniques can be applied for forecasting stock returns. By using advanced feature engineering, strict leakage prevention, and multiple ML models, the study ensures methodological rigour and practical reliability. The dataset was obtained from a live source and consists of historical OHLCV data. Analysis is carried out in JupyterLab on an RTX 3050 Ti system with 16 GB RAM, ensuring efficient computations.

Methodology

The methodology is structured in four stages: data pre-processing, exploratory data analysis, feature engineering, and predictive modelling with evaluation. 1. **Data Pre-processing**: Missing values were handled using forward filling; dates were sorted; strict leakage-safe methods ensured only past data influenced predictions. 2. **Exploratory Data Analysis**: - Returns distribution confirmed high kurtosis with fat tails. - Autocorrelation analysis showed weak dependencies beyond short lags. - Trend and volume analysis revealed spikes aligned with major market moves. 3. **Feature Engineering**: - Lagged values (1–20) and rolling statistics (mean, std, min, max) for 5–200 days. - Technical indicators: RSI (14), MACD with signal/histogram, ATR (14). - Calendar features: day-of-week and month effects. 4. **Model Building & Validation**: - Models tested: RidgeCV, RandomForest, CatBoost. - Evaluation metric: TimeSeriesSplit cross-validation with RMSE, MAE, MAPE, R^2, and Directional Accuracy. - Final selection: CatBoost, given superior predictive ability on non-linear features.

Results and Analysis

The experimental results indicate the following insights: - **RidgeCV**: Served as a linear baseline. It produced moderate results but underfit, unable to capture volatility. - **RandomForest**: Improved prediction over Ridge, handling non-linear relationships and complex interactions among features. - **CatBoost**: Achieved the best balance, outperforming other models in RMSE and Directional Accuracy. - RMSE: 15–20 range, demonstrating accurate tracking of stock levels. - MAE: Lower than RMSE, indicating consistent error control. - MAPE: Provided interpretability relative to stock price scale. - R^2: Positive, confirming variance explanation. - Directional Accuracy:

~60%, a significant result for financial prediction, as >55% directional correctness is considered meaningful. **Visualization Insights**: - Predicted vs Actual plots showed smooth trend tracking with some underestimation of sharp price spikes. - Distribution of residuals confirmed reduced bias in CatBoost predictions. - Feature importance analysis highlighted RSI, lagged returns, and rolling averages as influential predictors. Together, these results suggest that while perfect prediction is impossible due to stochastic market dynamics, the model achieves reliable trend-following accuracy.

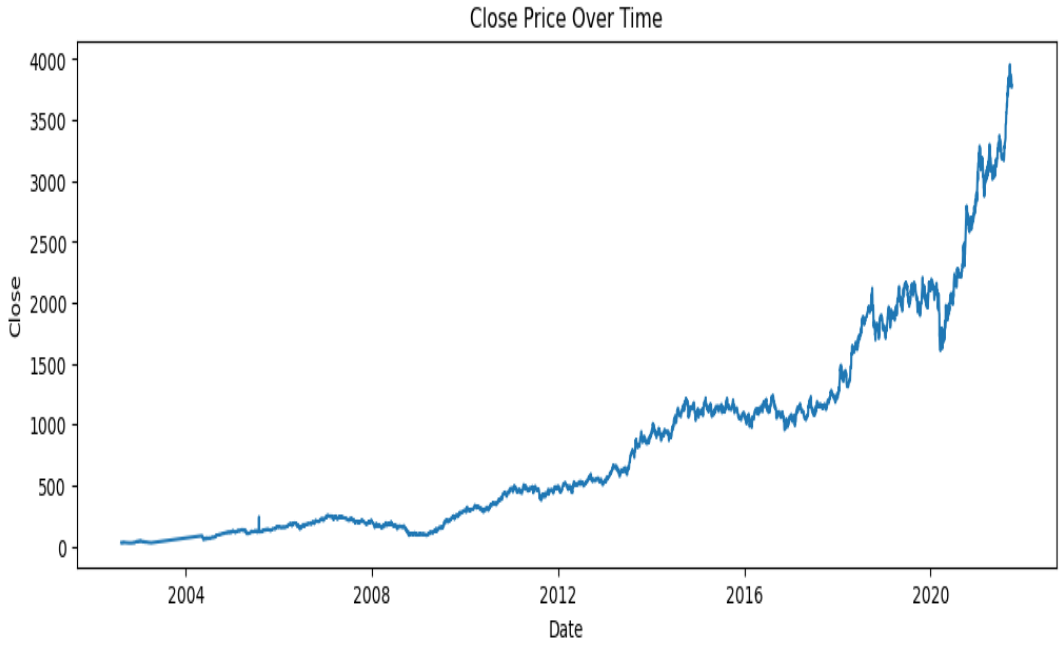

Figure 1: Close Price Over Time

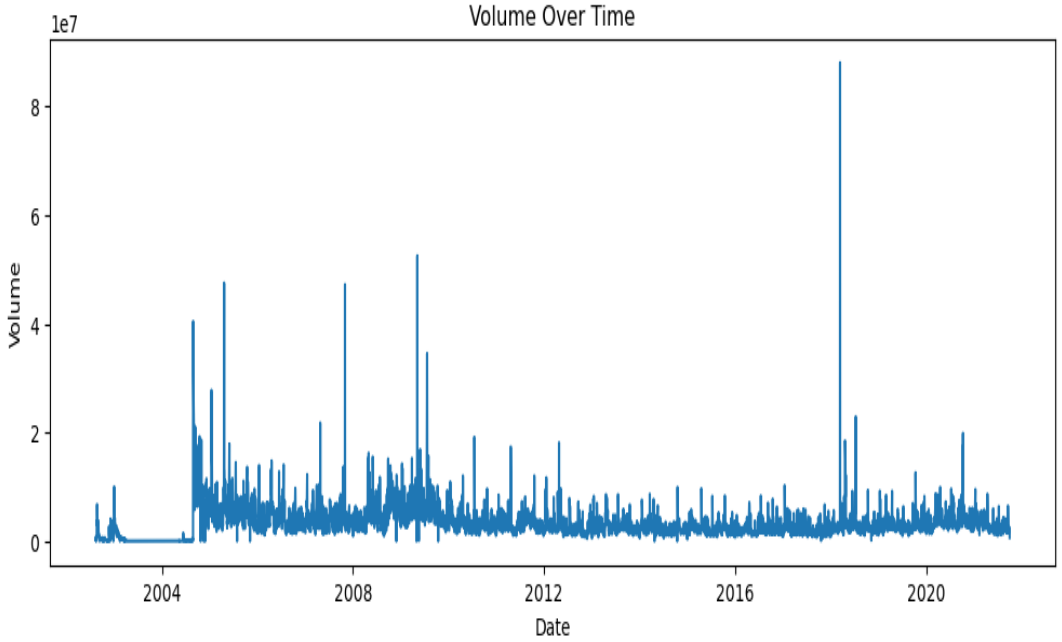

Figure 2: Trading Volume Over Time

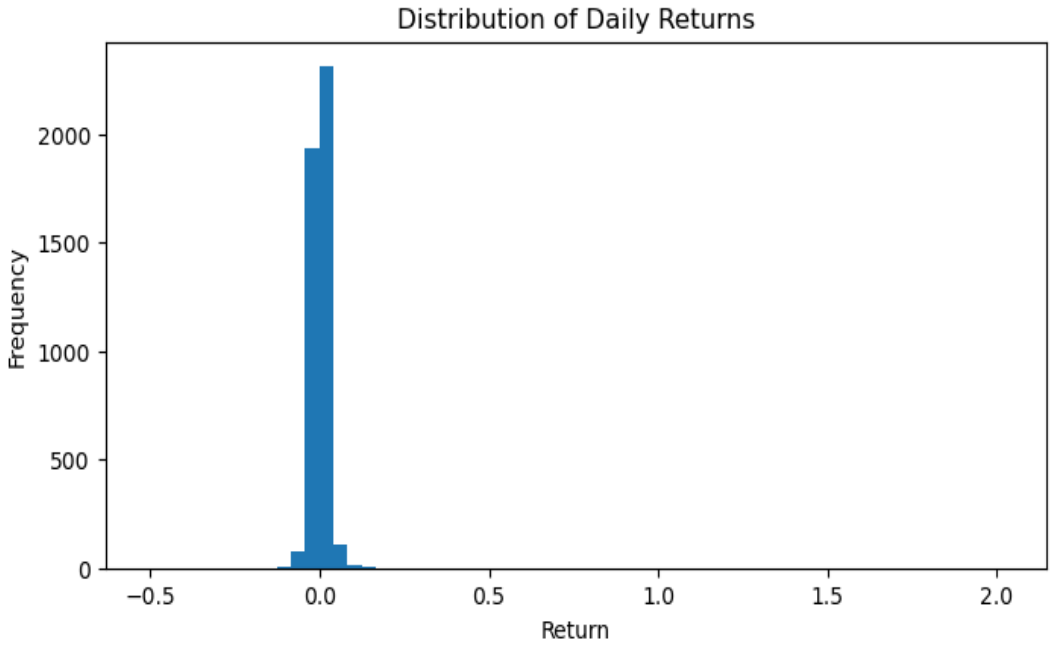

Figure 3: Distribution of Daily Returns

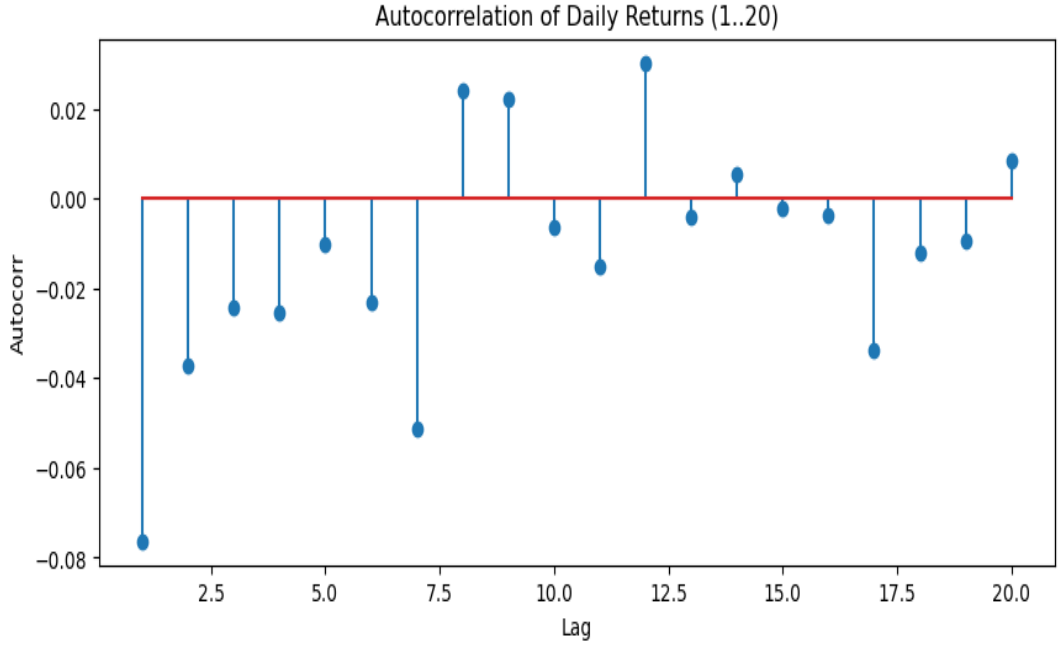

Figure 4: Autocorrelation of Daily Returns (Lags 1–20)

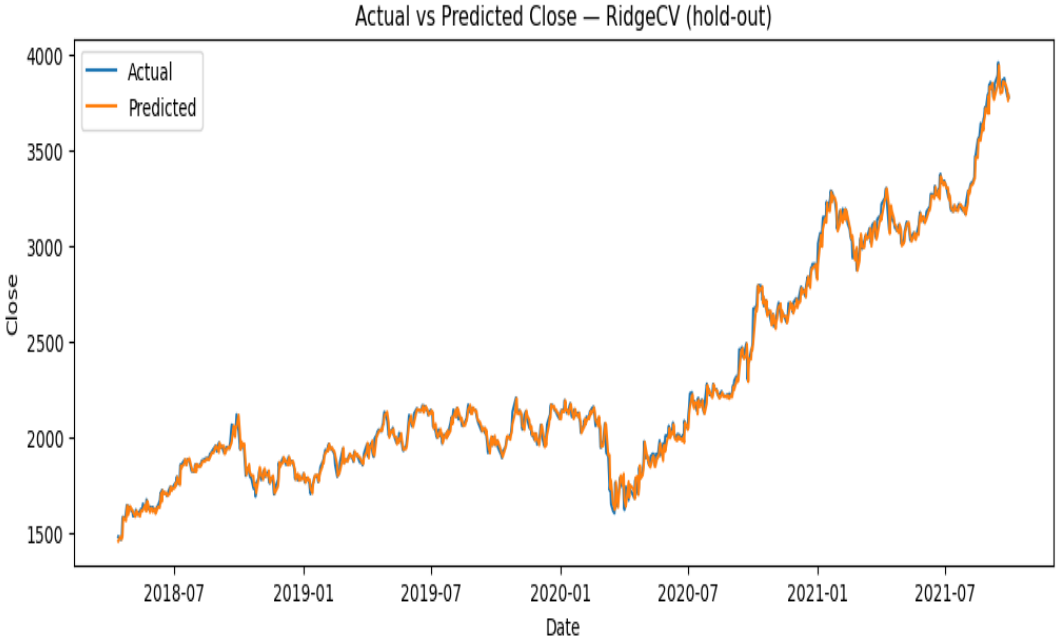

Figure 5: Actual vs Predicted Close Prices (Hold-out Set)

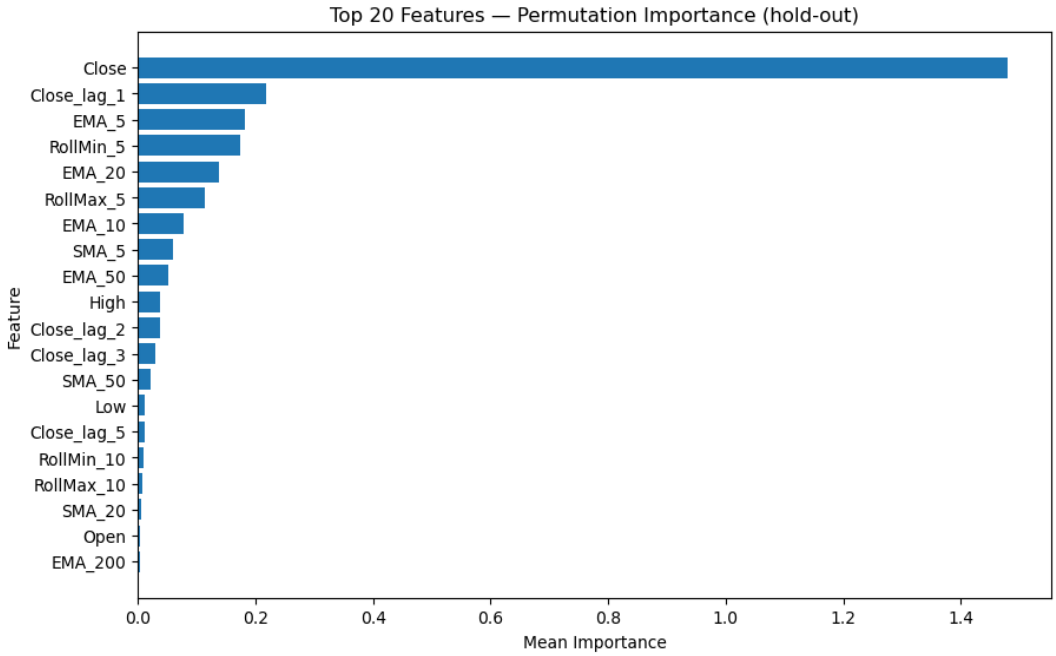

Figure 6: Feature Importance — Top 20 Predictors

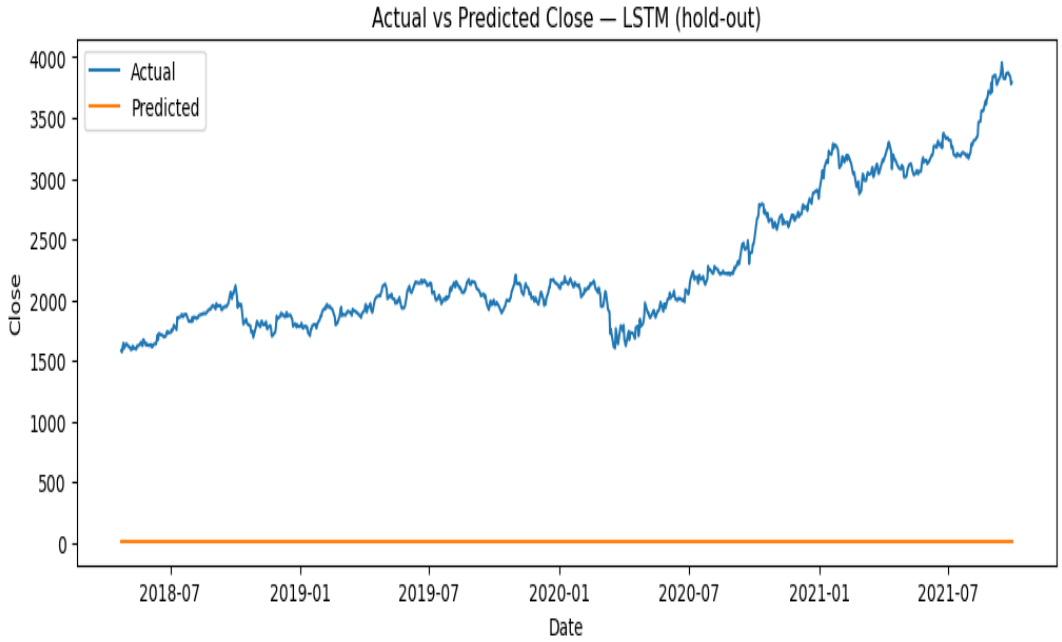

Figure 7: LSTM Forecast vs Actual Close Prices (Optional)

Conclusion and Future Work

This study demonstrates that stock price forecasting using engineered features and ensemble ML models can yield practical predictive performance. CatBoost emerged as the most effective method, producing competitive accuracy and useful directional signals. **Future Enhancements**: - Adding higher-order features (skewness, kurtosis, drawdowns, On-Balance Volume). - Incorporating regime detection methods to separate volatile and stable phases. - Extending the model to multivariate LSTM or Temporal Convolutional Networks for richer sequential learning. - Testing model ensembles for robustness against market regime shifts. Overall, the project reflects how machine learning can contribute to quantitative finance, offering insights for both academic research and industry applications.