Assignment 9: Pricing Loan Spread Using ML

Overview (Weight: 10%)

Objective: Perform feature engineering and construct ML model for pricing loan spread

The goals of this assignment are for the students to

- Perform feature engineering using financial ratios, market variables, and macro variables for corporate finance ML tasks.
- Understand the use of auxiliary tasks in ML models.
- Critically evaluate model outputs, focusing on drawing meaningful conclusions and insights from the analysis, rather than just implementing the programming.

Required Files for Submission

- A **Jupyter notebook** (.ipynb) containing your code, clearly commented and organised.
- An **HTML or PDF** output generated from the notebook showing the results.
- A **five-page PDF** summarising your analyses and findings.

You do **not** need to submit the raw data files.

Data

You need to use data from 1996 to 2022.

1. Loan Data

- The name of the file is *loan_pricing_dealscan.csv*
- Description of the variable is provided in *Facility - Variable Definitions.pdf* and *Package Variable Definitions.pdf* files
- It has "PERMNO" and "gvkey" identifier to connect with COMPUSTAT and CRSP data
- See the definitions of industry code provided at the end of the assignment
- "allindrawn" (spread on the drawn amount in bps) is variable we are trying to estimate and "allinundrawn" (spread on the undrawn amount in bps) will serve as auxilary task variable so please **do not** use them as feature for model.

2. COMPUSTAT Annual Fundamental Data

- The name of data file is *COMPUSTAT_funda_annual.csv*
- "gvkey" is unique identifier for a company in COMPUSTAT data
- "fyear" is financial year of the company
- Description of all the variable is provided in *COMPUSTAT_Variable_Definition.pdf*

3. CRSP Monthly Stock File (MSF) Data

- The name of data file is *MSF_1996_2023.csv*
- "PERMNO" is unique identifier for a company in MSF data
- It is important that you use the variables you computed from previous assignments. For example, systematic volatility, idiosyncratic volatility, beta, etc.

4. Macro-economic data

- Use fed fund rate and other macro-economic indicators from FRED. Collect variables beyond what you collected for the SIFMA assignment.

Note: To reduce work (computational effort), for each year, instead of analyzing the entire set of firms, select a sample of 100 firms randomly. Use industry categories as dummy feature variable. - Divide the sample into in-sample estimation period (1996 to 2017) and out of sample estimation period (2018 to 2022)

Assignment

Feature Engineering

- **Qualitative Feature Selection:** From each of the four datasets listed above, make a list of 30-40 variables that you believe would be most valuable for estimating loan spread.
  - Select 5-10 variables from loan data.
  - Select 20-25 variables based on FinRatio assignment (AS3).
  - Select 2-3 variables based on MSF data and your work for AS4, AS5, and AS6.
  - Select 3-5 macro economic indicators (also used in AS2).
- Provide reasoning why you selected them.
- Optional Step: Read the suggested papers in the "Supplementary Readings" section to develop better understanding of loan pricing.
- **Quantitative Feature Selection**
  - Calculate the correlation matrix for the variables you selected. (do not include macro economic indicator for correlation)
    - Would you group them differently?
    - Based on the analyses what variables you would think of dropping? Provide reasoning
  - Get NBER recession data. Compute the correlation matrix for NBER recession = 1 and NBER recession = 0. What do you notice?
- **Feature Scaling**: Based on your observation about the distribution of each feature, think about which features need feature scaling.
  - standard feature scaling methods are:
    - Normalization: Scale features to a range (e.g., 0 to 1)
    - Standardization: Transform features to have zero mean and unit variance
    - Log Transformations: Apply log transformations to skewed features to reduce the impact of extreme values

Construct MLP to Estimate Drawn Spread

- In the previous "Neural Beta" assignment, we covered the basics of constructing a Multilayer Perceptron (MLP).
- For this task, use **"allindrawn"** as the outcome variable to train and test an Artificial Neural Network (ANN) model.
- Perform hyperparameter tuning for the following configurations:

  - **Number of neurons** in the hidden layer (consider three different values).
  - **Learning rate** for the optimizer (consider three different values).
  - **Activation functions** for the hidden layer (`'linear'`, `'sigmoid'`, `'tanh'`, `'relu'`).

- Create a plot showing the loss over epochs for the best-performing hyperparameters.
- Finally, evaluate the MLP model on the test dataset and report the performance metrics.

Introducing an Auxiliary Task

Now, keep the model architecture and features almost the same as the previous part, but include an **auxiliary task** using the **"allinundrawn"** variable as an additional output. This auxiliary task will allow the MLP to predict both "allindrawn" (primary outcome) and "allinundrawn" (auxiliary outcome) simultaneously.

- **Implementation Steps**:
  - Modify the existing MLP model to have two neurons at output layer:
    - One for predicting **"allindrawn"**.
    - One for predicting **"allinundrawn"** as the auxiliary task.
  - Ensure that the loss functions for both outputs are optimized jointly. Use a weighted sum approach to balance the importance of each task.
  - Report the best calibrated weight for each loss function.
- **Run the Model**: Train the modified MLP with the auxiliary task included, and document the performance for both the main and auxiliary outputs.

Analysis: Impact of Auxiliary Task on the Main Task

After training the MLP with the auxiliary task, compare the results with the model trained on the main task alone ("allindrawn").

- **Questions to Address**:
  - How did including the auxiliary task affect the performance of the main task?
  - Did the loss for "allindrawn" improve or degrade?
  - Are there any noticeable changes in the feature importance or learned representations?
  - Discuss the potential reasons behind the observed impact and provide your conclusions.
- Create a comparative table and visualizations to clearly show the differences between the two models.

Compare the performance of various ML models

- For the original task of just estimating **"allindrawn"**, train the following models:
  - LASSO regression
  - XGBOOST
  - LIGHTGBM
  - KNN
- Run the regression model for the test data
- Compare the performance of MLP algorithm with the models built. Which one has the best performance?

Supplementary Details

Readings

- Lending to Innovative Firms (Link)
- Environmental Externalities and Cost of Capital (Link)
- Do Shareholder Rights Affect the Cost of Bank Loans? (Link)

Industry code:

SIC Code	Industries
1 – 999	Agriculture, Forestry and Fishing
1000 – 1499	Mining
1500 – 1799	Construction
2000 – 3999	Manufacturing
4000 – 4999	Transportation and other Utilities
5000 – 5199	Wholesale Trade
5200 – 5999	Retail Trade
6000 – 6799	Finance, Insurance and Real Estate
7000 – 8999	Services
9000 – 9999	Public Administration