Predicting Probability of Default

Data Mining and Machine Learning Techniques to Optimize Loan Origination for Banca Massiccia

Group Indigo:

Ilias Arvanitakis
Yasi Asgari
Alex Herron
Joseph Schuman
Alexandre Vives

11.22.2022

Business Understanding

What is the Problem?

Problem: Banca Massiccia loan default risk

- Banca Massiccia is an Italian lender, making loans for businesses for a long period of time
- Goal is to optimize the the loan origination process, to better estimate the default risk of the borrower
- Machine learning will be introduced as the main tool used to address this problem

Loan origination process

- Borrower applies for a loan, providing financial statements
- The statements are evaluated to estimate the default risk
- The interest rate and fees are determined using the default risk
- If the loan is considered profitable it is eventually originated
- In the event of default the shortfall is recovered through various procedures

Deliverable

- Utilizing financial statements, a machine learning model should be able to determine the probability of default of the firm within the next 12 months
- The model will be able to determine the probability of default over the next 12 months using the financial statements from previous years for corporate clients

Data Understanding

What does the data look like?

Data Overview

- Financial statement data with the following information:
  - Each row consists of a company's financial statement
  - A total of 44 variables, including both numerical and categorical
  - All quantities are expressed in euros
  - Only firms with > €1.5MM in assets are included
  - Only non-finance/non-insurance firms are included

Financial statement data

- Balance sheet data include assets, liabilities, equity and some of their components such as bank debt, accounts payable, tangible and intangible assets, cash and equivalents, etc.
- Income statement data include operating income, gross profit, taxes and net income as well as sub components such as interest, expenses, cost of goods sold etc.

Non Financial Data

- Headquarter city, legal structure and ATECO sector

Data Understanding

What is the meaning of our data?

Understanding Balance Sheet

- A balance sheet shows how a company finances its assets. The composition of debt and equity is crucial to determine the default probability.
- Sub-components of the balance sheet like fixed assets vs current assets can help us understand if the company is in a position to satisfy its liabilities by having easy access to liquidity.
- The composition of long term vs short term debt is crucial, since it indicates the need for immediate liquidity to satisfy liabilities.

Understanding Income Statement

- Profitability in its various levels is a major indicator outlining the capability of the company to generate income.
- Net profit and EBITDA are two very important measures.
  - Net profit is the total income generated by the company for its shareholders.
  - EBITDA is the profit of the company from its main direct and indirect operations. It is an indicator of an efficient cost structure.
- For our project objective, interest expense is highly important since it indicates the capability of the company to pay interest

Data Understanding

The use of ratios in financial data

Comparability of Data

- Comparing financial statements of different companies is intractable due to the potential difference in size.
- To address the issue above, financial ratios are introduced (also useful to compare data within the same firm).

Major Ratios

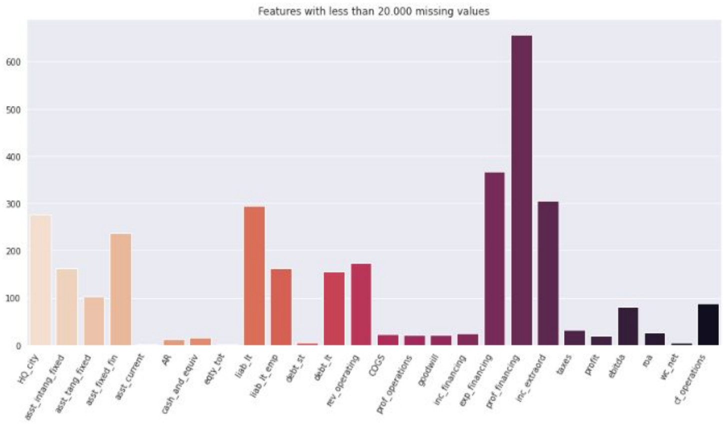
- **Net interest margin:** Represents the ability of a firm to cover interest expenses. A ratio above 12.5 is considered excellent.
- **Debt-to-equity ratio:** Represents the firm’s exposure to creditors. High levels indicate reduced ability to generate capital.
- **Leverage ratio:** The portion of the assets financed by equity. Low levels of leverage indicate the company uses its own capital for asset purchases.
- **Current ratio:** The firm’s capability to cover short term liabilities from liquid assets.
- **Return on equity:** The firm’s capability to compensate its shareholders.
- **Net profit margin:** The profit of the company as a percentage of sales. Indicates the efficiency of the company to generate income through all of its financial and operational activities.
- **Cash flow-to-debt ratio:** The ratio of a company’s cash flow to its total debt. Can be used to determine how long a company would need to repay its debt.

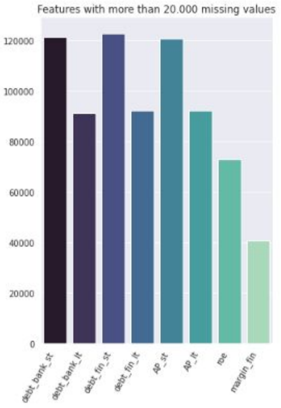
Data Preparation

Handling missing data

Treating NaN Values

- Several features had missing values, ranging from <1% to 12% of the total.
- We decided to follow the steps below to fill in the missing values:[1]
    a. Use accounting equations to fill them in with an exact value.
    b. Impute the mean of the values from *previous* statements within the same company.
    c. Impute the median of the values from similarly sized companies within the same ATECO sector.





Data Preparation

Dealing with the look-ahead bias

Defining a Firm year

- A firm year starts in January and ends in December, but financial statements are not reported until March or April.
- If that is not taken care of, we will inadvertently peek into the future. I.e., use data that we would not actually have had at the time of the prediction.

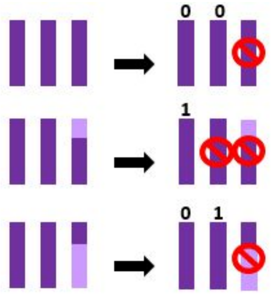
Preventing look-ahead bias

- To prepare the dataset for training, we identified the firms that defaulted and flagged the row that preceded the default. In order to emulate the timeline of available data in a live environment, we needed to identify what would have been the most recent financial statement for a firm, given that they would subsequently default within the next year.
- For companies that defaulted before May, the financial statement of the previous year was dropped (it would not have been published at the time of default) and the default flag was assigned to the previous financial statement.
- For companies that defaulted in May of afterwards, the default flag was assigned to the financial statement of the year before the default (the reports would have been published by that time, so they are able to be used by the model).

Company did not default:

Company defaulted before May:

Company defaulted between May and Dec:



▮ Represents a statement

▮ Represents a statement that defaults before May

▮ Represents a statement that defaults in May or afterwards

🚫 Statement dropped

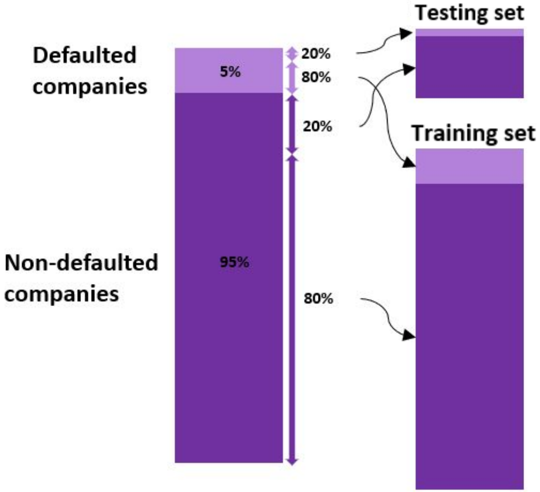
0: Non-default flag assigned to statement
1: Default flag assigned to statement

Modeling the Data

Types of Models

Train Test Split:

- Taking into account that the dataset is unbalanced, we decided to keep the proportion of defaults/non-defaults on the train/test split. The split was based on company IDs, following the figure on the right. Specifically, the training set was assembled from the statements from 80% of the defaulting companies and the statements from 80% of the non-defaulting companies. On the other hand, the testing set was created using all statements from 20% of the defaulting companies and the statements from 20% of the non-defaulting companies.



Models Used:

- Since this is a classification problem with a categorical dependent variable and mainly numeric independent variables, we trained the following models: Logit, Decision Tree Classifier, and XGboost.[2] According to Wang, Y.J. (2011), prior to the introduction of probability-based models, MDA models were the most commonly used on default predictions papers even though it is a static model (can only predict one step ahead), followed by Logit models.

Modeling the Data

Choices for the data mining algorithm

Alternatives:

- One alternative model we considered was Random Forest since it is robust to outliers and can be deployed in a classification setting. However, the Random Forest algorithm can be computationally intensive for large datasets so we decided not to use it.
- We chose to not use neural networks either due to a lack of explainability and loss of economic intuition, which could make our model hard to explain to our key stakeholders.

Final Decision:

- We experimented with a Logit Model[2] because it provides complete visibility into its coefficients, a Decision Tree Classifier because it is easily interpretable, and finally XGBoost.
- We ultimately chose **XGBoost** because it avoids overfitting and has the best combination of prediction performance and processing time.

dmlc
XGBoost

Modeling the Data

Variables Used

Prediction and Dependent variables:

- We curated our prediction variables based off of the financial ratios we created in our feature engineering step. They were: 'gross margin rate', 'current ratio', 'debt to equity ratio', 'net profit margin', 'leverage', 'interest coverage ratio', 'quick ratio', 'cash ratio', 'receivables turnover ratio', 'debt to asset ratio', 'debt to capital ratio', and ‘cash flow to debt ratio’.
- Our dependent variable was set to be a 1 if a company defaulted within a year and a 0 if it did not.

Benefits of Feature Engineering and Economic Intuition:

- By creating financial ratios to utilize as our predictor variables, we enable a more consistent comparison between companies
- For example, a larger company is bound to have more debt than a smaller company, and so feature engineering through ratios helps us account for this variability.
- Instead of simply using “debt” as a predictor variable, we are using the ratio of a company’s size to it’s debt.

Evaluation

Evaluation of Data Engineering and Model

What we Evaluated our Data on:

- The results from our models were evaluated based off of their performance on our test data set, which had 20 percent of our total data (split based on company IDs, and stratified to ensure an equal distribution of defaulting/non-defaulting companies between our train and test sets)
- This helped us confirm we were not overfitting on just our training data

Useful Metrics:

- In this setting, we have an imbalanced dataset as we do not have nearly as many defaulting companies as we do non-defaulting ones
    - Hence, accuracy is not the best metric to evaluate model performance
- Instead, we primarily look at the recall, which can be a better representation of performance, given the potential financial loss associated with giving loans to riskier companies
- Additionally, we also evaluated our models using AUC

Evaluation

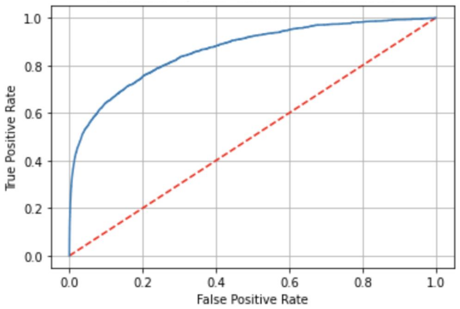
Benchmarks for Performance

Performance:

- For all three models, we took a look at the F1 score, precision, recall, and the area under the curve.
- AUC tells us the ability of our classifier to distinguish between classes (default and non-default) and is used as a summary of the ROC curve.
- We found that the XGBoost model had the best F1 score and AUC, followed by the decision tree classifier.

Metrics for best-performing model (XGBoost):

- Precision: 0.052
- Recall: 0.631
- F1 score: 0.0958
- AUC: 0.817



$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}.$$

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

Evaluation

How Can the Business Case be Developed?

Application of our Model:

- The most important metric that Banca Massiccia should take into account is recall, defined as the number of correctly classified statement defaults divided by the total statement defaults. The reason why this metric is important is based on the risk of making a type I error versus a type II error.
- Giving a loan to a company that was wrongly classified as a non-defaulting company is much riskier (in terms of potential financial loss) than refusing to give a loan to a company wrongly classified as a defaulting company.
- We can move forward with creating a business case by setting a threshold via cost-benefit analysis to find the value of each model.
- Next, we can assess how our model improves Banca Massiccia's overall revenue by examining how their interest rates change as well as the amount of loans they end up giving out.
- Additionally, we can compare the cost of our XGBoost model to the previous methodology the Bank was using.

Deployment

How will these results be deployed?

When can these results can be used:

- The results of this model will be deployed by Banca Massiccia to assist in estimating risk associated with new loan origination
- This tool can be used to enhance the Bank's corporate loan pricing strategy, both in setting interest rates and determining underwriting fees for loan origination
- The results can also be used to inform decisions related to approval or denial of loan applications

How to deploy the model:

- When deploying the model, input data must be pre-processed via the methodology applied to the training data (e.g., filling missing values using prior financial statements)
- The harness function will implement all necessary steps to align new input data with the model and produce a corresponding output

Deployment

Risks and limitations Results:

Limitations:

- **Consider Impact & Pricing**: The results of this model should not be used as a standalone tool to determine whether a loan application should be approved or denied. The model is agnostic to whether or not a specific loan should be approved. Pricing and loan approval decisions must take into account factors such as the financial impact of loan origination.
- Banca Masciccia must independently determine its risk appetite, capital requirements, and strategic risk profile in order to determine how to apply these results.
- **Out of sample estimates:** The results may not be viable if applied in situations outside the scope of the original data mining exercise
- The modeling results were developed with historical data from medium and large corporate clients. The validity of results applied to potential clients that fall outside the purview of the parameters of the training data cannot be substantiated. This includes:
    - Firms with < €1.5MM in assets
    - Financial firms & Insurance companies
    - Non-Italian corporate clients
- **Short Horizon:** The model has been developed to predict default probabilities in the near term (within the next year). These results have not been validated to inform risk based decisions extrapolated beyond 1 year.

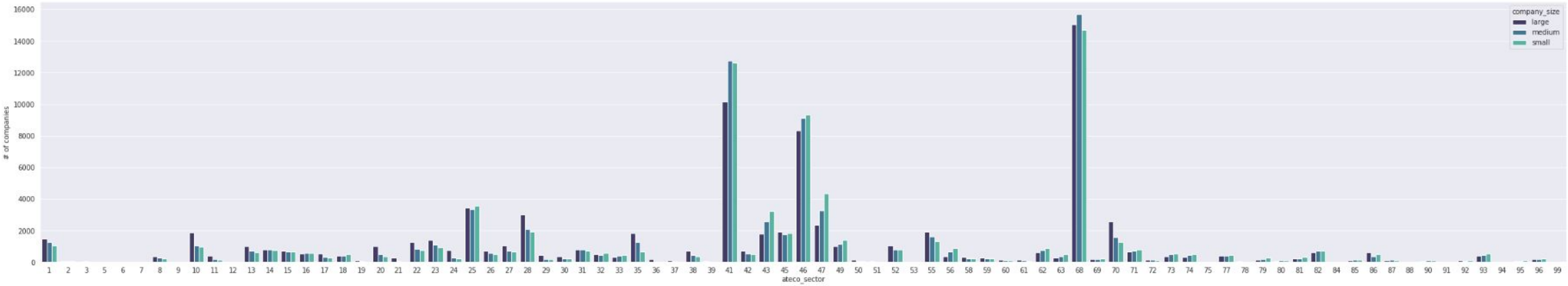
Appendix

Contributions of each team member

Individual contributions:

- Ilias Arvanitakis: Feature engineering with ratios, Data understanding & data preparation writeup
- Yasi Asgari: Harness, testing the harness, modeling and evaluation write up
- Alex Herron: Exploratory data analysis, train_model file, harness file, modeling and evaluation write up
- Joseph Schuman: Supporting documentation for Modeling and Data Preparation methodology, Deployment writeup
- Alexandre Vives: Dealt with the data cleaning (filling in missing value and column formatting), worked on the stratified train/test split and the one-hot encoding of the features.

Appendix



This graph shows the quantity of companies in each ateco sector and company size group pair. This groups were used as the last step to fill in any missing values.

Bibliography:

1. Blanchet, J., Hernandez, F., Nguyen, V. A., Pelger, M., & Zhang, X. (2022). Bayesian Imputation with Optimal Look-Ahead-Bias and Variance Tradeoff. arXiv preprint arXiv:2202.00871.

2. Wang, Y.J. (2011). Corporate default prediction : models, drivers and measurements.

3. Bryzgalova, S., Lerner, S., Lettau, M., & Pelger, M. (2022). Missing Financial Data. Available at SSRN 4106794.

4. Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning, 77(1), 103-123.

5. Hand, D. J. (2006). Classifier technology and the illusion of progress. Statistical science, 21(1), 1-14.

6. “Precision and Recall.” Wikipedia, Wikimedia Foundation, 16 Nov. 2022, https://en.wikipedia.org/wiki/Precision_and_recall.

7. “F-Score.” Wikipedia, Wikimedia Foundation, 6 Nov. 2022, https://en.wikipedia.org/wiki/F-score.

8. “XGBoost Documentation” XGBoost Documentation - Xgboost 1.7.1 Documentation, https://xgboost.readthedocs.io/en/stable/.