Optimizing Investment Portfolios with the Super Learner Model

Eugene Jang[1] and Xin Jin[1]

[1]Department of Mathematics, The University of Tampa

1 Project Overview (written collaboratively) (250 words)

This project aims to develop an advanced model for predicting future financial returns and optimizing a long-term portfolio (10+ year) using a combination of machine learning techniques. The core of the project is the creation of a Super Learner model that combines multiple machine learning algorithms—Linear Regression, Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS), and Neural Networks—to generate more accurate predictions. The model will then be used to help optimize portfolio allocations based on these predictions. Additionally, the project will incorporate regularization methods to select the most relevant variables from high-dimensional financial data and use Bayesian Optimization to fine-tune the hyperparameters in the model.

The student will play an active role in implementing and testing the individual models (XGBoost, SVM, Neural Networks, etc.) and will be involved in applying regularization techniques to the data. They will also assist with portfolio optimization and the evaluation of model performance. The faculty member, as the project leader, will guide the student through the development of the Super Learner model, provide expertise on model integration, and ensure the application of correct machine learning methodologies. The faculty member will also supervise the Bayesian Optimization process for hyperparameter tuning. Together, we will assess the overall performance of the models and analyze their ability to predict financial returns and optimize investment portfolios. This collaborative effort combines both practical and theoretical aspects of machine learning in finance.

2 Project Description (written by student with faculty mentor guidance) (1000 words)

The accurate prediction of financial returns and the optimization of portfolio weights are central goals in modern finance. Portfolio optimization refers to the process of selecting the best mix of assets to maximize returns while managing risk. Traditional finance models often use linear relationships and static assumptions, which may not fully capture the complexity of the financial markets. Recent advancements in machine learning offer promising solutions by incorporating non-linear patterns and adapting to new data. However, building a reliable and accurate prediction model remains a challenging task, particularly when dealing with high-dimensional data—where the number of variables exceeds the number of observations.

One innovative approach to enhancing model accuracy is through the use of ensemble models, such as the Super Learner. The Super Learner combines the outputs of multiple machine learning models to improve prediction performance. In this project, we aim to develop a Super Learner model that combines five different machine learning techniques—Linear Regression, XGBoost, Support Vector Machines (SVM), MARS (Multivariate Adaptive Regression Splines), and Neural Networks—to predict future returns and optimize portfolio weights. By combining these models, we hope to leverage the strengths of each algorithm and improve overall accuracy, a key consideration in financial decision-making.

The use of machine learning in finance has already proven to be valuable in predicting asset prices, returns, and other key financial metrics. This research seeks to further explore how different models, when combined in an ensemble approach, can provide enhanced accuracy. Additionally, regularization methods and Bayesian Optimization will be employed to ensure that the models are both efficient and effective, preventing overfitting and finding the best model parameters.

The project will employ several theoretical frameworks and methodologies from both finance and machine learning.

- Super Learner Model: The Super Learner is an ensemble learning technique that combines multiple predictive models to improve prediction accuracy. Each base model (in our case, Linear Regression, XGBoost, SVM, MARS, and Neural Networks) will provide an individual prediction, which will be combined to form the final prediction. The idea behind the Super Learner is that combining different models compensates for the weaknesses of each model, leading to better overall performance. This method has been successfully applied in various domains but has not been widely explored in the context of portfolio optimization.

- Machine Learning Models:

1. Linear Regression: A statistical method used to model the relationship between a dependent variable and one or more independent variables. It will serve as a baseline model for comparison.
2. XGBoost: A powerful, gradient-boosted decision tree model that is widely used in machine learning competitions due to its ability to handle both regression and classification tasks efficiently.
3. SVM: A supervised learning algorithm that is effective in high-dimensional spaces, making it well-suited for financial data with many features.
4. MARS: A non-parametric regression method that models non-linear relationships between variables. It is flexible and can capture complex patterns in data.
5. Neural Networks: Deep learning models that consist of layers of neurons. Neural Networks can model very complex relationships in the data, and their ability to learn from large datasets makes them particularly powerful in financial prediction.

- Regularization for Variable Selection: Financial data is often high-dimensional, meaning that there are many potential features that could influence the outcome. Regularization techniques, such as Lasso or Ridge regression, help select the most relevant variables by penalizing the inclusion of irrelevant features. This is crucial for ensuring that the model remains interpretable and avoids overfitting, which occurs when a model becomes too complex and fits noise in the data rather than underlying patterns.

- Bayesian Optimization: Bayesian Optimization is a global optimization technique used to find the best hyperparameters for machine learning models. It is especially useful when the search space is large and expensive to explore exhaustively. By constructing a probabilistic model of the objective function, Bayesian Optimization efficiently narrows down the range of possible hyperparameters, ensuring that the models are fine-tuned for optimal performance.

- Portfolio Optimization: Once the Super Learner model has been trained to predict future returns, the next step is to use these predictions to optimize portfolio weights. This will be done using classical portfolio optimization methods, such as Mean-Variance Optimization, Markowitz Optimization, or Genetic Algorithms, which balances the trade-off between expected return and risk. The goal is to allocate capital across different assets in a way that maximizes the overall portfolio performance.

The project aims to answer the following research questions:

- What assets should be included in the portfolio? How does regularization for variable selection impact model performance in high-dimensional financial data?

- How do we assess the risk and return of a portfolio using predicted returns, asset correlations, and risk metrics like volatility and the Sharpe ratio?
- Can the Super Learner model outperform individual machine learning models in predicting future financial returns and optimizing portfolio weights?

By addressing these questions, we aim to contribute new insights into the application of machine learning for financial decision-making, particularly in the area of portfolio management.

This research is significant for several reasons: While machine learning has been widely applied in finance, the use of the Super Learner model for both predicting returns and optimizing portfolio allocation is relatively novel. This project has the potential to provide a more robust approach by combining multiple models, each contributing different strengths to the prediction task. Financial datasets often contain many variables, and selecting the most relevant ones is a critical challenge. Regularization techniques will ensure that the models remain efficient and avoid overfitting, which is a common issue in high-dimensional datasets. The use of Bayesian Optimization to fine-tune hyperparameters will help achieve optimal performance for each individual model, contributing to the overall success of the Super Learner model. The ability to accurately predict returns and optimize portfolios is crucial for investors and financial managers. This research could lead to more effective portfolio management strategies and improve decision-making in financial markets.

In this project, the student will take an active role in implementing the machine learning models, including Linear Regression, XGBoost, SVM, MARS, and Neural Networks, as well as applying regularization methods for variable selection. The student will also assist in the development of the Super Learner model and will be responsible for evaluating model performance and conducting portfolio optimization.

The faculty mentor will guide the student through the research process, providing expertise in machine learning techniques, portfolio optimization, and the overall development of the Super Learner model. The faculty member will also assist with the application of Bayesian Optimization for hyperparameter tuning and ensure the proper implementation of financial theory in the project.

This project seeks to combine state-of-the-art machine learning techniques with classical financial theories to create a more robust and accurate method for predicting financial returns and optimizing portfolios. By employing an ensemble approach, regularization methods, and Bayesian Optimization, we aim to advance the field of machine learning in finance and provide practical tools for portfolio management. The collaborative nature of the project will enable the student to gain valuable experience in both machine learning and financial modeling, while contributing to the ongoing development of this promising research area.

3 Proposed Activities and Benefits to Student(written collaboratively) (1000 words)

This project offers an opportunity for the student to engage in advanced research at the intersection of machine learning and finance. The project involves building an ensemble machine learning model, Super Learner, to predict future financial returns and optimize portfolio weights. The activities will focus on model development, data analysis, portfolio optimization, and model evaluation. Throughout this process, the student will be able to develop a variety of academic and practical skills while collaborating with the faculty mentor.

Proposed Activities:

- Literature Review and Background Research: The first task for the student will be conducting a comprehensive literature review on the application of machine learning in financial forecasting and portfolio optimization. The student will investigate the use of various machine learning techniques, including Linear Regression, XGBoost, SVM, MARS, and Neural Networks, and explore previous research that has utilized Super Learner models. Additionally, the student will examine regularization methods, such as Lasso and Ridge regression, and the use of Bayesian Optimization for hyperparameter tuning.

  This activity will help the student develop skills by learning how to access, evaluate, and synthesize existing literature to gain insights into the state of the field. It will also enhance their understanding of the theoretical foundations underpinning the project, particularly the intersection of finance and machine learning.

- Data Collection and Preprocessing: Once the student has a solid understanding of the existing literature, they will begin collecting relevant financial data. This will include historical asset returns, macroeconomic variables (interest rates, inflation, etc.), and other relevant features for portfolio optimization. The student will use R to clean the data, addressing issues such as missing values, scaling, and normalizing the data. The student will also ensure that the data is suitable for machine learning models by transforming variables and splitting the data into training and testing sets.

  This task will enhance the student's quantitative skills, as they will apply statistical and data manipulation techniques to prepare data for analysis. It will also provide hands-on experience with data management, a critical skill in data science and finance.

- Model Development and Training: The core of the project involves developing and training five machine learning models: Linear Regression, XGBoost, SVM, MARS,

and Neural Networks. The student will implement each model in R, gaining an understanding of the nuances of each technique and its application to financial data. For example, they will work with XGBoost by fine-tuning the number of trees and their depth and develop Neural Networks by adjusting the number of layers and neurons.

Through this activity, the student will develop creative thinking and problem-solving skills as they adapt models to handle financial data and address challenges such as overfitting and underfitting. Additionally, the student will enhance their machine learning techniques and coding skills, both of which are critical tools in data science and statistics.

- Building the Super Learner Model: The student will then focus on combining the individual models into a Super Learner model. The Super Learner will take the predictions from each base model and learn how to optimally combine them for improved performance. This task will involve experimenting with different ensemble techniques, evaluating their effectiveness, and selecting the best-performing Super Learner model.

  Building the Super Learner will provide the student with practical experience in model aggregation and ensemble learning. It will also foster critical thinking as they evaluate the performance of different combinations of models and adjust the architecture to improve predictions. The student will gain insight into how ensemble methods can be used to improve predictive accuracy and will learn how to assess model performance using cross-validation techniques.

- Portfolio Optimization: With the Super Learner model trained, the next step will be to use the predictions from the model to optimize portfolio weights. The student will implement portfolio optimization methods, such as Mean-Variance Optimization, which balances expected return and risk.

  This task will help the student develop a practical understanding of finance by applying machine learning predictions to real-world financial problems. It will also enhance their quantitative skill, as they will work with concepts such as expected returns, covariance matrices, and risk measures. The student will gain hands-on experience in portfolio management and learn how to evaluate the performance of different portfolio strategies.

- Evaluation and Results Analysis: Once the models are developed and the portfolio is optimized, the student will evaluate the performance of the Super Learner model and compare it to the individual base models. The student will assess the models based on metrics such as prediction accuracy (e.g., Mean Squared Error), portfolio performance (e.g., Sharpe Ratio), and risk-adjusted returns.

This activity will improve the student's critical thinking as they analyze the results, identify patterns, and draw conclusions based on quantitative data. It will also develop their communication skills, as the student will need to present their findings clearly and effectively, both in writing and in oral presentations.

- Writing and Dissemination: The final activity for the student will be writing up the results of the research and preparing a report or academic paper for submission to a peer-reviewed journal or conference. This will involve summarizing the methodology, presenting the results, and discussing the implications of the findings. The student will also create visualizations and figures to communicate the key insights from the analysis.

  This process will improve the student's communication skills, particularly their ability to write clearly and concisely about complex technical topics. It will also help the student develop academic writing skills, preparing them for future research projects and publications. Additionally, presenting the findings at a conference or seminar will help the student hone their public speaking and presentation skills, essential for their future career in academia or industry.

Throughout the project, the student will be responsible for the hands-on work of implementing models, preprocessing data, applying regularization, tuning hyperparameters, optimizing the portfolio, and evaluating results. The student will also be responsible for documenting the process, writing reports, and presenting their findings.

The faculty mentor will guide the student in terms of conceptualizing the project, ensuring that the methodologies are sound, and helping the student navigate challenges in the research process. The mentor will provide expertise in both machine learning techniques and portfolio optimization, helping the student refine their approach to data analysis and model evaluation.

By participating in this project, the student will gain valuable skills in machine learning, financial modeling, and data analysis, as well as experience in applying these techniques to real-world problems. The project will foster the student's critical thinking, problem-solving, and communication skills, preparing them for future academic or professional endeavors. Additionally, the student will gain hands-on experience with cutting-edge research techniques, such as the Super Learner model and Bayesian Optimization, and will contribute to advancing the use of machine learning in financial decision-making.

4 Mentoring Plan (written collaboratively, no more than 500 words)

The goal of this mentoring plan is to provide the student with the guidance, resources, and support necessary to successfully complete the proposed research project. The faculty mentor will work closely with the student throughout the project to facilitate their academic development, enhance their research skills, and ensure the project's success. The mentoring process will be structured, with clear communication, regular check-ins, and active guidance in both technical and conceptual aspects of the project.

- Meeting Schedule and Communication: To ensure steady progress, the student and faculty mentor will meet weekly to discuss the project's progress, address any challenges, and provide feedback on the student's work. These meetings can take place either in person or virtually, based on the schedules and preferences of both the mentor and the student. From June 1 to July 15, the mentor will not be physically present but will continue to offer guidance and feedback through virtual sessions on Zoom. During these meetings, the mentor will review the student's findings, offer direction, and help refine the next steps of the research. In addition to the weekly meetings, the mentor will be available via email to provide timely feedback on specific aspects of the project, such as coding issues, model development, and data analysis. The mentor will also utilize platforms like Overleaf to facilitate communication, allowing the student to present any challenges and receive guidance on potential solutions.

- Training and Skill Development: The mentor will ensure that the student develops proficiency in the technical skills required for the project. This will include:

  1. R Programming and Machine Learning Techniques: The mentor will guide the student in R, ensuring they are comfortable with data manipulation, model development, and visualization. The student will learn how to implement machine learning models (XGBoost, SVM, Neural Networks, etc.) and apply them to financial data.
  2. Portfolio Optimization: The mentor will provide insights into finance-related aspects, such as portfolio optimization methods, risk-return trade-offs.

  The mentor will also encourage the student to engage with supplementary resources, such as academic papers, online tutorials, and external workshops, to broaden their understanding of the project's methods and the research field.

- Support Student Ownership: Initially, the mentor will guide the student through the project, helping them set up the framework, familiarize themselves with the tools and data, and provide hands-on assistance during the model development process. As the student becomes more confident and capable, the mentor will gradually reduce the level of oversight and encourage the student to take more ownership of the project. This will involve:

    1. Encouraging Independent Problem-Solving: The mentor will foster an environment where the student can independently tackle challenges, ask questions, and find solutions, thereby increasing their problem-solving ability and confidence.
    2. Reviewing Progress and Setting Milestones: The mentor will help the student set clear milestones for each phase of the project and review their progress regularly. This will include evaluating the models' performance, adjusting strategies based on results, and discussing potential improvements.
    3. Providing Constructive Feedback: The mentor will offer timely, constructive feedback on the student's written work, coding approaches, and analytical techniques, allowing the student to refine their skills and thinking. This feedback will promote self-reflection and continuous improvement.
    4. Encouraging Scholarly Engagement: Toward the end of the project, the mentor will assist the student in preparing a scholarly report or paper for publication or conference presentation. The mentor will guide the student in structuring their findings, writing clearly and concisely, and preparing for peer review.

By gradually transferring responsibility for the project to the student, the mentor will help the student develop ownership of their research, increasing their independence, and preparing them for future academic and professional opportunities.

Through this structured mentoring approach, the student will gain both the technical expertise required for the project and the critical thinking and communication skills necessary for academic growth. The mentor will provide consistent guidance and foster a collaborative learning environment, ensuring the student is well-equipped to complete the project and thrive as a scholar.

5 Anticipated Dissemination and Reach (written collaboratively; maximum 250 words)

The student will aim to present the results of the research at a relevant academic conference or seminar, such as *The American Statistical Association Joint Statistical Meetings* (JSM),

International Conference on Machine Learning (ICML), *The IEEE International Conference on Machine Learning and Applications* (ICMLA). These conferences focus on the intersection of finance and machine learning and attract scholars and professionals from diverse fields. Presenting at these conferences will provide an excellent platform for showcasing innovative research and engaging in discussions on the latest methodologies in predictive modeling and portfolio optimization.

In addition to conference presentations, the student will also explore the possibility of submitting the work to a peer-reviewed journal. Potential journals include those focused on finance, machine learning, or their intersection, allowing the student to share their research with a broader academic audience. The mentor will guide the student through the manuscript preparation process, ensuring the work meets the standards of academic publishing. If the student is unable to present at a conference or face delays in submission, the contingency plan includes presenting the work in SURF Symposium at UTampa. These local academic events will provide the student with an opportunity to share their findings with the academic community, fostering valuable feedback and scholarly discussion.

These dissemination outlets are significant to the field as they offer opportunities to showcase the innovative application of machine learning to finance, contributing to both academic knowledge and real-world financial practices.

6 Itemized and Justified Budget

This budget outlines the necessary expenses to support the student's full-time research project during the summer under SURF program at UTampa. The requested funds will ensure that both the student and faculty mentor have the required resources for successful completion of the project.

- Student Stipend: $3,500
  Justification: The stipend will support the student's full-time engagement in the research project throughout the summer. This will allow the student to focus entirely on the development, testing, and optimization of machine learning models, including their application to portfolio optimization. The stipend also encourages the student's intellectual growth and commitment to high-quality, sustained research.

- Faculty Mentor Stipend: $1,000
  Justification: The faculty mentor stipend compensates for the time and effort dedicated to supervising the student's research, providing guidance on technical and conceptual aspects, and supporting the student's professional development.

- Faculty Mentor Fringe Benefits: $83.30
  Justification: Fringe benefits of 8.33% are applied to faculty stipends and are calculated as a percentage of the $1,000 stipend. This amount covers the cost of fringe benefits.

- Travel to Conference Presentation: $2000
  Justification: The student will present the research findings at a relevant academic conference such as *The American Statistical Association Joint Statistical Meetings* (JSM). The requested funds will cover the registration fee, transportation, and housing costs associated with attending and presenting at these conferences. This opportunity will allow the student to gain exposure to the broader research community and receive feedback on the project.

This budget supports the successful completion of the proposed project by providing the necessary resources for both the student and faculty mentor to carry out high-quality research. The requested funds will help develop the student's critical thinking, communication, and technical skills, while also advancing their academic and professional development. Through this project, the student will gain hands-on experience in cutting-edge techniques such as machine learning and portfolio optimization, and the mentor will continue to grow as a researcher and educator.

7 Faculty Evaluation of Participating Student (500-word max)

I have had the pleasure of working with Eugene Jang this semester in my MAT 272 Applied Statistics course. Throughout the course, Eugene has demonstrated a strong interest in applying advanced statistical methods to real-world problems. He approached me expressing an interest in conducting research on portfolio optimization using machine learning models in R, and after discussing the details of his ideas, I found his enthusiasm and initiative to be impressive.

While Eugene has not participated in formal research projects before, his academic background in Data Science and commitment to learning have made him a strong candidate for this research endeavor. He has consistently shown an ability to grasp complex statistical concepts quickly and apply them to both theoretical and practical problems. His proficiency with R, coupled with his understanding of machine learning techniques, will provide a solid foundation for his contributions to this project.

In collaborating with Eugene on this grant proposal, I observed his ability to think critically and approach problems with a problem-solving mindset. He was proactive in

researching various machine learning models and their applications in portfolio optimization. Eugene contributed thoughtfully to discussions regarding the selection of models, such as XGBoost, Neural Networks, and Support Vector Machines. Furthermore, Eugene's ability to break down complex tasks, such as model training and data preprocessing, has been a clear asset in the development of this proposal.

One of Eugene's strengths is his attention to detail and his ability to engage with the material at a deep level. He also demonstrates a willingness to learn from feedback, incorporating suggestions and refining his understanding of concepts and techniques. This collaborative approach will undoubtedly serve him well in the context of this research project.

That being said, Eugene has room to grow in his experience with independent research, particularly in handling the challenges that arise during the research process. Although he has the necessary theoretical understanding, I believe the practical aspects of managing a complex research project—such as debugging code, fine-tuning machine learning models, and interpreting results in a way that advances the field—will provide valuable opportunities for growth. As with any emerging researcher, Eugene may face moments of uncertainty, but I am confident that his strong work ethic and eagerness to learn will allow him to navigate these challenges effectively.

In summary, Eugene is a highly motivated and capable student with the potential to make significant contributions to this research project. His strengths in statistical modeling and machine learning, coupled with his enthusiasm and critical thinking skills, make him a promising candidate for this research opportunity. I am confident that the experience gained through this project will further develop his skills and help him grow as a scholar.

8 Student's Letter of Intent

Please answer the following three questions in paragraph form (500-word limit): 1. How did your interest in this research project develop? 2. What skills do you possess that will help you complete this project? What skills will this project help you develop? 3. How does the completion of this research project fit within your future plans? The letter should be uploaded as a pdf file.

9 References