Accounting Manipulation in Banks

An Exploratory Project

Table of contents

Before we begin…

A summary of our raw data exploration

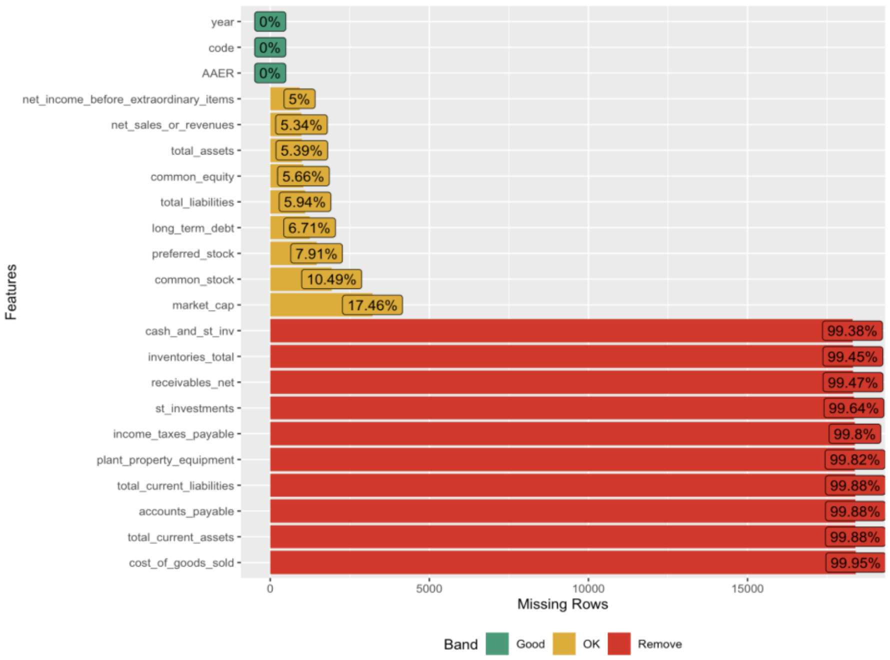
	df	df_aaer
Filtering	General Industry Classification == 4 (for banks) Year!=2023	
Results	❑ **46k** observations ❑ **3053** unique companies	❑ **18941** observations ❑ **1262** unique companies
Findings	❑ **125 -> 39 countries** after merging (AAERs released by SEC - based in USA) ❑ Accounting line figures in **different currencies** (raises comparability issues) ❑ **% of AAER is very low each year**, which makes sense since frauds are far and few between and companies which commit fraud tend to keep it well hidden. ❑ **% of AAER decreases over the years**, suggesting that there could be stricter regulations over time or companies got better at hiding.	

01 Research: Academic Papers

Raw Financial Data Items:

- **Foundational elements** of the accounting framework
- Converting raw data into financial ratios based on incomplete behavioural theories could result in **loss of useful predictive info**
- Fraud prediction models based on raw data can take on **more flexible & complex functional forms**

Journal of Accounting Research: Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach
(Bao et al., 2015)

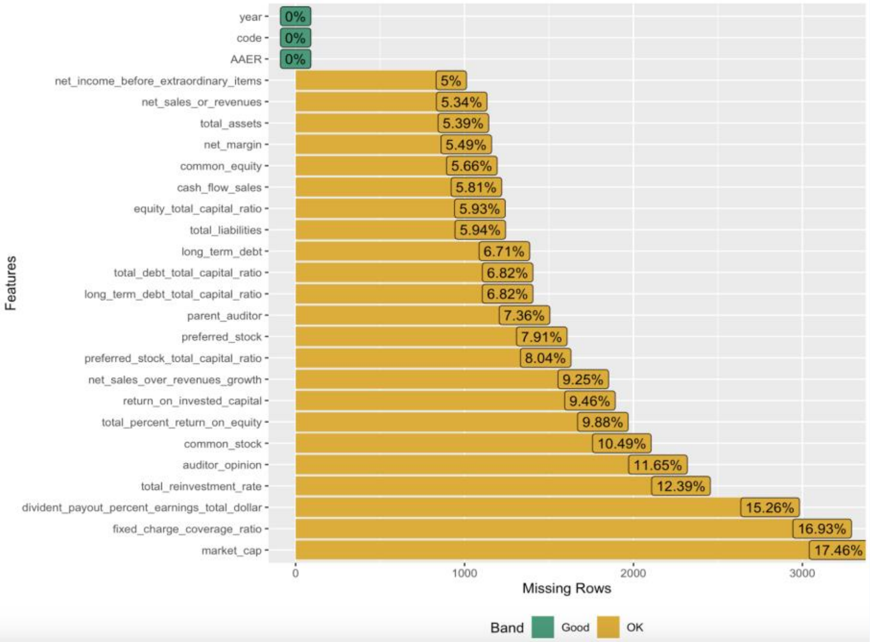


01 Research: Academic Papers

Financial Ratios:

- ❑ Fraud prediction models based on **financial ratios are powerful** because they are identified by human experts that **offer sharp prediction** on when corporate managers have incentives to engage in fraud.

- ❑ Fraud prediction models based on raw financial data may be less powerful as they are not directly linked to theory



Predicting Material Accounting Misstatements
(Dechow et al., 2011)

01 Research: Domain Knowledge

Financial Ratios:

- ❑ Due to data limitations, we could not adopt the ratios recommended in the research papers.
- ❑ Leveraged on our domain knowledge to derive 12 ratios, provided by Worldscope database.

1. Equity % Tot Capital	Banks manipulate ratio via income smoothing, moving liabilities off the B/S or overvaluing assets to inflate equity for a higher ratio.
2. Preferred Stock % Tot Capital	Banks manipulate ratio by selectively issuing/redeeming preferred stock strategically for a higher Tier 1 capital ratio.
3. LT Debt % Tot Capital	Banks manipulate ratio by issuing/retiring LT debt or refinancing existing debt for the desired ratio that matches their capital management objectives.
4. Tot Debt % Tot Capital	Similar to point 3, banks manipulate for a lower ratio.
5. Return on Equity	Banks manipulate ratio by adjusting reserves or recognising certain gains or losses to inflate net income and deflate equity for a higher ROE ratio.
6. Return on Invested Capital	Banks manipulate ratio by adjusting asset valuations or liabilities to inflate net income for a higher ratio.

01 Research: Domain Knowledge

Financial Ratios:

- ❑ Due to data limitations, we could not adopt the ratios recommended in the research papers.
- ❑ Leveraged on our domain knowledge to derive 12 ratios, provided by Worldscope database.

7. Net Margin	Banks manipulate this ratio by selectively recognising revenues or expenses to inflate profitability ratio.
8. Fixed Charge Coverage Ratio	Banks manipulate ratio by adjusting fixed charges or operating CF through timing of payment or others to inflate ratio.
9. Dividend Payout (% Earnings)	Banks manipulate ratio by adjusting earnings figures/dividend payouts to inflate ratio with the goal of influencing investors' perception of co's financial health.
10. Cash Flow/Sales	Banks manipulate ratio via premature revenue recognition, fictitious sales or understate expenses to inflate sales and cash flow, respectively.
11. Reinvestment Rate	Banks manipulate ratio by selecting where to allocate capital to inflate/deflate investment figures for a more favorable picture of their growth prospects.
12. Net Sales/Revenue Growth	Banks manipulate ratio via fictious sales, premature revenue recognition, round-trip transactions to inflate ratio.

01 Research: Topic Modelling

- ❑ AAERS from 2004 – 2023
- ❑ Minimum 3 pages
- ❑ 608 documents

- ❑ Coherence metric

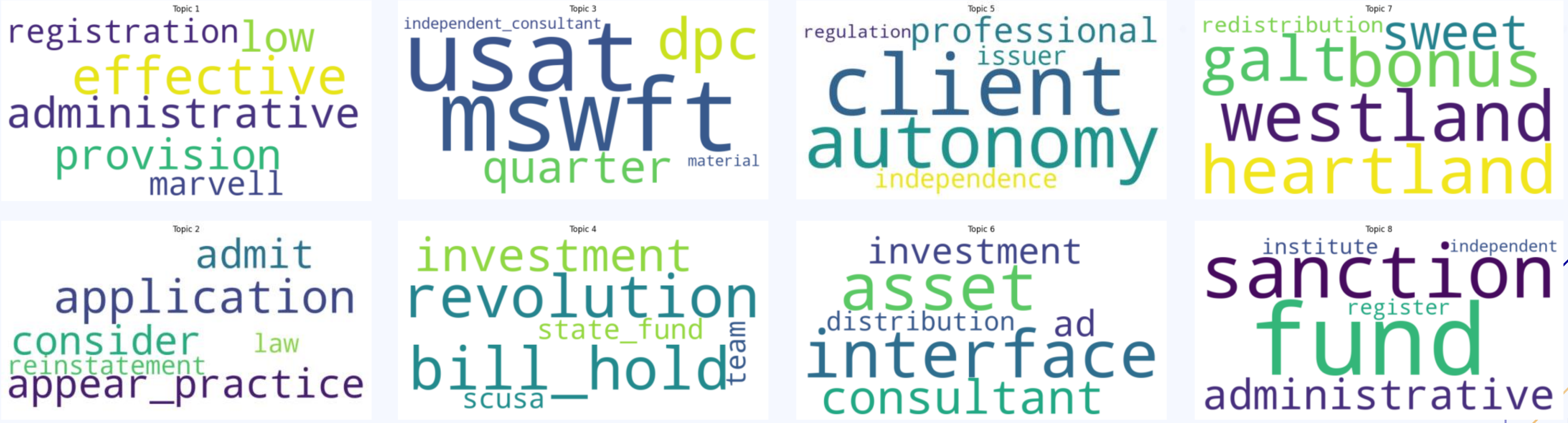
Min Threshold	Max Threshold	Number of Topics	Passes	Perplexity	Coherence Score	Silhouette Score
2	600	5	15	-8.05069418	0.506488152	0.756528009
2	600	5	20	-8.01339457	0.493743787	0.708322436
2	600	6	15	-8.02299412	0.432268016	0.710764141
2	600	6	20	-8.02166974	0.398219825	0.715155299
2	600	7	15	-7.97652222	0.472714616	0.757562658
2	600	7	20	-7.96724046	0.458064039	0.749114573
2	800	5	15	-7.9525447	0.466285961	0.734449138
2	800	5	20	-7.96669513	0.421432737	0.747798329
2	800	6	15	-7.91998034	0.523022708	0.677892421
2	800	6	20	-7.92203999	0.485967425	0.661778811
2	800	7	15	-7.91783414	0.443765253	0.705670671
2	800	7	20	-7.89470623	0.479950767	0.665280453
2	1000	5	15	-7.88375062	0.364446056	0.770238191
2	1000	5	20	-7.85195421	0.476270773	0.686376159
2	1000	6	15	-7.84818818	0.392541961	0.745006059
2	1000	6	20	-7.82514187	0.459108175	0.729564494
2	1000	7	15	-7.84029401	0.369669138	0.729264712
2	1000	7	20	-7.81793031	0.377287669	0.665743483

```python
auditors = ['kpmg', 'pwc', 'deloitte', 'ey', 'bdo', 'berkower']

companies = ['pascale', 'moduslink', 'comscore', 'iconix', 'weatherford',
             'westland', 'magnachip', 'vmware', 'newell', 'microtune',
             'cambrex', 'netease', 'mcafee', 'wagework', 'akorn',
             'oppenheimer', 'qualcomm', 'broadwind', 'valeant', 'marcum',
             'kcap', 'norvatis', 'tidewater', 'huron', 'soyo',
             'apple', 'uhp', 'pareteum', 'voxeljet', 'dxc',
             'bruker', 'wowjoint', 'gtt', 'wex', 'psi',
             'novartis', 'compass', 'galena', 'ppg', 'gentex',
             'philidor', 'mswft', 'usat', 'galt']

banks = ['jp', 'morgan', 'citigroup', 'sbb', 'southwestern',
         'scusa', 'heartland']

pharma = ['musclepharm', 'herbalife', 'alexion']

names = ['maxwell', 'grace', 'stonemor', 'mcneeley', 'stryker',
         'bednar', 'boyle', 'connor', 'culpepper', 'crowe',
         'koeppel', 'pattison', 'winemaster', 'bertuglia', 'matta',
         'davy', 'doody']

countries = ['gibraltar']

unsure_terms = ['company', 'pcaob', 'crs', 'osg', 'official', 'government']

custom_stopwords = set(['commission', 'accountant', 'board', 'auditor', 'audit',
                        'release', 'section', 'pursuant', 'shall', 'would',
                        'make', 'one', 'rule', 'i', 'ii',
                        'iii', 'iv', 'v',  'au', 'see',
                        'whether', 'cpa', 'respondent', 'order', 'include',
                        'proceeding', 'approximately', 'appear', 'describe', 'make',
                        'also', 'include', 'audits', 'however', 'armour',
                        'alc', 'become', 'thereunder', 'wilfully', 'come',
                        'show', 'need', 'take', 'willfully', 'serve',
                        'involve', 'hereby'])

#Updating custom stopwords
custom_stopwords.update(auditors)
custom_stopwords.update(companies)
custom_stopwords.update(banks)
custom_stopwords.update(pharma)
custom_stopwords.update(names)
custom_stopwords.update(countries)
custom_stopwords.update(unsure_terms)


removal_texts = ['securities and exchange commission', 'united states of america', 'securities
                 'accounting and auditing enforcement', 'file no.', '"commission"',
                 'united states', 'security act']
```

01 Research: Topic Modelling - General

LDA hyperlink

01 Research: Topic Modelling - Bank

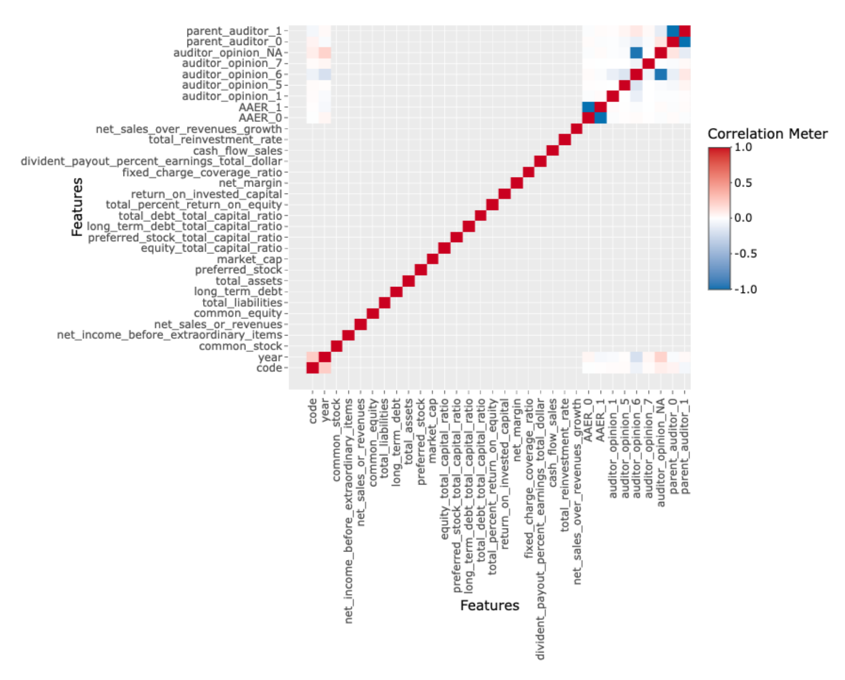


LDA hyperlink

01 Final Variables and their Correlation

Variables:

- ❑ No significant correlation between any 2 variables
- ❑ Keep all 22 variables



01 Finalised Variables

	Quantitative Variables		Qualitative Variables
	Raw Accounting Items	**Financial Ratios**	**Topic Modelling**
X-Variables	8 variables: ❑ Total Assets ❑ LT Debt ❑ Total Liabilities ❑ Common Equity ❑ Preferred Stock ❑ Net Sales or Revenue ❑ Common Stock ❑ Net Income before Extraordinary Items	12 variables: ❑ Equity % Tot Capital ❑ Preferred Stock % Tot Cap ❑ LT Debt % Tot Capital ❑ Total Debt % Tot Capital ❑ ROE ❑ ROIC ❑ Net Margin ❑ Fixed Charge Coverage ❑ Dividend Payout (%) ❑ Cash Flow/Sales ❑ Reinvestment Rate ❑ Net Sales/Rev Growth	2 variables: ❑ Auditors ❑ Auditors' Opinion
Y-Variable	❑ AAER: binary. ❑ 1 indicates potential accounting manipulation ❑ 0 indicates no accounting manipulation		

02 What did we do with the NAs?

01

Backfill

- ❑ Group by company code
- ❑ Backfill based on last recorded value
  - ➢ Used last recorded instead of mean because we **do not want to distort trends**

02

Impute Market Capitalization

- ❑ For variables that cannot be imputed by backfilling
  - ➢ NA data in every record
- ❑ Group by market cap and take the mean
  - ➢ Imputed market cap values will follow size of the company

03

auditor_opinion NAs

- ❑ Categorical
  - ➢ Cannot backfill or impute mean
- ❑ Conservative Approach:
  - ➢ Impute with 1 (“not audited”)

02 Check for Skewness

```r
log_and_drop_skewed_numeric_columns <- function(df, threshold, small_number) {
    for (col in names(df)) {
        if (is.numeric(df[[col]])) {
            skew <- skewness(df[[col]])
            if (abs(skew) > threshold) {
                log_col_name <- paste0("log_",col)
                # Log-transform positive values with the addition of a small number
                df[[log_col_name]] <- ifelse(df[[col]] > 0, log(df[[col]] + small_number), df[[col]])
                df[[col]] <- NULL
            }
        }
    }
    return(df)
}

# Set the skewness threshold (absolute value)
skewness_threshold = 10
small_number = 0.01
# Call the function and get the list of skewed columns
df_aaer3 = log_and_drop_skewed_numeric_columns(df_aaer2, skewness_threshold, small_number)
```

- Created a function
    1. For each numeric column, R calculates the skewness using the skewness function. If the absolute value of the skewness is greater than the provided threshold, it indicates that the column is skewed.
    2. Logs skewed values
    3. Appends "log_" to the original column name. Then, it drops the original column
- Applied this function to df_aaer2 → Stored results in df_aaer3

02 Improving Comparability

- ❑ Different companies release accounting figures in different currencies
- ❑ Affects the comparability of figures between rows
- ❑ Captured in ITEM6099, "Currency Of Document"

To fix this problem, we came up with 2 more models:

1. **Percentage Change Model**, where each record is a percentage change relative to the previous year

- ❑ Potential flaw: Percent change model would not be able to differentiate between large companies and small companies, as the absolute values/magnitude of values are not captured.

2. **Currency Adjusted Model**, where each record has been transformed from its respective foreign currency into USD by applying an exchange rate

03 Final 4 Model Loadouts

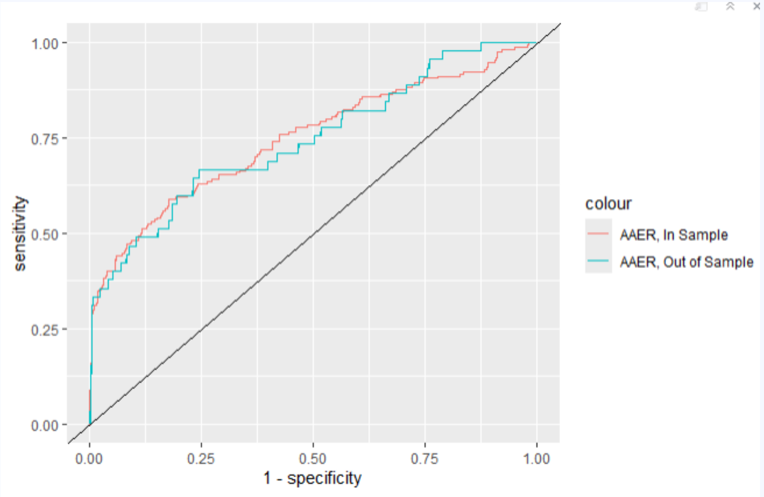
Model	Raw Model [df_aaer2]	Log-Transformed [df_aaer3]	Log-Transformed Percentage Change [df_aaer4]	Log-Transformed Currency Adjusted [df_aaer5]
Variables	22 variables: ❑ 8 raw financial data ❑ 12 financial ratios ❑ Auditors ❑ Auditors’ Opinion	22 variables: ❑ Logged all numeric variables which are skewed	22 variables: ❑ % change of 8 raw financial data ❑ 12 financial ratios ❑ Auditors ❑ Auditors’ Opinion	22 variables: ❑ 8 currency-adjusted financial data ❑ 12 financial ratios ❑ Auditors ❑ Auditors’ Opinion

We will be focusing on only 1 model based on AUC….

03 Comparison of AUCs

	XGBoost		lambda.min		lambda.1se	
	In	Out	In	Out	In	Out
Raw model	0.9756983	0.6607232	0.6974038	0.6733187	0.6653501	0.6987844
Log Model	0.9757430	0.8446033	0.7155628	0.6839778	0.6888045	0.6995785
% Change Log Model	0.9404513	0.7922464	0.6858110	0.7047249	0.6549695	0.6966743
Currency Adjusted Log Model	0.9694841	0.8264798	0.7038530	0.6973123	0.6834450	0.6864822

03 Log Model: Logistic Regression



In-sample AUC	Out-of-sample AUC
0.7475618	0.7451041

Significantvariables

- ❑ total liabilities
- ❑ long-term debt
- ❑ total assets
- ❑ long term debt to total capital ratio
- ❑ total dollar percent earnings dividend payout
- ❑ log preferred stock
- ❑ log preferred stock to total capital ratio
- ❑ log total debt to total capital ratio
- ❑ log net sales over revenues growth

03 Log Model: LASSO

Why LASSO?

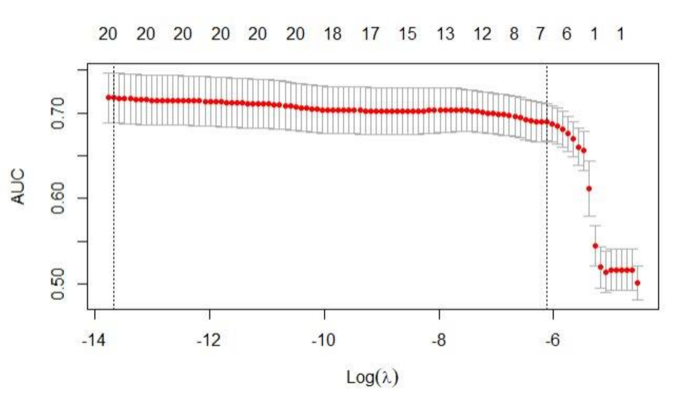
- ❑ Helps with **variable selection**
    - ➢ While we have done research into which variables could possibly help to predict misstatements, we don't have good judgement as to which variables would be more effective/ineffective.
    - ➢ 2 types of models: lambda.min and lambda.1se

03 Log Model: LASSO

Difference between lambda.1se and lambda.min:

Lambda.min imposes a lower penalty, hence retaining more variables, to give the best performing model which maximises AUC.

Lambda.1se trades model performance for explainability, having a higher penalty to retain less variables, to create a simpler model that still performs well.



Analysis:

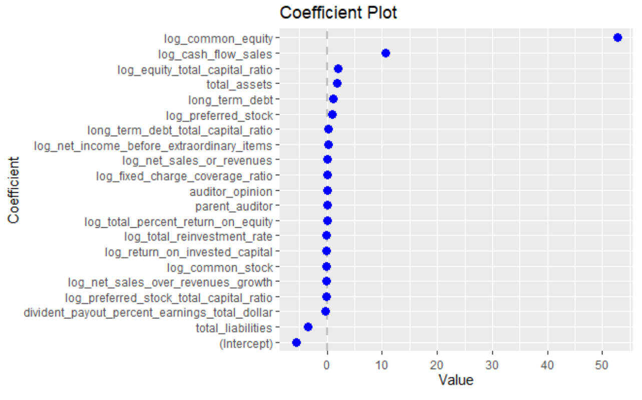
lambda.min (1.173294e-06) -> <u>best performance</u>

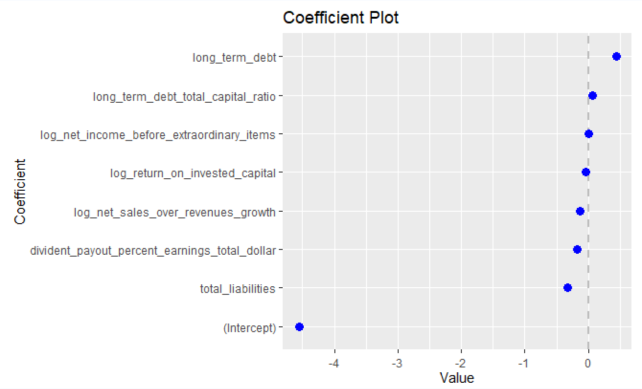
- ❑ AUC peaks at approximately 20-21 variables

lambda.1se (0.002198539) -> <u>simplest model</u> within 1 standard error of lambda.min

- ❑ AUC peaks at approximately 6-7 variables

03 Log Model: LASSO





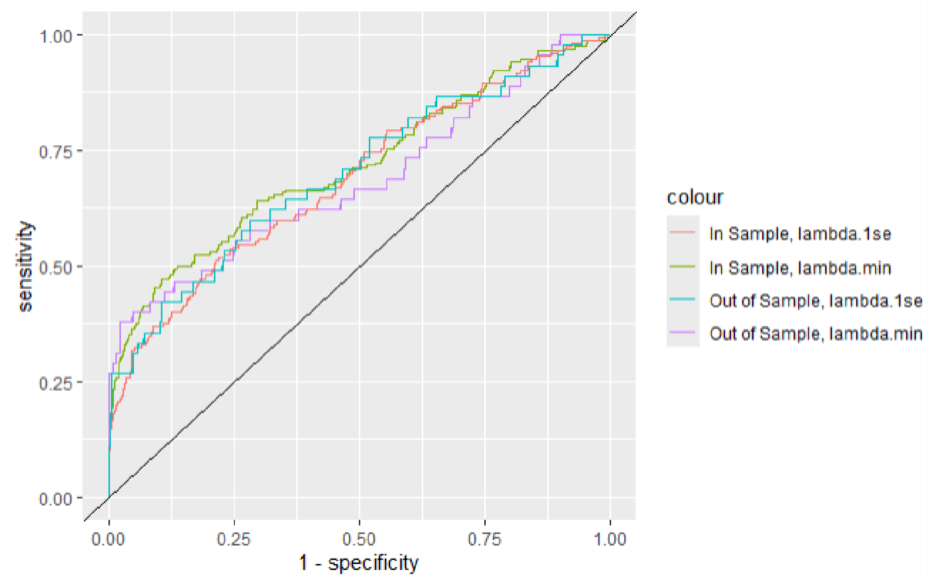
lambda.min

- ❑ Most variables are close to zero with log_common_equity being abnormally large and positive.
- ❑ This may indicate the model is too dependent on this variable
  - ➢ Could become inaccurate with the addition of new data

lambda.1se

- ❑ Most variables are closer to each other in absolute value
  - ➢ Could indicate a better fit for the data

03 Log Model: LASSO



lambda.min		lambda.1se	
In-sample	**Out-of-sample**	**In-sample**	**Out-of-sample**
0.7155628	0.6839778	0.6888045	0.6995785

Analysis:

lambda.min

AUC: in-sample > out-sample

- **Best performing model** at the expense of potentially selecting a more complex model

lambda.1se

- AUC: out-sample > in-sample
- A simpler model with reduced predictive performance but **better at taking in new data**
- More explainable than lambda.min as well, due to having less variables.

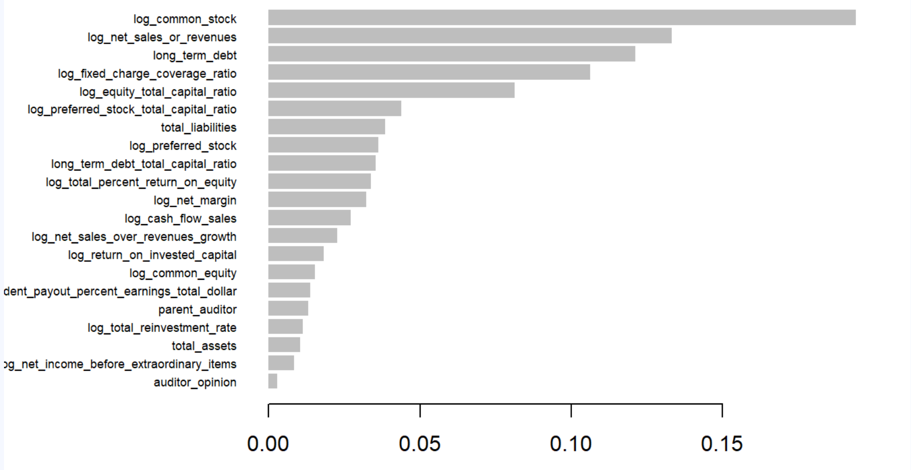
Conclusion

- lambda.1se superior to lambda.min due to better explainability & out-of-sample performance
- lambda.1se out-sample AUC > lambda.min out-sample AUC

03 Log Model: XGBoost

Why XGBoost?

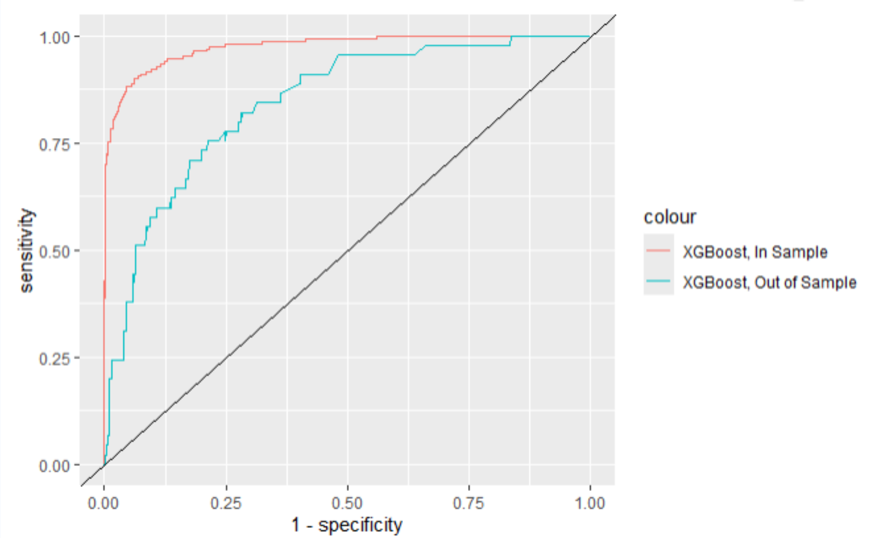
- ❑ Improves performance by running iterative models
  - ➢ Each successive model run improves on previous model



Variable Importance of XGBoost

- ❑ Log Common Stock
- ❑ Log Net Sales or Revenues
- ❑ Long Term Debt
- ❑ Log Fixed Charge Coverage Ratio
- ❑ Log Equity Total Capital Ratio
- ❑ Log Preferred Stock Total Capital Ratio
- ❑ Total Liabilities
- ❑ Log Preferred Stock
- ❑ Long Term Debt Total Capital Ratio
- ❑ Log Total Percent Return on Equity

03 Log Model: XGBoost



XGBoost	
In-sample	**Out-of-sample**
0.9757430	0.8446033

Analysis:

- ❑ AUC: In-sample > Out-sample
- ❑ In-sample AUC significantly high
- ❑ Out-of-sample AUC of 0.845, which shows that model performs well even on unseen data
- ❑ However, difference between in and out sample AUC is relatively large
  - ➢ Could possibly be an indicator of slight overfitting

04 Comparison of Models

	Raw Model		XGBoost		lambda.min		lambda.1se	
	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample	In-sample	Out-of-sample
Log Model	0.7475618	0.7451041	0.9757430	0.8446033	0.7155628	0.6839778	0.6888045	0.6995785

The best model is the **XGBoost** model, as it has the highest AUC scores.

This makes sense as XGBoost is able to iteratively run better models which builds upon previous models.

However, the model might be overfitted due to an extremely high In-sample AUC score (0.976),

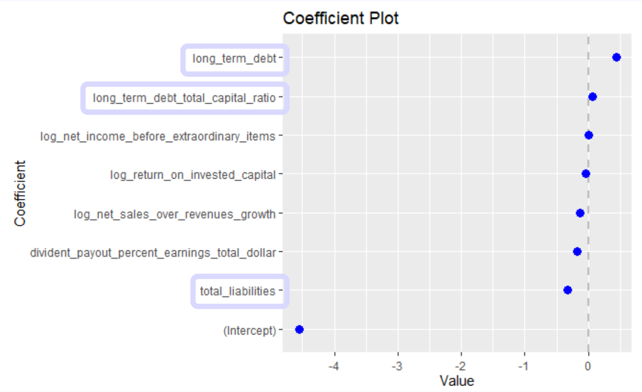
But still performs very well on unseen data (AUC of 0.845).

04 Comparison of Models

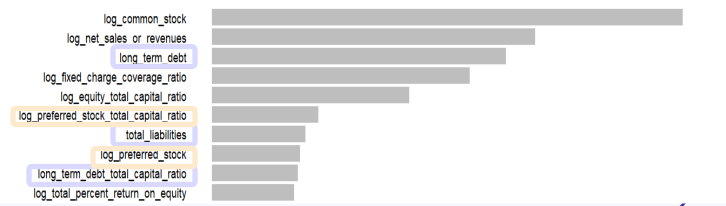
Logistic regression significant variables

- ❑ total liabilities
- ❑ long-term debt
- ❑ total assets
- ❑ long term debt to total capital ratio
- ❑ total dollar percent earnings dividend payout
- ❑ log preferred stock
- ❑ log preferred stock to total capital ratio
- ❑ log total debt to total capital ratio
- ❑ log net sales over revenues growth

lambda.1se LASSO variables



Top 10 XGBoost variables



Analysis:

- ❑ Total Liabilities
- ❑ Long-term Debt
- ❑ Long-term Debt to Total Capital Ratio
  - ➢ All dealing with debts and liabilities
  - ➢ Higher debts balance tends to increase the probability of accounting misstatements

- ❑ Log Preferred Stock
- ❑ Log Preferred Stock to Total Capital Ratio
  - ➢ Increased Preferred Stock tends to increase probability of accounting misstatements
  - ➢ Another form of raising money for the bank

Can research more into the effects of high debts and preferred stock on banks and accounting misstatements

Thank you!

Q&A