Bankruptcy Prediction with Machine Learning Models

SZA

Seonkyu Kim
Ziyue Zhang
Anto Fredric Henry Mohan Dass

Contents

Final Model 1

[Neural Network]
Multilayer Perception
# of Hidden Units=6

[Logistic Regression]
Selection Model=[None, Forward]

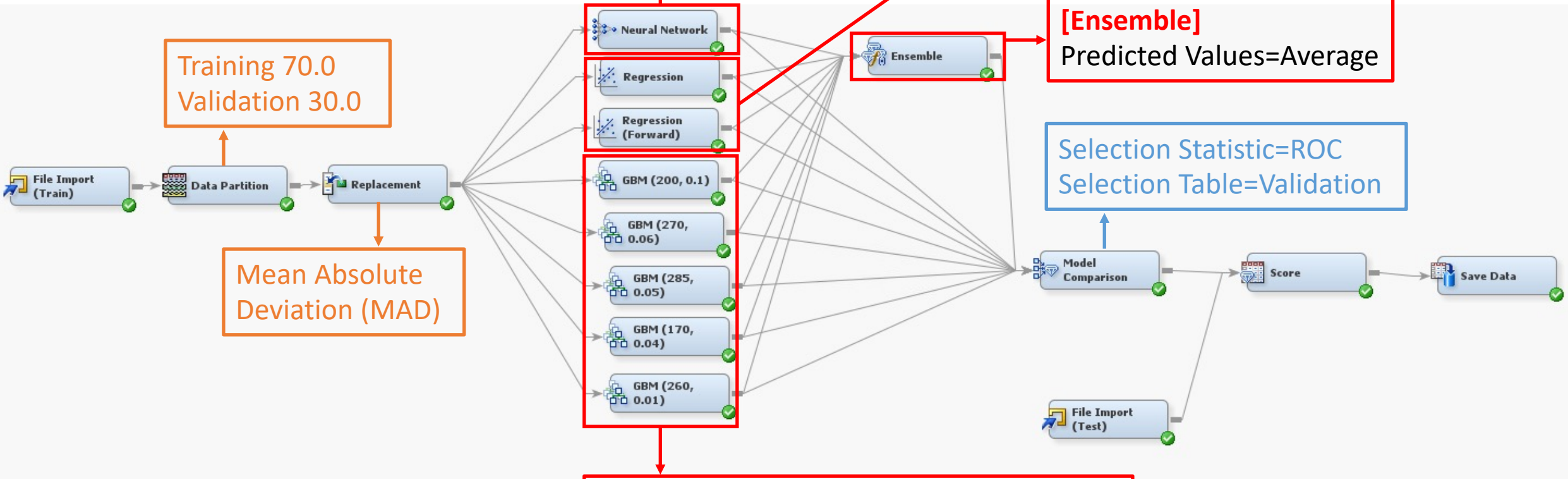
[Ensemble]
Predicted Values=Average

Training 70.0
Validation 30.0

Selection Statistic=ROC
Selection Table=Validation

Mean Absolute Deviation (MAD)

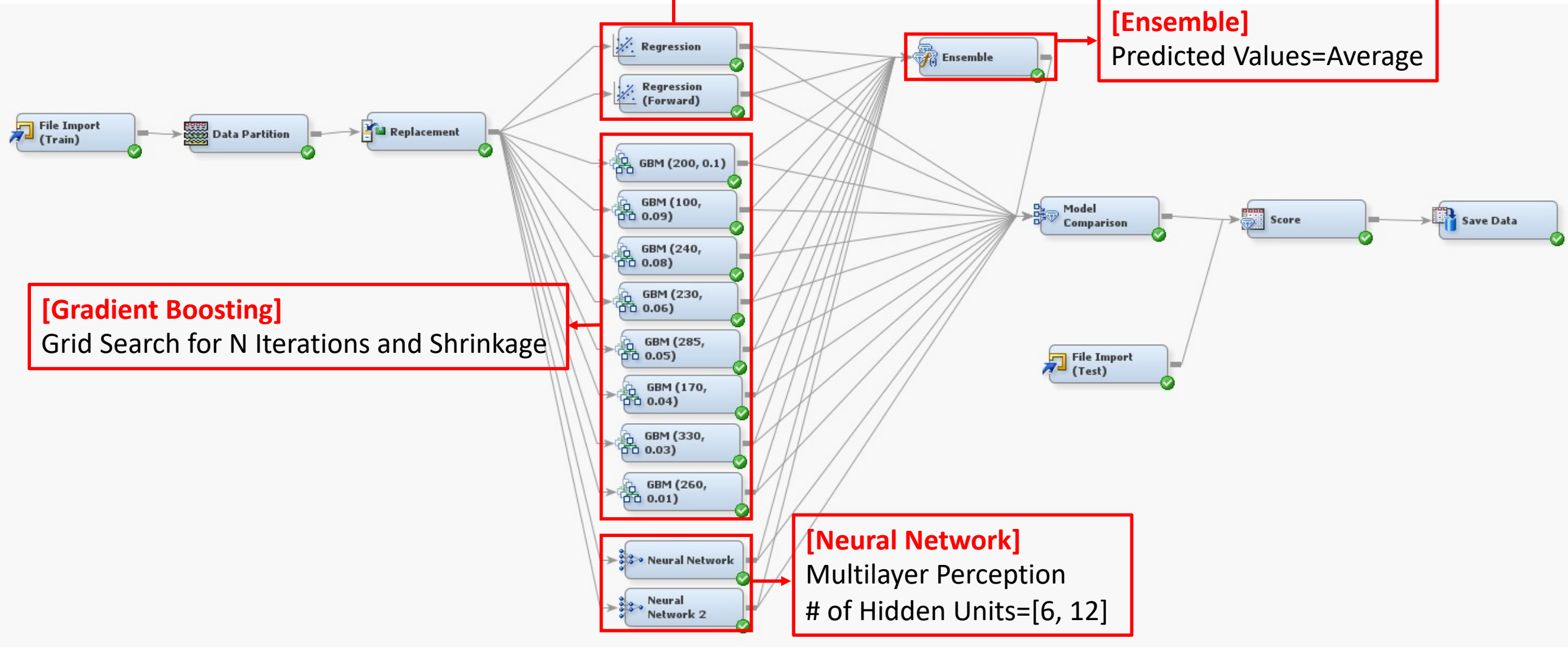
[Gradient Boosting]
Grid Search for N Iterations and Shrinkage



Final Model 2

[Logistic Regression]
Selection Model=[None, Forward]

[Ensemble]
Predicted Values=Average



[Gradient Boosting]
Grid Search for N Iterations and Shrinkage

[Neural Network]
Multilayer Perception
# of Hidden Units=[6, 12]

Model Selection Criteria

1. Highest Public Score (Final Model 1)

2. Highest ROC in the validation set (Final Model 2)

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index
Y	Ensmbl	Ensmbl	Ensemble	class		0.937
	Neural	Neural	Neural Net...	class		0.933
	Neural3	Neural3	Neural Net...	class		0.929
	Boost3	Boost3	GBM (285, ...	class		0.913
	Boost7	Boost7	GBM (230, ...	class		0.909
	Boost13	Boost13	GBM (330, ...	class		0.909
	Boost5	Boost5	GBM (240, ...	class		0.908
	Boost9	Boost9	GBM (100, ...	class		0.904
	Boost	Boost	GBM (170, ...	class		0.904
	Boost4	Boost4	GBM (200, ...	class		0.903
	Reg	Reg	Regression	class		0.895
	Boost6	Boost6	GBM (260, ...	class		0.887
	Reg3	Reg3	Regression...	class		0.882

Result

	Public Score	Private Score
Model 1	0.94311	0.91819
Model 2	0.94277	0.91792

The model was overfitted to the Public set.

Part III. Reason and Model Improvement

Common Reasons Lead to Significant Performance Difference between Public and Private Leaderboard Scores:

- Data Leakage
- Overfitting
- Model Instability
- Difference in Evaluation Metric

The Most Possible Reason for Our Project:

- Overfitting >> Regularization

Part III. Reason and Model Improvement

- Future Improvement
- Prevent focusing too much on Public Leaderboard Score
- Regularization
- Cross-validation

Thanks for Listening!