IE 7300 Statistical Learning for Engineering Spring 2023

Group 3

Project Report

Companies Bankruptcy Data

Group Members

Mukul Agrawal

Abstract

The purpose of this project was to gain an understanding of how machine learning algorithms can be applied to analyze real-world datasets. To achieve this goal, we were given the task of selecting a dataset from the UCI Machine Learning Repository and using Python programming language to explore the data. Our main objective was to identify patterns and trends in the data and generate insights from our findings.

Throughout the project, we employed various machine-learning techniques and algorithms to explore the dataset. We also implemented data preprocessing techniques to clean and prepare the data for analysis. By applying these methods, we were able to identify significant patterns in the data, which we presented in the final report.

Overall, the project provided us with valuable experience in utilizing machine learning algorithms to analyze real-world datasets. It also allowed us to gain insight into the potential applications of machine learning in various industries, such as healthcare, finance, and marketing.

Introduction

In recent years, machine learning techniques have gained much attention due to their ability to provide accurate predictions and insights. One of the areas where machine learning has shown significant potential is in predicting the bankruptcy of companies. The ability to accurately predict bankruptcy can help investors, financial institutions, and regulatory bodies make informed decisions and take necessary steps to mitigate financial risks.

In this report, we present the results of a machine-learning project that focused on predicting the bankruptcy of Polish companies. The dataset used in this project includes data on bankrupt companies from 2000-2012.

The main objective of this project is to analyze the data thoroughly and develop a machine learning model that can accurately predict the likelihood of bankruptcy of Polish companies. We use various machine learning algorithms, including Logistic Regression, Naive Bayes, and support vector machines, to train and evaluate our models. We also conduct feature selection and engineering to identify the most important features that contribute to the prediction of bankruptcy.

The findings of this project can provide valuable insights to investors, financial institutions, and regulatory bodies interested in assessing the financial health and stability of Polish companies. The results can also help companies identify potential financial risks and take necessary measures to prevent bankruptcy.

Data Description

This project is focused on predicting the bankruptcy of Polish companies using a dataset that was collected from the Emerging Markets Information Service (EMIS), a database that contains information on emerging markets from around the world. The dataset contains financial information on both bankrupt and still-operating companies, with a total of 10173 instances (financial statements) included.

Of these 10173 instances, 400 represent bankrupt companies, while 9773 represent firms that did not go bankrupt in the forecasting period. The data shows the bankruptcy status of these companies after 4 years of operating, providing valuable insight into their financial health and stability.

The financial rates of the companies are described by 64 attributes, each of which provides a unique perspective on the company's financial situation. These attributes include *net profit / total assets, net profit/sales, total assets, working capital, sales/inventory*, and more. All the independent features in the dataset are numerical in nature, allowing for easy analysis and comparison between different companies and financial metrics.

The dependent feature in the dataset is a boolean, with a value of 1 indicating bankruptcy and a value of 0 indicating no bankruptcy. This allows for the creation of a predictive model that can use the various independent features to accurately predict whether a given company is at risk of bankruptcy in the future. Overall, the dataset provides a valuable resource for researchers and analysts interested in understanding the financial health of Polish companies and predicting their likelihood of bankruptcy.

Descriptive Analysis

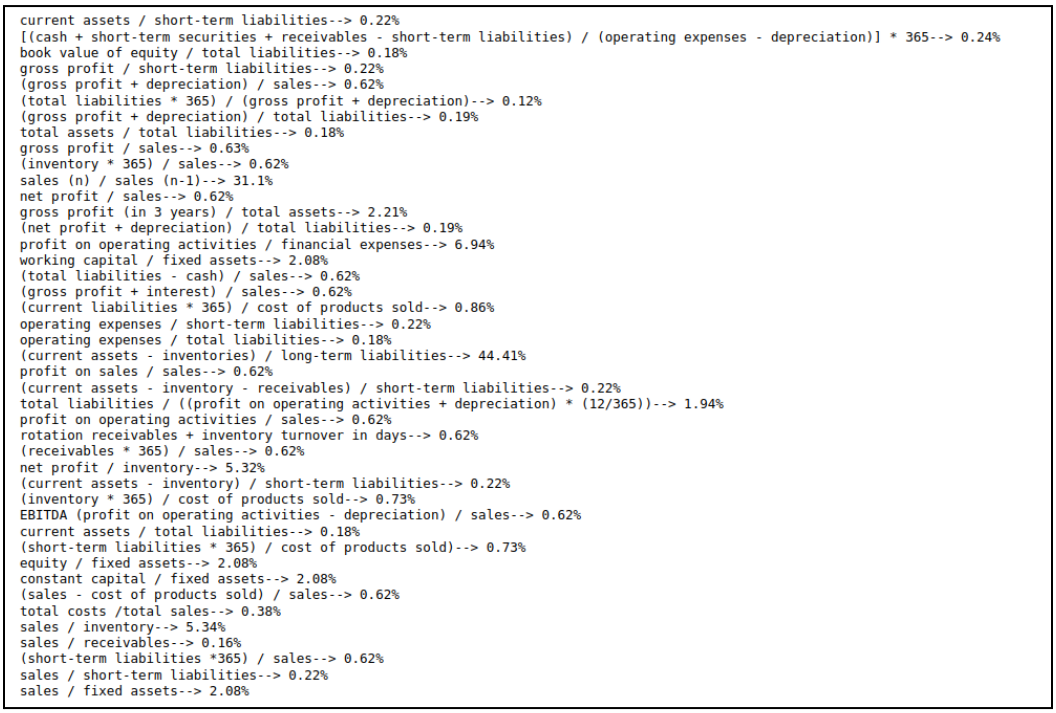
Descriptive analysis is a statistical method used to describe and summarize a set of data. It involves examining the data to identify patterns, trends, and relationships among variables, as well as summarizing the data using measures such as mean, median, mode, range, and standard deviation. This type of analysis is used in many different fields to gain a better understanding of the data and to inform subsequent analyses and decision-making.

While applying descriptive analysis to the dataset, we figured out that the following steps can be taken before moving on to the next step of the project i.e. Predictive analysis:

- Deal with null values
- Deal with correlated variables
- Deal with unscaled data
- Deal with Imbalanced data

1. Deal with null values

There are many columns with null values (as shown in the below image).



There could be many ways to deal with null values. Some of them are:

a) Delete rows containing null values
b) Delete columns containing null values, and
c) Fill null values with mean, median, or mode

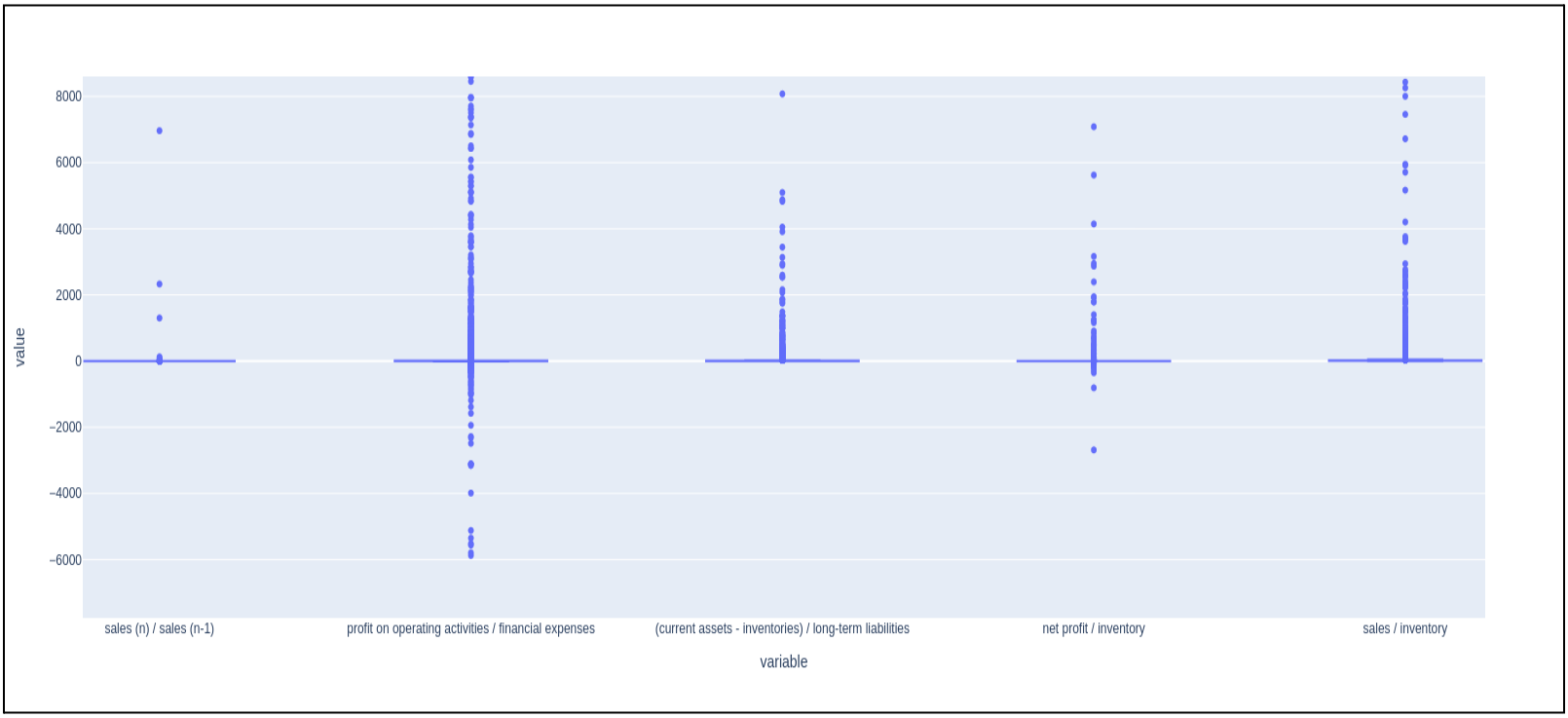
For this project we have divided our approach to deal with null values into 2 ways:

A. Delete rows containing null values

If in a column the null value percentage is below 3, we have deleted those rows. After this step, we removed almost 7% of the data.

B. Fill null values median

There are 4 columns that have more than 3% null values. Since these columns contain outliers (as shown the below figure) too which might be useful for predictive analysis. Therefore, instead of the mean, we have filled null values with the median of the respective column.

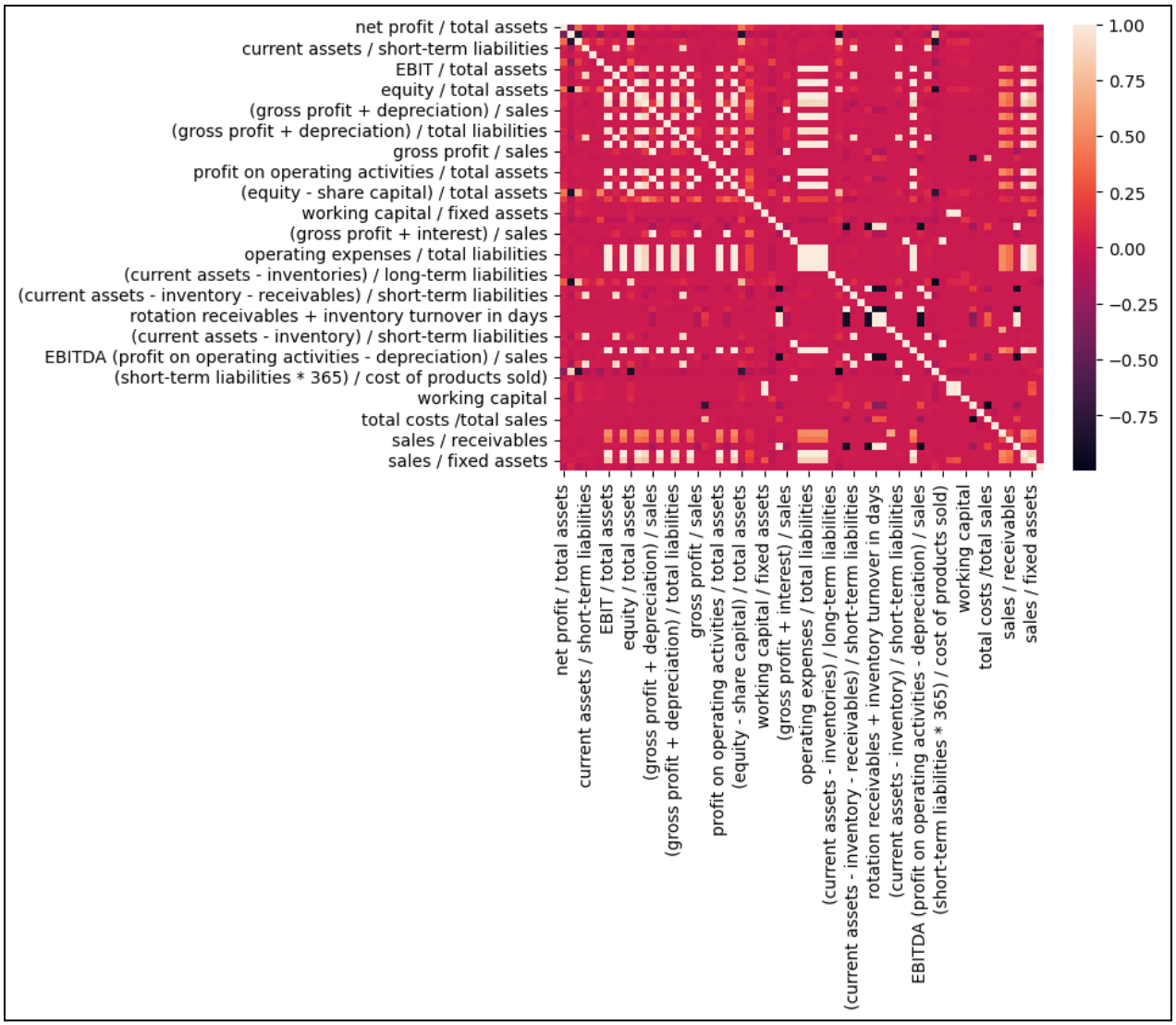


2. Deal with correlated variables

A **collinearity** is a special case when two or more variables are exactly correlated.

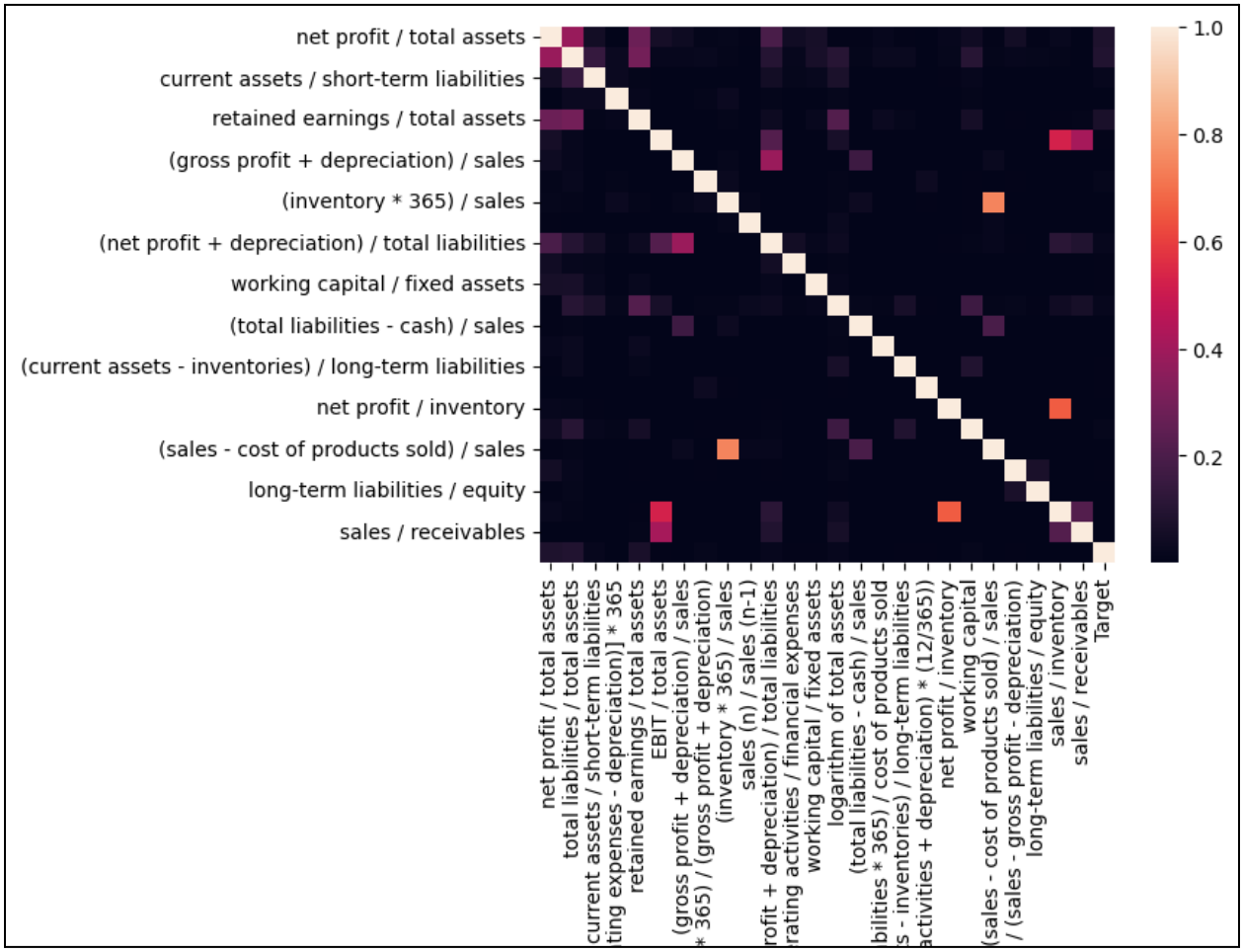
This means the regression coefficients are not uniquely determined. In turn, it hurts the interpretability of the model as then the regression coefficients are not unique and have influences from other features.

In the dataset, there are many columns that are correlated to each other (below heat plot)



It becomes important to delete any of the two correlated variables so that it doesn't affect our model in predictive analysis. There are 39 columns that can be dropped to remove collinearity from the dataset.

The heat plot after dealing with collinearity looks like the below plot. The number of columns in the dataset is 25.



3. Deal with unnormalized data

Feature scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model. Feature scaling is essential when working with datasets where the

features have different ranges, units of measurement, or orders of magnitude.

Standardization is a scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

4. **Deal with imbalanced data**

The term "imbalanced data" describes a type of dataset where there is an uneven distribution of observations within the target class. This means that one of the class labels has a significantly higher number of observations than the other, resulting in an imbalance.
To deal with imbalanced data we are using SMOTE (Synthetic Minority Over-sampling Technique). This works by randomly selecting the value of a column from its nearest k neighbors.

The steps followed by SMOTE are:

a) Select all minority class observations.
b) For each minority class observation, find k neighbors
c) To create a new observation, go through all the observations one by one and select a random value from its neighbors for each column.
d) Repeat this process the number of times the majority class is to the minority class

In our dataset, the companies that filed for bankruptcy are 379 compared to 9071 companies that did not.

After using SMOTE, the companies that filed for bankruptcy are 6600 while companies that did not are 6340 in number.

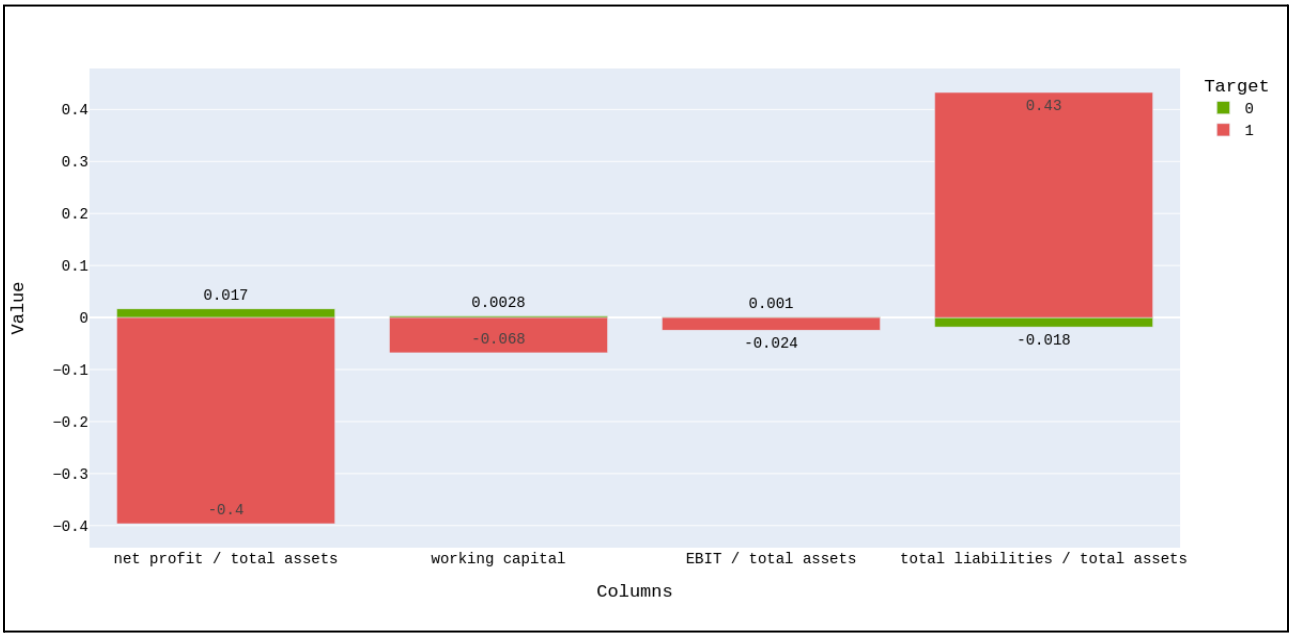
Exploratory Data Analysis

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables.

Below are some insightful observations that came out of EDA:

I. **Bar plot for the mean value of different features** (0: No Bankruptcy, 1: Bankruptcy)

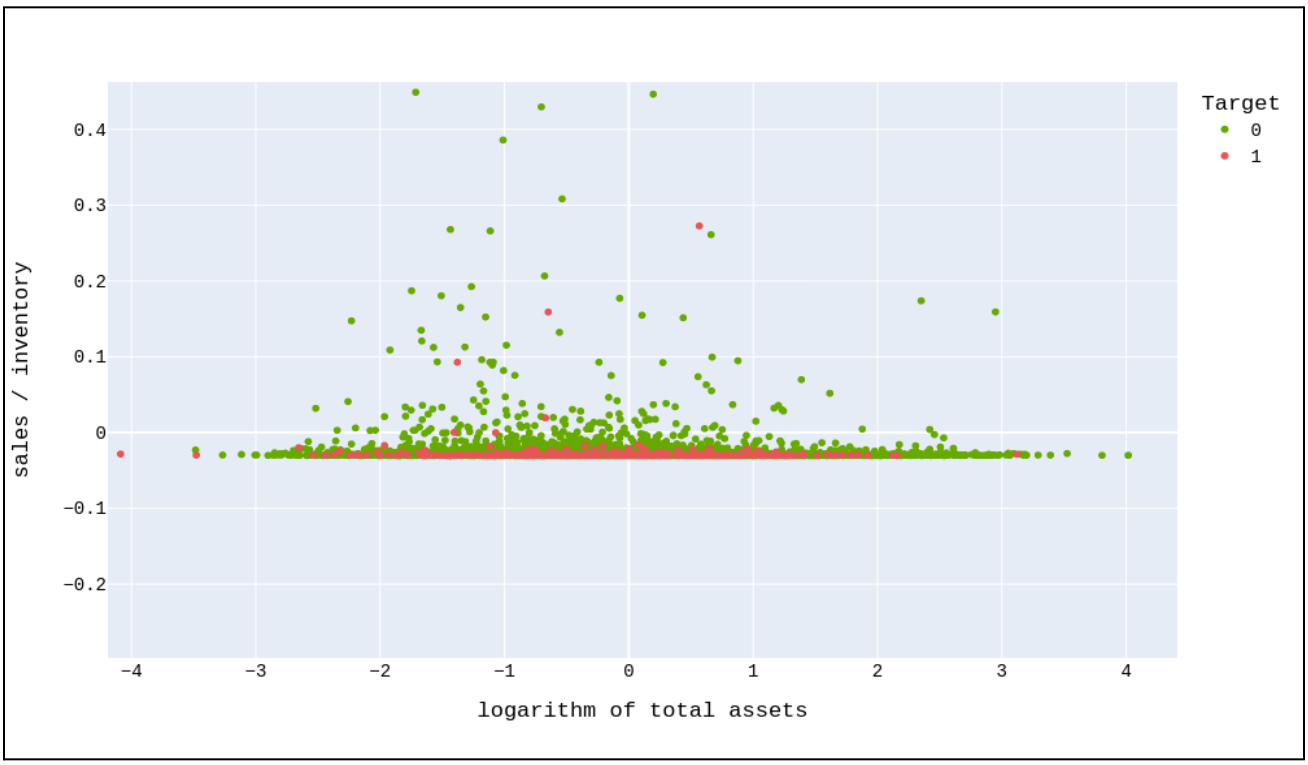


The financial health of a company is crucial for its success and sustainability. A company's balance sheet is an important indicator of its

financial well-being, as it reflects its assets, liabilities, and equity. When a company's liabilities exceed its assets, or when it is unable to generate sufficient revenue to cover its expenses, it may face financial difficulties, which could lead to bankruptcy.

The above bar plot suggests that bankrupt companies had negative net profit, working capital, and EBIT in relation to their total assets. Negative net profit means that the company's expenses exceeded its revenue, resulting in a loss. Low working capital indicates that the company was unable to meet its short-term obligations, such as paying suppliers and employees. Low EBIT implies that the company was not generating sufficient profits to cover its operating expenses. In addition, bankrupt companies had high total liabilities in relation to their total assets, indicating that they had a significant amount of debt that they were unable to repay.

II. Scatter plot between *Sales/inventory* and *total assets*

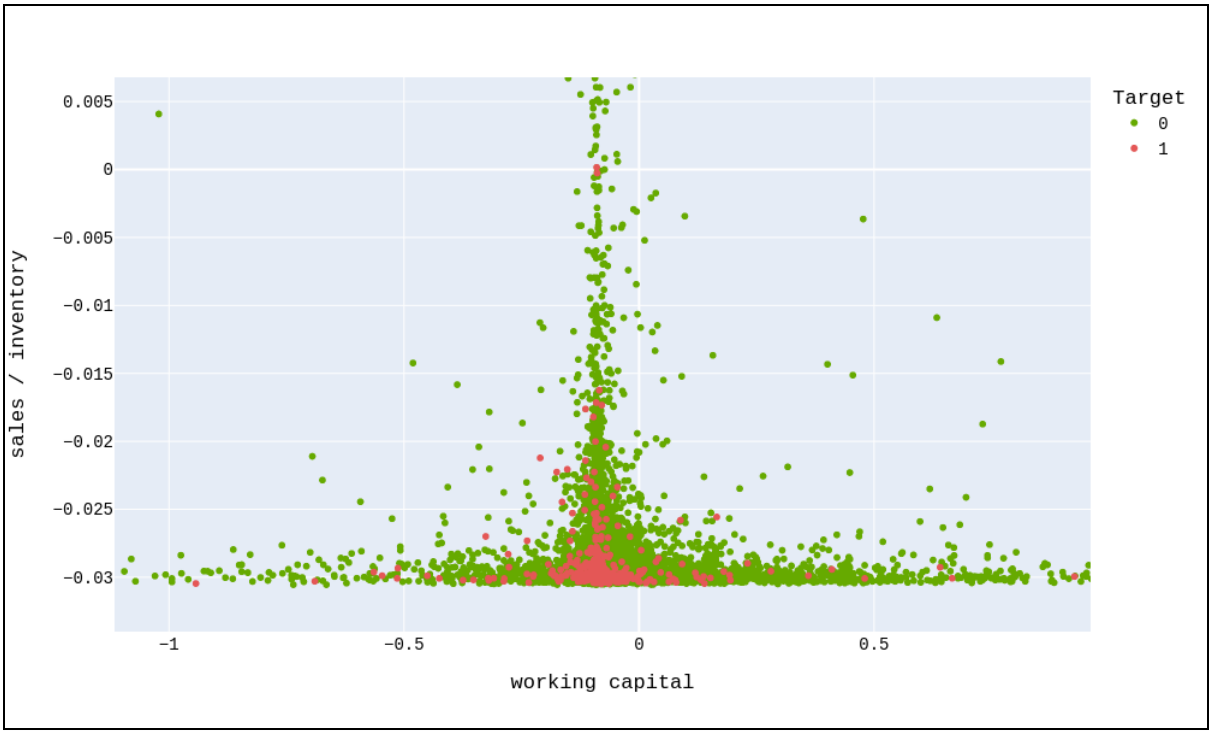


The plot suggests that there is a correlation between a company's sales and its inventory levels, and how this might affect its financial health. Specifically, the plot indicates that companies that went bankrupt had lower sales compared to their inventory levels, meaning that they were holding onto more inventory than they were able to sell.

When a company has excess inventory, it ties up capital that could be used for other purposes, such as investing in new products, expanding operations, or paying off debts. If the company is unable to sell its inventory, it may result in lower revenue and cash flow, which can ultimately lead to financial difficulties and bankruptcy.

Therefore, the plot underscores the importance of effective inventory management and sales strategies to maintain a healthy balance sheet and avoid bankruptcy. Companies should strive to keep their inventory levels in line with their sales to ensure that they are not holding onto excess inventory that they cannot sell, while also ensuring that they are meeting customer demand and generating sufficient revenue to sustain their operations.

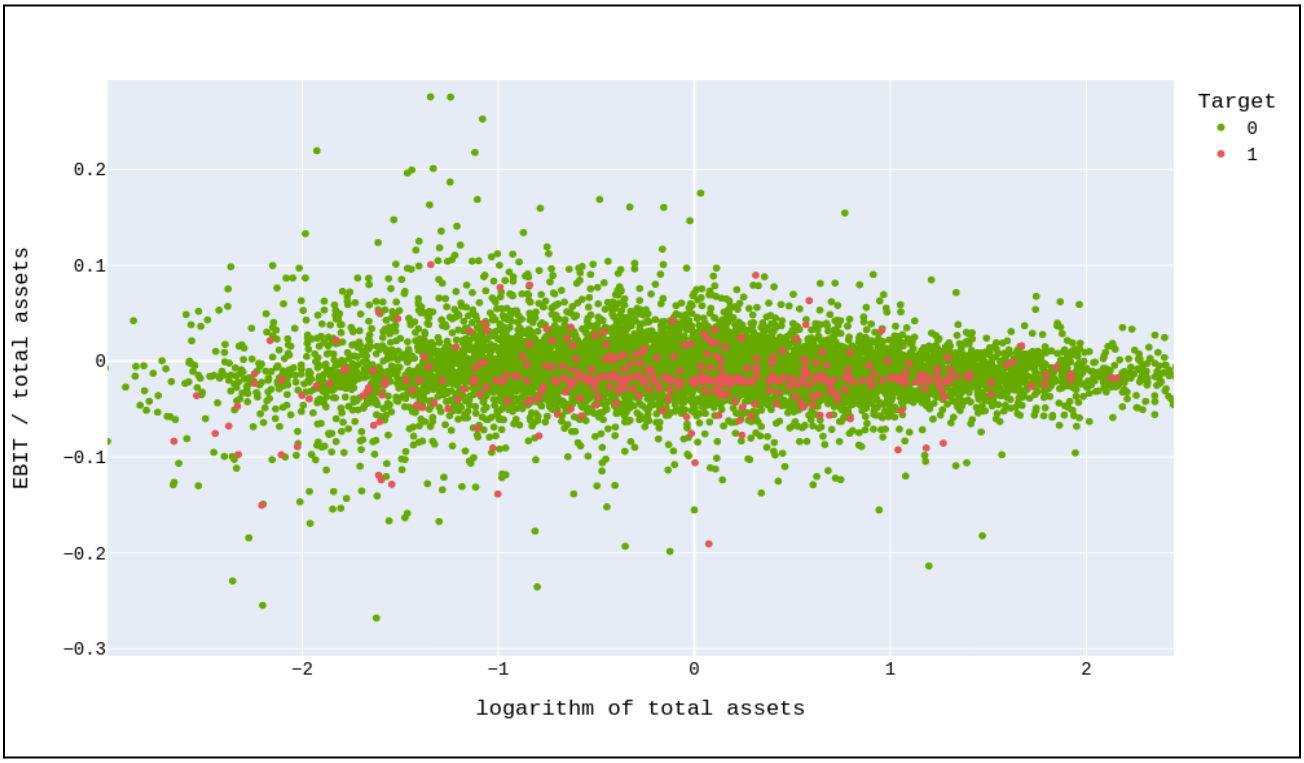
III. Scatter plot between *Sales/inventory* and *working capital*



Working capital is a financial metric that indicates a company's ability to meet its short-term obligations. It is calculated by subtracting current liabilities, such as accounts payable and short-term debt, from current assets, such as cash, inventory, and accounts receivable. A positive working capital indicates that the company has enough short-term assets to cover its short-term liabilities.

The plot suggests that many bankrupt companies had negative working capital, which means that their current liabilities exceeded their current assets. This implies that these companies were unable to meet their short-term obligations, such as paying suppliers and employees, and were struggling with cash flow issues. Furthermore, as indicated by the plot, these companies had low sales compared to their inventory, which suggests that they were unable to sell their inventory and generate revenue to cover their expenses.

IV. Scatter plot between *EBIT/total assets* and *total assets*



EBIT (Earnings before Interest and Tax) is a financial metric that represents a company's operating profit before deducting interest expenses and taxes. EBIT is used to analyze a company's profitability and operating efficiency, as it provides insight into how much profit a company generates from its core operations.

The plot suggests that almost all of the companies had low EBIT/total assets, which implies that they were not generating sufficient profits from their operations. Low EBIT/total assets could indicate that a company's expenses are higher than its revenues, resulting in a loss or a very low-profit margin. It could also indicate that a company's assets are not being utilized effectively to generate revenue.

Predictive Analysis

Predictive analytics is an area within advanced analytics that utilizes statistical modeling, data mining techniques, and machine learning to make predictions about future outcomes based on historical data.

For our project, we have used the following models:

a) **Logistic Regression**: Logistic regression involves developing a model that predicts the probability of a specific outcome based on an input variable. The most prevalent type of logistic regression involves modeling a binary outcome, which means an outcome that can only take one of two possible values, such as true/false or yes/no.

b) **Gradient Descent**: Gradient descent (GD) is an iterative algorithm used in machine learning (ML) and deep learning (DL) to minimize a cost/loss function by finding a local minimum/maximum of a given function. This first-order optimization method is often applied to linear regression problems and other types of models in which the goal is to minimize the difference between predicted and actual values.

c) **Stochastic Gradient Descent**: Stochastic gradient descent (SGD) is an iterative algorithm used to optimize an objective function with certain smoothness properties, such as differentiability or sub-differentiability. It is a stochastic approximation of the gradient descent optimization method because it replaces the actual gradient, calculated from the entire dataset, with an estimate derived from a randomly selected subset of the data. This is particularly useful in high-dimensional optimization problems because it reduces the computational burden, allowing for faster iterations at the expense of a lower convergence rate.

d) **Ridge Regression**: Ridge regression is a technique used for model tuning that is applicable to data sets with multicollinearity issues. This method employs L2 regularization to address the problem. In situations where multicollinearity is present, the least-squares approach can lead to biased results and large variances, causing predicted values to deviate significantly from actual values. Ridge regression helps mitigate these issues, leading to more reliable predictions.

e) **Lasso Regression**: Lasso regression is a regularization method utilized in regression analysis to achieve more accurate predictions. This technique involves shrinkage, which is the process of pulling data values toward a central point such as the mean. Lasso regression encourages the creation of simple and sparse models with fewer parameters. By constraining the coefficients of the model to be closer to zero, Lasso regression can help identify the most important features in the data while ignoring those that are less significant.

f) **Naive Bayes**: Naïve Bayes is also known as a probabilistic classifier since it is based on Bayes' Theorem. It would be difficult to explain this algorithm without explaining the basics of Bayesian statistics. This theorem, also known as Bayes' Rule, allows us to "invert" conditional probabilities. As a reminder, conditional probabilities represent the probability of an event given some other event has occurred, which is represented with the following formula.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

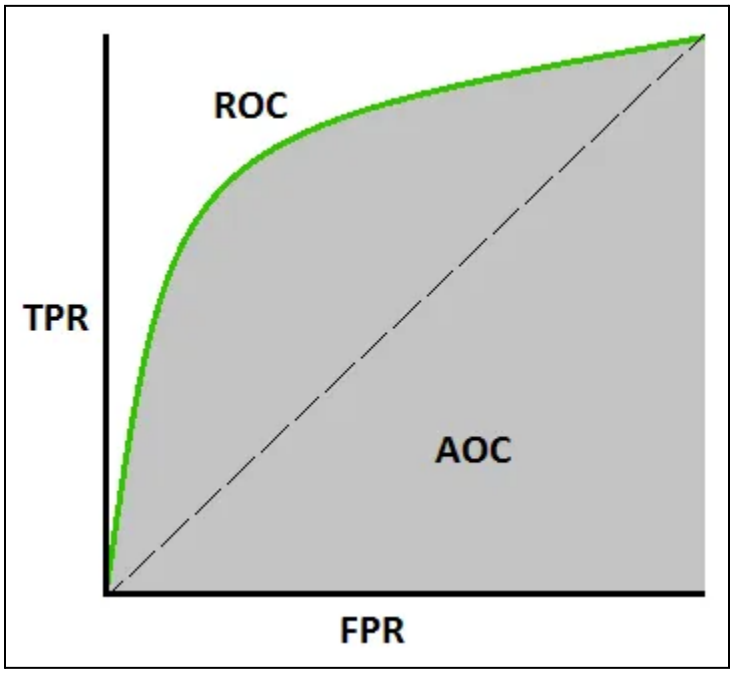
g) **Support Vector Machine**: The support vector machine algorithm aims to identify a hyperplane in an N-dimensional space (where N represents the number of features) that effectively separates the data points into distinct classes. Numerous hyperplanes can be used to separate the two classes of data points. However, the goal is to identify the hyperplane with the maximum margin, which refers to the maximum distance between the data points of both classes. By maximizing the margin distance, we can increase confidence in future data point classification.

The metrics taken for evaluation purposes are:

a) **Accuracy**: Classification accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions.

b) **F1-Score**: is a machine learning evaluation metric that assesses the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by accuracy. F1 score combines two competing metrics- precision and recall scores of a model, leading to its widespread use in recent literature.

c) **AUC ROC**: is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 class as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.



Results

After establishing a model and providing it with training, one outcome is to evaluate its performance with unseen test data. This involves assessing the speed and accuracy of the different models and their resulting outcomes.

We started our data modeling process with the use of logistic regression. However, the results obtained were not satisfactory, with all metrics showing low scores. This led us to experiment with different approaches, such as gradient and stochastic gradient descent, which showed some improvement in the results.

Despite these improvements, we noticed a significant difference between the performance on the training data and the test data, which could be attributed to overfitting. Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data.

To combat overfitting, we decided to use regularization techniques like ridge and lasso regression. These techniques penalize large coefficients in

the model, effectively reducing the complexity and preventing overfitting. However, we did not achieve the expected improvement in results despite the use of these regularization techniques.

	Logistic Regression		Gradient Descent		Stochastic Gradient Descent		Ridge Regression		Lasso Regression	
	train	**test**	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**
Accuracy	45.07	52.98	54.9	32.9	54.9	41.0	48.4	52.38	55.0	44.7
Precision	0.46	0.52	0.54	0.03	0.55	0.04	0.49	0.04	0.55	0.03
Recall	0.39	0.04	0.76	0.63	0.67	0.58	0.46	0.49	0.63	0.52
F1_Score	0.42	0.07	0.63	0.06	0.60	0.07	0.47	0.07	0.59	0.06
AUC	0.53	0.62	0.54	0.46	0.54	0.47	0.48	0.53	0.55	0.45

We decided to experiment with additional models after not seeing satisfactory results on the test data using previous modeling techniques. Decided to test the performance of Naive Bayes and SVM models on the same dataset.

Upon analyzing the results, we observed that the recall metric for both Naive Bayes and SVM models showed good results. Recall is a metric that measures the percentage of actual positive cases correctly identified by the model. This implies that the models are correctly identifying most of the companies that may go bankrupt in the future.

However, the f1 score, which is a weighted average of precision and recall, was very low, indicating that there are many false positives. False positives refer to cases where the model incorrectly predicts that a company will go

bankrupt in the future, leading to false alarms and potentially unnecessary interventions.

Despite the low f1 score, we acknowledge the importance of correctly identifying all the companies that may go bankrupt in the future. Therefore, we may decide to use these models with caution, possibly in combination with other techniques that can further reduce false positives, to achieve a balanced trade-off between sensitivity and specificity.

	Naive Bayes		**SVM**	
	train	**test**	**train**	**test**
Accuracy	53.23	9.62	51.0	3.66
Precision	0.52	0.90	0.51	0.03
Recall	0.97	0.035	1.0	1.0
F1_Score	0.68	0.06	0.67	0.07
AUC	0.98	0.91	1.0	1.0

Conclusion:

The primary objective of the project is to employ advanced machine learning techniques to analyze the financial records of a company and determine if it is likely to go bankrupt. To achieve this goal, several preprocessing steps have been taken to transform the data and improve the predictive accuracy of the models used. These preprocessing steps could involve data cleaning, normalization, scaling, feature engineering, or feature selection.

We also conducted a thorough exploratory data analysis to gain insights into the data and the relationship between the different features. This involved creating various plots and visualizations to help understand the patterns and correlations in the data. By examining these plots, we were able to identify important features and gain a deeper understanding of how the different variables impact the likelihood of bankruptcy.

Finally, we employed various machine learning models such as Naive Bayes and Support Vector Machines (SVM) to predict the likelihood of bankruptcy. These models were trained on a portion of the data and then tested on a separate set of data to determine their accuracy. After analyzing the results, we concluded that these models are effective in predicting whether a company would go bankrupt or not.

Overall, the project aimed to use machine learning techniques to help financial analysts and investors identify companies that may be at risk of bankruptcy, allowing them to make more informed decisions and potentially avoid financial losses. The project's success could have significant implications for the financial industry and contribute to the development of more advanced and accurate predictive models.

Future Research

We have applied and used every primary and possible step in this project but there are many other complex steps that might not only give more understanding of the data but may improve the prediction results as well. Following are some of the steps which we would like to use to improve our results:

a) ***Apply Neural Network***: A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

b) ***Hyperparameter Tuning***: Hyperparameter tuning works by running multiple trials in a single training job. Each trial is a complete execution of your training application with values for your chosen hyperparameters, set within the limits you specify.

c) ***Feature Selection***: Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

d) ***Use Validation data***: Validation data provides an initial check that the model can return useful predictions in a real-world setting, which training data cannot do.