Bankruptcy Analysis

Goal| Develop a predictive model that combines various econometric measures and allows one to foresee a financial condition of a firm

Data Exploration

- Perform EDA and understand the characteristics of the data
- Identify the most important variables that affect the outcome

Data Cleaning

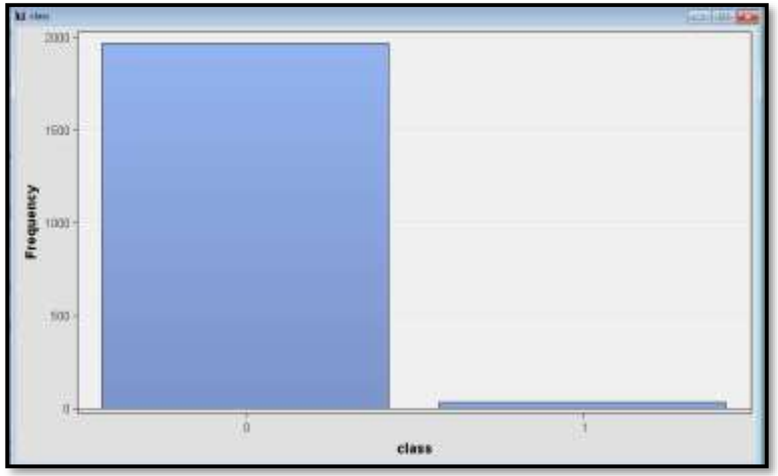
- Checking missing values and impute if needed
- Outliers analysis
- Transformations of variables

Model Building

- Developing a baseline model
- Comparison with multiple models
- Choosing the best model based on the evaluation criteria

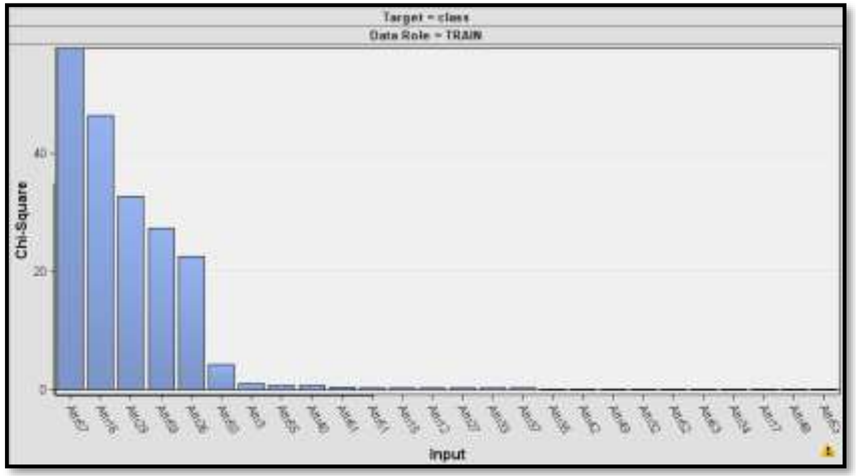
Data Exploration| Distribution of target variables and chi-square values are explored to understand the data better

Target variable distribution



- There are 65 columns in our dataset out of which there are 64 interval variables and one binary variable- class.
- We have a highly imbalanced dataset with only 2.2% of values having bankruptcy
- There are no missing values in our dataset

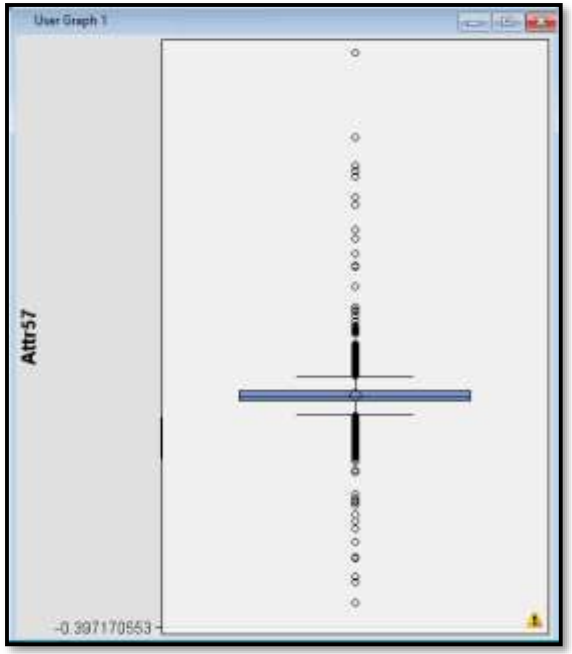
Chi-square values

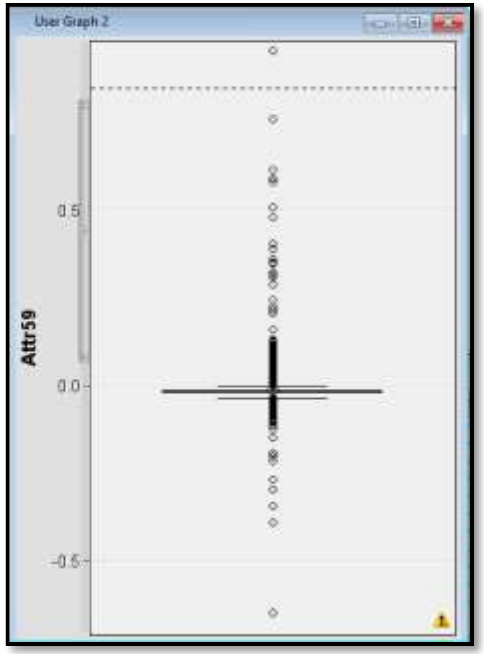


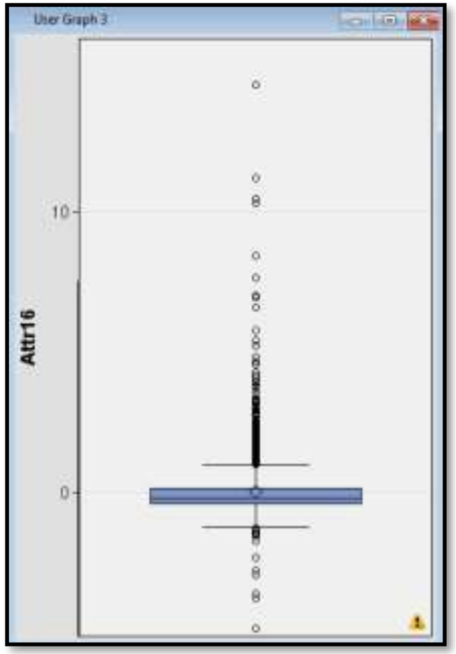
Input	Cramer's V	Prob	Chi-Square
Attr57	0.096646	<.0001	93.4038
Attr16	0.068142	<.0001	46.4327
Attr29	0.057247	<.0001	32.7718
Attr59	0.052345	<.0001	27.4003
Attr26	0.047525	0.0002	22.5866
Attr50	0.02042	0.3835	4.1700

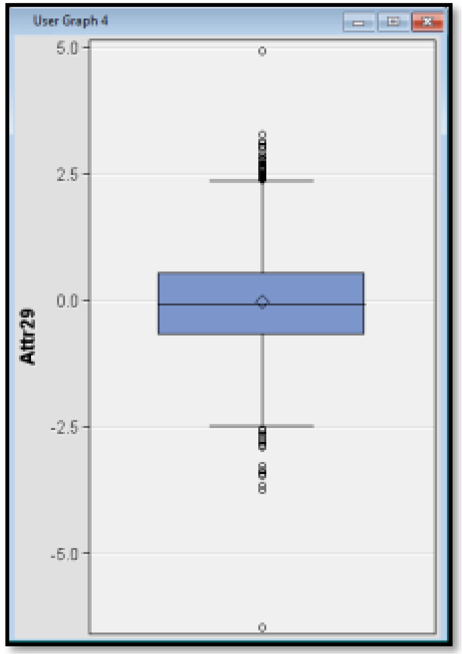
- Attributes 57, 16, 29, 59 and 26 are the important variables that influence the target variable based on the chi-square plot and the p-values

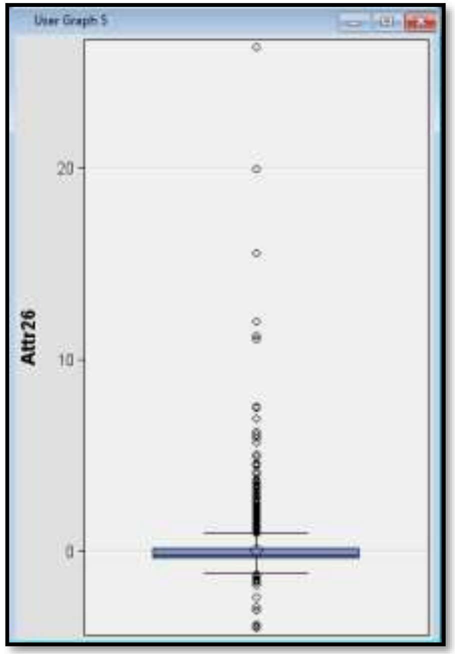
Data Exploration| Outliers are observed across all the variables from the results of chi-square test











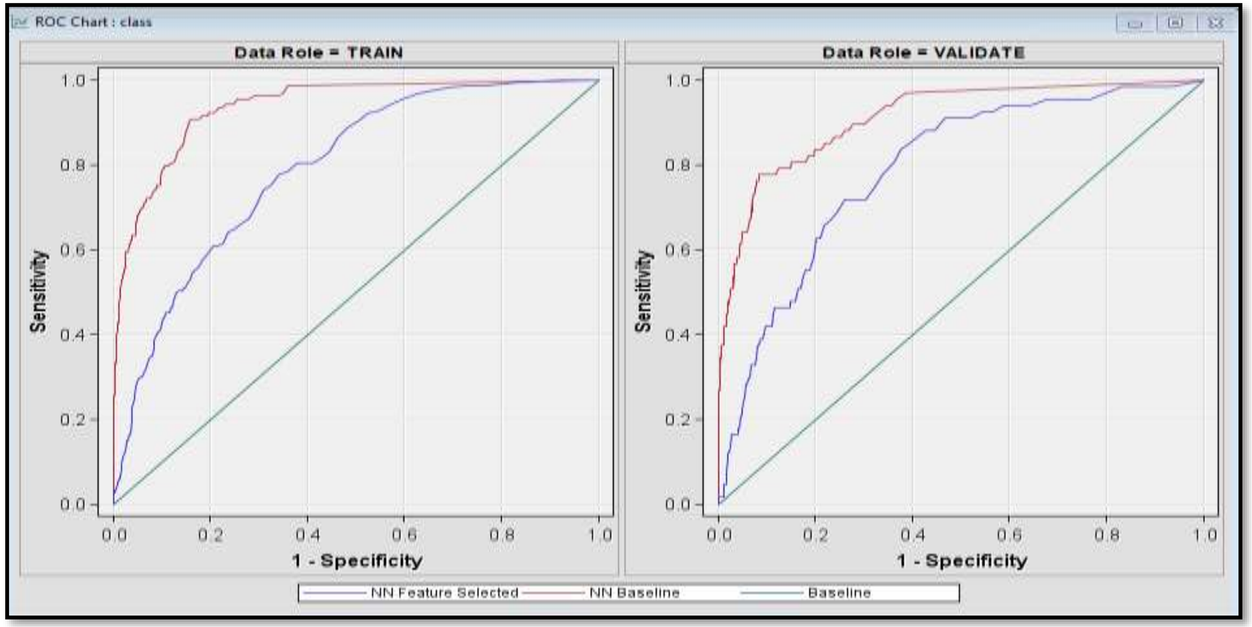
- Based on the results of the chi-square values, outliers are checked for the 5 most important variables
- There are outliers present for all the variables but would need to explore them further in order to understand their impact on the model

Baseline model| Neural network with all features performs better than neural network with features selected based on chi-square

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Roc Index	Selection Criterion: Valid: Misclassification Rate	Tr De Fr
Y	Neural2	Neural2	NN Baseline	class		0.914	0.01866	
	Neural3	Neural3	NN Feature...	class		0.787	0.022326	

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Neural2	NN Baseline	TRAIN	class		121	6839	7	32
Neural2	NN Baseline	VALIDATE	class		49	2927	7	18
Neural3	NN Feature Selected	TRAIN	class		151	6846	0	2
Neural3	NN Feature Selected	VALIDATE	class		67	2934	0	.



Model Selection Criteria:
I used ROC index, and number of false negatives as the two most important metrics given that the dataset is highly imbalanced

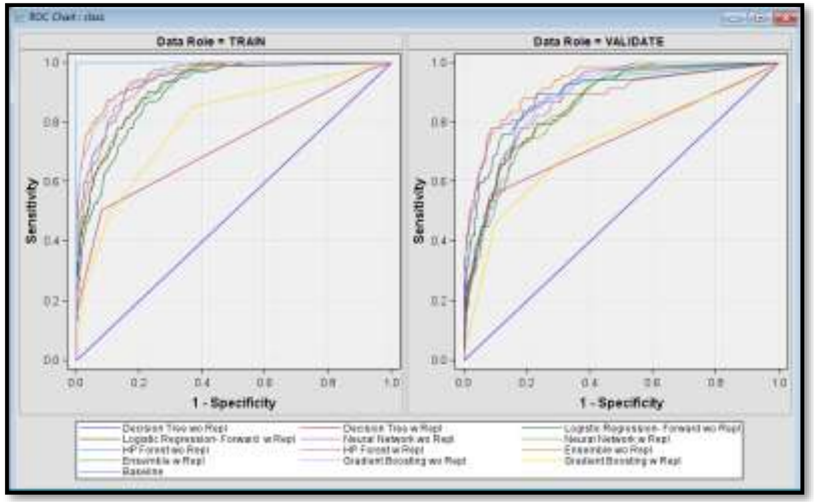
- Based on the chi-square values, I built 2 baseline models- neural network with feature selection and neural network without feature selection
- I observed that neural network without feature selection has a better ROC, misclassification rate, and lesser false negatives indicating that all the variables are needed to accurately predict the output

Model Comparison| Ensemble model with logistic regression, gradient boosting and neural networks gave the best ROC score among all models

Model 1- Baseline Neural Network → Model 2- Model comparison → Model 3- arriving at ensemble

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable ▲	Target Label	Selection Criterion: Valid: Roc Index	Valid: Roc Index
Y	Ensmbl2	Ensmbl2	Ensemble wo Repl	class		0.926	0.926
	Neural4	Neural4	Neural Network wo Repl	class		0.914	0.914
	Reg2	Reg2	Logistic Regression- Forward ...	class		0.899	0.899
	Boost2	Boost2	Gradient Boosting wo Repl	class		0.883	0.883
	HPDMFore...	HPDMFore...	HP Forest wo Repl	class		0.866	0.866
	Ensmbl	Ensmbl	Ensemble w Repl	class		0.862	0.862
	Neural	Neural	Neural Network w Repl	class		0.858	0.858
	Reg	Reg	Logistic Regression- Forward ...	class		0.853	0.853
	HPDMForest	HPDMForest	HP Forest w Repl	class		0.852	0.852
	Tree	Tree	Decision Tree w Repl	class		0.737	0.737
	Boost	Boost	Gradient Boosting w Repl	class		0.726	0.726
	Tree2	Tree2	Decision Tree wo Repl	class		0.5	0.5

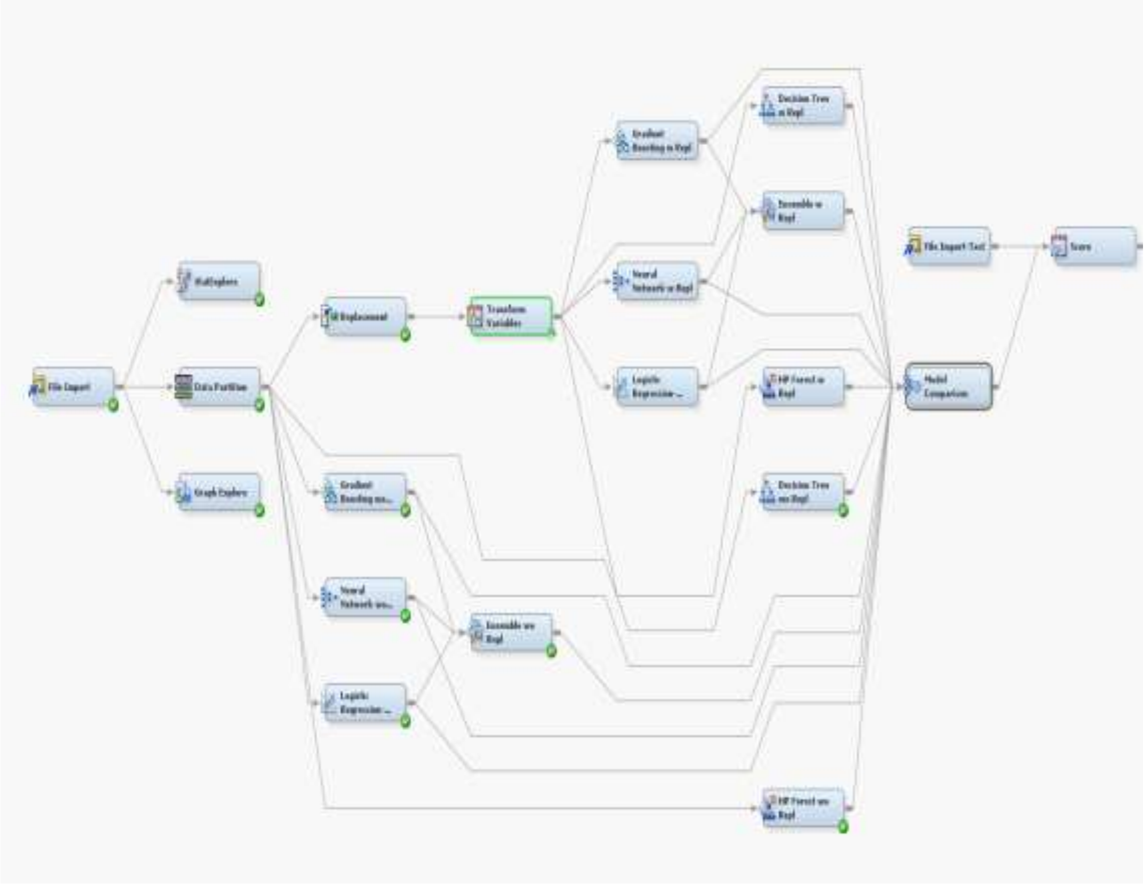


Event Classification Table
Model Selection based on Valid: Roc Index (_VAUR_)

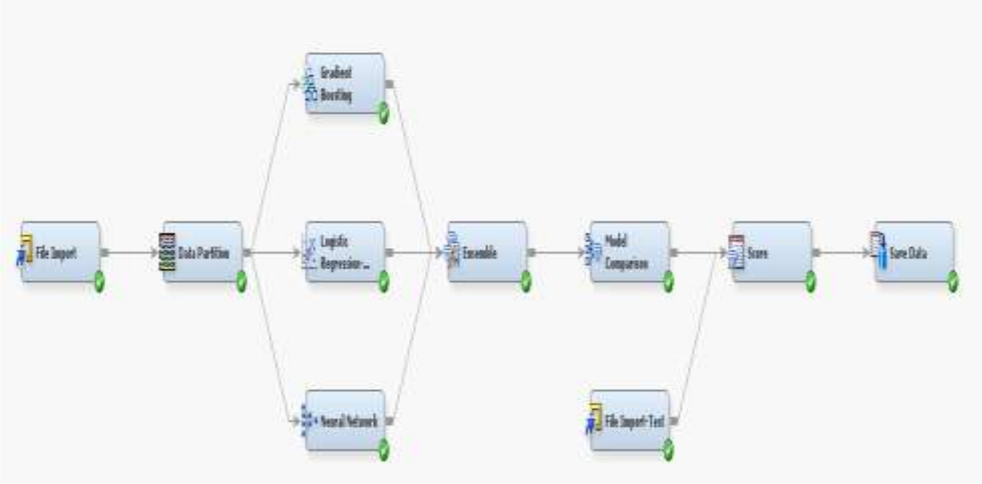
Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Ensmbl	Ensemble w Repl	TRAIN	class		150	6846	0	3
Ensmbl	Ensemble w Repl	VALIDATE	class		67	2934	0	.
Boost	Gradient Boosting w Repl	TRAIN	class		153	6846	0	0
Boost	Gradient Boosting w Repl	VALIDATE	class		67	2934	0	0
Neural	Neural Network w Repl	TRAIN	class		150	6846	.	3
Neural	Neural Network w Repl	VALIDATE	class		65	2933	1	2
Reg	Logistic Regression- Forward w Repl	TRAIN	class		126	6839	7	27
Reg	Logistic Regression- Forward w Repl	VALIDATE	class		60	2924	10	7
Tree	Decision Tree w Repl	TRAIN	class		130	6835	11	23
Tree	Decision Tree w Repl	VALIDATE	class		56	2926	8	11
HPDMForest	HP Forest w Repl	TRAIN	class		25	6846	0	128
HPDMForest	HP Forest w Repl	VALIDATE	class		66	2934	0	1
Boost2	Gradient Boosting wo Repl	TRAIN	class		131	6842	4	22
Boost2	Gradient Boosting wo Repl	VALIDATE	class		63	2932	2	4
Neural4	Neural Network wo Repl	TRAIN	class		121	6839	7	32
Neural4	Neural Network wo Repl	VALIDATE	class		49	2927	7	18
Reg2	Logistic Regression- Forward wo Repl	TRAIN	class		140	6836	10	13
Reg2	Logistic Regression- Forward wo Repl	VALIDATE	class		61	2925	9	6
Tree2	Decision Tree wo Repl	TRAIN	class		153	6846	0	0
Tree2	Decision Tree wo Repl	VALIDATE	class		67	2934	0	0
HPDMForest2	HP Forest wo Repl	TRAIN	class		1	6846	.	152
HPDMForest2	HP Forest wo Repl	VALIDATE	class		66	2931	3	1
Ensmbl2	Ensemble wo Repl	TRAIN	class		136	6846	.	17
Ensmbl2	Ensemble wo Repl	VALIDATE	class		62	2931	3	5

- **Model 2**:
  - The baseline model is compared against decision tree, logistic regression, gradient boosting and random forest
  - The top 3 models obtained are Neural Network, Gradient boosting and logistic regression with forward propagation based on the ROC index
- **Model 3/ Final Model**:
  - The top 3 models ie logistic regression, gradient boosting and neural network is combined to arrive at the ensemble model
  - Ensemble model gave an ROC score of 92.6% which is better than all the models

Model Tuning| Models without outlier replacement, transformations perform better than models with outlier and transformations



Final Model Obtained:



- Models are also tested with outlier treatment and transformations to check if we obtain better results in terms of ROC and false negatives, however this is not the case
- Model parameters for Neural Network like no of hidden units, activation function were changed to test the models but the default parameters gave the best results.

Final Results and Learnings

- **Data Partition:** Starting at a 60:40 ratio, as I increased the partition to 70:30, the validation ROC index improved since there was more data available for the model to train, but further increasing to 80:20, the validation ROC index decreased as the model was overfitting to the training dataset.

- **Neural Networks:** NN may be prone to overfitting since there are thousands of parameters to learn in a complex model and if the training data is limited then it may also be learning the noise in the data. Consequently, NNs work best for large datasets with lots of variables.

- **Ensemble methods:** While individual methods did not yield a significant improvement in and of themselves, ensembling the best-performing methods yielded the best possible model. Changing the parameters in the individual models may have improved the performance of the ensemble model.