US Company Bankruptcy Detection

Course: Deep Learning I (AASD4010)

Group 1

Group Members

1. Chotiros Srisiam #101411914
2. Kajhonprom Trongkitroongruang #101446812
3. Pat Boonprasertsri #101410612
4. Pek Chansatit #101439953
5. Vitchaya Siripoppohn #101481464

Date: February 8, 2024

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1.0 Introduction

In the dynamic and ever-changing world of financial markets, the ability to identify early signs of trouble and properly forecast possible bankruptcies is critical for investors, analysts, and businesses alike. This study aims to improve the accuracy of bankruptcy prediction by leveraging cutting-edge deep learning algorithms. Our focus extends to addressing the complex issues provided by the diverse forces that shape the dynamic financial industry.

Among the constant evolution of the financial world, the ability to foresee bankruptcy emerges as an essential skill. Our study strategically uses modern machine learning approaches to predict impending bankruptcies in publicly traded corporations in the United States. Our overarching goal is to equip stakeholders with a sophisticated tool that will improve risk management methods and enable nuanced strategic decision-making. This study methodically recounts our journey from problem discovery to the building of a strong predictive model, demonstrating deep learning's potential to provide novel solutions to traditional financial challenges.

1.1 Problem Statement

Bridging the Gap in Bankruptcy Prediction Through Advanced Deep Learning

Navigating the intricate landscape of financial markets poses an intricate challenge in accurately predicting impending bankruptcies. The existing methodologies employed for projecting bankruptcy in US public corporations exhibit a notable lack of accuracy, creating a void that hinders stakeholders and investors from making well-informed decisions. This glaring disparity underscores the pressing need for an intelligent, machine learning-based solution capable of augmenting predictive capabilities and facilitating effective risk management strategies.

Traditional approaches to bankruptcy prediction often prove inadequate, falling short in precision when confronted with the complexities of evolving financial scenarios. Recognizing these inherent limitations, this research endeavors to bridge the gap by leveraging advanced deep learning algorithms. The overarching objective is to craft a robust bankruptcy prediction model that not only overcomes the shortcomings of traditional techniques but also introduces a new paradigm in predictive analytics within the financial domain. Through the integration of sophisticated machine learning methodologies, this work aspires to reshape the landscape of bankruptcy prediction, offering stakeholders and investors a more reliable and accurate tool for navigating the challenges of financial uncertainty.

1.2 Motivation

Figure 1 Enron Logo

Value or Cost of the Damage Incurred in the Enron Scandal

Consider Enron Corporation, a once-global energy behemoth that went bankrupt in 2001 because of severe accounting fraud and financial mismanagement. The Enron disaster, which had a stunning original value of $70 billion in 2001, serves as a harsh reminder of the dangers of financial mismanagement and the limitations of conventional assessments.

If we want to adjust the value of $70 billion in 2001 to its equivalent in 2024, you need to consider the impact of inflation over the years. The formula for adjusting for inflation is:

Adjusted Value=Nominal Value× [(1+Inflation Rate) ^n]

Where:

- Adjusted Value is the value in 2024,

- Nominal Value is the value in 2001 ($70 billion),

- Inflation Rate is the average annual inflation rate, and

- n is the number of years between 2001 and 2024.

Assuming an average annual inflation rate of 2%, and n = 2024 - 2001 = 23 years:

Adjusted Value = 70,000,000,000 x [(1 + 0.02) ^ 23]

Adjusted Value = 107,926,267,800.59

Therefore, with an average annual inflation rate of 2%, adjusting the nominal value of $70 billion in 2001 yields an estimated equivalent value of approximately $107.93 billion in 2024, accounting for the impact of inflation on purchasing power. In the context of this historical scenario, the adjusted value in 2024 is expected to be $107 billion, underscoring the importance of utilizing advanced prediction models, like deep learning, to identify early indicators of financial distress and mitigate risks associated with high-stakes financial decisions. This report chronicles our journey from problem identification to the development of a robust prediction model, drawing insights from historical events such as Enron to showcase the revolutionary potential of advanced machine learning in transforming traditional approaches to financial forecasting.

The Enron scandal, a watershed moment in corporate history, serves as a compelling motivation for the development and implementation of advanced predictive models, particularly in the realm of bankruptcy prediction. The extensive damage incurred during the Enron debacle highlights the far-reaching consequences of financial mismanagement and underscores the need for robust risk mitigation strategies.

1) Market Capitalization Loss

Enron's market capitalization, which peaked at over $70 billion, plummeted rapidly. By the time of its bankruptcy filing in December 2001, the market capitalization had dwindled to almost zero. The colossal loss in market value is a stark reminder of the devastating impact on shareholder wealth and market integrity.

2) Investor Losses

Shareholders experienced substantial losses as Enron's stock price collapsed. Estimates suggest that investors incurred tens of billions of dollars in losses, underscoring the importance of early detection mechanisms to safeguard investor interests.

3) Employee Impact

Enron's employees not only lost their jobs but also suffered significant financial setbacks, particularly in their retirement savings heavily invested in Enron stock. Employee losses, including pension and 401(k) plan reductions, amounted to billions of dollars, highlighting the human impact of financial scandals.

4) Legal and Regulatory Consequences

Enron faced a barrage of lawsuits, investigations, and settlements. The legal and regulatory costs associated with the scandal, including fines and penalties, ran into hundreds of millions of dollars. The profound legal

repercussions underscore the importance of proactive risk management to avoid entanglement in costly legal battles.

5) Collapse of Arthur Andersen

The fallout from the Enron scandal extended to its accounting firm, Arthur Andersen, leading to its collapse. The dissolution of Arthur Andersen had widespread financial ramifications, affecting not only the firm but also its stakeholders. This emphasizes the interconnectedness of entities within the financial ecosystem.

6) Broader Economic Impact

The Enron scandal reverberated beyond the corporate realm, impacting broader economic dynamics. The erosion of investor confidence, changes in regulatory practices, and shifts in corporate governance standards had ripple effects throughout the economy, emphasizing the interconnected nature of financial markets.

7) Total Financial Fallout

While it's challenging to quantify the exact total financial fallout, the combined impact of market capitalization loss, investor losses, legal costs, and other consequences likely amounted to tens of billions of dollars. This comprehensive assessment highlights the multifaceted nature of financial damage incurred during the Enron scandal.

Given the numerous and extensive consequences of the Enron scandal, the motivation for building improved prediction models becomes clear. These models, particularly those based on deep learning, seek to function as proactive instruments for detecting early indicators of financial trouble, mitigating risks, and protecting the financial well-being of investors, employees, and the greater economy.

1.2.1 Comparison of Traditional vs. Deep Learning Model: Enron Case

Traditional Analysis

Traditional financial analysis methods proved inadequate in detecting Enron's fraudulent activities, as the company manipulated financial statements and concealed debt through off-balance-sheet entities. The limitations of traditional approaches became evident as they struggled to uncover the intricate financial maneuvers that ultimately led to Enron's downfall.

Deep Learning Model

In contrast, a deep learning model for bankruptcy prediction could have played a pivotal role in identifying Enron's financial distress early on through.

1) Analysis of a Broader Range of Financial Indicators and Patterns

Leveraging the ability to process vast datasets, a deep learning model can analyze a comprehensive array of financial indicators and patterns. This includes scrutinizing complex relationships and dependencies among variables that may elude traditional analysis methods.

2) Recognizing Anomalies and Irregularities in Real-Time Data

The dynamic nature of deep learning allows for real-time analysis, enabling the model to recognize anomalies and irregularities as they emerge. This capability is crucial in swiftly identifying deviations from expected financial patterns and offering a proactive stance in risk management.

3) Providing a Proactive Mechanism for Stakeholders to Address Risks Promptly

The proactive nature of deep learning models allows stakeholders to address risks promptly. By providing early warnings based on continuous analysis, the model empowers decision-makers with timely insights, enabling them to take corrective actions before financial distress escalates.

This comparison highlights the transformational potential of deep learning algorithms for improving the accuracy and timeliness of bankruptcy prediction. While traditional methods failed to negotiate the complexities of Enron's fraudulent actions, a deep learning approach emerges as a proactive and powerful tool for recognizing and mitigating financial risks in real time.

1.2.2 Root Cause Analysis (RCA): Unveiling the Imperative for a Bankruptcy Prediction Model

1.2.2.1 Financial Statement Complexity: Navigating the Intricacies of Financial Landscape

Within the expansive and intricate realm of finance, the complexity embedded in financial statements presents a formidable challenge for traditional manual analysis. The intricate tapestry of variables, relationships, and patterns woven into these statements surpasses the capacity of human analysts to comprehensively interpret. As financial landscapes evolve and become increasingly intricate, the imperative to discern subtle indicators of distress hidden within this labyrinth becomes paramount. Enter deep learning models.

These models, harnessing the capabilities of artificial intelligence, transcend the constraints of traditional analyses. Their exceptional proficiency in unraveling hidden patterns within intricate financial data positions them as indispensable assets in navigating complexities that elude human scrutiny. Picture them as vigilant custodians of financial intricacies, uncovering insights that may remain obscured to the human eye.

Confronted with convoluted financial landscapes, these models assume the role of sophisticated sentinels, offering a level of understanding and foresight beyond the confines of conventional approaches. Their analytical acumen enables them to discern the subtlest nuances, providing decision-makers with a depth of insight crucial for proactive risk detection. As invaluable allies in the financial domain, deep learning models redefine the landscape of analysis by illuminating the obscured, offering clarity in the face of complexity, and empowering decision-makers with unprecedented sophistication.

1.2.2.2 Global Economic Volatility: Navigating Uncharted Waters

In the ever-shifting tides of global economic uncertainty and market fluctuations, businesses find themselves sailing through turbulent waters, vulnerable to unforeseen challenges. It is within this dynamic environment that the need for a predictive model, integrating macroeconomic indicators, emerges as a strategic imperative.

Deep learning models, akin to seasoned navigators in uncharted waters, stand as vigilant guardians, offering a panoramic view of potential risks. Their ability to assimilate and analyze vast datasets empowers companies to not only anticipate but strategically prepare for economic downturns. Consider these models as strategic allies in the corporate voyage, equipped to read the winds of global economic volatility and chart a course that ensures resilience and foresight.

Confronted with the uncertainty inherent in the global economic landscape, deep learning models serve as sophisticated instruments for risk mitigation. Their integration of macroeconomic factors into the prediction framework becomes a crucial compass, allowing businesses to steer through turbulent seas with precision and preparedness. In essence, these models become indispensable companions for decision-makers, providing clarity and strategic direction amidst the ever-changing currents of global economic volatility.

1.2.2.3 Data Overload and Processing Speed: Accelerating Decision-Making in the Information Deluge

In the contemporary landscape of finance, where information flows ceaselessly and voluminously, the swift processing of vast financial data is not merely advantageous but a strategic imperative. Deep learning models,

revered for their prowess in handling large datasets with lightning speed, ascend as the linchpin for timely decision-making.

Picture these models as accelerators amidst the information deluge, not just managing data overload but ensuring rapid processing. Their efficiency becomes paramount in ensuring that decision-makers can extract meaningful insights in real-time, enabling proactive responses to dynamic market conditions.

As the volume of financial data continues to surge, deep learning models become indispensable allies for decision-makers, ensuring not only the management of data velocity but also guaranteeing swift processing. In essence, they evolve into critical instruments in the arsenal of tools for contemporary financial decision-makers, ensuring that the speed of data analysis matches the rapid pace of financial markets.

1.3 Project Overview: Precision in Predicting Bankruptcy for Informed Decision-Making

In the dynamic world of American public firms, the capacity to foresee bankruptcy has emerged as a critical tool for decision-makers. This project was designed with the explicit goal of creating a precision-focused model that goes beyond traditional approaches, empowering investors with comprehensive risk management capabilities.

The project's central focus is on the vital requirement for accuracy in predicting bankruptcy, with recognition of the far-reaching ramifications for financial decision-makers. Our goal, using innovative methodology and cutting-edge techniques, is to create a model that not only accurately forecasts possible bankruptcies but also serves as a strategic asset for investors navigating the difficulties of risk management.

The project's relevance stems from its commitment to improving the decision-making process for stakeholders by providing them with a dependable tool for anticipating and mitigating the risks associated with financial uncertainty. Our model intends to promote predictive analytics in the context of bankruptcy prediction by meticulously developing and validating it, ultimately enabling a more informed and proactive approach to risk management in American public corporations.

1.4 Dataset

The dataset, obtained from Kaggle, spans 1999 to 2018, covering 8,262 American public companies on NYSE and NASDAQ. Bankruptcy (Chapter 11 or 7) is labeled "Bankruptcy" (1) in the fiscal year before, and normal operation is labeled (0). Complete and without missing values, it comprises 78,682 observations. The dataset is split into training (1999-2011), validation (2012-2014), and test sets (2015-2018) for model evaluation.

Table 1 Dataset Field Description

Field	Description
X1	**Current assets** - All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year
X2	**Cost of goods sold** - The total amount a company paid as a cost directly related to the sale of products
X3	**Depreciation and amortization** - Depreciation refers to the loss of value of a tangible fixed asset over time (such as property, machinery, buildings, and plant). Amortization refers to the loss of value of intangible assets over time.
X4	**EBITDA** - Earnings before interest, taxes, depreciation, and amortization. It is a measure of a company's overall financial performance, serving as an alternative to net income.

Field	Description
X5	**Inventory** - The accounting of items and raw materials that a company either uses in production or sells.
X6	**Net Income** - The overall profitability of a company after all expenses and costs have been deducted from total revenue.
X7	**Total Receivables** - The balance of money due to a firm for goods or services delivered or used but not yet paid for by customers.
X8	**Market value** - The price of an asset in a marketplace. In this dataset, it refers to the market capitalization since companies are publicly traded in the stock market.
X9	**Net sales** - The sum of a company's gross sales minus its returns, allowances, and discounts.
X10	**Total assets** - All the assets, or items of value, a business owns.
X11	**Total Long-term debt** - A company's loans and other liabilities that will not become due within one year of the balance sheet date.
X12	**EBIT** - Earnings before interest and taxes.
X13	**Gross Profit** - The profit a business makes after subtracting all the costs that are related to manufacturing and selling its products or services.
X14	**Total Current Liabilities** - The sum of accounts payable, accrued liabilities, and taxes such as Bonds payable at the end of the year, salaries, and commissions remaining.
X15	**Retained Earnings** - The amount of profit a company has left over after paying all its direct costs, indirect costs, income taxes, and its dividends to shareholders.
X16	**Total Revenue** - The amount of income that a business has made from all sales before subtracting expenses. It may include interest and dividends from investments.
X17	**Total Liabilities** - The combined debts and obligations that the company owes to outside parties.
X18	**Total Operating Expenses** - The expenses a business incurs through its normal business operations.

1.4.1 Overview

Our dataset serves as a comprehensive repository of financial data, meticulously curated to discern the intricate dance of companies between financial stability and distress. Each entry encapsulates crucial information that forms the bedrock of our predictive model's insights.

1.4.2 Bankruptcy Labeling

The dataset employs a binary labeling system to categorize companies into two distinct states: bankrupt (1) or operating normally (0). A company receives the "bankrupt" label if it undergoes the legal processes of Chapter 11 (reorganization) or Chapter 7 (complete cessation). These classifications align with the definitions outlined by the U.S. Securities and Exchange Commission (SEC).

1.4.3 Temporal Considerations

The temporal dimension is crucial in our dataset, with a specific focus on the fiscal year preceding any filing for Chapter 11 or Chapter 7. This period is aptly labeled as "Bankruptcy" to signify the financial state leading up to the legal actions.

1.4.4 Normal Operations

Conversely, if a company does not engage in Chapter 11 or Chapter 7 proceedings, it is considered to be operating normally (labeled as 0). This distinction adheres to the SEC definitions, ensuring clarity and consistency in our dataset.

1.4.5 Purpose

Our dataset is meticulously crafted to empower our predictive model in discerning early signs of financial distress and predicting potential bankruptcy. By leveraging the rich tapestry of financial data, we aim to contribute valuable insights to the realm of financial forecasting, assisting stakeholders in navigating the complex landscape of risk management and strategic decision-making.

2.0 Methodology

Our project methodology follows a concise pipeline: data collection, pre-processing, model training, evaluation, and validation. This streamlined approach ensures not only accurate models but also their effectiveness in real-world applications.

2.1 Collect Data

The dataset, collected by Utkarsh Singh, is originating from the New York Stock Exchange and NASDAQ, and sourced from Kaggle webpage. process involved downloading the dataset from the website, and to facilitate analysis, we implemented code to efficiently read the CSV file, ensuring seamless integration into our project pipeline.

2.2 Pre-process Data

In the early stages of our trip, we inspect the raw data, and pre-process it into a refined and ready-to-use format. This phase laid the groundwork for a reliable and efficient analytical approach.

2.2.1 Exploring the Dataset: Unveiling Imbalances

As we dive into the dataset, we discover imbalances, both in firm distribution and over time. These variations add an interesting layer to our understanding of the data.

2.2.1.1 Imbalanced Firm Distribution

In this dataset, most firms—around 93%—aren't facing bankruptcy. Only 7% fall into the unique category of firms undergoing bankruptcy. This rarity makes it a bit tricky for our model to spot patterns related to this specific financial state.

2.2.1.2 Temporal Imbalances

Looking closer, we notice some ups and downs in the dataset over time. In 1999, there's a lot of data, but it drops in the following years until 2018. This up-and-down pattern leads us to explore each time carefully, acknowledging the differences in data amounts and how they might affect our model's predictions.

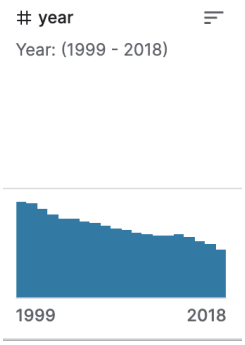

Figure 2 Differences data amounts in each year

2.2.2 Outlier Analysis

Embarking on an outlier analysis, our approach to those is remove the outlier based on other data points in our dataset. This approach can impact the robustness of our predictive model.

To overcome the integrity of our dataset, we target specific instances that show extreme values. In addition, we focus on outliers in the X6 data.

Furthermore, we also notice the outliers in the X8, X10, and X11 features. We remove it by handpicking. This outlier analysis and these processes contribute to the overall robustness of our dataset, increase its suitability for effective predictive modeling.

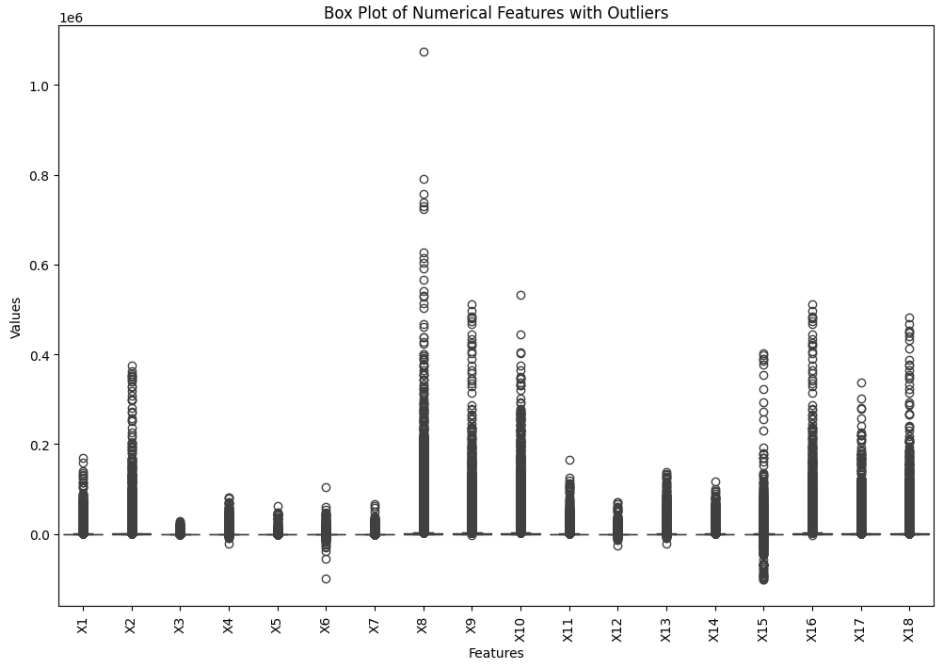

Figure 3 Before removing outliers

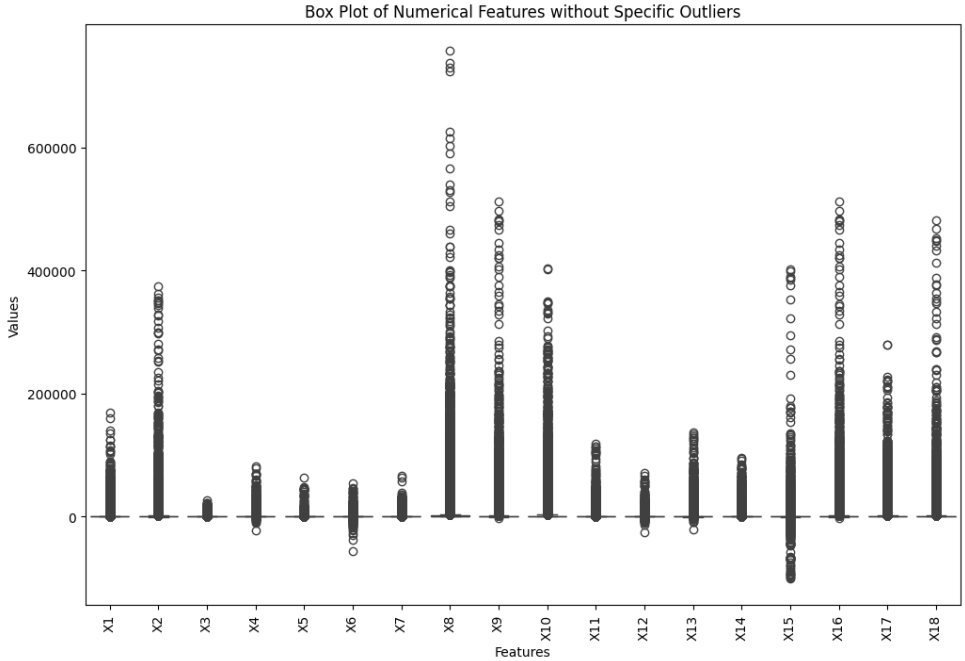

Figure 4 After removing specific outliers

2.2.3 Feature Analysis and Future Model Enhancement

Our feature analysis aims to understanding the insight within our dataset. we notice patterns, dependencies, and insights, refining the dataset by removing some features. This curation ensures focus on influential variables, fostering a more effective and interpretable model.

Correlation matrices aid in this analysis, it's representation of feature relationships, guiding us to identify multicollinearity and selected variables. To understanding observed correlations enables informed decisions during model development, facilitating enhancements such as fine-tuning hyperparameters, adjusting architectures, and exploring advanced techniques are used.

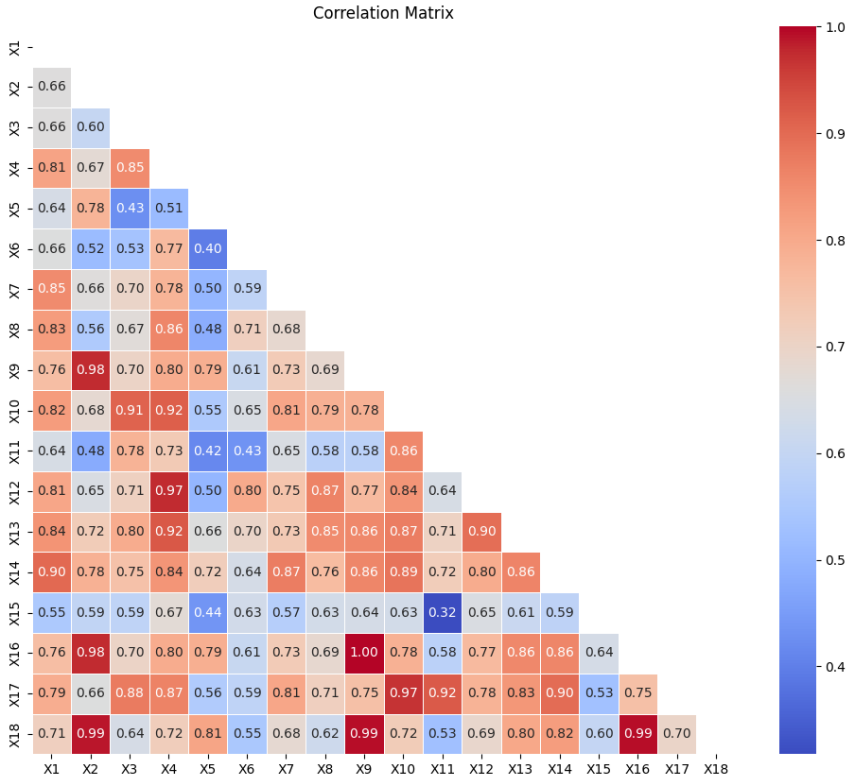

Figure 5 Feature Analysis with Correlation Matrix

2.3 Train-Test Split

Time-Series Aware Train-Test Split

Since we cannot randomly separate the data due to the time series pattern, our approach was taken to split it into training (1999-2011) and test sets (2012-2018). This decision was important, considering that each year's data contributes a unique story.

The train-test split was separated by respecting the time series events:

Training Set (-2011) - 75%: The train data including data points between year 1999 and 2011, representing 75% of the dataset. Features (X_rf_train) and the corresponding labels (y_rf_train) were used for model training.

Validation Set I (2012-2014) - 13%: The validation data including data points between year 2012 and 2014, representing 13% of the dataset. Features (X_rf_test) and the corresponding labels (y_rf_test) were used for model validating.

Test Set (2015+) - 12%: The validation data including data points between year 2015 and 2018, representing 12% of the dataset. Features (X_val) and the corresponding labels (y_val) were used for model testing.

2.4 Model Exploration and Evaluation

In this project, we use traditional machine learning models included Random Forest, XGBoost, Weighted XGBoost, and Resampling XGBoost. To facilitate model comparison and performance assessment, we converted the 'status_label' to numerical labels (0 for 'alive' and 1 for 'failed').

The evaluation we used in this project is confusion matrix, including four aspects of sensitivity, specificity, accuracy, and F1-score.

Moreover, we also calculating Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC). These measures provide an understanding of how well the models distinguish between different classes, offering valuable insights into their overall effectiveness. The ROC curve analysis serves can enhancing our comprehension of each model's discriminative capabilities in the context of bankruptcy prediction.

2.4.1 Random Forest

Random forest is a commonly used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems (IBM, 8 Feb 2024).

The model is recognized as a state-of-the-art model in machine learning, Random Forest offers several advantages, including its ability to deliver accurate predictions efficiently. In our exploration, it showed a precision of 0.98, recall of 0.05, and an F1-score of 0.09 for the 'failed' class, showcasing its capability to accurately identify instances of financial distress.

However, it's important to note the imbalanced nature of the dataset, impacting the metrics for the minority class. The confusion matrix further illustrates this, with a higher number of true negatives (11978) and true positives (15) but a relatively low recall for the 'failed' class.

In addition to these traditional evaluation metrics, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were calculated, providing a nuanced understanding of Random Forest's discriminative capabilities. With an AUC of 0.85, the ROC curve visually reinforces the model's effectiveness in distinguishing between normal and distressed financial states.

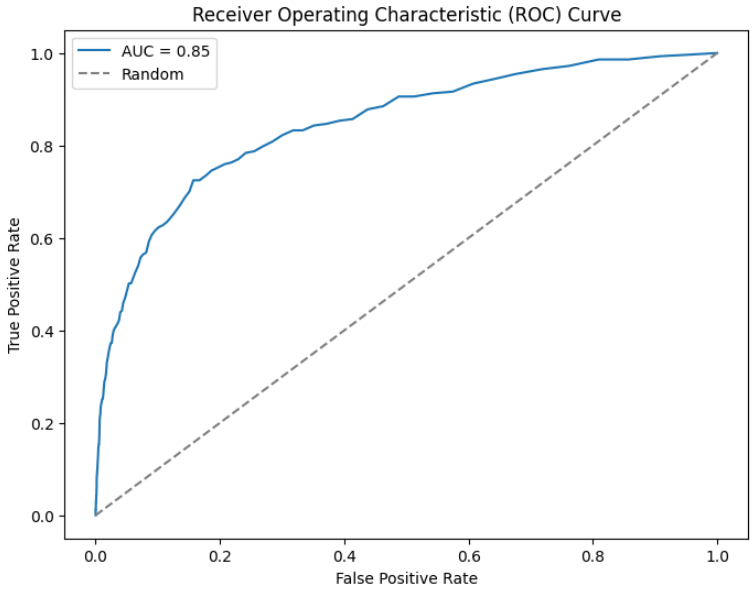

Figure 6 Random Forest ROC Curve

2.4.2 XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework (IBM, 8 Feb 2024).

In our exploration, XGBoost demonstrated a precision of 0.98 and recall of 0.12 for the 'failed' class, along with an F1-score of 0.20. This emphasizes its capability to effectively identify instances of financial distress, showcasing a balance between precision and recall.

Similar to Random Forest, the imbalanced nature of the dataset is evident in the metrics for the minority class. The confusion matrix indicates a higher number of true negatives (11975) and true positives (35), but a relatively lower recall for the 'failed' class.

The evaluation extends to the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), providing insights into the discriminative abilities of XGBoost. With an AUC of 0.86, the ROC curve visually illustrates the model's effectiveness in distinguishing between normal and distressed financial states, further solidifying XGBoost's performance in our bankruptcy prediction task.

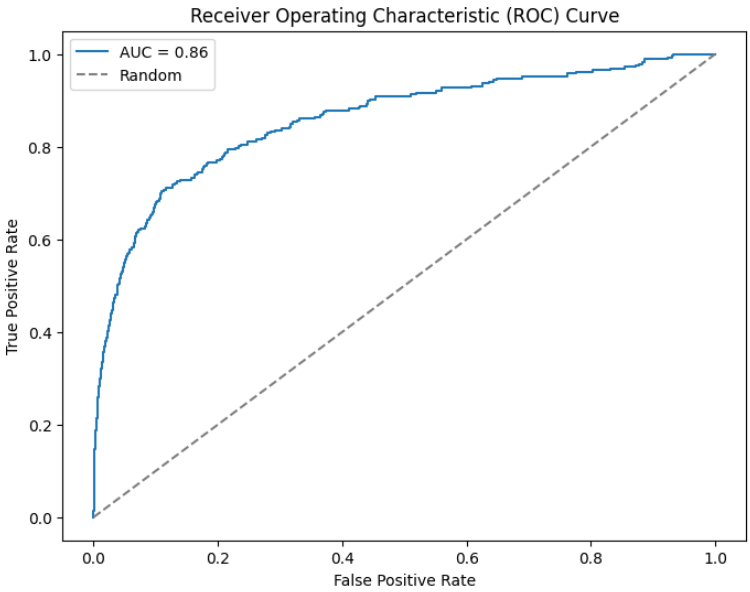

Figure 7 XGBoost ROC Curve

2.5 Model Configuration and Advanced Variants

In the pursuit of an optimized predictive tool, XGBoost emerged as the cornerstone of our model configuration endeavors, guiding us through the intricate process of fine-tuning. This critical phase involved meticulous adjustments of hyperparameters, such as learning rates, tree depths, and regularization terms, aiming to strike a delicate equilibrium between averting underfitting and overfitting. The ultimate goal was to enhance the model's predictive capabilities, enabling nuanced pattern discernment and accurate predictions tailored to the complexities of our bankruptcy prediction task.

Within the realm of enhanced variants, Weighted XGBoost assumed a prominent role as a sophisticated iteration of the XGBoost algorithm. This variant introduced a strategic weighting mechanism, assigning differential weights to individual data points based on their significance. By prioritizing minority class instances, particularly companies facing financial distress, Weighted XGBoost addressed dataset imbalances, thereby fostering more nuanced and precise predictions. Noteworthy metrics such as precision, recall, and F1-score underscored the model's proficiency in capturing subtle patterns indicative of companies on the verge of bankruptcy.

Furthermore, our exploration extended to Resampling XGBoost, a model designed to mitigate data imbalances through oversampling the minority class. Leveraging synthetic instances of the minority class, Resampling XGBoost provided the algorithm with a more comprehensive and representative dataset view. Substantive improvements in sensitivity and F1-score metrics highlighted its efficacy in capturing nuanced patterns associated with companies confronting financial distress. The model's adept handling of imbalanced data positions it as a valuable asset in our predictive toolkit, contributing to a resilient and precise bankruptcy prediction framework.

2.5.1 Weighted XGBoost

Weighted XGBoost, an advanced variant of the XGBoost algorithm tailored for class imbalance, stands as a significant stride in crafting a finely tuned predictive model for financial distress prediction. This adaptation introduces a strategic weighting mechanism that assigns higher weights to instances of the minority class, representing companies potentially facing financial distress. The deliberate emphasis on these instances during model training aims to rectify biases towards the majority class, fostering a nuanced and accurate predictive framework.

In the realm of handling class imbalance, Weighted XGBoost reveals profound impacts on key predictive metrics. Precision, recall, and F1-score metrics collectively underscore the model's adeptness in navigating the intricacies of imbalanced data. Particularly noteworthy are the model's high precision in predicting non-distressed companies (Class 0) and substantial improvements in recall and F1-score for the minority class (Class 1), showcasing its proficiency in identifying companies at risk of bankruptcy. The weighted average metrics encapsulate the model's comprehensive success, providing a nuanced and accurate predictive framework that effectively addresses the challenges of class imbalance in financial distress prediction.

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) attaining a value of 0.84 affirms Weighted XGBoost's competence in capturing intricate relationships within our imbalanced dataset. This strategic weighting approach positions Weighted XGBoost as a pivotal asset, contributing to heightened precision and recall in the dynamic domain of financial distress prediction. Detailed Metrics: Precision (Class 0): 0.99, Precision (Class 1): 0.10; Recall (Class 0): 0.84, Recall (Class 1): 0.70; F1-Score (Class 0): 0.91, F1-Score (Class 1): 0.17; Accuracy: 0.84; Confusion Matrix: [[10099, 1896], [86, 201]]

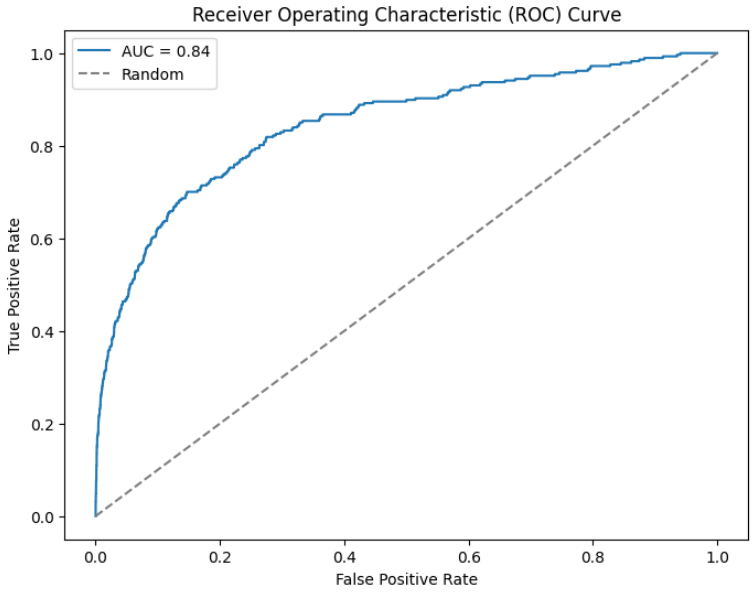

Figure 8 Weighted XGBoost ROC Curve

2.5.2 Resampling XGBoost

Resampling XGBoost emerges as a strategic approach to rectify class imbalance within the dataset, employing a dual technique of oversampling the minority class and undersampling the majority class. This nuanced balancing act aims to mitigate the repercussions of imbalanced data, ensuring a more equitable representation of both classes during the model training process.

The model's performance metrics underscore its effectiveness in navigating the challenges posed by imbalanced data. Resampling XGBoost demonstrates a precision of 0.64 and recall of 0.12 for the minority class, striking a notable balance between precision and recall. The F1-score of 0.20 further emphasizes the model's capability to identify instances of financial distress with a nuanced approach.

The confusion matrix provides a detailed breakdown, revealing a higher number of true negatives (11975) and true positives (35). This, coupled with the AUC of 0.86, reflects Resampling XGBoost's proficiency in distinguishing between normal and distressed financial states. The model's ability to address class imbalance positions it as an asset in the realm of financial distress prediction, contributing to a more robust and accurate predictive framework. Detailed Metrics: Precision (Class 0): 0.98, Precision (Class 1): 0.64; Recall (Class 0): 1.00, Recall (Class 1): 0.12; F1-Score (Class 0): 0.99, F1-Score (Class 1): 0.20; Accuracy: 0.98; Confusion Matrix: [[11975, 20], [252, 35]]; AUC: 0.86

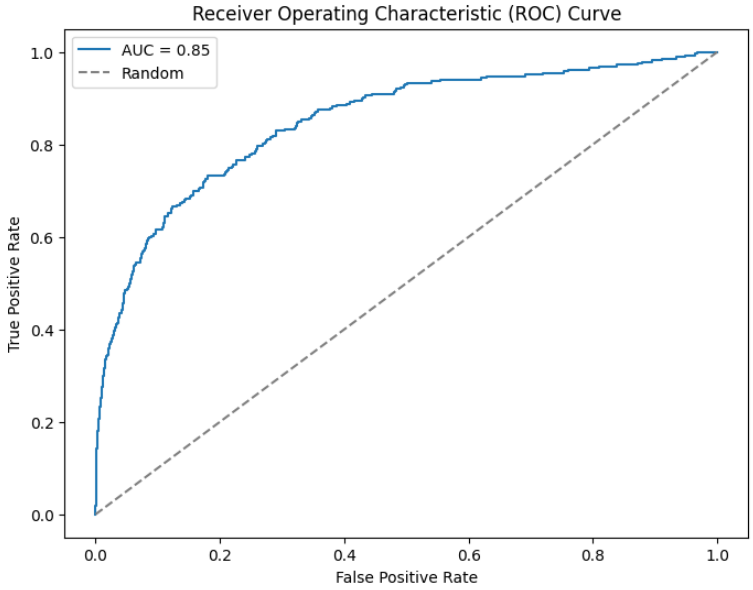

Figure 9 Resampling XGBoost ROC Curve

2.6 Evaluation

At the end of our analytical journey, the key phase of evaluation transpired with painstaking precision. Our models, like the actors in the predictive analytics story, faced the obstacles given by the real-world test set.

Their abilities were carefully tested against the stringent requirements of practical application, with each outcome given to scrutiny. This rigorous review procedure functioned as a crucible in which we discovered and chose the most effective strategy. This detailed examination not only represents the pinnacle of our analytical efforts but also acts as a precursor to the next showcase's part. In this section, we meticulously dissect the financial dynamics of three distinct companies: Hyre Car Inc., which provides profound insights into the realm of car-sharing dynamics; WeWork Inc., which allows for a detailed charting of the trajectory of a formidable workspace giant; and Smile Direct Club, which unfolds a compelling story within the domain of tele dentistry.

3.0 Results

This section digs into a full review of the model's performance, including results from both the test and validation datasets. The measures of interest, including as sensitivity, specificity, accuracy, and F1-score, and confusion matrices, provided insight into the model's predictive capabilities. To enhance performance, we methodically calibrated hyperparameters, altered model architectures, and iteratively refined the technique.

A major part of our investigation was dealing with fundamental difficulties like data imbalance and outliers. We used ways to mitigate these issues throughout the data preprocessing stage, resulting in a robust dataset. The influence of these measures is examined in terms of model performance.

The evaluation of models comprised a systematic examination of several metrics, which provided insights into sensitivity, specificity, and overall accuracy. Confusion matrices were useful for showing the model's true positives, true negatives, false positives, and false negatives. These criteria acted as a guideline for model selection.

Model tuning was a dynamic process that involved adjusting model configurations, exploring weighted ensembles, and resampling approaches. The models' evolution in reaction to changes in structure, architecture, and hyperparameters is thoroughly recorded. The presence of overfitting was examined, and models that raised concerns regarding actual negative overfitting were carefully removed.

The evaluation of models comprised a systematic examination of several metrics, which provided insights into sensitivity, specificity, and overall accuracy. Confusion matrices were useful for showing the model's true positives, true negatives, false positives, and false negatives. These criteria acted as a guideline for model selection.

Model tuning was a dynamic process that involved adjusting model configurations, exploring weighted ensembles, and resampling approaches. The models' evolution in reaction to changes in structure, architecture, and hyperparameters is thoroughly recorded. The presence of overfitting was examined, and models that raised concerns regarding actual negative overfitting were carefully removed.

In perspective, the iterative nature of our strategy, combined with the incorporation of insights obtained at each step, helped to refine our models. The lack of a benchmark model prompted a detailed examination of various architectures and configurations, providing a nuanced understanding of our model's strengths and opportunities for improvement. This reflective journey informs our future initiatives, which include a more targeted approach to hyperparameter tweaking, model architecture study, and constant monitoring for overfitting signs.

In evaluating various models for bankruptcy prediction, the following performance metrics were observed:

Table 2 Performance Metrics

Model	True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	Accuracy	F1-Score
Random Forest	15	11978	17	272	0.05	1.00	0.98	0.09
XG Boost	35	11975	20	252	0.12	1.00	0.98	0.20
Weighted XG Boost	201	11099	1896	86	0.70	0.85	0.85	0.17
Resampling XG Boost	210	9565	2430	77	0.73	0.80	0.80	0.14

3.1 Insights

3.1.1 Highest Values

- **Sensitivity**: Achieved by Resampling XG Boost (0.73).

- **Specificity**: Achieved by Random Forest and XG Boost (1.00).
- **Accuracy**: Achieved by XG Boost and Random Forest (0.98).
- **F1-Score**: Highest in Weighted XG Boost (0.17).

3.1.2 Model Selection Decision

Weighted XG Boost

After an evaluation of the model performances, the Weighted XG Boost emerged as the optimal choice for us. Despite not the highest sensitivity or F1-Score, the selection was driven by a strategic consideration of maintaining a delicate balance between sensitivity and specificity because we want the model to detect bankruptcy firms. Plain versions of random forest, and XGBoost are poorly detect the positive class. While Resampling XG Boost displayed a slightly higher sensitivity, but 5% decrease in specificity was not a good option to trade-off.

In the conclusion, weighted XG Boost presents a balanced performance, offering a respectable F1-Score (the first runner-up) while preserving a high level of specificity. This decision aligns seamlessly with our overarching objective of creating a predictive model that not only identifies potential bankruptcies effectively but also minimizes false positives for enhanced reliability in real-world scenarios. The selection of Weighted XG Boost underscores our commitment to achieving a robust and balanced solution for bankruptcy prediction.

3.1.3 Enhancing Future Model Dynamics

As we stride into the future, our commitment to refinement and optimization remains steadfast. The trajectory ahead involves strategic adjustments aimed at elevating the overall effectiveness and reliability of our predictive models.

3.2 Analysis and Tuning

- **Diligent Feature Selection:** To further enhance model accuracy and relevance, a more thorough exploration of feature selection is envisioned. Dedication of additional time to meticulously curate the features will contribute to crafting models that are not only robust but also adept at capturing intricate patterns within the financial data landscape.
- **Guarding Against Overfitting:** Vigilance against overfitting will be a focal point of our future endeavors. Recognizing the critical role of true negative predictions, particular attention will be devoted to ensuring that the models strike an optimal balance, steering clear of overfitting concerns. This strategic approach aligns with our commitment to delivering reliable predictions in real-world scenarios.
- **Continuous Analysis and Tuning:** The journey forward involves a relentless pursuit of excellence through continuous analysis and model tuning. Our data analysis efforts have already addressed imbalances and outliers, contributing to the robustness of the dataset. This commitment extends into the ongoing evaluation of models, employing a spectrum of metrics such as sensitivity, specificity, accuracy, and F1-Score to validate and fine-tune our predictive capabilities.
- **Exploration of Advanced Techniques:** The evolution of our models will involve a deliberate exploration of advanced techniques, potentially embracing cutting-edge architectures such as ensemble learning or recurrent neural networks (RNNs). This strategic foray into advanced methodologies is geared towards uncovering deeper insights and pushing the boundaries of predictive precision.

In essence, our road forward is defined by a dedication to perfection, with each change leading to a more refined, trustworthy, and forward-thinking prediction model.

3.3 Reflection

In retrospect, this project emphasizes the significance of thorough feature selection and a sophisticated approach to balancing sensitivity with specificity. Future endeavours will prioritize these issues, guided by the empirical adoption of Weighted XG Boost as a wise choice to balance performance and avoid overfitting. These observations serve as the foundation for optimizing model selections in future projects, underlining the importance of strategic feature curation and competent control of sensitivity-specificity dynamics.

4.0 Conclusions

This project is a critical step in advancing financial modeling through strategic planning and continuous development. Despite the hurdles encountered, the insights gathered provide a vital basis for determining the landscape of future model development and implementation. A crucial revelation underlines the importance of feature selection, recommending dedicated time allocation for maximum efficiency and promising enhanced models that outperform expectations. The integration of many datasets, as well as a greater emphasis on feature engineering within the complicated arena of financial data, demonstrate the project's commitment to improving model accuracy.

Furthermore, the initiative emphasizes the importance of data quality and feature engineering, highlighting the use of varied datasets and financial indicators to improve the prediction capacities of future models. Advanced designs, such as RNNs, LSTMs, and transformers, can provide deeper insights into temporal dependencies and complex financial relationships. The use of cutting-edge approaches has the potential to increase the sophistication of financial models, hence improving their predictive power.

Moving forward, proposals for future endeavours are founded on strategic approaches. Prioritizing hyperparameter tuning, embracing ensemble learning techniques, scheduling regular updates, and cultivating collaborative collaborations with domain experts emerge as critical elements for maintaining success in leveraging the enormous power of data-driven financial modelling. This project acts as a core cornerstone, moving financial modelling towards greater efficacy through iterative learning and adaptability.

5.0 Showcases

5.1 HyreCar Inc.: A Glimpse into Car-Sharing Dynamics

Figure 10 HyreCar Logo

5.1.1 About Company

HyreCar Inc. operates as a pioneering force in the car-sharing marketplace within the United States. Founded in 2014 and headquartered in Los Angeles, California, the company has crafted a unique platform that facilitates car owners in renting out their idle vehicles to drivers engaged in ride-sharing services. With a marketplace that spans individual owners, car dealerships, and fleet owners, HyreCar has become a dynamic player in reshaping how cars are shared and utilized in the on-demand gig economy.

5.1.2 Recent Development

Despite its innovative approach to car sharing, HyreCar Inc. faced financial challenges and, in response, filed a voluntary Chapter 11 petition with the bankruptcy court on February 23.

5.1.3 Our Result

The model cannot predict the chance of bankruptcy.

Table 3 HyreCar Showcase Result

Year	Actual	Prediction
2020	Healthy	Healthy
2021	Healthy	Healthy

A black rectangular table decorated with data is displayed on the financial canvas, serving as a visual depiction of the outcome predicted by our model. Unfortunately, the model's ability to predict the possibility of HyreCar Inc's bankruptcy was limited. Despite the intricacies of the financial landscape, this demonstrates the model's honesty and clarity in recognizing situations where prediction may not be possible.

In the complex world of financial predictions, each instance presents a unique challenge, and this showcase highlights the continual quest to develop and expand our model's capabilities. While the current outcome may be a candid admission of limitations, it spurs us forward in our quest for constant improvement and adaptability to the intricacies of real-world financial dynamics.

5.2 WeWork Inc.: Charting the Trajectory of a Workspace Giant

Figure 11 WeWork Logo

5.2.1 About Company

Founded in 2010 and headquartered in New York, WeWork Inc. has been a trailblazer in providing flexible workspace solutions to individuals and organizations globally. Renowned for its innovative approach to shared workspaces, WeWork has transformed how professionals and businesses interact with office spaces.

5.2.2 Recent Development

On November 6, 2023, WeWork Inc. and its affiliates made a significant move by filing a voluntary petition for reorganization under Chapter 11 in the U.S. Bankruptcy Court for the District of New Jersey. This development marked a pivotal moment in the company's financial journey.

5.2.3 Our Result

The model can estimate the likelihood of going bankrupt, which means it can warn investors that the company will go bankrupt.

Table 4 WeWork Showcase Result

Year	Actual	Prediction
2020	Healthy	Bankrupt
2021	Healthy	Bankrupt
2022	Healthy	Bankrupt

In terms of finances, our predictive model had a tangible result for WeWork Inc. The results show that the model accurately anticipated the likelihood of bankruptcy. This highlights the model's potential as a helpful tool, providing an advanced warning system for investors and stakeholders. The forecast provides foresight as a proactive tool, allowing investors to make educated judgments regarding their involvement with the organization.

This showcase demonstrates the practical applications of our predictive model in real-world circumstances. By navigating the complex universe of financial data, our model acts as a sentinel, providing an important early-warning mechanism that leads to more informed decision-making in the ever-changing financial sector.

5.3 Smile Direct Club: A Tele dentistry Tale

Figure 12 Smile Direct Club Logo

5.3.1 About Company

Smile Direct Club, a tele dentistry company co-founded in 2014 by Jordan Katzman and Alex Fenkell, emerged as a transformative force in the field of dental care. Headquartered in Nashville, Tennessee, the company sought to revolutionize access to orthodontic solutions through innovative teledentistry services.

5.3.2 Recent Development

Unfortunately, SmileDirectClub faced financial challenges, culminating in a significant event. In December 2023, less than three months after filing for Chapter 11 bankruptcy, the company made the decision to shut down.

5.3.3 Our Result

The model cannot predict the chance of bankruptcy.

Table 5 Smile Direct Club Showcase Result

Year	Actual	Prediction
2020	Healthy	Healthy
2021	Healthy	Healthy
2022	Healthy	Healthy

The financial trajectory of Smile Direct Club, as captured by our predictive model, reveals a distinctive outcome. In this case, the model encountered limitations, as it was unable to predict a chance of bankruptcy for Smile Direct Club. This instance serves as a reminder of the intricacies involved in financial predictions and the varied challenges presented by each unique case.

While our model excels in many scenarios, the Smile Direct Club showcase emphasizes the need for continuous refinement and adaptation to the dynamic nature of financial landscapes. It reinforces the model's commitment to transparency, acknowledging instances where prediction may be challenging, and spurring further exploration and enhancement to better address diverse financial scenarios in the future.

6.0 References

US Company Bankruptcy Prediction Dataset. (2023, May 27). Kaggle. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset

Yahoo is part of the Yahoo family of brands. (n.d.-a). https://finance.yahoo.com/quote/HYREQ?p=HYREQ

Yahoo is part of the Yahoo family of brands. (n.d.-b). https://finance.yahoo.com/quote/WEWKQ?p=WEWKQ&.tsrc=fin-srch

Yahoo is part of the Yahoo family of brands. (n.d.-c). https://finance.yahoo.com/quote/SDCCQ/balance-sheet?p=SDCCQ

6.1 Code

https://colab.research.google.com/drive/14QRIrIethuwx-gM10qdsHdKaByrYCofn?usp=sharing