DATA MINING

FINAL PROJECT

Team Members

SOHINI CHAKRABORTHY

MADHULIKA CHILLA

SHASHANK REDDY MANDA

Business Problem

Firm collapse prediction has been a subject of interest for almost a century and it still ranks high among the hottest topics in economics. The aim of predicting financial distress is to develop a predictive model that combines various econometric measures and allows one to foresee a financial condition of a firm. The purpose of bankruptcy prediction is to assess the financial condition of a company and its future perspectives within the context of long term operation on the market.

Implementation

Data Files can be found at Kaggle

- bankruptcy_Train.csv — the training set with 64 predictors and 1 target variable
- bankruptcy_Test_X.csv — the test set with ID and 64 predictors.
- bankruptcy_sample_submission.csv — the sample submission with ID and the predicted probability of firm bankruptcy

Exploratory Data Analysis

The second step in our model was to analyze and understand the current data. In order to do this, we used the multiplot node for the following:

- To analyze the kind of skewness of each variable in the data set.
- To identify the outliers and patterns for each variable.
- To study the correlation between the variables.

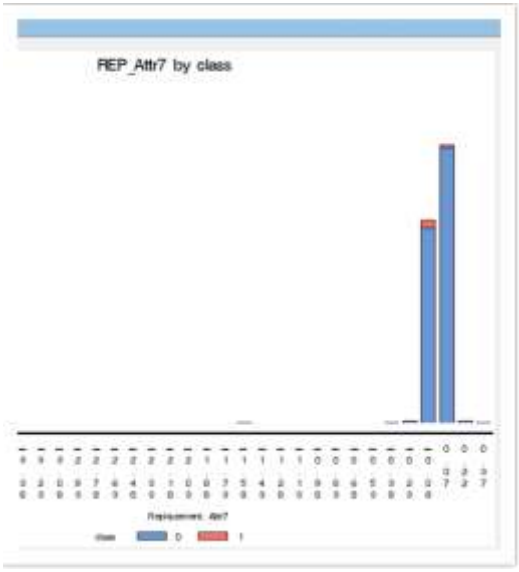


Fig. Multiplot

Data Manipulation:

After analyzing the data, we have performed two types of modification on our data.

- We used replacement node to replace the outliers with the next best value. For this, we performed both standard deviation and MAD methods on the variables and got the best result using a cut-off of two standard deviation on each variable.
- Having observed that our source data consists of variables that are both left and right skewed in different degrees in the multiplot analysis after replacement, we performed different transformation depending on the skewness of that variable. No transformation for 3,14,29

Methods used:

- We handled the left skewness using square method of transformation.
- We performed natural log transformation for moderate right skewness and log10 for highly right skewed variables.

Name	Method	Number of Bins	Role	Level
REP_Attr45	Square	4	Input	Interval
REP_Attr46	Default	4	Input	Interval
REP_Attr47	Default	4	Input	Interval
REP_Attr48	Square	4	Input	Interval
REP_Attr49	Square	4	Input	Interval
REP_Attr5	Square	4	Input	Interval
REP_Attr50	Default	4	Input	Interval
REP_Attr51	Default	4	Input	Interval
REP_Attr52	Default	4	Input	Interval
REP_Attr53	Default	4	Input	Interval
REP_Attr54	Default	4	Input	Interval
REP_Attr55	Default	4	Input	Interval
REP_Attr56	Square	4	Input	Interval
REP_Attr57	Square	4	Input	Interval
REP_Attr58	Default	4	Input	Interval
REP_Attr59	Default	4	Input	Interval
REP_Attr6	Default	4	Input	Interval
REP_Attr60	Default	4	Input	Interval
REP_Attr61	Default	4	Input	Interval
REP_Attr62	Default	4	Input	Interval
REP_Attr63	Default	4	Input	Interval
REP_Attr64	Default	4	Input	Interval
REP_Attr7	Square	4	Input	Interval
REP_Attr8	Default	4	Input	Interval
REP_Attr9	Default	4	Input	Interval
class	Default	4	Target	Binary

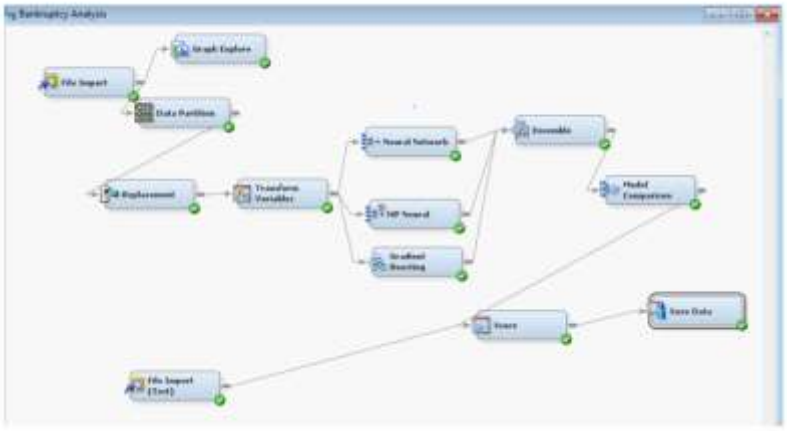
Initial model after performing the above modifications:

Model 1: Tried modelling the data using different predictive modelling techniques like logistic regression, decision tree, HP Random Forest, bagging but couldn't achieve the desired outcome.

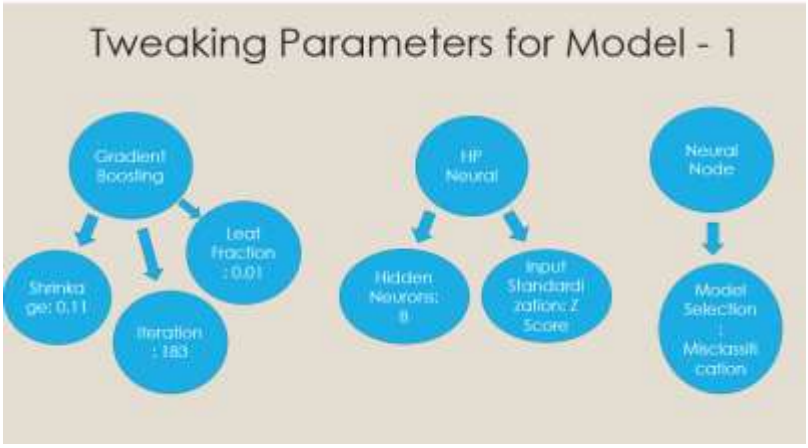
The following models were able to give us the desired outcome partially:

- Gradient boosting
- HP Neural
- Neural Network

We observed that the Neural Network was working good in predicting the 1's and Hp neural was performing good with 0's. Hence, we used an ensemble of these three models.



Tweaking Parameters for Model – 1 to improve the performance



OPPORTUNITIES OF IMPROVEMENT IN MODEL1

- Observed that the misclassification rate was high. And on analyzing the errors we noticed that our models were having high False negative rate.
- The valid ROC index for the model was 0.922.
- To further improve the misclassification rate of the first model, worked on second model by tweaking the parameters, changing the data partition and modified the transformation method on the variables.

Model Description	False Negative	True Negative	False Positive	True Positive
Train	110	8797	4	87
Validation	5	977	2	8

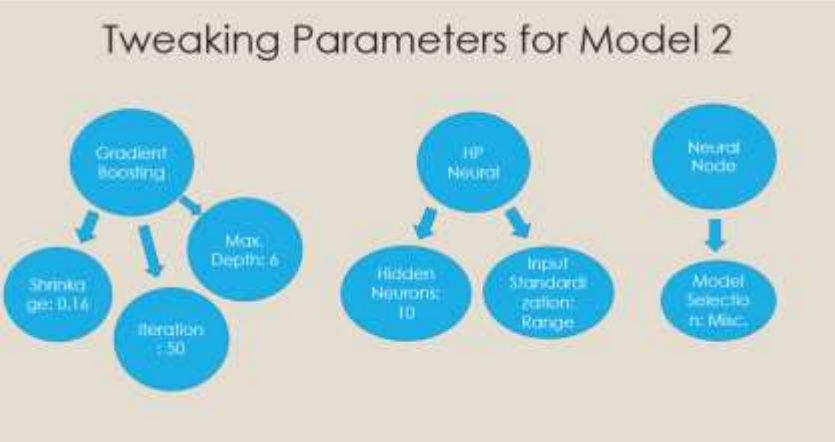
Enhancements in Model 2:

- First, we modified the data partition to 90% training and 10% validation.
- The variables in the transformation node were modified according to the graph explorer node that tells the skewness for each of them.
- For left skewness, we used square and for right skewed inverse transformations.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Roc Index	Train: Misclassification Rate	Valid: Misclassification Rate	Target Variable ▲
Y	Ensmbl	Ensmbl	Ensemble	0.971	0.010669	0.01996	class

Final model parameters used for attaining better accuracy for model 2:



Key Learnings:

- Learnt about how the transformation techniques can improve the fit of a model by stabilizing the variables, correcting the non-normality and removing the non-linearity in the variables.
- Learnt about how the parameters matter in different models that we have tried.
- Learnt how tweaking the parameters can lead us to an optimized model.