Fall 2024 MGMT 571

Final Project - Bankruptcy Prediction

Predict bankruptcy using econometric measures via data mining algorithms.

Tools: SAS Enterprise Miner, Kaggle competition platform.

By FAST Analytics

Problem Statement

1 Goal

Develop models to identify firms at risk of bankruptcy.

2 Dataset

Brief description of features and size.

3 Evaluation

ROC-AUC on the private leaderboard.

Initial Finding(EDA) and Challenges

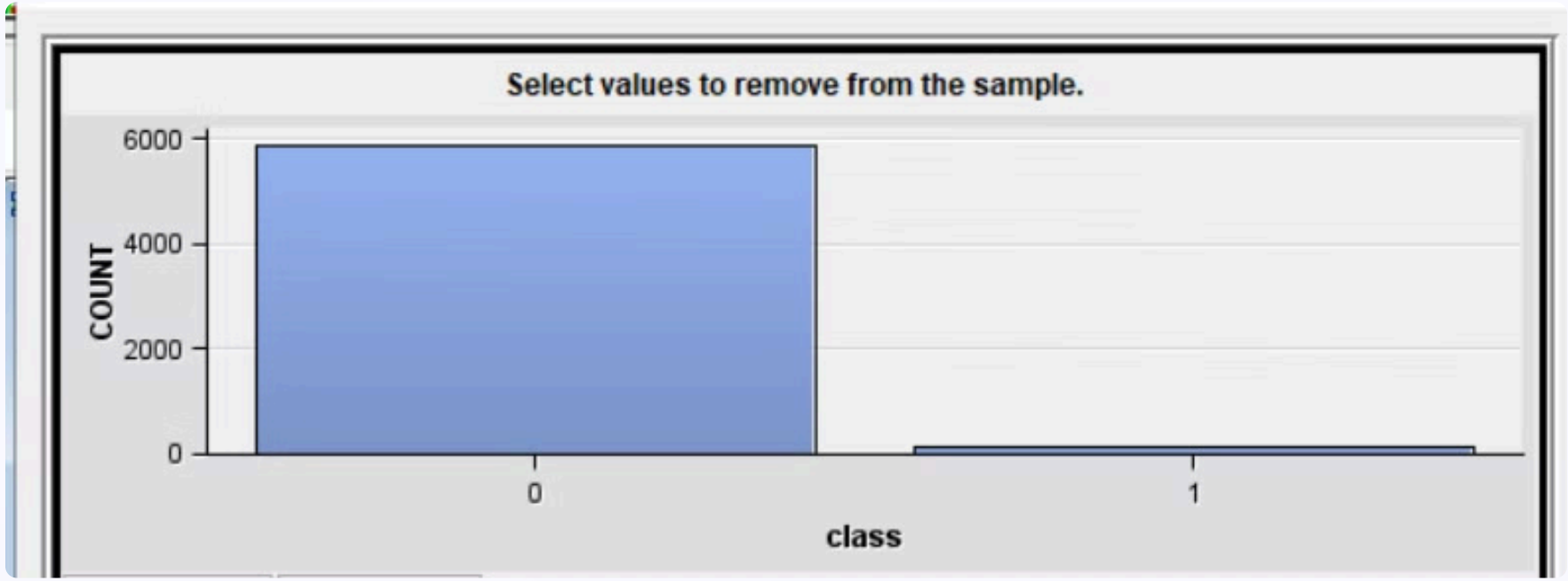
High Dimensionality: The dataset has numerous attributes (60+), which could lead to the curse of dimensionality. Many attributes have extreme values, suggesting that not all features may be equally important or informative, making dimensionality reduction essential.

Class Imbalance: A significant skew in the distribution of target classes

Multicollinearity: Attributes with similar means and ranges (e.g., Attr33, Attr63) seemed to be highly correlated, leading to redundancy. This could distort model interpretation and requires a correlation matrix analysis to confirm.

Extreme Skewness and Kurtosis: Attributes like Attr1 (Skewness: -10.26, Kurtosis: 341.54) and Attr13 (Skewness: 77.25, Kurtosis: 5977.77) had significant skewness and kurtosis. These distributions suggest potential outliers and non-normality, which could impact the performance of machine learning models.

Class Imbalance



DATA PROCESSING

We attempted to address data imbalances and skewness through techniques such as sampling (Sample Node), outlier handling (Replacement/Filter Node), and standardization (Transform Node). However, these preprocessing steps did not improve the model's performance. Consequently, we opted to rely on models like Gradient Boosting, which are inherently robust to such data challenges, and skipped explicit data preprocessing

DATA PARTITION (60%, 40%)

Models Explored

Model 1

Ensemble of various algorithms

Combination of:

- Logistic Regression (with regularization and polynomial features)
- HP GLM
- HP SVM (polynomial kernel)
- Gradient Boosting (100 iterations, depth 2)
- Neural Network (3 hidden layers)
- LARS (adaptive lasso)

Performance: ROC-AUC = 0.94216

Model 2

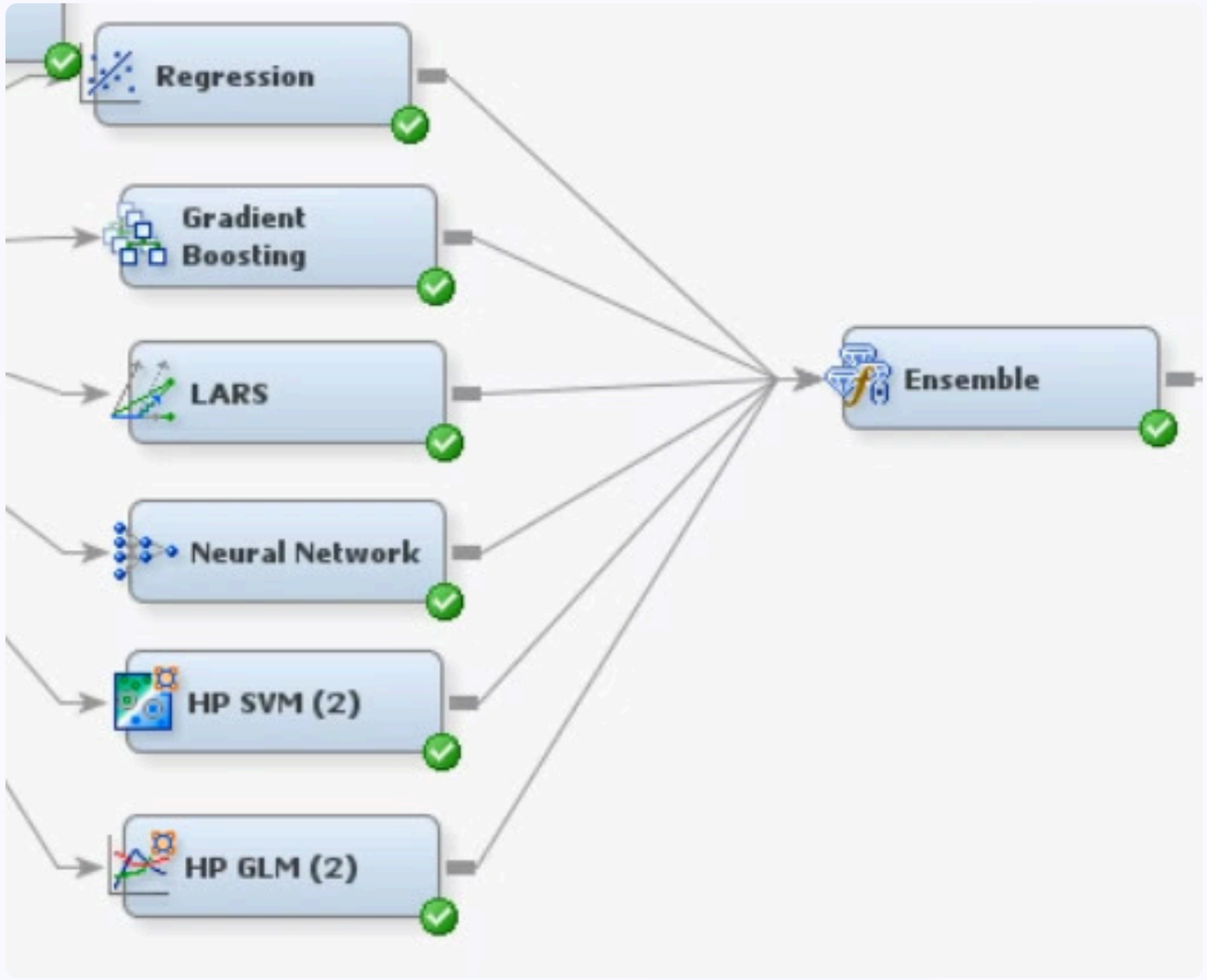
Similar ensemble without HP GLM.

Performance: ROC-AUC = 0.94150

MODELS EXPLORED

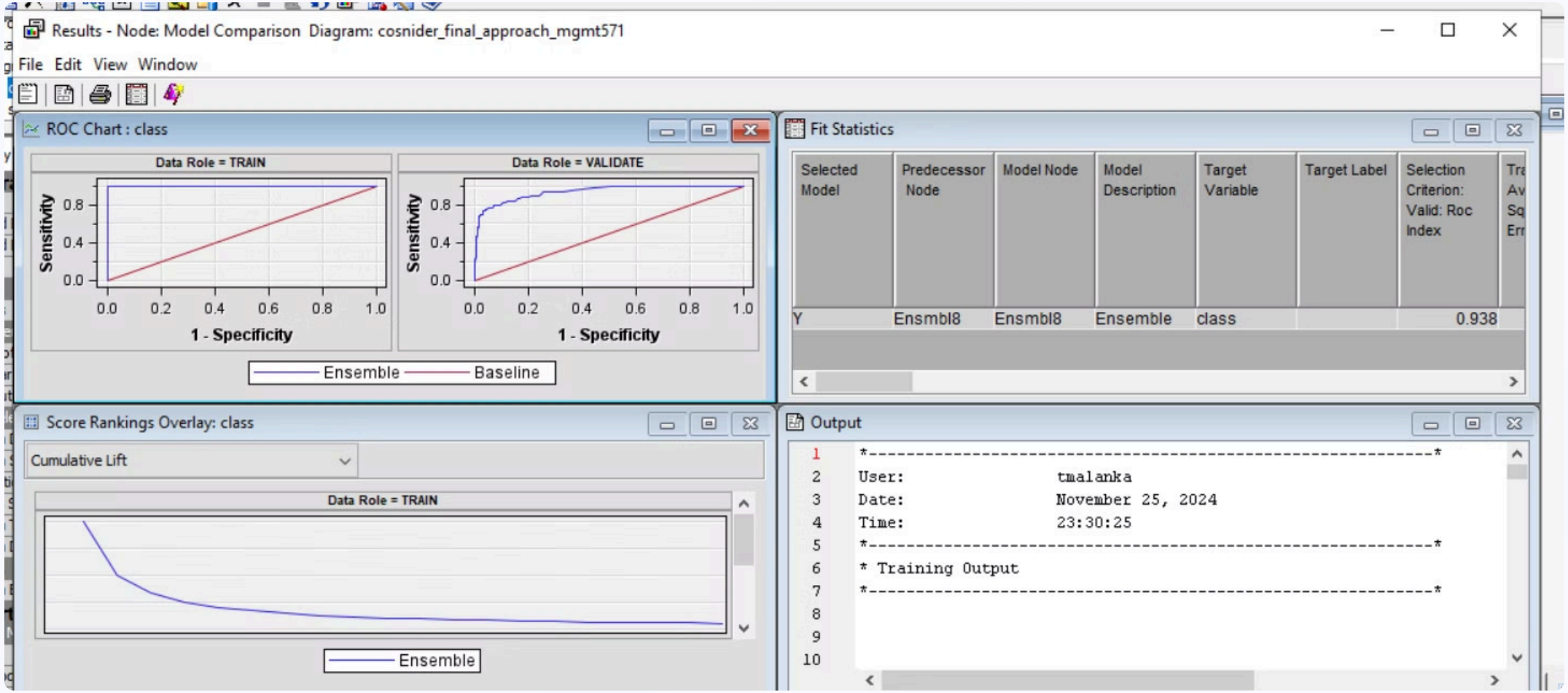
To address overfitting observed in standalone Gradient Boosting, we transitioned to ensemble techniques to leverage model diversity and reduce variance for improved generalization and stability.

FINAL MODEL:



ROC TEST: 0.94466

MODEL RESULTS



Rationale

Each model contributes unique strengths to the ensemble:

- Gradient Boosting and Neural Networks are excellent for capturing complex, non-linear patterns.
- Regression and GLM provide interpretability and robustness for simpler relationships.
- SVM excels in high-dimensional spaces with good margin separation.
- Adaptive Lasso efficiently handles high-dimensional datasets, focusing on relevant features.

Key Decision where we increased our scored

1. Transitioned to ensemble models for improved accuracy and robustness.
2. Adjusted GB parameters from 50 iterations with max depth 3 to 100 iterations with max depth 2, enhancing generalization.
3. Increased neural network hidden units from 5 (default) to 10 for better learning capacity.
4. Switched to a polynomial kernel from a linear kernel in SVM to capture non-linear patterns.
5. Incorporated TRUREG and CONGRA optimization in techniques for more efficient convergence.

OTHER APPROACHES

Decision Tree in Ensemble: Limited model diversity may lead to overfitting or insufficient performance improvement in the ensemble.

Gradient Boosting with Identical Node Weighting: Over-emphasis on certain nodes may reduce model flexibility, limiting overall improvement.

LARS for Feature Selection: Sparse feature selection may exclude relevant features, reducing model performance for complex relationships.