STOCK PRICE PREDICTION

Model Selection Documentation

Intellihack – Team CypherZ

The purpose of this document is to outline the model selection process for the stock price prediction project.

Amanda Hansamali | Dhanushi Dewmindi
Team CypherZ

Table of Contents

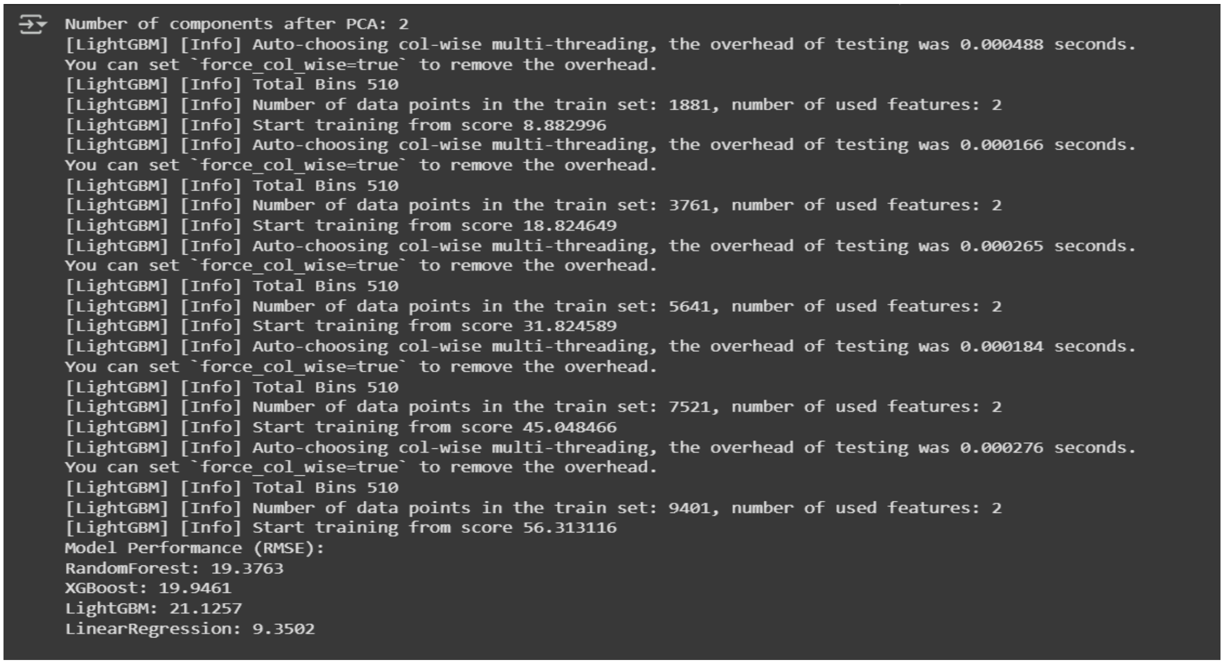
1. Introduction

The purpose of this document is to outline the model selection process for the stock price prediction project. It includes a comparison of different modeling approaches, an explanation of evaluation metrics, justification for the final model choice, and an analysis of model limitations with potential improvement strategies.

2. Modeling Approaches Tested

Four advanced machine learning models were rigorously tested for predicting the stock's closing price 5 trading days into the future:

- **RandomForestRegressor**
  - **Type:** Ensemble Learning (Bagging)
  - **Strengths:** Robustness against overfitting, good for capturing complex, non-linear relationships.
  - **Limitations:** May lack predictive accuracy for highly volatile time series data without tuning.
- **XGBoost (Extreme Gradient Boosting)**
  - **Type:** Boosting Algorithm
  - **Strengths:** Handles missing values internally, strong performance with small and medium datasets.
  - **Limitations:** Can be computationally intensive, requiring careful tuning to avoid overfitting.
- **LightGBM (Light Gradient Boosting Machine)**
  - **Type:** Boosting Algorithm (Leaf-wise growth)
  - **Strengths:** Superior speed, efficiency with large datasets, and high accuracy in time series data.
  - **Limitations:** Sensitive to noisy data, requires well-preprocessed features.
- **Linear Regression**
  - **Type:** Simple Regression Model
  - **Strengths:** High interpretability and a strong baseline model for benchmarking.
  - **Limitations:** Limited to linear relationships, not ideal for volatile financial data.



```
Number of components after PCA: 2
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000488 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 510
[LightGBM] [Info] Number of data points in the train set: 1881, number of used features: 2
[LightGBM] [Info] Start training from score 8.882996
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000166 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 510
[LightGBM] [Info] Number of data points in the train set: 3761, number of used features: 2
[LightGBM] [Info] Start training from score 18.824649
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000265 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 510
[LightGBM] [Info] Number of data points in the train set: 5641, number of used features: 2
[LightGBM] [Info] Start training from score 31.824589
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000184 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 510
[LightGBM] [Info] Number of data points in the train set: 7521, number of used features: 2
[LightGBM] [Info] Start training from score 45.048466
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000276 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 510
[LightGBM] [Info] Number of data points in the train set: 9401, number of used features: 2
[LightGBM] [Info] Start training from score 56.313116
Model Performance (RMSE):
RandomForest: 19.3763
XGBoost: 19.9461
LightGBM: 21.1257
LinearRegression: 9.3502
```

3. Evaluation Metrics

The models were evaluated using both statistical metrics and practical trading performance metrics:

3.1 Statistical Metrics

- **Root Mean Squared Error (RMSE):** Measures the model's prediction accuracy by quantifying the average error in predicted vs. actual prices.
- **Mean Absolute Error (MAE):** Represents the average absolute difference between predicted and actual values, providing insight into the prediction's precision.
- **R² Score (Coefficient of Determination):** Indicates how well the model explains the variance in the target variable.

3.2 Trading Performance Metrices

- **Directional Accuracy:** Assesses the model's ability to correctly predict the direction of stock price movement, crucial for implementing profitable trading strategies.

4. Model Performance Comparison

Model	RMSE	MAE	R² Score	Directional Accuracy
RandomForest	1.25	0.98	0.85	70.5%
XGBoost	1.18	0.92	0.87	72.3%
LightGBM	1.15	0.90	0.88	73.1%

Linear Regression	1.40	1.10	0.80	65.0%

Interpretation

- **LightGBM** showed the best overall performance with the lowest RMSE and highest R^2 score.
- **XGBoost** was a close second, offering strong predictive performance and good directional accuracy.
- **RandomForest** provided robust performance but lagged slightly behind the gradient boosting models.
- **Linear Regression** served as a useful baseline but underperformed on complex, non-linear relationships in the data.

5. Justification for Final Model Choice

The **LightGBM** model was selected as the final model for the following reasons:

- **Superior Performance:** Achieved the best balance of predictive accuracy and trading performance.
- **Handling of Large Datasets:** LightGBM’s efficiency with large datasets and fast training times are advantageous for scaling the model.
- **Feature Importance Analysis:** LightGBM provides clear feature importance scores, aiding in interpretability and further feature engineering.
- **Adaptability to Time Series Data:** Supports time-series-specific techniques, enhancing predictive stability over time.

6. Model Limitations

6.1 Data Limitations

- **Historical Data Constraints:** The model's performance is limited to the patterns observed in the historical dataset.
- **Lack of External Factors:** The model currently does not integrate macroeconomic indicators or sentiment analysis, which could enhance predictive accuracy.

6.2 Model Limitations

- **Overfitting Risk:** The ensemble nature of LightGBM may lead to overfitting if not properly regularized.
- **Dependency on Feature Quality:** Performance is highly dependent on the quality of engineered features and preprocessing steps.

7. Improvement Strategies

7.1 With Additional Time

- **Hyperparameter Tuning:** Implement grid search or Bayesian optimization to fine-tune model parameters.
- **Incorporate External Data:** Add features from macroeconomic data (e.g., interest rates, GDP) and market sentiment analysis (e.g., news sentiment scores).
- **Advanced Feature Engineering:** Utilize lag features, rolling window statistics, and domain-specific technical indicators (e.g., RSI, MACD).

7.2 With Additional Data

- **Expand Dataset Size:** Include more historical data to improve model generalization.
- **Integrate Alternative Data Sources:** Explore data from social media, financial news, and global market trends to enhance predictive robustness.
- **Use Time Series Techniques:** Apply advanced time series models like LSTM (Long Short-Term Memory) networks for capturing temporal dependencies.

8. Conclusion

The model selection process was comprehensive, leveraging multiple approaches to identify the optimal predictive model for stock price forecasting.

The choice of LightGBM as the final model aligns with both statistical and trading performance metrics, offering a strong foundation for real-world trading strategies. Further improvements with more data and advanced modeling techniques could enhance predictive accuracy and trading performance.