Proposal: Predicting Corporate Bond Credit Spreads Using Machine Learning

Team members: Yuxi Geng, Yutung Lin, Qinqin Huang

1. Objective

The objective of this project is to predict corporate bond credit spreads 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 XGBoost. We aim to identify the key drivers of credit spreads and evaluate how well these models capture the cross-sectional variation in credit risk across firms.

2. Methodology

2.1 Data Description

We will obtain and merge the following datasets from WRDS:

Source	Dataset	Key Variables	Purpose
TRACE	Bond trade data	Bond ID (CUSIP), trade price, yield, maturity, coupon	Used to compute market-based bond yields and spreads
LSEG Mergent FISD	Bond characteristics	Issue date, rating, issue size	Bond-level descriptors
Compustat	Firm fundamentals	Leverage, profitability, total assets, interest coverage	Issuer financial strength
CRSP	Stock data	Stock return volatility, beta, recent returns	Market-based firm risk
Federal Reserve Board Reports	Interest rates	Treasury yields (1Y–10Y), term spread, VIX	Macroeconomic indicators
TBD	Macroeconomic data	Inflation rate, GDP growth	Economic environment

The target variable / outcome is corporates bonds' credit spread, defined as:

$$\text{Spread}_i = \text{Bond Yield}_i - \text{Treasury Yield (same maturity)}$$

Notice:

- We only consider bonds without embedded options to avoid complexities in pricing.
- If there is autocorrelation in the residuals, we may consider taking differences of the credit spreads to ensure stationarity.
- The key columns used to merge datasets will be determined based on availability but will likely include firm identifier, bond identifier and date.
- Data frequency is expected to be monthly or quarterly, depending on availability.
- Data range also depends on availability.

2.2 Feature Construction

The features used are still under discussion, but we will include below types of variables:

- Bond-specific features: maturity, coupon rate, issue size.
- Issuer-level features: leverage ratio, coverage ratio, ROA, total assets.
- Market-based features: stock return volatility, recent performance.
- Macroeconomic variables: Treasury yields, inflation rate, GDP growth.

We will standardize continuous features to have zero mean and unit variance. Categorical variables, such as credit ratings, will be one-hot encoded.

We may consider using dimensionality reduction techniques like PCA if the feature space becomes too large.

2.3 Modeling Framework

We will compare models introduced in class:

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
XGBoost	Gradient-boosted decision trees

Evaluation metrics:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R^2 (Out-of-sample)

Additionally, a linear model will be used as a benchmark for comparison.

3. Implementation Plan

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

Steps to be followed:

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 RMSE, MAE, and R^2 on the test set.