BIG DATA (CSGY-6513)

Section D, Fall '23

Team Members

Anubhav Ghildiyal (ag8766)

Noel Nebu Panicker (nnp5666)

Yogya Sharma (ys5250)

Project Proposal

Title: Predictive Analysis of Nasdaq Stock Exchange Closing Prices Using Big Data Technologies

1. Project Abstract

The stock market, especially during its closing minutes, poses a challenging and high-stakes environment for traders and market participants. This project seeks to leverage advanced data analytics and big data technologies to predict the closing price movements of Nasdaq-listed stocks during the crucial final ten minutes of trading. By developing sophisticated predictive models and harnessing the power of big data tools, our goal is to contribute to market efficiency and provide valuable insights to investors.

2. Problem Statement and Objectives

Problem Statement:

Predicting the closing price movements of Nasdaq-listed stocks during the last ten minutes of trading is a complex task that requires advanced data analysis techniques and big data technologies.

Objectives:

- Predict the closing price movements of Nasdaq-listed stocks during the last ten minutes of trading.

- Develop models that merge traditional order book data and auction book data to enhance predictive accuracy.

- Utilize big data technologies such as Hadoop, Apache Spark, and related tools for efficient processing and analysis of large-scale stock market data.

3. Data Source

Data Source Name: Optiver Trading at the close

Data Link: https://www.kaggle.com/competitions/optiver-trading-at-the-close/data?select=train.csv

Data File Size: 646.7 MB

Approximate Number of Records: 500000+

4. Proposed Technologies

Programming Language: Python

Big Data Technologies:

- Hadoop: Used for storing and managing large volumes of structured and unstructured data.

- Apache Spark: Utilized for real-time processing and analysis of large datasets.

- Apache Hive: Provides an SQL interface for querying and analyzing the data stored in Hadoop.

- Apache Kafka: Employed for real-time data streaming and processing.

5. Methodology

Data Preprocessing:

- Cleanse, transform, and preprocess raw stock market data.

- Handle missing values, outliers, and inconsistencies.

Feature Engineering:

- Extract relevant features from order book and auction data, such as bid-ask spreads, trading volumes, and volatility indicators.

Model Development:

- Utilize machine learning algorithms (e.g., XGBoost, Random Forest) and deep learning techniques (e.g., LSTM networks) for predictive models.

- Train models using historical data and validate their performance.

Real-time Data Streaming:

- Use Apache Kafka to stream real-time market data for dynamic model adjustment.

Evaluation:

- Evaluate models using accuracy, precision, recall, and F1-score.

- Perform backtesting and simulation to assess the models' performance in a real trading scenario.

6. Expected Outcomes

- Predictive models capable of accurately forecasting Nasdaq-listed stock closing price movements during the last ten minutes of trading.

- Insights into supply and demand dynamics, price adjustments, and trading patterns during the critical closing auction period.

- Enhanced understanding of big data technologies and their applications in financial forecasting and market analysis.

7. Conclusion

By leveraging big data technologies and advanced analytics, this project aims to provide valuable insights into stock market behavior during the intense final minutes of trading. The project's outcomes can be valuable for traders, investors, and market analysts, enhancing their decision-making processes and contributing to a deeper understanding of stock market dynamics.