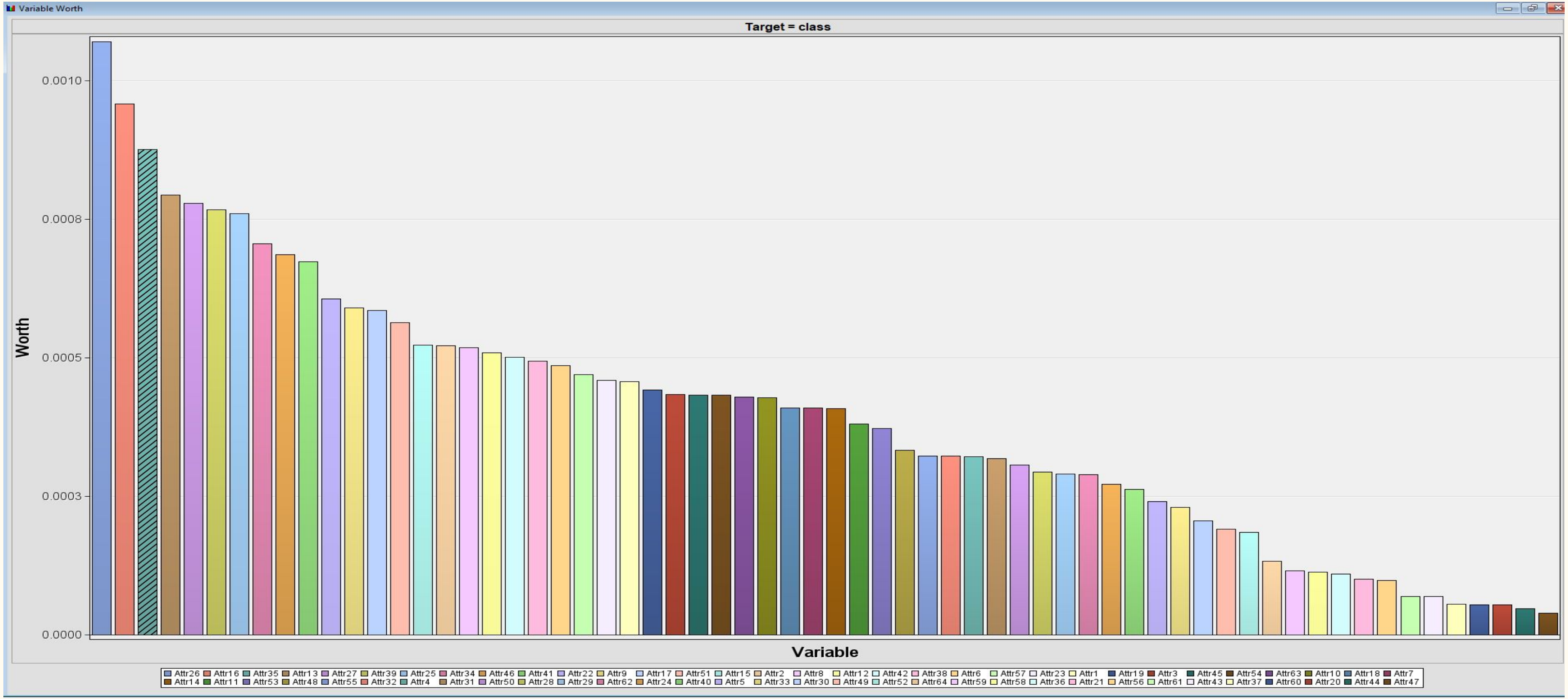
BANKRUPTCY CLASSIFICATION MODEL

Timothy S.

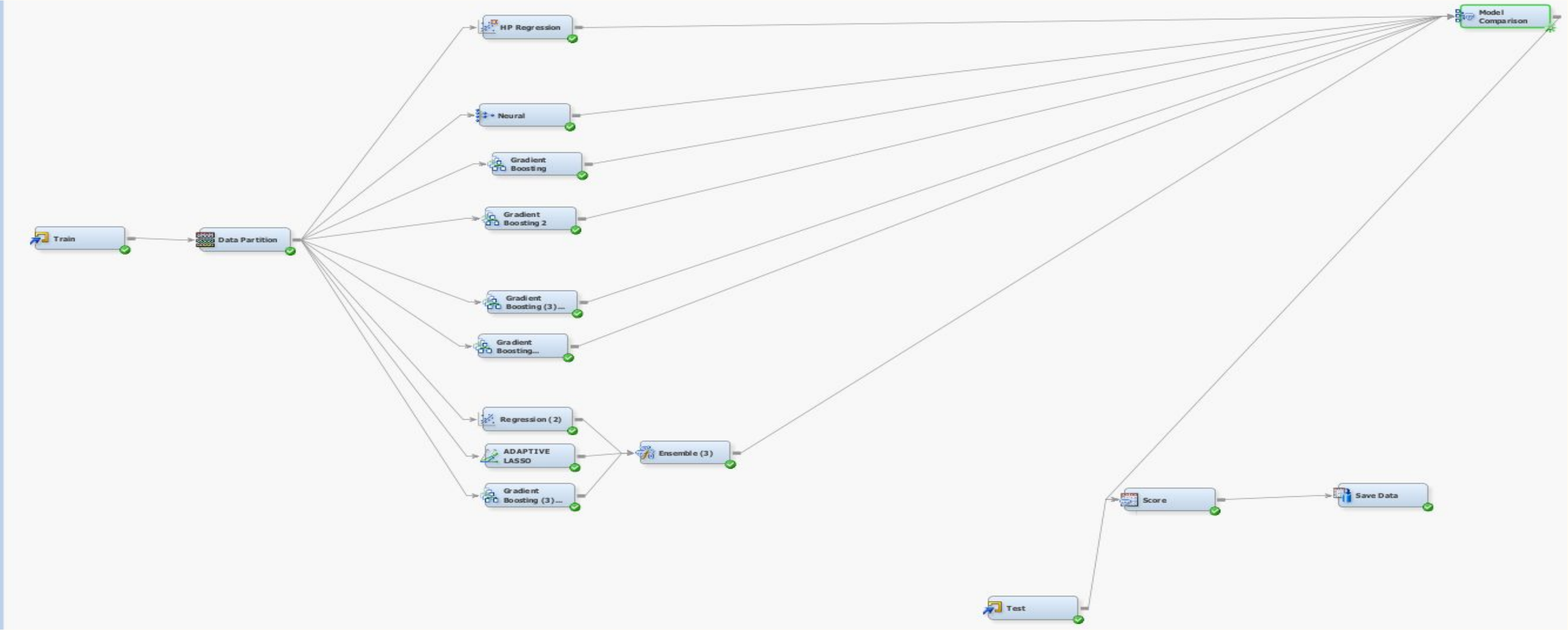
OBJECTIVES

- To use econometric data and financial information to predict the likelihood of bankruptcy for any given firm
- Run machine learning models in SAS EM to identify best possible model on training set.
- Maximize binary classification accuracy metric- ROC AUC score

EDA(Variable Worth)



MODEL DIAGRAM



ROC RESULTS

- Gradient boosting yielded best results
- Room for improvement of neural network by optimizing number of hidden units through trial & error
- Regression improvements were marginal

Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index
Boost3	Gradient Bo...	class		0.932
Boost2	Gradient Bo...	class		0.92
Boost	Gradient Bo...	class		0.918
Boost4	Gradient Bo...	class		0.915
Ensmbl3	Ensemble (...	class		0.903
Neural	Neural	class		0.855
HPReg	HP Regres...	class		0.853

Model Interpretation

The Gradient boosting method outlined the following features as most important in the model generation:

- Attr34-Operating Expenses
- Attr56-Profit Margin
- Attr44-Receivables*365/sales
- Attr58-Cost/sales
- Attr46-Current assets-inventory/short-term liabilities

MODEL REASONING

- LOGISTIC REGRESSION
  - Logit models are a go to for binary classification problems.
  - Good starting point for understanding the relationship between 64 features and target variable
  - High interpretability using coefficients to understand feature influence
  - **Adaptive Lasso** was also tested as it tends to reduce the impact of less important features
  - Stepwise selection was selected
- GRADIENT BOOSTING
  - Seemed to be the most accurate in classification tasks for complex data based on results and constant re-iteration
  - Uses weaknesses to strengthen prediction accuracy sequentially
  - 800 iterations(trees), 0.2 shrinkage(learning rate); 850 iterations, 0.06 shrinkage; 700 shrinkage, 0.02 shrinkage; **800 iterations & 0.05 shrinkage with log transformation** yielded my best results with **ROC of 0.938**

MODEL REASONING

- NEURAL NETWORK
  - Attempted this model due to complexity of the dataset
  - Multilayer Perceptron is advantageous as it factors non-linear relationships that may exist between the 64 financial metrics
  - Deep Learning advantages as well
- ENSEMBLE
  - Used this to compare between Gradient Boosting, Logistic Regression and Adaptive Lasso
  - Leveled the playing ground for model comparison by analyzing the three different algorithms using weight

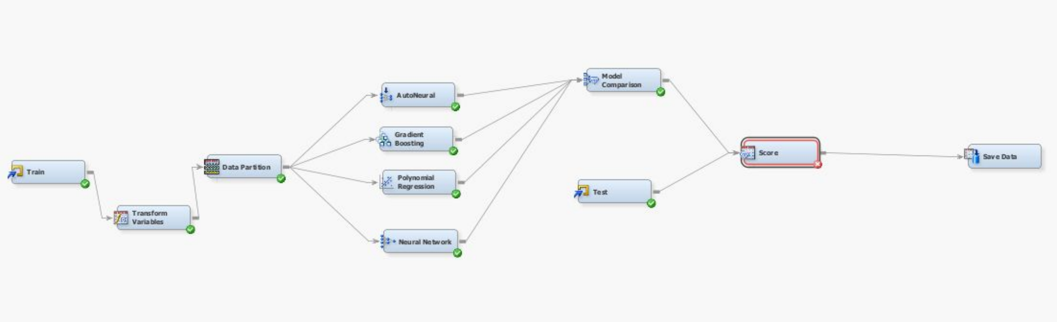
Overall, constant trial error is what helped to identify improvements in various models. I choose my model because of a relatively high ROC. x

ALTERNATIVE MODEL RESULTS

Overall, constant trial & error is what helped to identify improvements in various models.The chosen model got a relatively high ROC at time of submission.

The alternative model included polynomial regression to capture non-linear relationships. It yielded a promising ROC, however optimal number of terms needed to be achieved. I only used **2 polynomial terms**. Furthermore, Gradient boosting results seemed to benefit from transformation.

Was able to improve on the model using log transformation(0.938 ROC).



Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index	Train: Total Degrees of Freedom
Boost	Gradient Bo...	class		0.938	6999
Neural	Neural Net...	class		0.881	6999
Reg	Polynomial ...	class		0.869	6999
AutoNeural	AutoNeural	class		0.5	6999

LESSONS LEARNT

- Lots of trial & error is required
- Different perspectives will help discover useful models or approaches quicker
- Research on similar problems will reduce time to solution discovery
- Thoughtful pre-processing and standardization usually leads to better accuracy
- There are lots of different algorithms to be understood
- It is helpful to take notes of different parameters before re-iterating through models; saves time and helps to rule out non-useful approaches