Technical Report: AI Agent for Stock Return Prediction

CIS 9660 – Data Mining
August 13, 2025
Heeje Yoo

1. Business Problem and Importance

This project addresses the Stock Price Prediction problem by developing a regression-based AI agent to predict the next-day percentage return of a stock ticker. The goal is to provide an educational tool demonstrating an end-to-end machine learning workflow—from data collection and feature engineering to model training and deployment in a real-time, interactive web app. The agent applies data mining principles to a complex, real-world financial problem, with a disclaimer that it is not financial advice.

2. Data Source and Preprocessing

The primary data source is the Yahoo Finance (yfinance) API, providing historical daily stock data. The preprocessing pipeline (in Regression_AI_Agent.ipynb) includes:

- **Data Loading** – Last two years of daily price data (Open, High, Low, Close, Volume) for a chosen ticker (e.g., TSLA).
- **Feature Engineering** – Seven technical indicators (e.g., SMAs, RSI, volatility, MACD) derived from raw prices.
- **Data Splitting** – Chronological 80% training / 20% hold-out split to avoid leakage.
- **Missing Values** – NaNs from rolling calculations are dropped to maintain model integrity.

3. Model Selection Process and Results

Two regression models were evaluated: Ridge Regression (enhanced linear baseline) and Random Forest Regressor. Using 5-fold cross-validation on training data, Ridge was chosen for its simplicity, interpretability, and stable performance. On the hold-out set, Ridge achieved MAE = 3.30%, RMSE = 4.52%, and $R^2 = 0.002$.

4. Key Insights and Recommendations

Predicting short-term returns from technical data alone proved extremely difficult, consistent with the efficient market hypothesis. The near-zero R^2 shows the features explain negligible variance. This AI agent should be used strictly for educational purposes to illustrate a data science workflow, not for financial advising.

5. Limitations and Future Improvements

- **Limitations** – Relies only on technical indicators, excluding fundamentals, macroeconomic variables, and sentiment data.
- **Future Improvements** – Incorporate broader features (fundamentals, NLP-based sentiment), and explore more advanced models like Gradient Boosting (XGBoost) or LSTMs.