Data Mining – Credit Default Prediction

Using SAS Enterprise Miner

Cheeky Miners

Rohit Soans | Nalin Singh

PROJECT MAP

Our approach to the bankruptcy classification problem

Classification Problem - Project Workflow

Exploratory Data Analysis (EDA) –
to identify different trends in the raw data

Data Modification –
preparing our data for modelling

Data Modelling –
running different models to train our data

Cross Validation –
public v. private leaderboard score

Way Forward –
major learnings from the project

Project Objective

APPLY CLASSIFICATION ALGORITHMS TO FORECAST IF A CLIENT WILL DEFAULT

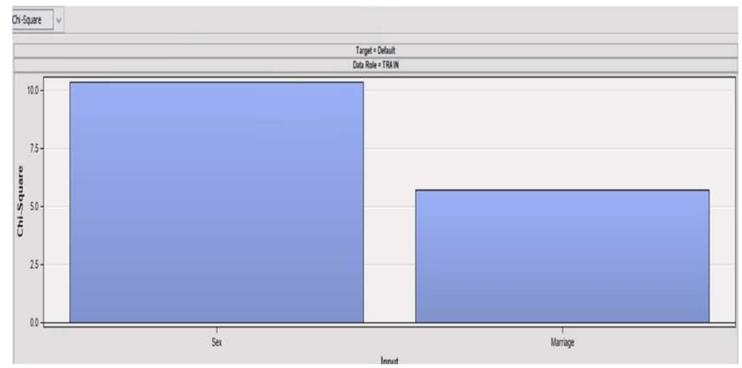
Data Pre-Processing

Exploratory Data Analysis (EDA)

Identifying patterns in our data

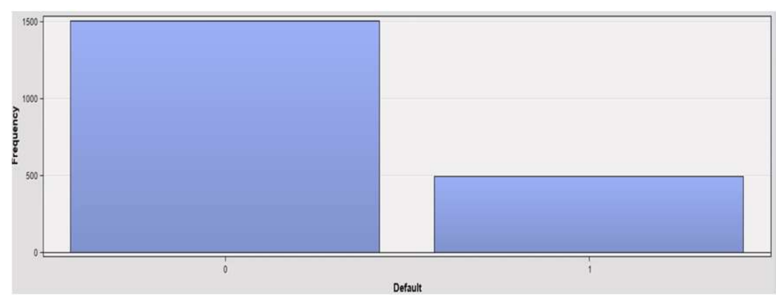
Limit	Sex	Education	Marriage	Age	Status_1	Statement_1	Payment_1	Status_2	Statement_2	Payment_2	Status_3	Statement_3	Payment_3	Status_4	Statement_4	Payment_4	Status_5
1.0203808...	1	11		0.2732249...	1	105935.38	5138.86	2	100815.86	30022.02	2	103744.47	5008.69	2	103423.25	10017.1	
-0.5203582...	1	32		-1.1369961...	2	76068.75	2896.01	2	69335.51	2090.18	2	54884.9	1514.48	2	55385.28	1124.87	
-0.21221041	1	22		0.0562678...	2	58165.94	2503.78	2	59695.81	3002.56	2	61732.14	2303.06	2	62412.93	3001.5	
0.1729743...	0	42		-0.5946034...	2	198098.28	4515.53	2	194580.41	15466.96	2	202870.47	4702.43	2	198938.2	47450.88	
0.7122330...	1	12		1.0325747...	1	6062.43	0	1	-899.31	0	0	-899.19	2808.72	1	1906.53	6.83	
0.2500113...	1	21		1.3580104...	2	205153.31	0	2	0	0	0	0	0	0	0	0	
0.1729743...	0	51		0.4901820...	2	58783.09	3101.07	2	59881.41	80000.33	2	136994.15	5003.64	2	118269.45	4000.92	
1.3285286...	0	12		-0.0522106...	1	320.78	3075.45	1	3077.29	3003.84	1	3002.83	2613.9	1	2615.23	1807.83	
2.5611199...	0	32		1.1410533...	1	24239.84	15000.9	2	36676.83	1560.35	1	1562.52	310852.35	1	310852.56	10001.92	
0.0189004...	1	12		-0.5946034...	2	129848.98	5115.75	2	89690.14	2500.39	2	31050.39	0	2	0	0	
-0.4433212...	0	12		-0.7030819...	0	7399.51	2001.04	0	-16.91	0	0	-12.03	0	0	-16.88	0	
-0.9055429...	0	21		-0.1606892...	2	2793.46	2001.74	2	4565.44	1201.21	2	5000.99	0	2	0	0	
0.3270482...	0	22		-0.1606892...	1	8548.34	2998.79	1	2990.9	4585.58	1	4572.53	3274.72	1	3268.39	8649.54	
2.2529721...	0	12		-0.1606892...	1	14461	8869.89	1	8869.47	23665.03	1	21494.64	12660.61	1	12657.8	5752.89	
2.5611199...	0	21		0.5986605...	2	131721.34	3900.4	2	103746.44	3201.43	2	91805.07	2768.28	2	76152.08	2103.86	
-0.21221041	0	21		-0.1606892...	1	56705.67	3000	2	53584.81	3003.81	2	54794.81	3000.76	2	55952.7	3000.1	
-0.9055429...	0	21		-0.3776463...	1	391.66	390.97	1	390.17	391	1	390.32	394.21	1	391.65	933.67	
-1.0596169...	1	12		-1.2454747...	2	28440.38	1801.25	2	28848.92	3302.33	2	31072.89	0	4	30249.86	1501	
-0.0581365...	1	21		1.5749675...	1	269.97	0	1	0	698.74	1	697.12	697.96	1	698.92	897.51	
-0.2892473...	1	22		-1.0285176...	4	137204.82	7002.38	4	139190.92	0	4	133523.62	5133.7	2	133183.45	5002.76	

Understanding Data – We noticed that there was no missing values in the data



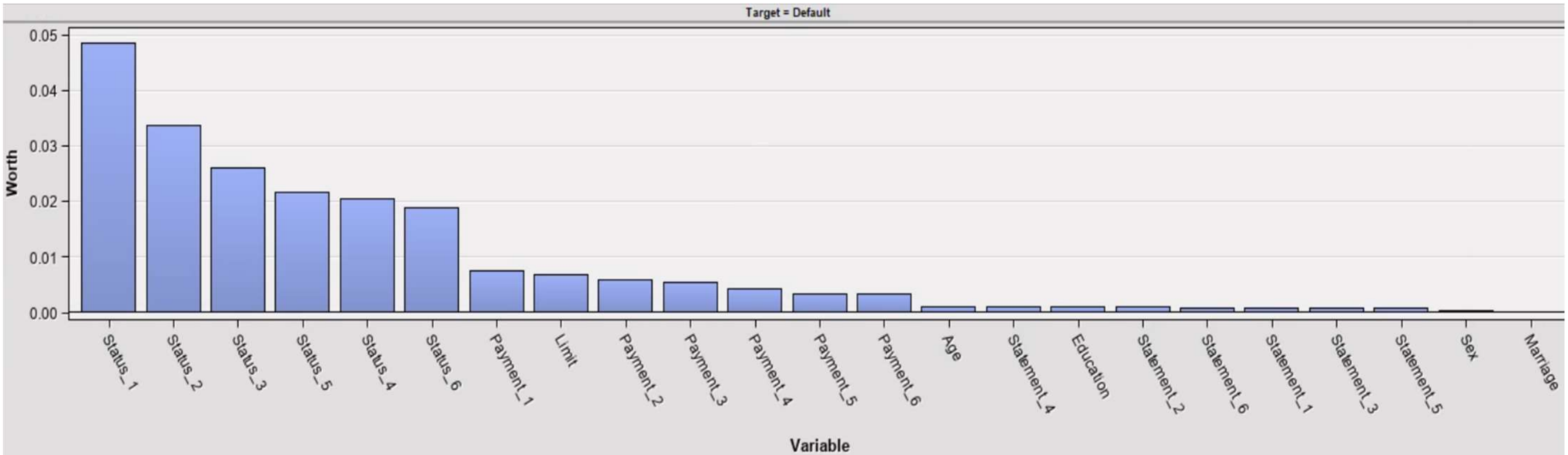
Important Class Attributes – As per the Chi Square test we notice that Sex is strongly associated than Marriage with our target variable.

Imbalanced Target Variable – Frequency of 0 in the target variable is significantly higher than that of 1, creating an imbalanced data. It will be an **important factor** when we create our model.



Exploratory Data Analysis (EDA)

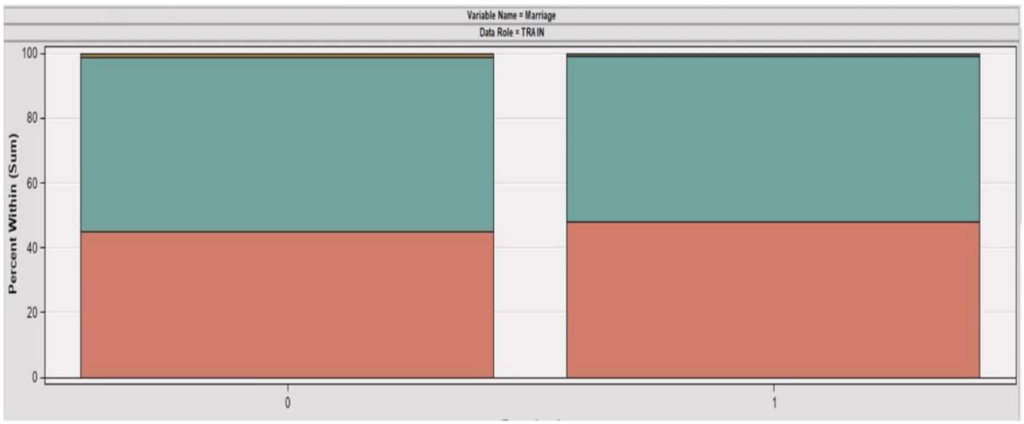
Identifying patterns in our data

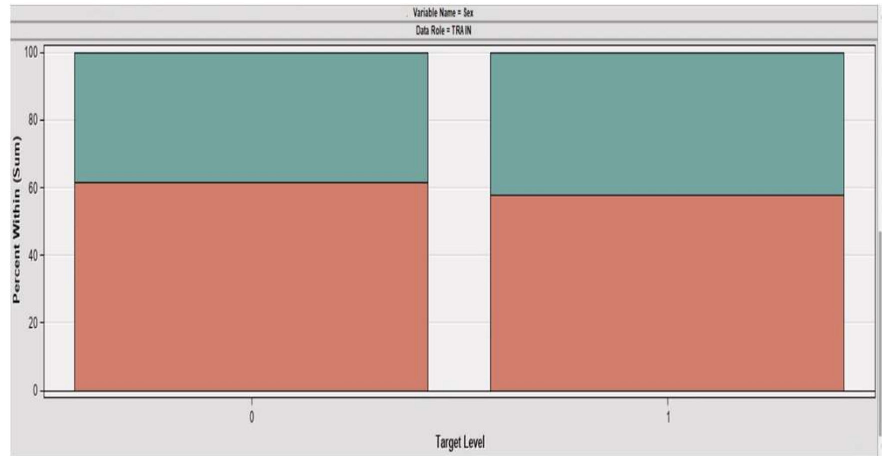


Important Attributes – As per the variable worth graph, most important attributes are: **Status attributes**

EXPLORATORY DATA ANALYSIS (EDA)

Identifying patterns in our data





Marriage and Sex appear to not have a clear separation with our target variable

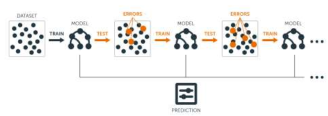
Data Modelling & Cross Validation

Data Modelling

Running different models to fit our data

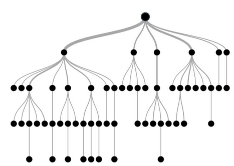
Total Models Deployed

1 Gradient Boosting



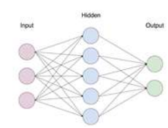
- Tuning parameters tried: **no. of iterations, shrinkage, leaf fraction, max depth & reuse variables.**

2 HP Forest



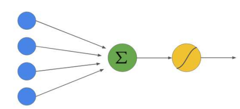
- Tuning parameters tried: **max number of tress, max depth, split size & min category size.**

3 Neural Network



- Tuning parameters tried: **number of hidden layers, max iterations, max time & model selection criterion.**

4 Logistic Regression

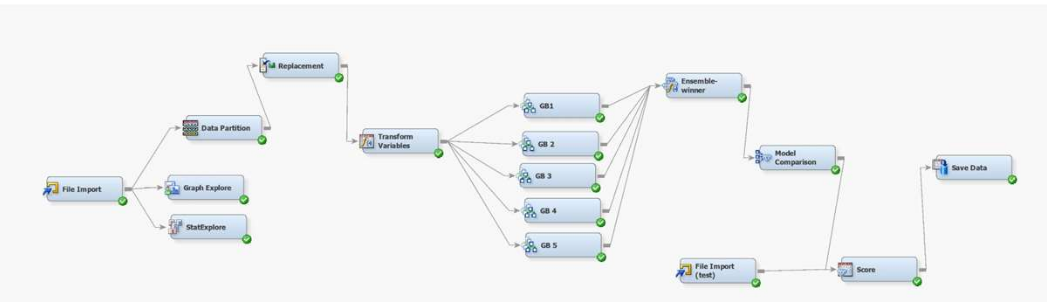


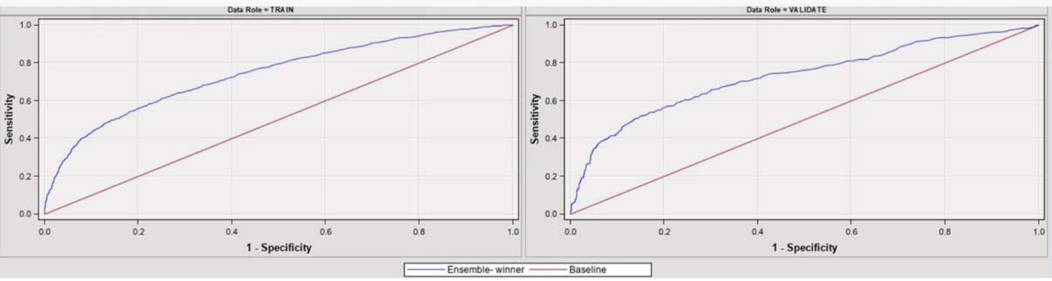
- Tuning parameters tried: **input coding method**. In addition, used both **impute & transform** nodes before running logistic regression.

DATA MODELLING

Running different models to fit our data

Best Model #1



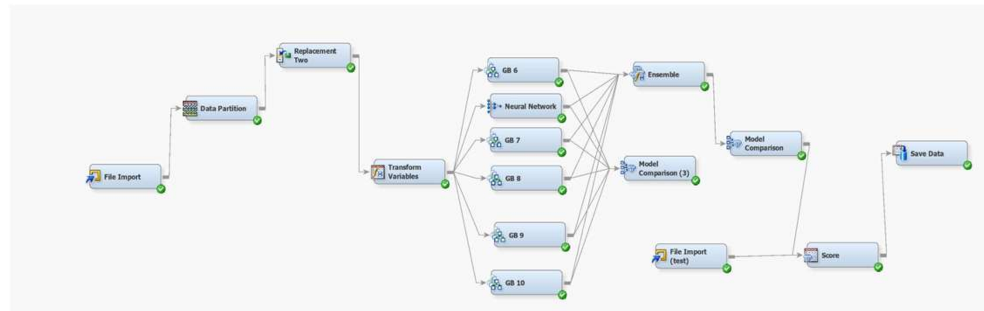


Valid ROC Index: .727

Public Score: .74631

Private Score: .74843

Best Model #2



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	Train: Misclassification Rate
Y	Boost7	Boost7	GB 7	TI_Default1	Default 1	0.734	8998	17996	0.212269
	Boost8	Boost8	GB 8	TI_Default1	Default 1	0.732	8998	17996	0.210158
	Boost6	Boost6	GB 6	TI_Default1	Default 1	0.732	8998	17996	0.210936
	Boost9	Boost9	GB 9	TI_Default1	Default 1	0.731	8998	17996	0.211714
	Boost10	Boost10	GB 10	TI_Default1	Default 1	0.73	8998	17996	0.211825
	Neural	Neural	Neural Net...	TI_Default1	Default 1	0.727	8998	17996	0.206935

Valid ROC Index: .735

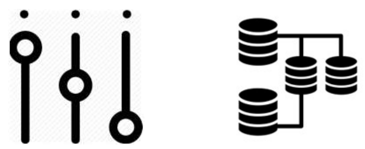
Public Score: .74617

Learnings & Way Forward

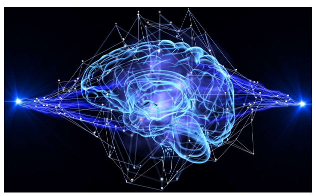
WAY FORWARD

Major learnings from public & private leaderboard score difference

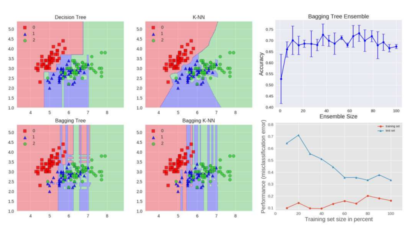
Gradient Boosting Tuning



Neural Network



Model Ensemble



1. Shrinkage seems to be a **balance parameter** as we increased number of iterations. Increasing number of iterations can lead to a better model fit if it is balanced with a decrease in shrinkage.

2. In addition to number of iterations, max tree depth can be used as a **regularization parameter**.

3. Better combination of **combination and activation functions(Combination = Linear, Activation= mlogistic)** for target layer in neural network can be used for better fit.

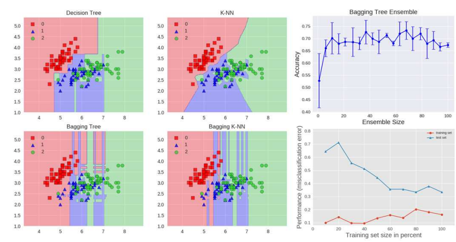
4. Adjusting the **number of hidden layers/ units** can also be used as a balancing parameter for better fit.

5. Combining different models together can help **reduce variance** in our model fit. For classification, voting could have been used as a judging parameter

6. Enables to train diverse models and use a combination of models to stabilize the overall performance and hence **increase accuracy**.

Way Forward

Major learnings from public & private leaderboard score difference



Model Ensemble

1. Combining different models together can help **reduce variance** in our model fit. For classification, voting could have been used as a judging parameter

2. Enables to train diverse models and use a combination of models to stabilize the overall performance and hence **increase accuracy**.

3. **Most Important Learning** – In order to get a high score on the Public Leaderboard, we tried different models making changes in iterations, shrinkage, data partition, etc., which led to overfitting of the data causing high variations in the final leaderboard.

Thank You!