Group Assignment: Predicting Monthly Excess Returns of Market Index

Overview

In this group project, you will apply machine learning techniques to predict monthly excess returns of the S&P 500 using the Welch and Goyal (2008, 2022) dataset. The focus is on developing predictive models, comparing their performance, and deriving financial insights.

Deliverables (3 files to be submitted)

Preliminary Presentation (December 9, 2024)

- Duration: 15-2 minutes (incl. Q&A).
- Content: Data overview, methodology, findings, and model comparison.
- Include visualizations and ensure all team members participate.
- You should submit **a slide deck** on the day of the presentation.

Submission of Final Reports (January 9, 2025)

- You should submit **a code script** that contains all the code used in the analysis.
- You should submit **a report** that includes the following sections:
    1. Executive Summary
    2. Introduction (Problem statement, objectives)
    3. Data Description (Overview of dataset)
    4. Methodology (Preprocessing, model selection, validation approach)
    5. Results (Performance comparison, predictor analysis)
    6. Financial Insights (Predictor importance, investment recommendations)
    7. Appendices (Code, visualizations, contribution statement)

Technical Requirements

- **Recommended Language**: Python
- **Key Libraries**: `pandas`, `numpy`, `scikit-learn`, `matplotlib`, `seaborn`
- **Submission Format**: One slide deck (.ppt or .pdf), one code script (.ipynb or .py), and one report (.pdf).

Dataset

- **Name**: `data.csv`
- **Source**: Welch and Goyal Financial Indicators
- **Time Period**: 1926–2023
- **Target Variable**: The last column, R, representing monthly excess returns of the CRSP value-weighted market index.
- **Predictors**: All other columns, which include 15 financial indicators.

Objectives

1. Predict the monthly excess returns of the market index using machine learning models.
2. Compare two distinct models from different families: one from linear and one from tree-based family.
3. Extract financial insights.

Grading Criteria

- **Technical Implementation**: 40%
- **Model Performance**: 25%
- **Financial Interpretation**: 20%
- **Presentation**: 15%

Data Description (Detailed)

Column	Description
date	Monthly time index.
R	Monthly excess return of the CRSP value-weighted index (R = CRSP_SPvw - Rfree). Calculated using the risk-free rate (`Rfree`), not `tbl`.
mr	One lag of the market return (CRSP_SPvw, including dividends). Represents the previous month's value-weighted market return.
dfy	Default Yield Spread: The difference between BAA and AAA-rated corporate bond yields (dfy = BAA - AAA).
dfr	Default Return Spread: The difference between long-term corporate bond and long-term government bond returns (dfr = corpr - ltr).

infl	Inflation: Based on the Consumer Price Index (CPI). Since CPI is released with a one-month lag, this value is lagged for use in regressions.
svar	Stock Variance: Computed as the sum of squared daily returns on the S&P 500 index.
de	Dividend Payout Ratio: The difference between the log of dividends and the log of earnings.
lty	Long-Term Government Bond Yield: Yield on long-term U.S. government bonds.
tbl	Treasury Bills: Monthly risk-free rate approximated by the 3-month Treasury bill rate (secondary market rate from the Federal Reserve Economic Database - FRED).
ltr	Long-Term Rate of Returns: Returns on long-term government bonds.
tms	Term Spread: The difference between the long-term government bond yield and the Treasury-bill rate (tms = lty - tbl).
dp	Dividend-to-Price Ratio: The difference between the log of 12-month moving sums of dividends (D12) and the log of prices.
dy	Dividend Yield: The difference between the log of 12-month moving sums of dividends and the log of lagged prices.
ep	Earnings-to-Price Ratio: The difference between the log of 12-month moving sums of earnings (E12) and the log of prices.
b/m	Book-to-Market Ratio: The ratio of book value to market value for the Dow Jones Industrial Average.
ntis	Net Equity Expansion: The ratio of 12-month moving sums of net issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks.

Recommended Steps

Data Preparation

- Perform exploratory data analysis (EDA) and handle missing values or data type inconsistencies.
- Use data from **1926–2018** for training and validation.
- Reserve data from **2019–2023** for testing.

Modeling

- Implement and compare **two distinct models**:
  - One **linear model** (e.g., Ridge, Lasso, Elastic Net).
  - One **tree-based model** (e.g., Random Forest, Gradient Boosting).
- Tune hyperparameters to optimize out-of-sample performance using **time-series validation**.

Evaluation

- Use at least one metrics such as:
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - R-squared (R^2)
  - Mean Absolute Error (MAE)
- Compare predictions to actual returns and visualize results.

Financial Analysis

- Discuss the economic significance of key predictors.
- Analyze the strengths and limitations of each model.
- Provide actionable investment insights.

More Hints and Actionable Tasks

- You are not expected to finish all these tasks. They are provided as a guide to help you structure your analysis.

1. Data Preparation	Hints and Actionable Tasks
Dataset Preview	Inspect dataset structure, types, and missing values using `df.head()`, `df.info()`, and `df.shape()`. Explain the role of each predictor and its financial relevance (e.g., `dfy` measures risk premia).
Summary Statistics	Use `df.describe()` to calculate mean, median, variance, and detect anomalies. Consider standardizing features to avoid feature scalinig issues.
Missing Data	Visualize and calculate missing values using heatmaps and percentage calculations. Select the proper imputation techniques like forward-fill (`ffill`) or backward-fill (`bfill`).

Correlation Analysis	Compute and visualize correlations between predictors using a heatmap. Highlight potential multicollinearity issues.
Visualization	Create scatterplots and residual plots to analyze relationships between predictors and the target.
Feature Distributions	Plot histograms for individual predictors to check for skewness or outliers. Consider standardizing, log-transforming or clipping extreme values to avoid distorting models.
Feature Engineering	Create lagged variables and interaction terms to capture underlying dynamics (e.g., dp/dy). Justify new features based on financial theories or hypotheses.
Data Splitting	Split the data into training (1926–2018) and testing (2019–2023) sets, ensuring no data leakage. Use a time-based split instead of random sampling to maintain temporal integrity.

2. Modeling	**Hints and Actionable Tasks**
Model Selection	Choose two distinct models: one linear (e.g., Ridge, Lasso) and one tree-based (e.g., Decision Trees, Random Forest, Gradient Boosting). Highlight strengths and limitations of each model type.
Time-Series Validation	Use time-series cross-validation for hyperparameter tuning and model validation. Illustrate the importance of not training on future data when testing past performance.
Hyperparameter Tuning	Optimize key parameters for each model using techniques like GridSearchCV or RandomizedSearchCV. Provide parameter ranges for tuning (e.g., `alpha` for Ridge, `n_estimators` for Random Forest).
Regularization	Apply regularization techniques (e.g., Ridge, Lasso) to the linear models or post-pruning to tree-based models to address overfitting. Explain the trade-off between bias and variance.
Residual Analysis	Analyze residuals to check for patterns or autocorrelation. Plot residuals vs. time or predicted values to ensure no systematic patterns remain unmodeled.

Economic Interpretation	Link model results (e.g., feature importance, residuals) to financial theories or practical investment strategies. Explain why certain predictors consistently rank highly and their financial relevance.

3. Evaluation	Hints and Actionable Tasks
Performance Metrics	Evaluate models using MSE, RMSE, R^2, and MAE. Interpret metrics financially (e.g., RMSE indicates average prediction error in percentage points of excess return).
Out-of-Sample Performance	Test models on the reserved test set (2019–2023) to evaluate generalizability. Compare out-of-sample performance to training/validation results to check for overfitting or underfitting.
Visualization of Results	Plot actual vs. predicted returns for both models. Use line plots or scatterplots to visually assess model alignment with actual returns.
Comparative Analysis	Compare performance of the linear and tree-based models using metrics and visualizations. Highlight trade-offs (e.g., interpretability of linear models vs. accuracy of tree-based models).

4. Financial Analysis	Hints and Actionable Tasks
Key Predictor Analysis	Identify the most important predictors from the models and explain their economic relevance. Relate predictors (e.g., `tms`, `dp`) to financial theories like risk premia or market efficiency.
Predictor Stability	Assess the stability of predictor importance over time or economic regimes. Compare predictor ranks in expansion vs. recession periods to identify regime-specific insights.
Investment Strategies	Propose strategies based on model predictions, such as the timing strategy of asset allocation. Suggest actionable plans, e.g., long the S&P 500 index when excess returns are predicted to be positive.
Visualization of Insights	Use charts to illustrate the accumulated returns from the proposed investment strategies.