Predicitve Modelling For Bankruptcy Predicting Using SAS EM

Alekya Rama
Namitha Ramakrishnan
Uttish Narayan

Project Overview and Dataset Characteristics

1 Objective

Predict bankruptcy probabilities for firms using machine learning models and maximize performance on the leaderboard.

2 Dataset

64 predictors derived from financial and operational metrics, with an imbalanced distribution (only ~2% bankrupt firms).

3 Preprocessing

Replacement of missing values and variable transformation for uniform scaling.

4 Evaluation

Validation ROC (Receiver Operating Characteristic) score as the primary metric.

Journey to the Best Model: Exploratory Phase

1

Polynomial Regression

Leveraged non-linear interactions between variables. Provided stable performance but had limitations in handling highly complex feature interdependencies.

2

Gradient Boosted Trees (GBT)

Offered high interpretability and feature importance rankings, but showed signs of overfitting. Fine-tuning partially mitigated this issue.

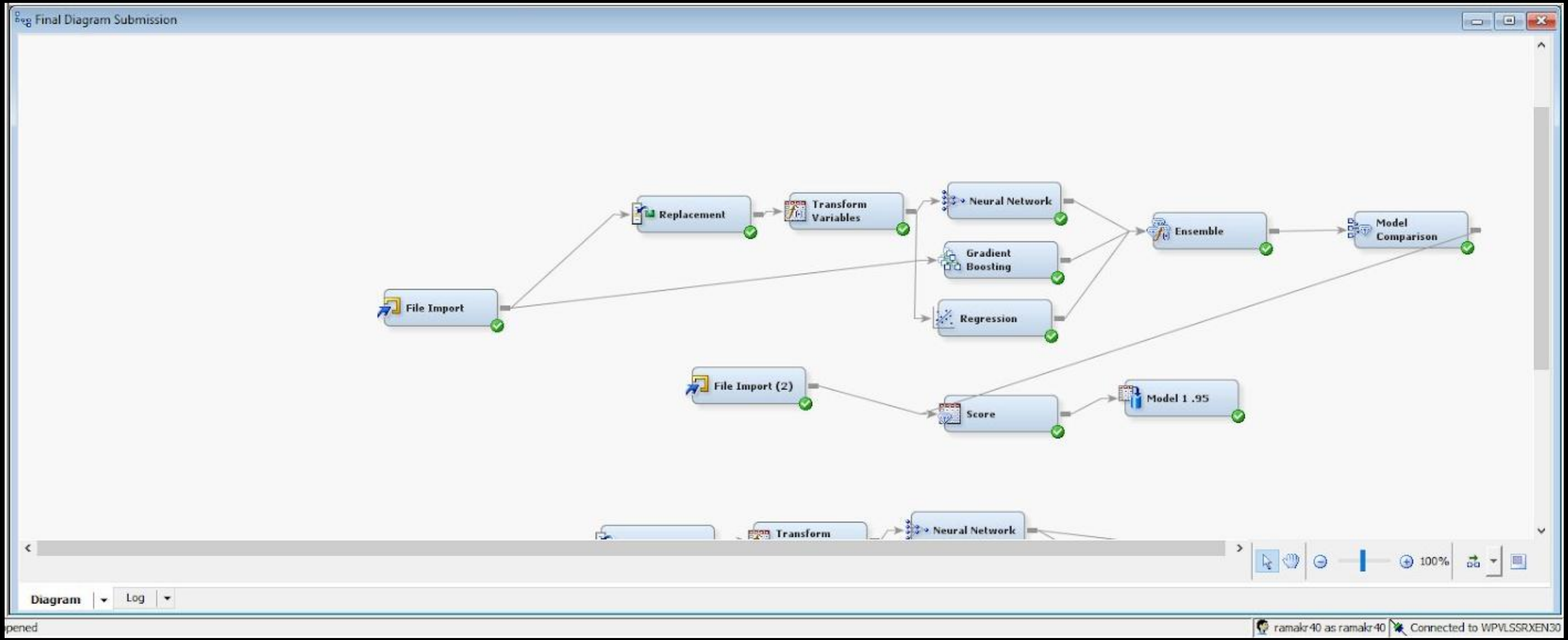
3

Neural Networks (NN)

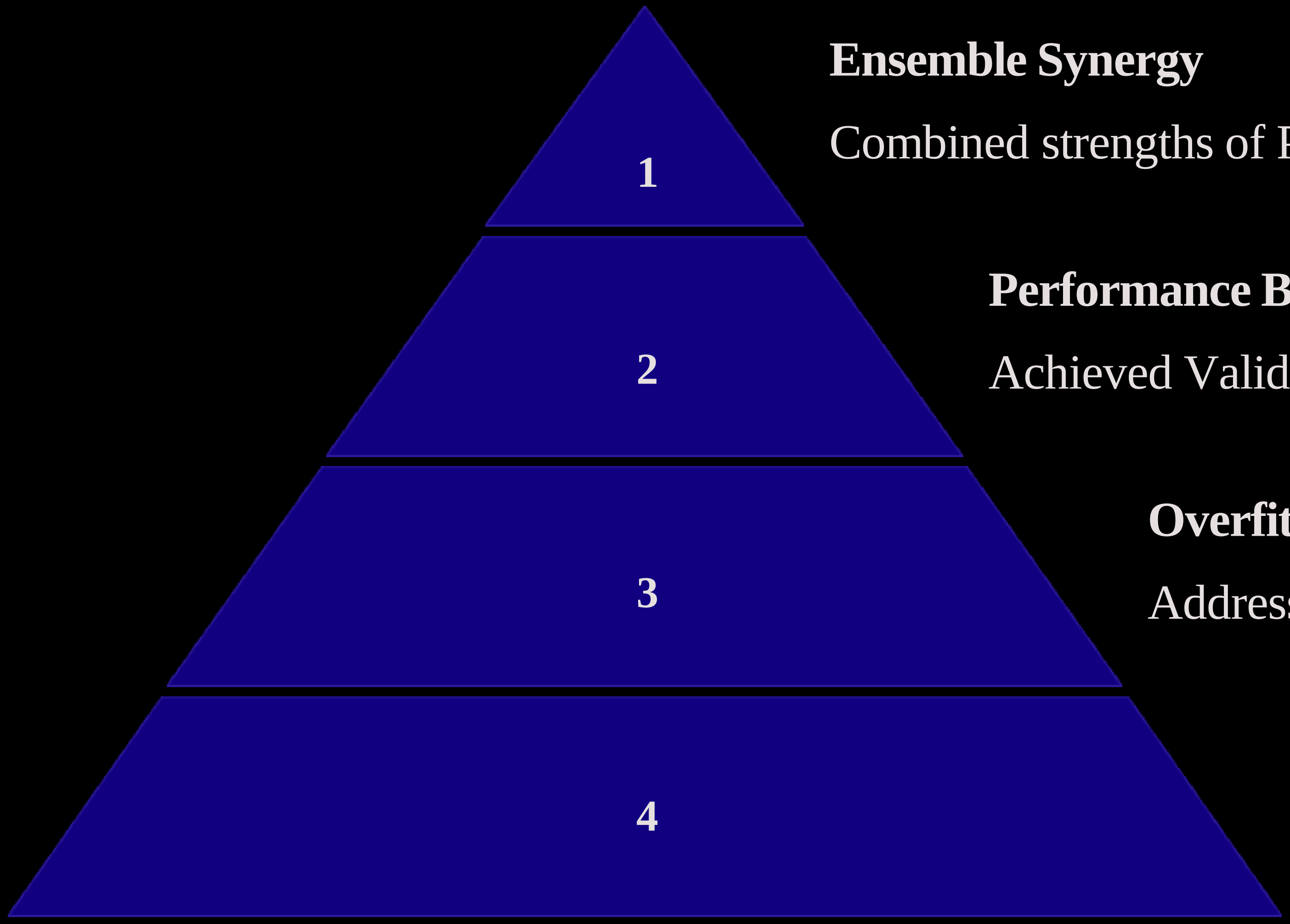
Capable of learning complex, non-linear relationships. Faced challenges in balancing training and validation performance.

OUR MODEL

Our ensemble model strategically integrates Gradient Boosting, Neural Network, and Polynomial Regression, combining their individual strengths to create a robust and reliable predictive framework.



Ensemble Model: The Final Selection



Ensemble Synergy

Combined strengths of Polynomial Regression, Gradient Boosted Trees, and Neural Networks

Performance Boost

Achieved Validation ROC ~0.94+, minimizing train-validation gap

Overfitting Mitigation

Addressed generalization challenges of individual models

Robust Framework

Balanced approach for complex data patterns

Model Performance Statistics

0.936

Ensemble ROC

Outperformed individual models by harmonizing strengths and balancing bias-variance tradeoff.

0.911

GBT ROC

Strong generalization with balanced performance across metrics.

0.897

NN ROC

Excellent pattern recognition but slightly prone to overfitting.

0.859

Poly Reg ROC

Effectively captured non-linear trends, providing valuable support in the ensemble.

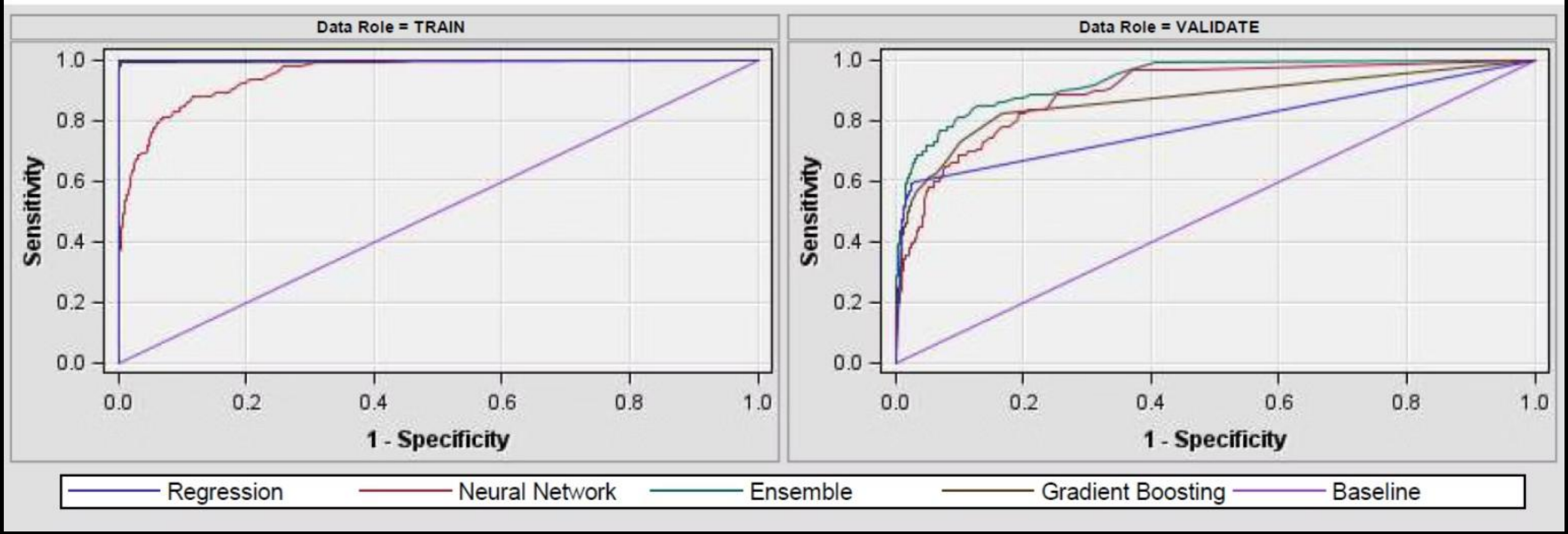
Performance Insights: ROC Curve Analysis

Training vs. Validation Performance

The ROC curve demonstrates the ensemble's exceptional balance between learning and generalization. With near-perfect training accuracy (~0.99) and strong validation performance (ROC: 0.931), our model showcases its ability to learn complex patterns without overfitting.

Model Comparison

The ensemble consistently outperforms individual models across various threshold levels. This superiority highlights the robustness of our approach, leveraging the strengths of each component model to achieve optimal predictive power in bankruptcy prediction.



Conclusion: The Power of Ensemble Modeling

Balanced Strengths

Our ensemble leverages the unique strengths of each model: Polynomial Regression for simpler non-linear trends, Gradient Boosted Trees for robust feature interactions, and Neural Networks for deep non-linear pattern extraction.

Mitigated Weaknesses

The ensemble approach effectively addresses the limitations of individual models, such as overfitting in GBT and NN, and the inability of Polynomial Regression to capture complex relationships alone.

Superior Performance

By achieving the highest validation ROC of 0.936, our ensemble model exemplifies the power of collaboration among diverse algorithms, solidifying its position as the leaderboard champion in bankruptcy prediction.

In conclusion, our model ranked 5th on the public leaderboard with 50% of the test data and maintained its position on the private leaderboard with 50% additional unseen data. This consistent performance highlights the robustness and reliability of our predictive framework. This project underscores the importance of collaborative modeling techniques in addressing complex financial forecasting challenges.