Data Mining Project

Problem Statement

* Global fixed income market totals about $130 trillion in outstanding debt (compared to $42 trillion in global equities market)
* Fixed income products (bonds) are denoted with a rating given to them by one of 3 credit agencies (Moody’s, S&P, and Fitch)
* These ratings provide investors with a gauge of the level of risk associated with investing in a particular bond
* Higher rated bonds are classified as Investment-Grade and lower rated bonds are classified as High Yield
* For bond traders and financial market participants, being able to predict corporate bond rating changes and corporate bond ratings for newly issued bonds is very important
  + Can lead to being on the correct side of trades (Ex: By being able to predict a ratings downgrade or upgrade can position oneself ahead of time to profit as the rest of the market reacts in real-time)
* Objective: Using attributes regarding the financial health of a company, we want to create models to accurately predict and classify bond ratings within the High-Yield bond market
* Importance: By being able to predict and classify bond ratings we can systematically determine rating changes and new issuer ratings using company financial information

Main Problem: Ratings Prediction

* Using attributes regarding the financial health of a company, we want to create models to accurately predict and classify bond ratings within the High-Yield bond market
* Skills: Machine Learning Classification Models (basic techniques and more advanced techniques)

Sub-Problem 1: Feature Importance

* Given the large set of attributes and features from the dataset, can we obtain a smaller subset of attributes that provide the most information and are most important in predicting a bond’s rating
* This is important as we can see which qualities and attributes of a company need to be paid attention to/accounted for most when evaluating a bond for its rating
* Skills: EDA, Data Cleaning, Feature Engineering, Random Forest, Logistic Regression, PCA

Sub-Problem 2: Bond Clustering

* Given the subset of attributes we deem most important, can we create an unsupervised machine learning clustering model to group bonds based on these characteristics? Do these groups correspond to the ratings classification and how accurate are these clusters in terms of showing rating groups?
* This is important as we can potentially have another method at determining a bond’s rating by grouping similar bonds together that we can compare to our supervised machine learning models
* It also shows if our features are aligned in a manner that matches with bond ratings well
* Skills: Unsupervised learning, clustering analysis

Data

* The data is collected from Intercontinental Bond Data Index (ICE) from BoFa which tracks more than $100 trillion in global market debt
* We specifically are using HY Index data from September 1, 2023 (HY Index includes only bonds rated BB+ and lower according to S&P rating system
* Index includes 1866 unique bonds (rows) and 138 unique features (columns)
* Glossary of data fields can be found here: https://www.ice.com/publicdocs/data/Bond\_Index\_Methodologies.pdf (page 73)

Data Preprocessing

1. Feature Selection and Cleaning
   1. Drop Columns with high percentage of missing values (> 50%)
      1. 5 columns with 100% NaN (WAM, MUNI\_STATE, WAC, WALA, CPR)
   2. Drop Constant Features (columns with same value in each row)
      1. INDEX\_NAME (All values are H0A0)
      2. INTEREST\_CASH (All values are 0)
      3. KDR\_TRSY\_50\_YRS (All values are 0)
      4. CASH (All values are 0)
      5. MLINDLVL1\_CODE (All values are CORP)
      6. KDR\_SWAP\_50\_YRS (All values are 0)
      7. FACTOR (All values are 1)
      8. ML\_INDUSTRY\_LVL\_1 (All values are Corporate)
      9. ISO\_CURRENCY (All values are USD)
      10. PRINCIPAL\_CASH (all values are 0)
      11. DESCRIPTION2 (All values are ICE BofA US High Yield Index)
      12. REDEMPTION\_CASH (All values are 0)
      13. PAYDOWN\_RETURN\_MTD (All values are 0)
      14. AS\_OF\_DATE (All values are 9/1/2023)
      15. CASH\_VALUE (All values are 0)
   3. Drop CUSIP and ISIN\_NUMBER columns
      1. CUSIP and ISIN\_NUMBER are unique identifiers for a bond. Every bond has a unique CUSIP and ISIN\_NUMBER, therefore this does not provide us with any relevant information for our models
   4. Drop columns with Duplicate Information in a different form
      1. DESCRIPTION and TICKER have the same information (drop DESCRIPTION)
      2. ML\_INDUSTRY\_LVL\_2 (3,4) and MLINDLVL2\_CODE (3,4) same information, drop the ML\_INDUSTRY\_LVL columns
   5. Deal with Missing Values
      1. For numerical columns: fill with column average
      2. For categorical columns: fill with most occurring value (mode)
   6. Check for 100% correlation among independent variables (multicollinearity)
      1. Multicollinearity is a major issue in machine learning and arises when features are highly correlated to each other
      2. This is a problem because it undermines the significance of an independent variable and inflates standard errors of regression coefficients
      3. First check: See which pairs of features have 90% direct or inverse correlation or more and then remove one feature from each unique pair (went from 138 features including target to 49 features including target)
   7. One Hot Encoding to transform categorical features to numerical features
      1. One hot encoding converts categorical features to numerical features by adding new columns for each different categorical value and using binary variables to fill them
      2. Most machine learning algorithms require categorical data to be converted to numerical data (now have 2415 columns)
2. Feature Scaling
   1. Using one-hot encoded data, perform a min-max normalization
3. Feature Engineering and Dimensionality Reduction
   1. Perform Train-Test Split (70% Train, 30% Test)
   2. Logistic Regression Feature Importance
      1. Run a basic logistic regression on train data and look at coefficients
      2. Larger coefficients (absolute value) have the most effect on the prediction
      3. Reduce dimensionality by extracting 10% of features with largest coefficients (now have 214 columns)
   3. PCA (Principal Component Analysis)
      1. From just looking at most important features, we still have high dimensionality (214 columns)
      2. We can reduce dimensionality and still explain a high amount of the variance of the data while simultaneously reducing multicollinearity using a PCA
      3. Set an explained variance threshold of 80% and see how many principal components the data will reduce to (Reduces to 44 components)

Main Problem

Baseline Models

1. K-Nearest Neighbors

* Cross-Validation (Elbow method) for choosing optimal K
* Grid Search for Optimal Hyperparameters (weights, distance metric)
* Error Metrics: Confusion Matrix, Accuracy, Precision, Recall, F1

2. Gaussian Naïve Bayes

* Error Metrics: Confusion Matrix, Accuracy, Precision, Recall, F1

3. Classification Tree

* Grid Search for Hyperparameter Tuning
* Error Metrics: Confusion Matrix, Accuracy, Precision, Recall, F1

Advanced Models

1. Random Forest Classifier

2. SVM Classifier

3. Neural Network

Final Subproblem

* Clustering analysis – can we group bonds by the most important features and see if our groupings correlate with the actual bond ratings
* Important because we can look at summary statistics of groups and subsequently determine what characteristics make up each rating group
* Gives another method at predicting rating groups