ORIE 5741 Project
Advanced Machine Learning Applications in Morningstar Fund Rating Predictions

Yuxuan Wang, Xiaoxuan Liu, Quanyi Li

May 2024

Abstract: This study investigates the application of advanced Machine Learning (ML) techniques, including Support Vector Machine (SVM), Bagging, and hybrid models integrated with Principal Component Analysis (PCA), to efficiently use data to predict Morningstar fund ratings. Utilizing a comprehensive dataset of over 13,900 investment funds, we explored the effectiveness of these models in a dynamic financial market environment. The PCA was implemented to address multicollinearity and reduce dimensionality, leading to enhanced model performance and computational efficiency. Our findings indicate that while the standalone Bagging model serves as a sturdy baseline, its integration with PCA considerably improves prediction efficiency without compromising its predictive power. This study highlights the potential of combining dimensionality reduction techniques with ML to tackle complex predictive tasks in finance, thereby offering insights that could enhance real-time, data-driven fund evaluations and investment decision-making processes.

GitHub Link: <https://github.com/Howl101/ORIE5741_Project_Fund_Ratings>

1. Introduction

This report investigates the efficacy of advanced ML techniques in enhancing the precision of Morningstar fund ratings. Utilizing a list of investment funds, including Mutual funds and Exchange-Traded Funds (ETF), the study compares the performance of SVM, Bagging, and a hybrid model combining PCA to these two approaches in predicting the fund ratings.

1.1. Our mission

The complexity of the global financial market necessitates robust tools for investors to evaluate investment funds. [1] The traditional methods, although comprehensive, could fail to adapt quickly to dynamic market conditions. [2] This project leverages ML to address these challenges, aiming to provide a scalable, accurate data-driven alternative to existing rating methods, with more efficient use of data to reduce operational fees and thereby increase revenue.

We aim to leverage the fund rating outcome to refine our investment strategies, ensuring they are responsive and informed. This advancement aligns with our core objective: optimizing investment intelligence, elevating our investment acumen, and adding substantial value to our business operations and stakeholders.

1.2. Data Description

The Morningstar dataset, updated as of January 2024, encompasses detailed records of more than 13,900 investment funds worldwide. It provides a comprehensive view of fund characteristics and performance, including quantitative and fundamental financial metrics and categorical variables. Key financial metrics include net flow, annual and multi-year returns, fund size, SEC yields, turnover ratio, expense ratios, volatility indicators, and risk-adjusted returns such as the Sharpe ratio. Categorically, the dataset delineates industry sectors, and management styles, and notably, integrates an ESG (Environmental, Social, Governance) binary classification, reflecting the growing emphasis on sustainable investing. Most of the financial metrics have 1-year, 3-year, and 5-year monthly values. For our study, we have chosen to measure all metrics using a 3-year timeframe on a monthly basis. This allows for a mid-to-long-term perspective on performance, while also minimizing the risk of overfitting short-term fluctuations. [3]

The chosen features and the interpretations are as follows:

Term	**Description**
Total Return (%)	Total return of the fund.
Standard Deviation	Measures the return volatility.
Sharpe Ratio	Measures the risk-adjusted return.
Alpha	Measures the fund's performance relative to the benchmark.
Beta	Measures the fund's correlation with the market.
R-Squared	Measures synchronization of fund movements with the market.
Downside Capture Ratio	Reflects fund performance when the market declines.
Upside Capture Ratio	Reflects fund performance when the market rises.
Information Ratio	Measures the excess return of the fund relative to the benchmark against the tracking error.
Tracking Error	Measures the deviation of fund performance from the benchmark.
Turnover Ratio (%)	Reflects the activity level of fund management.
Net Expense Ratio (%)	Reflects the cost of fund management.
ESG portfolio	Indicates if the fund is an ESG investment portfolio.
Management Fee Ratio (%)	Measures the proportion related to the Management fee.
Fund Size ($)	Measures the Asset Under Management of the fund.

The data includes the Morningstar Fund Rating, our selected response variable, which is normally distributed as illustrated in Figure X. In summary, the dataset provides a comprehensive overview of the funds, meets the necessary criteria for model training and testing, and can be used effectively for classification purposes.

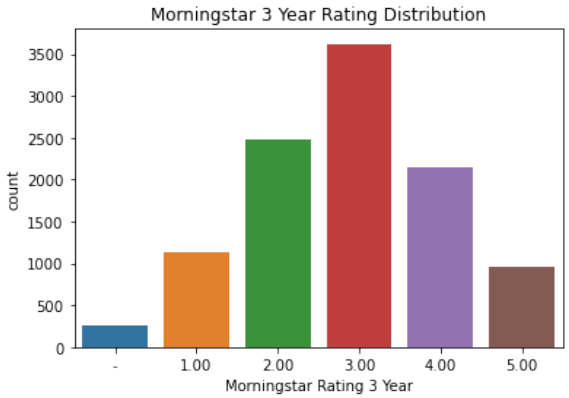

Figure 1: Response Variable (Morningstar 3 Year Rating) Distribution

2. Data Preparation

The performance of predictive models is highly dependent on the quality and accuracy of the input data. In this section, we outline the specific steps we took to clean and transform the Morningstar dataset.

2.1. Data Cleaning

The dataset was initially cleaned to remove any instances with missing values in key financial indicators. This involved filtering out any funds that lacked complete data entries such as net flow, return rates, or ESG scores. Outliers, particularly in financial metrics, were identified and filtered using Interquartile Range (IQR) techniques.

2.2 Data Transformation

Categorical variables such as fund type were encoded using One-Hot Encoding. The portfolio's ESG characteristic was converted into binary format. Numerical features such as fund size, returns, and expense ratios were normalized using the MinMaxScaler to bring them onto a uniform scale. This prevents any single metric from disproportionately influencing the model due to scale differences and can help us in the later PCA and SVM.

2.3. PCA

With the prepared features, we further generated a feature correlation map to see if certain features demonstrate a strong correlation.

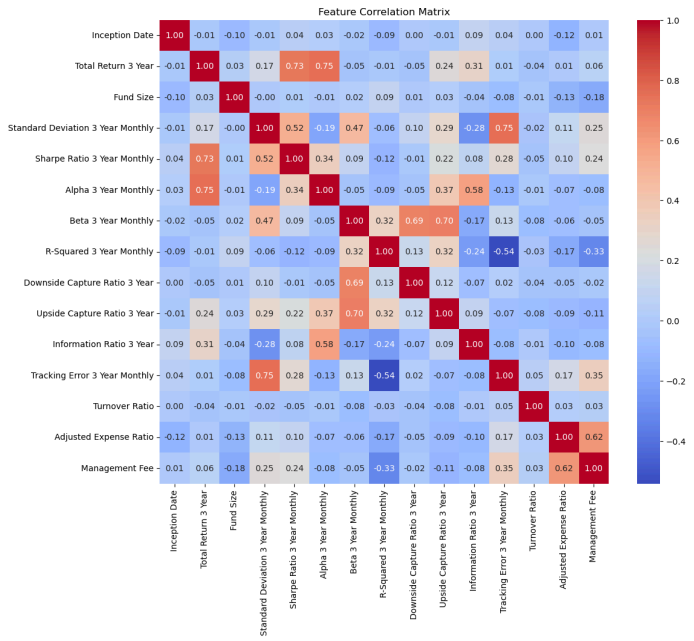

Figure 2: Feature Correlation Matrix

Given the high dimensionality of the original dataset and to address multicollinearity. PCA was applied to reduce the dimensionality of the data after normalization and transformation. This step was crucial to avoid overfitting and to enhance model performance on unseen data. Since we want to acknowledge the trade-off between dimensionality reduction and pre-

serving important information, we set the explained variance threshold to 95%.

As indicated in Figure 3, the chosen 10 principal components were able to explain 95.06% of the data variance.

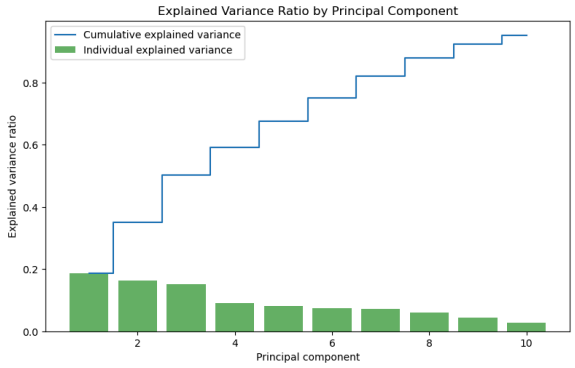

Figure 3: Explained Variance Ratio

2.4 Data Splitting

The dataset was then methodically split into training and test sets to ensure a robust validation of the models' performance. Initially, 20% of the entire dataset was taken to form the test set. This test set remains completely untouched during the model selection and tuning phases and will serve as an objective benchmark to assess the models' performances. The remaining 80% of the data was further divided into training and validation sets. Specifically, 25% of this subset (approximately 20% of the original dataset) was set as the validation set, with the remaining 60% being the training set. This structured approach ensures that the training set is used to build and train the models, the validation set aids in fine-tuning parameters and selecting the best model without the risk of overfitting, and the test set provides an unbiased evaluation of the final model's performance.

3. Methodology

Two ML models, SVM and Bagging, were initially developed. Following parameter tuning and performance evaluation, we identified the optimal settings for each model. To further refine our approach and address the issue of multicollinearity, we integrated PCA with the two models.

3.1. SVM

SVM finds the optimal hyperplane that maximizes the margin between different classes, even in complex and high-dimensional spaces. With the use of kernel functions, the model can interpret data structures and relationships better by mapping data into higher-dimensional spaces where classes can be separated linearly. [4]

Parameter Tuning: The choice of Kernel and Parameters including the regularization para meter (C), the kernel coefficient (gamma), and the degree of freedom (df) were tuned using a cross-validated grid search (GridSearchCV) to determine the best combination, which is reflected below.

Parameter	**Value**
C	100
Kernel	RBF
Gamma	scale
df	3

Specifically, the RBF kernel handles the non-linearity in the relationship between fund characteristics and their ratings efficiently.

The optimal parameters yielded the following result on our test set. Despite the numerical performance not being exceptionally high, the results are sufficient for the operational fund settings, where perfect and not overfitting predictions are often unattainable due to inherent unpredictability and variability of financial data. [5]

Precision	0.61
Recall	0.56
F1-score	0.58
Accuracy	0.59

Below is the Confusion matrix of the model.

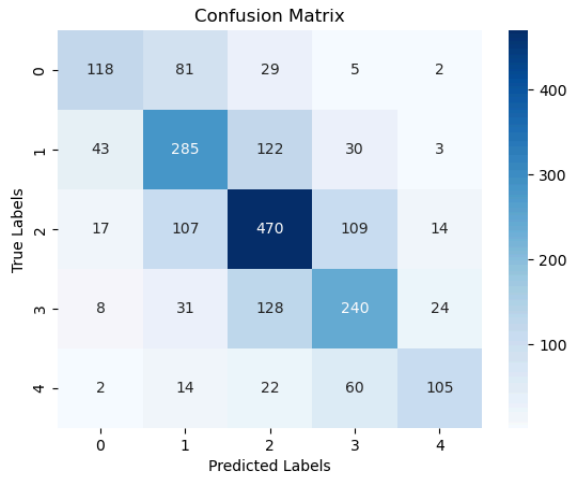

Figure 4: Confusion Matrix for SVM

As shown in Figure 4, SVM has the capability to detect whether a fund possesses a high rating (4 or 5), medium rating (3), or low rating (1 or 2). Nonetheless, its proficiency in distinguishing between 1 and 2 as well as 4 and 5 is limited.

3.2 Bagging

Following the SVM, we implemented a Bagging ensemble model, which enhances the model accuracy and stability by combining the predictions from multiple decision trees. Given the large Morningstar datasets with numerous features, Bagging is an effective choice as it handles overfitting. Unlike Random Forest, which constrains each decision tree to a random subset of features, Bagging allows each tree in the ensemble to consider all features during training, offering a freer, potentially more robust modeling approach that can adapt better to complex data structures.

Hyperparameter Tuning: Similar to the SVM, we conducted a grid search with GridSearchCV to tune the Bagging model's parameters, such as the number of base trees (n_estimators), the maximum number of data (max_samples, a ratio indicating the maximum proportion of data used by each base tree), and features (max_features, a ratio indicating the maximum proportion of features used by each base tree) sued to build each base tree. The best combination is as follows.

Parameter	**Value**
n_estimator	200
max_samples	1.0
max_features	0.5

As we can see from the table below, the result is significantly better, compared to the SVM model. The aggregation of multiple decision trees is more robust against our diverse and complex financial data.

Precision	0.83
Recall	0.80
F1-score	0.81
Accuracy	0.82

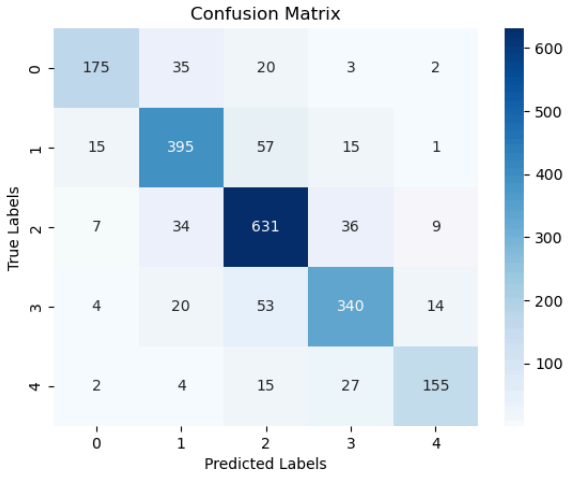

Figure 5: Confusion Matrix for Bagging

Bagging performs better overall at classifying each rating, with an enhanced ability to distinguish between the 1 and 2 and 4 and 5 pairs, as shown in Figure 5.

Comparing the results in the two tables and Figures 4 and 5, the Bagging model exhibited improved precision, recall, F1-score, and accuracy metrics compared to the SVM model.

3.3 Hybrid Model with PCA

In this section, we explored the integration of PCA with our two models. The rationale behind this approach was motivated by the significant dimensionality reduction (from 16 to 10) achieved through PCA, which simplifies the feature space while retaining the most informative aspects of the data.

3.3.1. PCA-SVM

For the PCA-SVM model, we used the principal components as input features and the best SVM hyperparameters from previous tuning efforts. The model was trained on the PCA-transformed training set and subsequently evaluated on the PCA-transformed test set.

As shown in the table below, the PCA-SVM model demonstrated an overall accuracy of 57%, with precision, recall, and F1-scores varying moderately across different classes.

This repre- sents a slight drawback of applying the feature dimension reduction method to the models, as compared to using the full feature space.

Precision	0.61
Recall	0.55
F1-score	0.57
Accuracy	0.57

Below is the corresponding confusion matrix.

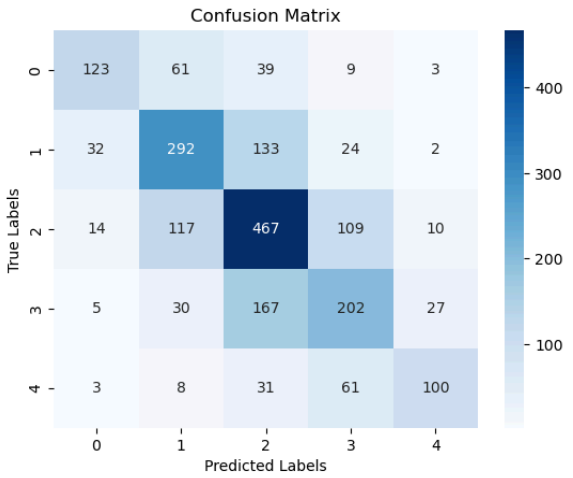

Figure 6: Confusion Matrix for PCA + SVM

3.3.2. PCA-Bagging

Following the PCA-SVM model, we extended PCA to the Bagging ensemble model. By using the same PCA-transformed features, we trained the Bagging model under the best Bagging hyperparameters previously calculated.

As shown in the table below, it is evident that the PCA-Bagging model outperformed the PCA-SVM model when evaluated on the transformed test set. The former model showed an overall accuracy of 74%, which although slightly lower than the previous model with the 16 full feature space, still demonstrated high accuracy. This shows that the incorporation of PCA into ensemble methods for increased robustness without compromising too much on accuracy.

Precision	0.80
Recall	0.71
F1-score	0.74
Accuracy	0.74

Below is the corresponding confusion matrix.

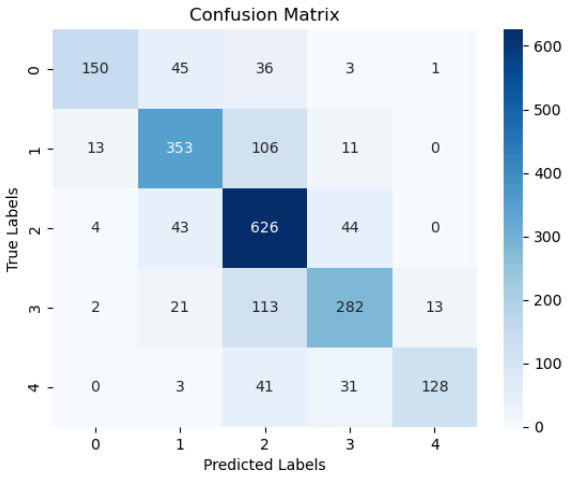

Figure 7: Confusion Matrix for PCA + Bagging

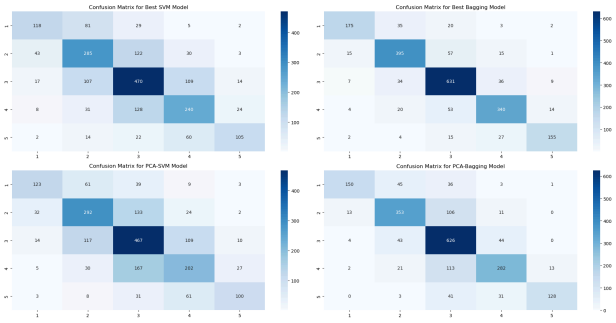

Figure 8: Comparison between models Confusion Matrices

After combining the four confusion matrices in Figure 8, it can be observed that the performance on each rating score is not significantly impacted by using PCA.

4. Discussion

4.1. Model Complexity vs. Performance

The integration of PCA with SVM and Bagging models showcases a strategic compromise between model complexity and performance efficiency. While the PCA-SVM and PCA-Bagging models did not achieve as high accuracy as the standalone SVM and Bagging models, this is a reasonable trade-off. PCA reduces the dimensionality of the model, which not only speeds up the model training process but also potentially decreases the risk of overfitting, making the model potentially perform better on new, unseen datasets.

Although the PCA-based models exhibit slightly diminished performance metrics, the benefits of reduced computational complexity and enhanced generalizability underscore the value of this approach. The slight reduction in model accuracy is outweighed by gains in training efficiency and a lower tendency for the model to overfit, which is crucial for maintaining robust performance across varied datasets.

4.2. Fairness and Ethical Implications

The implementation of ML within the financial sector, specifically in the evaluation of fund ratings, remains a relatively new and innovative approach. However, due to the dependence on historical data, which may contain inherent biases, it is imperative to maintain continual precision to guarantee impartiality. This necessitates regular updates and testing of the models against fresh data, to prevent the perpetuation or formation of any discriminatory patterns. [5] Additionally, it is crucial to prioritize transparency in both the construction and operation of these models, to uphold the trust of investors and other interested parties.

4.3. Weapons of Math Destruction

The concept of "Weapons of Math Destruction" highlights the dangers of opaque algorithms that scale and perpetuate biases. In sectors like finance where they can have profound personal and societal impacts. Our use of ML for predicting fund ratings could fall into this problem if not carefully managed. Ensuring that

these models do not become black boxes involves clear documentation, open methods, and provisions for audibility. Furthermore, the models should be designed to be as inclusive as possible and regularly reassessed for unintended consequences. [5]

The integration of advanced techniques such as PCA does mitigate some risks by simplifying the models and focusing on the most significant features, which can help reduce the occurrence of unintended bias. However, as we enhance model sophistication, we must also advance our strategies for managing these tools responsibly. Ensuring ethical usage, preventing discrimination, and maintaining transparency must be ongoing priorities as we continue to develop and use these techniques in the financial industry. [4]

5. Conclusion and Business Insights

Our project successfully demonstrated the power of integrating PCA with ML models SVM and Bagging to predict Morningstar 3-Year Ratings for investment funds. The use of PCA reduces the dimensionality of the dataset, which decreases the computational complexity and potentially enhances model training speed. This reduction can also mitigate the risk of overfitting, possibly leading to better performance on unseen data. Therefore, even though the PCA-based models exhibit a slight decrease in performance, the trade-offs are justified from a broader perspective of model management and deployment, especially in scenarios where computational resources are a constraint.

Looking forward, there is potential to explore other dimensionality reduction techniques, such as LDA to compare their effectiveness against PCA. Developing user-friendly tools that incorporate these models could also make sophisticated fund rating predictions more accessible to investors and financial analysts.

Despite these innovations, the reliance on historical data may add existing biases into the models, requiring continuous monitoring and updates to ensure fairness. Moreover, while PCA helps in improving model efficiency, it could reduce the interpretability of the results, which is critical in financial decision-making.

In conclusion, this study highlights the significant impact that ML and dimensionality reduction can have on finance, particularly in investment fund ratings. By leveraging these tools, we can optimize investment decision-making processes, achieving equally precise results with fewer features, which saves the computation resources as well as the data-collecting resources needed, cutting down operating fees for our business. This approach empowers us to maintain robust and responsible models. Going forward, we will integrate these techniques into our fund rating operations, streamlining our data collection and model processing efforts while continuing to deliver high-standard service to our clients.

References

1. Yu, P., Li, Z. & Wang, Y., 2009. Can Institutional Investors Outperform Individual Investors? Journal of Financial Research, (08), pp.147-157.
2. Sirri, E.R. & Tufano, P., 1998. Costly Search and Mutual Fund Flows. Journal of Finance, 53, pp.1589-1622.
3. Luo, Y., 2005. Comparative Study of Fund Rating Methods. Statistics and Decision, (20), pp.39-41.
4. Ding, H., 2021. Application Research of Enterprise Credit Rating Based on Support Vector Machine with Grid Optimization Model. Era of Financial Technology, 29(10), pp.63-66.
5. Feng, H. & Li, S., 2021. Research on Multilevel Cascade Classifier Based on Various Support Vector Machines and Its Application in Credit Scoring. Data Analysis and Knowledge Discovery, 5(10), pp.28-36.

6. Contribution

Yuxuan Wang: Dataset preparation, coding (debugging and validating), proposal and presentation composing, and presentation.

Xiaoxuan Liu: Literature review, coding (data processing and model building), proposal and presentation enhancing, and presentation.

Quanyi Li: Proposal enhancing, presentation script composing and video editing, and presentation.