**Homework 8. Decision Tree, Bagging, and Boosting using Lending Dataset Modeling**

Author: Rohit V. Akole

UConn ID: 3133252

University of Connecticut

OPIM5604 Predictive Modeling

Professor: Pankaj Prakash

Abstract

This homework is about the Lending Dataset based on year 2014 data. I used the previously cleaned dataset to perform Decision Tree, Bagging (Bootstrap Forest), and Boosting on 1st iteration of the logistic regression model and compared these models with the Homework 5 final model.

## Decision Tree

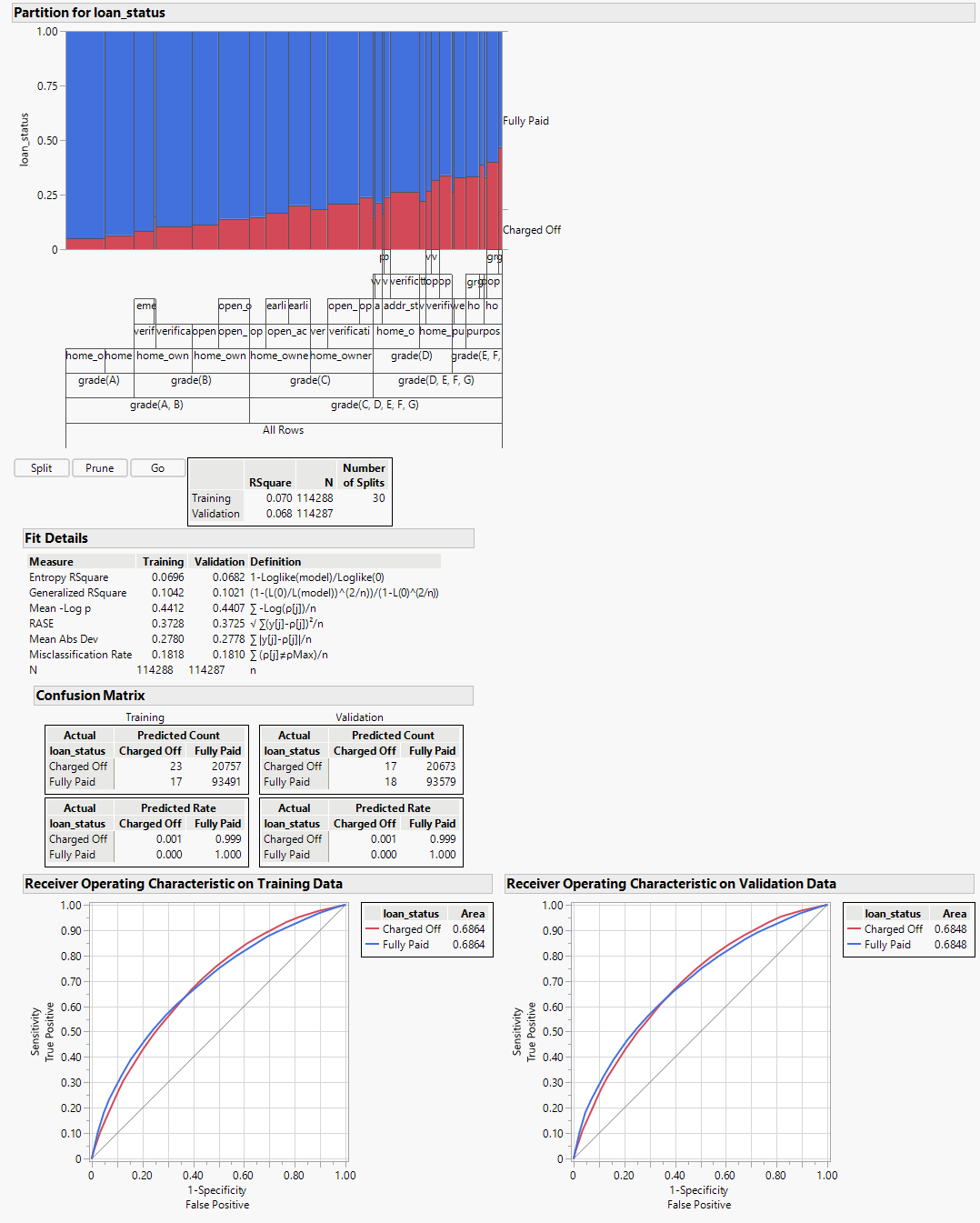


Figure 1. Decision Tree Model

Using the decision tree model, I get the best AUC for Validation data in the decision tree at 30 splits. After 30 splits, the AUC starts decreasing and the same happened when I started pruning the decision tree. I ended up with the conclusion that the 30-split model gave the best output for my decision tree model using the 1st iteration variables of the Logistic Regression model from homework 5.

## Bagging (Bootstrap Forest) Model:

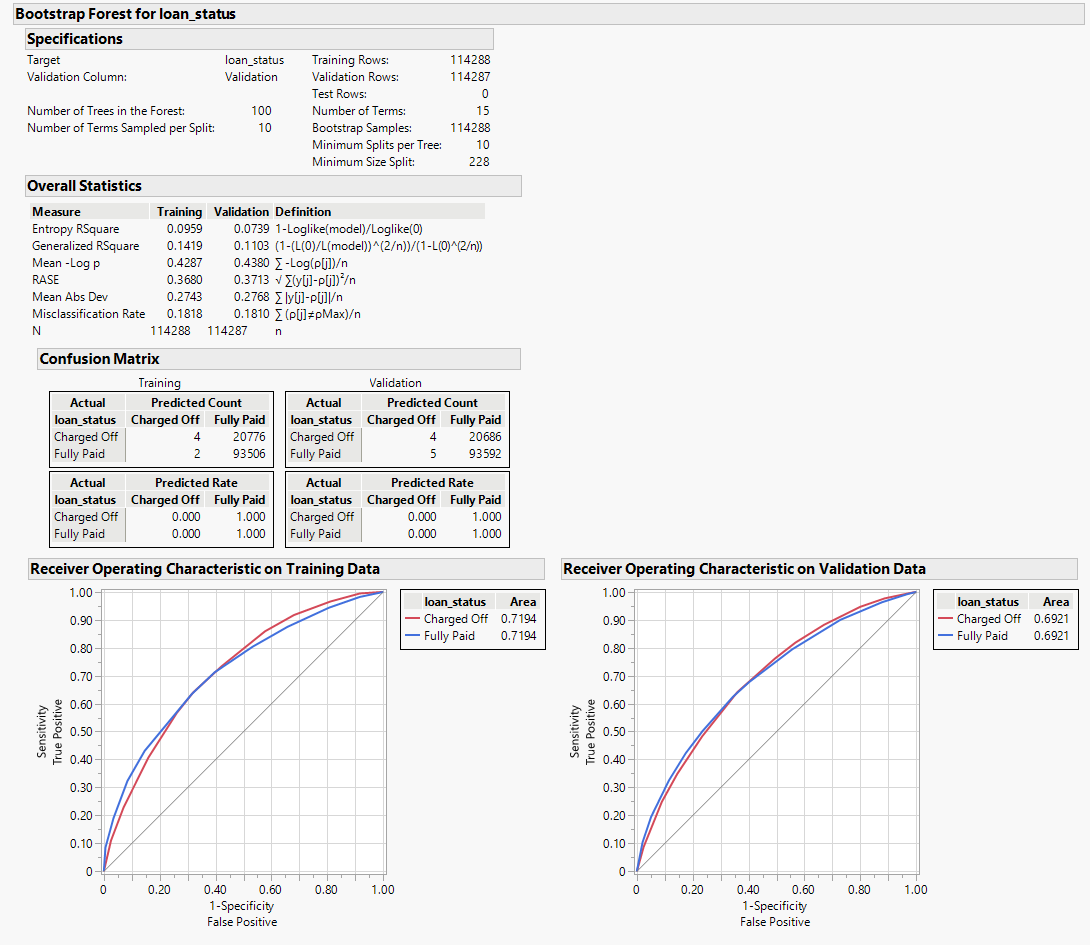


Figure 2. Bootstrap Forest (Bagging) Model

Using the default values of bootstrap forest, I got the AUC for the validation dataset at 69.21%. The difference between the validation and training dataset AUC is not much, so it is safe to say that the model is not overfitting. With the default values, the model ran for 100 trees in the forest with 10 number of terms sampled per split.

## Boosting (Boosted Trees) Model:

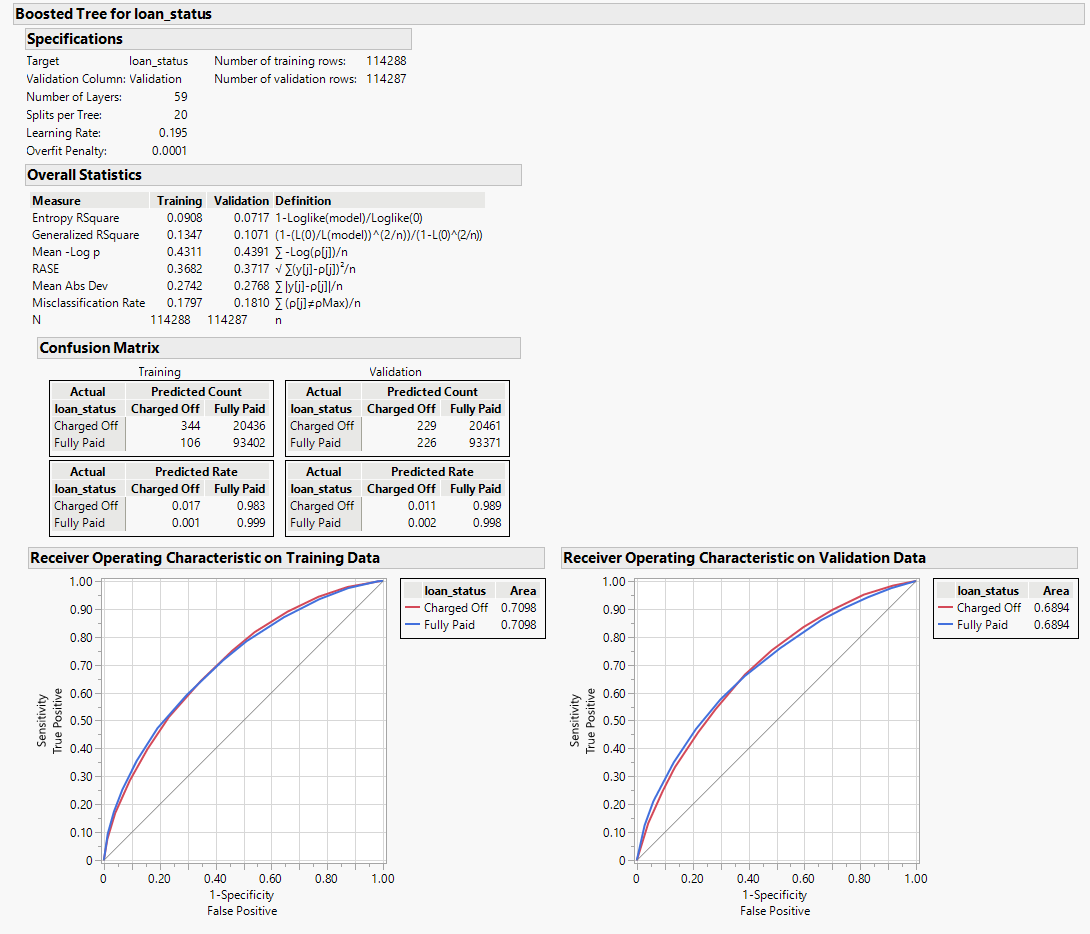


Figure 3. Boosted Tree (Boosting) Model

Using the default boosting model option, I got an AUC of 68.94% for validation data. As per the model, the difference between training and validation AUC is not much. There are no signs of overfitting in this model.

## Model Comparison

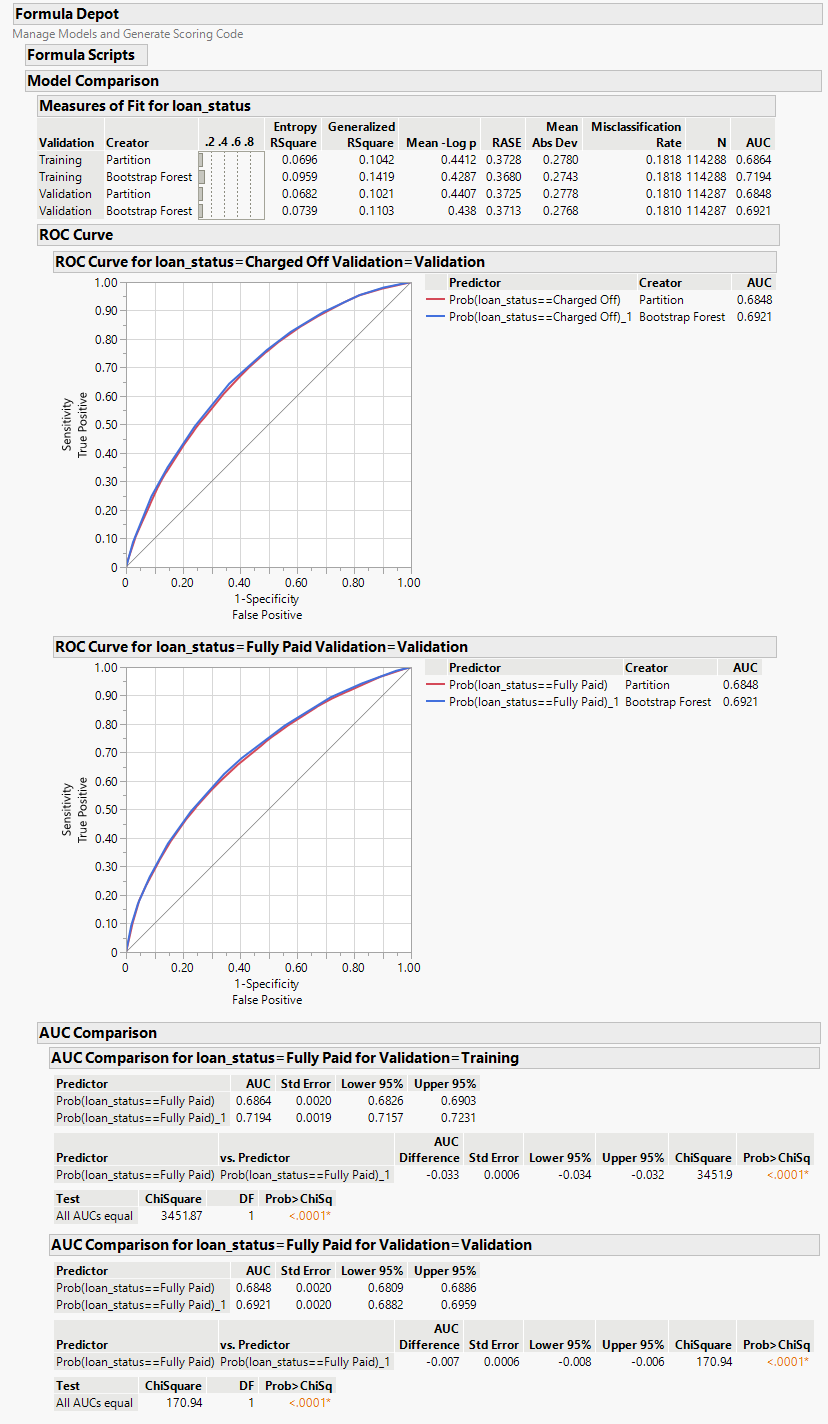


Figure 4. Model Comparison Decision Tree vs. Bagging (Bootstrap Forest)

Above is the model comparison between Decision Tree and Bagging Model. As discussed, the Boosted tree was giving error in formula, so I compared the above 2 with the boosted tree model itself. The table with the comparison is given below.

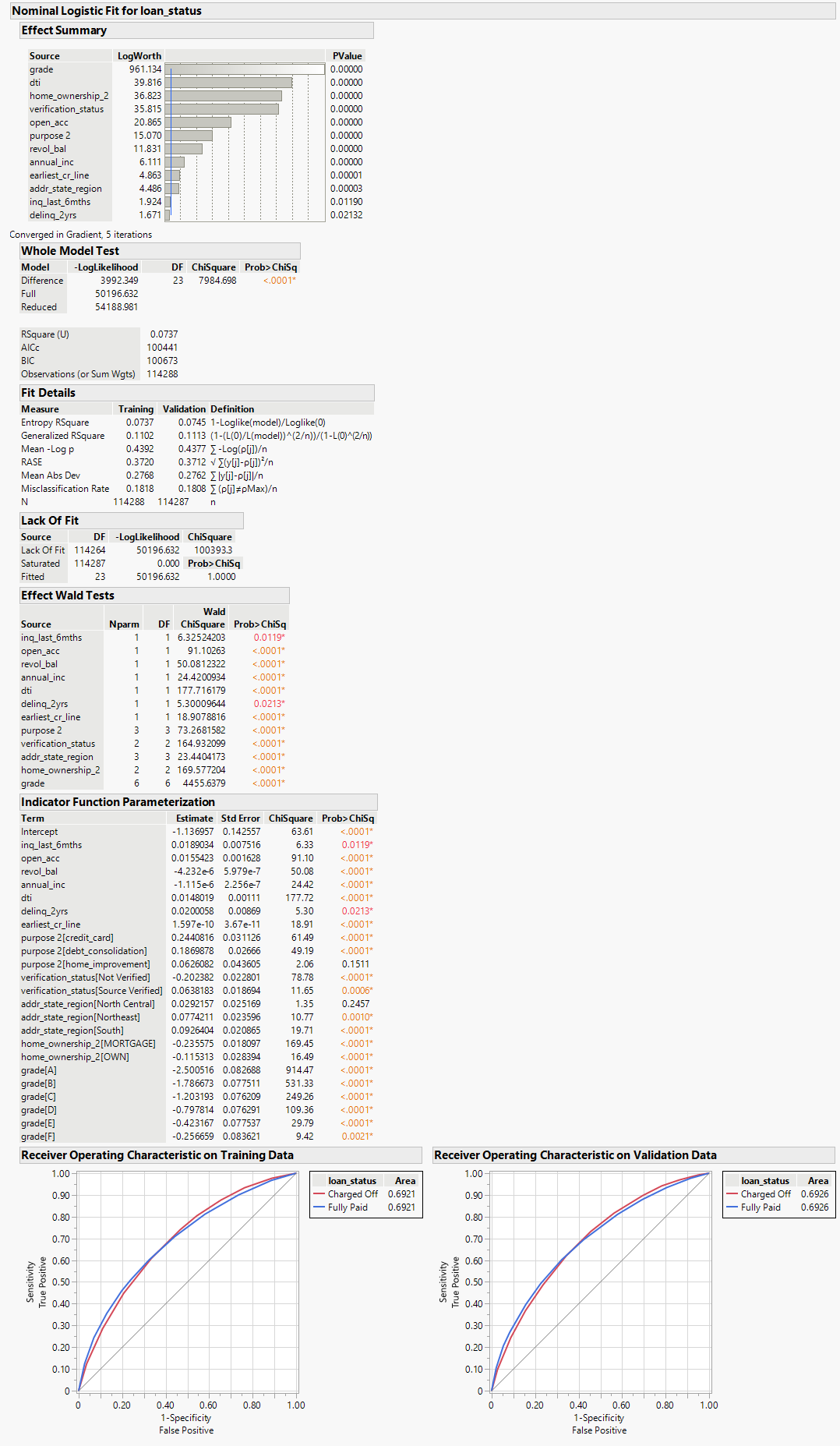


Figure 5. Final Logistic Regression Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Validation Dataset** | **Old Model** | **Decision Tree** | **Bootstrap Forest** | **Boosted Tree** |
| **AUC** | 69.26% | 68.48% | 69.21% | 68.94% |
| **Entropy R2** | 7.45% | 6.82% | 7.39% | 7.17% |
| **Generalized R2** | 11.13% | 10.21% | 11.03% | 10.71% |
| **RMSE** | 37.12% | 37.25% | 37.13% | 37.17% |

As per the table, the old logistic regression model is performing the best of all four models followed by Bootstrap Forest. The AUC for the logistic regression model (69.26%) is the highest of all 4 followed by Bootstrap Forest (69.21%). The entropy R2 is also the highest for the logistic regression model (7.45%) of all 4 followed by Bootstrap Forest (7.39). Generalized R2 is also the highest for the old logistic regression model (11.13%) followed by bootstrap forest (11.03%). Similarly, RMSE is lowest for the logistic regression model (37.12%) followed by the bootstrap forest model (37.13%). All the quantitative factors suggest that the logistic regression model is better. Similarly, the decision tree is performing the worst of all 4.