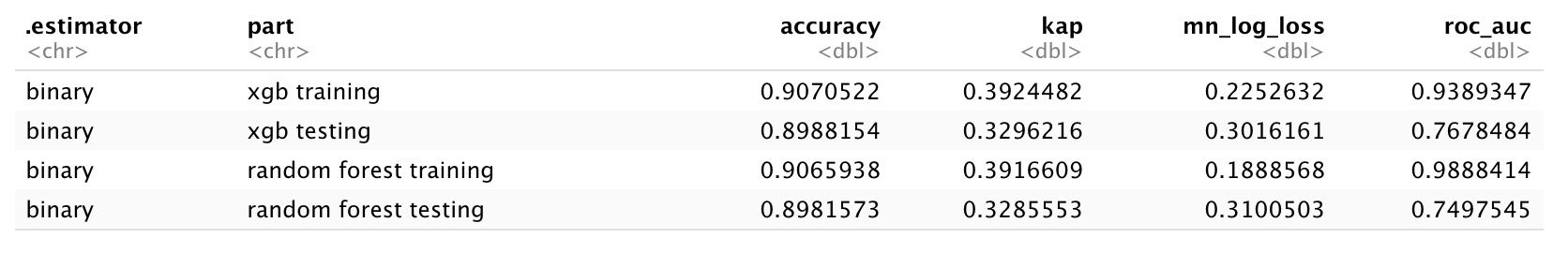
Yinan Guo

guoy221@wfu.edu

2021/11/14

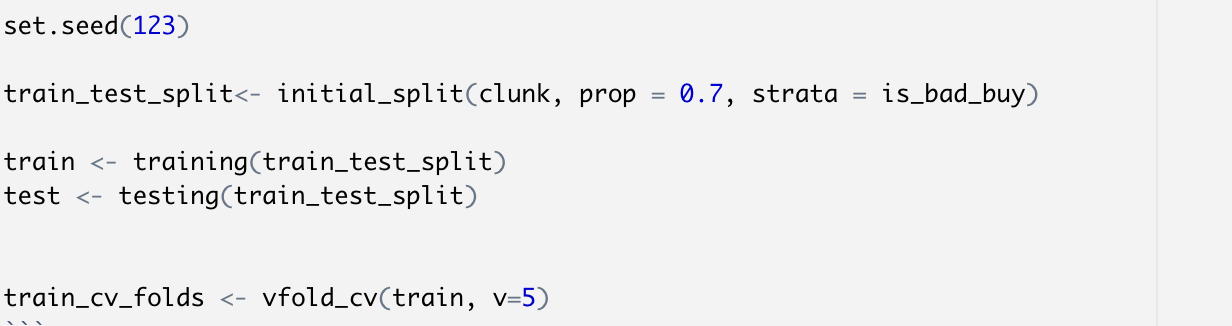
# **Model Performance**



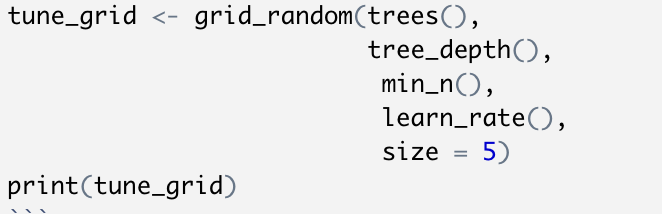
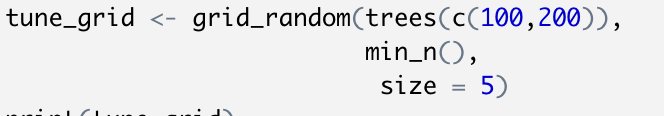
First of all, according to the metrics of the training and test dataset of the two models, both of them have serious over-fitting problem. In comparison, xgb performs better since the difference on AUC and log loss between test and training set are smaller.

Also, xgb has higher accuracy, higher AUC and lower log loss, which means that xgb will have stronger ability to tell the bad buy from not bad buy and give out a more accurate probability of whether the car is a bad purchase.

# **K-Fold Tuning**

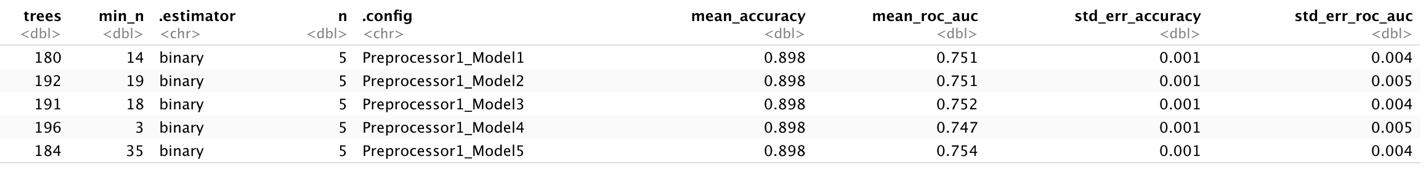


The train set is divided by 5 folds to make cross validation.



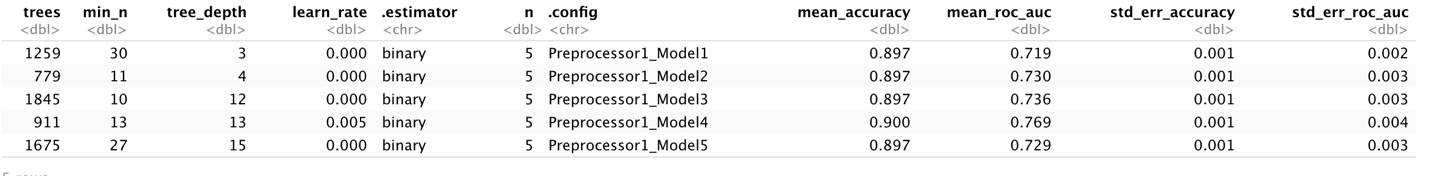
And in the tune grid part, the algorithm randomly selected 5 groups of different parameters for rf and xgb to build the model. The tuning results are shown as following:

* Random Forest



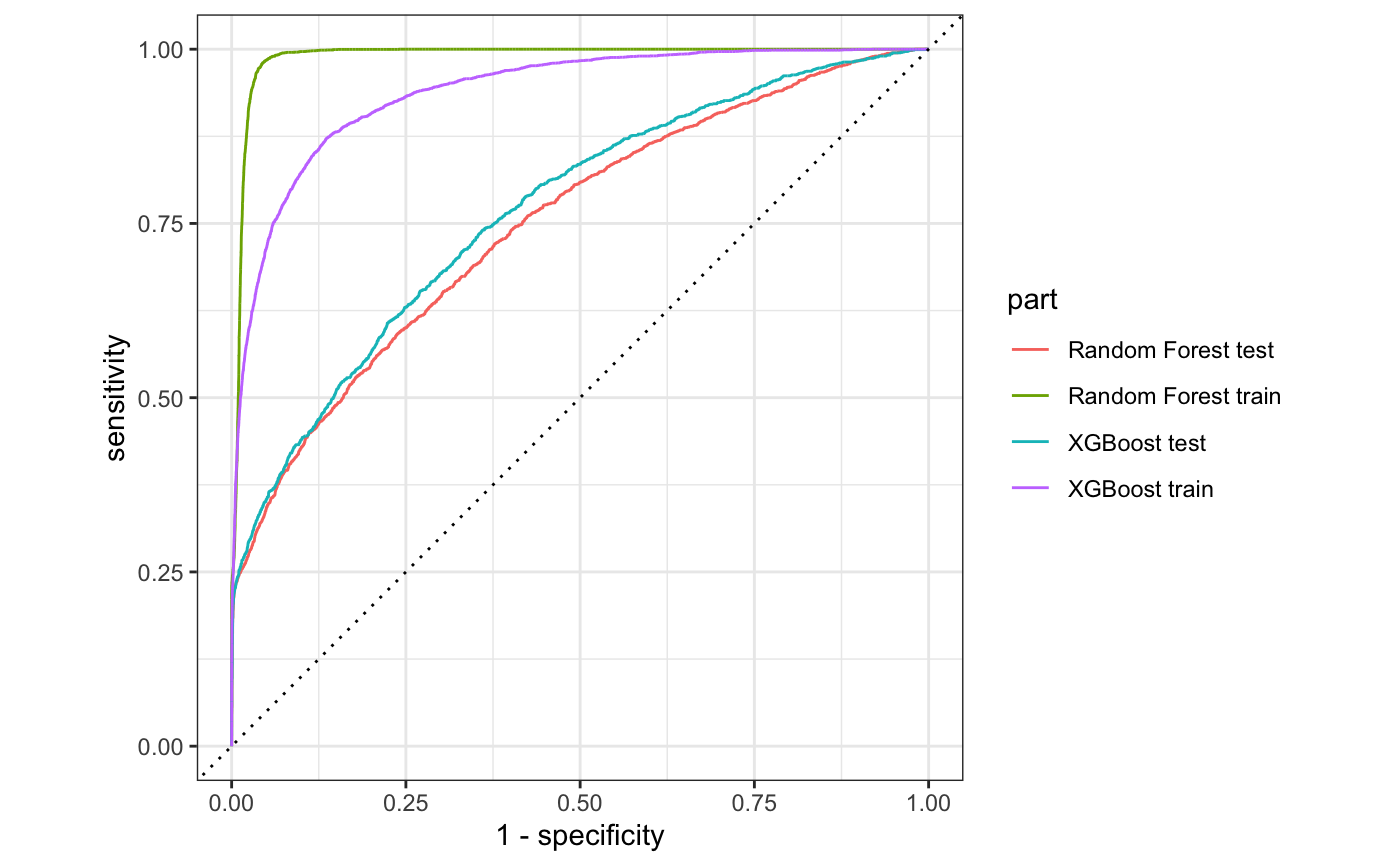
The best model in random forest is the one with 184 trees and min\_n of 35, which gives out a mean auc of 0.754 in 5 folds and mean accuracy of 0.898.

* XGBoost



The best model in XGBoost is the one with 911 trees, min\_n of 13, tree depth of 13 and learn rate of 0.00537, which gives out a mean auc of 0.769 in 5 folds and mean accuracy of 0.900.

# **ROC chart**



As shown in the ROC, XGBoost has better curves whose area under the curve are closer to 1 for both training and testing set. However, the overfitting problems is still severe.

# **Variable Importance**

|  |  |
| --- | --- |
| Random Forest | XGBoost |
|  |  |

The most important predicters in random forest model are wheel type (whether it is NULL), purchase date (which is shown in the picture as purchday, which is the day from purchase date till now) and buying cost. The most important predicters in XGBoost model are wheel type, vehicle age and purchase date.