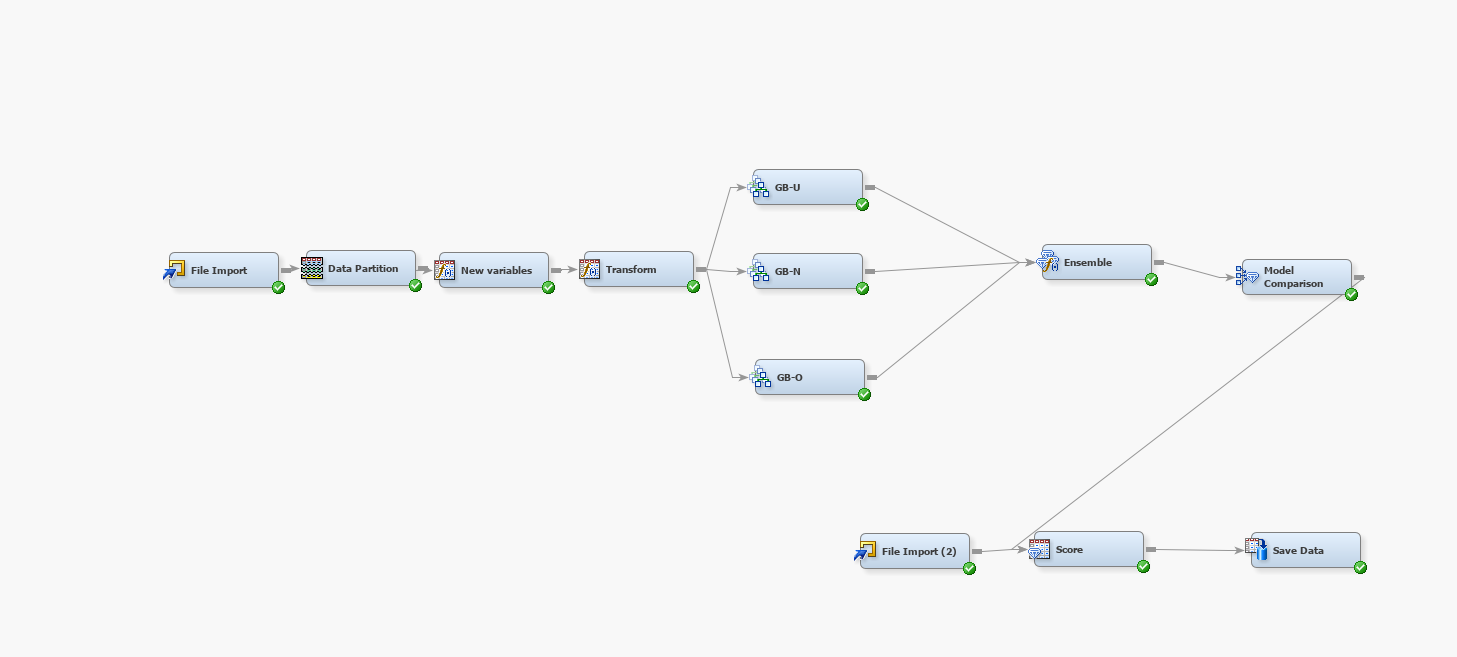
# **DATA MINING PROJECT**

* **Team: Minotaur**

**Winner (Model 1):**



For the best model we used the following nodes-

File Import 🡪 Data Partition 🡪New Variable Generator 🡪 Transformation 🡪3 Gradient Boosting Models 🡪 Ensemble 🡪 Model Comparison 🡪 Score 🡪 Save Data

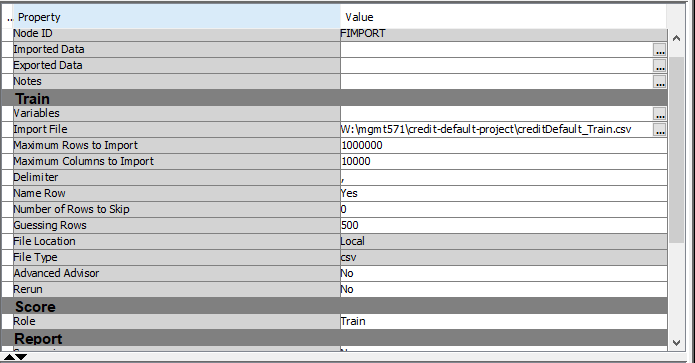
We created a new variable “Percentage Paid” which is the ratio of Payment to Statement. Thus, we created 6 new variables which played an important variable during model training.

Furthermore, we used an ensemble of 3 different gradient boosting algorithms –

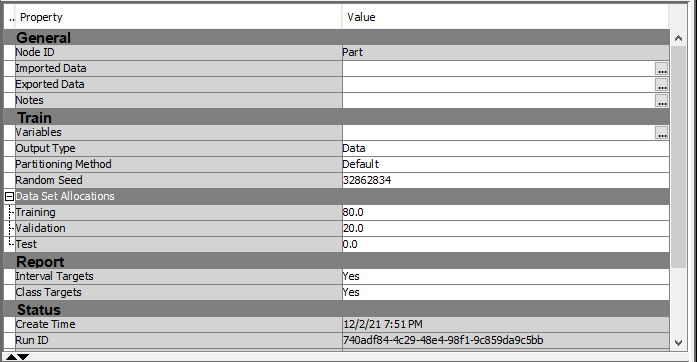
1. An overfitting Algorithm
2. An underfitting Algorithm
3. A normal Algorithm

The following screenshots are provided for each and every node :–

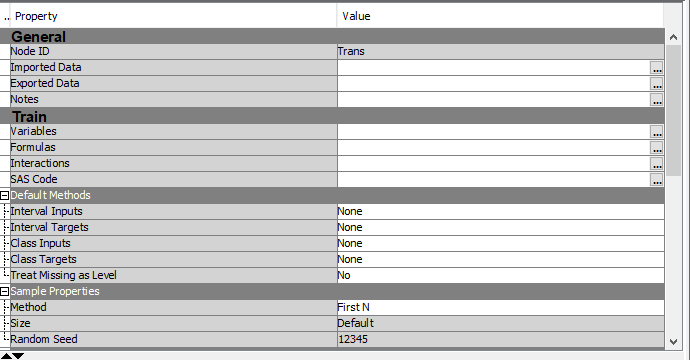
**File Import Node:**



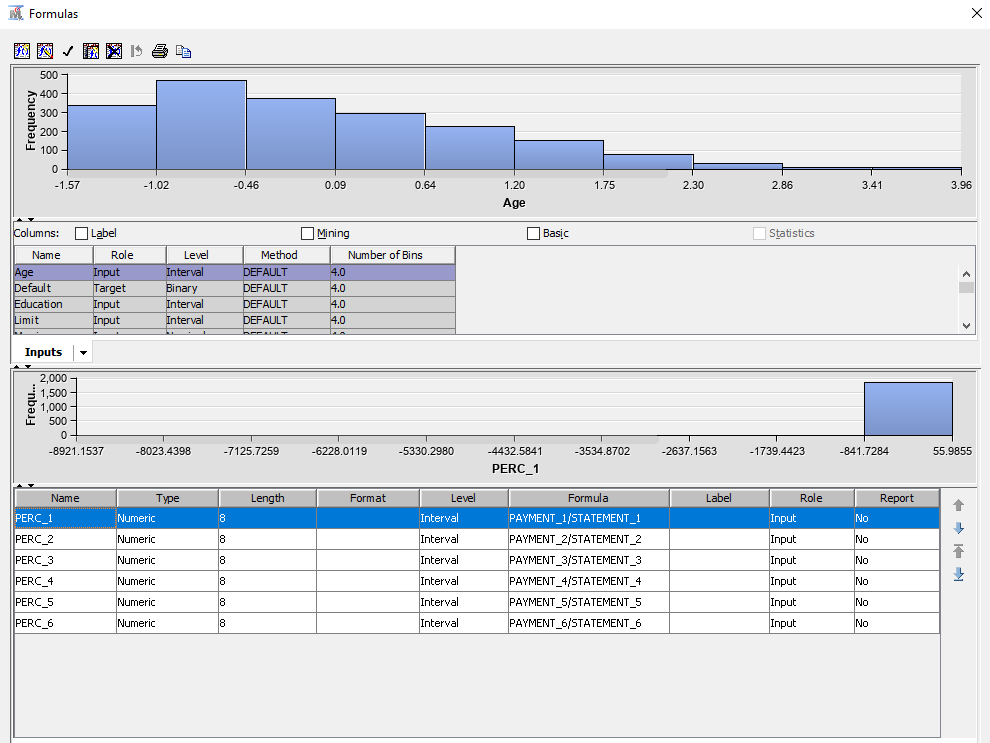
**Data Partition Node-**

****

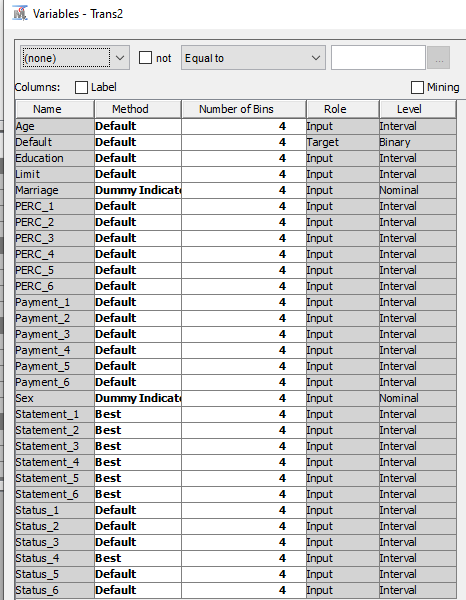
**Transformation node-**



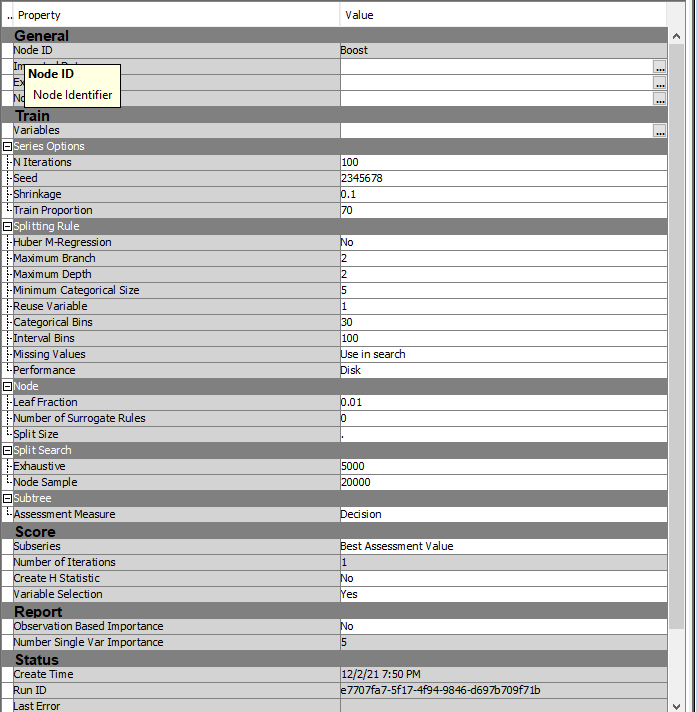
**Created new variables as shown below-**



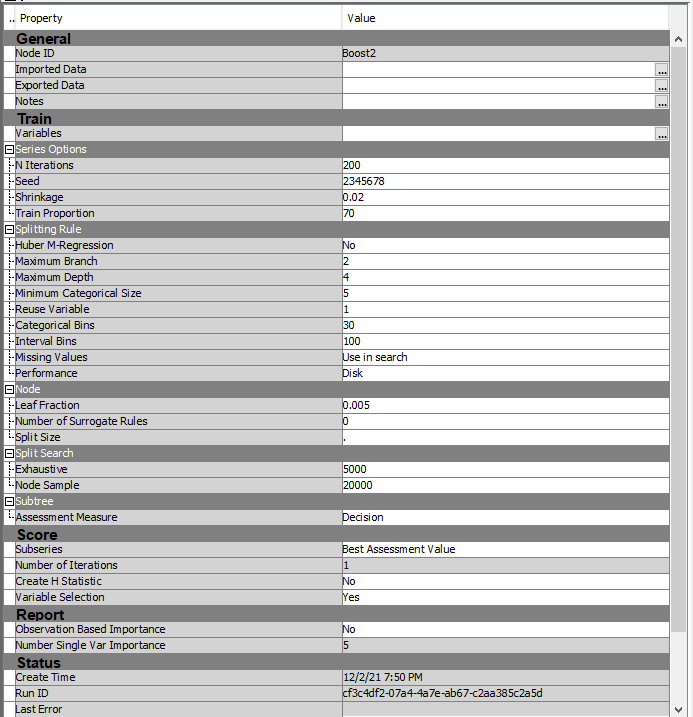
**Second Transformation Node -**



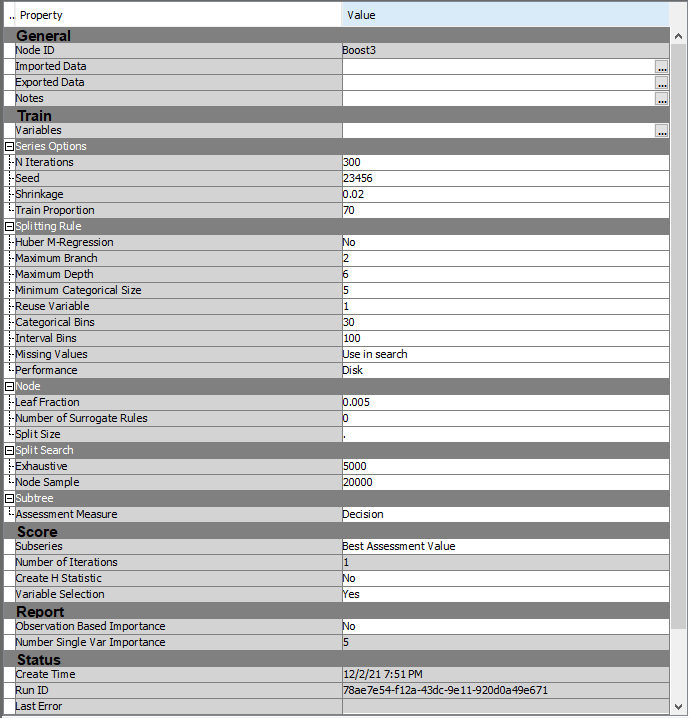
**Underfitting Gradient Boost Model**



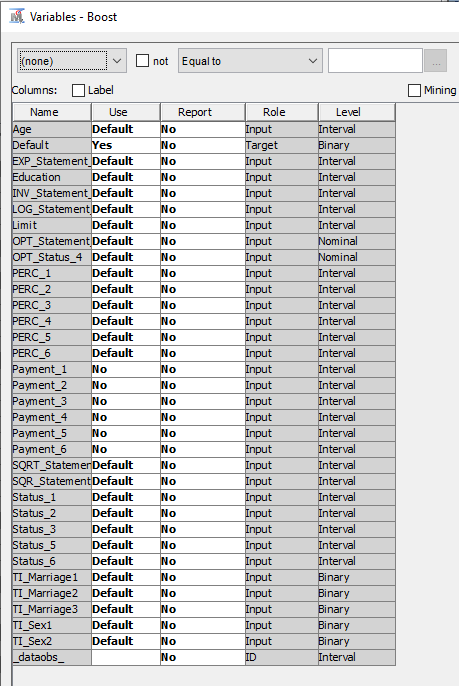
**Normal Fit Gradient Boosting Model-**



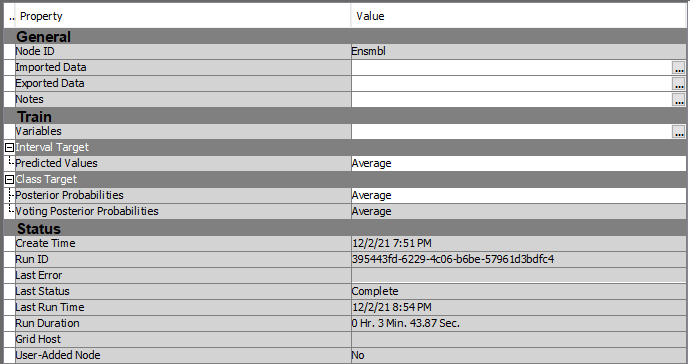
**Overfitting Gradient Boosting Model-**



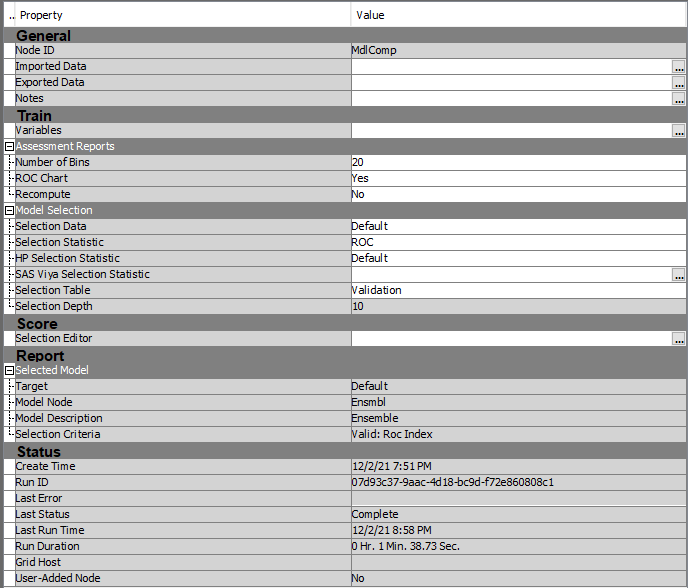
**Variable selection for the 3 Gradient Boosting models**



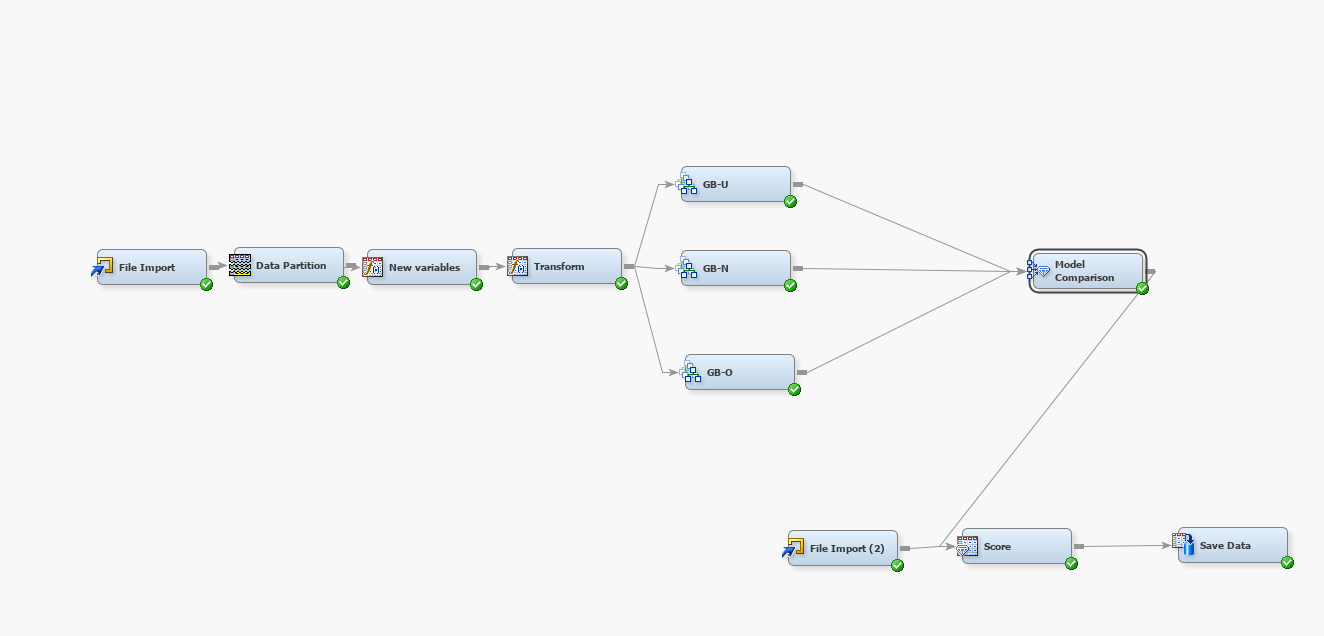
**Ensemble-**



**Model Comparison**



**Model 2**



For the second best model we used the following nodes-

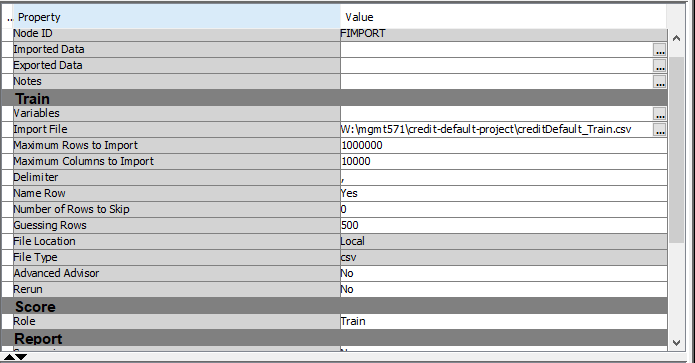
File Import 🡪 Data Partition 🡪New Variable Generator 🡪 Transformation 🡪3 Gradient Boosting Models 🡪 Model Comparison 🡪 Score 🡪 Save Data

We created a new variable “Percentage Paid” which is the ratio of Payment to Statement. Thus, we created 6 new variables which played an important variable during model training.

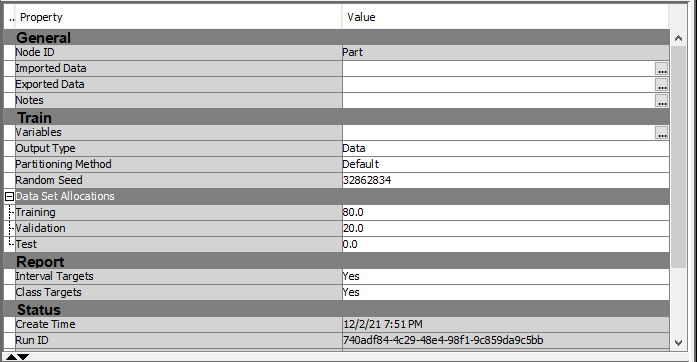
In this model we didn’t use an ensemble with the best model being the normal fitting model.

The following screenshots are provided for each and every node: –

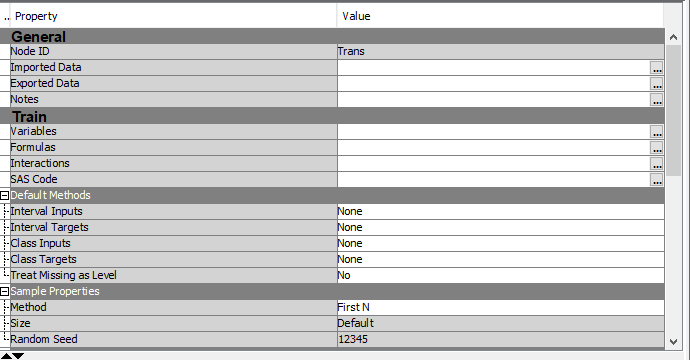
**File Import Node:**



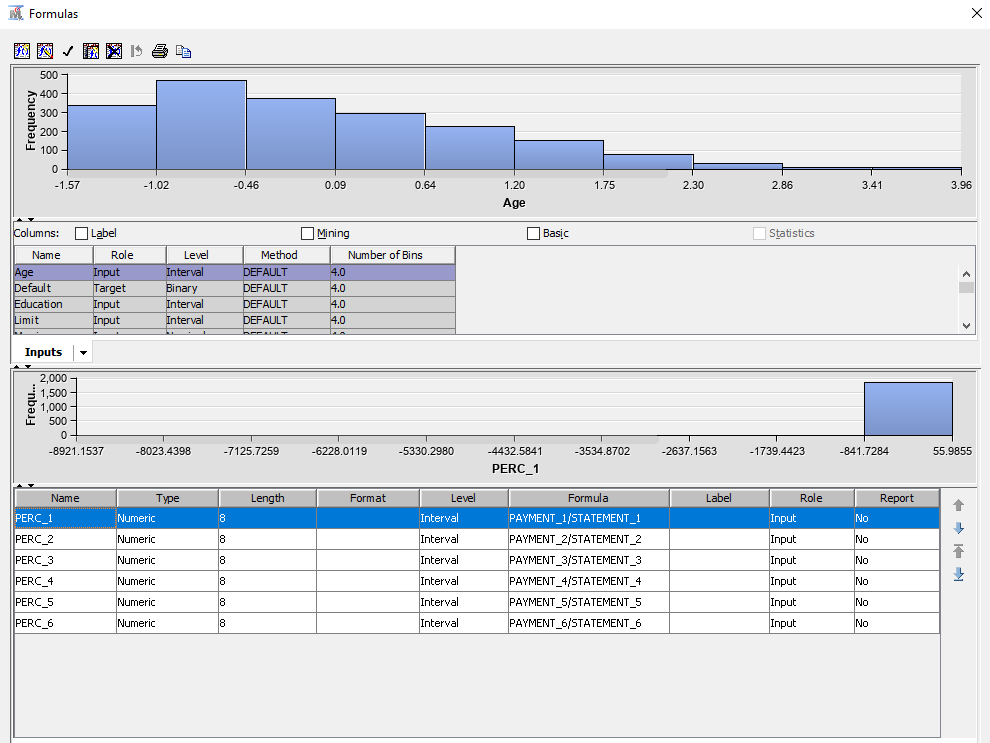
**Data Partition Node-**

****

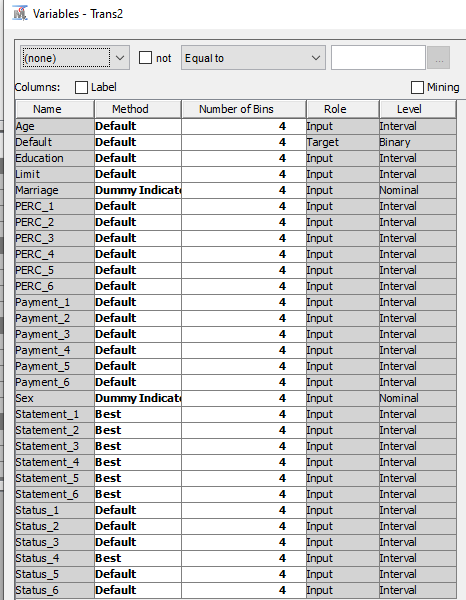
**Transformation node-**



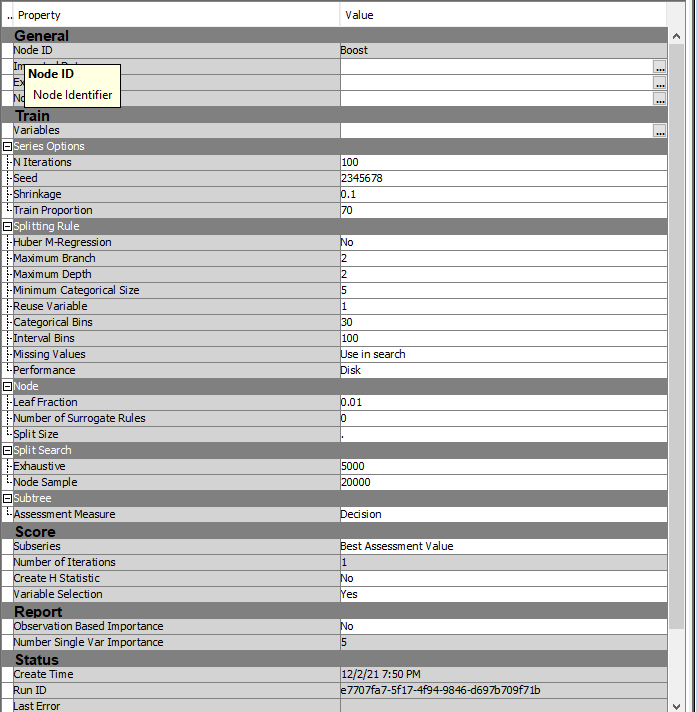
**Created new variables as shown below-**



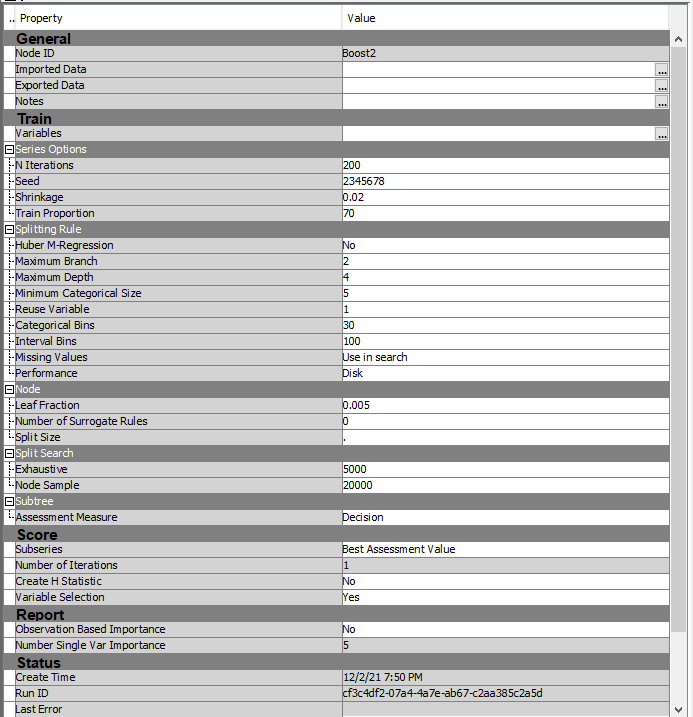
**Second Transformation Node -**



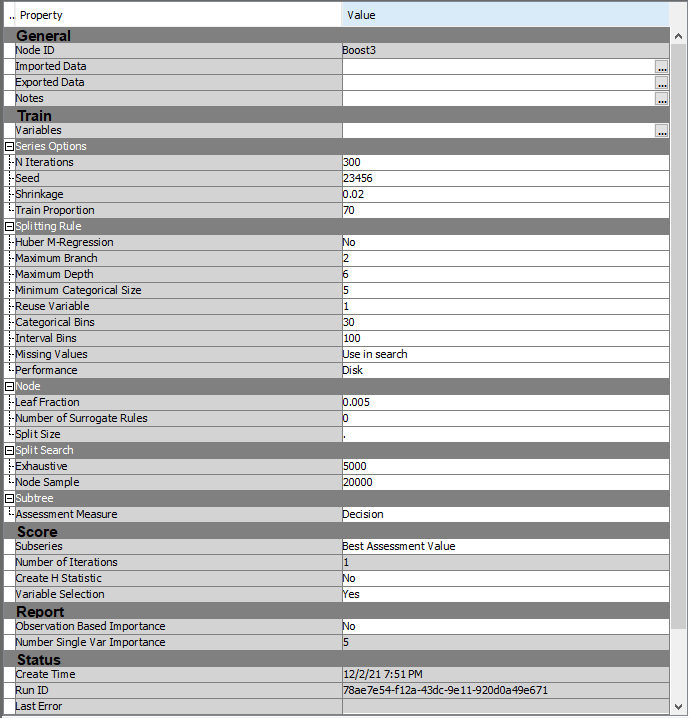
**Underfitting Gradient Boost Model**



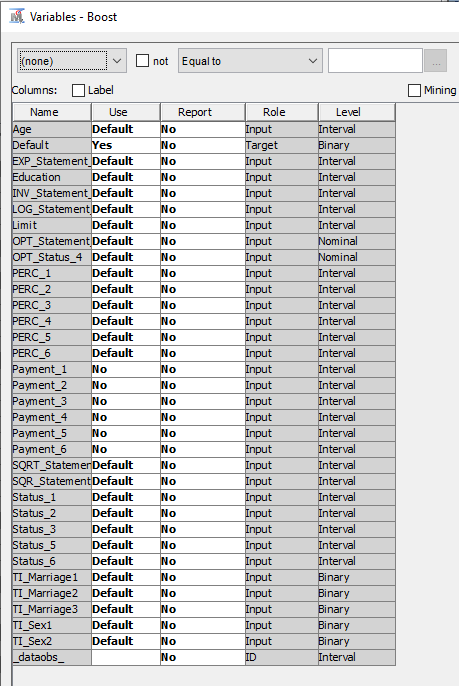
**Normal Fit Gradient Boosting Model (Best Model)-**



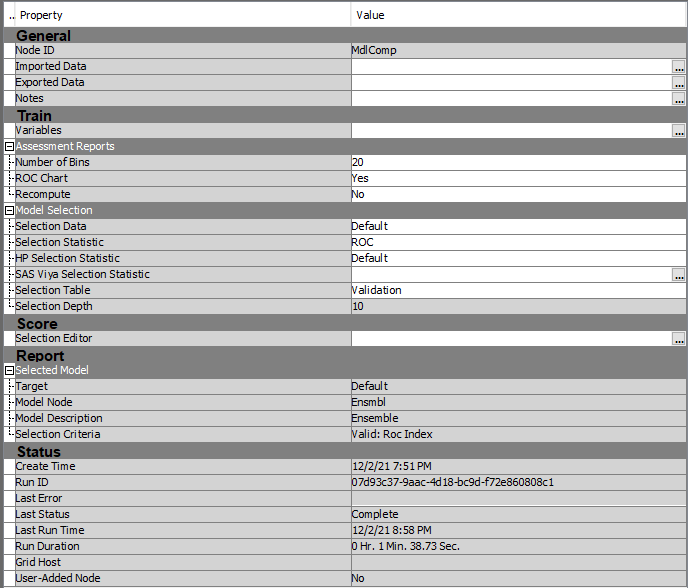
**Overfitting Gradient Boosting Model**



**Variable selection for the 3 Gradient Boosting models**



**Model Comparison**



Conclusion-

1. By using feature engineering, we increased our test ROC by ~0.3%.
2. We tried multiple model starting from Logistic regression to Random Forests and finally used GradientBoost.
3. Hyperparameter tuning allowed us train a robust model
4. By using a ensemble model we created the best model performance