NORWEGIAN SCHOOL OF ECONOMICS

FIE453

Big Data with Applications to Finance

Fall 2023

Candidates: 32, 40, 51

Table of Contents

Introduction

In this comprehensive report, our primary objective is to explore the potential of machine learning in predicting corporate bankruptcies within a year, based on current financial health indicators. This subject is interesting since it leads to several different real world practical applications. It could for instance contribute to better strategies in investment realms, such as short selling, as well as elevate the efficiency of portfolio and risk management. To this end, we have designed two distinct machine learning models. The first model adopts a broad, generalized approach to data analysis, while the second delves into industry-specific financial and macroeconomic insights. This comparative method allows us to ascertain if specialized, industry-focused data enhances bankruptcy prediction accuracy compared to a more generalized model. One significant challenge in this study is the inherent nature of bankruptcy, namely it is already a rare event. This rarity creates a substantial imbalance in our dataset from compustat, skewing heavily towards financially stable companies over bankrupt ones. The issue amplifies for the industry-specific model, where data subdivision further narrows the available dataset for model training. Dealing with these issues, we have employed several tactics in the realm of finance and machine learning. In this report we will explain how we used financial data and ratios to analyse model performance. Additionally, we look to textual analysis, and macro data to gather more predictors for our model, while also reducing the number of missing values through unsupervised learning. We use time series cross validation and deploy the SHAP framework to do feature selection, we then use these methods to enhance further testing of our model. In summary, our report tries to dig into the realm of machine learning and financial analysis, and as we get further into the report we will elaborate more on our methodology and our findings.

Variables

Definition of bankruptcy

While the exact legal definition of a bankruptcy is hard to grasp and might change depending on where the company is registered, the general concept is easily defined. A bankrupt company is unable to pay what it owes, and have had control of their financial matters given, by a law court, to a person who sells your property to pay your debts. (Camebridge dictionary, 2023)

Due to finding it hard to find good data on bankruptcy, we decided to define our own bankruptcy variable based on parsed 8K's filed to the SEC, specifically 8K's including item 1.03. Form 8-K is known as a "current report" and must be filed by public companies to inform shareholders of major events shareholders should know about. (SEC, 2023). Item 1.03 is one of these events and must be filed by every company that goes into bankruptcy or receivership. Whilst receiverships are different from bankruptcies, we find that it fits under the scope of the broad bankrupt definition above. Lastly, to increase practical application, we decided to lag the observations by a year. Our final dataset defines a company as bankrupt if they filed an 8-k including item 1.03 in the next fiscal year. The final distribution of bankruptcies is shown in the table to the right.

Year	Bankruptcies
2000	0
2001	0
2002	0
2003	23
2004	49
2005	39
2006	34
2007	57
2008	95
2009	59
2010	46
2011	42
2012	34
2013	32
2014	34
2015	57
2016	36
2017	29
2018	45
2019	68
2020	20

Figure 1- Bankruptcy distribution in the data

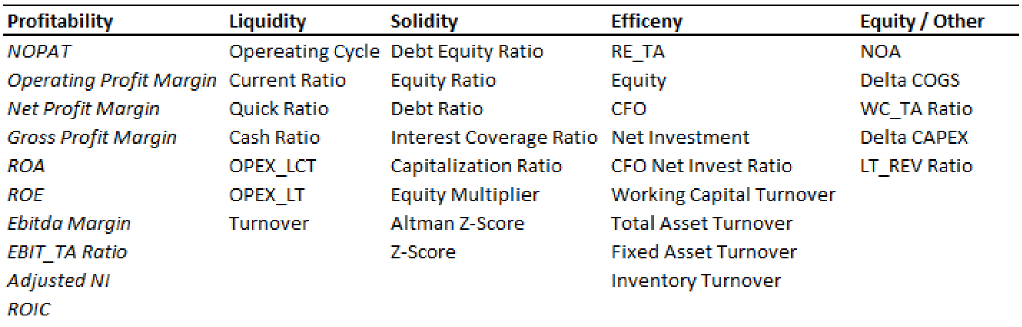
Financial Variables

Like we introduced in the introduction, we are trying to use accounting data from compustat to predict whether a firm goes bankrupt or not. To make sure that the model can grasp the full extent of the data available, implementing and creating ratios for comparing, analyzing and understand the data is quite crucial before we introduce it to the machine learning model.

When conducting any type of financial analysis of a firm today, one would always tend to use ratios as a measure of gauging the performance of a certain firm. The use of ratios goes far back in the literature, and the common theme is that ratios manage to catch the snapshot of the current situation of a company's, health when considering factors such as, liquidity, solidity, efficiency, and profitability. Thus, it has become the norm when conducting fundamental analysis, but also provide easy access to information to any shareholder that wants to invest on the stock market. (Nadar, 2019)

Ratios for financial data is widely used due to its simplicity, not just in terms of computing, but also to compare performance across years within each firm as a management and control tool. This enables firm to track their performance and their trends, giving decision makers ample data to perform necessary adjustments to increase performance. Furthermore, ratios play a pivotal part in being the best benchmarking tool, enabling the comparison of performance across firms in a certain industry, or across the entire market. For machine learning purposes, ratios serve as a normalization tool. This insight is crucial, because ML can be prone to large outliers, and by using ratios one can reduce the amount of noise when training the model. Furthermore, ratios can better explain certain relationships when trying to predict an outcome due to its inherent relationship with the data. Lastly, for our purpose of building a model that wants to predict bankruptcy, ratios serve as a great tool for assessing the risk of a company. The risk of the company can't be drawn out only by ratios, but it provides a great foundation for further analysis when assessing the snapshot of their financial health.

In the table below we have listed all the financial ratios calculated using the compustat dataset. In older studies like (Altman, 1968) and (Zmijewski, 1984), They include traditional operational and management and control ratios, typically used in the 20th century. The development of the Altman Z scores and Zmijewski score for predicting bankruptcy, considers traditional financial ratios and are reliant on the use of a correctly calculated asset class. In our model we have tried to develop more modern and new ratios as well as using the empirically proven ratios. Since we are predicting bankruptcies, including two notable bankruptcy predictors were adequate for the purpose of the report. Both the Altman Z- Score and Z-score are proven in the literature to give insight into whether a firm will go bankrupt given the results of the scores themselves. As introduced earlier, these scores are closely related to the asset class, being derived from; Working capital, Return on assets, total assets, debt ratio, current assets, and current liabilities. While these models are notable and proven predictors, these studies were conducted when accounting standards where different, meaning that firms structured their capital differently than today. With changing accounting standards and firms adapting by using assets and leverage differently these models become simpler than what the reality represents. The financial analysis practices have changed a lot since Altman and Zmijewski, and some of the data sourced from the compustat restricted the use of some of the more modern solutions. However, we would argue that many of the ratios used in our model still has relevance in today's financial landscape. To highlight some of the key ratios we have ROIC, Altman Z score, NOA, and interest coverage ratio.



Profitability	Liquidity	Solidity	Efficeny	Equity / Other
NOPAT	Opereating Cycle	Debt Equity Ratio	RE_TA	NOA
Operating Profit Margin	Current Ratio	Equity Ratio	Equity	Delta COGS
Net Profit Margin	Quick Ratio	Debt Ratio	CFO	WC_TA Ratio
Gross Profit Margin	Cash Ratio	Interest Coverage Ratio	Net Investment	Delta CAPEX
ROA	OPEX_LCT	Capitalization Ratio	CFO Net Invest Ratio	LT_REV Ratio
ROE	OPEX_LT	Equity Multiplier	Working Capital Turnover	
Ebitda Margin	Turnover	Altman Z-Score	Total Asset Turnover	
EBIT_TA Ratio		Z-Score	Fixed Asset Turnover	
Adjusted NI			Inventory Turnover	
ROIC				

Figure 2: Overview of Ratios used.

When dealing with large datasets, there are often large amounts of missing data present. In the compustat dataset, we experienced this for multiple data entries. To deal with the missing values, we decided to create synthetic observations from the data. These synthetic observations significantly reduced the number of missing values we had for certain ratios. However, we are aware that this impact the data, but we would argue that this adjustment is not enough to skew the results of model significantly.

Textual Variables

Textual analysis has seen an increased use in finance application over the last decade, and one of the more widespread uses is using word counts and dictionaries to catch sentiment in a document. One of the more established articles in this field was published by Tim Loughran and Bill McDonald in 2011. (Loghran & McDonald, 2011) In this study they created their own dictionaries to catch different sentiments in annual reports. In their study they found that their wordlists were correlated with market reactions around the 10-K filing date. They also found that some of the wordlists were related to firms accused of accounting fraud and to firms reporting material weaknesses in their accounting controls.

Similar studies have been conducted to explore the impact of textual data when predicting financial distress. A study published in the journal forecasting found that the inclusion of textual data can have a significant role in increasing predictive performance by supplementing the traditional financial features. (Tang et al. 2020) By using the Loughran-McDonald dictionaries to implement sentiment scores to our data, we hope to see a similar increase in predictive performance.

The following sentiment variables have been included in our model: Negative, Positive, Litigious, Constraining, Strong Modal, Weak Modal. These are presented as a percentage of the overall words in the annual report. While this is not the most accurate way of calculating sentiment, we find deeper exploration to be beyond the scope of this report.

Macro Variables

Bankruptcies among firms can be attributed to an array of causes, including macroeconomic factors like changing market conditions, competitive pressures, inflation rates, and credit availability, as posited by du Jardin (2009). To understand the business environment affecting a firm's health, analyzing these macro conditions is crucial. Our predictive model incorporates these macroeconomic variables, revealing a predominant trend of bankruptcies in the U.S. occurring among unlisted companies. Approximately 4,144 companies were publicly traded in the U.S. in 2020, but there may also be an estimated 27 million unlisted companies, according to Dobridge, John, & Palazzo (2022), and Biery (2013). The number of bankruptcies included in our model represent a mere 0.02% of the U.S. total for 2020 out of 89642 total bankruptcies the same year, a rate similar across all modeled years. The limited size of our dataset might introduce some biases or distortions in our model's outputs, which is a concern when working with samples that may not fully represent the broader population.

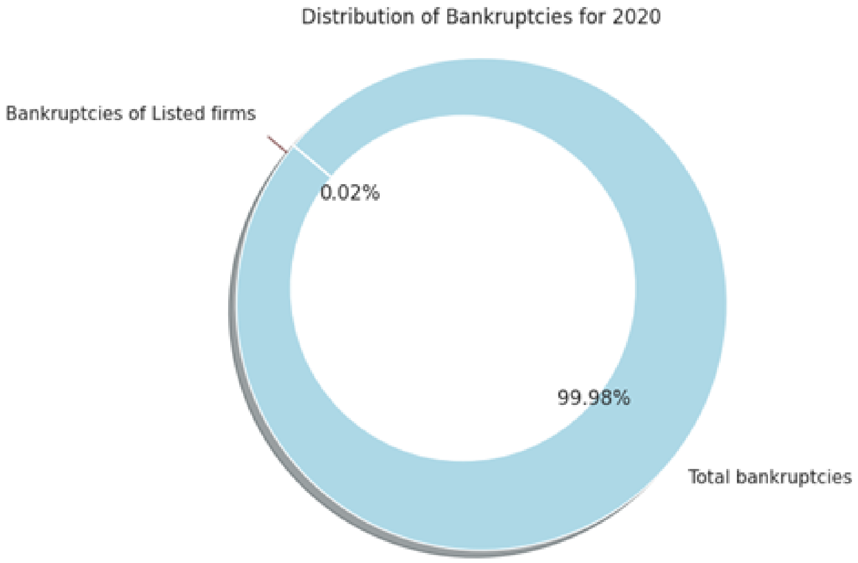


Figure 3: Distribution of Bankruptcies for 2020

Correlation between the macro variables

The correlation table on the right shows the macro variables displays the macro variables in our model, except for S&P 500 (SP_Close). The correlation between unemployment rates and total bankruptcies in the table is calculated at 85%, significantly supporting the theory that unemployment affects domestic demand (Jacobsen & Kloster, 2005).

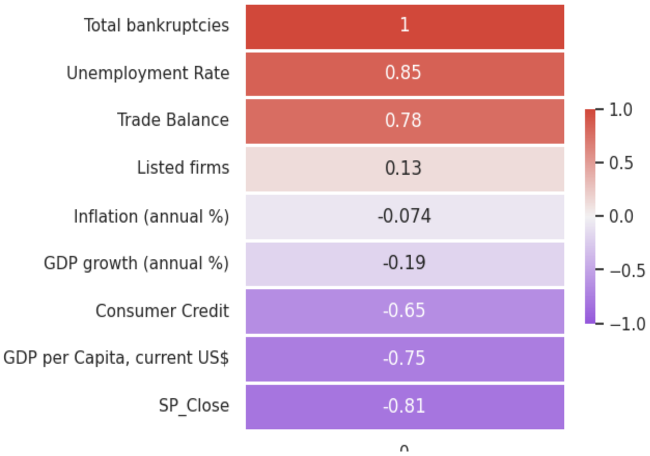


Figure 4: Correlation between the macro variables

Conversely, the S&P 500 index (SP_close) demonstrates a strong negative correlation, suggesting that when the stock market is doing poorly, which the S&P 500 index reflects, it might be due to or result in economic downturns (recessionary signals) or be a response to rising unemployment rates.

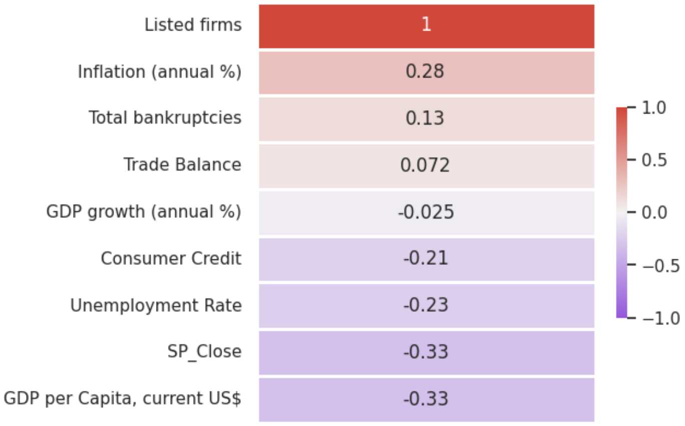


Figure 5: Correlation between the macro variables

Figure 5 on the right is a correlation table of listed bankrupt companies correlated with the same macro variables. Interestingly, the second table shows that inflation has a notable correlation at 28 %, despite a low correlation in the first table, figure 4. Furthermore, the negative correlation between unemployment rates and bankruptcies in listed firms in figure 5 contradicts established theory, calling for cautious interpretation of the data for listed bankruptcies. Moving forward, the relationship between listed firm bankruptcies and macroeconomic factors warrants careful consideration due to the inconsistent results observed.

To provide a more nuanced analysis of U.S. bankruptcy trends from 2003 to 2020, we present two histograms below. The light blue histogram tracks total bankruptcies, peaking notably during the post-2008 financial crisis period, which saw heightened unemployment and reduced consumer spending, as suggested by figure 1 in the correlation analysis above.

The dark blue histogram, illustrating bankruptcies of listed firms, demonstrates greater volatility, potentially due to the smaller sample size. A significant uptick in bankruptcies in 2015 may correspond with the oil price slump driven by U.S. shale oil production surges and OPEC's strategy not to cut output despite low global demand (Stocker, Baffes og Vorisek, 2018). This drop in oil price likely contributed to bankruptcies in the energy sector.

The two graphs show that larger public companies often handle tough times like economic downturns and other challenges better than smaller private ones. This is likely because they have more regulations to follow, which may help them prepare and manage these difficulties. Smaller companies often make up the majority of bankruptcies, possibly because they have fewer resources to deal with hard times.

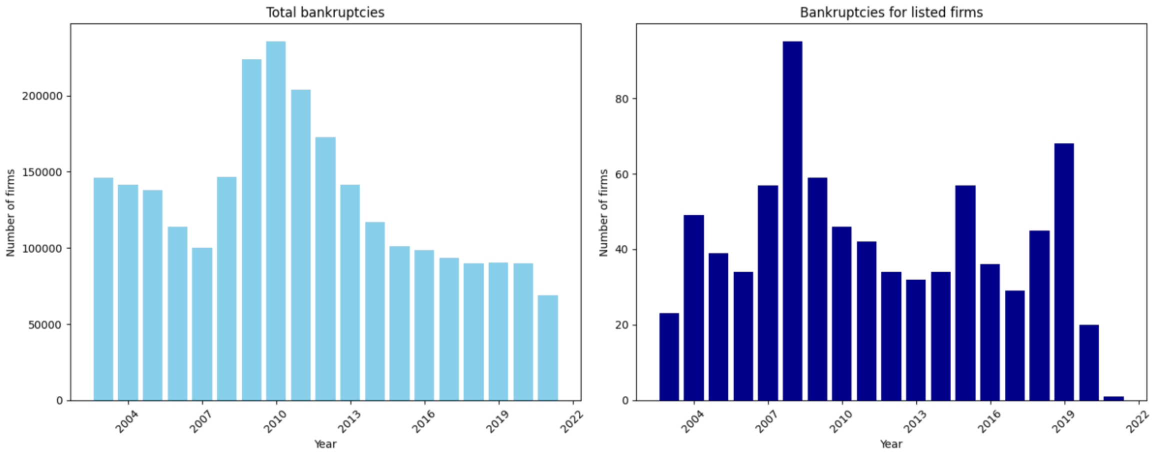


Figure 6: Total Bankruptcies vs Listed Bankruptcies

Model Design

We have chosen XGBoost as the machine learning algorithm for all models developed in this report. The reasoning behind this choice is a study benchmarking different machine learning models for bankruptcy prediction. (Alanis et al, 2022) The results of these benchmarks concluded with XGBoost being the best performing algorithm. Performing our own analysis on which algorithm performs the best is considered beyond the scope of this report.

Data Preprocessing

When analyzing big datasets, it is important to first process the data before one does any form of analysis on the data itself. There are several reasons as to why, and in machine learning specifically it is critical that no information leakage happens between the train, test, validation sets to ensure that the model is consistent.

Firstly, we removed column and rows with more than 30% of the values being Nas from the dataset. Secondly, we filtered the dataset to be within the years 2003 and 2020. This was done due to either lacking or incomplete data on bankruptcies and variables used that can be seen in table 3. Furthermore, to deal with the remaining Na values from the initial 30% split, we used the unsupervised learning algorithm K-Nearest Neighbors.

K-nearest Neighbors or KNN is a non-parametric classification method that is often used in machine learning, because of its intuitive and straightforward approach with many practical applications (Gongde Guo et al.,2004). KNN tries to classify an object, that is not known to the algorithm at the start. By using the objects nearest neighbors, it classifies the object to the majority class it finds closets to the object. This selection process is done by choosing a distance for the algorithm to branch out or search for. Choice of distance can lead to different results of the model, and due to its basic structure, increasing the amount of distance will improve its prediction strength but increase its computational power need (Gongde Guo et al.,2004). Consequently, one must sometimes choose between computational efficiency and prediction strength. However, there is a risk of pushing the distance to far leading to information loss, this implies spending time on tuning the k, can be beneficial when dealing with large dataset. This is done to not miss out on valuable details and be protected against imbalance in the data introducing biases. In table X we can see how the algorithm is intended to work. It locates its neighbors and assigns weights points to the observations by calculating the distance from the object to the closest observations. Meaning, that observations that are further out will be assigned a smaller score, and observations that are located more closely to the object. When, this is done it checks and verifies that the majority class around the object indeed is correct and outputs the value it believes to be the closest fit. This illustration is a stylistics example of how it is intended to work.

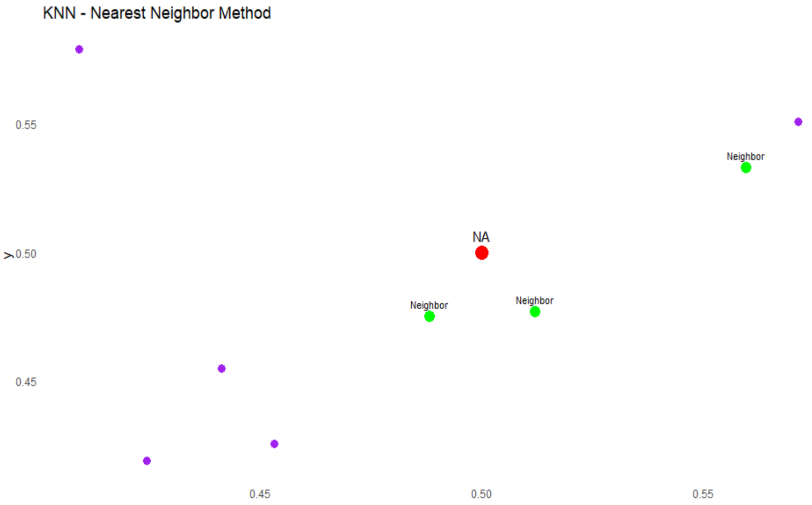


Figure 7 - KNN Impute visualization

The final dataset included a total of 92 346 observations, with a total of 57 variables including ratios, textual and macro data.

Imbalanced Data

In machine learning classification problems imbalanced classes are a common phenomenon. By the very nature of bankruptcies, one would expect the sample of non-bankrupt companies to be significantly larger than the sample of bankrupt companies. This imbalance is even more apparent when analyzing publicly traded companies as discussed in the macroeconomic section above.

In the case of very imbalanced data this will have implications for the model results. Due to the minimal occurrence of the minor class, the model will skew towards predicting the majority class simply because it has a higher probability of being right.

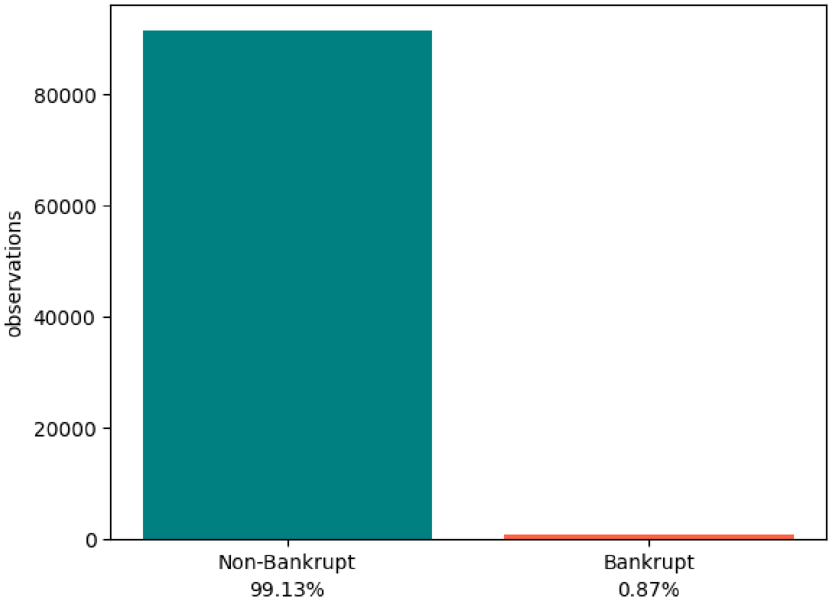


Figure 8: Illustration of the Imbalanced Data

As you can see from figure 8, we have a very imbalanced dataset with bankruptcies representing <1% of the total observations. In theory this makes the model 99 % accurate by predicting 0 bankruptcies, which is problematic.

The literature has come up with many techniques to tackle the imbalanced data problem. For our model we have chosen to use the synthetic minority oversampling technique (SMOTE) first introduced by Chawla. (Chawla et al. 2002) SMOTE creates synthetic observations of the minority class by filling in the segments between two existing observations until the desired distribution is achieved. We have oversampled bankruptcies to represent 15% of the observations in our training sets by using a variation of SMOTE called SMOTE-NC. SMOTE-NC is a variation of SMOTE that allows for both continuous and categorical variables in the samples. This technique is only used during the training of the model, and not for validation or testing. Below is a visualization of data before and after SMOTE is implemented.

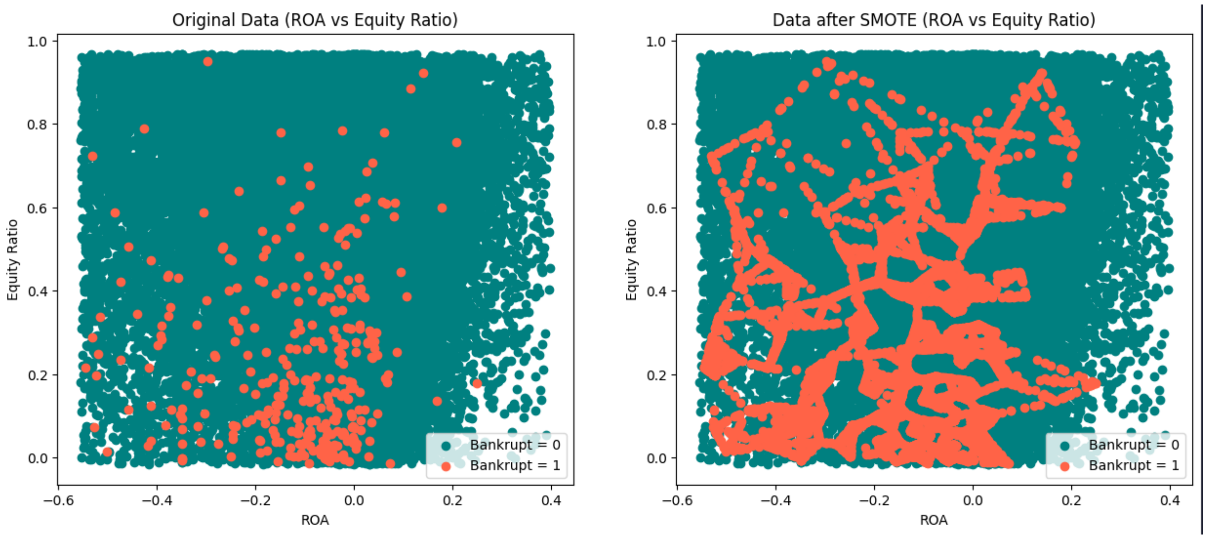


Figure 9: Illustration of pre and post - SMOTE

Data Split

In the field of machine learning it is considered best practice to split the data into separate training and testing sets. The idea is that the training set is used to train and tune the model, which is then tested on the test set. This is to get a realistic metric on how the model performs on unseen data. However, simply training the data on a single training set usually introduces strong overfitting in the model, making it less able to predict unseen data. A way to eliminate this bias is to split the training set into k-folds and train the model k times using a new fold as the validation set each time as shown in figure 10 below.

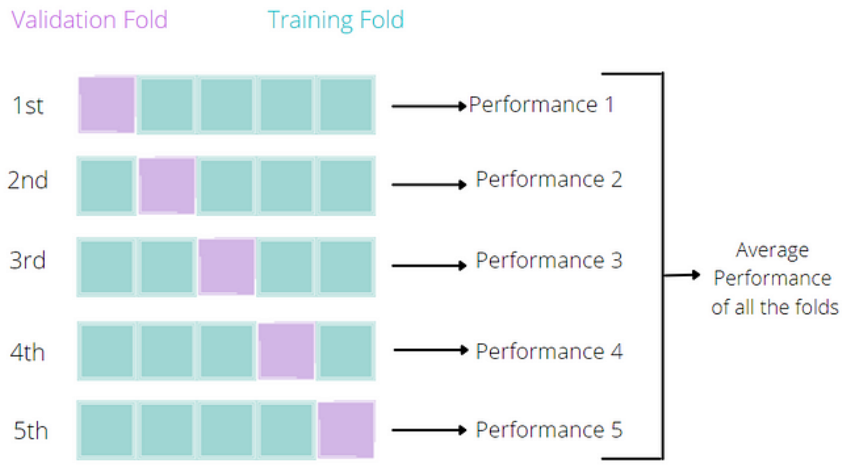


Figure 10: Illustration of the data split

On timeseries data however, where data close in time are highly correlated, the k-fold approach would result in unreasonable correlation between training and testing sets. Therefore, we apply a timeseries split to our model. This technique also splits the data into k folds. However, it now adds new data for each fold making the new fold a "superset" of the previous. This helps prevent data leakage, where future observations are used to predict the past. The timeseries split of our data is visualized in figure 11 below.

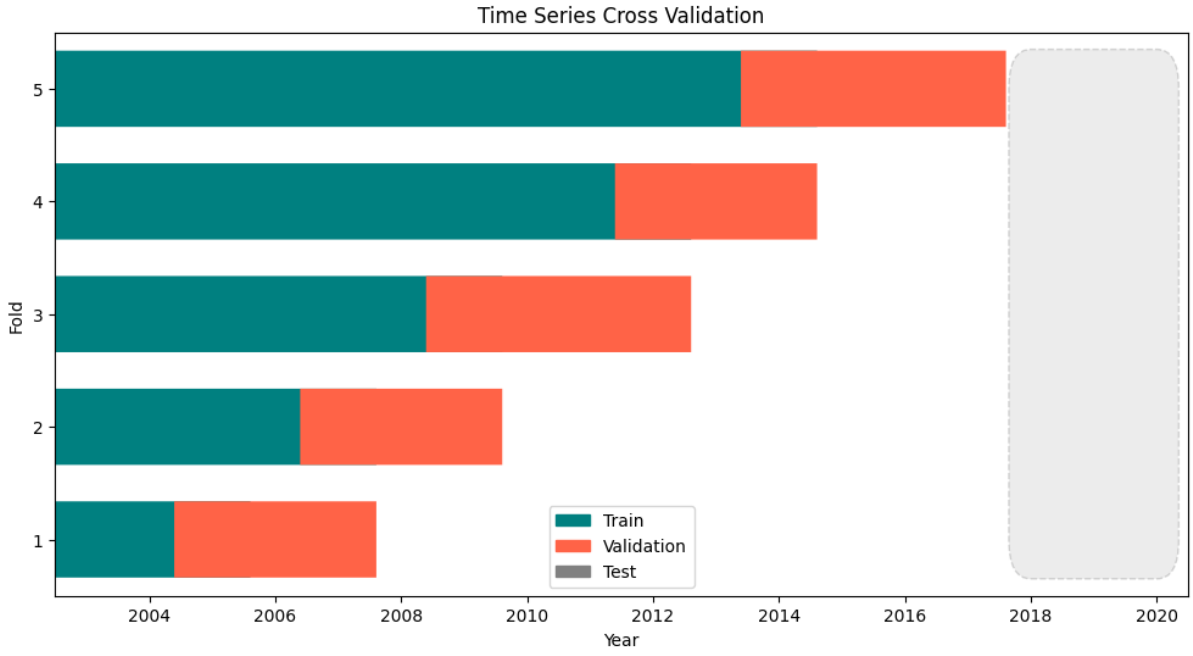


Figure 11: Time series cross validation

Model evaluation

To evaluate and compare the predictive performance of our model's, we will use a collection of metrics measuring the model's performance on the test set. One must also consider the weighting between the classes when measuring a model's performance. This is typically done by either prioritizing recall (True positive rate) over precision (False positive rate) or vice versa. To decide which one to weight more one must decide which type of error would be most costly. If type1 errors are less desired one would typically give more weight to precision, and if type2 errors is less desired it calls for a prioritization of the Recall score. Given the nature of the report, where we have limited our model to include publicly traded companies, this weighting would depend on how one intends to use the forecast information. If we were to use it to implement a shorting strategy, one would typically weight precision higher, as the potential downside incase of a false positive is infinite. However, if it is intended as an early warning sign for companies in your portfolio, one would typically weight recall higher. For this report, we have decided to put a higher weight on recall when constructing our models.

In this section we will briefly introduce some of the metrics we will use to measure the performance of our models: F2, AUC and Brier Score.

F2

F2 score is a metric not commonly used in the literature but is essentially a modification of the more widely used F1 score. Where F1 score is a metric that gives equal weight to recall and precision, F2 score is a modification of the formula to put more weight on the Recall-score when evaluating models.

$$F1 = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} \qquad F2 = \frac{3}{\frac{2}{Recall}+\frac{1}{Precision}}$$

Figure 12: Formula for calcualting F1, F2 Score

Due to our choice of weighting recall higher than precision, we find this to be an excellent measure of performance. Therefore, we have used this as the scoring measure when hyper-parameter tuning our models.

Brier Score

While the F2 score is strictly measured by the final classifications of the models, brier score is a measure to capture the quality of the predictions. Every prediction is assigned a probability, where >50% probability is classified as positive(bankrupt). The brier score is a measure of the mean squared error of the differences between the predictions and the actual outcome.

$$Brier\ Score = \frac{1}{N}\sum_{t=1}(f_t - o_t)^2$$

AUC

A scoring metric that is commonly used to measure performance in the literature is the AUC score. This is a metric that measures the tradeoff between the true positive rate and the false positive rate. Because of its ability to measure accuracy well, even in imbalanced data predictions, it will provide valuable insights to the performance of our models.

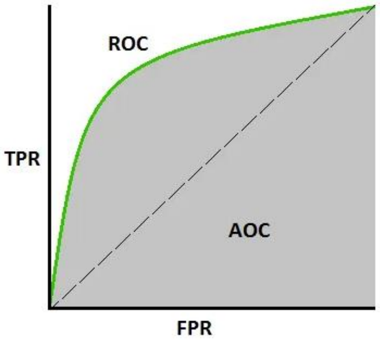


Figure 13: Auc score illustration

Feature Importance

To further explain the predictions of our model, we decided to go with the SHAP framework to be able to analyze our features and the impact of them and use this framework to analyze whether testing best performing feature influences model performance.

SHAP values is a method that uses the framework of game theory, where there are as many players as there are features. Specifically, it uses cooperative game theory, and in that extent the shapely value. This value is made to ensure that each participant in the cooperative game is attributed a fair value. All features and observations will be seen thus, each value will be assigned the fair value proportional to their total contribution of the game (Shapley, 1951). Meaning, that for each single observation, SHAP will give that observation a value of how much it contributes to the output of the model. However, the value it gives says nothing about the quality of the prediction itself, only how much that single observation contributes. Additional, SHAP values are efficiently approximated and calculated in gradient boosting models like XG boost, due to it being able to approximate the shapely value through its tree structure. This aligns well, with the properties of cooperative game theory, and approximation of the shapely value (Lundberg & Lee., 2017) and makes for a great use case in machine learning. Furthermore, the SHAP framework helps in solving one of the many issues with machine learning, namely the black box problem. This black box problem stems from the complexity of large tree structures; therefore, it is hard to gauge what is impacting the decisions that leads to the output. Like we stated earlier, SHAP attributes a value to each feature based on the change in model prediction on that specific feature. This Value explains the change the feature has, when it was observed when the model knew nothing about the features used (Lundberg & Lee., 2017)

As seen in table X: each point in the bee swarm diagram is considered a single observation. Here we can analyze the features that have the most contribution to the prediction results; Notably, return on assets, equity, and adjusted net income, are features that scores the highest in this model output. The diagram arranges the features such that the features with the most effect is ranked highest, and each step-down shows less and less contribution to the model. Additionally, the X- axis shows the SHAP value for each data point, and the color on the Y axis shows the value of its contribution.

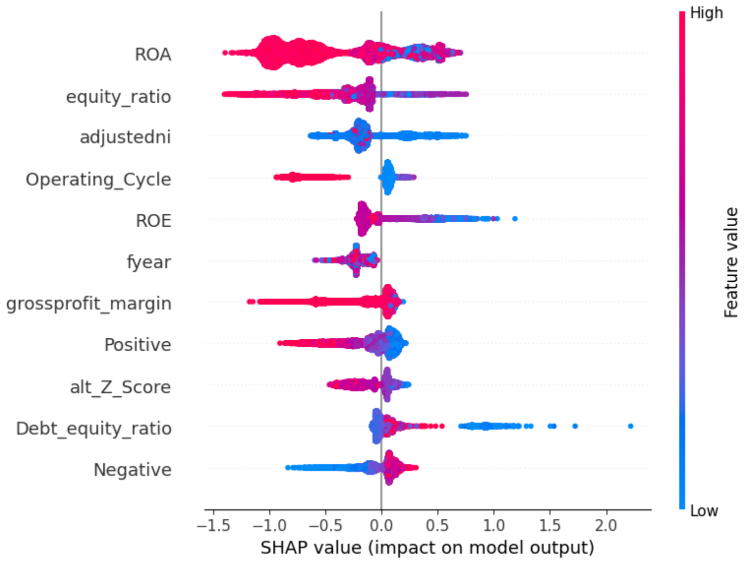


Figure 14: Beeswarm diagram of the SHAP results

General Model

The first model constructed will be a generalized model aimed at predicting bankruptcies across all industries. Using the model methodology explained in the previous sections we have constructed five different models for this purpose. The models are trained on the same observations but differs in the features included in the model. We start at a basis model with only financial ratio features, before training new models after adding text and macro data.

Model	AUC	Precision	Recall	F2	Brier Score
Model with financial ratios	0,8826	0,1578	0,4088	0,3101	0,0180
Model with textual data	0,8875	0,2067	0,3899	0,3312	0,0154
Model with macro data	0,9026	0,1607	0,4528	0,3321	0,0198
SHAP-40	0,8997	0,1414	0,4403	0,3095	0,0216
SHAP-20	0,8974	0,1534	0,3522	0,2797	0,0185

Figure 15: Results for the general model

As you can see from the results above, we saw an improvement at each stage after adding features to the model. Because of the large number of features in the model is derived from company specific data, it is reasonable to expect high multicollinearity between observations. This might produce unwanted noise in the model, and a subset of the most important features might improve the results of the model. We therefore trained two new models limiting the features to the top 40 and top 20 features measured in mean absolute SHAP values. We saw no improvement in our model by limiting the features indicating that the model makes use of most, if not all the features we have included. Overall, we find that the model including all the features provides the best results and will be the one used for the final comparison.

Industry specific model

To enhance our analysis, we aim to compare our general model with a specific industry model, determining which model best represents our data. This section will focus on selecting an industry based on the Fama French 48 industrial classification (FFInd) for in-depth examination. Key variables particularly relevant to the chosen industry will be identified to ensure a customized and more precise modeling approach.

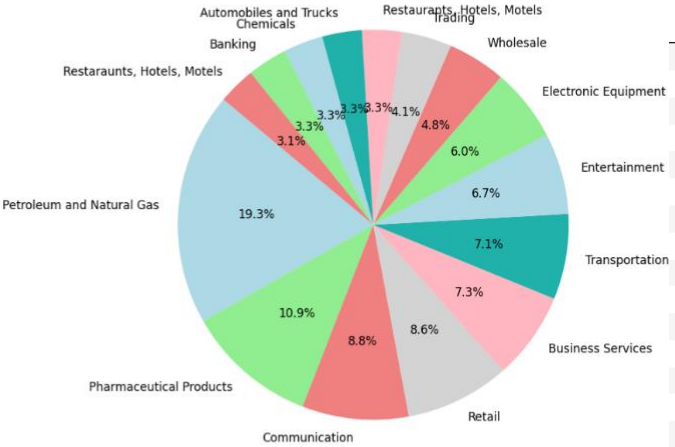
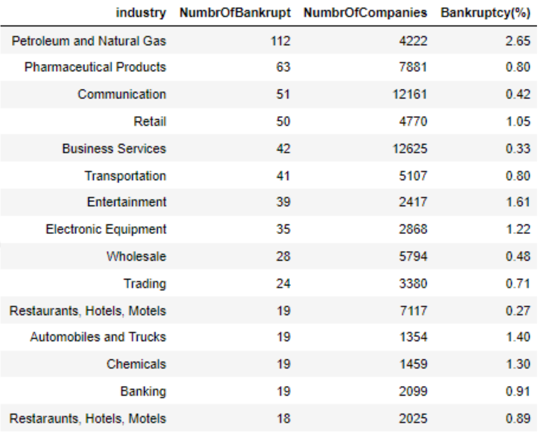


Figure 16: Distribution of industry



industry	NumbrOfBankrupt	NumbrOfCompanies	Bankruptcy(%)
Petroleum and Natural Gas	112	4222	2.65
Pharmaceutical Products	63	7881	0.80
Communication	51	12161	0.42
Retail	50	4770	1.05
Business Services	42	12625	0.33
Transportation	41	5107	0.80
Entertainment	39	2417	1.61
Electronic Equipment	35	2868	1.22
Wholesale	28	5794	0.48
Trading	24	3380	0.71
Restaurants, Hotels, Motels	19	7117	0.27
Automobiles and Trucks	19	1354	1.40
Chemicals	19	1459	1.30
Banking	19	2099	0.91
Restaraunts, Hotels, Motels	18	2025	0.89

Figure 17: List of all Bankruptcies by industry

The analysis revealed that the petroleum and gas industry with a rate at 2,6% have the highest rate of bankruptcies compared with other sectors in the dataset, as illustrated in figure 16 and 17 above. Therefore, to gain a deeper understanding, the report will focus on this sector and compare it with our general bankruptcy machine learning model. Figure 18 below illustrates the trend of bankruptcies in the petroleum and gas industry from 2007 to 2020.

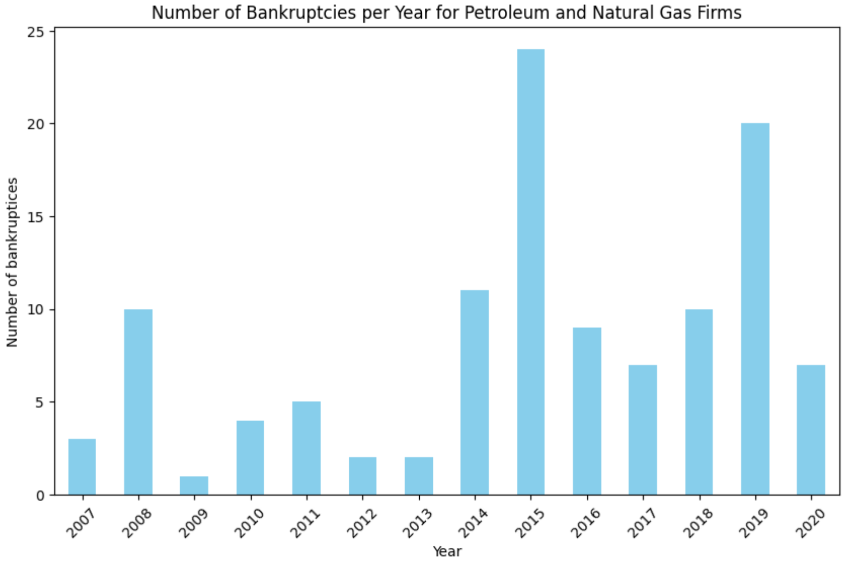


Figure 18: Number of bankrupcises per year for Petroleum and Natural gas firms

Introducing new variables to the industry-Specific Machine Learning Model

Key variables unique to the industry have been incorporated to construct a predictive model tailored for this sector. These include the average yearly West Texas Intermediate (WTI) oil price, oil rig costs, and the annualized volatility in WTI prices. Other potential variables relevant to the petroleum sector, such as lifting cost, break-even analysis, exploration expenses, legal disputes, and varying categories of oil reserves (proven, probable, and possible) has been considered but not included here due to data accessibility and the scope of this assessment.

Bankruptcy Trends Across Industries

The histogram below presents a comparative analysis of bankruptcies across various industries. The four sectors with the highest bankruptcy rates are highlighted: Petroleum and Gas, Pharmaceuticals, Communication, and Retail, giving them distinct colors for a more transparent trend analysis, while other industries are represented in a uniform grey shade. This visualization underscores the heightened volatility of bankruptcies in the Petroleum and Gas sector, particularly during specific periods like the 2007-2009 financial crisis and the 2014-2016 oil crisis, with a peak in 2015 (data from the pandemic is not included). The financial crisis seems to have influenced many industries, not only the oil and gas sector. However, the oil-crises appear to be uniquely influenced by the oil and gas industry, unlike other industries.

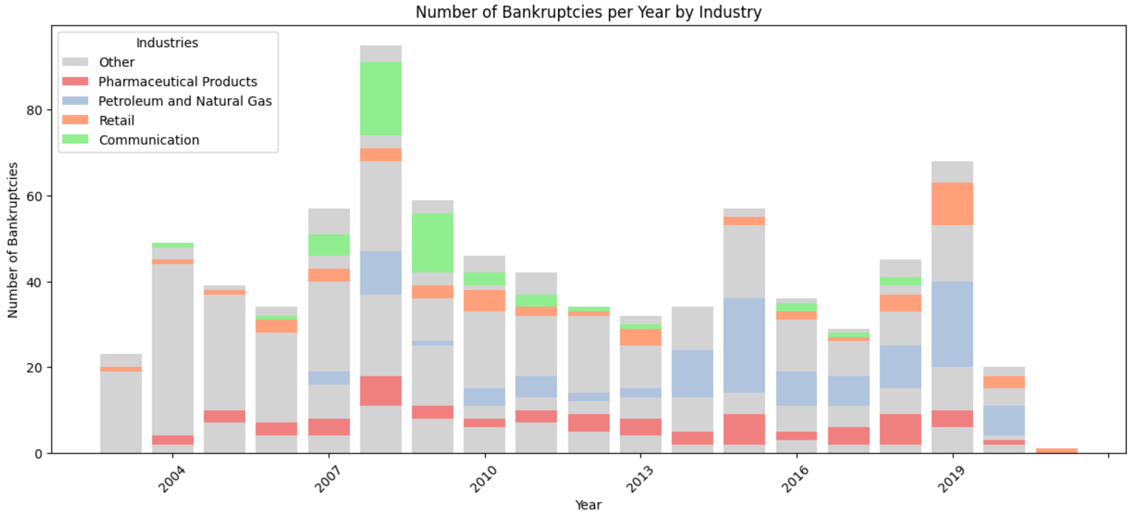
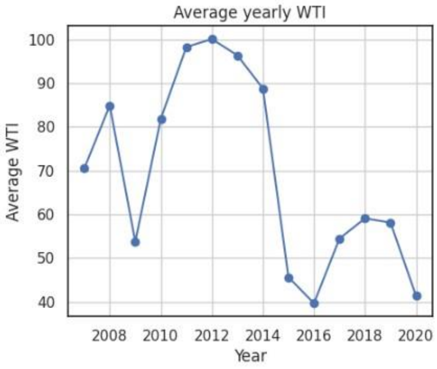


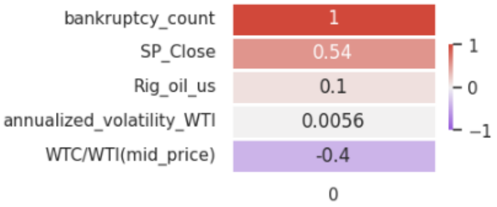
Figure 19: Number of bankruptxies per year by indsutry

Note that some companies in the Petroleum and Natural Gas sector filed for bankruptcies between 2000 and 2007, but these are not included in the analysis due to a lack of corresponding data in Compustat.

Industry specific variables

The figure below illustrates the fluctuation of WTI oil prices over time. This indicator is very significant to the oil and gas industry, as it directly impacts the revenues and profitability of companies operating within this sector. As shown in Figure below, the correlation between WTI oil prices and bankruptcy rates is -0.40, suggesting that a decrease in oil prices is associated with an increase in bankruptcies. Interestingly, there is almost no correlation between annualized volatility and bankruptcy rates. Furthermore, the correlation data indicates a bias towards the S&P 500, an increase in bankruptcies tends to coincide with an increase in the S&P 500, indicating that companies in other sectors benefits from lower energy prices, thus a positive correlation with the share index.





The observation in the correlation figure above may indicate trouble specific in the oil and gas sector, but not necessarily for the stock market as a whole. To gain a broader understanding of the phenomenon, we will examine some financial variables relevant to this industry.

Some of the financial variables and trends

In the analysis of bankruptcy consequences, key financial indicators from companies' balance sheets and income statements have been assessed. This report highlights one crucial variable from each financial statement, believed to be indicative of a company's health and robustness. Return on Invested Capital (ROIC) has also been included as a vital metric to connect the income statement and balance sheet, reflecting the efficiency of capital usage.

Debt-to-equity ratio

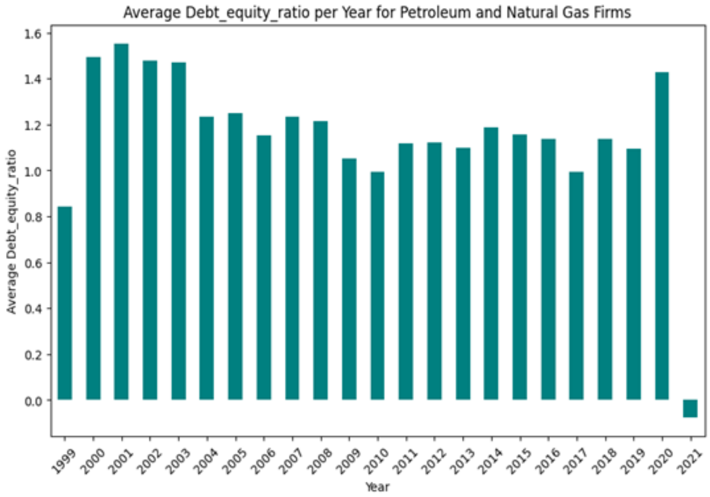
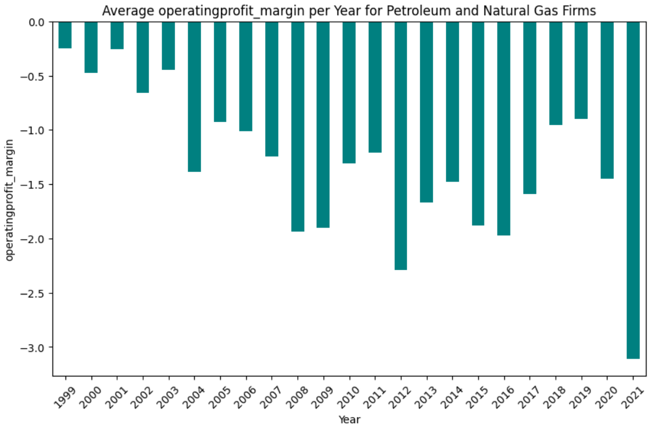


Figure 22: Avereage Debt/equity ratios per year for Petroleum and Natrual gas firms

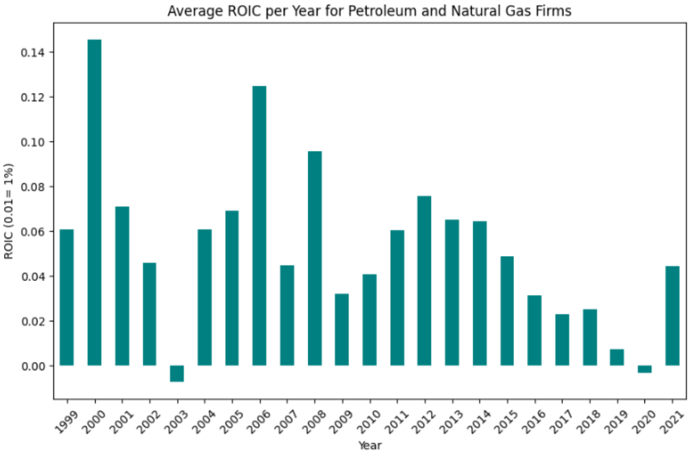
The analysis of the debt-to-equity ratio in Figure 22, adjusted for extreme values (5th and 95th percentiles), revealed a relatively stable average over time. This suggests that companies' leverage levels have not increased significantly. A comprehensive understanding of the substance of the equity calls for a further investigation into the nature of the debt (operational vs. financial) to separate between operational debt and interest-bearing debt, and the composition of equity (tangible vs. intangible). However, this is not included here due to limitation of data.

Operational Margin Trends



The trend of average operational margins in the oil and gas sector indicates a progressive decline in profitability over time. Notably, the margins dipped further during the financial and oil crises, only to recover post-2017, before plummeting again during the 2020-2021 pandemic period.

ROIC and Industry Volatility

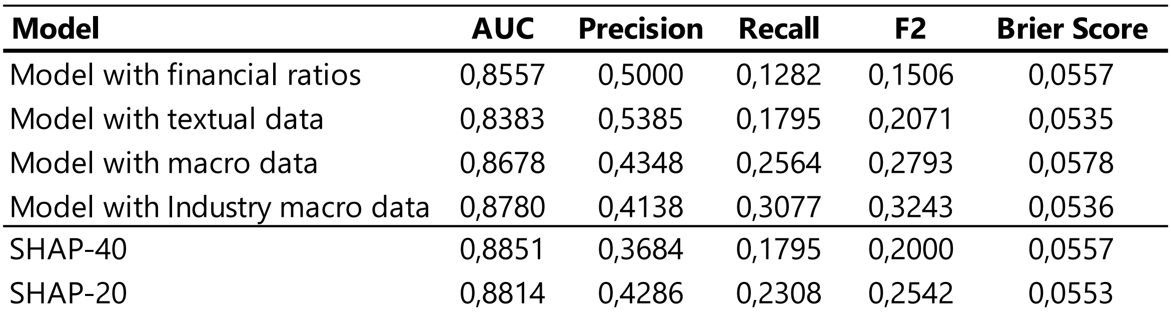


The ROIC offers insights into the industry's return trends over time. As an important financial indicator for value creation, a higher ROIC often correlates with better credit ratings and valuation (Petersen, 2017, p. 142). The ROIC trend in the industry appears to be declining, possibly due to factors such as intense competition, product pricing and the cost and financial structures of firms. Here it's important to recognize that oil and gas as products are homogeneous and prices are sensitive to the supply and demand balance, significantly impacted by OPEC+, U.S. and major consumers.

The analysis indicates a low correlation between average ROIC to bankruptcies, debt-to-equity ratios, or operational margins within the industry. However, the significant volatility observed in ROIC within this sector suggests a heightened risk profile, potentially explaining the elevated bankruptcy rates in the petroleum and gas industry. This volatility also implies challenges for our machine learning model, as it must contend with less stable data patterns.

Results

Our model implementations for the industry specific models are shown below. The methodology used is identical to the one used in the general model, with a new inclusion of a model with the industry specific macro features discussed in the section above.



Model	AUC	Precision	Recall	F2	Brier Score
Model with financial ratios	0,8557	0,5000	0,1282	0,1506	0,0557
Model with textual data	0,8383	0,5385	0,1795	0,2071	0,0535
Model with macro data	0,8678	0,4348	0,2564	0,2793	0,0578
Model with Industry macro data	0,8780	0,4138	0,3077	0,3243	0,0536
SHAP-40	0,8851	0,3684	0,1795	0,2000	0,0557
SHAP-20	0,8814	0,4286	0,2308	0,2542	0,0553

Figure 25: Results for the industry specific model

As you can see from the results in the table above, we see clear improvements at each step of additional features added to the model. Compared to the general model however, the industry specific model seems to gain more from the addition of macro specific variables. This might be to the significant volatility in the petroleum industry across the data, but we can’t say for sure without further exploration which we consider beyond the scope of the report. We see no further improvements by limiting the number of features in the model. Overall, the model with all features included provide the best results and will be used for the final comparison.

Comparing Models

In this section we will analyze and compare the general and the industry-specific model against each other. Before concluding on the best approach.

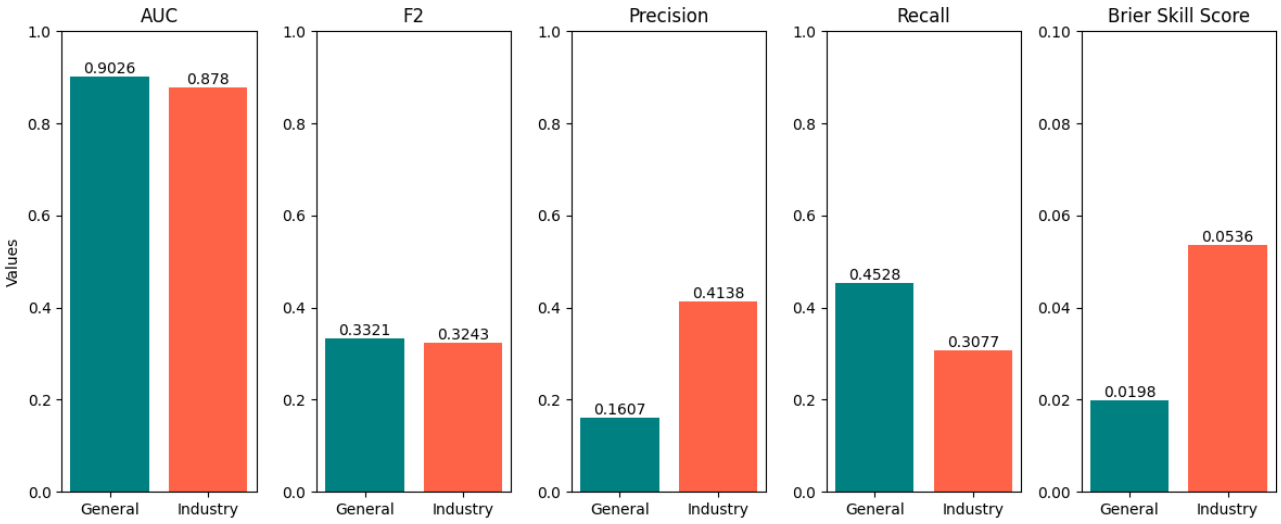


Figure 26: Illustration of the best results from both models

When comparing the metrics of the two models it becomes apparent that they excel in different areas. While an initial look at the AUC and F2 scores makes you think they perform similarly, we see that its not the case when we take a closer look at the precision and recall scores of the models. The industry model scores significantly better on precision, whilst the general model scores significantly better on recall. Taking the significantly larger precision-score of the industry model into account, one would assume that the industry model performs better also on the brier scale. However, the results indicate that the general model performs better in this metric. This indicates more overfitting in the industry model, which is to be expected given the reduced data input.

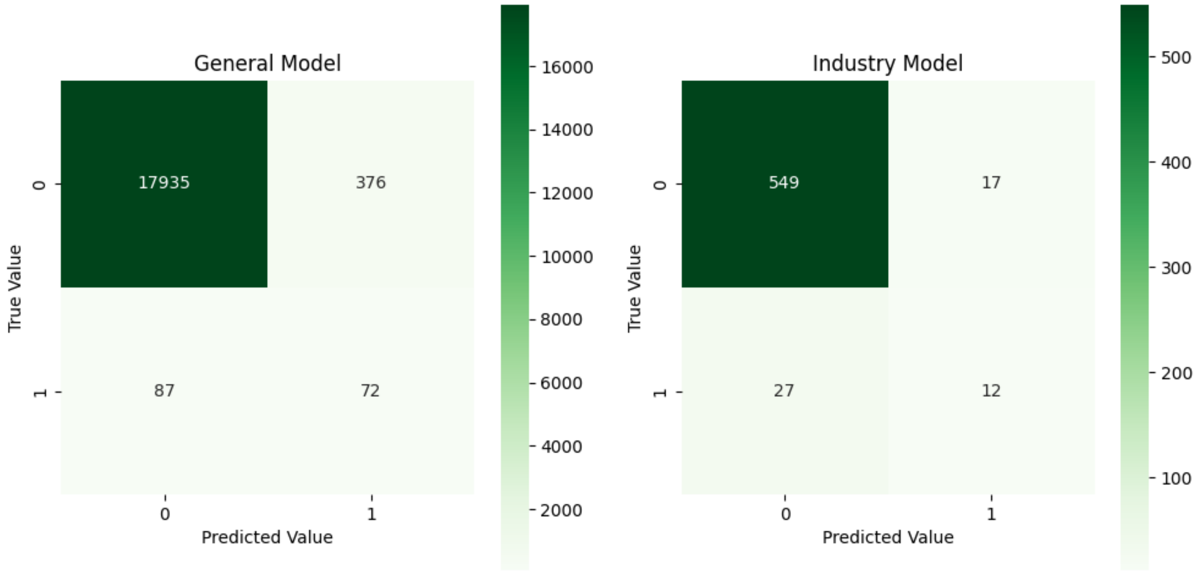


Figure 27: Confusion matrix, that compares the two models.

As expected from the metrics previously analyzed, the industry model has a higher degree of type 2 errors, while the general model has a higher degree of type 1 errors. However, it becomes apparent that the limited data in the industry model can cause the results to be overestimated. The metrics will be extremely sensitive to just a small adjustment in the predictions. Given small data sample, and the fact that the industry model performs worse on the brier scale, one must be careful when interpreting the results from the industry-model.

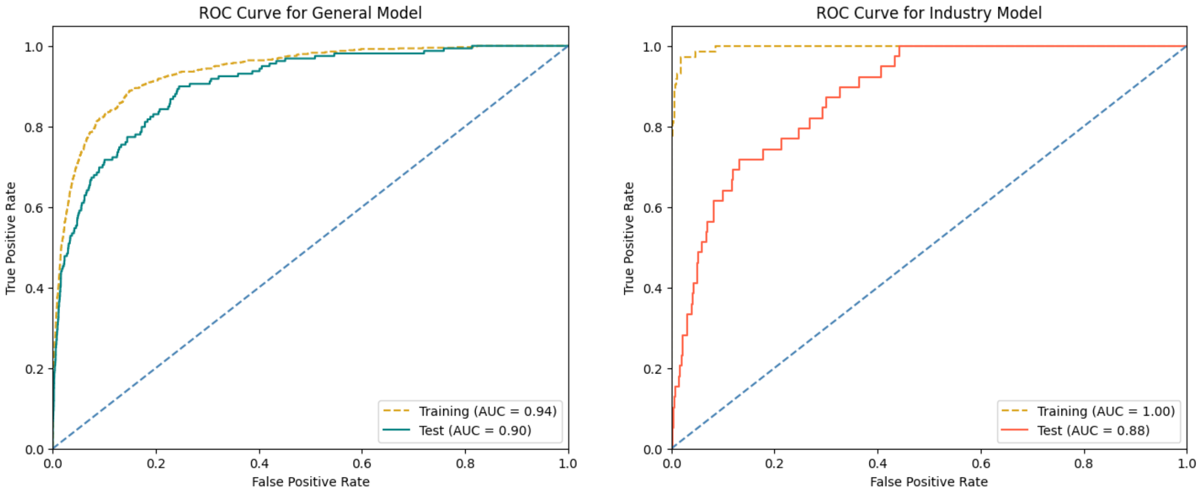


Figure 28: Roc curve for the respective models

After plotting the ROC-Curves of the models on both the training and test set it becomes clear that the industry model has a substantially higher degree of overfitting. While the general model performs only marginally better on the test set, the industry model predicts perfectly on the test set.

Conclusion

In conclusion, our exploration into the application of machine learning for bankruptcy prediction has given us insight on how such a model can be made, but also how one can improve it going forward.

Our approach, which combined traditional financial analysis with machine learning techniques, proved effective in addressing key issues such as missing and imbalanced data. This methodology not only kept the integrity of our analysis but also showcased the flexibility of using machine learning combined financial principles and data together.

Despite the inherent complexities and challenges associated with implementing this model, our findings reveal that such models can indeed produce reasonably accurate results. This is particularly noteworthy given the scope and limitations of this smaller-scale study. We also found that both textual and macro-economic features had a significant effect on the performance of the models.

We did however not find any improvements in the industry specific model over the general model. Our industry specific model was severely affected by the reduced data-sample causing a significant increase in overfitting, whilst also reducing the % of bankruptcies identified. Due to the extreme volatile nature of the petroleum industry, which might have had a large impact on our model, we cannot conclude that this will be the case for all industries. Further analysis is needed to make a definite conclusion.

Weaknesses

Financial variables

Unfortunately, our analysis encountered obstacles in the variables we intended to include in our model. This issue prevented us from calculating all the desired metrics such as consistent data on interest-bearing debt, both long-term and short-term. This data provides better insights into how companies finance their operations and enable relatively novel variables to assess the financial health of companies more accurately. As an example, the relationship between financial assets and short-term interest-bearing debt indicates a company's short-term financial resilience in case of diminishing profits. Furthermore, this would have enabled us to calculate Return on Invested Capital (ROIC) according to theoretical principles, distinguishing operational debt from financial debt, and likewise with assets.

The model includes Return on Assets (ROA) and Return on Equity (ROE). However, from an accounting point of view these variables require cautious interpretation. Equity, for instance, can be significantly influenced by accounting practices and may not always reflect the firm's true financial state. It's important to note that equity can derive from both intangible and tangible assets. For example, trademarks or goodwill are intangible assets. These can inflate a company's equity, especially in cases of mergers and acquisitions where overpayment occurs. Our attempts to extract data on goodwill and other intangibles from Compustat frequently resulted in 'NA' responses, leading us to exclude these variables from our model. It's worth mentioning that negative equity is often seen as a strong indicator of potential bankruptcy. Adjusting for intangible assets, particularly goodwill, could have provided more valuable insights on this matter. Additionally, ROA has its limitations, as net income is also susceptible to manipulation. The further down the income statement one goes, the more likely the results are to be distorted or unrepresentative.

Availability of specific financial data is not easily available, hindering creation of more specific variables as initially intended. Consequently, focus changed to calculate more traditional metrics such as the current ratio, ROE, ROA, gross profit, turnover, and EBITDA and EBIT margins as relevant accounting information such as inventory, current liabilities, total assets, and similar categories is more readily available.

Furthermore, as accounting standards evolve over time this will impact on the reliability of accounting variables. A recent example is the implementation of IFRS 16, which mandates that leasing assets and debts be recorded on the balance sheet, with leasing costs divided into financial expenses and depreciation. Prior to this, leasing costs were typically only included in operational expenses. This change has affected the company's profit margins, EBITDA, EBIT, and balance sheets. After the implementation, many companies may appear less financially robust due to an increased asset base against unchanged equity, thereby lowering the equity ratio. Our project could not adjust for these changes due to data limitations and time constraints (Kinserdal, 2019).

Macro Variables

Macro variables here are calculated annually, based on the averages of monthly data and then matched with company annual financial reports. Audited annual company reports are available only in the second quarter of the following year, thus analysis based on this data will be less timely. Companies do provide quarterly information; however, this information is more generic and less accurate as it is not subject to external audit. A trade-off with accuracy of date will make the analysis more timely but then less accountable.

A further enhancement would be to incorporate other critical factors, such as competitiveness, which Jacobsen & Kloster (2005) from Norges Bank identified as a significant driver for bankruptcies. However, this has not been introduced here due to lack of data availability.

A longer time frame would have provided a more extensive dataset and encompassed several business cycles, offering deeper insights and more robust predictive power. However, long term historical accounting data is limited and constrain the ability to fully leverage the potential of our machine learning model.

Model design

The weaknesses of our approach to the machine learning implementation are mainly caused by the quality of data accessible to us. Because of the poor data quality, we have used unsupervised learning techniques to artificially increase the quality of the data. This comes with the downside of potentially introducing large biases into our model.

Industry specific model

The average rate of bankruptcies is generally low, and even for petroleum and gas industry with the highest industry ratio the number is low, which is a significant source of inaccuracy in our analysis. Additionally, the oil and gas sector's heavy reliance on political factors and oil prices poses challenges in developing a robust and accurate model. The analysis revealed high volatility in bankruptcy trends within this industry, a stark contrast to sectors like pharmaceuticals, which seems to have been more stable. However, selecting the pharmaceutical industry for investigation might have also resulted in biased outcomes due to low level of bankruptcy data in this sector.

Looking ahead, it would be intriguing to compare different industries using machine learning models. In future studies, expanding the scope to encompass multiple sectors could yield a more comprehensive understanding of bankruptcy trends and patterns. This approach would facilitate the development of more refined machine learning models and enhance our grasp of industry-specific drivers.

References

Alanis, E., Chava, S., & Shah, A. (2022). Benchmarking machine learning models to predict corporate bankruptcy. Retrieved from https://arxiv.org/abs/2212.12051

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.

Biery, M. E. (2013, May 26). 4 things you don't know about private companies. Forbes. Retrieved from: https://www.forbes.com/sites/sageworks/2013/05/26/4-things-you-dont-know-about-private-companies/

Dobridge, C., John, R., & Palazzo, B. (2022, June 17). The post-COVID stock listing boom. Federalreserve. Retrieved from https://www.federalreserve.gov/econres/notes/feds-notes/the-post-covid-stock-listing-boom-20220617.html

Du Jardin, P. (2009). Bankruptcy prediction models: How to choose the most relevant variables? EDHEC Business School. Retrieved from https://mpra.ub.uni-muenchen.de/44380/

Gongde, G., Hui, W., David, B., Yaxin, B., & Greer, K. KNN Model-Based Approach in Classification. Retrieved from https://www.researchgate.net/publication/2948052_KNN_Model-Based_Approach_in_Classification

Jacobsen, D. H., & Kloster, T. B. (2005). What influences the number of bankruptcies? Norges Bank (Brage). Retrieved from https://norges-bank.brage.unit.no/norges-bank-xmlui/bitstream/handle/11250/2504364/jacobsen.pdf

Kinserdal, F. (2019). Dårligere analyser med nye leasingregler. Norges Handelshøyskole.

Koller, T., Goedhart, M., & Wessels, D. (2020). Valuation – Measuring and managing the value of companies (7th edition). McKinsey & Company.

Lundberg, S. M., & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Retrieved from https://scholar.google.com/citations?view_op=view_citation&hl=en&user=ESRugcEAAAAJ&citation_for_view=ESRugcEAAAAJ:dfsIfKJdRG4C

Nadar, D. S. (2019, June 30). Theoretical review of the role of financial ratios.

Shapley, L. S. (1951). Notes on the n-Person Game - II: The Value of an n-Person Game (ASTIA Document No. ATI 210720). Retrieved from https://www.rand.org/pubs/research_memoranda/RM670.html

Stocker, M., Baffes, J., & Vorisek, D. (2018, January 18). What triggered the oil price plunge of 2014-2016 and why it failed to deliver an economic impetus in eight charts. Retrieved from https://blogs.worldbank.org/developmenttalk/what-triggered-oil-price-plunge-2014-2016-and-why-it-failed-deliver-economic-impetus-eight-charts

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models.

Xiaobo T, Shixuan L, Mingliang T, Wenxuan S (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods

Cawhla N, Bowyer K, Hall L. (2002) SMOTE: Synthetic Minority Over-sampling Technique.

Loughran T, McDonald B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks

"Form 8-K." U.S. Securities and Exchange Commission, https://www.sec.gov/answers/form8k.htm. Accessed [08.12.2023].

"Bankrupt." Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/bankrupt. Accessed [08.12.2023].