FT5005 Machine Learning For Finance

First Report

Group 5

CHULABUTRA CHUENCHOKSAN A0297551U

NEIL HEINRICH BRAUN A0298256M

POH JIA JUN A0146735Y

SUHUT MICKEY LIN A0298198A

TAN CHEEN HAO A0298197E

06 April 2025

Contents

* Page Count Excluding Contents, Reference, and Appendix: 8

Introduction

Financial analysts have access to similar datasets and data sources. Thus, it is how data is used and interpreted that separates the truly exceptional from the mediocre. Machine learning has become a common buzzword in the industry, but without domain knowledge, its application offers little to no value. Fully understanding all of this, our project aspires to improve the predictive accuracy of two key financial metrics by leveraging domain-informed features learnt from academic literature and industry reports.

1 Research Objective and Overview

This project focuses on the US banking industry (SIC 6021), specifically publicly listed banks. The research objective is to refine the predictive accuracy of 2 key financial metrics: (1) quarterly revenue and (2) Cumulative Abnormal Returns around Earning Announcement (CAR). While traditional forecasting models lean heavily on historical financial statements, this approach often overlooks critical macroeconomic dynamics and sentiment-driven variables that significantly influence performance in the highly cyclical, macro-sensitive banking industry.

1.1 Revenue

To bolster machine learning methodologies in forecasting quarterly revenue, we are incorporating 4 additional datasets: (1) macroeconomic indicators, (2) sentiment indices derived from financial news, (3) customer reviews analysis, and (4) insights derived from Earning Calls. These datasets are designed to encapsulate real-time market conditions, regulatory expectations, and forward-looking sentiments—components frequently absent from backward-looking models anchored solely in past financial data. By incorporating these variables, the model is expected to interpret business cycles and shifting market expectations better, thereby enhancing its ability to assess lending volumes, net interest margins, and operational efficiency.

The macroeconomic indicators we plan to use include the Federal Funds Rate, Treasury Yields, GDP growth, CPI inflation, unemployment rate, loan growth, and deposit volumes. These variables will be sourced from the Federal Reserve Economic Data (FRED), the Federal Reserve's H.8 release, and statistical agencies such as BEA and BLS. The inclusion of Bank of America's Prime Rate trends also gives us a firm-specific benchmark for commercial lending costs and can supplement our engineered macro indicators[1].

For sentiment features, we will incorporate the daily news sentiment using the financial sentiment scores generated through NLP tools like the Loughran-McDonald dictionary and FinBERT, and insights extracted from earnings call transcripts. Empirical evidence has shown the direct impact of sentiment from earning calls on shaping investor expectations and gauging credit risk (Pastor & Lamers, 2023; Zhiqiang Ma et al., 2020).

Furthermore, there are studies which highlight the positive correlation between high customer satisfaction levels and brand equity, which subsequently leads to better financial performance (Boonlertvanich, 2019). To achieve this, we will leverage Google Maps reviews as a proxy for assessing each bank's brand image over time. We extracted the dataset via

web scraping techniques, and in order to ensure methodological consistency, the reviews will be processed and utilized similarly to other textual data.

To integrate these datasets into predictive models, we propose several domain-informed feature engineering approaches. First, we will construct an interest rate sensitivity score based on historical earnings and interest rate movements. Second, a yield curve sentiment index will be constructed by correlating the slope of the yield curve with internal bank ratios, such as the Loan-to-Deposit Ratio. Third, momentum indicators will be created by calculating moving averages and lagged growth rates of GDP and CPI. Fourth, lagged sentiment features, consisting of the average sentiment score from the previous quarter, will serve as predictors for subsequent quarterly performance. Lastly, a composite economic stress index which synthesizes unemployment, inflation, and loan charge-offs will quantify macroeconomic pressure on banks into one consolidated metric, thereby enhancing the model's interpretability.

We also draw insights from risk management literature, such as the Silicon Valley Bank (SVB) case study (Bao et al., 2024), which illustrates the detrimental impact of interest rate risk mismanagement with deposit growth, contributing to SVB's unexpected failure. These findings are crucial in shaping our own macro sensitivity indicators, serving as a guiding principle to mitigate similar vulnerabilities.

In addition, industry perspectives from influential firms, such as McKinsey, emphasize the significance of external forward-looking variables as revenue determinants, as seen in McKinsey's 2023 report on unlocking value in banking through technology.[2]

1.2 Cumulative Abnormal Returns around Earning Announcement (CAR)

Our second goal is to predict the CAR of publicly traded bank stocks. Given the daily Abnormal Returns (AR), we define CAR as

$$CAR(T) = \sum_{t=1}^{T} AR_t.$$

This strategy assumes a hypothetical long position initiated on the announcement day and exited at the close T days following the announcement. In particular, we will use $CAR(5)$ the 5-day total abnormal return after the earnings announcement date.

Academic research has shown a strong linkage between stock returns and the earnings announcement cycle (Linnainmaa & Zhang 2018). Therefore, we expect the lagged CAR to be a reliable feature in predicting the CAR.

Another study (Savor and Wilson, 2016) indicates a risk-return dynamic for companies announcing their earnings earlier as they face higher risk. Earlier announcements can indicate greater uncertainty about earnings, leading to an associated risk premium from investors. Leveraging this insight, we aim to create a feature representing the difference in the number of days between the required earnings publication date and the actual earnings announcement date.

Building upon the research indicated, market sentiment will be assessed using data extracted from financial news and the earning calls transcript to predict the CAR. Complementary quantitative metrics will include technical indicators tied to the bank's stock price and broader market performance, such as movements in the S&P 500 index.

2 Literature Review

2.1 Macroeconomic Indicators and Bank Performance

As outlined in Section 1.1, macroeconomic variables such as interest rates, inflation, and GDP growth are central to our revenue prediction model. Extensive research supports the strong correlation between these factors and bank profitability, particularly through their influence on net interest income, lending volumes, and deposit behavior. For instance, rising interest rates generally boost banks' net interest margins (NIM), as lending rates increase more rapidly than deposit costs. Watkins et al. (2022)[3] demonstrate this dynamic by illustrating that banks' disclosures of interest rate sensitivity are predictive of revenue fluctuations, supporting the inclusion of central bank rate levels and bank-specific interest sensitivity as engineered features.

Cepni et al. (2022)[4] incorporate interest rate uncertainty into dynamic panel models to improve the forecasting of pre-provision net revenue (PPNR). Their methodology validates using uncertainty indicators such as yield curve spread volatility or Fed policy path dispersion for features engineering. Windsor et al. (2023)[5] further support this finding by demonstrating that a 100 bps decrease in benchmark rates corresponds to a reduction in NIM by an average of 5 bps for Australian banks, emphasizing the revenue impacts due to rate adjustments.

As indicated earlier in section 1.1, we draw practical insight from the SVB case (Bao et al., 2024)[6] in risk management, particularly how mismatches in duration between long-term securities and deposit funding can ultimately lead to liquidity crises. Such failures underscore the need to concurrently track the interconnected variables of risk management – loan growth, deposit sensitivity, and interest rate risk.

2.2 News Sentiment and Bank Performance

Complementing our use of macroeconomic variables, we incorporate sentiment-based indicators to capture forward-looking expectations. Consistent with our methodology, this section reviews how financial news sentiment has been shown to enhance forecasting accuracy in macro and financial domains, especially when integrated with traditional financial metrics. Buckman et al. (2020)[7] show that the San Francisco Fed's News Sentiment Index has a strong correlation with real GDP movements and improves macroeconomic nowcasting. Cerchiello et al. (2022)[8] demonstrate that blending sentiment features with financial ratios improves the prediction of bank distress, achieving a usefulness score of 43.2% with sentiment compared to 31.1% without sentiment.

These findings validate incorporating lagged sentiment scores, document embeddings, or economic news indices as predictive features for banking models.

2.3 Service Quality and Bank Revenue

Customer sentiment is a key driver of long-term revenue. Boonlertvanich et al. (2019) demonstrated that service quality—measured through reliability, assurance, tangibility, and responsiveness—positively influences bank profitability by enhancing trust and reducing attrition. High-net-worth clients tend to prioritize overall satisfaction above specific service offerings, whereas customers with entrenched banking relationships exhibit lower sensitivity to service quality improvements. As mentioned in section 1.1, we leverage Google Maps reviews as a proxy for assessing each bank's brand image over time.

2.4 Earning Calls

Pastor & Lamers (2023)[9] quantify the predictive power of bank executives' tone during earnings calls, finding that 1 standard deviation (SD) in optimistic tone correlates with a 1.85% increase in loan growth. Zhiqiang Ma et al. (2020)[10] go further by applying deep learning to the Q&A portion of earnings calls and finding that unscripted sentiments are particularly useful in predicting stock direction. This supports the use of Earning Calls for CAR prediction.

Ma et al. (2020) demonstrate that earnings call transcripts, particularly in the unscripted Answer sections, contain valuable predictive signals for financial outcomes. The authors' deep learning model, which combines GloVe embeddings with attention mechanisms, successfully predicts stock price movements with a 52.45% accuracy rate by extracting meaningful patterns from management commentary. This methodology can be adapted specifically for revenue prediction through 3 key steps: (1) extracting lexical features and forward-looking statements using fine-tuned financial embeddings like FinBERT, (2) applying hierarchical attention models to identify revenue-relevant phrases (e.g., "Q3 growth accelerated" or "demand softened"), and (3) integrating these textual features with quantitative fundamentals and industry-specific embeddings (GICS-based) to improve accuracy.

2.5 Earning Announcement Time and CAR

As mentioned earlier in section 1.2, a study by Savor and Wilson (2016) shows that companies announcing their earnings earlier in the quarter experience higher abnormal returns compared to those that publish later.

However, another study by Adams and Neururer (2020) shows that there are many conflicting results regarding the earning announcement time and its associated risk. Their findings through the use of implied volatility suggest that companies that publish later in the quarter have higher risk profiles than companies which publish earlier, directly contradicting the earlier study's conclusions.

Although the exact dynamics between the announcement time and its effect on CAR remains unclear, there is sufficient evidence to suggest that there is a certain correlation between the two. Therefore, the announcement time is a valid feature for our model.

2.6 Social Media Sentiment and CAR

Social media sentiment is a strong predictor of stock price movements around earnings announcements. Recent studies have confirmed this connection. For example, Chen et al. (2022) found that a 1 standard deviation increase in positive Twitter sentiment for US banks before earnings announcements correlated with a 1.2% increase in cumulative abnormal returns (CAR). Similarly, Broadstock and Zhang (2019) showed that StockTwits sentiment data improved CAR prediction accuracy by 18% compared to traditional financial data alone.

Other research further supports the link between social media sentiment and stock performance. Oliveira et al. (2017) found that positive pre-announcement Twitter sentiment led to a 1.8% higher CAR for European banks. Sun et al. (2020) demonstrated that incorporating sentiment from investment forums reduced CAR forecast errors by 23.5%, further emphasizing the value of social media in predicting stock movements.

Based on these findings, even though extracting social media data is challenging, integrating social media sentiment into our CAR model could be beneficial. This would involve analyzing features such as mention volume, sentiment scores, and pre-announcement activity to capture investor attention and broader market sentiment.

2.7 Previous Modelling Approaches and Benchmarks

Different predictive modelling approaches have been tested across this domain, from traditional regressions to advanced machine learning. Cepni et al. (2022) use a dynamic panel machine learning framework to forecast PPNR under macro uncertainty and show improved performance over static models. Cerchiello et al. (2022) compare Doc2Vec neural networks with baseline financial ratio models and show large gains in predictive accuracy.

From an applied perspective, McKinsey (2023) argues that forward-looking signals, such as tech investment and market sentiment, are increasingly driving valuation multiples in banking. Integrating such signals requires adaptable modelling strategies — like ensemble methods or LSTM models — which are adept at processing both structured financial data and unstructured sentiment-driven inputs.

These insights support our decision to pursue hybrid models that combine macroeconomic indicators, firm-specific fundamentals, and sentiment analytics.

2.8 Summary Table

Study	**Best Method**	**Performance Metric & Value**	**Important Features**
Watkins et al. (2022)	Disclosure analysis	Predictive alignment between disclosed sensitivity and earnings	Interest rate sensitivity (conceptual + firm disclosure)
Cepni et al. (2022)	Dynamic panel ML	Reduced forecast error, especially for non-interest income	Interest rate uncertainty index

Pastor & Lamers (2023)	Panel regression	+1.85% credit growth per 1 SD sentiment increase	Forward-looking tone from earnings call transcripts
Ma et al. (2020, S&P Global)	Deep learning + attention	Outperformed ML baselines on stock direction	Sentiment in Q&A section of earnings calls (unscripted tone)
Linnainmaa & Zhang (2018)[11]	Return cycle analysis	71 bps/month alpha for EARC strategy	Analyst optimism cycles and earnings announcement timing
Windsor et al. (2023, RBA)	Panel regression	5 bps drop in NIM per 100 bps rate decline	Net interest margin, macro rate levels, provision reversal effects
Boonlertvanich (2019)[12]	Structural equation model	Strong indirect effect of quality via trust and satisfaction	Customer satisfaction, trust, service quality, main-bank status
Cerchiello et al. (2022)	Neural net (Doc2Vec)	43.2% usefulness with news, 31.1% without	Combined financial ratios + news sentiment embeddings
Buckman et al. (2020)	NLP sentiment index	Improved nowcasting, strong correlation with economic cycles	News tone from 24 newspapers

Table 1: The summary table of other previous research

These papers satisfy the theoretical requirement by showing that engineered macroeconomic and sentiment features causally affect revenue or related financial outcomes. They also provide concrete benchmarks and modelling comparisons relevant to our forecasting objective.

3 Important Data Summary Statistics

This section summarizes key data from macroeconomic indicators, bank earnings call transcripts, and Google Maps customer reviews. Due to space constraints, further data analysis can be found in the Appendix.

3.1 Macroeconomics Indicators

We downloaded external macroeconomics data from 2000 to 2024 from the Federal Reserve Economic Data (FRED). This includes: quarterly GDP, monthly unemployment rate, daily average bank prime loan rate, weekly commercial bank deposit, monthly yearly change in Sticky Price Consumer Price Index less Food and Energy (inflation), and Quarterly Net Saving as a percentage of Gross National Income. For non-percentage values, we transformed the data into the percentage change (such as the percentage change in GDP). We also

aggregated the data into a quarterly format to match our prediction frequency. The table below shows the summary of the macroeconomics indicator.

Metric	Quarterly GDP Change (%)	Quarterly Average Unemployment rate (%)	Quarterly Average Prime Loan Rate (%)	Quarterly Average change in weekly deposits (%)	Quarterly Average of the yearly change in CPI (%)	Quarterly Savings per Gross Incoming (%)
Mean	1.11	5.70	5.01	1.67	2.66	2.08
Std	1.44	1.95	2.00	1.46	1.09	1.87
min	-8.25	3.53	3.25	-1.85	0.70	-2.90
25%	0.81	4.17	3.25	1.01	2.10	1.05
50%	1.16	5.08	4.25	1.62	2.44	2.60
75%	1.47	6.34	6.47	2.10	2.87	3.13
max	8.77	13.00	9.50	12.37	6.47	6.30

Table 2: The summary statistics of the macroeconomics indicators.

3.2 Earning Calls Transcripts

As earnings calls often correlate with market movements, we extracted all bank earnings call transcripts from 2008 to 2024. We processed a total of 3,868 earnings call transcripts from various banks spanning from 2008 to 2024. The textual data underwent preprocessing, including tokenization, stopword removal, and lemmatization, before computing the Term Frequency - Inverse Document Frequency (TF-IDF) scores.

Name	Value
Total number of transcripts	3868
Total Number of Words	14,734,386
Average number of words in the transcript	3809.30
Vocabulary size	125,698

Table 3: Basic summary statistics of the earning calls data

The word clouds below highlight the extensive textual data analyzed, capturing a broad vocabulary indicative of financial discussions in earnings calls. The computed TF-IDF scores help identify the most relevant terms across the corpus, providing insights into key themes and market sentiment.

Figure 1: The word cloud of the earning calls. The left is the raw word count, while the right is the average word TF-IDF of each word.

3.3 Customer Reviews

As discussed previously, we will use customer reviews on banks which have physical branches on Google maps. In order to do this, we utilized the Google Map Extractor[13] software to extract all review data for the banks from Google Maps.

However, the data contains certain issues. Firstly, even though the data includes a lot of data on popular banks such as Bank of America and JP Morgan Chase, there are certain public banks which do not have many branches such as OptimumBank Holdings Inc. Furthermore, branches which are already closed down are unlikely to be present on Google Maps as of 2025. Regardless, the data should be comprehensive enough for our use case.

We will be performing sentiment analysis on the review data and averaging across branches for the whole quarter. Below are the basic statistics of the review data.

Name	**Value**
Total number of reviews	606,330
Number of words in the vocabulary	296,851
Average number of words per review	22.44
Number of unique locations	66,192

Table 4: Basic summary statistics of the Google reviews data

The figure below also shows the count of reviews over time. This could arise from the older branches being discontinued, an increase in the popularity of Google Maps, and an increase in spam reviews. In fact, we can see clearly that there was a drop in the number of reviews in Q2 of 2024, which arose from the change in the Google spam detection algorithm. This could affect our testing result as we will be using 2021 to 2024 data for testing.

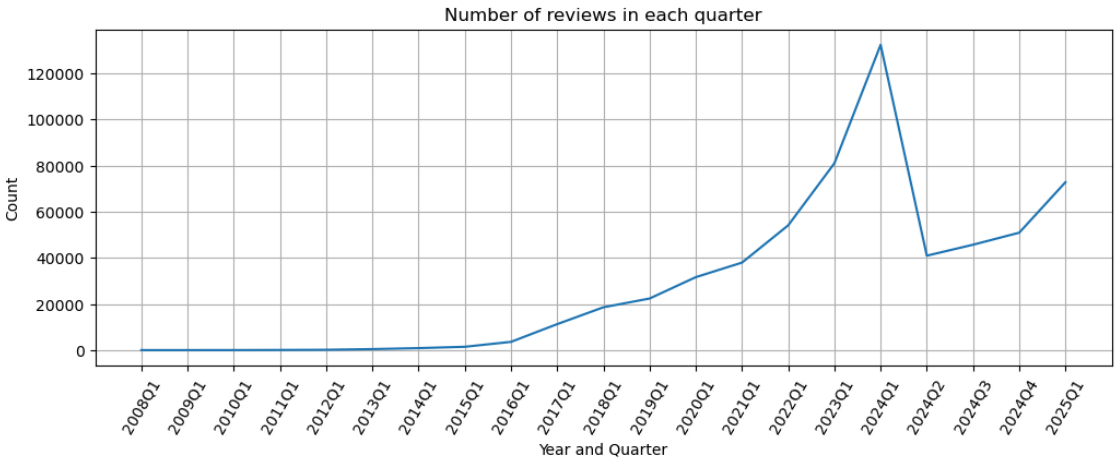

Figure 2: Number of Google Map reviews across different quarters.

The figure below shows the word cloud of the raw word count and the word cloud of the average TF-IDF value of each of the words after basic preprocessing such as lemmatization and case-folding.

Figure 3: The word cloud of the reviews data. The left is the raw word count. The right is the average word TF-IDF of each word.

Conclusion

We have carefully selected over 20+ features in order to improve the predictive accuracy of quarterly revenue and CAR. Among these, certain features are extremely common, such as Treasury Yields and CPI inflation. While others, like sentiment analysis, are gaining traction within the industry.

Taking a fresh approach, we gathered additional data in the form of reviews as a proxy for customer satisfaction and the quality of the banks' services.

Additionally, there are also new features we engineered ourselves, such as utilizing one metric to combine unemployment, inflation, and loan charge-offs. It is these last sets of domain-specific features that are unique to our model and reflect the culmination of our research efforts. Consequently, we are optimistic about the model's predictive performance and have high expectations for its success.

References

1. Bank of America. (n.d.). *Prime rate information*. Bank of America Newsroom. https://newsroom.bankofamerica.com/content/newsroom/home/prime-rate-information.html
2. McKinsey & Company. (2023, May 23). *Unlocking value from technology in banking: An investor lens*. https://www.mckinsey.com/industries/financial-services/our-insights/unlocking-value-from-technology-in-banking-an-investor-lens
3. Hardy, M. (2022, Spring). *The ups and downs of interest sensitivity*. *Mendoza Business Magazine*. https://bizmagazine.nd.edu/issues/2022/spring-2022/research-the-ups-and-downs-of-interest-sensitivity
4. Çepni, O., Demirer, R., & Gupta, R. (2022). Interest rate uncertainty and the predictability of bank revenues. *Journal of Forecasting*, 41(5), 923–943. https://doi.org/10.1002/for.2884
5. Windsor, C., Jokipii, T., & Bussiere, M. (2023). *The impact of interest rates on bank profitability: A retrospective assessment using new cross-country bank-level data* (RDP 2023-05). Reserve Bank of Australia. https://doi.org/10.47688/rdp2023-05
6. Bao, J., Campbell, T., & Stocker, J. (2023). *Silicon Valley Bank: The role of risk (mis)management* (Case No. W34036). Harvard Business Publishing. https://www.hbsp.harvard.edu/product/W34036-PDF-ENG
7. Buckman, S. R., Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). *News sentiment in the time of COVID-19* (Economic Letter No. 2020-08). Federal Reserve Bank of San Francisco. https://www.frbsf.org/wp-content/uploads/el2020-08.pdf
8. Cerchiello, P., Nicola, G., Rönnqvist, S., & Sarlin, P. (2022). *Assessing banks' distress using news and regular financial data*. *Frontiers in Artificial Intelligence*, 5, Article 871863. https://doi.org/10.3389/frai.2022.871863
9. Pastor y Camarasa, P., & Lamers, M. (2023). *Do actions follow words? How bank sentiment predicts credit growth*. SSRN. https://doi.org/10.2139/ssrn.4549755
10. Ma, Z., Bang, G., Wang, C., & Liu, X. (2020). *Towards earnings call and stock price movement*. In *KDD Workshop on Machine Learning in Finance (MLF '20), August 24, 2020, Virtual Event, USA*. ACM. https://arxiv.org/abs/2009.01317
11. Linnainmaa, J. T., & Zhang, C. Y. (2018). *The earnings announcement return cycle* (Working Paper). Retrieved from https://www.aeaweb.org/conference/2019/preliminary/paper/YrbKK6Zn
12. Boonlertvanich, K. (2019). Service quality, satisfaction, trust, and loyalty: The moderating role of main-bank and wealth status. *International Journal of Bank Marketing, 37*(1), 278–302. https://doi.org/10.1108/IJBM-02-2018-0021
13. GMAPS Extractor. (n.d.). *Google Maps Extractor*. https://gmapsextractor.com/

14. Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, 125-144. https://www.sciencedirect.com/science/article/abs/pii/S0957417416307187
15. Broadstock, D. C., & Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. *Finance Research Letters*, 30, 116-123. https://www.sciencedirect.com/science/article/abs/pii/S1544612318307888
16. Chen, H., De, P., Hu, Y., & Hwang, B. H. (2022). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 35(4), 1868-1906. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1807265
17. Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, 84, 25-40. https://www.sciencedirect.com/science/article/abs/pii/S0378426617301589
18. Sun, L., Najand, M., & Shen, J. (2020). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, 114, 105744. https://www.sciencedirect.com/science/article/abs/pii/S0378426616301595

Appendix

A1. Important Company Fundamentals

From the main quarterly reports Compustat dataset, we will also be using the company fundamental values to predict both the CAR and Revenue. The values that we deemed useful are:

1. Net Interest Income
2. Net Interest Margin
3. Net Charge-Offs
4. Cash and Short-term Investments
5. Net Income

Following a 2-tailed Winsorization, we get the following results.

Statistic	Net Interest Income	Net Interest Margin	Net Charge-Offs	Cash and Short-Term Investments	Net Income
mean	514.70	3.61	-41.24	12845.10	192.12
std	2026.76	0.65	171.42	75339.42	865.04
min	2.06	2.11	-1194.00	5.64	-84.89
25%	10.87	3.20	-5.40	44.64	1.78
50%	35.04	3.56	-0.94	135.04	6.48
75%	119.17	3.96	-0.13	509.06	30.37
max	24177.00	5.32	323.00	1056407.00	9179.00

Table A1.1. Important company fundamentals statistics after winsorizing.

As the Net Interest Income, Net Charge-Offs, Cash and Short-Term Investments, and Net Income have very extreme values, we will also perform a sign log transformation of the form

$$\widehat{X} = sign(X) \times log(1 + X).$$

This transformation will be able to deal with potential negative values.

Statistic	Net Interest Income (log)	Net Interest Margin	Net Charge-Offs (log)	Cash and Short-Term Investments (log)	Net Income (log)
mean	3.90	3.61	-1.27	5.35	2.29
std	1.80	0.65	1.68	2.18	2.26
min	1.12	2.11	-7.09	1.89	-4.45
25%	2.47	3.20	-1.86	3.82	1.02
50%	3.58	3.56	-0.66	4.91	2.01
75%	4.79	3.96	-0.12	6.23	3.45
max	10.09	5.32	5.78	13.87	9.12

Table A1.2. Important company fundamentals statistics after winsorizing and sign-log transformation.

These values now have a better distribution spread suitable for predictors and are ready to be min-max scaled. The box plot for the data is shown below.

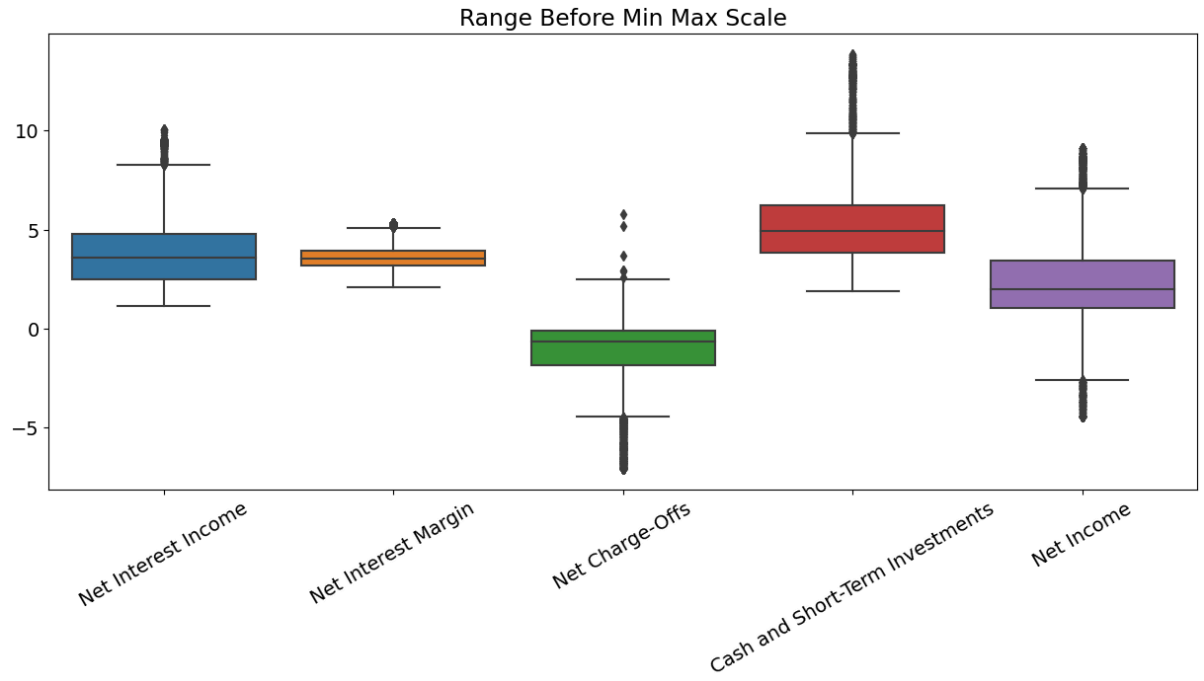

Figure A1.1. Box plot of important company fundamentals statistics after winsorizing and sign-log transformation.

A2. Important Additional Banking-Specific Fundamentals

From the banking-specific Compustat dataset, we deemed the following useful:

1. Deposits - Interest Bearing
2. Total Savings Deposits
3. Deposits - Total
4. Invested Capital - Total
5. Interest Income - Total

After winsorization, we will get the following statistics:

Statistic	Deposits - Interest Bearing	Total Savings Deposits	Deposits - Total	Invested Capital - Total	Interest Income - Total
mean	129868.37	54633.3821	31160.09	7493.14	464.26
std	256638.16	135493.8252	149955.18	41339.93	2466.67
min	5210.50	3309.94	64.96	14.11	1.26
25%	8565.02	4310.32	428.28	84.90	6.61
50%	26485.78	12049.28	1004.86	199.95	14.87
75%	97323.75	36886.00	3814.72	776.34	53.89
max	1812264.00	1229243.00	2561207.00	764834	79459.63

Table A2.1. Other company fundamentals statistics after winsorizing.

Again, as these values contain many extreme values, we will use the sign log transformation. The results are shown below.

Statistic	Deposits - Interest Bearing	Total Savings Deposits	Deposits - Total	Invested Capital - Total	Interest Income - Total
mean	10.44	9.66	7.38	5.78	3.28
std	1.58	1.40	1.96	1.97	1.84
min	8.56	8.11	4.19	2.72	0.81
25%	9.06	8.37	6.06	4.45	2.03
50%	10.18	9.40	6.91	5.30	2.76
75%	11.49	10.52	8.25	6.66	4.01
max	14.41	14.02	14.76	13.55	11.28

Table A2.2. Other company fundamentals statistics after winsorizing and sign-log transformation.

The box plot for each value before the min-max scale is shown below.

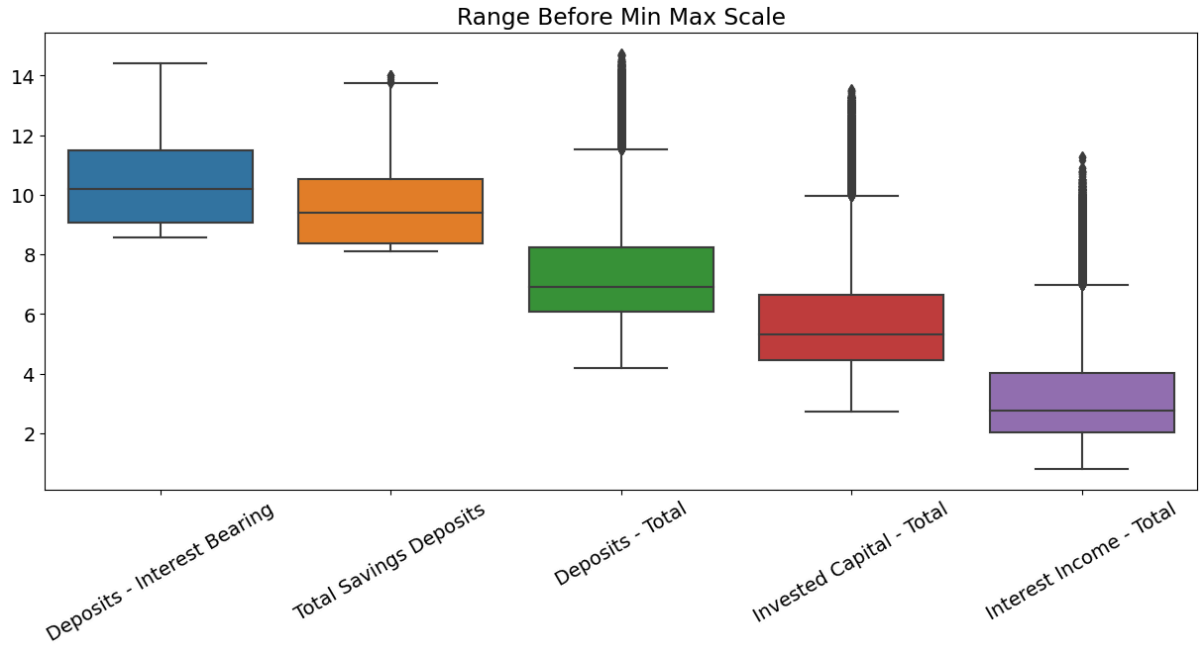

Figure A2.1. Box plot of important company fundamentals statistics after winsorizing and sign-log transformation.

A3. Y Values - Revenue and CAR

The revenue we will be predicting is the total current operating revenue. Furthermore, as there are many extreme values, we will predict the log of the revenue.

As for the CAR, we will be using the CAR(5) as given. Even though there are some outliers, these outliers are actual events and should not be removed or winsorized out. The distributions of the two Y values are shown below.

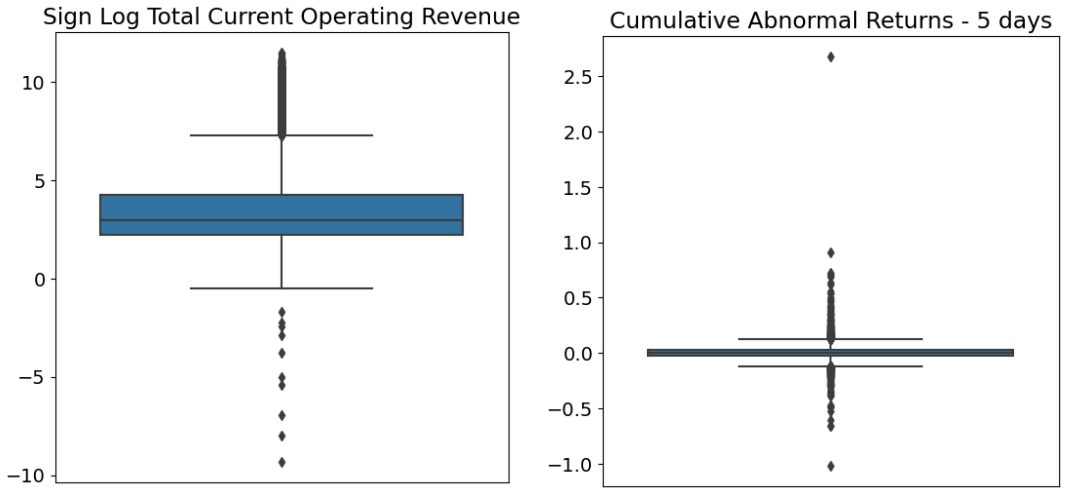

Figure A2.1. Box plot of the y values. Left is the Sign-Log of the Total Current Operating Revenue. Right is the $CAR(5)$.