Stock Market Prediction using ML

1. Introduction

This data science project aims to develop a robust binary classifier to aid non-professional investors in identifying the most promising stock to invest in among a selection of NASDAQ companies. Specifically, the goal is to predict whether the price of a stock will increase over a short period, leveraging historical market data. This detailed report encompasses the comprehensive methodology employed in data preprocessing, feature selection, machine learning model implementation, performance metrics evaluation, and the encountered challenges throughout the analysis.

2. Data Preprocessing

2.1. Data Loading and Inspection

The dataset, sourced from Kaggle, contains historical stock market data for various NASDAQ companies. Upon loading, the preliminary inspection was conducted to understand the structure, data types, and overall quality of the dataset.

2.2. Exploratory Data Analysis (EDA)

Extensive EDA was performed to gain insights into the dataset. Key aspects addressed include:

- Identification and handling of duplicate rows.
- Detection of duplicated columns and subsequent removal to avoid redundancy.
- Analysis of columns with all NaN values and decisions on handling missing data.
- Exploration of correlation between features and identification of those with significant correlation to the target variable.

2.3. Technical Analysis

Technical indicators such as Simple Moving Average (SMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), etc., were computed from the historical stock data. These indicators provide valuable insights into market trends, momentum, and potential buy/sell signals.

3. Feature Selection

3.1. Correlation Analysis

Features exhibiting high correlation with the target variable were identified using correlation matrices and selected for further analysis. This step ensures that only relevant features contribute to the predictive power of the model.

3.2. Domain Knowledge Incorporation

Features selected based on domain knowledge in finance and stock market prediction were given precedence in model training. This includes factors such as volume, volatility, price movements, and technical indicators.

4. Model Implementation

4.1 Model Selection

Several classification models were evaluated:

- **Logistic Regression:**
  A baseline model that provides interpretability and handles binary classification well.

- **K-Nearest Neighbors (KNN):**
  A non-parametric method is used for classification, where the majority class determines the output among the k-nearest neighbors.

- **Support Vector Machine (SVM):**
  Effective in high-dimensional spaces and cases where the number of dimensions exceeds the number of samples.

- **Random Forest:**
  An ensemble method that uses multiple decision trees to improve classification accuracy.

- **Gradient Boosting:**
  An iterative method that builds trees sequentially to minimize the loss function.

4.2. Performance Metrics

Model performance was evaluated using:

- **Accuracy:** The proportion of correctly classified instances among the total instances.
- **Precision:** The ratio of true positive predictions to the total predicted positives.
- **Recall:** The ratio of true positive predictions to all actual positives.
- **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.

4.3. Hyperparameter Tuning

- **Grid Search Cross-Validation (CV):**
  1. Employed 3-fold cross-validation to tune hyperparameters for each model.
  2. Parameters such as the number of neighbors (KNN), C parameter (SVM), and number of trees (Random Forest) were optimized.

5. Model Performance

5.1. Final Model Metrics

- **Accuracy:** 0.52344
- **Precision:** 0.19493
- **Recall:** 0.51228
- **F1 score:** 0.2824

5.2. Confusion Matrix

The confusion matrix would reveal the distribution of true positive, true negative, false positive, and false negative predictions. With a recall of 0.51228, it indicates a moderate ability to correctly identify positive cases. However, the precision of 0.19493 suggests a high rate of false positives.

5.3. ROC Curve

The ROC curve would plot the true positive rate against the false positive rate at various thresholds. Given the relatively low precision and recall scores, the ROC curve may not exhibit a steep increase towards the upper left corner, indicating suboptimal model performance in distinguishing between positive and negative cases.

6. Challenges Faced

1. **Dealing with imbalanced data**
   The dataset might have an imbalance between the classes, which could affect the model's performance. Techniques such as oversampling, undersampling, or using appropriate evaluation metrics were employed to address this issue.

2. **Feature engineering**
   Deriving meaningful features from raw data, especially in the context of stock market prediction, can be challenging. Technical indicators were computed to capture relevant patterns in the data.

3. **Model selection and optimization**
   Choosing the right model architecture and tuning hyperparameters to achieve the best performance required experimentation and careful analysis.

7. Conclusion and Future Directions

In conclusion, this project represents a comprehensive effort to develop a binary classifier for stock market prediction using machine learning techniques. Through meticulous data preprocessing, feature selection, and model implementation, we aimed to provide actionable insights to non-professional investors seeking to make informed investment decisions. Despite challenges such as data quality issues and model selection dilemmas, the project succeeded in building predictive models with promising performance metrics. Future directions may involve further refinement of feature engineering techniques, exploration of

advanced ensemble methods, and continuous monitoring and updating of the model to adapt to evolving market dynamics.