Date: 11/22/2022

$ XXXXXXXXX

DOLLARS

Banca Massiccia: Credit Risk Modeling

Group: Chartreuse

Members: Anthony Chen, Khevna Parikh, Adeet Patel

12345678 12345678 12345 Authorized Signature

Imagine that...

Michelle needs a loan. She goes to the bank, Banca Massiccia and provides her financial information.

The bank needs to assess her financial standing and whether or not she will be able to pay them back.

How should the bank set her interest rate and underwriting fees? How likely is it that Michelle will be unable to pay the bank back, or defaults?

If Michelle can make timely payments and pay them back fully, great! The bank will make a profit. If Michelle can't pay them back, then the bank will take a loss.

No bank wants to lose money...

It’s important for the bank to understand the cost of doing business with a Michelle or any particular firm:

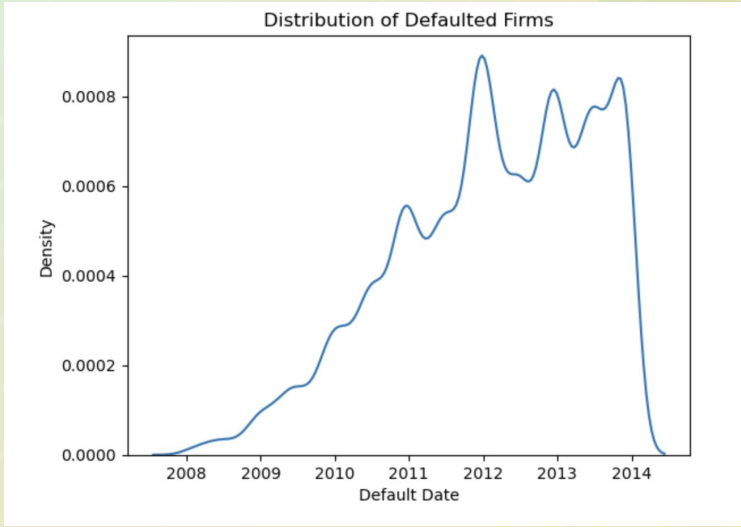
- Am I lending to a credible source?
- What is the amount I may lose, what do I have to gain from lending the money?
- How much risk am I taking if the obligor does default?

To assist Banca Massiccia in deciding whether an applicant is creditworthy and eligible for a loan:

- We estimate their probability of default (the probability that they will fail to make payments on their principal or interest payments over the next 12 months).

How is default defined?

- Occurs when an obliger is unable to pay its scheduled payments of principal or interest on time
- Is usually caused by lack of liquidity.
- If a firm defaults, its assets could be sold to repay debt, or given to the lender
- Could occur immediately upon a missed payment, or after a grace period
    - For our model, we calculate the probability that a firm could default within the next 12 months.
- Defaults are usually uncommon but not rare. 8% of the firms in our data defaulted between 2008 and 2014.



A data mining problem

- Our financial data is noisy and incomplete.
- Will require distributed approaches to effectively extract data from our voluminous dataset
- Should be able to able to handle new data and easily predict future probabilities

How can we use the raw financial/transactional data to find patterns or trends that could help Banca Massiccia reduce bad debtors? What mathematical algorithms can we use to evaluate the probability of default?

Past Models

- Structural/Causal Models:
  - Reason default through economic reasoning based on various quantities that describe the firm value or capital structure, the market value of the borrower’s assets falls

- Data-drive models:
  - Statistical irregularities of data relationships, without considering any economic theory

- Reduced-form Models:
  - Do not assume causal economic relationships, instead reason default as unpredictable event

Our Discrete-choice Model

- In addition to our economic intuition, we reason default by observing historical data to infer quantities relating to default likelihood of a firm
- We do not assume the availability of market information
- This grants us the flexibility to model events separately over different periods of time
- Involves probabilities of class membership

Data

- 44 features, include information on
  - Type of company, the industry the firm belongs to, the date when firm defaulted, geographic location
  - Firm's financial statements such as current and total assets or short-term and long-term debt

- 1.2 million records, annual observations where each record represents one firm-year

- Only non-finance/insurance companies with at least €1.5 Million in assets are included

Data Preparation

- Remove certain variables such as total equity for entire group ("family") since all records are NaNs, providing no significant information
- Change certain data types, e.g. convert the date of the financial statement from a string into a datetime object
- Substitute NaNs using interpolation - a statistical method used to estimate missing values by finding a line of best fit. For example, if we have a sequence of [1, 2, 3, NaN, 5], it would estimate 4.
- To prevent look-ahead bias, we sort the data by firm identifier then fiscal year and used this pairing as our index

Data Preparation Cont.

- Fiscal year in Italy is from January 1st to December 31st
- We assume that we would sit down Jan 1, 2009 to predict probability of default between for Jan 1, 2009 to Dec. 31, 2009. We would likely not have FY 2008 as it takes a few months to for firms to realistically release the data. So, we would only use data up prior and up to FY 2007
- Establish a target/dependent variable: an 0/1 indicator describing that a firm experienced a default event by a certain point in time. It would be 1 if the firm defaulted by time t+1, within the 12-month period, otherwise 0.

Feature Selection: Categorical Variables

- Distributions of the categorical variables (city of main branch, legal structure of the firm, and industry sector code) showed no significant difference between defaulted firms and non-defaulted firms

Top HQ Cities	Defaulted Firms	Non-defaulted Firms
58.0	0.099241	0.091272
15.0	0.088486	0.118400
63.0	0.049850	0.034036
1.0	0.037871	0.033343
17.0	0.033692	0.034251
48.0	0.027677	NaN

Top Industries	Defaulted Firms	Non-defaulted Firms
41.0	0.170164	0.147847
68.0	0.136651	0.196224
46.0	0.127839	0.111314
47.0	0.050372	0.041403
25.0	0.040389	0.043801
43.0	0.035092	NaN

Legal Structures	Defaulted Firms	Non-defaulted Firms
SRL	0.717531	0.735350
SRU	0.214067	0.159091
SPA	0.055567	0.084654
SAU	0.012631	0.020685
SAA	0.000153	0.000206
SRS	0.000051	0.000014

Feature Selection: Multicollinearity

- Computed Pearson's correlation to determine which factors are highly correlated or competing to explain the same variance

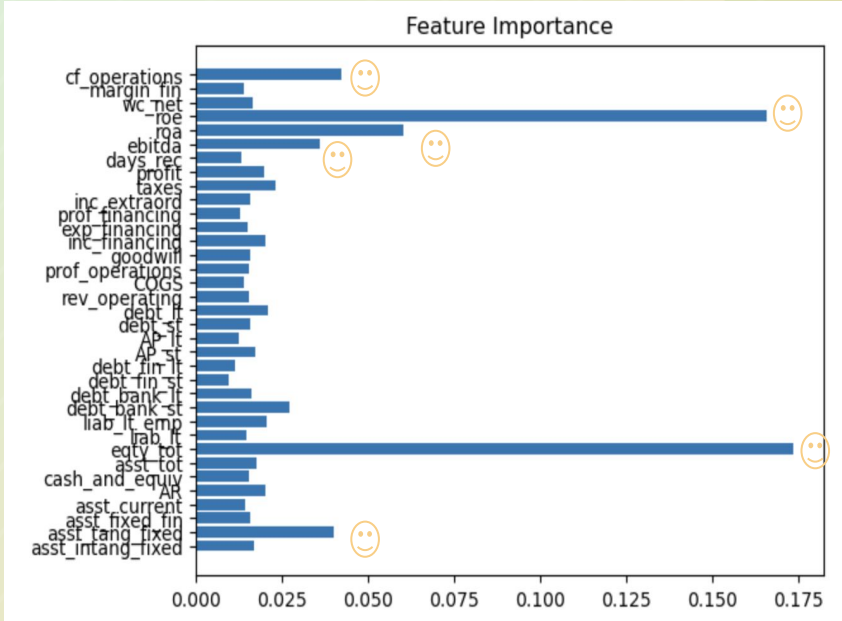
- Observe Operating Revenue & Cost of Goods Sold (COGS) and Taxes & Profit are highly correlated amongst others.
  - The calculation of operating revenue is gross income minus SG&A and other expenses, and gross income is sales minus COGS. Hence, COGS is already included in the calculation for operating revenue, which explains the nearly perfect correlation between the two variables.

	level_0	level_1	correlation
5	rev_operating	COGS	0.996275
7	taxes	profit	0.968024
1	asst_tang_fixed	eqty_tot	0.933470
3	asst_current	debt_st	0.933309
0	asst_tang_fixed	asst_tot	0.932832
2	asst_current	AR	0.932226
4	asst_tot	eqty_tot	0.929916
6	prof_operations	ebitda	0.928359
8	roa	ebit_ta	0.920224

XGBoost Feature Importance

- Gradient Boosting Ensemble, multiple decision trees combined to produce one optimal result, allows us to easily retrieve importance of a certain variables
- Using the Gini Index, the model provides a score on how useful each variable was in construction for each of the boosted decision tree. The score is averaged across all the decision trees within the model.
- This is a completely randomized model, so we threw all our non-categorical variables and found 6 important features for our target variable

Gini Index calculates the probability of a specific feature incorrectly classified when randomly selected



Feature Engineering

It’s also important to look at features in comparison with others. For example, debt without assets doesn’t provide much information. Observing them together provides a better picture of whether the firm will be able to pay off its debt in the near future with its current assets. Thus, we derive the certain ratios displayed below.

- Beaver claimed that cash flow/total debt (**CF/TD**) provided significant discriminatory power in differentiating future defaulting from nondefaulting firms. Thus, we include this ratio in our analysis. Is there enough cash on hand to pay of the debt?
- The working capital to total assets (**WC/TA)** ratio gives us an idea of whether a firm will experience difficulties meeting its short-term liabilities, indicated by negative or low working capital.
- The **EBIT/TA** ratio is a credible measure of a firm's profits from its assets before deducting factors like interest and tax.
- We define **leverage** as 1 - (total equity / total assets). Intuitively, leverage tells us the percent extent to which the company is borrowing against their total assets.

Feature Engineering Cont.

- Due to the large scalability of our data, we transform certain variables like short-term accounts payable using log to reduce this skewness
- Additionally, we found a 16% difference in median profit between defaulting and non-defaulting firm - the highest across all features
- Since we are predicting a one-year probability, we assume that variables associated with short-term debt and current assets would be useful hence we include a ratio of this in our analysis

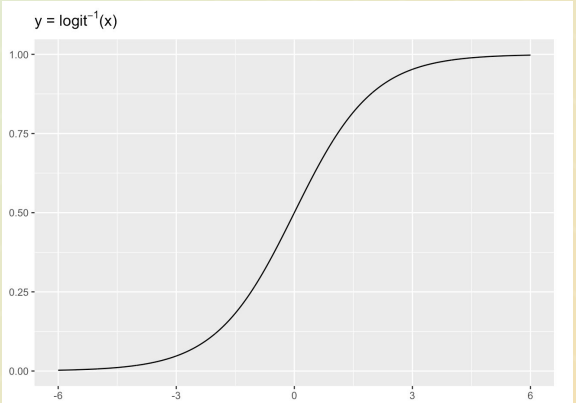
Subset of features considered in the following models are:

Operating cash flow, return on equity, return on assets, EBITDA, Total Equity, Tangible assets,profit, financial profit, leveral, short-term debt to current assets, short-term accounts payable, working capital to total assets, cash flow to total assets, and EBITDA to total assets

Baseline: Logistic Regression

- Models the probability of a firm experience a default event within the following 12-month period depending on the values of various independent variables discussed in the earlier slide
- Since our dependent variable is binary in essence, 1 if the firm defaults by time t or 0 otherwise, we can use logistic regression, a classification algorithm and use the predict_proba function to predict the probability of default
- This special type of linear regression predicts the probability of a default by fitting data to a logit function

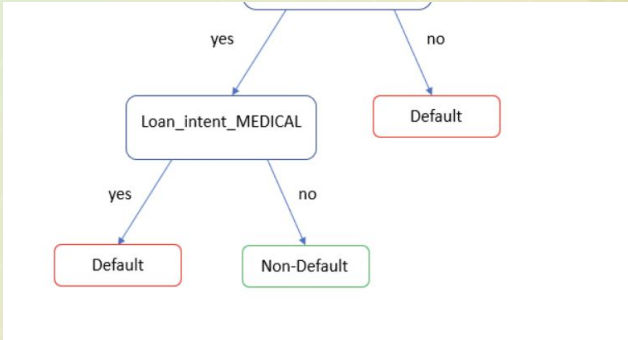
The ease of feature explainability in linear fashion allows us to choose logistic regression as our baseline model



- Linear Regression estimates a function such that each feature as linear relationship to the target variable. Each detail in financial statements is in the form of column-ized features. Thus, we assume every exponentiated coefficient corresponding to each feature increases or decreases the default rate.

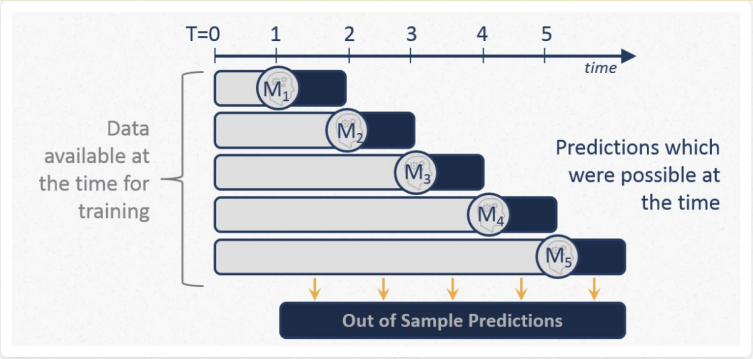
Final Model: XGBoost

- Decision Trees create predictions similar to logistic regression. XGBoost instead uses an ensemble of decision trees.
- These splits of our financial information give us an intuitive sense of probabilities of each event
- Flexible ability to take in a various features and use the predict_proba function to predict the probability of default
- Can capture the nonlinear separation in the data
- Pros: Parallelized tree building, efficient handling of missing data, built in cross-validation techniques, regularization techniques to avoid overfitting



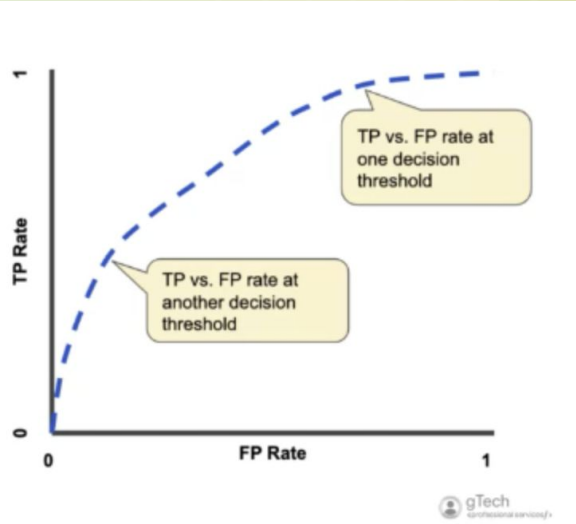
Training Model with Walk-Forward Analysis

1. Train the model using all observations available on or before the selected year, t
   a. We start at 2007, the earliest training observation available
2. Generate a test set consisting of observations for next year, t+1
3. Predict on the this test and save results into an array
4. Increment year by 1 and repeat the above steps
   a. Next iteration would have a train set consisting of observations from 2007 and 2008 and test set of 2009 observations
5. Evaluate the model using all predictions



Evaluation Metrics: AUC

- Goal is to find model parameters so that the predicted values are most similar to actual values, i.e want to minimize the loss
- Receiver operator characteristic (ROC) curve: plots True Positive Rate (TPR) and False Positive Rate (FPR)
  - TPR = TP / (TP + FN)
  - FPR = FP/ (FP + TN)
- Area Under the ROC Curve: measures the entire 2-D area underneath the entire ROC curve from (0,0) to (1,1)
- We use AUC as a measure to evaluate our models, which tells us how well the true positive, or default is predicted vs. the true negative, or non-default.



Baseline: Logistic Regression

	Variables Used:		AUC:	
	Profit		All variables used	0.67
6 features extracted through XGBoost	EBITDA		4 ratios only	0.75
	Total Equity			
	Return on Equity		4 ratios plus profit	0.8
	Return on Assets			
	Operating Cash Flow			
	Fixed Tangible Assets		Log of 6 features extracted through XGBoost	0.73
4 Ratios	Working Capital to Total Assets		Final Baseline	0.81
	Debt Ratio			
	EBITDA to Total Assets			
	Operating Cash Flow to Total Assets			
	Log Short Term Accounts Payable			
	Short-term Debt to Current Assets			

Even though the 4 ratios had satisfactory result, we thought it was important to include variables in our model which give us information about short term debt and assets - since we are predicting for the next 12-month period.

Our final baseline model consists of:

- Profit
- Debt Ratio
- EBITDA to Total Assets
- Operating Cash Flow to Total Assets
- Working Capital to Total Assets
- Log Short Term Accounts Payable
- Short-term Debt to Current Assets

And resulted in an AUC of 0.81.

Final Model: XGBoost

Combination of similar input features of the Logistic Regression model were used.

Note: even though distributions showed no statistically significant amongst our categorical variables, we still fitted a model using all variables and received much lower AUC in some case

The final Optimal XGBoost model includes

- Debt Ratio
- Financial Profit
- EBITDA to Total Assets
- Operating Cash Flow to Total Assets
- Working Capital to Total Assets
- Log Short Term Accounts Payable
- Short-term Debt to Current Assets

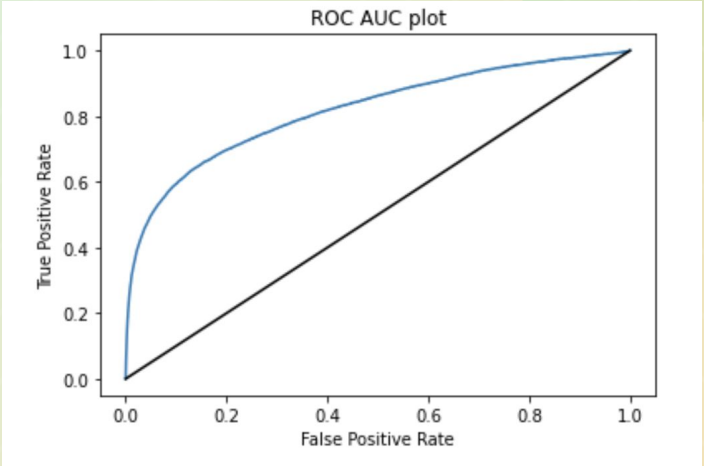
With an AUC of 0.82.

features= Index(['wc_ta', 'ebit_ta', 'leverage', 'cf_to_debt', 'profit', 'log_AP_st'], dtype='object')
AUC = 0.8202482573595342
features= Index(['wc_ta', 'ebit_ta', 'leverage', 'cf_to_debt', 'profit', 'ST_debt_to_cur_asst'], dtype='object')
AUC = 0.8159703752686267
features= Index(['wc_ta', 'ebit_ta', 'leverage', 'cf_to_debt', 'profit', 'log_cur_asst'], dtype='object')
AUC = 0.8145435486193747
features= Index(['wc_ta', 'ebit_ta', 'leverage', 'cf_to_debt', 'log_AP_st', 'ST_debt_to_cur_asst', 'financial_profit'],
AUC = 0.8244032686338129

What should the Bank do with this information?

- Our XGBoost has an AUC score is 0.824, meaning there is an 82% of chance that the classifier can distinguish between a firm defaulting within the next 12 months vs. a firm not defaulting within the next 12 months.

- When Banca Massiccia is assessing the creditworthiness of a prospective borrower, they can use this model to determine the probability or likelihood that the firm will default in the next 12-months.



- If the likelihood probability is high, the Bank can either reject a certain firm's application or lend the money with high interest rates and underwriting fees, in order to cover the costs in the case the firm does default

One more thing...

- In the future, Banca Massiccia can pass in new annual data into the model and the XGBoost can predict the probability of default that a prospective borrower will default in the next 12-month period

- Ideally, the model should be trained as the Bank accumulates more historical data to improve the accuracy of our model

- Concerns: The bias inherent in the dataset is selection bias because banks tend to lend to those who can afford to pay back. Additionally, how reliable is the data? Firms can inflate their income to look better on paper.

- Furthermore, it should not be used in times of economic uncertainties such as the pandemic. This model does not account for such financial changes.

- Even though XGBoost is known to be a state-of-the-art model, it can still be wrong. Hence, the model should not be the sole determinant of creditworthiness, human are still smarter than computers.

Appendix

Contributions of Members:

Khevna Parikh: Formulating the business problem and the data mining problem, exploratory data analysis, feature selection, presentation slides

Adeet Patel: Training and modeling, Vectorization, Object-Oriented Programming, Evaluation & Deployment

Anthony Chen: Feature selection, Modeling, Calibration, Evaluation

Calibration

The calibration curve shows that our model is pretty accurate.
If we were to further calibrate our generated probabilities we would follow the next steps:

- We calculated the sample default rate from the ratio of unique default firms occurrences in our dataset over the total number of unique firms.
- We chose a true default rate of 0.5% to represent a realistic probability of default.
- Then we calculated an adjusted default rate using Elkan's method