Outstanding job, Team Magenta!

Company Default Risk Prediction

100%

I have a number of questions that I would have liked to see addressed, and I think I disagree with some of your statements/assertions, but overall this is very solid work.

The three biggest gaps I saw were that you never actually showed the final model; you did not compare in-sample and out-of-sample results; and, most importantly, you included raw variables in your model, rather than ratios, but these are generally hard to interpret across companies or economic regiimes and thus make the algoritym work harder and be more susceptible to overfitting.

On the other hand, your formulation was very good (though, see prev. comment) and yougr discussion and interpretation of your results was superb!

I really appreciate the seriousness with which you approached the problem and the creativity you showed in your solution.

This was a great writeup of which you should be proud..

Mei Chen, Ken Zeng, Hongyi Zheng

Group **Magenta**

#FF00FF

Business Problem

Predict the probability that a potential borrower (with assets > €1.5MM) of Banca Massiccia will default on a principal or interest payment for a prospective loan over the next 12 months.

Why should we care?

- probability of default (PD) is the primary factor to determine the **Expected Loss** (a future credit loss)
- If predicting PD successfully, we could help Banca Massiccia **lower its lender cost** by stopping lending loans or charging higher loan interest rates to clients that have high probability of default.

AND lower rates to those that have low PDs, too..!

$$\mathrm{EL}=\mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD}$$

Ref: DS-GA 3001 Lecture 04 p.9

Past Approaches & Corresponding Limitations in this Dataset

- **Structural Models**
  - option-theoretic framework
  - (Merton, 1974)
- **Reduced Form**:
  - utilize market information
  - (Jarrow & Turnbull, 1975)
- **Econometric Models**
  - Early discrete choice model: using Accounting information (Beaver, 1966)
  - Hazard rate models (time until default)

- Assumptions on asset volatility(unknown); simplified **assumptions** on the corporate structure that could not accurately capture our dataset
- This dataset has less public trading information and more small companies; dealt little with illiquid tradings or **non-tradable debt**

"no" and "only"

- Accounting information could be restated by management; strong assumptions on **linear relationships** among features; high demand of data volumes

Our approach: An Ensembles of approaches

Core Idea: Use *open market information*, *financial statements* as features to train a *Machine Learning* model (CatBoost Classifier)

What we incorporated from previous work:

- Approximate asset volatility
- Utilize financial knowledge (e.g. Financial ratios)
- Utilize open market information that fits the target clients
- Benchmark against other popular ML method

What about other data-driven models from finance??

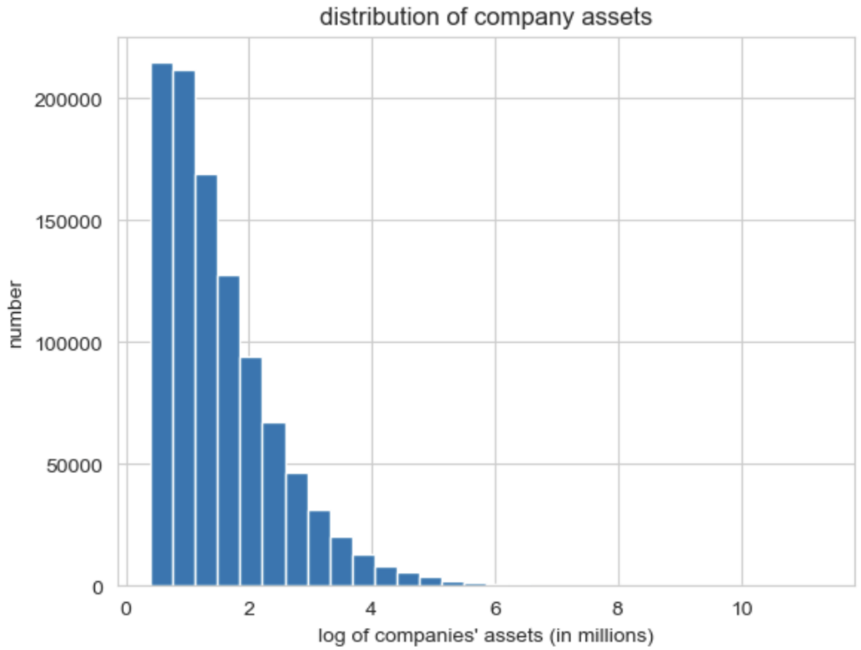
Why our approach is better/more innovative?

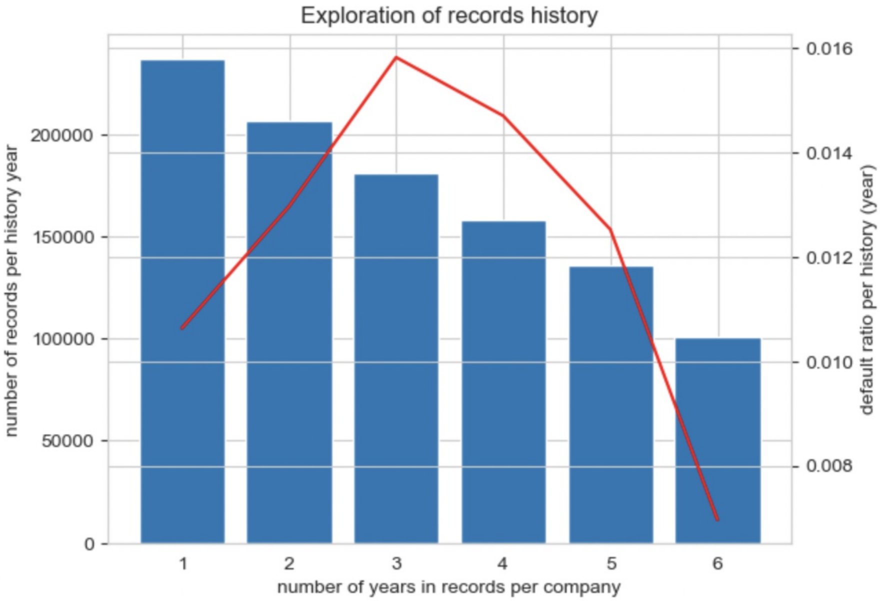
- Utilized strong machine learning model (CatBoost) with little distribution assumption that proven to be effective across many areas,
- Few have applied this model to predict default
- Kept interpretability using '**Attribute importance**' as it is usually rare among ML approaches, which could effectively transform the result into applications

AND YOUR FORMULATION !!! THIS WAS THE MOST IMOPORTANT CONTRIBUTION!!!

Dataset Initial Exploration

- Number of records: ~ 1M records
- Number of companies: 237117
- Avg. records per company: 4.29
- Company sizes: mostly small & medium sized (around 1.5M ~ 7.5M market value for 75% of companies)
- **Sampling bias**: the data distribution is manually cut off at total asset of 1.5M, which cause the statistical distribution of multiple features to be highly skewed

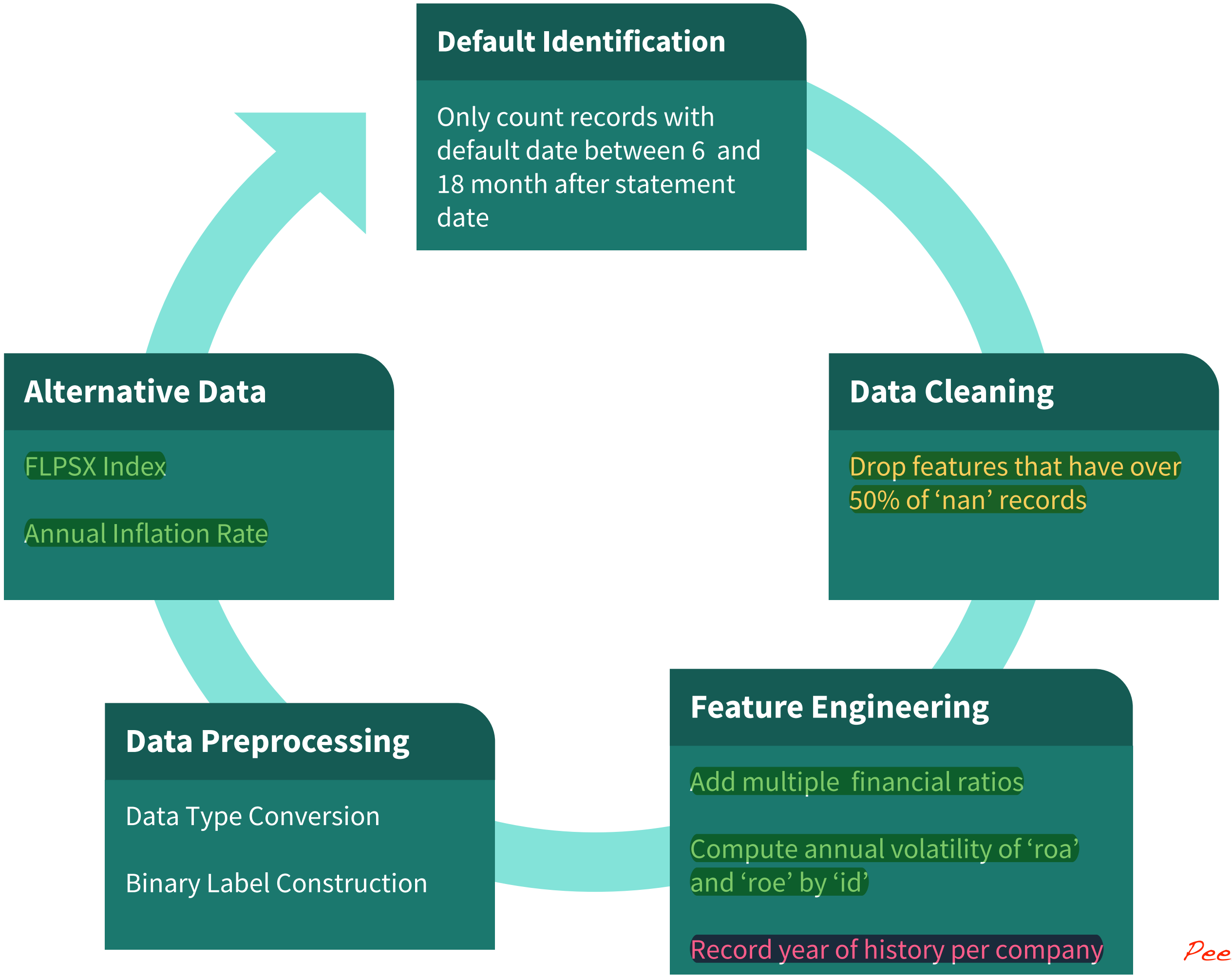




- 44 variables (including variables in financial statements)
- 3 categorical features, 2 of features with > 50% of null values
- Average sample default rate: 1.25%
- Financial statement record years: 2007 - 2012
- Company records histories: 1 - 6 years (and default most on 3rd year in record)

What is the intuition for this?
Hint what is the average number of statements in general??

Data Preparation: Overview



Default Identification

Only count records with default date between 6 and 18 month after statement date

Data Cleaning

Drop features that have over 50% of 'nan' records

Did you consider trying to reconstruct some of the missing values from other data based on, eg., financial knowledge?

Feature Engineering

Add multiple financial ratios

Compute annual volatility of 'roa' and 'roe' by 'id'

Record year of history per company

Peeking!!!

Data Preprocessing

Data Type Conversion

Binary Label Construction

Alternative Data

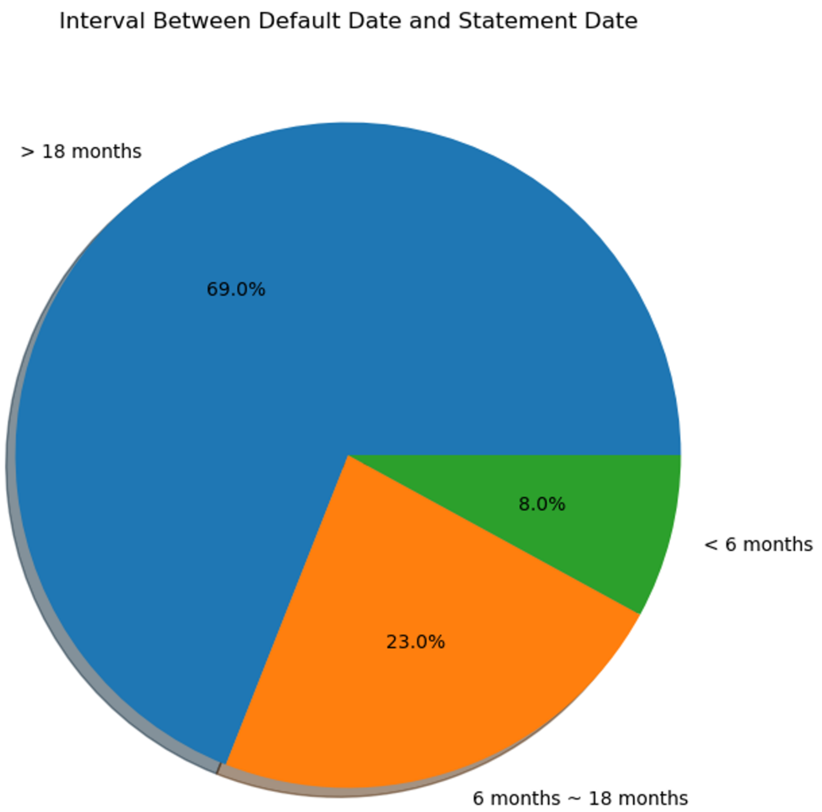
FLPSX Index

Annual Inflation Rate

Default Identification

As the actual release date of financial statements of one company usually lags in months, using data that has default date within **6 months** from statement date is peeking into future. Therefore, we exclude those records from our training set.

Besides, we regard records that have default date more than **18 months*** from statement as out-of-date records: they will be associated with negative '*is_default*' label.



*Ref: Chapter 4, Active Credit Portfolio Management in Practice (Jeffrey R. Bohn, Roger M. Stein)

Data Transformation & Training Set Construction

Added variables:

How did you control for time?

Volatility of return on assets: we calculates the rolling sample standard deviation on return on assets for companies that repeatedly appears in the record

is_default: a binary variable determined by *stmt_date* an *default_date* as the target variable of our model

Dropped variables:

After being used for data preprocessing and engineering , these following features themselves would not be put into the prediction ML model as a feature:

id: this will only be used for grouping companies and compute volatility

fs_year: redundant with 'stmt_date'

eqty_corp_family_tot & *days_rec*: variables that have over 50% of missing records

default_date: replaced by the binary target variable *is_default*

Did you consider trying to recon-struct some of the missin
values from other data based on, eg., financial knowledge?

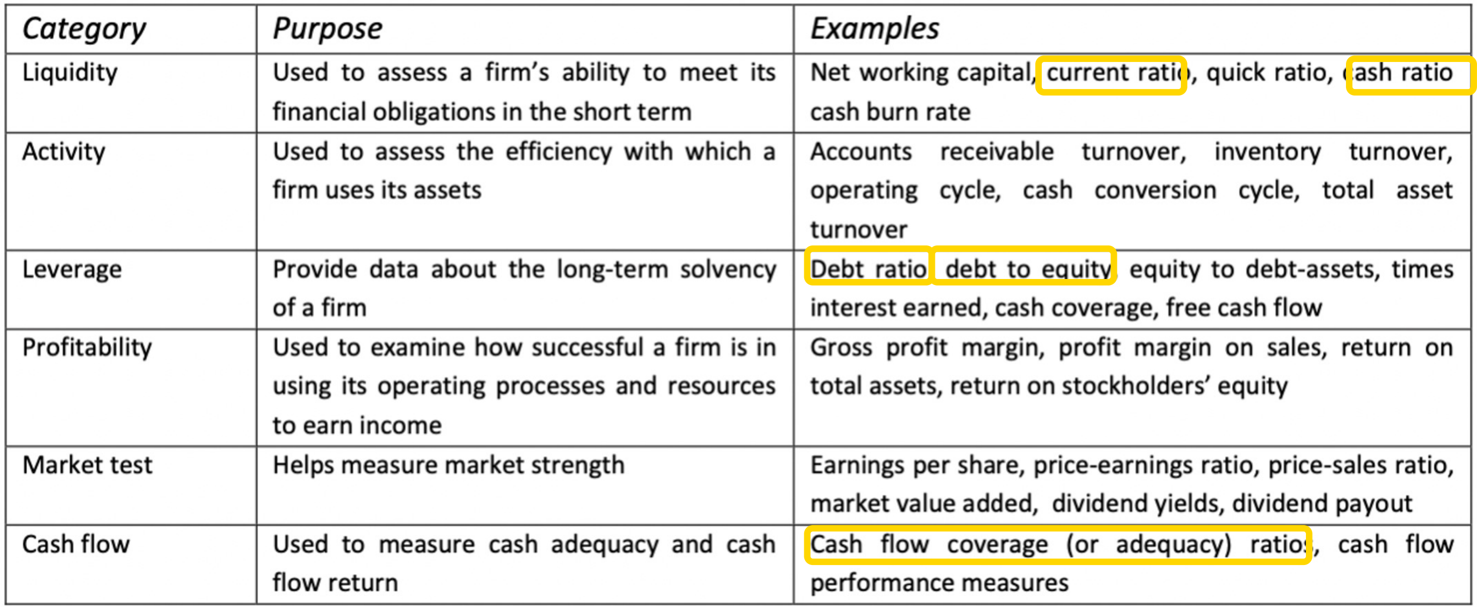
Deal with Missing Data:

Categorical variables: regard 'nan' as an individual category

Numerical variables: impute all 'nan' records with mean value of this variable

Note: These approaches mitigates but cannot completely eliminate potential bias stem from the missing data

Feature Engineering: Financial Ratios



Category	*Purpose*	*Examples*
Liquidity	Used to assess a firm's ability to meet its financial obligations in the short term	Net working capital, current ratio, quick ratio, cash ratio cash burn rate
Activity	Used to assess the efficiency with which a firm uses its assets	Accounts receivable turnover, inventory turnover, operating cycle, cash conversion cycle, total asset turnover
Leverage	Provide data about the long-term solvency of a firm	Debt ratio debt to equity equity to debt-assets, times interest earned, cash coverage, free cash flow
Profitability	Used to examine how successful a firm is in using its operating processes and resources to earn income	Gross profit margin, profit margin on sales, return on total assets, return on stockholders' equity
Market test	Helps measure market strength	Earnings per share, price-earnings ratio, price-sales ratio, market value added, dividend yields, dividend payout
Cash flow	Used to measure cash adequacy and cash flow return	Cash flow coverage (or adequacy) ratios, cash flow performance measures

Image Reference: G.I. White, A.C. Sondhi, D. Fried, The analysis and use of financial statements, John Wiley & Sons, Inc., 2003.

Why?? The data-driven models from the literature that you reference all use ratios!!!

- Financial ratios are often the first factors used to evaluate a company's performance in multiple aspects
- Ratios are traditionally difficult to capture using additive (logistic regression) and decision tree based (xgboost ,CatBoost, lightgbm, etc) approaches.
- Multiple literatures* about ML-methods in default prediction also have shown the importance of financial ratios in their model performance
- We mainly calculated ratios in ***Leverage***, ***liquidity***, ***cash flow*** categories highlighted above

*Ref: Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks … (see reference page)

Challenges: Collinearity

Data exploration shows that many features are highly similar and correlated. Using these features together could lead to lower model performance.

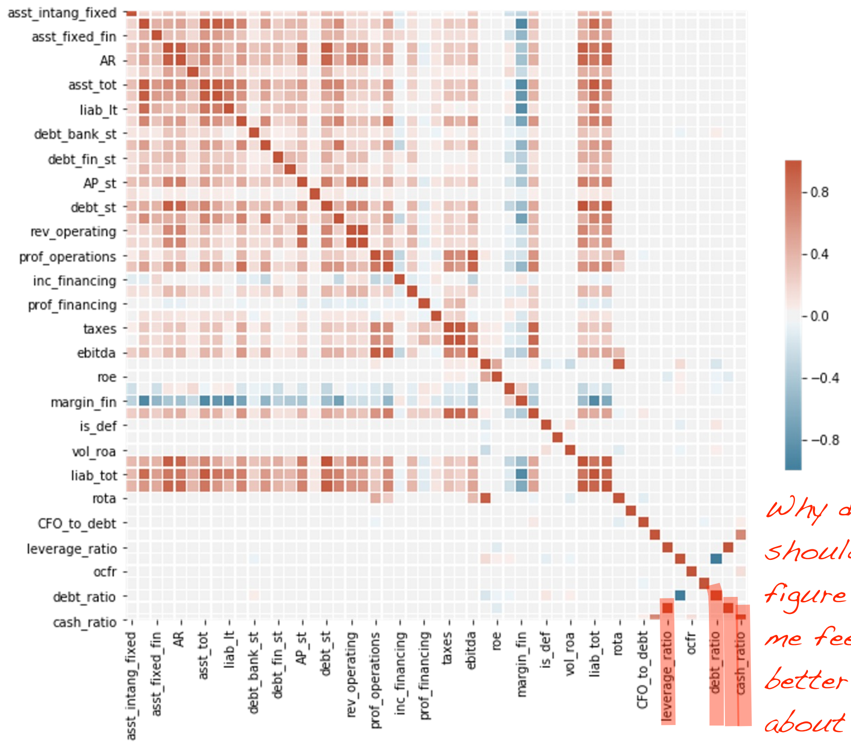
Possible Solutions

1. Use our **financial priors** to manually remove some dependent features
2. Use **variance inflation factor (VIF)** to identify features heavily influenced by multicollinearity
3. Use **feature selection** to identify independent features

How did you do this?

4. Use **PCA** to extract factors to use as features
5. Use algorithms that are less sensitive to collinearity, such as **decision tree-based classifiers**

Why do you think this?



Why do should this figure make me feel better about correlation?

Original features from the data tend to be show signs of multicollinearity,

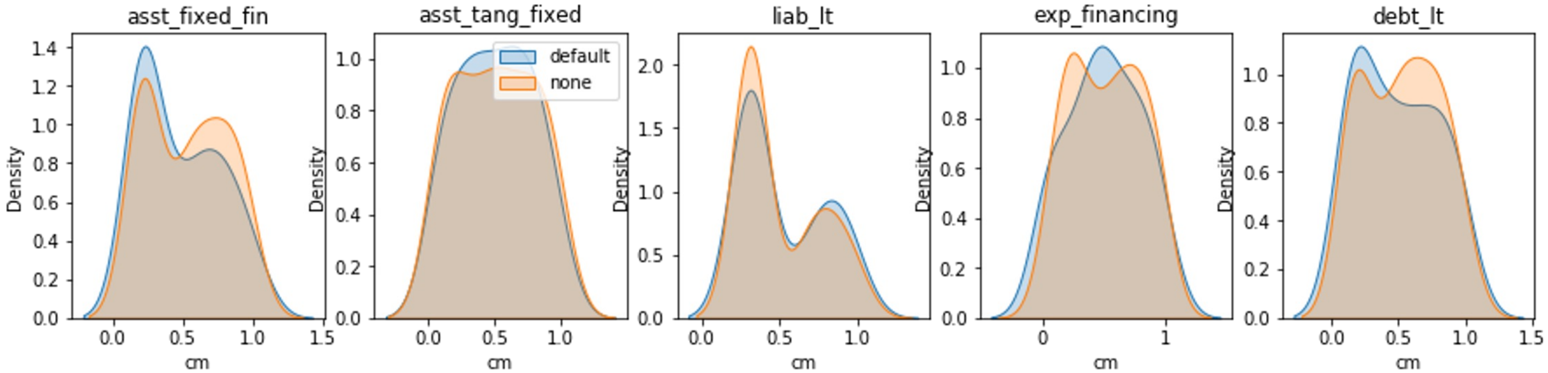
New engineered features tend to be uncorrelated with existing features

Feature Engineering: Observations

Outstanding!

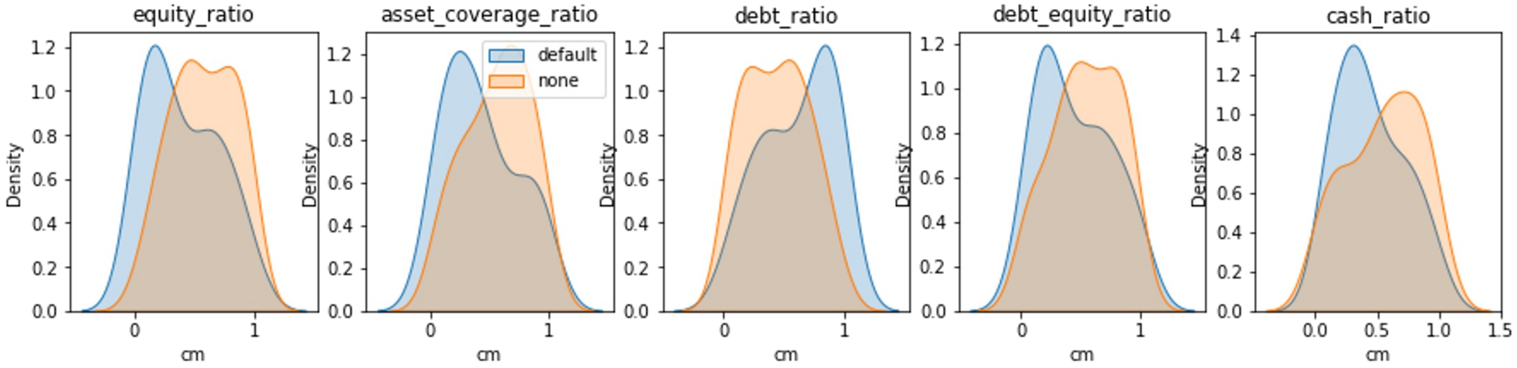
Original Features:

The distribution for defaulted and non-defaulted companies are almost identical, making the data **difficult to separate**



Engineered Features:

a list of financially meaningful ratios, which leads to **greater separability** within the two classes' distributions

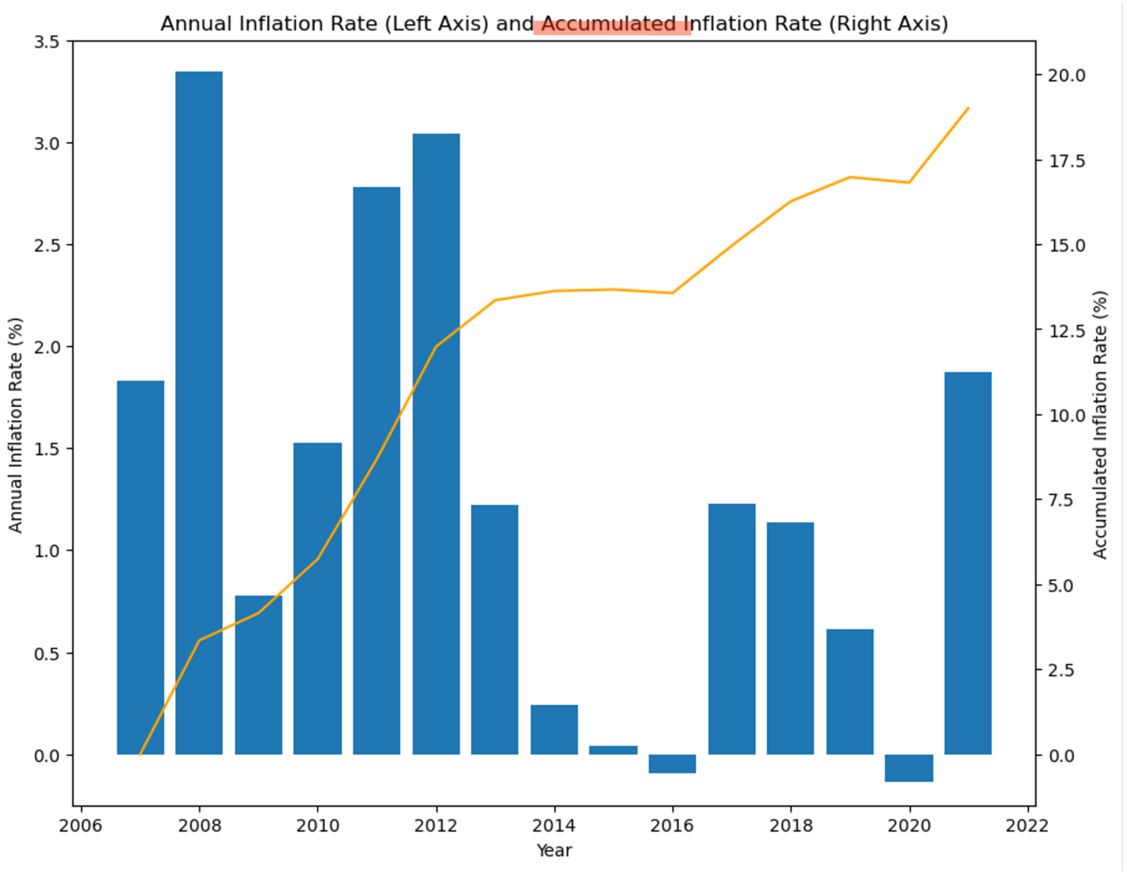


I would call this "external" data. "alternative" has a different meaning in general (see Lectures 6 and 9)

Alternative Data: Annual Inflation Rate

- The training dataset and testing dataset contains records spanning over 10 years
- **A Euro in 2007 worth much more than a Euro in 2017 in Italy**
- Therefore, we adjust all numerical columns (except ratios) with the accumulated **inflation rate** since 2007.

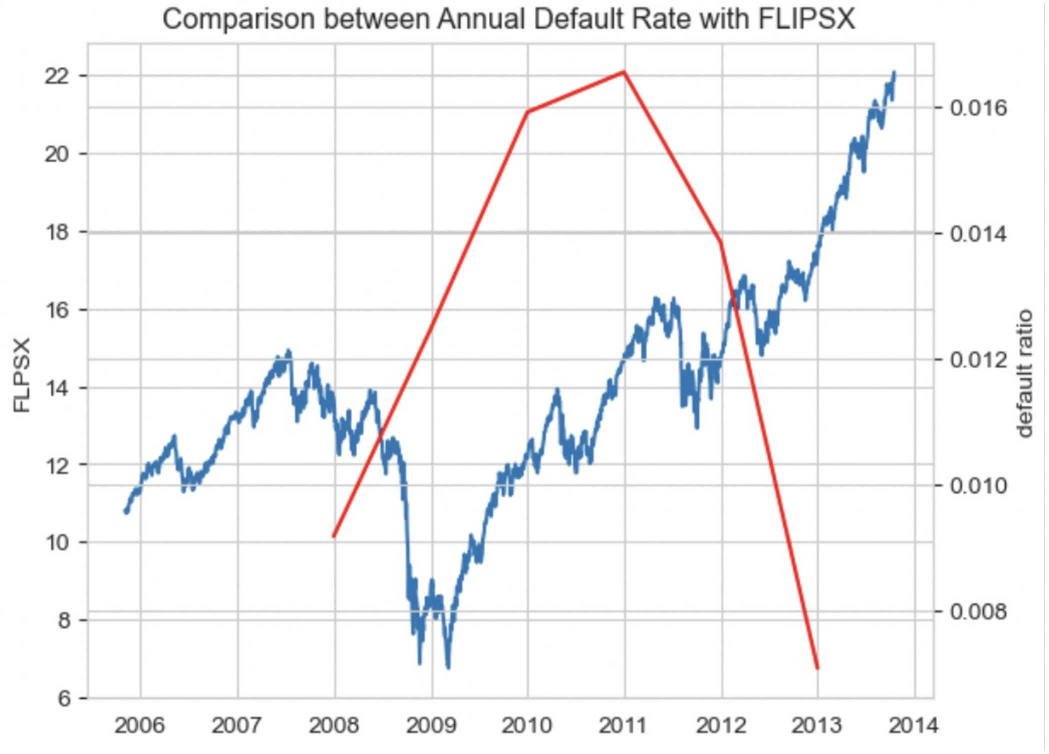
"cumulative"



I love this idea, but I worry that this is not a great proxy fo ritaly. (See next slide)

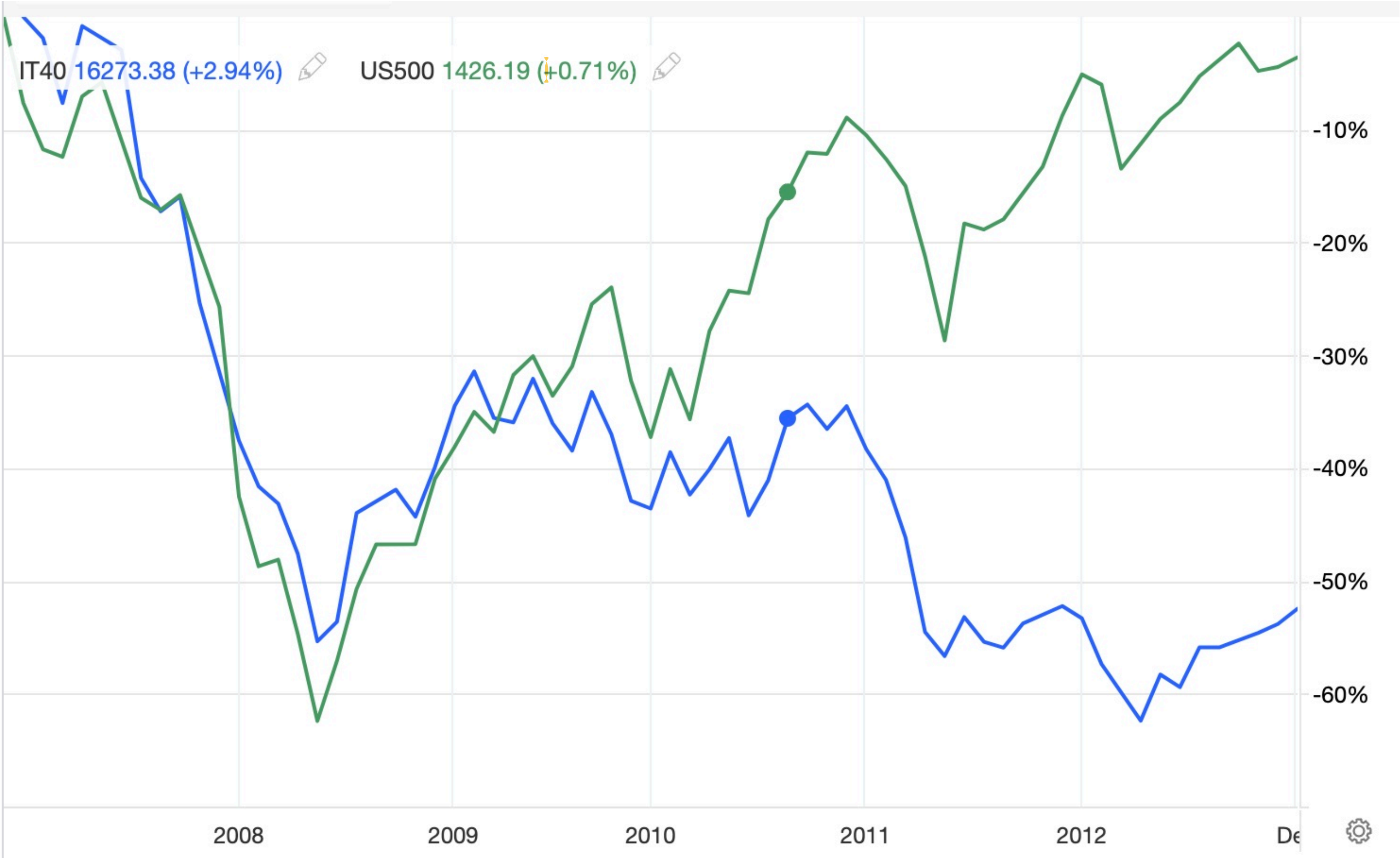
Alternative Data: Fidelity Low-Priced Stock Fund (FLPSX)

- A more fund that Normally investing at least 80% of assets in low-priced stocks, which can lead to investments in **small & medium-sized companies.**
- The company’s default trend roughly inversely proportional to FLIPSX trend with a little lag, and we believe this factor do affects clients’ default behavior in nature.



Ref: https://fundresearch.fidelity.com/mutual-funds/view-all/316345305

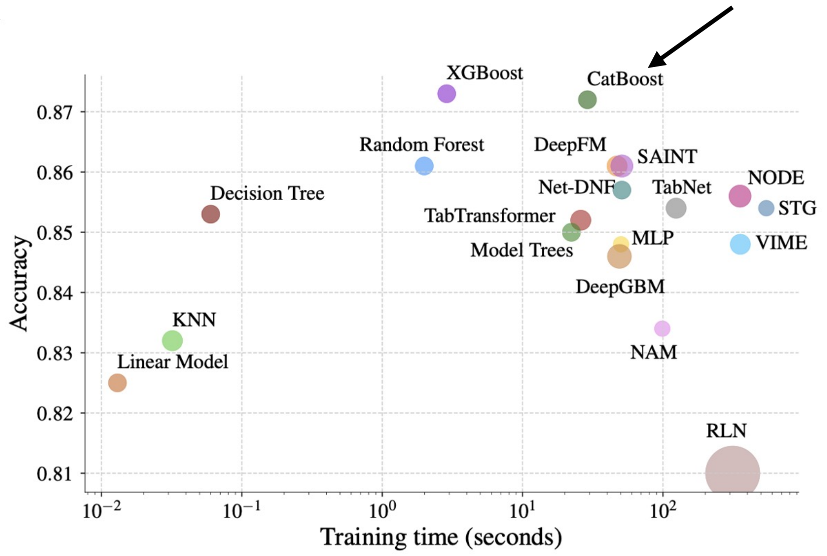
It seems that the returns on the Italian equity market may have diverged from the US in teh middle of your training window (and beyond)? Of course, these are larger companies, so it is unclear what would have happened with smaller ones, though these often tend to have lower correlation with the market overall)..

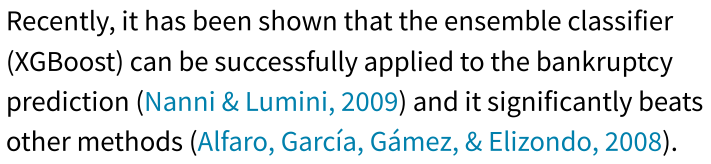


Our Choice: CatBoost

- One of the **best performing** out-of-the-box algorithms.
- Empirically shown to **outperforms** other common benchmark approaches.
- Converts categorical variables to statistically relevant numerical **representations**
- CatBoost is **NOT sensitive to multicollinearity** within the data.
- Offers **explainability** in the form of feature importance calculations and average feature prediction.

How do you know this?




Recently, it has been shown that the ensemble classifier (XGBoost) can be successfully applied to the bankruptcy prediction (Nanni & Lumini, 2009) and it significantly beats other methods (Alfaro, García, Gámez, & Elizondo, 2008).

As a result, we use most of variables as input **(37 original variables + 14 financial ratios + 1 alternative feature + 3 feature-engineered variables)**

Evaluation: model validation

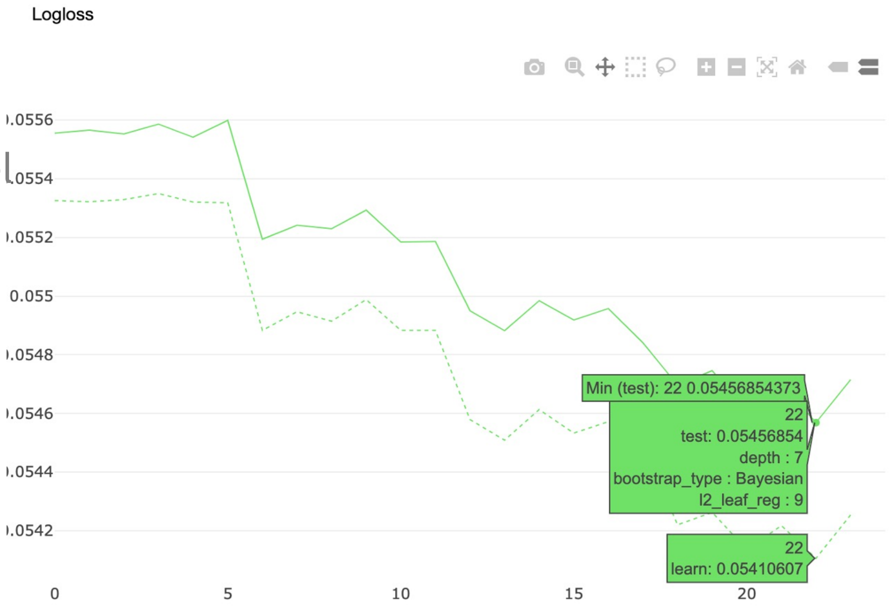
Train/Validation process:

- Split train & evaluation data by time (before/after 2011/01/01)
- Use out-of-time validation samples to evaluate model performance
- Tested on multiple model and select the best model
- Fix finalized hyperparameters and re-train on the entire dataset to generate final model

Hyperparameter Search Results:

- Depth (model complexity): **7**
- l2_leaf_reg (regularization strength): **9**
- Learning Rate: **0.1**
- Boosting Type: **Ordered** vs Plain
- Bootstrap Type: Bernoulli vs **Bayesian**

Were you able to drop some of the redundant features in the final model?



Visualization of how CatBoss loss decrease over hyperparameter search

Evaluation: Benchmark Comparison in Logistic Regression

Pros

- Logistic Regression is the most commonly used baseline machine learning model in default prediction.
- It’s **fast** to compute, **interpretable** and has decades worth of backing from both a theory and experimental standpoint.
- Some of these limitations can be effectively handled by newer machine learning frameworks, such as **CatBoost**

Cons

- Sometimes make **invalid assumptions** about the underlying data distribution
- **Requires feature engineering** and selection to capture non-linear relationships
- Low performance when there is significant **collinearity**
- Requires **interpolation** on null value

Evaluation: Model Performance

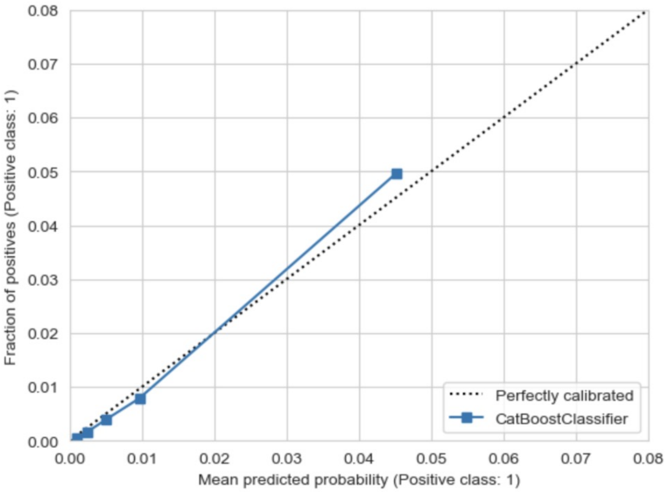
AUROC

- Catboost: **0.86**
- XGBoost: 0.84
- benchmark logistic regression: 0.79
- Ensemble: 0.85

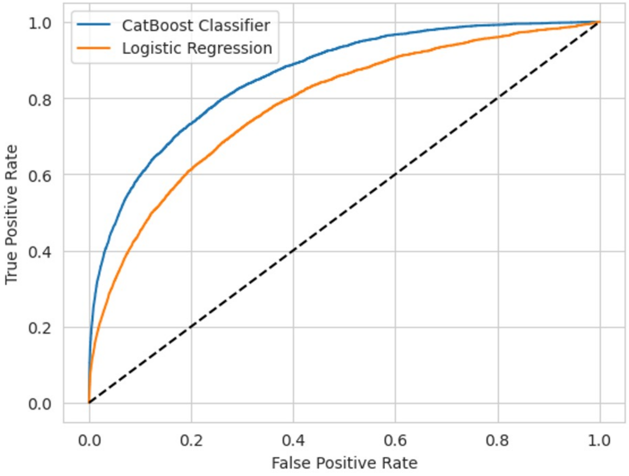
Analysis

- Since we do not have an available cost matrix, **AUROC** is used to evaluate the model performance with all possible thresholds
- AUROC is more robust to data imbalance compared to accuracy or recall based methods
- Calibration plot shows that model is close to sample probability of default (we assumed that this sample represent the true world PD distribution)

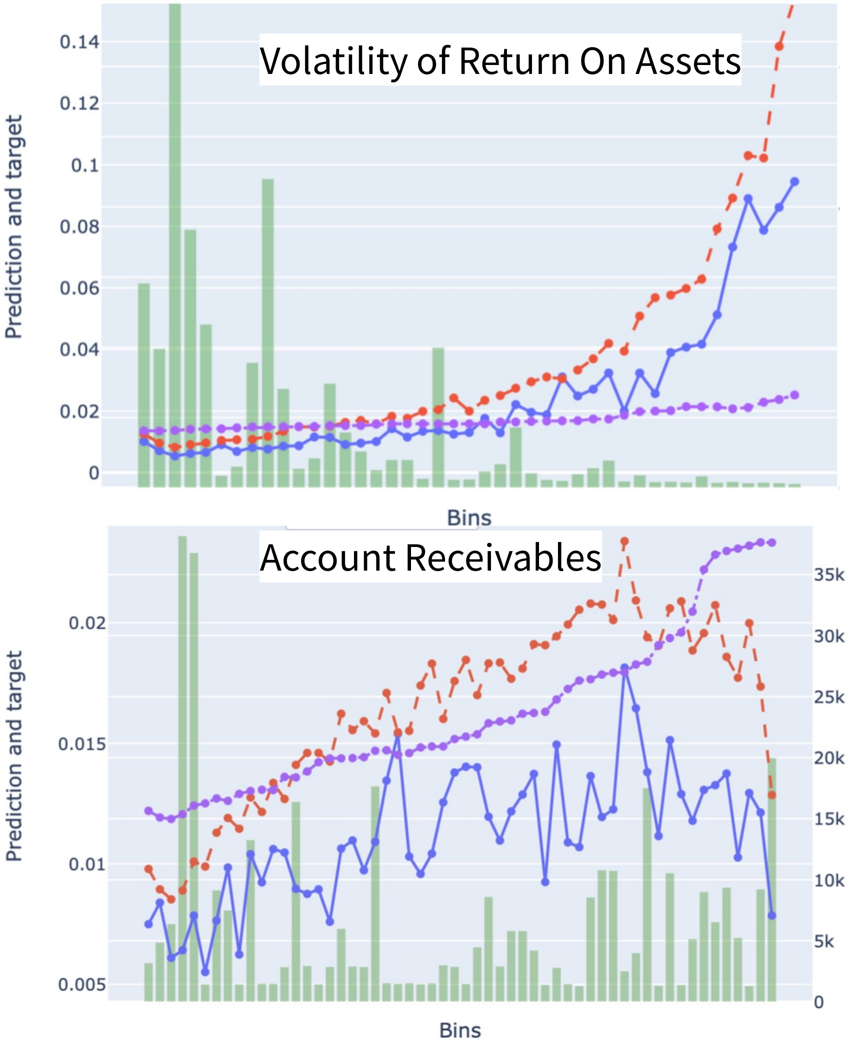
Model calibration curve



Receiver Operating Characteristic Curve



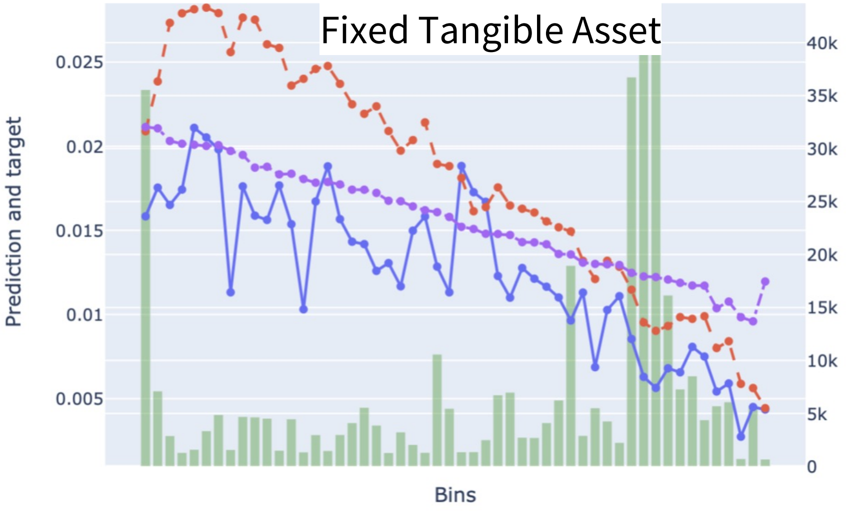
Evaluations: financial understanding (in top important features)

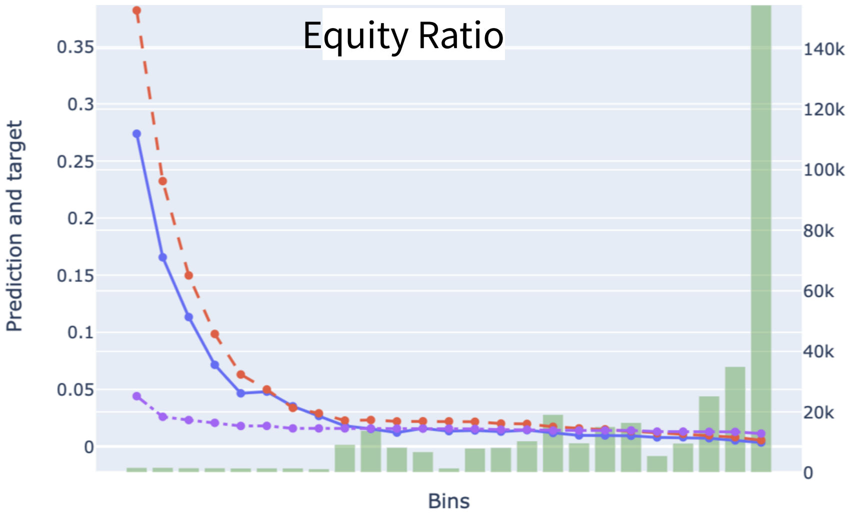


- **Approximated volatility of return on assets** and **Account Receivables** are two features that present rising trend in top important features
- Black-Scholes formula indicates that companies with higher *volatility on return* tends to default more, and our model shows the exact performance(see fig. on left) relative to equity!!!
- *Account Receivables* are accounts that haven't received debt from corporates' customers, and this trend fits with our nature instincts as only when corporate maintain a healthy financial state will they lent out debts to other companies.

●— Mean target
●-- Mean prediction on each segment of feature values
🟩 Objects per bin
●-- Mean prediction with substituted feature

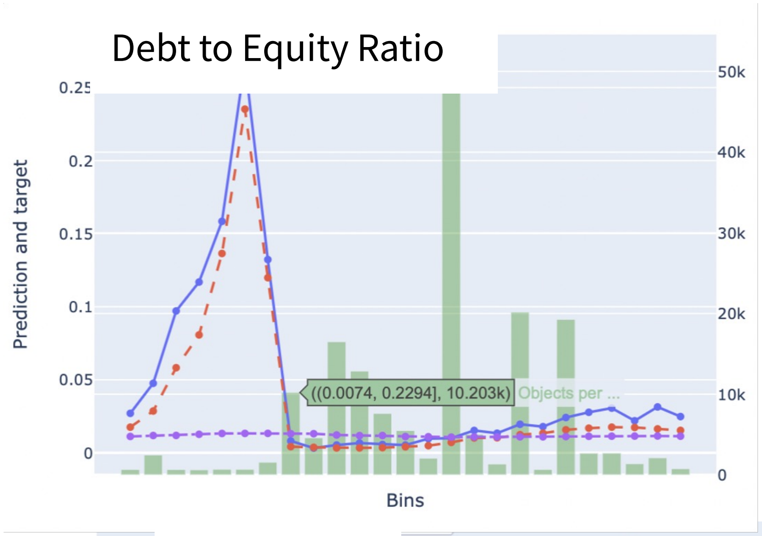
Evaluation: Financial Understanding (in top important features)



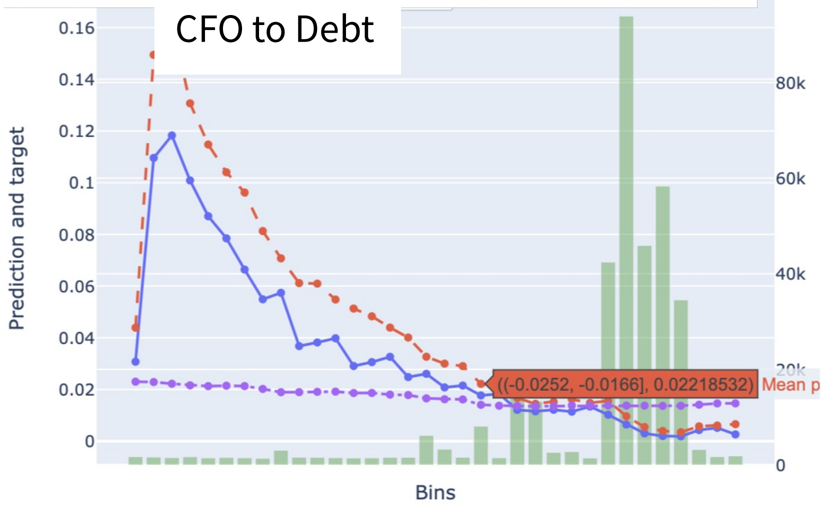


- Some features show a more continuous relationship: the probability of default decreases continuous as the company's **fixed tangible assets** and **Equity Ratio** increases.
- *Equity ratios* represents how much company is owned by shareholders and how much is leveraged by debt; the higher it is, the stronger solvency the corporate has, the better this corporate can met with long-term obligations, and it could a indirect sign on how company would default in one year as well.
- We suspect that the larger *physical assets* one company has, more stable this corporate could be, which results in a lower probability of default (Interestingly, *total assets* doesn't present this trend at all)

Evaluation: Financial Understanding (Features with Top Importance)



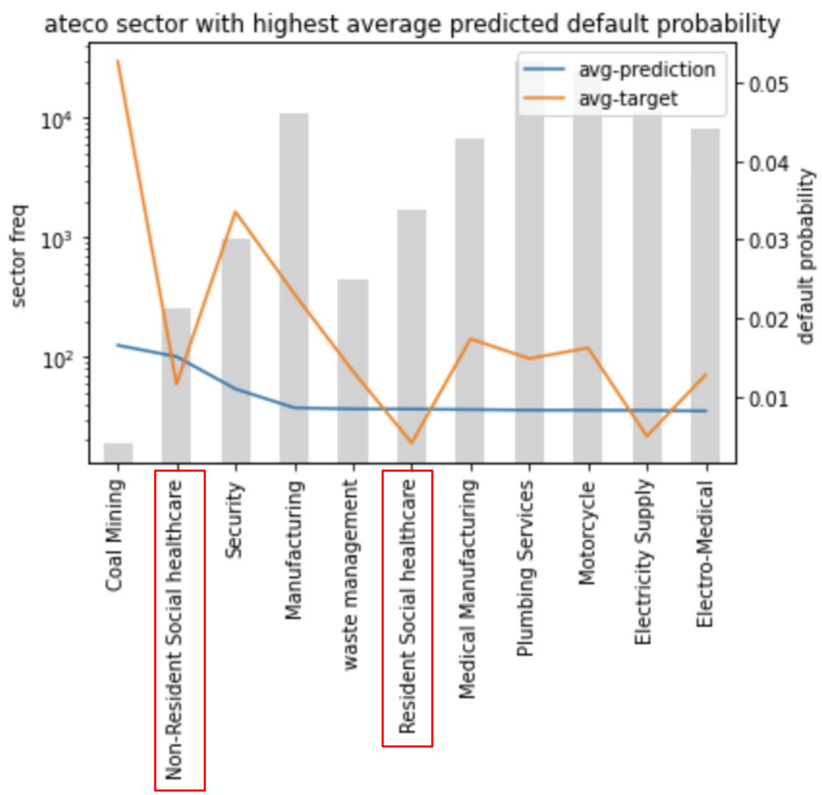
Debt to Equity Ratio



CFO to Debt

- Here we present visualizations with features that display similar trend: features have a clear **cutoff point** w.r.t to defaulting
- Company with negative **debt to equity ratio, CFO to debt, OCFR** (operating cash flow ratio) have a significantly higher chance of defaulting than those greater
- *Operating cash flow* can tell how much cash flow a business generates in a given time frame. If there are only negative cash flows, it wouldn’t be possible for company to pay off their liability in emergencies
- a negative *debt to equity ratio* indicates the equity of the company is negative - a sign that company has high chance of default

Evaluation: Model Biases



- We analyze the ATECO sectors with the top 10 average predicted default probabilities.
- Our model seem to **discriminate** against (residents & non-residents') healthcare companies despite their relatively low empirical default probabilities.

"be poorly calibrated" or "perform poorly" ?

- This model can mislead us into avoid lending to **financially safe** companies which benefits people.

Future Considerations & Limitations

- Currently we use snapshots of a company’s performance to estimate default likelihood. However, in the future we should try to evaluate this using the company’s **historical performance** across multiple years.
- Alternatively, we could also treat this as a company default as a **survival analysis** problem. This allows us to make estimates on when a company is likely to default.
- There are additional external features that we can also include into the dataset, including data specific to the Italian economy.
- We may also include more high-quality **alternative data**. For instance, indices that are more representative than FLPSX.

Deployment

When the client of the bank released its financial statement (usually April/March in the following year), the model can directly be applied to collect corresponding information and predict the future year's probability of default.

- **Potential risk:**
  - Our model doesn't assume **causal** predictions;
  - Training on imbalance dataset an cause model's output to be **biased**
- **Ethical/legal considerations:**
  - Our model's input requires slices of customers' **private data**, which could be illegal in some regions if the slice is too big and reveal certain information of customers
- **Setting that shouldn't applied:**
  - The default model should not be applied if there are sudden shifts in the data distribution, e.g. due to **systematic economic risk**
  - Cases such when economic recession, bank shouldn't only use results from our model to finalize the decision in making loans, as loans could **potentially stimulate economy**.
- **Risk with our proposed plan/how mitigate:**
  - Modifying output prediction based on class frequency

Contribution

- **Mei Chen**:
  - Feature Engineering: Volatility Calculation
  - Data Ingestion and preprocessing
  - Model understanding and evaluation
  - Preliminary Model Training and evaluation
- **Ken Zeng**:
  - Feature Engineering: Financial Ratios
  - Exploratory data analysis and visualization
  - Data Ingestion and processing
  - Logistic Regression Benchmark
- **Hongyi Zheng**:
  - Alternative Data collection
  - Feature selection and model evaluation
  - Hyperparameter Tuning
  - Developing Harness Function

Reference

Active Credit Portfolio Management in Practice (Jeffrey R. Bohn, Roger M. Stein)

G.I. White, A.C. Sondhi, D. Fried, The analysis and use of financial statements, John Wiley & Sons, Inc., 2003.

Esteban Alfaro, Noelia García, Matías Gámez, David Elizondo, Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks, Decision Support Systems, Volume 45, Issue 1, 2008, Pages 110-122,

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 36, 3028–3033.

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. Decision Support Systems, 45, 110–122.

Hyeongjun Kim, Hoon Cho & Doojin Ryu (2021) Predicting corporate defaults using machine learning with geometric-lag variables, Investment Analysts Journal, 50:3, 161-175, DOI: 10.1080/10293523.2021.1941554

Korangi, Kamesh, Christophe Mues, and Cristián Bravo. "A transformer-based model for default prediction in mid-cap corporate markets." European Journal of Operational Research (2022).

Lecture Notes, DS-GA 3001 Intro to Data Science in Finance: Discrete Choice

Fidelity® Low-Priced Stock Fund, https://fundresearch.fidelity.com/mutual-funds/view-all/316345305

Italy Inflation Rate,macrotrends, https://www.macrotrends.net/countries/ITA/italy/inflation-rate-cpi