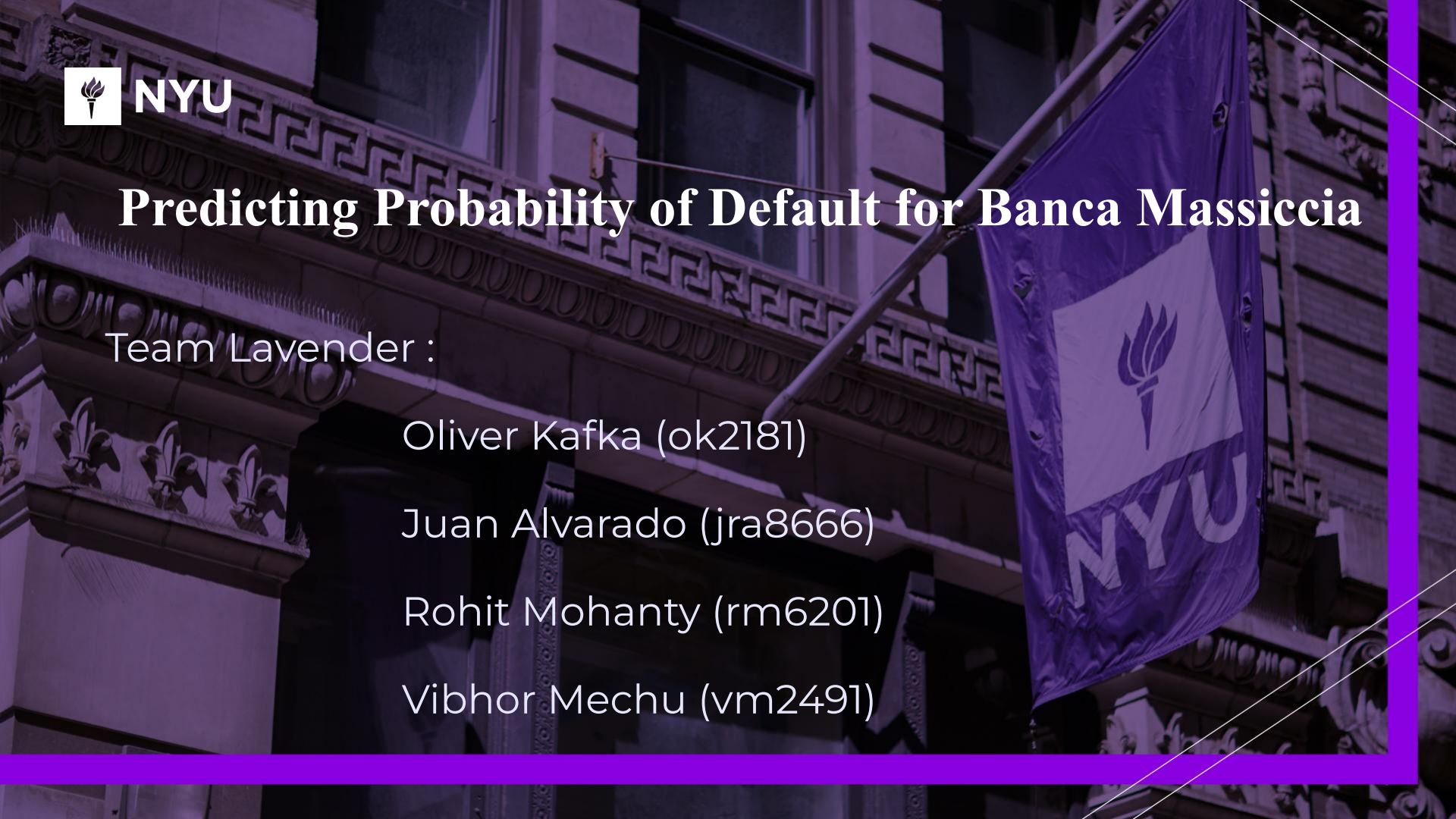


Predicting Probability of Default for Banca Massiccia

Team Lavender :

Oliver Kafka (ok2181)

Juan Alvarado (jra8666)

Rohit Mohanty (rm6201)

Vibhor Mechu (vm2491)

Business Understanding

- **Who are we Helping?**
  - **Banca Massiccia** , a leading Italian bank, has approached us to improve their loan approval process and reduce financial risk.
- **What is the Problem?**

  The bank struggles with:
  - Identifying which companies are risky to loan to and which are less risky.
  - Making accurate decisions about who to loan money to and determining appropriate interest rates.
- **Why does it matter?**
  - **Reducing Default Rates:** Accurate default predictions minimize financial losses.
  - **Enhancing Profitability:** Better decisions lead to optimized lending strategies and decreased costs.
- **How will we help?**
  - We will provide an independent probability for each firm predicting whether they will default within the next year, based on their most recent financials at the time of loan application.

NYU

How Our Solution Addresses the Problem

- **Recovering Missing Data:**
  - We use financial theory and balance sheet rules to reconstruct missing values in key financial metrics, ensuring consistency and accuracy in the dataset.
- **Selecting Ratios:**
  - We use financial theory to select ratios which inform us of a firm's liquidity, leverage, profitability, and growth.
- **Training the Model:**
  - We use these ratios as inputs into our machine learning model, which predicts the probability of default (PD) for each firm within the next 12 months.
- **Calibrating Predictions:**
  - We calibrate the model's output probabilities to match the population default rate, ensuring the predictions align with real-world default trends and are actionable for decision-making.

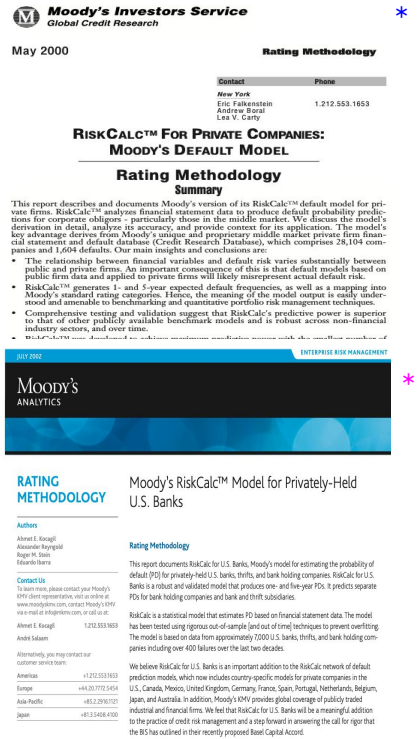
What Has Been Done Before & How We Are Different

Building on RiskCalc:

In the past, default risk was assessed using RiskCalc, which relies on financial ratios to estimate the probability of default. This approach provided a solid foundation based on financial theory.

Our Enhancements:

- **Added XGBoost:** We introduced a machine learning algorithm, XGBoost, to capture non-linear patterns and improve predictive performance.
- **Modeled New vs. Recurring Firms:** We separated new firms and recurring firms to account for their distinct financial characteristics, leading to more accurate predictions.
- **Used Custom Bins:** We used custom binning to gain finer resolution in the most critical parts of the distribution, where slopes are steepest.
- **Used Categorical Variables:** We make analogous PD transformations to categorical variables.



Moody's Investors Service
Global Credit Research

May 2000

Rating Methodology

Contact	Phone
New York	
Eric Falkenstein	1.212.553.1653
Andrew Boral	
Lea V. Carty	

RISKCALC™ FOR PRIVATE COMPANIES:
MOODY'S DEFAULT MODEL

Rating Methodology

Summary

This report describes and documents Moody's version of its RiskCalc™ default model for private firms. RiskCalc™ analyzes financial statement data to produce default probability predictions for corporate obligors – particularly those in the middle market. We discuss the model's derivation in detail, analyze its accuracy, and provide context for its application. The model's key advantage derives from Moody's unique and proprietary middle market private firm financial statement and default database (Credit Research Database), which comprises 28,104 companies and 1,604 defaults. Our main insights and conclusions are:

- The relationship between financial variables and default risk varies substantially between public and private firms. An important consequence of this is that default models based on public firm data and applied to private firms will likely misrepresent actual default risk.
- RiskCalc™ generates 1- and 5-year expected default frequencies, as well as a mapping into Moody's standard rating categories. Hence, the meaning of the model output is easily understood and amenable to benchmarking and quantitative portfolio risk management techniques.
- Comprehensive testing and validation suggest that RiskCalc's predictive power is superior to that of other publicly available benchmark models and is robust across non-financial industry sectors, and over time.

JULY 2002 ENTERPRISE RISK MANAGEMENT

Moody's ANALYTICS

RATING METHODOLOGY

Moody's RiskCalc™ Model for Privately-Held U.S. Banks

Authors

Ahmet E. Kocagil
Alexander Reyngold
Roger M. Stein
Eduardo Ibarra

Contact Us

To learn more, please contact your Moody's KMV client representative, visit us online at www.moodyskmv.com, contact Moody's KMV via e-mail at info@kmv.com, or call us at:

Ahmet E. Kocagil 1.212.553.1653
André Sabaam

Alternatively, you may contact our customer service team:

Americas	+1.212.553.1653
Europe	+44.20.7772.5454
Asia-Pacific	+852.2916.1121
Japan	+81.3.5408.4100

Rating Methodology

This report documents RiskCalc for U.S. Banks, Moody's model for estimating the probability of default (PD) for privately-held U.S. banks, thrifts, and bank holding companies. RiskCalc for U.S. Banks is a robust and validated model that produces one- and five-year PDs. It predicts separate PDs for bank holding companies and bank and thrift subsidiaries.

RiskCalc is a statistical model that estimates PD based on financial statement data. The model has been tested using rigorous out-of-sample [and out of time] techniques to prevent overfitting. The model is based on data from approximately 7,000 U.S. banks, thrifts, and bank holding companies including over 400 failures over the last two decades.

We believe RiskCalc for U.S. Banks is an important addition to the RiskCalc network of default prediction models, which now includes country-specific models for private companies in the U.S., Canada, Mexico, United Kingdom, Germany, France, Spain, Portugal, Netherlands, Belgium, Japan, and Australia. In addition, Moody's KMV provides global coverage of publicly traded industrial and financial firms. We feel that RiskCalc for U.S. Banks will be a meaningful addition to the practice of credit risk management and a step forward in answering the call for rigor that the BIS has outlined in their recently proposed Basel Capital Accord.

NYU

* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=236011 * http://www.rogermstein.com/wp-content/uploads/riskcalc-usbanks.pdf

Brief overview of dataset

Defaults within one year of loan application: **12,949**

Non Defaults within one year of loan application: **1,010,603**

Unique sectors: **83**

Data: **44 features + 1 Identifier**

Missing Values for features ——————>

Feature	Missing Values	% Missing
ateco_sector	0	0 %
def_date	0	0 %
fs_year	0	0 %
AR	12	0.0012 %
cash_and_equiv	17	0.0017 %
asst_tot	0	0 %
eqty_tot	1	0.0001 %
debt_st	5	0.0005 %
debt_lt	155	0.0151 %
rev_operating	174	0.0170 %
prof_operations	21	0.0020 %
exp_financing	367	0.035 %
profit	20	0.002 %
ebitda	81	0.0079 %

NYU

Problem formulation : Definition of Default

- **Definition of Default Label:**
  - We defined default as a default event occurring within **one year** of a firm's adjusted financial statement date, with a **150-day** buffer to account for processing delays.

- **Why This Definition ?**
  - **Industry Context:** In Italy, most firms generate their financial statements by **April*** and approve them by **July***, meaning our model must rely on the financial data available during this timeline.
  - **Relevance:** This definition captures default risk within the fiscal period, aligning with the bank's operational timelines and decision-making needs

NYU

*https://italianbusinessregister.it/en/annual-accounts

Financial Theory: Data Reconstruction

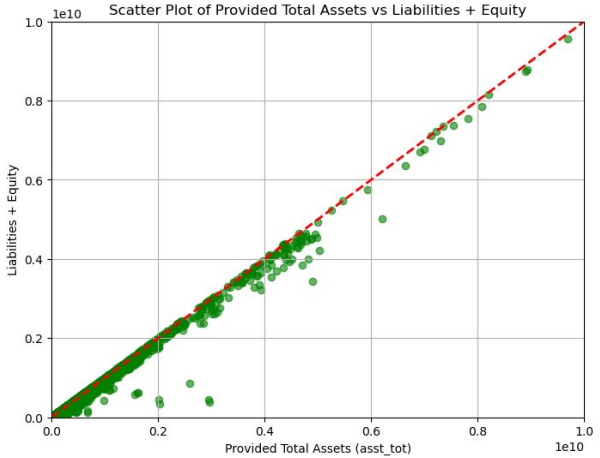
By using Balance Sheet Formulas we are able to reconstruct missing fields:
Ex: ROE

- ROE has 72937 null values in the training data (7.12% of data is missing)
- ROE (Return on Equity) is simply: profit / total equity
- By applying this formula, we are able to limit null values to only 0.2%

Ex: Finding Total Liabilities for use in features

- Balance Sheet States Assets = Equity + Liabilities
- Thus we can solve for liabilities, to find the best combination of fields to represent total liabilities
- We found that debt_st + debt_lt is a better approximation than debt_st + debt_lt + AP_st + AP_lt in our data

$$ROE = \frac{\text{Net Income (annual)}}{\text{Shareholders' Equity}}$$



NYU

Financial Theory: Ratios

We focused on critical ratios in **liquidity, profitability, debt coverage,** and **Growth**, based on their established use in assessing financial stability.

Leverage / Debt Coverage

- **Financial Leverage:** Measures to what extent debt is used to finance a firm's assets.
  Total liabilities / Total Assets
- **Debt Service Coverage Ratio (DSCR):** Assesses the firm's capacity to cover debt obligations using EBITDA.

Profitability

- **Profitability Ratio:** Ratio of net income to total assets, showing the firm's efficiency in generating profit.
- **Return on Equity (ROE):** Indicates profitability relative to shareholder equity.

Asset Management / Growth

- **Net Income Growth:** Tracks changes in net income over time, highlighting financial performance trends and risk appetite.
- **Sales Growth:** Measures growth in operating revenue, reflecting market share expansion and risk.

Liquidity

- **Quick Ratio (v2):** Measures the firm's ability to meet short-term obligations using liquid assets.
- **CFO Ratio (v2):** CFO measures how much cash flow from operations is available to cover short-term liabilities.

Growth Features

How we make them:

- Grouped data by **ID** and sorted by statement date.
- Computed year-over-year percentage change for metrics like **net_income** and **sales**.
- Filled missing values for first occurrences with 0% growth to avoid data gaps.
- Integrated historical data for recurring firms within harness to ensure consistent growth calculations.
- Applied quantile binning to handle outliers and rank firms relative to their peers.

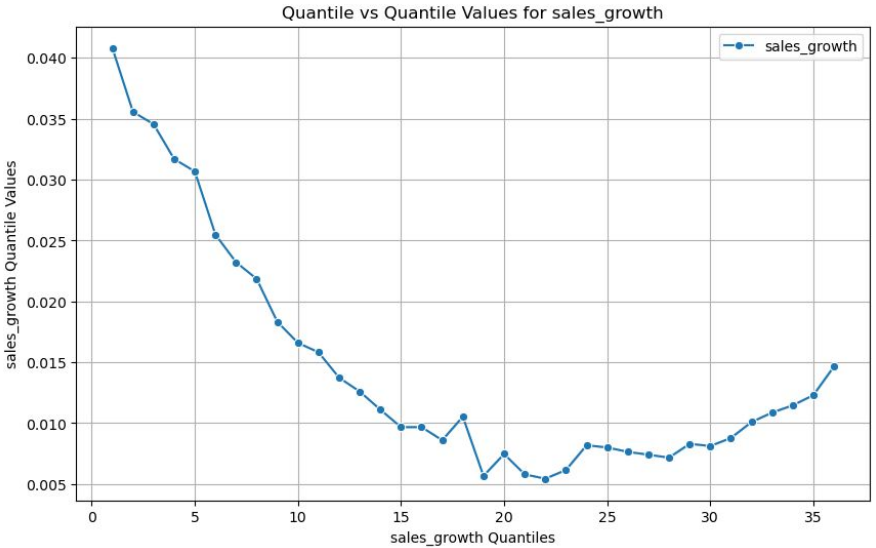
What they add:

- **Net Income Growth:** Tracks operational stability and financial trends over time.
- **Sales Growth:** Highlights revenue performance and market expansion.
- **Risk Insights:** Lower quantiles indicate stagnating or declining firms, flagged as higher risk.
- **Model Impact:** Quantile transformations reduce outlier impact, improving feature interpretability and model robustness.

Conditional Probability of Default

Idea comes from Risk Calc:

- Calculate univariate probability of default given our financial ratio features
- Use this conditional probability of default as and input into our model
- Helps us handle non-linear relationships between our chosen financial ratios and probability of default

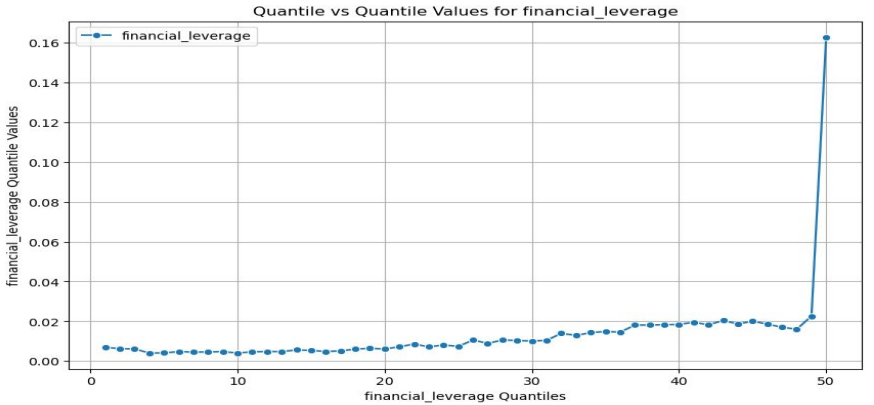


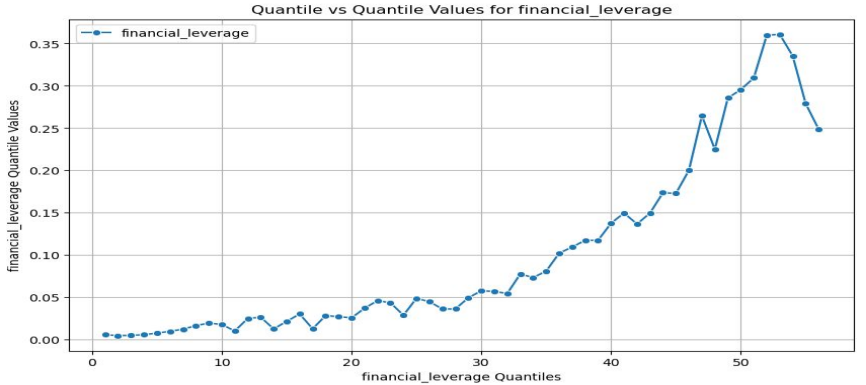
NYU

Conditional PD Bins

Custom Bins:

- We use custom binning to gain more resolution at more critical portions of the distribution
  a. (where slopes are most severe)
- This helps us from losing information that we would were we to use evenly spaced bins
- Custom bins were particularly effective in metrics like financial_leverage, where sharp transitions occur in higher quantiles, as shown here



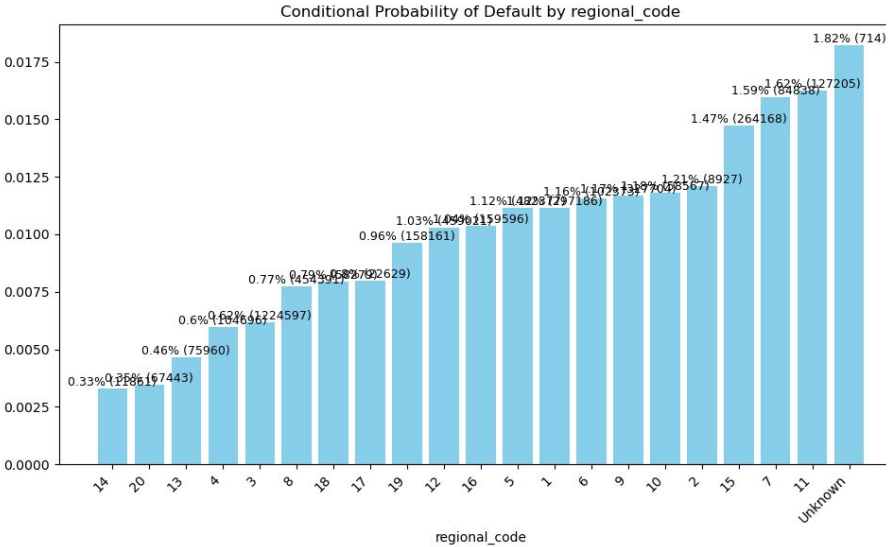


NYU

Conditional PD: Discrete Variables

Conditional Probability of Default:

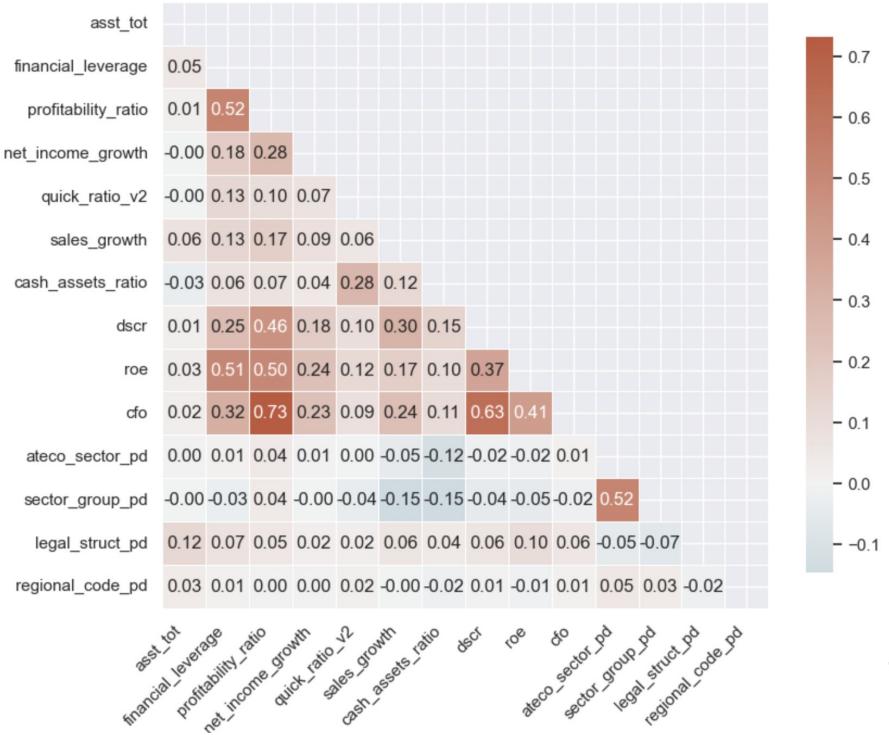
- We extend the use of conditional probability of default from continuous variables to categorical.
- We did so for 'ateco_sector', creating another field called: 'sector_group' which groups the more granular sectors into more broad categories.
- We creates conditional probabilities of default for 'ateco_sector', 'sector_group', 'regional_code', and 'legal_struct'.
- These fields are highly correlated, so we ultimately only used regional code in our model.
- Mapped HQ_city to broader regional_code groups using predefined regional mappings."
- Unmapped provinces were labeled as Unknown to handle missing or inconsistent data.
- Regional codes simplify analysis, reduce noise, and improve geographic risk interpretability.



https://www.istat.it/en/classification/codes-of-italian-municipalities-provinces-and-regions/

Feature Selection - Corr Matrix

- After using financial theory to decide which features we would develop for our model we looked at the correlation between features (multicollinearity) and the Variance Inflation Factor for our feature selection.
- From the correlation matrix we made the decision to keep all variables.
- Even though cfo had high correlation (0.73) with profitabilty_ratio and dscr we decided to keep it after experimenting and seeing performance decreases when removing it.



NYU

Feature Selection - Variance Inflation Factor

We used variance inflation factor (VIF) (which is a measure derived from a regression on variable X from the other variables) as another way of measuring multicollinearity between each individual feature and the rest.

The highest correlated feature was asst_tot which makes sense since its components or the feature itself is used to derive the other variables.

From the categorical (sector_group, regional, legal_struct) variables we kept regional_code since in experimentation it proved to give the highest increase in performance.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

Features	VIF (Pre-removal)	VIF (Post-removal)
asst_tot	51.26	*removed*
sector_group_pd	37.02	*removed*
regional_code_pd	28.82	7.91
legal_struct_pd	24.11	*removed*
cash_assets_ratio	17.46	*removed*
ateco_sector_pd	12.73	*removed*
quick_ratio_v2	7.72	6.38
cfo	4.21	4.20
sales_growth	3.94	3.71
profitability_ratio	3.51	3.47
dscr	3.16	3.14
net_income_growth	2.40	2.37
roe	2.27	2.26
financial_leverage	1.98	1.98

NYU

Modeling - New and Recurring Firms

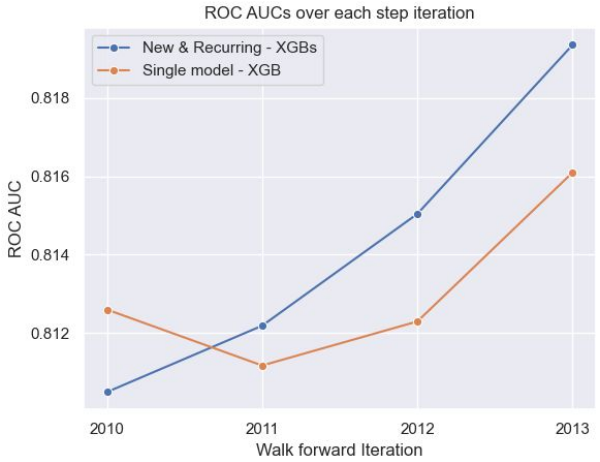
The bank may encounter two different types of firms that apply for a loan.

New firms: which are firms where the bank doesn't have any historical financial statements to user.

Recurring firms: firms where the bank has at least 1 historical financial statement from that same firm.

For recurring firms we develop "growth" features which allow us to capture how a firm has changes in sales and net income year over year.

We tested using 1 model without "growth" features for all firms versus 2 models one for new and one for recurring firms. The two split models perform best, and what we use in our final model.



NYU

Modeling

To develop our probability of default model we began by generating a naive baseline. The baseline assumed we only knew the Sector, Region and Legal structure of the firm and trained a LR with these variable. We compared against the following algorithm types:

GLMs:

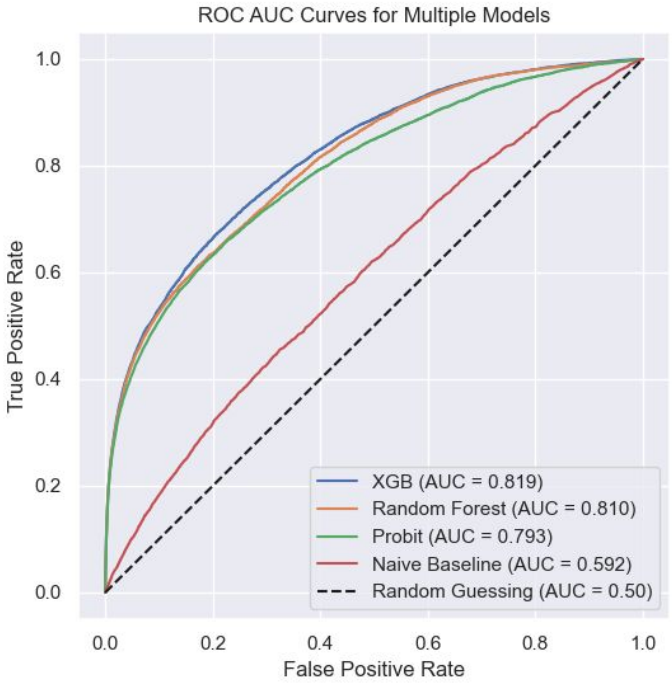
1. Logit
2. Probit

These models are more understandable, less complex and do not inherently capture non-linearity.

Tree Based Ensembles:

1. Random Forest
2. Boosting: LightGBM, XGBoost

Complex, robust to missing data and outliers and capture non-linearity. The tradeoff is that they can be less explainable but we can use marginal effects to explain their behavior.



Ultimately we decided to go with the **XGBoost model** which outperformed all the other models. We can conclude this since its ROC curve surpassed the curve of all the other models at every step.

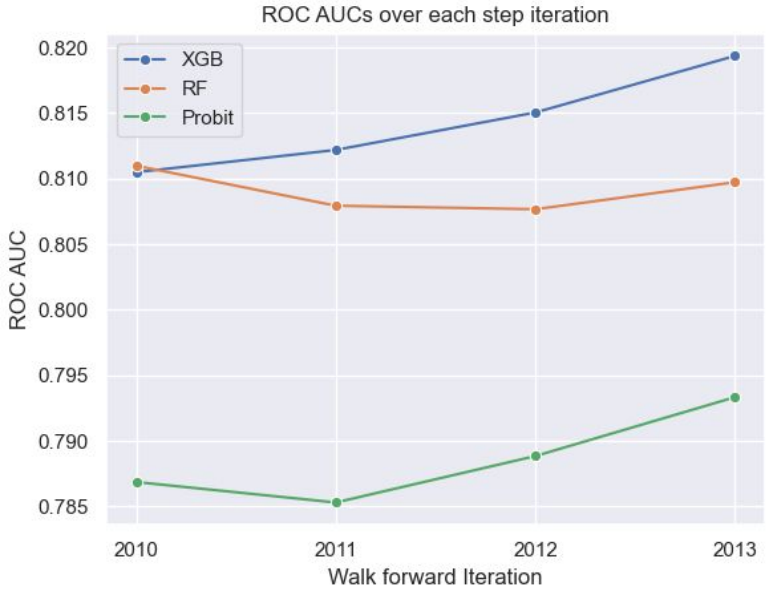
NYU

Evaluation - Walk Forward

We **assume that BM will be re-training their model on a yearly basis**. Using this assumption we run a walk-forward evaluation of the models iterating over each year we get access to new firm statements (simulating loan applications).

The XGBoost model increases its ROC AUC as it gets trained on more data. A characteristic we are looking for in our final model.

While for the Random Forest and Probit models there are instances where their ROC AUC decreases or stays the same despite gaining more data.



NYU

Final model - XGBoost Classifiers

Given our results from the evaluation and [illegible] we selected the XGBoost Classifier as our final model

- **XGBoost is an ensemble learning method based on decision trees that uses gradient boosting to improve predictive performance.**

ROC AUC: 0.819

From walk-forward results

Parameters chosen after optimization:

- *ETA*: 0.1 - Learning rate
- Tree based parameters:
  - *N_estimators*: 250 - uses 250 trees
  - *Max Depth*: 4 - decision tree depth
  - *colsample_bytree*: 0.5 - Choose half of the feature for each tree
  - *Subsample*: 0.8 - uses 80% of the data for each tree
- Regularization:
  - *Alpha*: 0.1 - L1 regularization
  - *Lamda*: 1 - L2 regularization

- Final features selected - *marginal effect size on next slide*
  - ROE
  - Regional PD
  - Financial Leverage
  - CFO
  - Sales Growth - (For recurring only)
  - DSCR
  - Profitability Ratio
  - Net Income Growth - (For recurring only)
  - Quick Ratio

Training Data	Defaults	Non-Defaults
New Firms	2,390 (0.01%)	235,321 (0.99%)
Recurring Firms	10,548 (0.013%)	775,293 (0.987%)

NYU

Final Model - Marginal Effects

Average Marginal Effects

In both models Region and ROE have a large average marginal impact.

This is followed by Financial Leverage, and CFO ratio

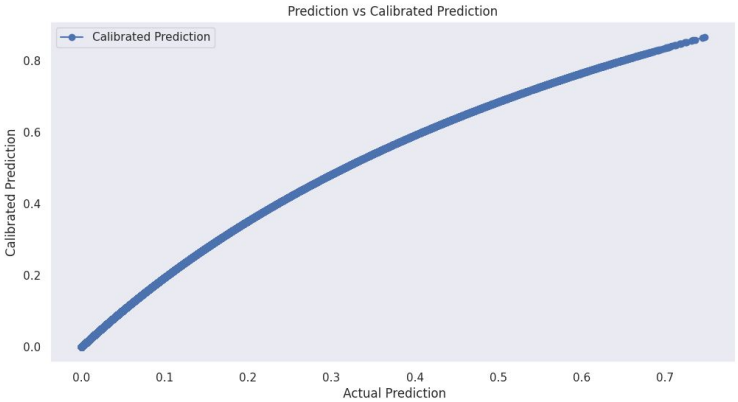
Features (First Model)	dy/dx
regional_code_pd	14.712380
roe_quantile_values	2.732044
financial_leverage_quantile_values	1.070400
cfo_quantile_values	0.809433
profitability_ratio_quantile_values	0.514005
dscr_quantile_values	0.368387
quick_ratio_v2_quantile_values	0.037000

Features (Re[illegible]	dy/dx
roe_quantile_values	4.436720
regional_code_pd	4.208318
financial_leverage_quantile_values	2.002530
cfo_quantile_values	1.983105
sales_growth_quantile_values	1.393474
dscr_quantile_values	0.961318
profitability_ratio_quantile_values	0.875804
net_income_growth_quantile_values	0.063407
quick_ratio_v2_quantile_values	-0.932354

NYU

Calibration

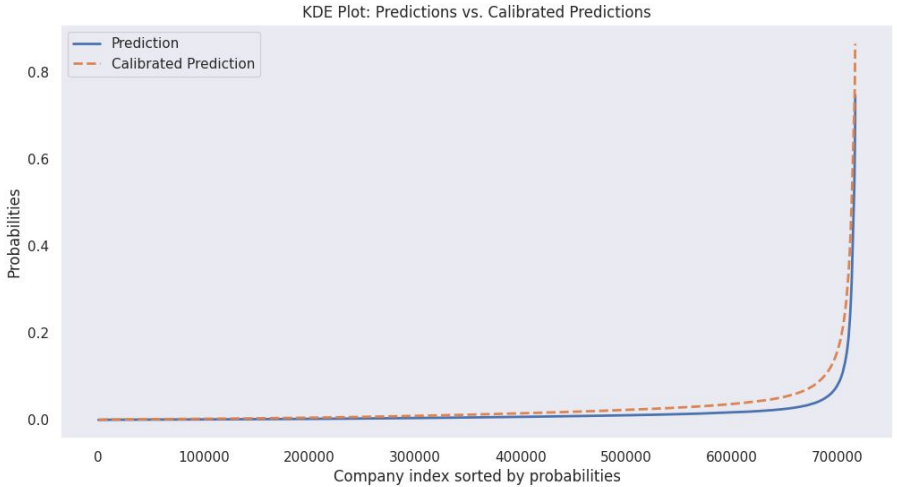
Sample population base rate : 0.126%
True population base rate: 2.7% (mean rate from 2010, 2011, 2012)



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

π_S : sample population base rate
π_T : true population base rate
p_i : model output for observation i
p_i^* : calibrated model output for observation i

Uncalibrated model mean probability: 1.27%
Calibrated model mean probability: 2.49%



True population base rate is derived from annual reports of **'BANCA D'ITALIA'**

https://www.bancaditalia.it/pubblicazioni/relazione-annuale/2012/en_rel_2012.pdf?language_id=1
pg(160)

Tech Stack & Deployment

Our final harness works by saving the following objects which are used in preprocessing and inference:

- The final XGB model is stored as a pickle file
- Custom bins for ratios which we defined from training data.
- Conditional Default Probability values for each bin (from training) which are used during inference.
- Historical values for sales and net income from all firm years in our training dataset. Used for growth calculation during inference.
- On calling the harness function, our user defined functions work on the input provided and a nx1 vector is generated (n: rows in the input_csv) and outputted as a csv file (Calibrated probabilities).

Dependencies:

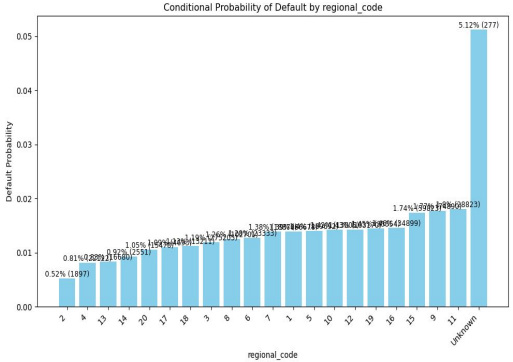
- **Python**
- **Pandas**
- **Numpy**
- **XGBoost**
- **User defined functions**

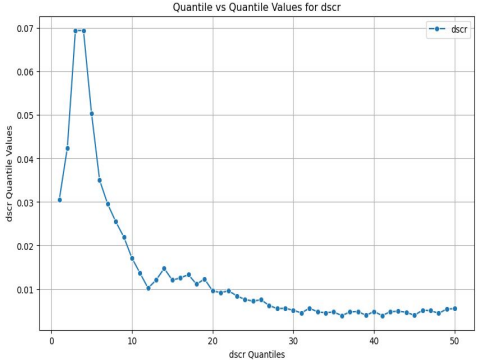
Appendix

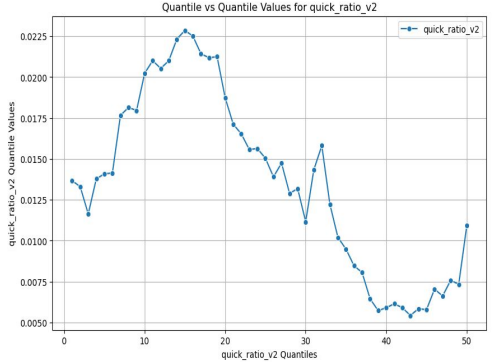
Contributions:

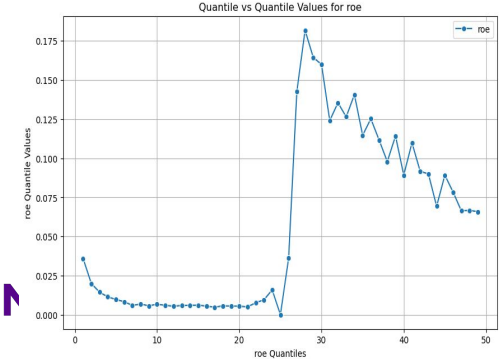
- **Oliver: Preprocessing, Conditional Probabilities (continuous and discrete)**
- **Juan: Modeling, Harness Function and testing**
- **Rohit: Calibration, Marginal Effects and testing**
- **Vibhor: Presentation,Visualizations and testing**
- **All Members: Literature Review, Planning, EDA**

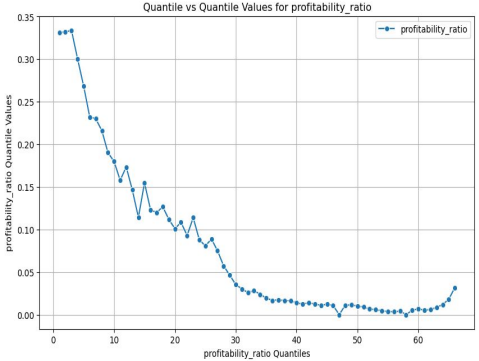
Appendix (Conditional PD)

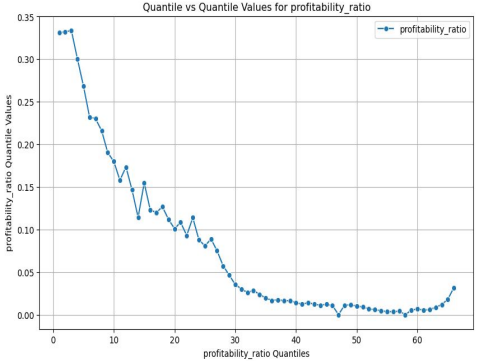




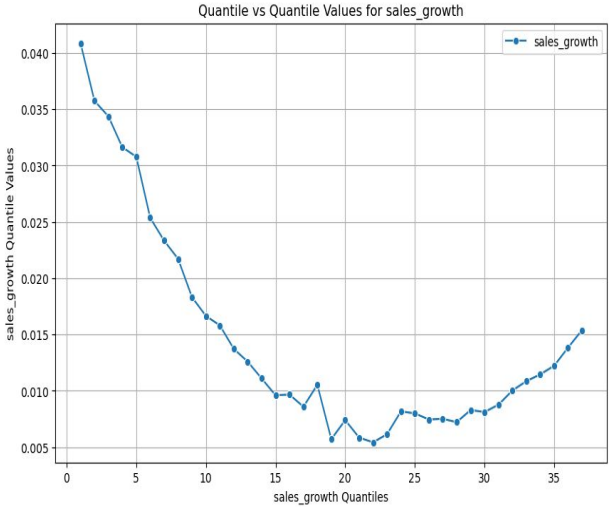


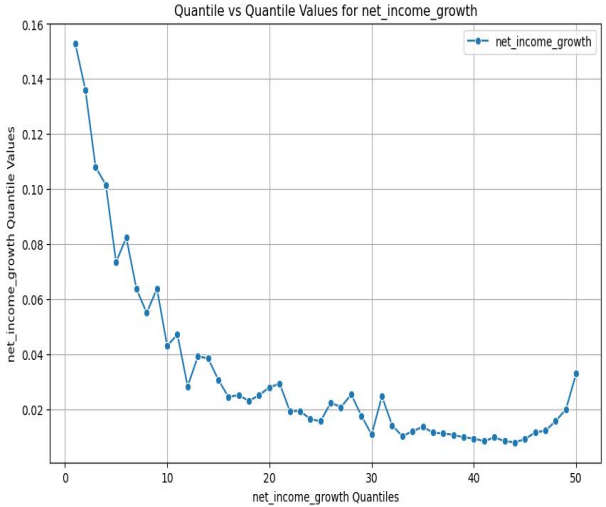






Appendix (Conditional PD)





Variance Inflation Factor (ViF): Recurring Firm

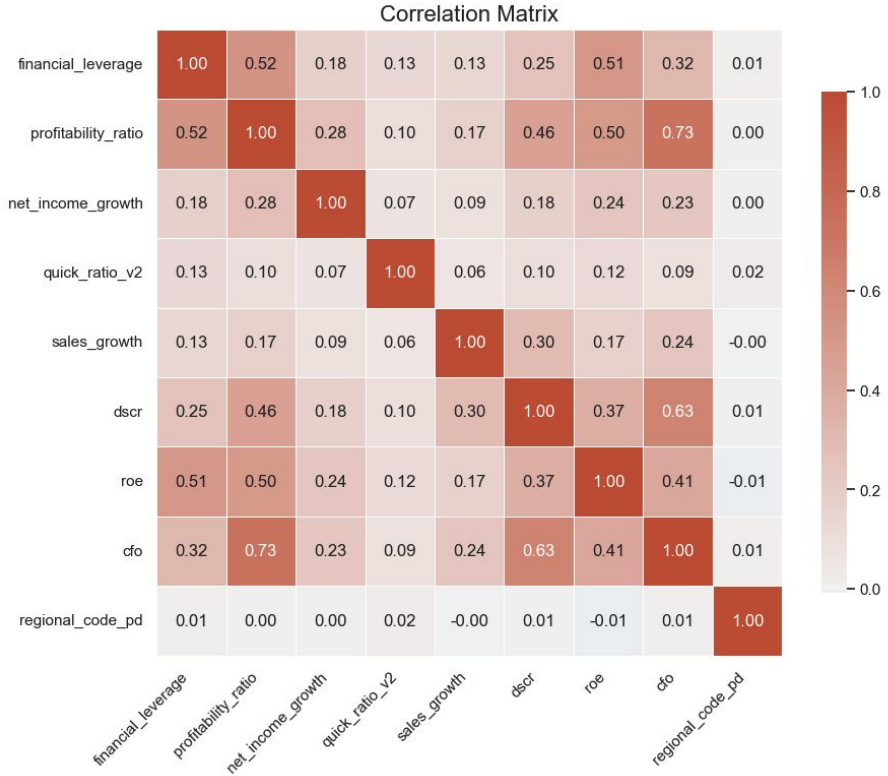
Variables	ViF
regional_code_pd	7.909020
financial_leverage_quantile_values	1.976428
profitability_ratio_quantile_values	3.473996
net_income_growth_quantile_values	2.374810
quick_ratio_v2_quantile_values	6.381658
sales_growth_quantile_values	3.711281
dscr_quantile_values	3.142687
roe_quantile_values	2.256744
cfo_quantile_values	4.197874

NYU

Variance Inflation Factor (ViF): First Time Firm

Variables	ViF
regional_code_pd	6.837453
financial_leverage_quantile_values	2.016479
profitability_ratio_quantile_values	3.182143
quick_ratio_v2_quantile_values	6.697786
dscr_quantile_values	2.754737
roe_quantile_values	2.011212
cfo_quantile_values	3.926684

Correlation matrix



NYU