WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Trayvonious Pendleton

Student ID # 011205284

10/17/2024

**A. Project Overview**

This proposal describes developing a machine learning-based system to recommend optimal options trading strategies to traders. By analyzing market data, the system aims to assist traders in making data-driven decisions to enhance profitability and minimize risk.

**A.1. Organizational Need**

The problem that needs to be solved is the difficulty traders face in identifying the most effective options trading strategies in a rapidly changing market environment. Traders require a solution to analyze market conditions and recommend suitable strategies to maximize profitability while minimizing risk. The existing manual process could be more efficient and prone to errors, and there is a need to automate the strategy selection process to make it more accurate and timelier.

**A.2. Context and Background**

Options trading involves making complex decisions based on implied volatility, stock behavior, and market conditions. Traders often struggle to continuously monitor these factors and make the best possible decisions in a timely manner. Machine learning has become increasingly prevalent in financial markets, providing a powerful tool to analyze large volumes of data and derive actionable insights. The proposed project aims to leverage machine learning to automate the decision-making process for options trading, providing traders with a data-driven approach to selecting optimal strategies.

**A.3. Outside Works Review**

1. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree" by Guolin Ke et al.** - This work presents LightGBM, an efficient implementation of Gradient Boosting Decision Trees (GBDT) that significantly improves training speed and memory consumption. The paper emphasizes the benefits of GBDT in handling large-scale datasets with high-dimensional features, which is directly applicable to the needs of financial market prediction. LightGBM's efficiency makes it an ideal choice for developing a real-time trading recommendation system, where quick and accurate predictions are crucial (Ke et al., 2017).
2. **"Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning" by Giuseppe Bonaccorso** - Bonaccorso's textbook provides a comprehensive overview of machine learning algorithms, including Gradient Boosting Decision Trees. The book highlights GBDT's ability to combine the strengths of multiple weak learners, resulting in high predictive accuracy and robustness. These characteristics are crucial for building a model that can provide reliable trading recommendations in a dynamic market environment (Bonaccorso, 2017).
3. **"Decision Trees for Intuitive Intraday Trading Strategies" by Prajwal et al.** - This preprint discusses the use of decision tree-based models for developing intraday trading strategies. The authors demonstrate how GBDT can effectively identify profitable trading opportunities by analyzing market trends and indicators. The study supports the use of GBDT in financial applications, showcasing its utility in capturing non-linear relationships between features, which aligns well with the objectives of this project (Prajwal et al., 2024).

**3a**. **Relation to Project Development**

Each of the reviewed works directly supports the development of this project by providing insights into the effectiveness of Gradient Boosting Decision Trees (GBDT) for financial applications, particularly options trading. The work by Guolin Ke et al. highlights the efficiency and scalability of LightGBM, an implementation of GBDT, which is crucial for handling large-scale, high-dimensional datasets typical in financial markets. This aligns with the project’s need for a model that can quickly analyze large volumes of data to provide real-time trading recommendations. Giuseppe Bonaccorso’s textbook offers foundational knowledge on GBDT, emphasizing its high predictive accuracy and robustness, which are essential for developing a reliable trading recommendation system. Lastly, the study by Prajwal et al. demonstrates the successful application of decision tree-based models, including GBDT, in financial trading strategies, showcasing its ability to identify profitable opportunities by capturing non-linear relationships in the data. These works collectively inform the model selection and development process, ensuring the proposed solution is well-suited to meet the project's goals.

**A.4. Solution Summary**

The proposed solution involves developing a supervised machine-learning model that analyzes options market data, implied volatility, and underlying stock behavior to recommend optimal trading strategies. By automating the decision-making process, the solution aims to provide traders with timely and accurate strategy recommendations based on current market conditions.

**A.5. Machine Learning Benefits**

Machine learning will improve traders' decision-making by providing data-driven recommendations based on a thorough analysis of market conditions. This will lead to improved profitability, reduced risk, and greater efficiency for traders, who can focus on higher-level strategic activities rather than manually analyzing market data. The automation provided by machine learning will save time, reduce human error, and ensure that traders have access to insights that are otherwise challenging to derive manually.

.

**B. Machine Learning Project Design**

**B.1. Scope**

**In Scope**:

* Development of a machine learning model for recommending options trading strategies.
* Integration of options market data, implied volatility, and underlying stock behavior.
* User interface for displaying recommended strategies.
* Model evaluation and testing on historical data.

**Out of Scope**:

* Actual financial trading or real-time execution of recommended trades.
* Automated real-money trading based on model recommendations.
* Customization of trading strategies based on individual trader preferences.

**B.2. Goals, Objectives, and Deliverables**

**Goals**:

* Improve traders’ ability to choose the right options strategy in different market conditions.
* Enhance decision-making efficiency and profitability.

**Objectives**:

* Achieve at least 85% recommendation accuracy based on historical data.
* Minimize risk levels by improving the predictability of strategies.

**Deliverables**:

* Trained machine learning model for options strategy recommendation.
* Web interface to present the recommendations.
* Documentation and a user guide for using the model.

**B.3. Standard Methodology**

The development will follow the methodology of CRISP-DM (Cross-Industry Standard Process for Data Mining).

* **Business Understanding**:
  + Identify trading goals, risks, and market behavior factors that impact strategy decisions. Define the specific requirements of the system to provide accurate trading strategy recommendations.
* **Data Understanding**:
  + Collect and explore options market data, stock behavior, and historical performance data. Analyze the data to understand patterns, trends, and anomalies affecting trading outcomes.
* **Data Preparation**:
  + Clean and format data for analysis. Handle missing values, remove outliers, and address any anomalies or inconsistencies in the data. Transform data into a suitable format for model development.
* **Modeling**:
  + Supervised learning models, such as Gradient Boosting Decision Trees (GBDT) or Random Forests, should be applied to train the recommendation system. Experiment with different model parameters to optimize performance.
* **Evaluation**:
  + Test the model's accuracy against historical data and validate results with domain experts. Evaluate the model's performance using accuracy, precision, and recall metrics to ensure it meets the project objectives.
* **Deployment**:
  + Deploy the model in a web interface that provides users with trading strategy recommendations. Ensure the system is user-friendly and accessible, allowing traders to understand and use the recommendations quickly.

**B.4. Projected Timeline**

* **Date 1(January 1, 2025)**: The proposal is accepted.
* **Date 2(January 15, 2025)**: A technical proof of concept is presented.
* **Date 3(February 1, 2025)**: Submitted for review.
* **Date 4**(February 15, 2025) Deliverables finalized.
* **Date 5(March 30, 2025)**: Project delivered.

**Sprint Schedule**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** | **Milestone** |
| 1 | January 1,2025 | January 14, 2025 | Business Understanding | Project plan approved |
| 2 | January 15,2025 | January 28,2025 | Data Collection | Data Collection completed |
| 3 | January 29,2025 | February 5,2025 | Data Preparation | Data pre-processing completed |
| 4 | February 6, 2025 | February 12, 2025 | Model Development | Initial model prototype completed |
| 5 | February 13, 2025 | March 9, 2025 | Evaluation & Testing | Model evaluation completed |
| 6 | March 10,2025 | March 30,2025 | Deployment & Interface | System deployed and operational |

**B.5. Resources and Costs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Resource** | **Description** | **Unit Cost** | **Quantity** | **Cost** |
| Data Scientist Hours | Personnel costs for data analysis and modeling | $100/hour | 150 hours | $15,000 |
| Software Licenses | Software tools for data processing and modeling | $1000/license | 3 licenses | $3,000 |
| Cloud Computing Resources | Compute power for training and testing models | $500 month | 3 months | $1,500 |
| Data Access Subscription | Access to financial data sources | $200 month | 3 months | $600 |
| Deployment Infrastructure | Web hosting and deployment services | $500 month | 3  months | $1,500 |
|  | **Total** |  |  | (Total cost) $21,600 |

**B.6. Evaluation Criteria**

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Ease of Use | |  | | --- | | User interface should be intuitive, allowing traders to easily understand and utilize recommendations with minimal training. | |
| User error rate reduction | Reduction in errors made by traders when selecting options strategies by at least 20% compared to the manual process. |
| Algorithm Efficiency | The algorithm should provide recommendations within 5 seconds for any given query, ensuring timely decision-making for traders. |
| Recommendation Accuracy | Achieve at least 85% accuracy in recommended trading strategies based on historical data. |
| User Satisfaction Rating | Attain a satisfaction rating of 90% or higher from  traders using the system. |
| Risk Reduction | Demonstrate a reduction in risk by at least 20% compared to non-AI-based approaches. |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

**Hypothesis of the Proposed Project:** The hypothesis is that using machine learning to analyze market data, such as implied volatility and underlying stock trends, can provide effective options trading strategies that maximize profitability and minimize risks.

**C.2. Selected Algorithm**

**Selected Algorithm**: Supervised learning, specifically Gradient Boosting Decision Trees (GBDT).

**C.2.a Algorithm Justification**

GBDT was selected for this project because it is well-suited to handle the complexity of financial data and can effectively model relationships between input features and trading outcomes. The ability of GBDT to work with mixed data types and capture non-linear relationships makes it an ideal choice for this application, where multiple market indicators influence trading strategies. According to research by Chen and Guestrin (2016), GBDT has shown strong performance in various predictive modeling tasks, particularly in scenarios with complex data interactions.

**C.2.a.i Algorithm Advantage**

One advantage of using GBDT is its ability to capture complex relationships in the data, leading to higher accuracy in predicting optimal trading strategies. This degree of confidence is further supported by the model's ability to handle different data types and its robustness in preventing overfitting through techniques such as boosting and regularization.

**C.2.a.ii Algorithm Limitation**

A limitation of GBDT is its computational intensity, particularly during the training phase. Training the model requires substantial computational resources, increasing costs, and longer training times. Additionally, GBDT models can be more challenging to interpret than simpler algorithms, which may complicate traders' understanding of the decision-making process.

**C.3. Tools and Environment**

**Hardware**: High-performance server for model training.

**Software**: Python with libraries like sci-kit-learn, pandas, and NumPy for data processing and model development.

**Development Environment**: Jupyter Notebook for exploratory analysis, PyCharm for coding.

**Third-Party Code**: APIs for accessing historical options data (e.g., Yahoo Finance, Quandl).

**C.4. Performance Measurement**

**Training Phase**: Evaluate model performance using accuracy, precision, recall, and F1 score metrics.

**Test Set Evaluation**: Measure against unseen test data to ensure the model's generalizability.

**Validation**: Perform cross-validation to minimize overfitting.

**Metrics**: Focus on the accuracy of strategy recommendations and risk level evaluation.

**D. Description of Data Sets**

**D.1. Data Source**

* Historical options market data from Yahoo Finance API.
* Stock behavior data from Quandl.

**D.2. Data Collection Method**

The data collection for this project will be conducted using API-based data collection from public financial data sources. The APIs will provide real-time options market data, historical stock behavior, and implied volatility information. This method allows for efficient, automated access to up-to-date data, which is critical for developing accurate machine-learning models.

**D.2.a.i. Data Collection Method Advantage**

One significant advantage of using an API-based data collection method is providing access to updated and reliable data directly from trusted platforms. This ensures that the data used for training the model is current and accurate, which is crucial for developing an effective trading recommendation system. Additionally, automating the data collection process reduces the time and effort required to manually gather large volumes of data, allowing the team to focus more on model development and analysis.

**D.2.a.ii. Data Collection Method Limitation**

A limitation of the API-based data collection method is that it may not provide access to proprietary or highly specialized data, which could add more value to the model. Public APIs may have limited coverage, particularly for data available only through premium services or proprietary databases. This limitation could impact the model's ability to generate optimal recommendations, as it may need to include valuable insights available through more exclusive data sources.

**D.3. Quality and Completeness of Data**

To ensure the quality and completeness of the data, the following steps will be taken during data preparation:

* **Formatting:** Convert categorical variables, such as option types or trading indicators, into numerical formats suitable for the machine learning model. One-hot encoding or label encoding will be applied as necessary.
* **Handling Missing Data:** Missing values will be addressed using techniques such as mean imputation, median imputation, or forward filling, depending on the data type and context. This ensures that the dataset remains consistent and complete.
* **Outliers:** Outliers will be identified and analyzed to determine their impact on model training. If deemed necessary, outliers that disproportionately affect model performance will be removed or transformed to maintain the integrity of the dataset.
* **Normalization and Scaling:** Features will be normalized or scaled to ensure that all variables contribute equally to the model's learning process, which is essential for algorithms such as GBDT to perform effectively.

**D.4. Precautions for Sensitive Data**

When working with and communicating about sensitive data, several precautions will be taken to ensure data privacy and security:

* **Data Anonymization:** The dataset's personally identifiable information (PII) will be anonymized to protect individuals' privacy. However, since this project primarily involves financial market data, PII is not expected to be collected.
* **Access Control:** Access to sensitive data will only be restricted to authorized personnel. Role-based access control (RBAC) will limit data access based on the user's role within the project.
* **Data Encryption:** Sensitive data will be encrypted both in transit and at rest to prevent unauthorized access. Encryption protocols such as TLS (Transport Layer Security) will be used to secure data transmission.
* **Communication Practices:** When discussing sensitive data, team members will use secure communication channels, such as encrypted emails or secure messaging platforms, to prevent data leakage. Additionally, sensitive information will not be shared in public forums or insecure communication channels.

**References**

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). https://github.com/Microsoft/LightGBM (NIPS-2017-lightgbm-a-hi…)

Bonaccorso, G. (2017). Machine Learning Algorithms: Popular algorithms for data science and machine learning. Packt Publishing. Packt Publishing

Prajwal, N., Balivada, D., Chandra, S. N., & Poornoday, T. (2024). Decision Trees for Intuitive Intraday Trading Strategies. Preprint. (2405.13959v1)