Banca Massiccia Default Prediction

Group Periwinkle: Asawari Badkundri (ab10445), Keegan Kelly (kmk9461), Samruddhi Lahoti (sl10152), Jay Rana (jpr8961)

NYU

Business Understanding

- Banca Massiccia, a major Italian bank, aims to enhance their loan underwriting procedures by leveraging ML techniques to predict the likelihood of prospective borrowers defaulting on their loans over a 1 year horizon
- **Loan Underwriting** is the process of evaluating and assessing the creditworthiness of a borrower before approving a loan application.
- Modeling this process is important as failing to capitalize on important client data such as leverage and profitability could lead to missed opportunities for the bank or the inadvertent approval of loans that are likely to default
- Leveraging the extensive historical data at Banca Massiccia's disposal, we have employed advanced analytics to uncover patterns and correlations that influence the 1-year probability of default
- Data driven approaches have been utilized since the 1960's (Altman 1968; Dwyer, Kocagil and Stein 2004) and we aim to draw from that wealth of finance and econometric theory

Problem formulation

- Banca Massiccia has used several methods, including statistical models of default and traditional credit scoring models to optimize loan underwriting
- The **infusion of financial intuition and theory** into our modeling is evident across various stages, including the selection of variables, exploration of feature transformations, multicollinearity evaluation, and the definition of the target variable.
- This approach is expected to be more successful than data mining without a finance context due to its **feature relevance, interpretability, alignment with business goals, and consideration of financial dynamics specific** to the domain of loan underwriting.
- Given that we are predicting the 1 year probability of default (henceforth, PD), our research suggests that we should incorporate measures of factors such as **liquidity, profitability, efficiency, leverage and size** in our model

Problem formulation

- The next step is to identify the records that indicate a default based on the **default date (def_date)** and **statement date (stmt_date)**
- Using these dates and adding a lag of 6 months to take into account the **difference between the date as-of which the statement is generated and when it is actually available for analysis**, we have defined our target variable as follows:
  - **1**, if the associated record has a non-null def_date that falls between 6 to 18 months of the stmt_date
  - **0**, if the def_date is null or the lag between the default date and the statement date is negative or more than 18 months
- Though we did not see any examples of firms coming back from default in our training dataset, we believe this definition will still capture such cases correctly.
- Our choice of 6 months to represent the appropriate lag was determined through research on the approval and filing process in Italy provided by the Italian Business Registrar.

NYU

Data Understanding

- The training dataset has 1,023,552 observations (with 237,711 unique firms), from year 2007 to 2012, where each record represents the financial snapshot of a firm for that year
- Since defaults are "uncommon but not rare", there are far less records in the dataset that indicate default than those that don't. Based on our default logic, **the sample default rate is 1.274%**
- One **limitation** of the dataset is that it contains **sampling bias**, since we only have data on firms that have been granted credit and lack information on potential clients who were not, which is characteristic of this type of finance problem.
- Some features of the dataset also had missing values. Since all features weren't relevant to our analysis, our next few slides describe how we filled in missing values for features of interest.

NYU

Data Preparation: Pre-processing

- We converted some features to the type required for analysis (for ex. datetime, categorical etc.)
- **Return on Assets (roa) and operating revenue (rev_operating)** were 2 features of interest that had missing data we were able to recover:
  - Missing ROA values were filled in using prof_operations / asst_tot * 100
  - Missing rev_operating were populated with prof_operations + COGS
- Calculated financial ratios for clusters of **profitability, liquidity, leverage and efficiency** (details on the next slides) and normalized total assets to account for the **size feature**
- Missing or incorrect values (NaN, ±∞), after calculating the ratios, were dropped which resulted in:
  - **3.722% records from the original dataset being dropped** (including dropped financials that occured in the 6 month window previously discussed)
  - This accounts for a **4.756% reduction** of default observations
- **ATECO sectors** (total 83) were grouped into 36 distinct categories based on the designations on page 47 of the Classificazione delle attività economiche Ateco 2007 to net out potential noise

NYU

Data Preparation: Defining Financial Ratios

Profitability: reflects a firm's ability to earn profits from sales, operations or assets

- gross_profit_margin_on_sales = prof_operations/rev_operating
- Net_profit_margin_on_sales = profit/rev_operating
- Cash_return_on_assets = cf_operations/asst_tot
- ROE = profit/eqty_tot
- ROA = prof_operations/asst_tot

Leverage: assesses whether a firm should be able to meet its debt obligations

- debt_assets_lev = asst_tot - eqty_tot/asst_tot
- debt_equity_lev = asst_tot - eqty_tot/eqty_tot
- financial_leverage = roe - roa

Liquidity: measures how quickly a firm's assets can be converted to cash

- current_ratio = asst_current/debt_st
- cash_ratio = cash_and_equiv/debt_st
- defensive_interval = (cash_and_equiv + AR) * 365 / (COGS + rev_operating - prof_operations)
- Wc_net = asst_current - debt_st

Efficiency: gives insight into how well a firm uses its resources

- receivable_turnover = rev_operating/AR
- Average_collection_receivables_day = 365/receivable_turnover
- asset_turnover = rev_operating/asst_tot
- Working_capital_turnover = rev_operating/wc_net

NYU

Feature Selection

Univariate AUC Feature Selection

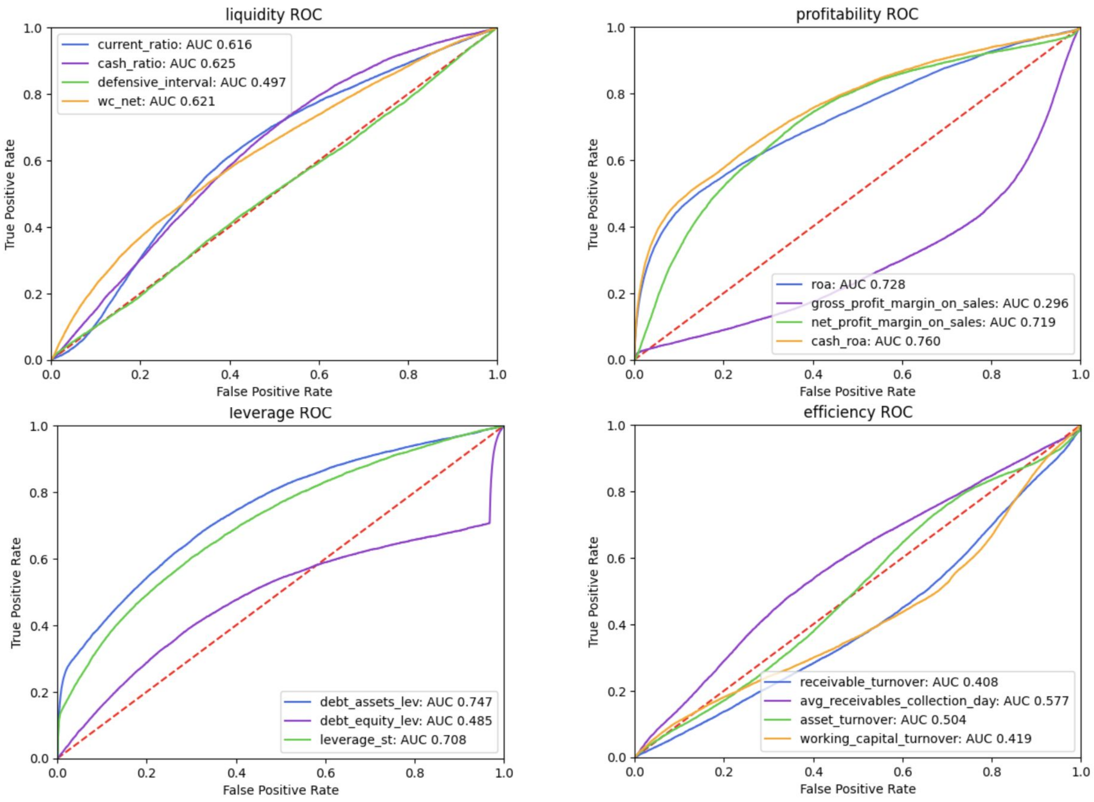
- Using our derived financial ratios, we selected one variable per cluster based off of an assessment of univariate predictive power using logistic regression.
- Along with the size and categorical features, the resulting feature set is: [cash ratio, cash return on assets, debt assets leverage, average collection receivables day, size, legal structure, ATECO sector, HQ city]

Recursive Feature Elimination

- For each feature, we fit a **decision tree classifier** with max depth of 4.
- RFE method of sklearn.feature_selection module returns the n most relevant features. We selected only best feature from each cluster so n=1.
- Along with the size and categorical features, the resulting feature set is: [current ratio, cash return on assets, debt assets leverage, asset turnover, size, legal structure, ATECO sector, HQ city]

Using **VIF**, we tested our non-categorical variables for noticeable signs of multicollinearity and detected minimal effect, with VIF under 2 for all features

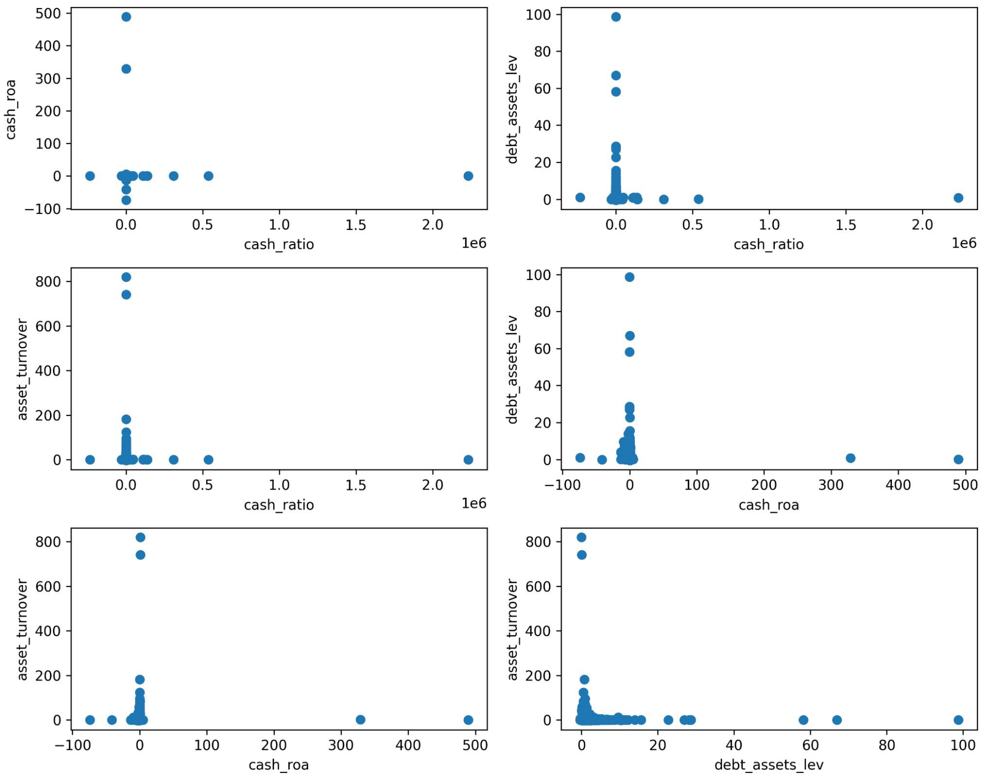
Feature Selection: Univariate AUC Analysis



NYU

Multivariate Outlier Analysis

- We performed outlier analysis on a selected subset of our features using techniques such as Cook's Distance, pairwise scatter plots etc.
- Although some data points do lie outside the concentrated range, excluding them from our data didn't change the performance considerably, so we chose to keep these values.



Modeling: Approach

- To model the one-year PD, our economic reasoning supports that as leverage rises and profitability, liquidity or efficiency fall, the PD will increase. Further, smaller size firms are associated with increased PD.
- Using the feature set chosen using univariate AUC, the statsmodel.formula.api implementation of **multivariate logistic regression** achieved a AUC of 0.774. This formed the baseline for our analysis, as any enhancements we would consider for modeling should ideally lead to better performance.
- We experimented with a **neural network** approach, with 2 hidden layers and 10241 trainable parameters and focal loss for loss function, which failed to outperform the baseline model, achieving an AUC of 0.726.
- Next, we implemented two **XGBoost** models, one each for the feature set obtained using AUC and RFE analysis.
- **Our final model is an ensemble of these two XGBoost models** such that the final, predicted PD is an average of their individual predictions.

Modeling: Why XGBoost?

- The **economic intuition** behind using XGBoost for probability of default prediction lies in its ability to address challenges presented by credit risk analysis such as missing features and large, imbalanced datasets.
- XGBoost provides a feature importance analysis, which can be crucial in credit risk modeling. This **interpretability** can be important for economic decision-making and regulatory compliance.
- Using **cost sensitive learning**, XGBoost performs well on unbalanced data ensuring that the model is not biased towards the majority class (non-defaulters in this scenario).
- The hold-out data could have missing features and XGBoost has a built-in mechanism for **managing missing values**.

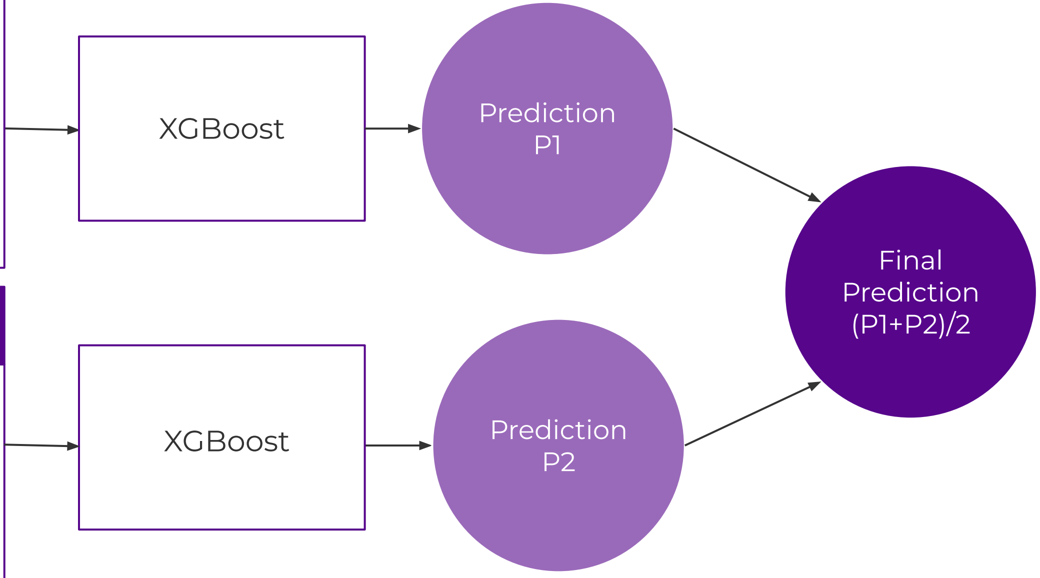
Modeling: Overall Flow

Univariate AUC Feature Set

- cash ratio
- cash return on assets
- debt assets leverage
- average collection receivables day
- size (normalized assets)
- legal structure, ATECO sector, HQ city

RFE Feature Set

- current ratio
- cash return on assets
- debt assets leverage
- asset_turnover
- size (normalized assets)
- legal structure, ATECO sector, HQ city



NYU

Modeling: XGBoost – Implementation Details

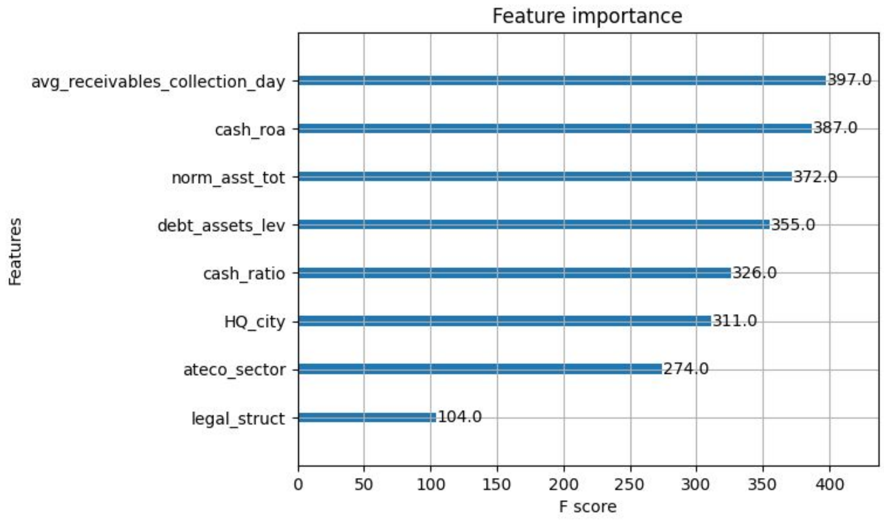
The XGBoost models are implemented as follows:-

- **Objective Function:** Binary Logistic Regression (since it is a binary classification problem)
- **Evaluation Metric:** Logarithmic loss (aka Binary Cross-Entropy)
- **Training Process:**
  - Features and target variable are used to create a DMatrix for both training and testing datasets.
  - XGBoost model is trained using the specified parameters and 100 boosting rounds.
- **Prediction:**
  - Model predicts the probability of the positive class for a given test dataset

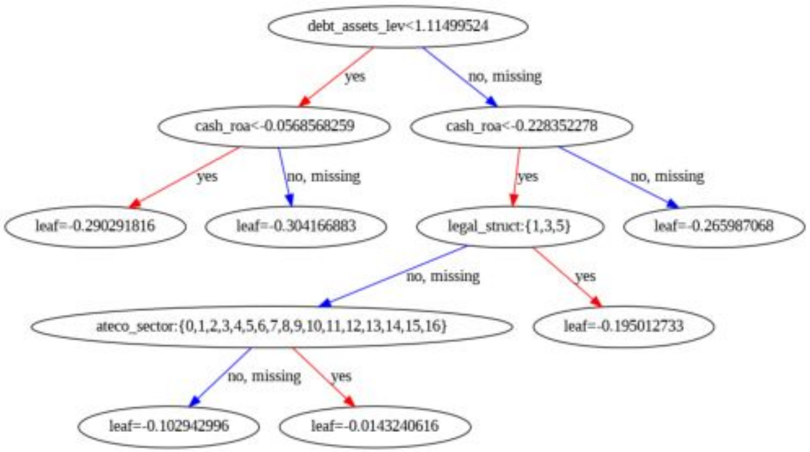
The final prediction is the average of the prediction from the two models

Limitation: Since our dataset is unbalanced, we tried to handle it using the scale_pos_weight parameter but with limited success.

Modeling: Model Specification



XGBoost Feature Importance

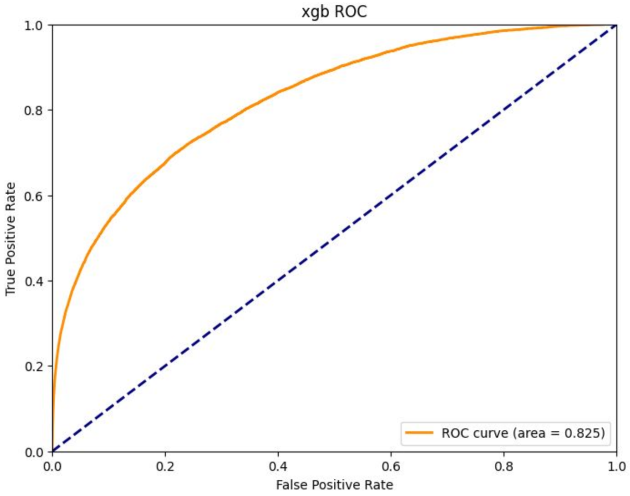


XGBoost Tree
(at tree_index=7)

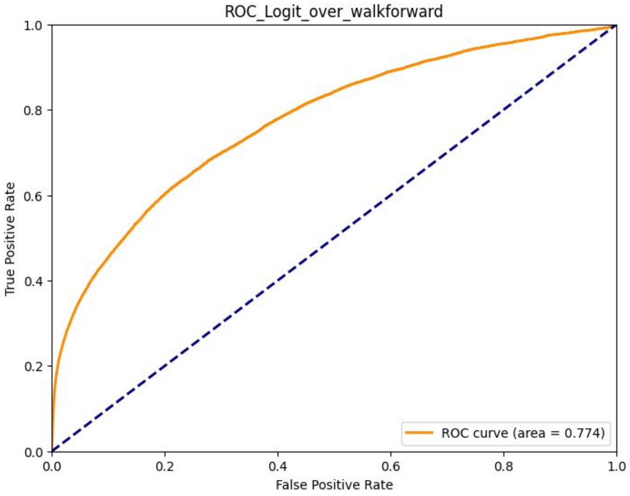
Evaluation

- We implemented a **walk-forward approach**, as discussed in our readings, to evaluate our model at each step as this method has the advantage of utilizing as much training data as possible to build the model while still generating **meaningful validation statistics**
- Our unit of analysis was the firm year and data up to **each year from 2008 to 2012 accounted for one step** of the walk-forward analysis
- At each step the, the model was tested on the records from the next firm year since we are modelling 1-year PD
- This helped us achieve **out-of-time** (testing on data in the next year, not included in training) and **out-of-sample** (potentially new companies in the records for next year) evaluation results.
- Predictions from all the walk-forward steps (4 in total) were concatenated to plot an **ROC curve that showcased how the model performed over all the steps**

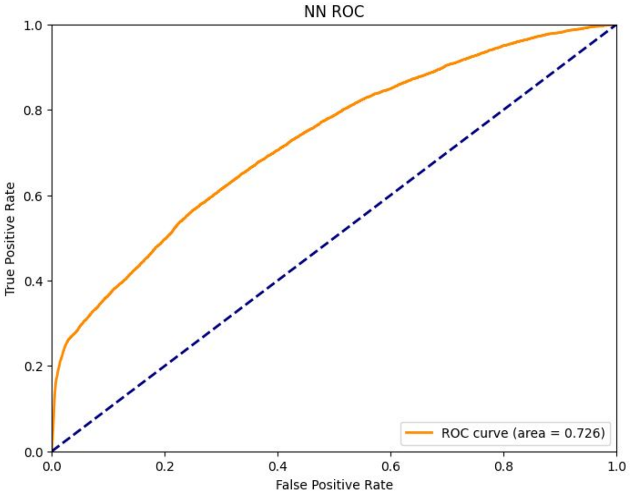
Evaluation: ROC Plots



XGBoost
(Univariate AUC
Features)



Logit
(Univariate Features)

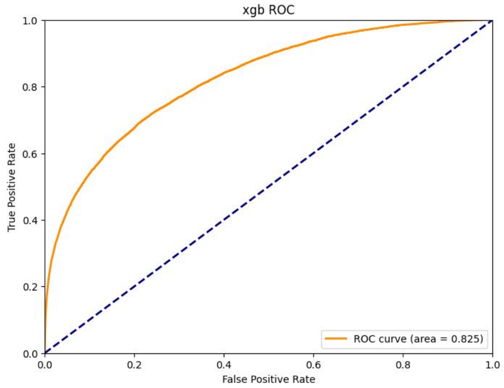


Neural Net
(Univariate AUC
Features)

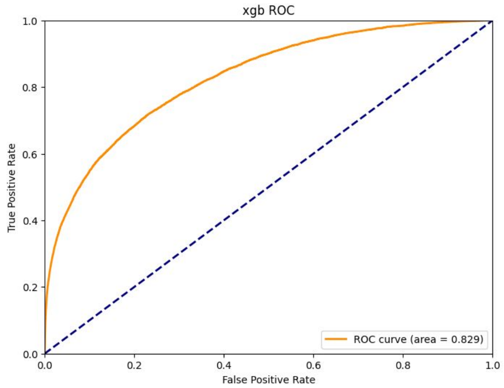
NYU

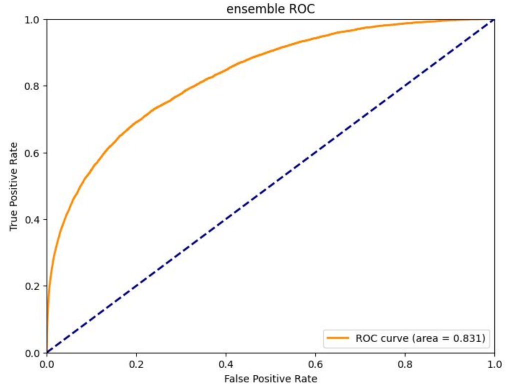
Evaluation: Ensemble Model

XGBoost
(Univariate AUC
Features)



XGBoost
(RFE Features)





Ensemble model
(average prediction)

NYU

Calibration

Step 1: Non-parametric mapping of model output to empirical probabilities

- Utilized LOWESS smoothing which fits a smooth curve through ordered data using weighted linear regression on a fraction of nearby data points to minimize noise and non-linearity
- The delta parameter can be used to decrease the computation time of the process; delta indicates how many estimations will be skipped and linearly interpolated
- Chose delta = 0.01 * length(model output) in accordance with statsmodel documentation suggestions and visualization of the fit

Step 2: Adjusting the smoothed empirical probabilities to real world probabilities by taking into consideration the difference between the sample and the population default rate

- Here, we used the Elkan (2001) approach as shown below where p_i^* is the adjusted probability, p_i is our model output and π_S and π_T are our sample and true probabilities of default:

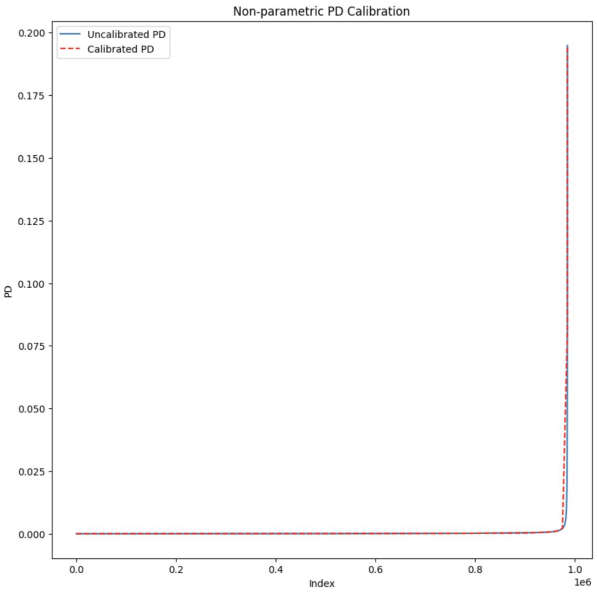
$$p_i^* = \pi_T \frac{p_i - p_i \pi_S}{\pi_S - p_i \pi_S + p_i \pi_T - \pi_S \pi_T}$$

- Using our train dataset, we calculated the sample probability (π_S: 1.27%)
- We derived the true probability by taking an average of the probabilities of default in Italy from 2007 to 2012 as reported by the International Monetary Fund (π_T: 9.52%).

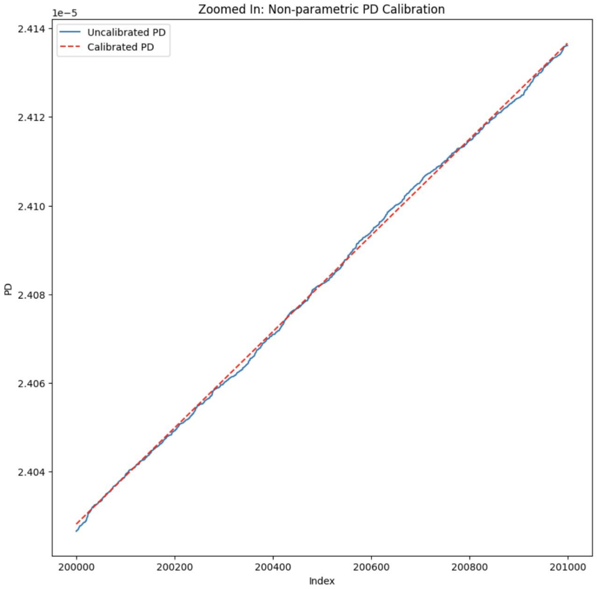
NYU

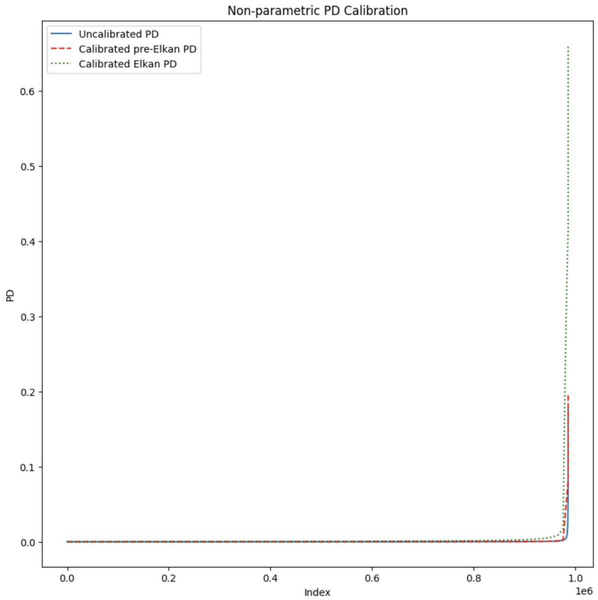
Calibration

- Non-parametric smoothing is more noticeable at the granular level
- PD generally shifted upward in the second step of calibration



Step 1 Calibration

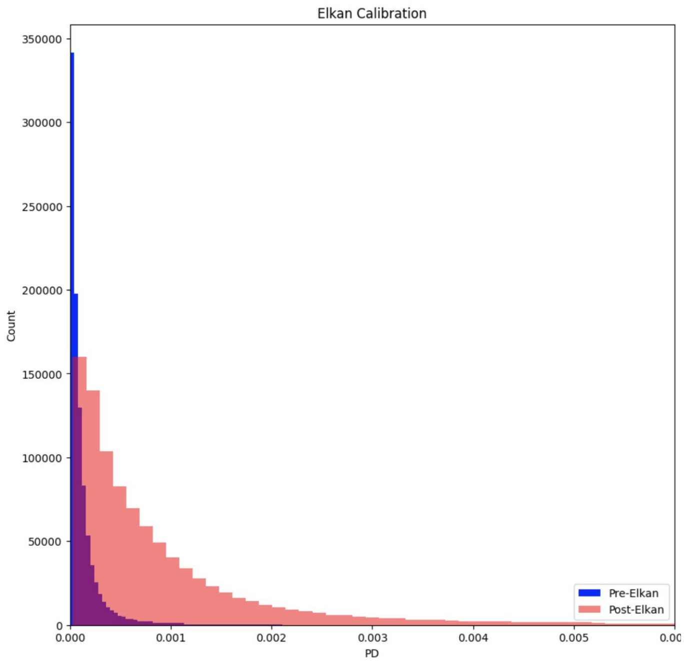




Step 2 Calibration

NYU

Calibration



Data on Italian Default from IMF:

Year	2007	2008	2009	2010	2011	2012
Default	5.78%	6.28%	9.45%	10.03%	11.74%	13.75%

Likelihood of our model output:

Data	Log Likelihood
Uncalibrated	-91,932
Calibrated Step 1	-86,502
Calibrated Step 2	-64,065

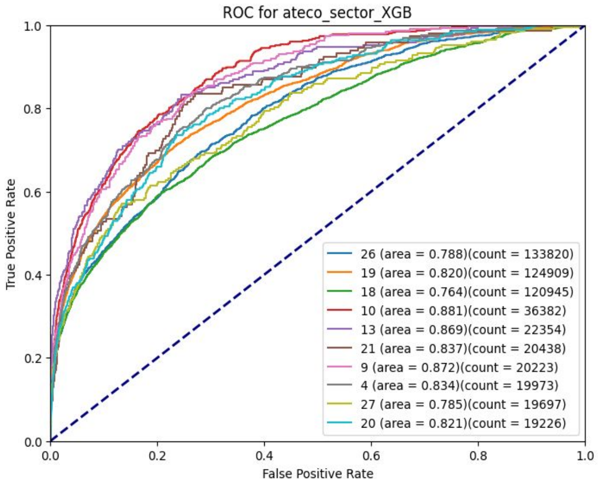
NYU

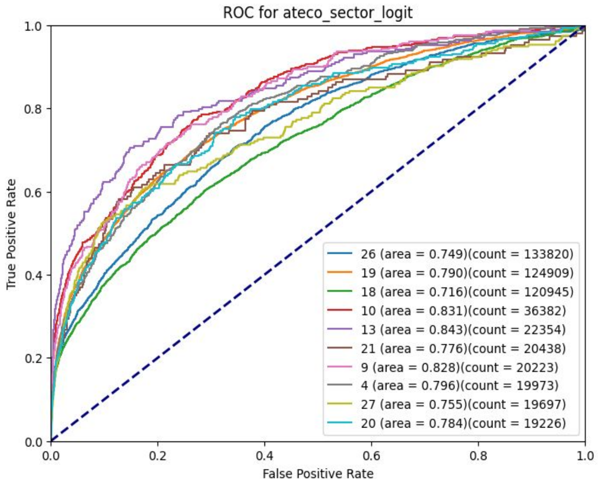
Deployment

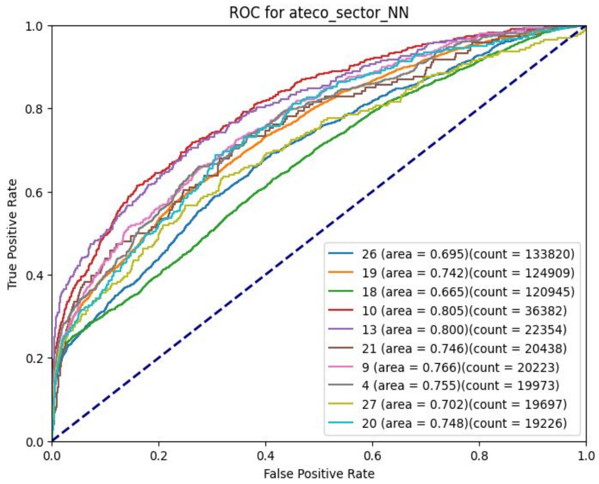
- Like many financial data mining problems, **our PD model is just one piece of a larger ecosystem**. Our model will return a 1 year PD for every financial supplied to it. Using these predictions, the **bank will employ a pricing strategy in order to limit bad loans.**
- HQ city is one of our independent variables. From a legal standpoint, Banca Massiccia should consider whether there are any laws restricting what kinds of loans can be made in those regions.
- On the **ethical side**, while our model is expected to discriminate between good and bad loans, it is possible that restrictions based on this output might lead to unfair denial of loans that are tied to attributes like firm size or legal structure. This could put newer/smaller firms at a disadvantage.
- The bank's implementation of a pricing strategy rather than optimal cutoffs **should mitigate legal/ethical ramifications**. Under this strategy, specific segments of the market that might otherwise be cutoff could be granted loans. A pricing strategy often leads to greater payoff for banks as well.
- There is still risk involved in utilizing the PD model. **The quality of the model output is dependent on the quality of the input.** While we have steps to clean data, missing and erroneous data could lead to unexpected results. If a firm's financials are missing key inputs, our model may not be able to discriminate well. This highlights the importance of continuous monitoring of model performance.

Deployment

- Our model doesn't do well for certain ateco sectors which might require alternative strategies to oversee credit approval.
- In evaluating the model, we noticed that certain (grouped) ATECO sectors had more predictive influence than others.
- This varied across the implemented models, with XGBoost still giving the best AUC.







Appendix: Work Distribution

Team Member	Worked on...
Asawari	Business Understanding, Problem Formulation, Pre-processing, Feature Selection, Models - Logit, Calibration
Keegan	Business Understanding, Problem Formulation, Financial Ratios, Feature Selection, Models - Logit, Calibration
Samruddhi	Business Understanding, Problem Formulation, Pre-processing, Multivariate Outlier Analysis, Models - Neural Net, XGBoost, Deployment
Jay	Business Understanding, Problem Formulation, Feature Selection, Models - LR, NN, Ensemble model, Validation, Evaluation, Deployment

Bibliography

- Active Credit Portfolio Management in Practice (Excerpts), by Stein, Roger M.; Bohn R., Jeffrey. Case No. WIL-9780470080184_ex3. 2009: John Wiley & Sons, Inc. 2009
  - Ch 4: Econometric Models
  - Ch 7: PD Model Validation
- Altman, E. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 22, 589-610
- Course cheat-sheets
- D.W. Dwyer, A.E. Kocagil, R.M. Stein, The Moody's KMV EDF RISKCALC v3.1 model, Technical documentation, 2004.
- Nonperforming Loans to Total Gross Loans, Financial Soundness Indicators, International Monetary Fund.
- https://italianbusinessregister.it/en/annual-accounts
- https://www.linkedin.com/pulse/ai-ml-navis-tech-forward-approach-loan-underwriting-sharth-mandan/
- https://www.linkedin.com/pulse/revolutionize-credit-underwriting-through-machine-amol-k-bahuguna/