Project Proposal: Analyzing Mutual Funds and ETFs for Investment Decisions

Brandon Ngo and Joshua Sanderson

Introduction

As students interested in finances, understanding investment vehicles like mutual funds and exchange-traded funds (ETFs) is critical for making informed financial decisions. These instruments represent significant investments, and selecting the right one requires careful evaluation of various metrics. Our project leverages a comprehensive dataset to analyze mutual funds and ETFs, focusing on their performance, risk, and cost. We aim to create visualizations, regression models, and clustering analyses to help investors identify funds that align with their financial goals. Specifically, we will examine differences between actively managed mutual funds and passively managed ETFs, explore the impact of fees on returns, and compare top-performing funds within specific categories (e.g., large-cap equity).

Dataset: Mutual Funds and ETFs

Data

The dataset, “Mutual Funds and ETFs” by Stefano Leone, contains 24.8 MB of data with 27 features describing 24,239 funds. Key features include:

- Fund Characteristics: Fund name, category (e.g., large-cap, bond), investment strategy (active/passive), inception date, and fund family.
- Performance Metrics: Net asset value (NAV), total net assets, annual returns (1-year, 3-year, 5-year, 10-year), and dividend yield.
- Risk and Volatility: Standard deviation, beta, alpha, Sharpe ratio, and Morningstar risk rating.
- Cost Metrics: Expense ratio, management fees, and load fees (front-end/back-end).
- Other Features: Minimum initial investment, fund size, and turnover ratio.

This dataset provides a rich foundation for analyzing the factors that drive fund performance and investor preferences.

Methodology Plan

We will employ the following analytical techniques to explore the dataset:

1. Linear Regression: We will use linear regression to investigate relationships between variables, such as the impact of expense ratios and risk (e.g., beta) on annual returns. For example, we hypothesize that lower expense ratios correlate with higher returns for ETFs compared to mutual funds.
2. K-Means Clustering: We will apply clustering to group funds with similar characteristics, such as risk profiles, returns, or asset classes. This will help investors identify funds within the same performance tier. For instance, we can cluster funds by category (e.g., equity, fixed income) to compare actively managed mutual funds against ETFs.

3. Comparative Analysis: We will compare actively managed mutual funds to passively managed ETFs to understand differences in cost, risk, and performance. Additionally, we will analyze top-performing funds within specific categories to identify trends (e.g., why certain large-cap ETFs outperform others).

Visualizations, such as scatter plots of risk vs. return and bar charts of expense ratios by fund type, will enhance interpretability.

Evaluation Plan

Our primary goal is to determine which mutual funds or ETFs offer the best risk-adjusted returns for investors. To evaluate this, we will:

- Survey Finance Students: Present our models and visualizations to finance students and collect feedback via a survey. The survey will assess whether our analyses help them identify suitable funds based on risk tolerance and investment goals.
- Secondary Questions:
  - Which funds perform best in specific asset classes (e.g., equity, bonds)?
  - How do expense ratios and management fees impact long-term returns?
  - What distinguishes top-performing ETFs from mutual funds in terms of risk and cost?
- Metrics: We will use regression coefficients to quantify the impact of features like expense ratios on returns. K-means clustering will help grade funds on a scale (e.g., A+ to F) based on performance, risk, and cost relative to peers.

GitHub Repository and Initial Tasks

[GitHub Repository Link: https://github.com/bngo03/data-mining-final-3162]

Initial Tasks

- Data Pre-Processing - Brandon Ngo and Joshua Sanderson (Due April 22, 2025)
- Linear Regression Analysis - Brandon Ngo and Joshua Sanderson (Due April 25, 2025)
- K-Means Clustering - Brandon Ngo and Joshua Sanderson (Due April 28, 2025)
- Insights and Conclusion - Brandon Ngo and Joshua Sanderson (Due April 30, 2025)
- Impact of Our Project - Brandon Ngo and Joshua Sanderson (Due May 3, 2025)
- Presentation Deck - Brandon Ngo and Joshua Sanderson (Due May 5, 2025)
- Uploading Code to GitHub Repository - Brandon Ngo and Joshua Sanderson (Due May 8, 2025)

Group Expectations

We are committed to meeting the deadlines outlined above. To ensure accountability:

- Lack of communication will result in a warning to the responsible group member.
- After two warnings, we will escalate the issue to the TA for mediation.
- Persistent issues post-TA intervention will lead to a letter grade reduction per additional warning.

This project will provide valuable insights into mutual funds and ETFs, empowering investors to make data-driven decisions.