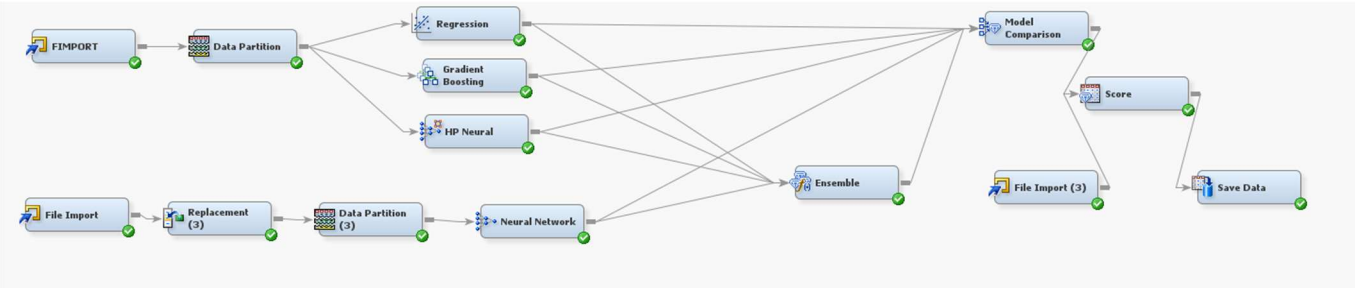
MGMT 571 Final Project- The Analytics Avengers

1.

The firm-level financial ratios, such as profitability, liabilities, and efficiency measures, are imported and prepared for further analysis. The preprocessing step ensures that attributes with high correlation (16 attributes) are removed upfront to reduce noise and redundancy in the dataset, improving model performance and interpretability.



Data Replacement (Replacement Node):

The Replacement node is included to handle missing or invalid data values, which are common in real-world financial datasets. In this scenario, incomplete or erroneous financial ratios (e.g., attributes related to net profit or liabilities) could introduce bias or reduce model accuracy. By replacing missing or outlier values with appropriate estimates, the node ensures the dataset is clean, reliable, and ready for training predictive models.

This was kept initially to train and validate the model, later disconnected.

- Why it's important here: Many financial ratios are interdependent, and missing values can disrupt the overall analysis. For example, missing data in net profit/total assets could skew the model's ability to assess profitability, a critical indicator of financial distress.

Data Partition:

The Data Partition node splits the dataset into training and validation. This is vital for this bankruptcy prediction problem because:

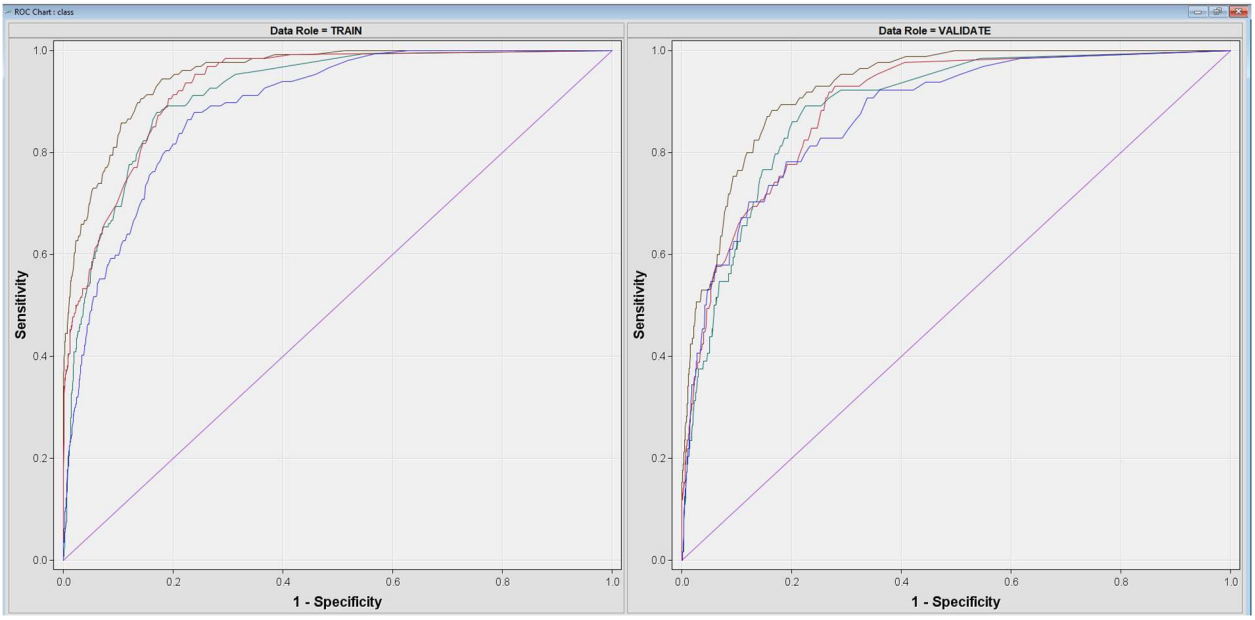
- Training data is used to fit the models and capture patterns in the financial attributes.
- Validation data is essential for fine-tuning model parameters and preventing overfitting, especially when using complex models like neural networks or gradient boosting.
- Test data allows for unbiased evaluation of the model's performance in predicting bankruptcy.
- Why it's important here: Predicting bankruptcy is a high-stakes problem where the model's generalizability is critical. By partitioning the data, we ensure that the model performs well not just on the training data but also on unseen firms, reflecting real-world performance.

Model Development:

Several predictive models are built to address the specific complexities of the bankruptcy prediction task:

- Regression: Logistic regression is included as a baseline model. It is well-suited for binary classification tasks like bankruptcy prediction, where the goal is to estimate the probability of a firm belonging to one of two classes (bankrupt or not). It provides interpretability, allowing stakeholders to understand which financial ratios significantly influence bankruptcy.
- Gradient Boosting: This model is chosen for its ability to handle non-linear relationships and interactions between financial attributes. Gradient boosting works well with tabular data and can capture complex patterns in financial distress signals that simpler models might miss.

- Neural Networks (HP Neural): High-performance neural networks are included because they can model intricate relationships between financial attributes. For example, they can capture subtle patterns in combinations of profitability and liquidity ratios that might be indicative of financial instability.
    - Why these models were used in this scenario:
        - Bankruptcy is often the result of complex interdependencies between financial metrics.
        - Logistic regression provides a clear baseline and interpretability.
        - Gradient boosting and neural networks provide the ability to capture non-linear patterns and interactions, which are critical for improving prediction accuracy.



Model Comparison:

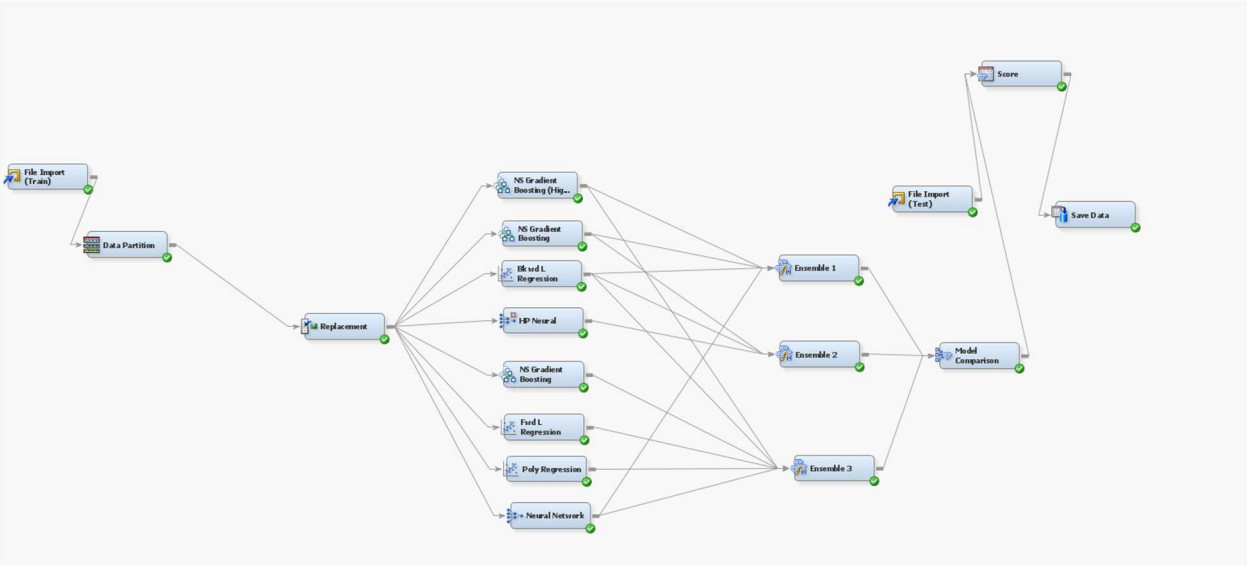
The Model Comparison node is used to evaluate all models based on the Area Under the Receiver Operator Curve (AUC), the competition's evaluation metric. AUC is critical for this problem because:

- It measures the model's ability to rank firms correctly based on their likelihood of bankruptcy.
- It is robust to class imbalance, which is common in bankruptcy datasets (typically, far fewer firms go bankrupt than stay solvent).
- Why it’s important here: For bankruptcy prediction, identifying firms at risk (true positives) is far more critical than identifying solvent firms. AUC helps ensure the model prioritizes this goal.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Roc Index
Y	Ensmbl	Ensmbl	Ensemble	0.927
	Boost2	Boost2	Gradient Boosting	0.914
	Neural3	Neural3	Neural Network	0.892
	HPNNA2	HPNNA2	HP Neural	0.889
	Reg	Reg	Regression	0.878

2.



Predictive Models

A variety of models are built to capture the different patterns and relationships in the data. Each model offers unique strengths that make it valuable for bankruptcy prediction:

1. Gradient Boosting (High and Default Settings)
    - Purpose: Gradient Boosting builds an ensemble of decision trees to capture non-linear relationships and interactions between financial attributes.
    - Why it's used here:
        - Firms' financial health is influenced by complex, non-linear interactions among variables.
        - Using different configurations of Gradient Boosting (high and default settings) allows exploration of trade-offs between complexity and generalization.
2. **Backward and Forward Stepwise Logistic Regression**
    - Purpose: These models iteratively include or exclude variables to find the best subset of predictors.
    - Why it's used here:
        - Provides a parsimonious model focusing only on the most significant financial ratios.
        - Helps identify the key predictors of bankruptcy, offering interpretability.
3. **Polynomial Regression**
    - Purpose: Incorporates polynomial terms to model non-linear relationships in the data.

- Why it's used here:
    - Captures more complex relationships between financial attributes and bankruptcy risk.

4. **High-Performance Neural Network (HP Neural)**
    - Purpose: Utilizes deep learning techniques to model intricate patterns in the data.
    - Why it's used here:
        - Captures non-linear dependencies and complex interactions among financial variables.
        - Useful for identifying subtle patterns that may indicate financial distress.
5. **Neural Network**
    - Purpose: A custom neural network configuration is included to test alternative architectures for bankruptcy prediction.
    - Why it's used here:
        - Provides additional diversity in neural network modeling, improving ensemble performance.

Ensembles (Ensemble 1, Ensemble 2, Ensemble 3)

- Purpose: Combines predictions from multiple models to create a more robust and accurate final model.
- Why it's used here:
    - Each individual model captures different aspects of the data. Combining them helps reduce variance and bias, leading to improved prediction accuracy.
    - Three different ensemble configurations are tested to identify the best combination of models for bankruptcy prediction.