Financial Distress Prediction

Final Project for Data Mining
ISM6136.901F22

Team Members

Prathyusha Thummalapalli (U90390196)
Akhil Vayunandan Reddy Ankireddypalli (U32430700)
Ramya Rani Tatiparthi (U95670175)

Table of Contents

Introduction

Financial distress occurs when an organization is unable to generate adequate revenues or money to satisfy or fulfill its financial commitments. If financial distress is not alleviated, it may result in bankruptcy. Financial distress is typically connected with some expenses incurred by the company, which are referred to as "costs of financial distress." The direct and indirect costs of financial distress, which could also be some of the attributes in our dataset, include higher capital costs, a high debt burden, operational issues, and bankruptcy costs like auditors' fees, legal fees, management fees, and other payments, etc. Additionally, the company has firm expenses that must be paid, which may include financing, such as paying interest on debts, and opportunity costs of projects.

Some of the well-known companies that went into financial distress are Apple, IBM, Marvel, L'Occitane, Revlon, etc.

Motivation

There were numerous discussions about choosing a project, and all ideas were aggregated to investigate different models and splits on the dataset and investigate new concepts that result in better modeling performance.

Background

We examined multiple datasets that fit our objective. We finally settled on the financial distress dataset which we believe has a broad scope.

We chose this dataset from Kaggle which deals with the financial distress prediction for a sample of companies.

https://www.kaggle.com/datasets/shebrahimi/financial-distress

Problem Statement

Analyze the various attributes of different companies over a period of time. Then train the data using predictive models to determine which attributes have the greatest impact and which model performs better on this dataset.

Objective

Investors and the government incur significant economic losses when a firm is in a financial crisis, in addition to the company itself as it might find it hard to secure new financing. Therefore, it is vital to use different prediction models for financial distress that are very precise and efficient. The primary objective is to analyze, predict, compare and visualize the data utilizing different prediction models.

Advantages

- Predicting financial distress provides early warning for the company.
- It is beneficial for both investors and creditors.

Summary

Implementation

- Each company has various period data.
- Subset the latest period data using python.
- We categorized healthy as 0 and distressed as 1
- Compared Logistic Regression and Neural Network
- Used Hyperparameter Tuning to increase the efficiency
- Used Azure ML Studio

Inference

- This dataset is imbalanced (Distressed-136, 286-Healthy).
- Considered F1-Score as an evaluation metric rather than accuracy as data is skewed.
- Compared 2 classification models by doing 2 splits (70-30 and 80-20).
- Used correlation plot to find the most affecting attributes to the target variable.

Dataset

There are 3673 records and 86 attributes in the actual dataset.

Below is the description of each column.
First column: Company has a unique ID that represents sample companies.
Second column: Time shows different periods that data belongs to. The period varies between 1 to 14 for each company.
Third column: The target variable is denoted by "Financial Distress".
Fourth to the last column: The features denoted by x1 to x83 are some financial and non-financial characteristics of the sampled companies.

Limitations

Since this data set is unbalanced and skewed, the performance evaluation criterion should be the f-score.

- This data could be viewed as a classification problem.

Methodology

As the first step, we transformed the dataset

- In the dataset, Financial Distress is a numerical and continuous value. For a better analysis, the Financial Distress column is converted to binary based on the following condition:

If Financial Distress>-0.5 then it is considered as Healthy and 0 is assigned else, it is considered as Distressed and 1 is assigned.

- Instead of using the entire dataset, it only made sense to subset the latest period data for each company, resulting in 423 records.

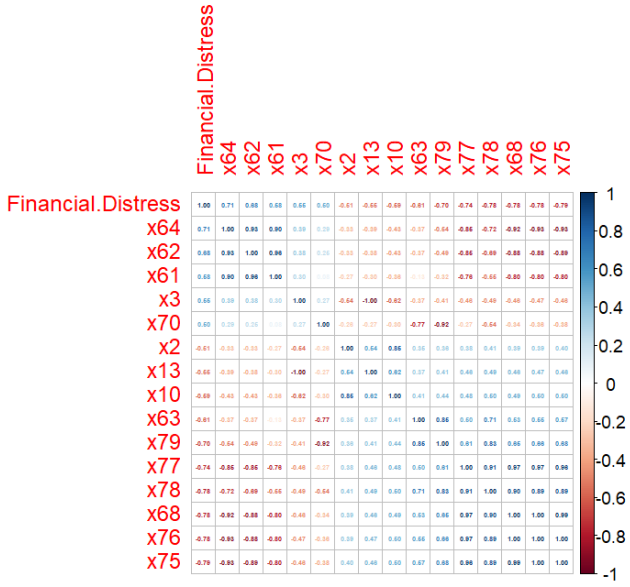
  To do so, the following Python code is implemented-

```
import pandas as pd
data = pd.read_csv("C:/Users/Downloads/Financial Distress.csv")
finaldata = data.sort_values('Time').drop_duplicates(subset='Company',keep='last')
def D(row):
    if row['Financial Distress'] > -.5:
        return 0
    else:
        return 1
data["Financial Distress"] = finaldata.apply(D, axis=1)
finaldata.to_excel('C:/Users/Downloads/DM.xlsx',index=False)
```

- To determine the attributes' degree of impact a sorted list of top correlated attributes is compiled, and a correlation plot is visualized.

```
cm = data.corr()
p=cm["Financial Distress"].sort_values(ascending=False)
p.to_excel('C:/Users/Downloads/corr.xlsx',index=True)
```

	Financial Distress
x64	0.714163992
x62	0.683952391
x61	0.582117669
x3	0.550011671
x70	0.501158142
x2	-0.505400386
x13	-0.550012296
x10	-0.591181169
x63	-0.611065136
x79	-0.704188997
x77	-0.741128225
x78	-0.782831329
x68	-0.783542608
x76	-0.784319764
x75	-0.791192714



- The top five are positively correlated to the target attribute and the bottom ten are negatively correlated.

Positive Correlation- As one variable increases, so does the other variable.

Negative Correlation- As one variable increases, The other variable decreases.

- Even after finding the impacting attributes, when a correlation plot is taken for all the attributes in the dataset it is observed that attributes are correlated between themselves and together are correlated to the target variable. So, we have decided to consider all the attributes as part of building models.
- And then on the transformed dataset, Two-Class Neural Network and Two-Class Logistic Regression models are built with 70:30 and 80:20 splits in Azure ML Studio.
- Later Hyper Parameter Tuning is applied to better the performance of models.

Two-Class Neural Network- Neural Network is nothing but a set of interconnected layers. Between the input and output layers, there are a few hidden layers. The no. of hidden layers is determined based on the predictive task assigned. A Two-Class Neural Network is adopted when the target variable has only two values, which is the case of our dataset. It is said that given any dataset, Neural Network always overfits it. So, we wanted to compare the outcomes of both models with and without Tuning the Hyperparameters and determine the better-performed model.

Two-Class Logistic Regression- Logistic Regression is well-known for classification tasks. It is used to predict the possibility of an outcome. As our dataset has a class column (Financial Distress) and is two-class i.e., 0 or 1, it is only logical to use this model.

Hyperparameter Tuning- Hyperparameters are adjustable parameters that have control over the training of a model. It is a process of finding the configuration of hyperparameters that results in the best performance. For example, in neural networks, you give the number of hidden layers and the number of nodes in each layer. Usually, this is a manual task and is time-taking. Azure ML has this process automated.

Evaluation Metrics

Since we have a skewed dataset, accuracy cannot be the evaluation metric. To better comprehend a model's performance, the F-Score is considered for the evaluation metric.

F-Score- It is a harmonic mean between Precision and Recall.

Precision- Within everything that has been predicted as a positive, precision counts the correct percentage.

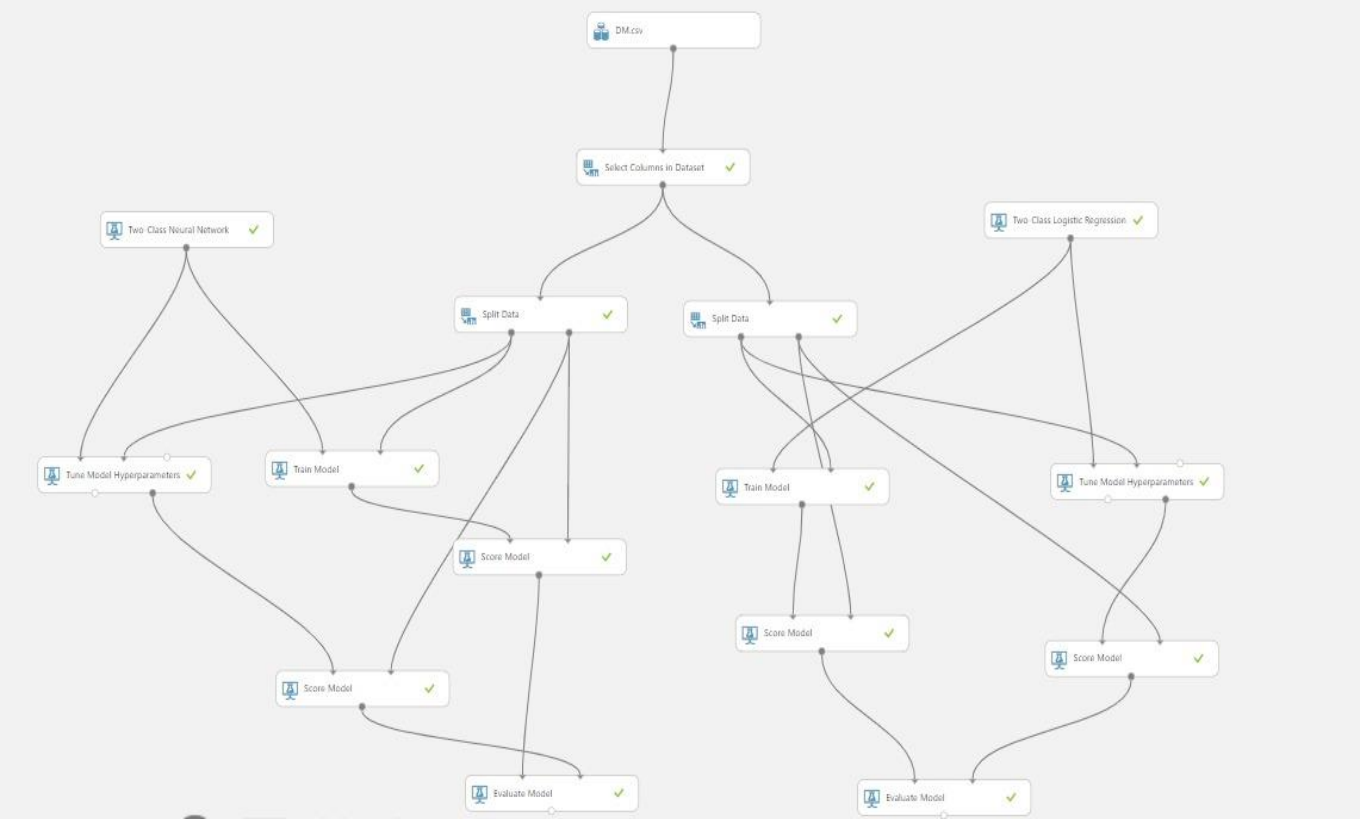
$$\text{Precision} = n(TP)/\ n(TP)+n(FP)$$

Recall- Within everything that is truly positive, how many did the model succeed to find?

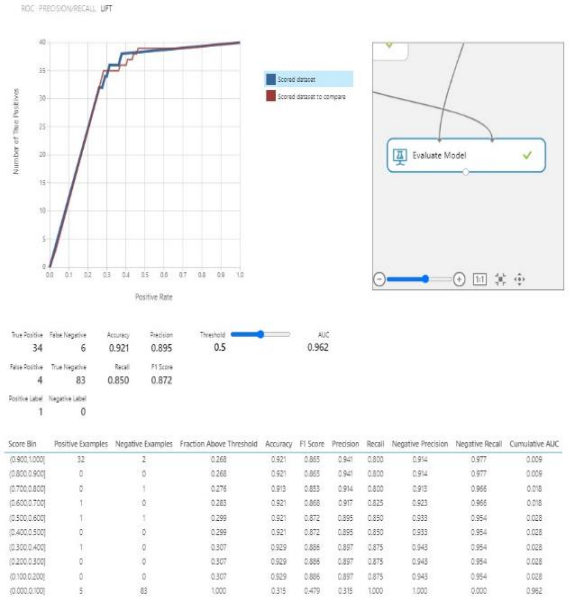
$$\text{Recall} = n(TP)/n(TP)+n(FN)$$

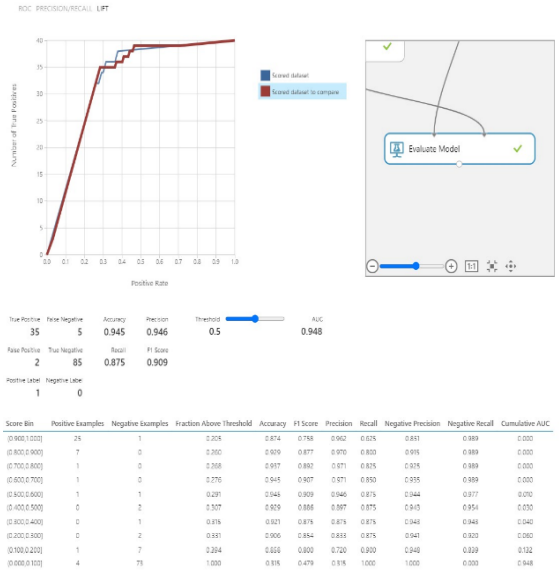
Results

Below is our Azure ML Experiment with a Two-Class Neural Network and Two-Class Logistic Regression models and Hyperparameter Tuning.

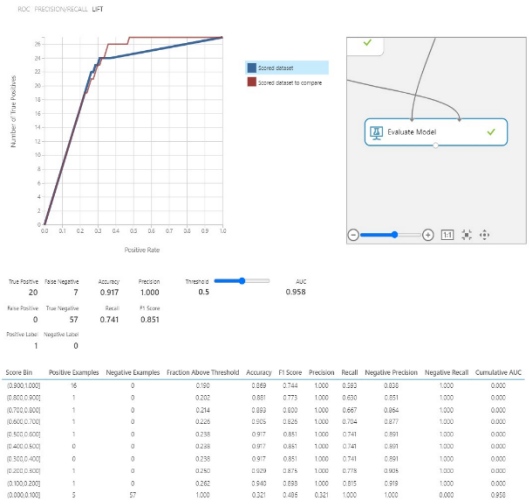


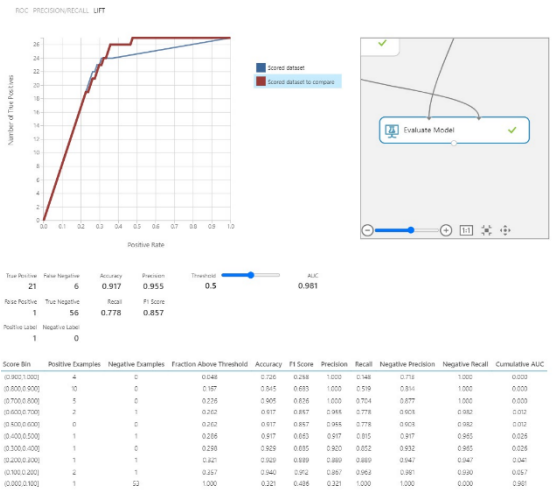
The following are the results of the Two-Class Neural Network for a 70-30 Split without and with applying Hyper Parameter Tuning respectively.



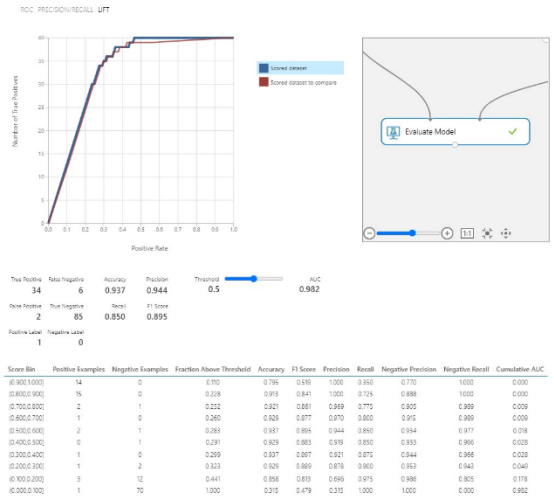


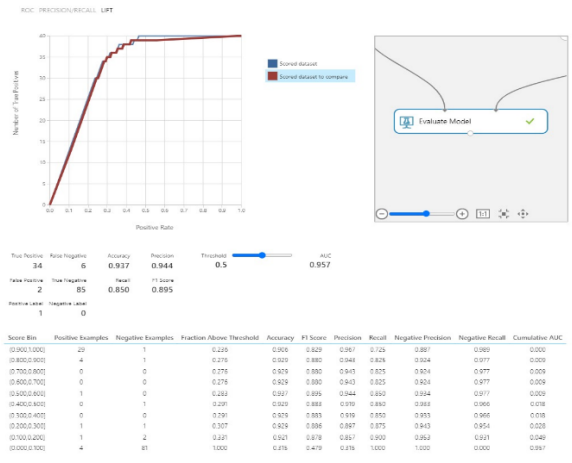
The following are the results of a Two-Class Neural Network for an 80-20 Split without and with applying Hyper Parameter Tuning respectively.



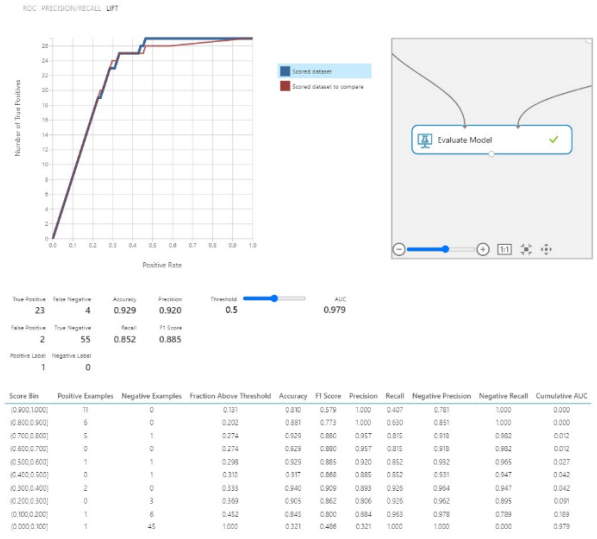


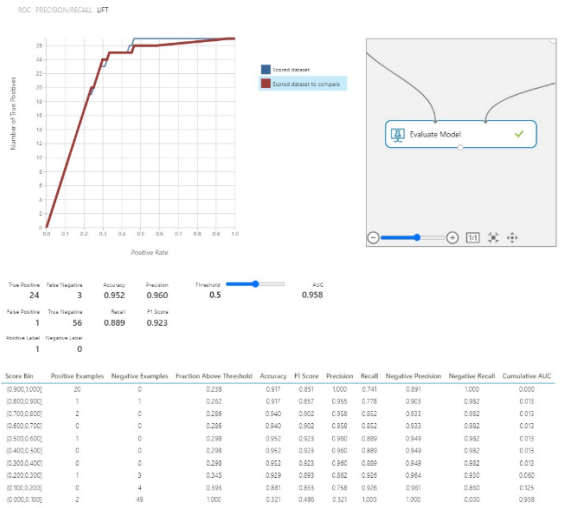
The following are the results of Two-Class Logistic Regression for a 70-30 Split without and with applying Hyper Parameter Tuning respectively.





The following are the results of Two-Class Logistic Regression for an 80-20 Split without and with applying Hyper Parameter Tuning respectively.





To summarize the above results, a table with the respective F-Score follows

Split	Model	Tuning	F-Score
70-30	Logistic Regression	No	0.895
70-30	Logistic Regression	Yes	0.895
70-30	Neural Networks	No	0.872
70-30	Neural Networks	Yes	0.909
80-20	Logistic Regression	No	0.885
80-20	Logistic Regression	Yes	0.923
80-20	Neural Networks	No	0.851
80-20	Neural Networks	Yes	0.857

Conclusion

It is observed that without Hyperparameter Tuning, Logistic Regression has the highest F-score in a 70-30 split and with Hyperparameter Tuning, Logistic Regression has the highest F-score in an 80-20 split. Considering both the splits, with Hyperparameter Tuning, Logistic Regression has the highest F-score in an 80/20 split. Overall, it can be concluded that Logistic Regression is a better model for this dataset. According to additional analysis of these results, the F-score for Logistic Regression in an 80/20 split after Hyperparameter tuning and for Neural Networks in a 70/30 split after Hyperparameter tuning improves significantly.

Future Scope

The dataset is interesting when accuracy and the f score are taken into consideration. In order to evaluate the attributes on real-time data, we would like to spend more time learning the names of the attributes to better comprehend the data.

Reference

1. https://en.wikipedia.org/wiki/Financial_distress
2. https://www.investopedia.com/terms/f/financial_distress.asp#:~:text=Financial%20distress%20is%20a%20condition,revenues%20sensitive%20to%20economic%20downturns
3. https://www.cbinsights.com/research/retail-apocalypse-timeline-infographic/
4. https://www.cbinsights.com/research/corporate-comeback-stories/
5. https://learn.microsoft.com/en-us/azure/machine-learning/v1/how-to-tune-hyperparameters-v1
6. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/two-class-neural-network
7. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/two-class-logistic-regression

Appendix

Financial Distress Actual.csv

Financial Distress Transformed.csv

Correlation Plot.jpeg