Title:

Initial Plan: Enhancing Stock Market Predictions through Integration of Market and Financial Data

Author: Artem Gavryushenko
Supervisor: Sylwia Polberg

Project Description:

In the rapidly evolving financial markets, the ability to accurately predict stock performance is invaluable for investors (Gandhmal and Kumar 2019), analysts, and financial institutions. Traditional stock market predictions have relied heavily on historical market data, often overlooking the profound impact of company-specific financial health on stock performance. With the availability of extensive financial datasets, there's a compelling opportunity to enhance prediction models by integrating detailed financial data (SEC.gov | EDGAR Application Programming Interfaces 2022) with traditional market indicators.

The primary challenge lies in the effective integration and ways to analyze of two distinct data sets: market data and company-specific financial data. Market data includes historical stock prices, trading volumes, and other market indicators. Whilst, company-specific financial data includes earnings reports, balance sheets, and cash flow statements, typically sourced from the company filings.

The hypothesis is that the combination of market data and detailed financial information from companies will lead to improved prediction accuracy. Data will be sourced from reputable platforms such as the UC Irvine repository (Datasets - UCI Machine Learning Repository 2024), Kaggle(Find Open Datasets and Machine Learning Projects | Kaggle 2024), and from the SEC (SEC.gov | EDGAR Application Programming Interfaces 2022) for company financial data, addressing the gap in available datasets combining both types of data. The project will involve comparing the performance of various algorithms, conducting cross-validation to ensure model robustness, and performing feature analysis to understand the impact of each feature on the prediction models.

This project addresses the lack of comprehensive datasets that combine these two data types and the need for predictive models that can interpret this integrated data effectively. By focusing on this aspect of the problem, the project aims to make a meaningful contribution to the field of stock market prediction, demonstrating how the integration of market and financial data can lead to more accurate and actionable insights.

Aims and Objectives

Aim:
Develop an integrated predictive model that enhances the accuracy of stock market predictions by combining market data with company-specific financial data, employing advanced techniques for analysis and feature importance evaluation.

Subaims:

1. **Data Aggregation and Integration:**
   Integrate stock market data with company-specific financial data from sources like SEC Edgar, UC Irvine repository, and Kaggle.
   Risk: Mismatch between the accessibility and compatibility of stock market data and company financial data, leading to potential issues in data integration and completeness.
   Mitigation: Apply a selective approach in choosing companies and features for analysis to ensure data integrity. Prioritize data sources that offer complete and compatible datasets, and reduce the number of companies or features if necessary to maintain data completeness.
2. **Algorithms Development and Evaluation:**
   Implement and compare the performance of multiple algorithms on the pure-market and integrated datasets, determining the most effective approach for stock prediction.
   Risk: False results due to the misinterpretation of the hypothesis validity because of data integration complexities or inappropriate choices in the model.
   Mitigation: Prepare for a broad initial survey of algorithms and focus on model transparency and interpretability to ensure the validity of hypothesis testing.
3. **Feature Analysis:**
   Conduct a feature analysis to determine the importance and impact of individual features from market and financial data on the prediction accuracy.
   Risk: Overemphasis or misinterpretation of feature importance due to data complexity or integration issues, resulting in misleading insights.
   Mitigation: Regularly cross-check results against domain knowledge and empirical evidence to avoid overfitting or misinterpretation.
4. **Develop Solution for Automated Financial Data Integration:**
   Develop a solution that automates the retrieval of company financial data from SEC API Edgar, addressing the lack of available datasets that combine stock market and company financial data.
   Risk: Technical issues with API stability that could disrupt data retrieval.
   Mitigation: Implement appropriate time-outs with the API; batch data retrieval with offline storage and access.

Feasibility

This project is tailored to address the specific problem of enhancing stock market predictions through the integration of market data and company-specific financial data within the constraints of available data, computational resources, and time. The focus will be on selecting a manageable yet representative set of stocks and financial indicators, ensuring that the problem is addressed comprehensively but remains feasible within the project's scope. The use of API Edgar for financial data and established repositories for market data ensures that the data is accessible and manageable. The project will utilize publicly available data from open governmental sources, ensuring ethical compliance. The potential challenge of hardware performance for training and evaluating multiple classifiers on large datasets will be mitigated by carefully selecting sample sizes.

Work Plan

Spring Week 1-2:

- Initial plan drafting and meeting with the supervisor.
- Collection and preliminary analysis of financial and market data.

Spring Week 3-4:

- Implement initial prediction algorithms and perform preliminary performance comparison.
- Begin feature analysis to identify key indicators from the combined dataset.
- Weekly supervisor meetings.

Spring Week 5-6:

- Deepen feature analysis, focusing on the impact of financial data on prediction accuracy.
- Optimize algorithms based on cross-validation outcomes and feature importance.
- Weekly supervisor meetings.

Spring Week 7-8:

- Finalize the integration of market and financial data in the prediction models.
- Conduct comprehensive performance comparisons to validate the hypothesis.
- Weekly supervisor meetings.

Spring Week 9-10:

- Develop integration tool.
- Continuation of feature analysis.
- Weekly supervisor meetings.

Spring Week 11:

- Collating findings and report writing.
- Compilation of data visualization results.
- Weekly supervisor meeting.

Spring Week 12:

- Finalizing the report and preparing for submission.
- Review with the supervisor and incorporation of feedback.

References

Datasets - UCI Machine Learning Repository. 2024. Available at: https://archive.ics.uci.edu/datasets [Accessed: 5 February 2024].

Find Open Datasets and Machine Learning Projects | Kaggle. 2024. Available at: https://www.kaggle.com/datasets [Accessed: 5 February 2024].

Gandhmal, D.P. and Kumar, K. 2019. Systematic analysis and review of stock market prediction techniques. *Computer Science Review* 34, p. 100190. doi: 10.1016/j.cosrev.2019.08.001.

SEC.gov | EDGAR Application Programming Interfaces. 2022. Available at: https://www.sec.gov/edgar/sec-api-documentation [Accessed: 5 February 2024].