Team Periwinkle Project Update

Sunny Yang, Alon Florentin, Xinchen Zhang, Jaylen Peng*

October 27, 2024

1 Introduction

The business problem is that Banca Massiccia needs to more accurately predict the probability that a single prospective borrower will default on a loan (either the principal or the interest payment) within the next 12 months. When a prospective borrower applies for a loan, the borrower will submit their financial data (assumed to be in the form of financial statements). The loan officer will then input the financial data into our model and our model will output a Probability of Default (PD). The loan officer must then determine the interest rates and underwriting fee based on the PD. More accurate PD estimates will allow the bank to better assess the true risk associated with each loan, and set suitable interest rates and underwriting fees. Our approach involves a multi-stage process to calculate the PD for firms, utilizing both unsupervised and supervised models.

2 Pre-processing

2.1 Data Imputation

The first pre-processing step we did was to impute the missing values based on several known balanced column relations from the data. These include:

1. *prof_operations = rev_operating - COGS*
2. *roe = profit / eqty_tot * 100*
3. *wc_net = asst_current - debt_st*
4. *margin_fin = eqty_tot - (asst_fixed_fin + asst_intang_fixed + asst_tang_fixed)*
5. *cf_operations = ebitda + profit + inc_financing + inc_extraord - taxes*

*New York University, Center for Data Science

On the full dataset, the match rate for all of these relations is above 92%, despite the presence of missing values. This is a strong indication that they are correctly implied. Note that this imputation works if there is only one missing value in a given row for a particular relation, otherwise it amounts to do nothing.

2.2 Dropping Unnecessary Columns

After imputing the missing values, we drop four columns, first one is the indexing column because it's another unnecessary ID column, then $eqty\_corp\_family\_tot$ column since it's completely empty, then we drop HQ_city because it is not relevant to our analysis, and finally $days\_rec$ given it has way too many missing values, and some of the values are not making sense (more than 25 percent have negative days receivable).

2.3 Calculating Financial Ratios

After dropping all columns we deemed unnecessary, we calculated the financial ratios for each row (see Table 1). Once we calculated the financial ratios, there was some erroneous values (e.g. *inf*, *-inf*, etc.) so we removed those values by first imputing NA values, then dropping all rows with NA values. Ultimately, this left us with a dataset with approximately 805,000 rows (lost about $\approx 20\%$ of the original dataset).

2.4 Target Variable

We assumed the financial statements for a fiscal year are not available until May of the following year. Therefore, to avoid peeking into the future, we shifted the fiscal year in such a way that the financial statements from one year only apply to predictions (of a one-year PD) starting in June of the following year. For example, the financial statements for a fiscal year ending on "2005-12-31" will be used to predict default from "2006-06-01" to "2007-05-31". Consequently, the binary target variable was defined accordingly, based on whether the def_date falls within the prediction period. Additionally, if the def_date value was NaT, we assumed that was a non-default. The target variable was encoded into a column called *label* (0 for non-default, 1 for default).

3 Modeling

3.1 Overall Model

For our benchmark, we wanted to train a model before we clustered the data. We chose to use a logistic regression model (*smf.logit* from *statsmodels.formula.api*). Using only the specified financial ratios, we first split the dataset into training and validation sets (70-30 split). Then we fit a scaler on the training set (only X_train, we didn't scale y_train since it is a binary variable) and transformed

Metric (Original Name)	**Formula**
Debt to Equity (debt_to_equity)	$\frac{\text{debt\_lt+debt\_st}}{\text{eqty\_tot}}$
Current Ratio (current_ratio)	$\frac{\text{asst\_current}}{\text{debt\_st+AP\_st}}$
Operating Margin (operating_margin)	$\frac{\text{prof\_operations}}{\text{rev\_operating}}$
EBITDA to Assets (ebitda_assets)	$\frac{\text{ebitda}}{\text{asst\_tot}}$
EBITDA Margin (ebitda_margin)	$\frac{\text{ebitda}}{\text{rev\_operating}}$
Return on Assets (ROA) (roa)	$\frac{\text{profit}}{\text{asst\_tot}}$
Return on Equity (ROE) (roe)	$\frac{\text{profit}}{\text{eqty\_tot}}$
NWC Ratio (nwc_ratio)	$\frac{\text{wc\_net}}{\text{asst\_tot}}$
Cashflow to Debt (cashflow_to_debt)	$\frac{\text{cf\_operations}}{\text{debt\_lt+debt\_st}}$
Profit Margin (profit_margin)	$\frac{\text{profit}}{\text{rev\_operating}}$
Days Receivable Turnover (days_receivable_turnover)	$\frac{\text{AR}}{\text{rev\_operating}} \times 365$
Liabilities to Assets (liab_lt_to_assets)	$\frac{\text{liab\_lt}}{\text{asst\_tot}}$
Interest Coverage Ratio (interest_coverage_ratio)	$\frac{\text{prof\_operations}}{\text{exp\_financing}}$
Asset Turnover (asset_turnover)	$\frac{\text{rev\_operating}}{\text{asst\_tot}}$
Equity Ratio (equity_ratio)	$\frac{\text{eqty\_tot}}{\text{asst\_tot}}$

Table 1: Financial Ratios and Their Formulas

both *X_train* and *X_test*. After scaling the data, we fit our logistic regression model and got an AUC score of 0.7113.

3.2 Clustering

We clustered *X_train*, then predicted cluster labels for *X_test*. We used the elbow method to determine the optimal number of clusters, which seemed to be around 4. However, although the elbow method suggests 4, when we tried to cluster the dataset according to k=4, the majority of the dataset was inside only 2 clusters; the other clusters had not enough data to train the models. After examining the distribution to ensure there was sufficient data to run the models, we decided to proceed with only 2 clusters. So we split the dataset into 2 clusters, and for each cluster, we then split them into training and validation sets.

3.3 Supervised Models

For each cluster, we then fit several supervised models: logistic regression and decision trees. We attempted to fit an XGBoost model, but ultimately decided

that we did not have enough time to properly tune the hyperparameters so that we could accurately represent the performance of the XGBoost model. We found the AUC score for each model, and the results are listed in Table 2. However, we note that it does not differ much from the baseline AUC of 0.7113, as reported in Section 3.1.

	Cluster 0	Cluster 1
Logistic	0.724	0.710
Decision Tree	0.512	0.514

Table 2: AUC Results by Cluster and Model

4 Hypotheses & Issues

4.1 Hypotheses

4.1.1 Expected Coefficients

Our current models have output some unexpected coefficients, and that the coefficients are mostly statistically insignificant; for example, if we refer to Figure 3, we would expect that *roe* would have a statistically significant effect on the PD. However, we observed that *roe* did not have a statistically significant impact on the PD. Our hypothesis is that there must be some ratios that we missed that would be better at predicting PD.

4.1.2 Clustering Method

Our current clustering method is based on financial ratios, and we have achieved middling results. Our hypothesis is that we may achieve better performance with our models if we group based on size or industry, rather than using financial ratios and size together.

4.2 Issues

4.2.1 "Imbalanced" Balance Sheet

So far, despite some successes in re-discovering the column relations (as mentioned in Section 2), we did not manage to recover arguably the most fundamental property a balance sheet has to satisfy, that is *Total Assets = Total Equity + Total Liabilities*. This is concerning coupled with the fact that we also could not decompose the *Total Assets* and *Total Equity* correctly from other related columns in the dataset. If we were to accept this truth, in the worst case we will only be able to recover partial data, which has a detrimental (cascading) effect on all the downstream calculations such as financial ratios and modeling.

4.2.2 Missing Data

Our current assumption is that if the company has provided a value for the row, then we will use that data even if our calculated value is different. However, so far, we have been mostly dropping missing data (instead of proactively imputing them) with the rationale that on one hand, it is very hard to back up most of the columns given the unclear underlying relations. For example, if the short-term account payable AP_st is missing, in our current understanding it is difficult to even guesstimate a figure for it. On the other hand, we think imputing with either the population mean/median can also be dangerous because of the ecological fallacy and it can only introduce unnecessary biases into the models without any obvious benefits. One potential solution to this is to define a more fine-grained sub-population and then impute the missing entries with their corresponding (group) average statistics. In our vision, such sub-population can be achieved by clustering based on the company size and/or its industry sector, however, whether it is tangent to the problem remains unsolved. The issue will recur in the real-world test set when we must predict even when the missing data presents.

4.2.3 Imbalanced data

So far, we have discovered a striking class imbalance in the labels produced in the way mentioned in Section 2.4. In particular, the proportion for default versus non-default companies is close to 0.008 vs. 0.992. We did not yet fully recognize what significant impact this class imbalance could have on our modeling choice. Although we know there are already established literature on how to deal with class imbalance such as under/over-sampling, or adjusted training-sample weights, we did not have a clear understanding on how these might influence the model interpretability (which is the topmost priority that our group has decided to put forward) and other unexpected consequences. Notably, the trust-worthiness of column weights produced by such model remains unclear.

4.2.4 Performance from Decision Tree Model and XGBoost Model

So far, the performance from our Decision Tree models and XGBoost models have been extremely subpar, much worse than we expected; there may be an issue with how the models are being trained. More specifically, it could be an issue with the data we are using to train it, or the hyperparameters for each model. This will require more investigation to fix.

5 Appendix

5.1 Logit Regression Results

Logistic Regression - Overall Model:

Logit Regression Results

Dep. Variable:	label	No. Observations:	563744
Model:	Logit	Df Residuals:	563731
Method:	MLE	Df Model:	12
Date:	Mon, 28 Oct 2024	Pseudo R-squ.:	0.03689
Time:	01:00:40	Log-Likelihood:	-26186.
converged:	True	LL-Null:	-27189.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.2339	0.025	-167.631	0.000	-4.283	-4.184
roa	-2.7568	0.102	-27.094	0.000	-2.956	-2.557
roe	-0.0002	0.000	-1.373	0.170	-0.000	8.04e-05
debt_to_equity	-4.446e-06	1.65e-05	-0.270	0.788	-3.68e-05	2.79e-05
current_ratio	-1.119e-05	4.07e-05	-0.275	0.784	-9.1e-05	6.87e-05
ebitda_margin	1.879e-07	1.17e-06	0.160	0.873	-2.11e-06	2.49e-06
nwc_ratio	-0.0469	0.046	-1.030	0.303	-0.136	0.042
cashflow_to_debt	0.0008	0.001	0.883	0.378	-0.001	0.003
days_receivable_turnover	7.588e-11	8.11e-11	0.936	0.349	-8.3e-11	2.35e-10
liab_lt_to_assets	1.2581	0.285	4.420	0.000	0.700	1.816
interest_coverage_ratio	3.063e-08	2.39e-07	0.128	0.898	-4.39e-07	5e-07
asset_turnover	-0.0254	0.017	-1.537	0.124	-0.058	0.007
equity_ratio	-2.5251	0.085	-29.853	0.000	-2.691	-2.359

Figure 1: Overall

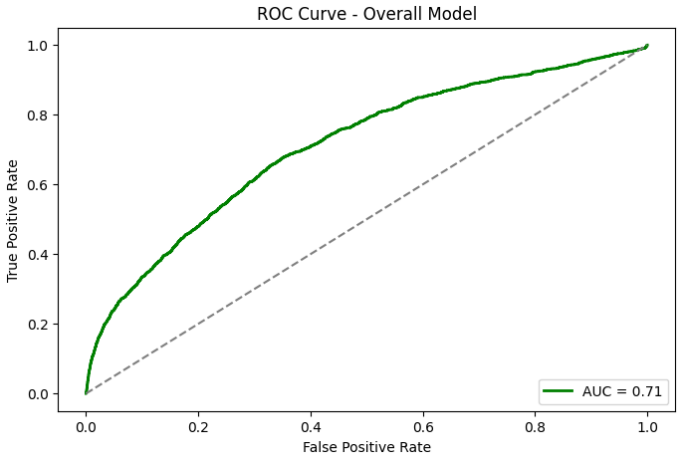

Figure 2: Overall AUC

Logistic Regression - Cluster: 0

Logit Regression Results

Dep. Variable:	label	No. Observations:	177059
Model:	Logit	Df Residuals:	177046
Method:	MLE	Df Model:	12
Date:	Mon, 28 Oct 2024	Pseudo R-squ.:	0.04011
Time:	01:20:37	Log-Likelihood:	-8467.3
converged:	True	LL-Null:	-8821.2
Covariance Type:	nonrobust	LLR p-value:	1.001e-143

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.1926	0.044	-95.744	0.000	-4.278	-4.107
roa	-2.8464	0.174	-16.315	0.000	-3.188	-2.504
roe	-8.496e-05	0.000	-0.418	0.676	-0.000	0.000
debt_to_equity	-3.854e-06	2.36e-05	-0.163	0.870	-5.01e-05	4.24e-05
current_ratio	-1.032e-06	1.09e-05	-0.094	0.925	-2.24e-05	2.04e-05
ebitda_margin	7.661e-08	2.14e-06	0.036	0.971	-4.12e-06	4.27e-06
nwc_ratio	-0.0794	0.080	-0.989	0.323	-0.237	0.078
cashflow_to_debt	-0.0006	0.005	-0.121	0.904	-0.011	0.010
days_receivable_turnover	-2.032e-09	2.69e-09	-0.756	0.450	-7.3e-09	3.24e-09
liab_lt_to_assets	1.4258	0.493	2.889	0.004	0.459	2.393
interest_coverage_ratio	-2.722e-07	3.52e-07	-0.773	0.439	-9.62e-07	4.18e-07
asset_turnover	-0.0069	0.028	-0.249	0.803	-0.061	0.047
equity_ratio	-2.6200	0.151	-17.385	0.000	-2.915	-2.325

Figure 3: Cluster 0

Logistic Regression - Cluster: 1

Logit Regression Results

Dep. Variable:	label	No. Observations:	386684
Model:	Logit	Df Residuals:	386671
Method:	MLE	Df Model:	12
Date:	Mon, 28 Oct 2024	Pseudo R-squ.:	0.03372
Time:	01:20:53	Log-Likelihood:	-17955.
converged:	True	LL-Null:	-18582.
Covariance Type:	nonrobust	LLR p-value:	6.716e-261

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.2482	0.030	-140.044	0.000	-4.308	-4.189
roa	-2.4993	0.132	-19.000	0.000	-2.757	-2.242
roe	-0.0012	0.000	-2.559	0.010	-0.002	-0.000
debt_to_equity	-3.161e-06	1.05e-05	-0.302	0.763	-2.37e-05	1.74e-05
current_ratio	-1.294e-07	3.58e-06	-0.036	0.971	-7.15e-06	6.89e-06
ebitda_margin	8.806e-08	1.41e-06	0.062	0.950	-2.68e-06	2.85e-06
nwc_ratio	-0.1270	0.055	-2.291	0.022	-0.236	-0.018
cashflow_to_debt	0.0007	0.001	0.689	0.491	-0.001	0.003
days_receivable_turnover	-2.83e-11	2.6e-10	-0.109	0.913	-5.38e-10	4.82e-10
liab_lt_to_assets	1.4413	0.340	4.244	0.000	0.776	2.107
interest_coverage_ratio	6.874e-08	1.96e-07	0.351	0.725	-3.15e-07	4.52e-07
asset_turnover	-0.0308	0.020	-1.557	0.119	-0.070	0.008
equity_ratio	-2.3843	0.101	-23.716	0.000	-2.581	-2.187

Figure 4: Cluster 1