NYU

Team Periwinkle

Xinchen Zhang, Sunny Yang,
Alon Florentin, Jaylen Peng

11.18.24

Table of Contents

Business Problem

- Accurately estimating the Probability of Default (PD) is the cornerstone of credit risk management, allowing for accurate loan pricing and portfolio management. Machine learning models offer banks like Banca Massiccia a more powerful way to effectively manage and minimize their credit risk.
- The business problem is that Banca Massiccia needs to more accurately predict the probability that a single prospective borrower will default on a loan (either the principal or the interest payment) within the next 12 months.
- More accurate PD estimates will allow the bank to better assess the true risk associated with each loan, and optimize loan pricing by setting suitable interest rates and underwriting fees.

Why a Data Mining Solution:

- A data-mining solution allows the bank to improve their identification of high-risk borrowers to mitigate default risks.
  - ML models are able to learn relationships that may be incredibly difficult to find using traditional methods of risk calculation (i.e. manual expert analysis determining rules based on experience / theory).
  - The mechanical nature of a data mining solution also reduces the effects of potential biases on decision-making, allowing more loan officers to be better aligned with bank guidelines for risk mitigation.
- Additionally, a data mining approach allows extremely large datasets to be handled easily and efficiently, which greatly speeds up the loan approval process with automated predictions.

Evolution of PD Models

- Horrigan (1968) utilized key financial ratios such as Debt-to-Equity and Current Ratio to assess a borrower's creditworthiness. This provided a static snapshot of financial health. However, these basic ratios lacked adaptability to changing market environments and complex borrower profiles. [1]
- Kocagil et al. (2007) of Fitch Ratings introduced a hybrid model combining option-based barrier models with financial and market information to estimate PD. They provided a forward-looking assessment of credit risk by incorporating market data. [2]
- Lopez (2002) analyzed the relationship between firm asset size and default probability, demonstrating that smaller firms tend to have higher probabilities of default. This enhanced our understanding of how firm size influences credit risk, leading to more accurate default prediction models. [3]

Our Approach:

- Comprehensive Financial Ratio Analysis:
  - We calculated financial ratios encompassing leverage, liquidity, profitability, and efficiency and eliminated collinear ratios to ensure robust and reliable predictive power. [1][4]
- Firm Size Segmentation:
  - We divided firms into three size-based groups using total assets, ensuring equal distribution across quantiles.
  - Identified that smaller firms have significantly higher PDs, validating the necessity of size-based segmentation. (see Figure 6 in appendix)
- Modeling:
  - Implemented logit model as a baseline model for its simplicity and interpretability.
  - Implemented ensemble models for each size group and the total dataset, resulting in nine models.

NYU

Data Understanding:

- We used the dataset provided by Banca Massica, which had ~1.02M rows and 44 columns. There was a substantial amount of missing data, and the balance sheet did not balance out properly for every row.
- After examining the data and conducting literature review, we introduced 14 financial ratios in addition to the existing columns. We will cover our pre-processing and our feature selection in more detail in the next section.
- There are some biases and limitations of this dataset: notably, only previous borrowers were included in the data, as the bank did not retain data on potential borrowers that were rejected. Additionally, we are unaware of what criteria the bank used to reject/accept potential borrowers previously. Thus, we may have missed information about important indicators of default.

Data Understanding (cont.):

- The dataset is comprised primarily of Italian firms (only 237 firms classified as extraterritorial firms). If the prospective borrower the bank wants to evaluate is extraterritorial (i.e. a U.S. firm), our model may be less effective.
- The dataset is also biased towards manufacturing (~25%), wholesale/retail (~19%), real estate firms (~18%), and construction firms (~18%); thus, the model may perform better for these sectors, and worse for other sectors.
- Finally, the time period for this dataset is from 2007-2012; notably, this is a state of a worsening economic environment, as Italy struggled to deal with the economic crisis in 2008 and had slowed economic growth in the following years. Thus, the model may perform worse when used in different states of the overall economy (i.e. may predict higher default rates in economic upturn). [5][6]

Target Variable:

- To generate the labels for our dataset, we assumed the financial statements for a fiscal year are not available until May of the following year.
- Therefore, to avoid peeking into the future, we shifted the fiscal year in such a way that the financial statements from one year only apply to predictions (of a one-year PD) starting in June of the following year.
    - For example, the financial statements for a fiscal year ending on “2005-12-31” will be used to predict default from “2006-06-01” to “2007-05-31”.
- Consequently, the binary target variable was defined accordingly, based on whether the default date falls within the prediction period.
    - Additionally, if the ‘def_date’ value was NaT, we assumed that was a non-default. The target variable was encoded into a column called ‘label’ (0 for non-default, 1 for default).

NYU

Feature Selection:

- Based on previous work from Pederzoli & Torricelli (2010), we have introduced these 14 ratios to quantify different areas of a firm's financial profile (leverage, liquidity, efficiency, profitability, and solvency), after checking for multicollinearity between these ratios. [4]
  - We tested each of these ratios using a univariate logit model, and all were found to be individually statistically significant.

Metric (formula)	Category
Debt to Equity $(\frac{debt\_lt+debt\_st}{eqty\_tot})$	Leverage
Equity Ratio $(\frac{eqty\_tot}{asst\_tot})$	Leverage
Long-Term Liabilities to Assets $(\frac{liab\_lt}{asst\_tot})$	Leverage
Cash & Securities to Assets $(\frac{cash\_and\_equiv}{asst\_tot})$	Liquidity
Current Ratio $(\frac{asst\_current}{debt\_st+AP\_st})$	Liquidity
NWC Ratio $(\frac{wc\_net}{asst\_tot})$	Liquidity
Quick Ratio $(\frac{cash\_and\_equiv+AR}{debt\_st})$	Liquidity
Asset Turnover $(\frac{rev\_operating}{asst\_tot})$	Efficiency
Current Assets to Sales $(\frac{asst\_current}{rev\_operating})$	Efficiency
Days Receivable Turnover $(\frac{AR}{rev\_operating} \times 365)$	Efficiency
Operating Margin $(\frac{prof\_operations}{rev\_operating})$	Profitability
EBITDA to Assets $(\frac{ebitda}{asst\_tot})$	Profitability
Cashflow to Debt $(\frac{cf\_operations}{debt\_lt+debt\_st})$	Solvency
Interest Coverage Ratio $(\frac{prof\_operations}{exp\_financing})$	Solvency

Table 1: Financial Ratios by Category

NYU

Data Preparation:

- Our first step was to try and balance the sheet; however, we found that many of the columns did not balance properly, either due to NA values or differences in calculations (i.e. companies may include different values based on their accounting practices). We were able to calculate the values for 5 columns consistently and accurately (>92% match rate):
  - ‘prof_operations’, ‘roe’, ‘wc_net’, ‘margin_fin’, ‘cf_operations’
- We then dropped columns that had substantial amounts of missing values, or could be accounted for using a different column:
  - We used ‘debt_st’ instead of {‘debt_bank_st’, ‘debt_fin_st’, ‘AP_st’}, and used ‘debt_lt’ instead of {‘debt_bank _lt’, ‘debt_fin_lt’, ‘ AP_lt’}.
  - We dropped ‘days_rec’ (> 72% NA) and ‘eqty_corp_family_tot’ (100% NA).
- We also dropped ‘HQ_city’ as it could be a source of bias (granting loans based on region may be discriminatory).

Data Preparation (cont.):

- Next, we quantized all numerical features for two reasons: first, it allows us to handle outliers without dropping data, and second, it increases our interpretability: we can see how each variable affects the PD across bins.
- We also one-hot encoded two categorical features (drop_first = True, to avoid multicollinearity): ‘ateco_sector’ was mapped with 21 categories (however, we only saw 19 during training), and ‘legal_struct’ was mapped to {0, 1}, based on whether firms were required to publish financial statements.
    - If some categories are not seen during the training (category K and T), it will be converted to 0’s across all the transformed columns.
- We then created two versions of the dataset, one for our baseline logit model and the other for our LightGBM model.
    - For the baseline dataset, we imputed NA values using KNN imputation (which also benefits from our quantization). No imputation is required for our LightGBM model as it handles NA values inherently.

Data Preparation (cont.):

- Our pre-processed datasets consist of ~1.02M rows and 62 columns, including the target column ('def_date').
- We then further split a copy of each version of the dataset into 3 groups, based on the quantized 'asst_tot' (size of firm): {Small, Medium, Large}.
- Overall, we have 8 datasets ready for our modeling approach:
  - 4 datasets for each model: one overall dataset, three grouped datasets
- Notably, there are several limitations for how we prepared the dataset:
  - Since we split our dataset on total assets (size), each grouped model did have less data to train on, which could affect performance.
  - Additionally, there is a class imbalance between the grouped datasets, given that smaller firms are more likely to default. (Figure 6 in appendix)
  - Finally, we imputed data to use for our logit model; this may lead to the model learning an erroneous relationship between some variables and ultimately affect model performance.

Modeling:

Like RiskCalc, before the actual modelling, a peek into how binned ratios are indicative of average PD already shows promising signs.

That is, the direction of change of most of them follows economic intuition (e.g., if ‘debt_to_equity’ is higher, then the company is more likely to default, regardless of size).

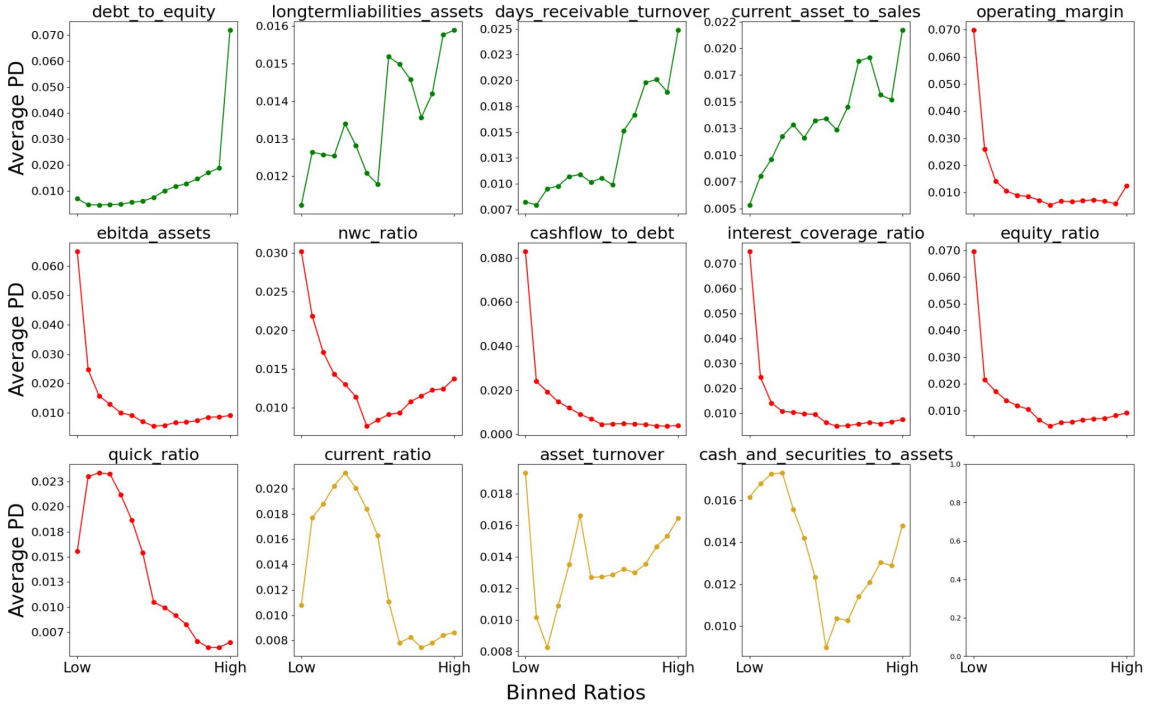

Figure 1: Binned ratios and PD

Modeling (cont.):

- Baseline model:
  - 3 logit models trained separately for Small, Medium and Large sized company
  - Pros/Cons: more interpretable but not as strong
- Proposed model:
  - An ensemble comprising of 9 LightGBM models: 3 different sizes x TimeSeriesSplit(3) within each size
  - Done in a walk-forward fashion to ensure there is no data leakage during training
  - Each sub-model is weighted by the number of samples it is trained on and the final prediction is the weighted average of the sub-predictions.
  - Pro/Cons: more performant, though less interpretable (see later for remediation!)

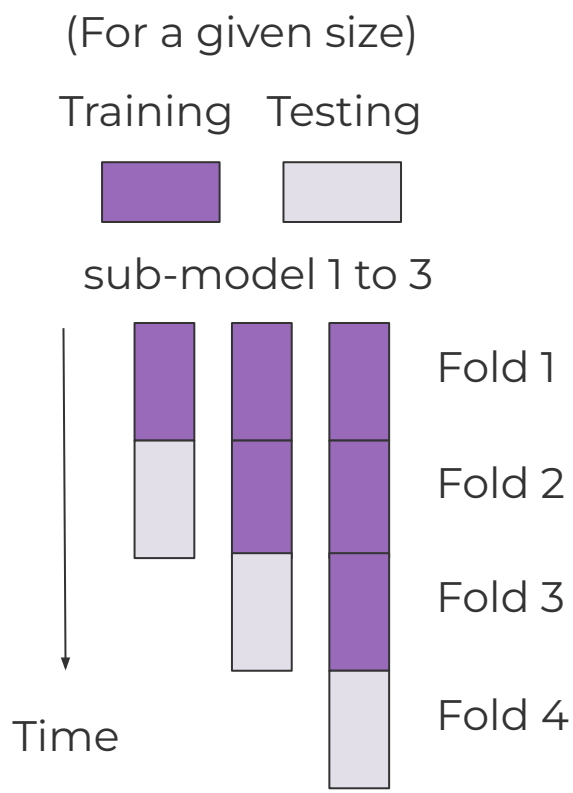

Figure 2: Walk-forward

Evaluation

To avoid overfitting, we first partially fit the model on a subset of data and test it on a separate holdout set that respects the temporal order:

- Training: all financial statements from 2007-12-31 to 2011-12-31
- Testing: all financial statements from 2011-12-31 to 2012-12-31
- Note: the final deployed model is trained on all the data

Then, we used AUC calculated respectively for Overall, Small, Medium and Large companies to compare the performance of the baseline logit model and the proposed LightGBM ensembles.

Lastly, SHAP is used to visually inspect whether what the model has learnt is meaningful. In particular, what are the key features to predict defaults.

Evaluation (cont.):

The proposed model demonstrates superior performance over logit model both overall and across company sizes!

Note: even though the improvement is smallest for Small companies, this category has the largest number of default cases, so the potential savings are still significant.

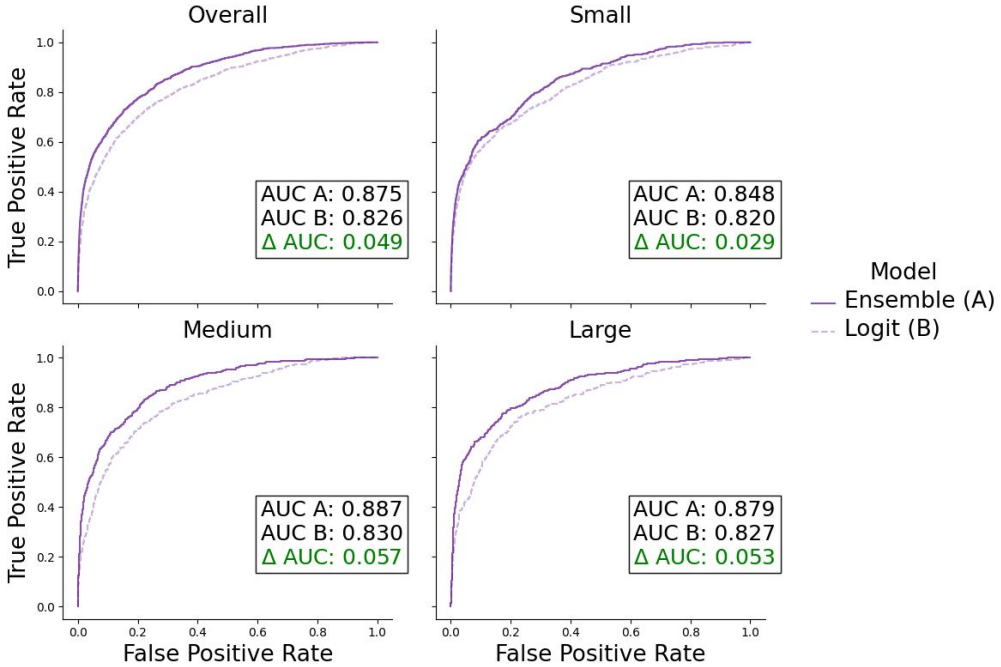

Figure 3: ROC curves of Ensemble model vs Logit model

Evaluation (cont.):

For different company sizes, the model is able to discover (different) most contributing factors that causes the company to default, and it makes economic sense.

For example, although we see higher ‘cashflow_to_debt’ (the red portion) significantly reduces the PD for Small companies, for Medium and Large companies, higher ‘profit’ or ‘equity_ratio’ have a stronger impact on them being less likely to default!

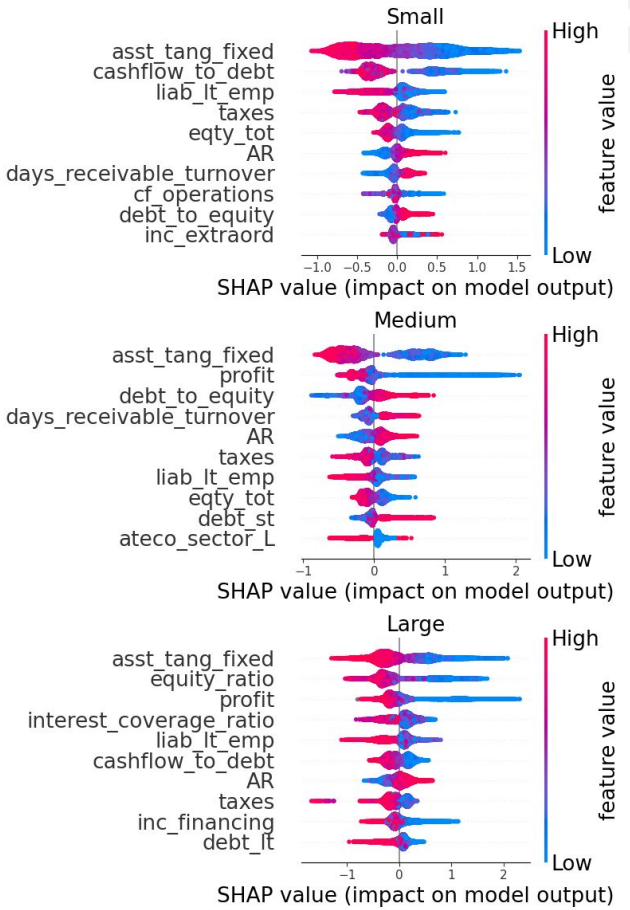

Figure 4: Beeswarm plots of grouped models

NYU

Evaluation (cont.):

What about a single new financial statement?

We know exactly how the prediction (in log odds) is made! We also know how perturbations of a single feature can affect the outcome, and in particular, by how much.

Note: samples can have very different feature weights!

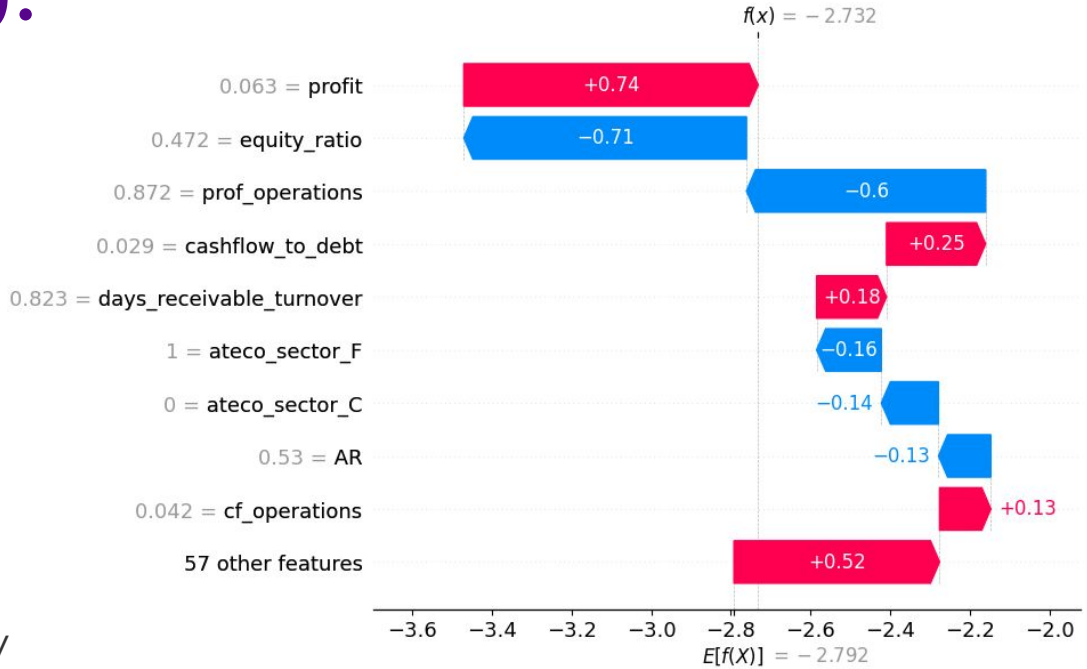

Figure 5: Waterfall plot of an individual prediction

Deployment:

- When a prospective borrower applies for a loan, the borrower will submit their financial data (assumed to be in the form of financial statements).
  - The loan officer will then input the financial data into our model and our model will output a PD. The loan officer must then determine the interest rates and underwriting fee based on the PD.
  - Since Banca Massiccia is subject to the EU GDPR, there may be legal obligations for ‘right to explanation’. If requested, the loan officer may generate a SHAP plot to explain the PD from our model.
- Updating our model:
  - The model should be re-trained quarterly with new data. This allows the model to remain up-to-date with market dynamics, and aligns with varying fiscal years across firms.
  - Additionally, model re-training is efficient and relatively simple, taking approximately 15 minutes to train on a 1M row dataset.

Appendix:

- Supplementary graphs to show the default rate of different sizes of companies across the years.
  - The default rate increases across years.
  - Smaller companies have a higher default rate.

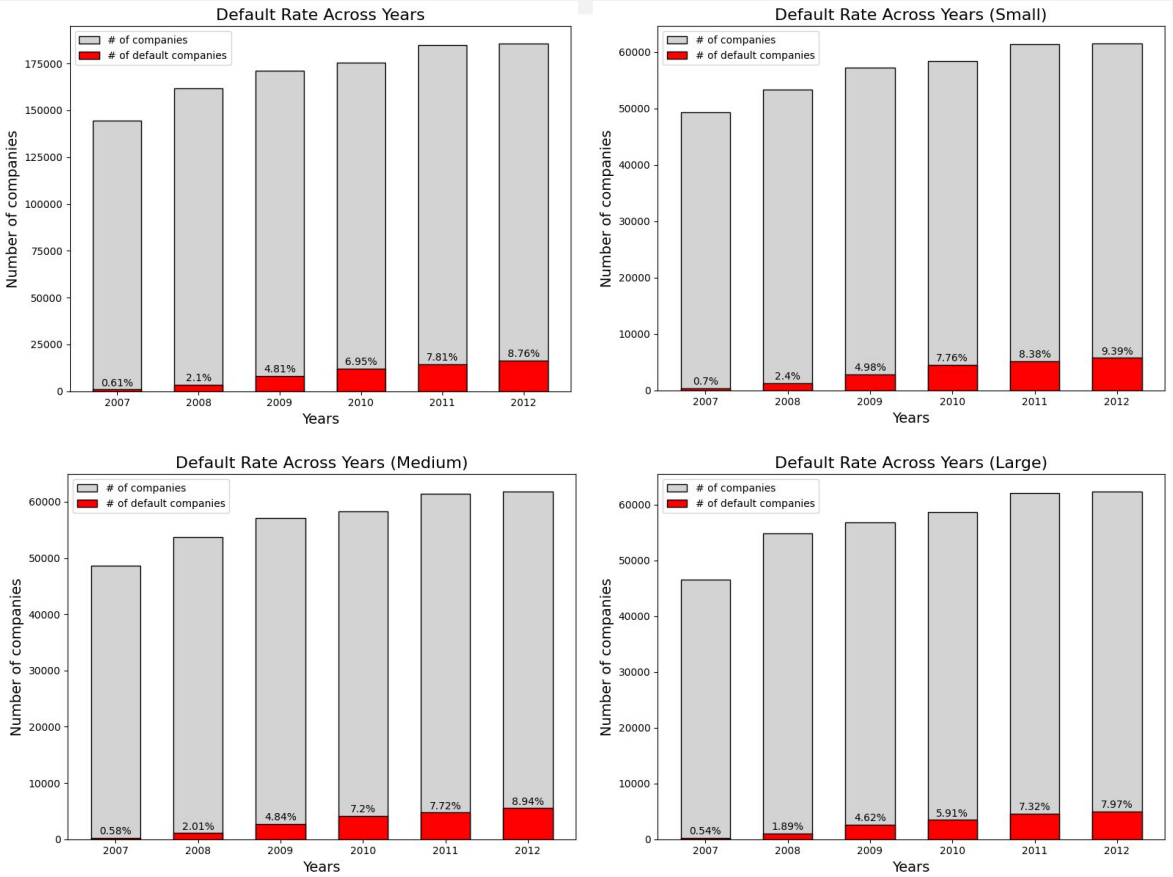

Figure 6: Default Rate Across Years for Different Size Firms

Appendix (cont.):

- Contributions:
    - Sunny Yang: Problem formulation, literature review, exploratory analysis, data preparation, baseline modeling, write-up of pitch deck.
    - Xinchen Zhang: Problem formulation, literature review, exploratory analysis, data preparation, baseline modeling and ensemble modeling, evaluation, visualization, write-up of pitch deck.
    - Alon Florentin: Problem formulation, literature review, exploratory analysis, data preparation, write-up of pitch deck.
    - Jaylen Peng: Problem formulation, literature review, exploratory analysis, write-up of pitch deck.

References:

1. Horrigan, James O. “A Short History of Financial Ratio Analysis.” The Accounting Review, vol. 43, no. 2, 1968, pp. 284–94. JSTOR, http://www.jstor.org/stable/243765. Accessed Nov. 2024.
2. Kocagil, Ahmet Enis et al. “Quantitative Research Special Report Fitch Equity Implied Rating and Probability of Default Model.” (2007).
3. Lopez, Jose. (2002). The empirical relationship between average asset correlation, firm probability of default and asset size. Journal of Financial Intermediation. 13. 265-283. 10.1016/S1042-9573(03)00045-7.
4. Pederzoli, Chiara & Torricelli, Costanza. (2010). A parsimonious default prediction model for Italian SMEs. Universita di Modena e Reggio Emilia, Facoltà di Economia "Marco Biagi", Centro Studi di Banca e Finanza (CEFIN) (Center for Studies in Banking and Finance). 5.
5. Coletto, Diego. “Effects of Economic Crisis on Italian Economy.” Effects of Economic Crisis on Italian Economy | European Foundation for the Improvement of Living and Working Conditions, Eurofound.eu. Accessed 5 Nov. 2024.
6. Erber, Georg. “Italy’s fiscal crisis.” Intereconomics, vol. 46, no. 6, Dec. 2011, pp. 332–339, https://doi.org/10.1007/s10272-011-0397-0.