Proposal: Predicting Corporate Bond YTM Changes Using Machine Learning

Team members: Yuxi Geng, Yutung Lin, Qinqin Huang

1. Objective

The objective of this project is to predict corporate bond yield to maturity (YTM) changes using firm-level fundamentals, market-based risk factors, and macroeconomic indicators.

By applying various machine learning regression models such as Lasso, Ridge, ElasticNet, Decision Tree, and Boosting. We aim to identify the most important predictors of bond yield changes and the best-performing model for this task.

2. Methodology

2.1 Data Description

We will utilize data from Wharton Research Data Services (WRDS) to construct our dataset. The data frequence is monthly and data range is from July 2002 to February 2025.

2.1.1 Target Variable The target variable is the corporate bond yield to maturity (YTM) change, defined as the difference in YTM over a one-month horizon:

$$\Delta \text{YTM}_i = \text{YTM}_{i,t} - \text{YTM}_{i,t-1}$$

where $\text{YTM}_{i,t}$ is the yield to maturity of bond i at the end of month t.

The Data is from the "WRDS Bond Returns" database, which provides monthly bond yields and other bond characteristics, sourced from TRACE and Mergent FISD. The dataset has been cleaned and more suitable for analysis.

We only consider senior bonds and exclude defaulted bonds. Additionally, the yields exceeding [-1, 1] are treated as outliers and removed from the dataset.

Total number of observations meeting the criteria is approximately 1760852, covering 74191 unique bonds and 3187 unique firms over the sample period.

2.1.2 Features

1. Key Columns

- Bond Identifier: CUSIP
- Firm Identifier: Stock Ticker
- Date: End of Month

We have checked the missingness and uniqueness of these key columns.

2. Bond-level features

- Time to Maturity (tmt)
- Coupon Rate (coupon)
- t_spread: the weighted average bid-ask spread of the bond which measures liquidity.
- return_eom: bond return of last month.
- Credit Ratings: One-hot encoded variables for ratings from AAA to D (rating_A, rating_AA, ..., rating_D). The ratings are weighted averages ratings from WRDS Bond Returns database rather than S&P, Moody's, or Fitch alone.
- Rating Change Indicators: Binary variables indicating whether there was an upgrade or downgrade in the bond's credit rating in the past month (upgrade, downgrade).
- 3-Month Constant Maturity Treasury Yield (gs3m): We may use the difference between two months rather than the level.
- Term Spread (term_spread): Difference between 10-year and 3-month Treasury yields.

Here, we used lagged values of features except time to maturity and coupon rate to avoid look-ahead bias.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
tmt	3.56e+6	6.35	8.67	0.00	1.61	3.54	6.89	102.13
coupon	3.56e+6	0.03	0.03	0.00	0.00	0.03	0.06	0.34
t_spread	1.79e+6	0.01	0.01	0.00	0.00	0.00	0.01	2.00
yield	3.39e+6	0.01	0.13	-1.00	0.00	0.03	0.05	1.00
ret_eom	3.42e+6	0.01	1.18	-1.00	0.00	0.00	0.01	2090.0
rating_A	3.42e+6	0.20	0.40	0.00	0.00	0.00	0.00	1.00
rating_AA	3.42e+6	0.04	0.21	0.00	0.00	0.00	0.00	1.00
rating_AAA	3.42e+6	0.01	0.09	0.00	0.00	0.00	0.00	1.00
rating_B	3.42e+6	0.03	0.16	0.00	0.00	0.00	0.00	1.00
rating_BB	3.42e+6	0.04	0.20	0.00	0.00	0.00	0.00	1.00
rating_BBB	3.42e+6	0.23	0.42	0.00	0.00	0.00	0.00	1.00
rating_C	3.42e+6	0.00	0.02	0.00	0.00	0.00	0.00	1.00
rating_CC	3.42e+6	0.00	0.03	0.00	0.00	0.00	0.00	1.00
rating_CCC	3.42e+6	0.01	0.09	0.00	0.00	0.00	0.00	1.00
rating_D	3.42e+6	0.00	0.02	0.00	0.00	0.00	0.00	1.00
upgrade	3.42e+6	0.00	0.04	0.00	0.00	0.00	0.00	1.00
downgrade	3.42e+6	0.00	0.05	0.00	0.00	0.00	0.00	1.00
gs3m	3.42e+6	0.02	0.02	0.00	0.00	0.01	0.02	0.06
term_spread	3.42e+6	0.01	0.01	-0.02	0.00	0.01	0.02	0.04

3. Firm-level features

4. Macroeconomic features

2.2 Feature Construction

We will standardize continuous features to have zero mean and unit variance.

We may consider using dimensionality reduction techniques like PCA.

2.3 Modeling Framework

We will use models introduced in class and pick the best-performing one based on out-of-sample evaluation metrics. Additionally, we will train a linear regression model as a benchmark.

Model	Description
Ridge Regression	Penalizes large coefficients with L2 norm
Lasso Regression	L1 penalty for feature selection
ElasticNet	Combination of L1 and L2
Decision Tree	Non-linear split-based model
Boosting (LightGBM)	Ensemble of weak learners

2.4 Model Evaluation

We will use a time-series split to create training, validation, and test sets. We will calculate the following evaluation metrics on the test set to evaluate model performance:

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R^2

3. Implementation Plan

3.1 Logistics

- Programming Language: Python
- Platform: Laptop/Local Machine
- Online Repositories: GitHub for version control and collaboration

3.2 Schedule

1. Data Acquisition: Extract and merge datasets from WRDS.
2. Data Preprocessing: Handle missing values, outliers, merge datasets and standardize features.
3. Feature Engineering: Check correlations, apply PCA if needed.
4. Model Training: Split data into training, validation, and test sets. Train models with hyperparameter tuning via validation set.

5. Model Evaluation: Compare models using MSE, MAE, and R^2 on the test set.