Credit Default Classification

SAS EM & Kaggle Competition

Team Machine: Li-Ci Chuang, Rachel Fagan, and Yi-Hsuan Hsu

~450	Attempts ran in SAS
48	Algorithms tested on kaggle
17	Model types tried
0.7401	Private leaderboard score
0.75251	Personal highest private score
3	Tears shed (in our hearts)

Data Pre-Processing Approaches

Filtering: (1) outside of 3 standard deviations from the mean, (2) mean absolute deviation

Replacement: (1) replace outliers with 3 standard deviations from the mean

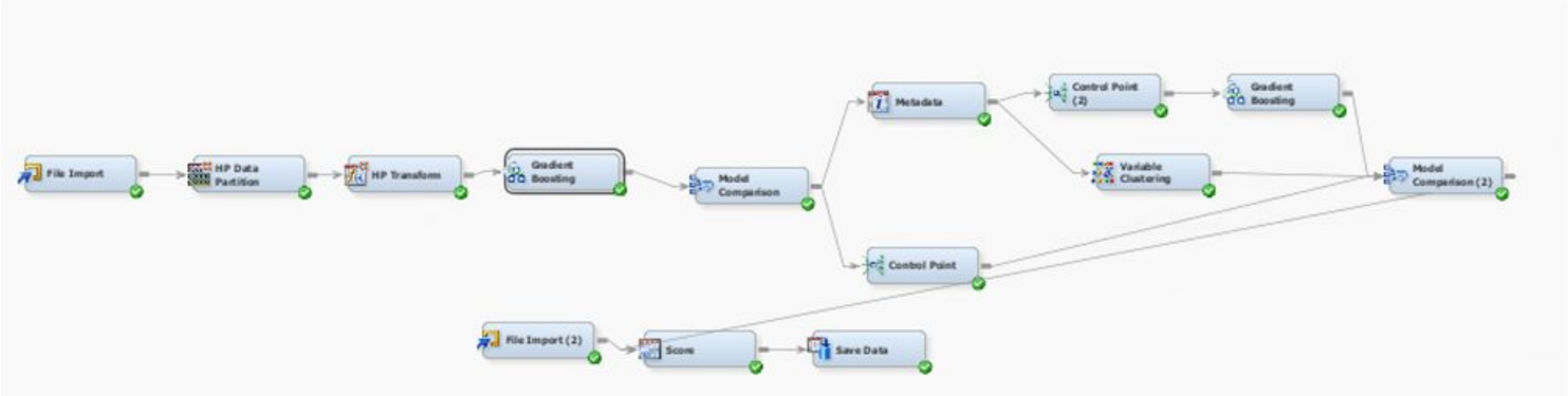
Transform: (1) exponential, (2) log 10, (3) log, (4) inverse, (5) optimal binning, (6) best

Variable Selection: (1) chi-square, (2) R-squared

Modeling Techniques Tried

- Regression with Variable Selection
- Decision Trees
- Bagging
- Random Forests
- **Gradient Boosting**
- Neural Network
- **Ensemble**
- HP BN Classifier
- HP Forest
- HP GLM
- HP SVM
- HP Neural
- HP Regression
- Lasso & Adaptive Lasso

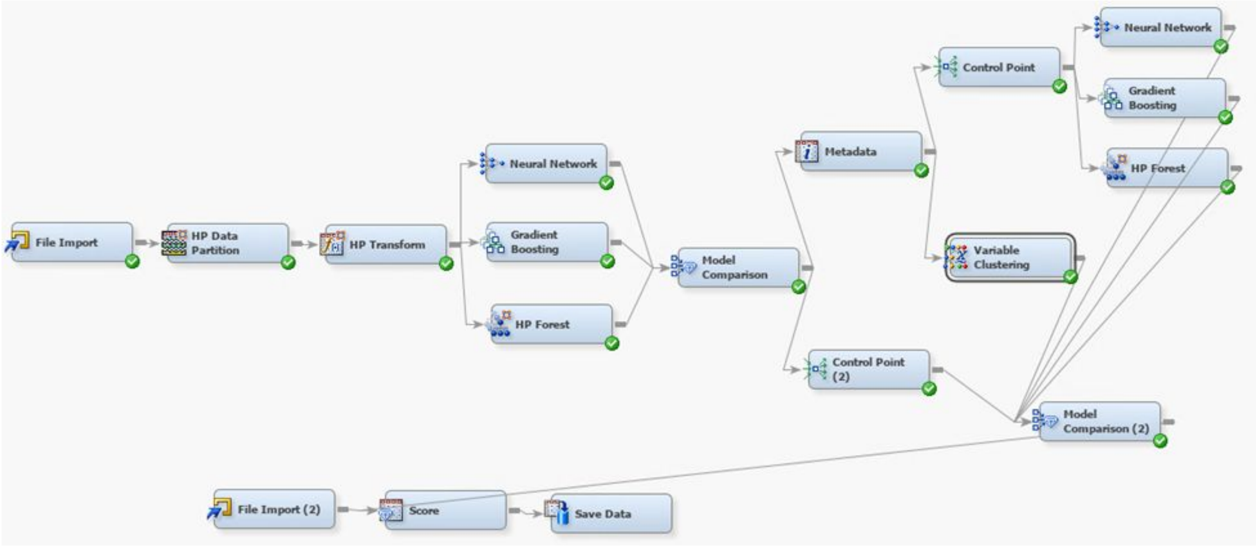
Final Algorithm 1



Why did we choose this model?

- It was our highest public score on the leaderboard
- Our Training ROC = Our Validation ROC
- Boosting keeps learning and avoids overfitting
- 50% of past winners used boosting

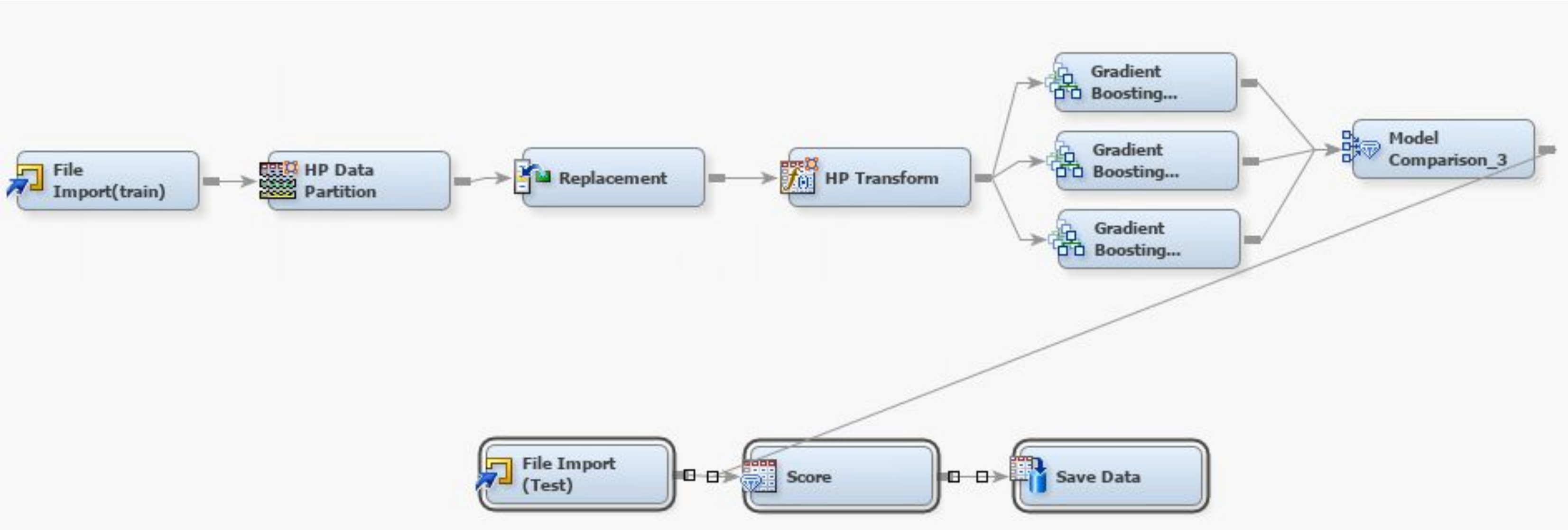
Final Algorithm 2



Why did we choose this model?

- It was our 4th highest score on the public leaderboard
- It was different from our Final Algorithm 1
- Applied different methods (HP Forest & Neural Network)

One of Our Better Models (0.75)



- We did not use this model because it scored 0.73842 on the public leaderboard.
- The private leaderboard ROC was 0.75189.
- It uses Gradient Boosting.

Learning & Takeaways

1. While our Final Algorithm 1 was not ranked high on the leaderboard (57th), it stayed at an ROC of 0.74.

2. Focusing on the robustness of the model instead of the slight differences between the training dataset ROCs is important.

3. Choosing an entirely different combination of models was more effective than tweaking the parameters of models.

Thank You