



Improving Loan Underwriting at Banca Massicca

Group Emerald

Abhilash Anand, Jin Choi,
Robin Wang & Shruti Wagle



Business Understanding

Problem Definition

- The core business problem at hand is to minimize the risk of loan defaults for Banca Massiccia.

Motivation

- By predicting each prospective borrower's one-year Probability of Default (PD), the bank aims to enhance its loan underwriting process.
- Improving prediction of loan default rates can lead to better risk management, improved profitability, and potentially more competitive loan offerings.
- A superior default prediction system can make Banca Massiccia a stronger contender in an increasingly competitive market.

Stakeholder Impact

- The high default rate directly impacts stakeholders including shareholders, investors, management, bank employees, and customers.
- Addressing this issue could improve the bank's market reputation and enhance customer trust.



Data Mining Solutions

Approach Overview:


- Using a model trained on key features of historical bank transaction data, the model will produce a probability of default for each new company records.

Specific Techniques:

- Employ feature engineering methods to extract meaningful indicators from raw data (e.g., debt ratio, leverage, cash return on assets, etc).
- Utilize logistic regression for default prediction and provide explanation of the model to the stakeholders.
- Assess the model through the application of finance-specific evaluation like walk-forward testing.

Outcome:

The data mining solution seeks to deliver PD predictions that mirror real-world financial results. The output of the model, which is the probability of default, will assist the bank in determining the appropriate interest rates and underwriting fees for that company.





Review of Past Approaches

Past approach in Finance:

These methods traditionally involved the use of accounting-based models which incorporate finance ratios in predicting defaults or macroeconomic-based models which link probability of defaults to economic conditions. These methods though useful do not capture the complete complexity of the borrowers profile.

Past approach in Machine Learning:

Machine learning has been utilized for risk evaluation, credit scoring, and predictive analytics in finance, which offers enhanced data processing and decision making capabilities. However, a drawback is that they lack integration with financial insights since they did not take Prof. Stein's ML in Finance course at NYU.



Problem Formulation

A typical pitfall of data scientists who approach these types of problem is that they do not consider business context and blindly apply the machine learning techniques to the problem. In the modeling framework triangle, we focused creating a better formulation, where we applied financial-context to the problem.

Financial Intuition

A thorough variable selection was conducted using the broad financial categories:

- Profitability
- Leverage
- Debt Coverage
- Liquidity
- Growth
- Activity
- Size

This method is more successful than relying too much on the machine to handle variable selection because incorporating the indicators amongst the broad categories allows us to handle missing values more effectively (by replacing with other business variables). Moreover, we can use feature engineering to capture important signals per categories (thereby reducing multicollinearity), and give us better explainability of the result.



Problem Formulation

Evaluation

Further finance domain context was used during the evaluation of the model performance. Rather than the typical train-test split, we applied Walk-Forward Analysis. This approach reduces the possibility of overfitting to the training dataset, and it also gives us a real-world simulation of how financial industries typically behave. We have split the data by statement year, and the walk forward analysis would train the model using current statement year and make predictions using next statement year's data.

After implementing walk-forward analysis, we used calibration to ensure that the default class probabilities are close to what the expected values are. We tested the model calibration within the walk-forward process before implementing it on the entire output class.

* Although walk-forward was implemented on all the models that we tested, calibration was only implemented on the one that provided the best auc during walk-forward.



Data Understanding:

Features engineered relied on a broad set of 6 categories

Profitability

Example: ROA

Firms which consistently make losses will erode equity and become insolvent

Leverage

Example: Liabilities/ Assets

Firms that are highly leveraged are less likely to be able to repay their debts

Debt Coverage

Example: operating cash flow/ interest expense

Firms with higher cash flow relative to debts payments are less likely to default.

Liquidity

Example: Cash/ Total Assets

A firm with poor liquidity is less likely to be able to repay its debts in the short-term

Size

Example: Total Assets

Small firms default more often

Activity

Example: Inventories/ Sales

Different activities have different relationships to default. A large stock of inventories relative to sales increases the default probability.

Some of the biases that need to be considered

Since we are uniformly implementing the data imputation and feature engineering process for the entire dataset, it is important to note that there may be some intrinsic bias within our methodology that would require further study to be mitigated.

- **Industry Specific Bias:**

The ratios may vary significantly between industries and it is important to select the right ratios based upon the company being taken into consideration.

- **Time Specific Bias:**

Some ratios might be affected by the timing, recognizing revenues and comparing ratios between different time periods might lead to inaccurate outcomes.

- **Assumption Bias:**

This bias comes into play when creating formulas for different ratios. Few variables were not readily available while creating financial ratios, which subsequently needed to be approximated.



Data Preparation:

Data Imputation

There were missing values presented in the features of the dataset. These variables were first handled using business knowledge, specifically replacing the variable using other variables via financial interpretation. Even if this replacement may not be an exact replacement, it is still better than using traditional Machine Learning methods.

The following variables were replaced with other variables we were given:

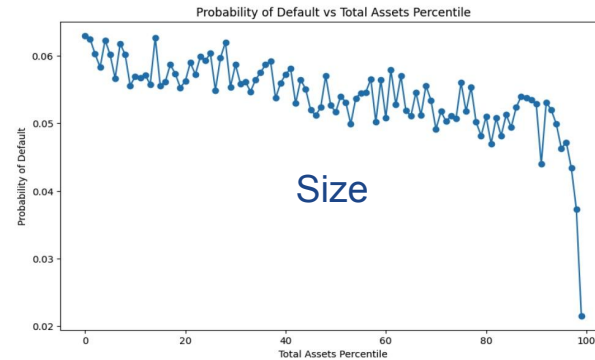
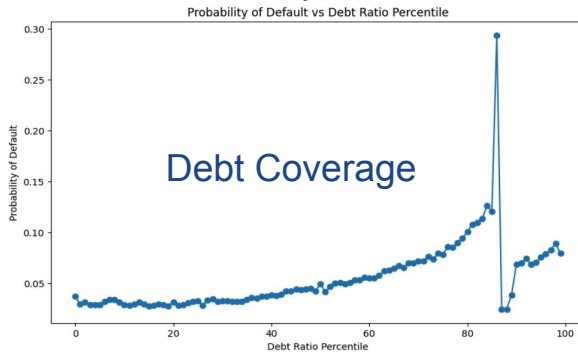
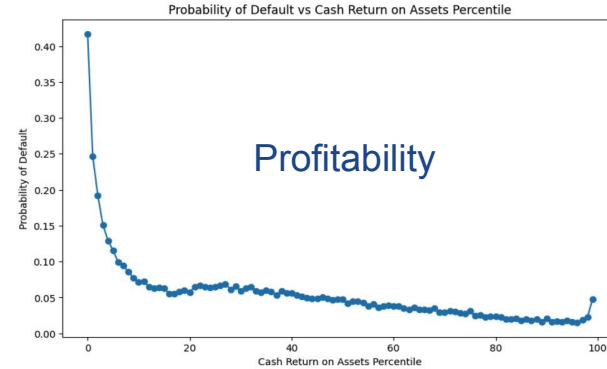
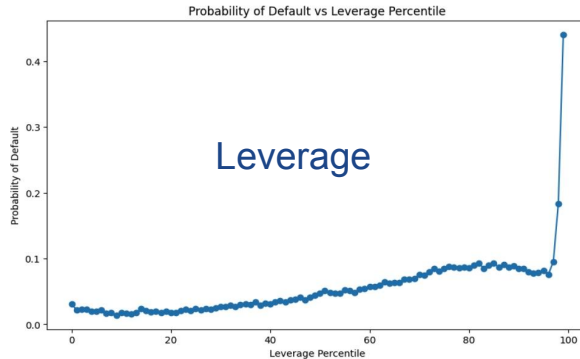
- $\text{Financial Expenses} = \text{Financial Income} - \text{Financial Profit}$
- $\text{Total Equity} = \text{total assets} - (\text{long term liabilities} + \text{short term debt} + \text{long term debt} + \text{short term debt other} + \text{long term debt other} + \text{Accounts payable short term} + \text{Accounts payable short term})$
- $\text{Cash flow operations} = \text{cash and equivalent holdings}$
- $\text{Accounts payable short term} = \text{short term debt}$
- $\text{Current assets} = \text{total assets} - (\text{tangible assets} + \text{intangible assets} + \text{financial assets})$
- $\text{Days receivable} = \text{Accounts receivable} / \text{operating revenue}$
- $\text{Return on equity} = \text{Profit} / \text{total equity}$
- $\text{Return on assets} = \text{Profit} / \text{total assets}$

* For some ratios, a small amount was added to the denominator to avoid division by zero error.

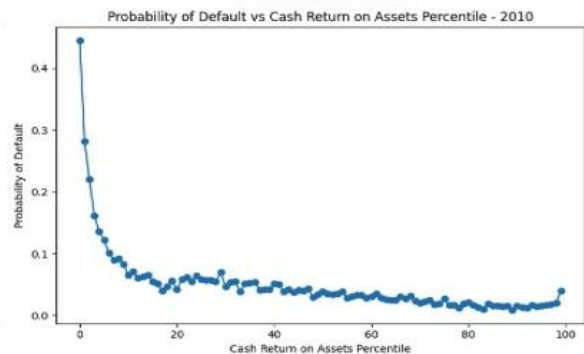
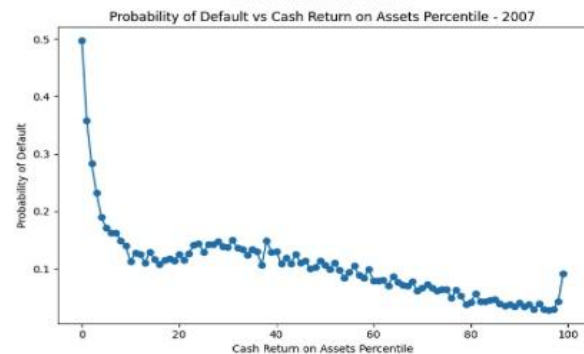
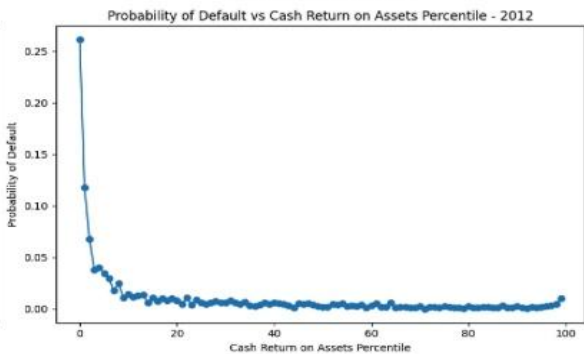
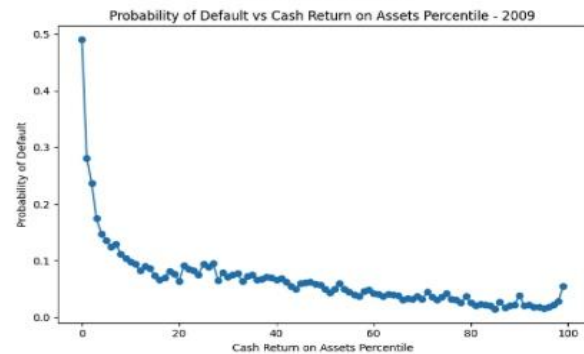
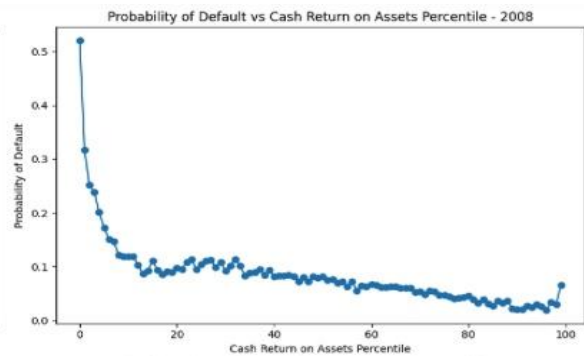
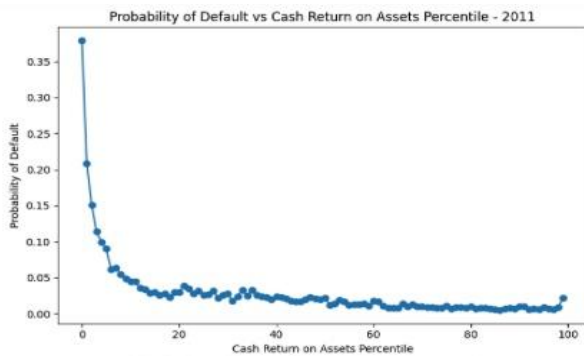
** After using business rule for imputation, the rest of the missing values were imputed using median of the distribution.

Some Data Visualizations:

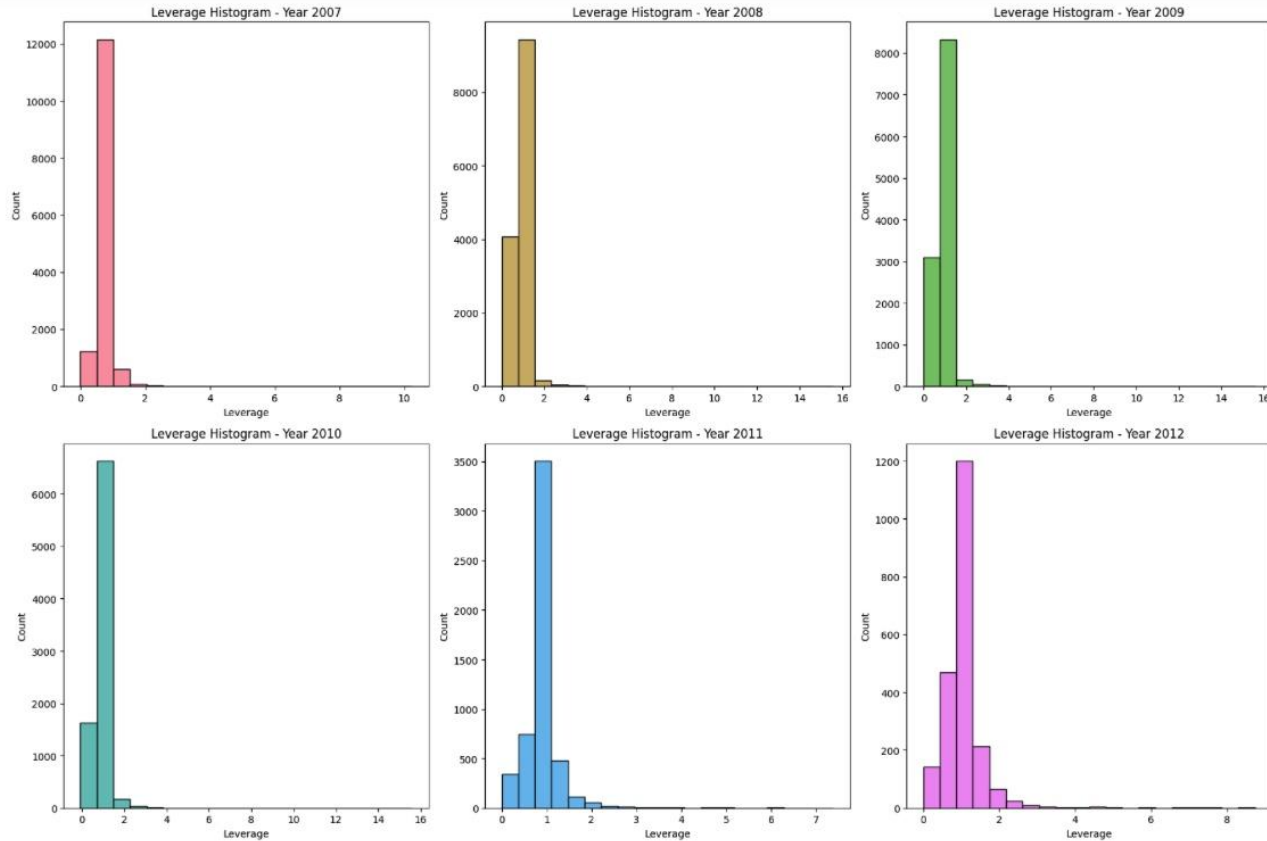
Ratios percentile versus Probability of default



Cash Return on Assets Percentile vs. Probability of Default for each year



Leverage vs default counts for different years



Feature Engineering

Initially selected ratios

Liquidity

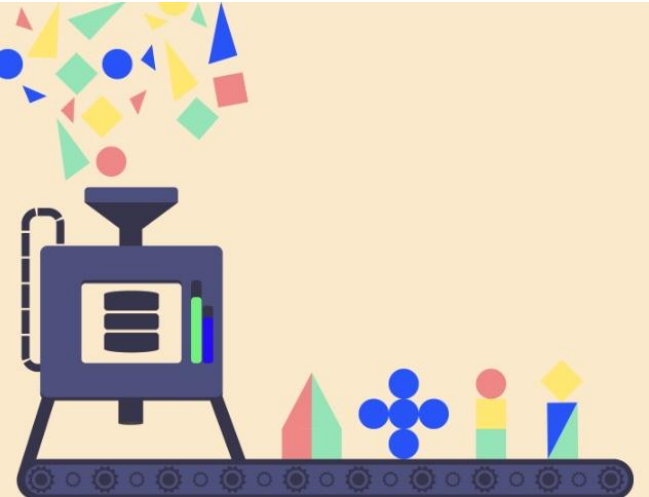
- $\text{Time interest earned} = \text{Earnings before interest, taxes, depreciation, and amortization (ebitda)} / \text{financial expenses}$
- $\text{Current ratio} = \text{Current Assets} / (\text{short term debt} + \text{short term accounts payable})$

Size

- Total assets

Debt Coverage

- $\text{Leverage ratio} = \text{total assets} / \text{total equity}$
- $\text{Debt ratio} = (\text{Total liabilities} + \text{short term debt} + \text{long term debt} + \text{short term debt other} + \text{long term debt other} + \text{short term accounts payable} + \text{long term accounts payable}) / \text{total assets}$



Feature Engineering

Profitability

- $\text{Net Profit Margin} = \text{Net Profit} / \text{operating revenue}$
- $\text{Cash return on assets} = \text{Operating cash flow} / \text{Total assets}$

Activity

- $\text{Asset Turnover} = \text{Operating revenue} / \text{total assets}$
- $\text{Net Working capital} = \text{current assets} / (\text{current assets} - (\text{short term debt} + \text{short term accounts payable} + \text{Long term debt other}))$

Leverage

- $\text{Leverage} = (\text{Long term debt} + \text{Short term debt}) / \text{total assets}$

* We applied data imputation techniques to the new features as well using business rules.

$\text{Debt ratio} = \text{asset total} + \text{equity total}$

** After using business rule for imputation, the rest of the missing values were imputed using median of the distribution.

These medians were calculated on the train dataset and have been hardcoded to be used in the test harness.



Effect of finance-based imputation

	Missing Values (Count)	Percentage (%)
days_rec	740211	72.32
debt_ratio	134145	13.11
nwc	133351	13.03
cash_ratio	120685	11.79
quick_ratio	120676	11.79
current_ratio	120636	11.79
roe	72937	7.13
debt_coverage	406	0.04
time_interest_earned	405	0.04
gross_profit	243	0.02
net_profit_margin	178	0.02
gross_profit_margin	178	0.02
asset_turnover	174	0.02
leverage	156	0.02
cash_return_assets	88	0.01
earning_power	81	0.01
roa	27	0.00
leverage_ratio	1	0.00
asst_tot	0	0.00
default_label	0	0.00



	Missing Values (Count)	Percentage (%)
debt_ratio	134144	13.11
nwc	92085	9.00
days_rec	778	0.08
quick_ratio	318	0.03
debt_coverage	310	0.03
time_interest_earned	309	0.03
gross_profit	190	0.02
gross_profit_margin	178	0.02
net_profit_margin	178	0.02
asset_turnover	174	0.02
leverage	156	0.02
earning_power	81	0.01
roe	21	0.00
cash_ratio	21	0.00
roa	19	0.00
current_ratio	6	0.00
leverage_ratio	1	0.00
asst_tot	0	0.00
cash_return_assets	0	0.00
default_label	0	0.00

Modeling Process

Feature Selection

Selecting the best combination of financial ratios from the 6 broad categories mentioned in the previous slides.



Model Selection

Selection of an intuitive and explainable yet powerful model (starting with logit).



Solution

Explainability of the model variables and their coefficients.



Feature Selection

Univariate Analysis

- **Objective:**
Analysis of individual feature to understand its contribution.
- **Implementation:**
Each financial ratio was assessed independently to evaluate its impact on the model's predictive capability.

Multivariate Analysis

- **Objective:**
Evaluation of the interactions between multiple features.
- **Implementation:**
We examined how combinations of features like ROA and Debt Ratio together influenced the logistic regression model's performance.

Variance Inflation Factor (VIF)

- **Objective:**
Measurement to identify multicollinearity among features.
- **Implementation:**
We used VIF to check for any interdependencies between variables like Total Asset and Leverage, ensuring the independence of features in our model.

P-value Inspection

- **Objective:**
Statistical method to determine the significance of individual features.
- **Implementation:**
For our analysis, features with a low p-value, indicating high significance, were prioritized. For instance, Cash Return on Assets showed a considerably low p-value, underscoring its importance in the model.

*Relevance to Project:

These techniques highlighted the importance of financial knowledge in feature selection. Helped in effectively reducing the feature set from 42 to 6 without compromising the model performance, as reflected in the AUC score.

Final Variables

By using a combination of greedy approach and grid search for selecting features for a logistic regression model we finalised on the below features.

The final selected features were:

- Cash return on assets (profitability)
- ROA (profitability)
- Cash ratio (liquidity)
- Days_rec (activity)
- Debt ratio (debt coverage)
- Leverage (leverage)
- Total Assets (size)

It is interesting to see that out of 42 features, 6 features are sufficient to give a high AUC score. In comparison, using VIF to obtain 20 features from the available features and imputing them using KNN, gave a significantly lower AUC score under the same evaluation criteria.

It is obvious that incorporating financial knowledge is a necessary tool to have while predicting on financial data.



Model Selection

We mainly selected two models:

- Logistic Regression
- Tree-based algorithm (XGBoost)

Logistic regression provide a simple and interpretable result that helps explaining our results to stakeholders. The understanding of coefficients helps us understand how each of the variables are performing. However, logistic regression is sensitive to correlated variables and therefore a careful analysis must be undertaken to avoid multicollinearity.

Given the nature of the model, XGBoost is robust to outliers and performs well in the presence of multicollinearity. While XGBoost or other tree-based models may perform better than GLM, the downside is that it has poor interpretability and explanation.

Quick Summary

Alternatives:

– Apart from XGBoost, onsidered alternatives included, Random Forest and Support Vector Machines (SVM).

– Pros and cons:

Random Forest – Good for handling unbalanced data but can be computationally intensive.

SVM – Effective in high-dimensional spaces but not as interpretable .

– Final Variable Set:

Selected features included Cash Return on Assets, ROA, Cash Ratio, Days_rec, Debt Ratio, Leverage, and Total Assets.

Criteria for selection: Statistical significance (P-value), low multicollinearity (VIF), and economic relevance.

– Economic Intuition Behind the Model:
Focus on variables that directly reflect risk factors and default chances.

The model encapsulates key aspects of a firm's profitability, liquidity, leverage, and operational efficiency.

– Business Problem Solution:
Aim to improve risk assessment and decision making in the financial sector.

The model can enhance predictive accuracy for credit risk, loan approvals, and investment decisions.

Intuition

For each of the variables selected we observe that:

- As Cash return on assets (profitability) increases probability of default decreases (non linear negative correlation).
- As debt ratio (debt coverage) increases probability of default increase (non linear positive correlation).
- Increase in leverage results in increase in probability of default (non linear positive correlation).
- As Total Assets increases probability of default decreases (non linear negative correlation)

All of the coefficients in the next slide are in line with our intuition.



Model Specifications

Optimization terminated successfully.

Current function value: 0.061024

Iterations 10

Logit Regression Results

```
=====
Dep. Variable:    default_label    No. Observations:    1023552
Model:            Logit           Df Residuals:        1023544
Method:           MLE             Df Model:            7
Date:            Fri, 24 Nov 2023   Pseudo R-squ.:       0.08480
Time:            04:36:17          Log-Likelihood:       -62461.
converged:        True             LL-Null:             -68248.
Covariance Type:  nonrobust        LLR p-value:         0.000
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
Intercept      -5.0212      0.022    -224.008      0.000      -5.065      -4.977
asst_tot       -3.17e-09    5.41e-10     -5.859      0.000     -4.23e-09     -2.11e-09
days_rec       4.808e-11    8.29e-11      0.580      0.562     -1.14e-10      2.1e-10
roa             -0.0314      0.001    -49.276      0.000      -0.033      -0.030
debt_ratio      3.824e-09    7.4e-10      5.170      0.000      2.37e-09      5.27e-09
cash_return_assets -1.1200      0.056    -19.916      0.000      -1.230      -1.010
leverage        0.8826      0.026     34.581      0.000        0.833        0.933
cash_ratio      1.391e-08    2.64e-08      0.528      0.598     -3.78e-08      6.56e-08
=====
```

default_label ~ asst_tot + days_rec + roa + debt_ratio + cash_return_assets + leverage + cash_ratio

Dimensions of Interest to the firm

- **Risk Management:** The model provides a quantitative measure of the probability of default, allowing the firm to assess its financial risk exposure.
- **Financial Planning and Decision-Making:** Understanding the financial variables that influence the probability of default can inform financial planning and decision-making processes.
- **Regulatory Compliance:** The model can be used to show compliance with regulatory requirements related to financial stability and risk assessment.
- **Credit Scoring and Lending Decisions:** By incorporating the predicted probability of default into credit scoring systems, the firm can make more informed lending decisions. This can lead to better risk-adjusted returns and a more robust credit portfolio.





Evaluation: Walk-forward Analysis

Our model was tested using a walk-forward approach which simulates a real-time trading environment. The analysis is carried out over 5 years beginning with 2008 as the test year and ending with 2012. For the WFA, all the data before a given test year is treated as the training data. Finally, we concatenate each year's predictions to create a new dataset upon which we run and performance metrics and get the optimal roc-auc. Through this analysis, we attempt to build a more powerful logit model, that is able to adequately discriminate between the two classes, i.e., default vs non-default.

* Using logit as the baseline, we used the six chosen features (with the same preprocessing) to test other models (XGBoost, Random Forest, SVM) with the the walk-forward method. However, since the logit model gave us the best results, we use this for our final analysis.





Evaluation: Interpreting Coefficients

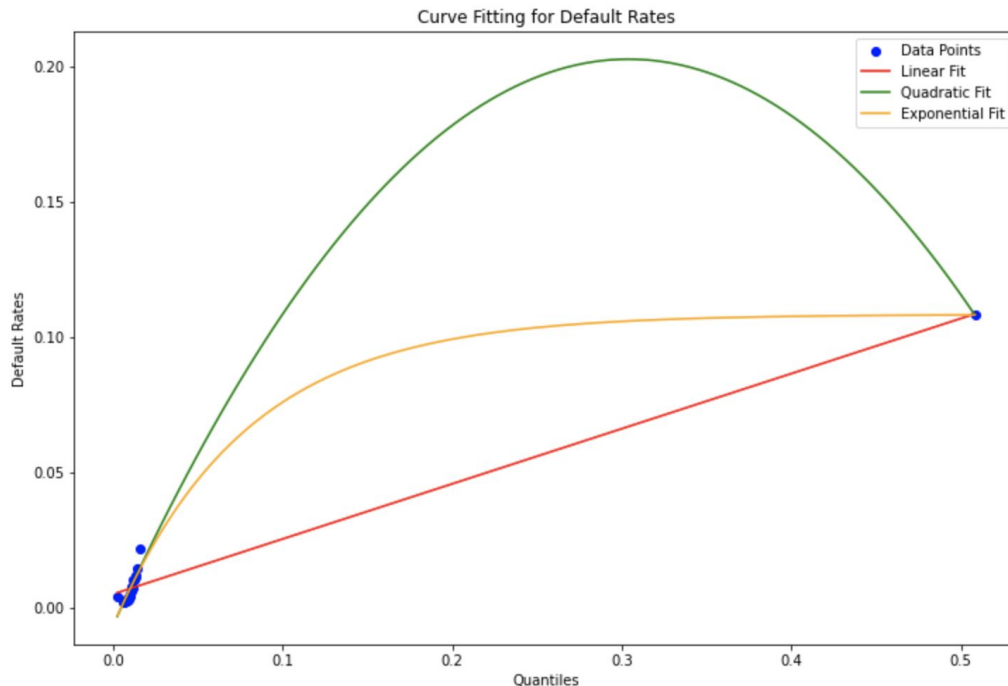
To ensure that the model is performing as we expected, we looked at the sign of the coefficients and verified that they adhere to our intuitive understanding of how each variables should behave.

- Assets and cash amount have inverse relationship with PD
- Debt, leverage, and days_rec (time it takes to pay back the product) have direct relationship with PD

	coef
Intercept	-5.0212
asst_tot	-3.17e-09
days_rec	4.808e-11
roa	-0.0314
debt_ratio	3.824e-09
cash_return_assets	-1.1200
leverage	0.8826
cash_ratio	1.391e-08



Evaluation: Calibration



- The model was calibrated by estimating a non-linear curve that maps quantiles to default rate. Several curves were fitted to the default rates per quantiles. Upon inspecting the curves, the most appropriate curve was the Quadratic Curve.
- We ensured that the coefficients of the curve are retrieved from train data only, and used the corresponding curve to map the predicted probabilities into calibrated probabilities

- Business use case – More reliable risk assessments, crucial for loan and investment decisions. Also, a more reliable model enhances the firm's reputation and customer trust.
- Challenges and alternatives – Alternative approaches like ensemble modeling or advanced machine learning could be considered if calibration does not yield desired results.

Deployment

- Integration with existing systems:
 - Ensure the prediction system's compatibility with the bank's IT infrastructure.
- User training and support:
 - Provide technical support to address any potential issues.
- Data privacy:
 - Adhere to regulation such as GDPR or other data protection laws.
- Bias and fairness:
 - Ensure that the deployment does not lead to unfair treatment or biases against certain customer groups.

Deployment

Privacy



Maintenance

Risk

- Monitoring:
 - Monitor the deployment for accuracy and efficiency.
- Updates:
 - Schedule periodic updates to the models/algorithms to adapt to changing banking trends and customer behaviors.
- Accuracy and Reliability:
 - Implement a backup system in case of failures or inaccuracies.
- Compliance:
 - Conduct regular audits to rectify any issues.
- Security:
 - Regularly update security protocols and conduct awareness training for staff.



Citations

- <https://www.elibrary.imf.org/view/journals/001/2006/149/article-A001-en.xml>
- FSA Note: Summary of Financial Ratio Calculations
- <https://www.bloomberg.com/professional/blog/probabilities-of-default/>
- <https://care-mendoza.nd.edu/assets/152336/dwyer.pdf>
- https://en.wikipedia.org/wiki/Mortgage_underwriting_in_the_United_States
- <https://gdpr-info.eu/>
- Cheatsheets, slide decks and class notes



Appendix A – Work Done

Abhilash Anand

- Target Labeling, Feature Ratio Engineering, Univariate Analysis, Multivariate Analysis, Logistic Regression, Visualizations, Slide Deck

Jin Choi

- Target Labeling, Missing Value Imputation, Feature Ratio Engineering, Univariate Analysis, VIF, Multivariate Analysis, Walk-forward Analysis, Calibration, Slide Deck

Robin Wang

- Target Labeling, Missing Value Imputation, Feature Ratio Engineering, VIF, Univariate Analysis, Slide Deck

Shruti Wagle

- Walk-forward Analysis; target labeling, missing value imputation; XGBoost, RFC, and SVM implementation; predictor harness; final harness & code structure; Slide Deck
- 