Machine Learning in Morningstar Fund Rating and Management

Project Proposal

March 14, 2024

Yuxuan Wang, Xiaoxuan Liu, Quanyi Li

Data: https://www.kaggle.com/datasets/mauriziogiorda/morningstars

Introduction & Importance of the Problem

This proposal aims to promote machine learning and data analysis innovation that aims to develop a predictive model for rating funds based on various financial attributes.

Investment funds are a crucial part of the global financial ecosystem, offering investors opportunities for wealth growth. However, the vast number of available funds makes it challenging for investors to make informed decisions. Therefore, fund ratings become crucial..

Traditional analytic methods, such as the Morning Star, while rigorous and professional, take time and can be slow to reflect market changes. Machine learning (ML) classification methods, while broadly applied in credit and other financial ratings, are less explored in fund ratings. Our team aims to investigate their potential in fund ratings, broadening their application and offering new insights. The use of ML models offers real-time updates and data-driven insights. Through analyzing vast datasets and market trends, the model promises more timely and accurate ratings.

By transitioning to a more data-driven approach, 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 fund performance through investment intelligence, elevating our investment acumen, and adding substantial value to our business operations.

Data Description

The Morningstar dataset, updated as of January 2024, encompasses detailed records of more than 13,900 investment funds worldwide. It is a rich repository of both quantitative and fundamental financial metrics and categorical variables, offering a comprehensive view of each fund's characteristics and performance. 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. The data also contains four response variables: Medalist Ratings, Morningstar Fund Rating, Morningstar Return Rating, and Morningstar Risk Rating. Altogether, the dataset offers a thorough presentation of the funds, satisfies the requirement for model training and testing, and can be effective in the classification and rating prediction task,

Objective & Research Questions

We seek to find a way to apply ML methods to accurately classify and predict the ratings of investment funds based on their diverse attributes.

We want to Identify the key variables that most significantly impact the classification and prediction of fund ratings. We will examine whether these influential variables are rooted in fundamental analysis (such as financial health and market position) or technical analysis (based on historical trading patterns) and whether they reflect short-term or long-term observation periods. We try to analyze the defining characteristics of funds based on their performance and risk profiles. This includes:

1. "Best-performing funds" - Funds with higher return ratings and lower risk ratings.
2. "Return-centered funds" - Funds with higher return ratings but also higher risk.
3. "Risk-averse funds" - Funds with lower risk ratings and modest return ratings.

We aim to evaluate which methods yield an accurate and efficient model. We will focus on the balance between model complexity and performance, aiming to achieve high accuracy without resorting to overly complex or numerous variables, to avoid underfitting and overfitting, ensuring that our models are both effective and efficient.

Methodology

We primarily aim to compare three models: Random Forest, SVM, and a hybrid Random Forest-SVM model. The Random Forest method helps us select key indicators, reducing complexity and refining our feature set. This subset is then used to build a new SVM model. We evaluate the hybrid model's performance against the individual Random Forest and SVM models, aiming to verify the effectiveness of our approach with a subset of funds as test samples. This streamlined methodology showcases the potential of machine learning in fund rating, offering a new, objective evaluation tool. Regularization methods, such as Lasso or Ridge, as well as Boosting and Bagging, may also be evaluated.

Potential Problems

1. Data Quality and Availability: incomplete data, outdated information, or inconsistent quality may impact model performance and reliability.
2. Model Interpretability: While machine learning models, especially complex ones like random forests and SVMs, offer high predictive accuracy, their decision-making processes are often "black boxes" lacking transparency.
3. Dynamic Market Conditions: the market is highly dynamic and unpredictable. As a result, the model may need regular updates to keep pace with market shifts.
4. Math Disasters: The complexity of ML models, especially when dealing with financial data, can lead to "math disasters." These occur when models, based on flawed assumptions or incorrect data interpretations, produce misleading ratings.
5. Moral and Fairness Considerations: Utilizing ML in fund rating also raises moral questions, particularly regarding fairness and bias. Algorithms might inadvertently favor certain types of funds based on historical data patterns, which could perpetuate existing biases or unfair practices in investment.