A blue and yellow logo

Description automatically generated

FLINT LINE REPLACEMENT PROJECT

DS 633 – DATA MINING FOR BUSINESS ANALYTICS

Professor: David Kaufman

Team:

Kasey Hunter

Ravi Teja Lakkakula

Padmini Poojari

Contents

[**1.**](#_heading=h.gjdgxs) **INTRODUCTION** 3

[**2.**](#_heading=h.30j0zll) **DATA COLLECTION & PROCESSING** 3

[**3.**](#_heading=h.1fob9te) **ANALYSIS** 3

[**4.**](#_heading=h.3znysh7) **MODEL BUILDING** 6

[**4.1.**](#_heading=h.2et92p0) **Logistic Regression** 6

[**4.2.**](#_heading=h.tyjcwt) **CART Analysis** 6

[**4.3.**](#_heading=h.3dy6vkm) **Bootstrap Forest** 6

[**4.4.**](#_heading=h.1t3h5sf) **Boosted Tree** 7

[**4.5.**](#_heading=h.4d34og8) **Neural Net** 7

[**5.**](#_heading=h.2s8eyo1) **COMPARING MODELS** 7

[**5.1.**](#_heading=h.17dp8vu) **Ensembles** 7

[**6.**](#_heading=h.3rdcrjn) **CONCLUSION** 8

[**8.**](#_heading=h.lnxbz9) **APPENDIX** 8

[**8.1.**](#_heading=h.35nkun2) **Fit Details** 9

[**8.2.**](#_heading=h.1ksv4uv) **ROC Curve** 10

[**8.3.**](#_heading=h.44sinio) **Confusion Matrix** 12

[**8.4.**](#_heading=h.44sinio) **Instructions for Scoring** 13

## **INTRODUCTION**

Tasked with helping to reduce the cost of inspecting for and replacing lead pipes in the city of Flint, Michigan, the analytics team present machine learning models for consideration. The team outline several methods, compare their performance, and suggest a model they believe will best predict the location of lead in order to guide the city as emergency management invests in line replacement.

To understand the goals of this study, one must understand the history of the issue. The Flint Water Crisis stands as a reminder of the devastating consequences stemming from neglect of critical infrastructure due to the worsening economic situation in the area. In attempts to ease the financial burden, the emergency city manager switched the water source to the Flint River. However, the failure to treat the new water source caused leaching of lead from aging pipes into the city's drinking water, resulting in a significant health hazard for thousands of residents.

In response to this urgent situation, part of the city’s comprehensive response continues to be to locate and prioritize the replacement of lead water service lines. Following the unethical behavior and subsequent firing of the previous firm, this study aims to leverage machine learning algorithms to implement an evidence-based solution for the city. This machine learning model will assist in optimizing the emergency manager’s infrastructure investment strategies, not taking advantage of the city for profit.

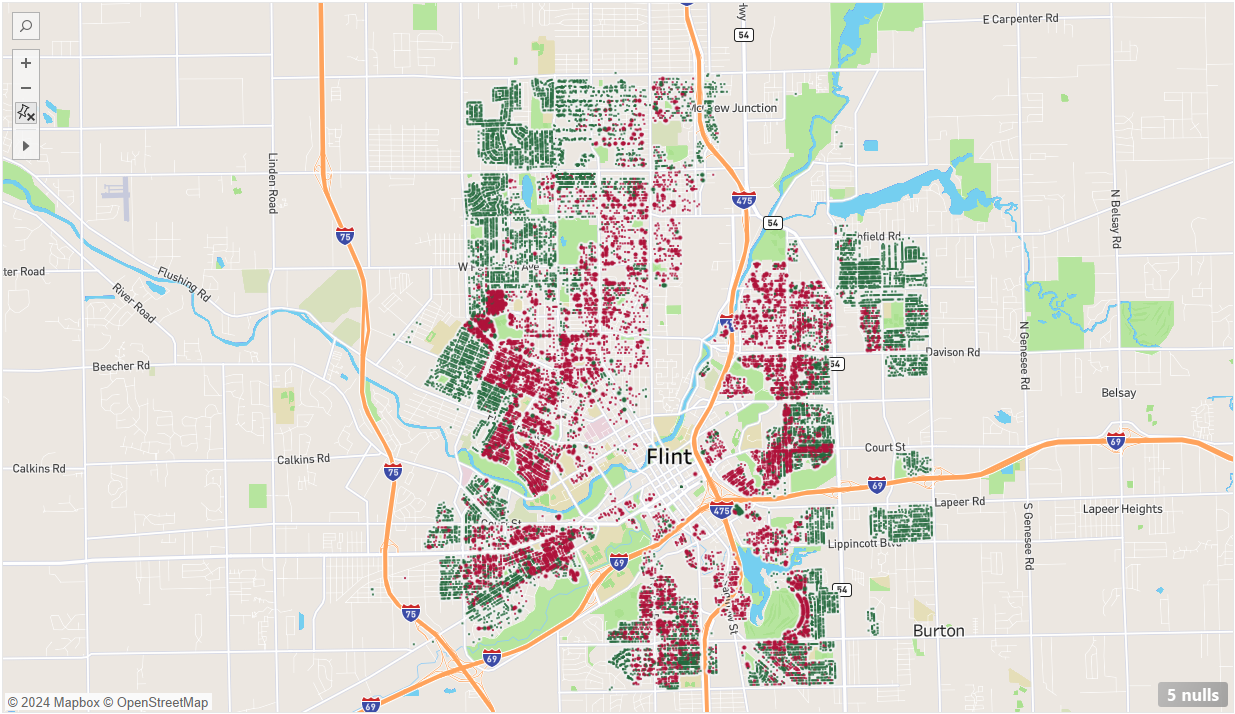
## **DATA COLLECTION & PROCESSING**

We combined data from publicly available sources such as government data, historical records and more recent inspections of service line type, and water sample test results. The governmental data contained a vast variety of variables detailing information such as zoning, value, use type, year built, and various survey and census data. The historical records represent an imperfect archival source of lead pipe location. More recent records were provided by the private home side inspections. Finally, water samples from some parcels were tested for lead. Combining several of these data sources will help us build better predictions of where the lead resides in order to prioritize further investigation in areas of high risk.

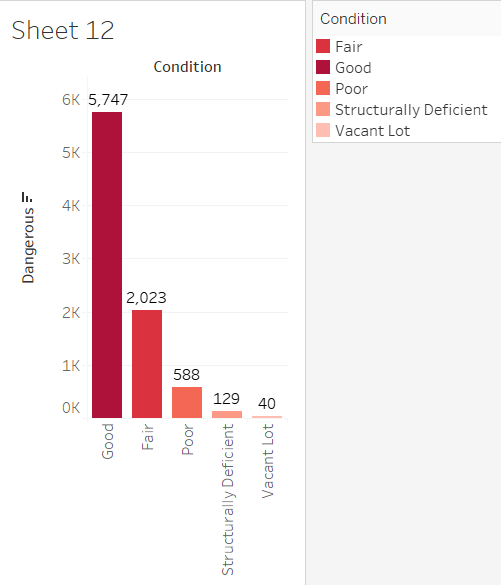
This data underwent extensive cleaning and processing before use in the models, including identifying missing values, combining variables, removing unneeded or redundant variables, and excluding records.

## **ANALYSIS**

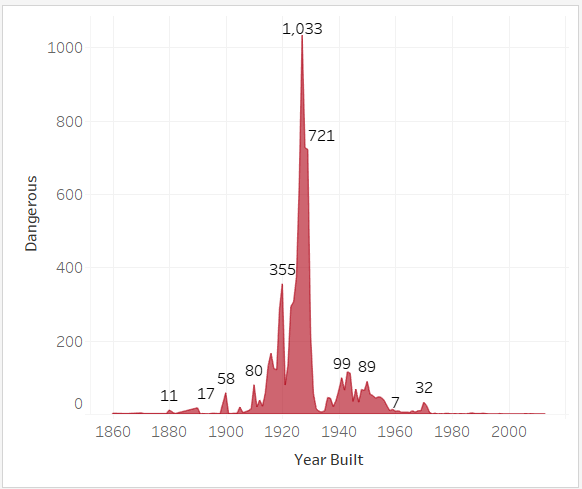
The machine learning models created were classification models, meaning each predicted if the parcel would be dangerous (likely to have a lead service line) or safe. Initial models include logistic regression, and CART analysis. Types of ensemble methods such as bootstrap forest model, and a boosted tree were then created. Using information on variable importance from our various tree models, we then created a neural network. Finally, we ensembled our best models, aiming to perform even better together than separately.



Where red indicates dangerous and green is not dangerous. This map indicates dangerous parcels in our training data. Finding the characteristics that predict these dangerous parcels will help us predict where to locate additional dangerous parcels.



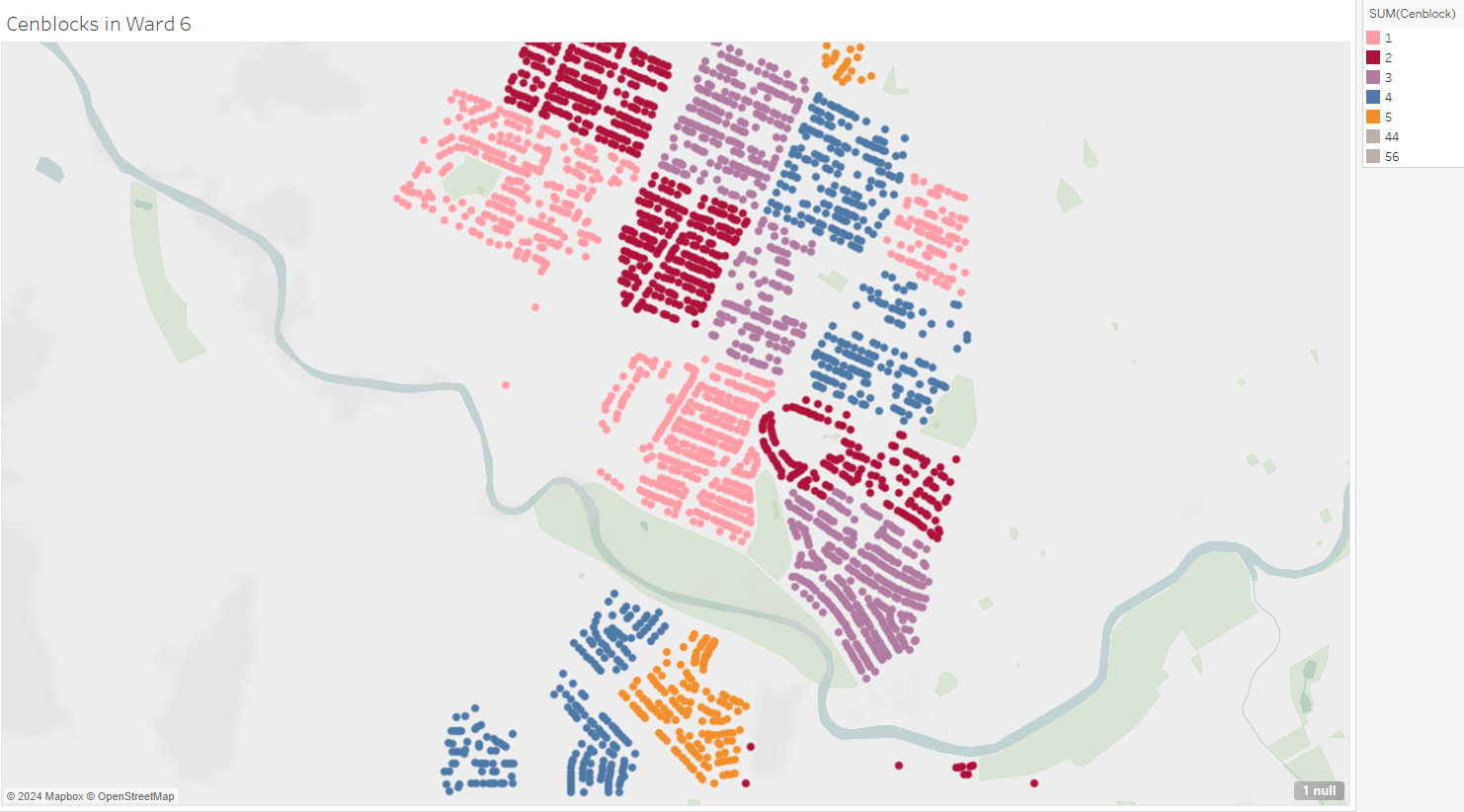
Why do we need to dig deeper into many predictors? This chart shows the number of dangerous samples found per condition of the property. Being in good condition does not mean the property will not have dangerous lead. We need to let the data mining methods and algorithms tell us which variables and their levels correspond to dangerous lead rather than making too many assumptions as the analysts.



For example, one of the variables we know from prior analysis is important is year built. There used to be widespread use of lead pipes until a ban in 1986. In combination with other predictors, year built can be a part of a good classification model.



With the aim of focusing more on the residential houses with USPS not being vacant which represents someone is staying at home compared to the residential houses with USPS vacant where there is no one staying. Based on the data the cenblock 2 has more lead pipes based on the dangerous field. Among the wards 6 stands out with the percentage of the houses that need lead pipes replaced. The purpose of this chart is to focus deeper in the areas where lead concentrated houses are more to better utilize the government budget.



When we look closer every ward is divided into cenblocks and based on the above data it is better to focus on the cenblock 2 from ward 6 and then followed by 1, 3 and so on according to the projections.

## **MODEL BUILDING**

### **Logistic Regression**

Logistic regression is a model that can be used for classification, especially where there are linearities in the data. However, the results given this data were not as compelling compared to other models. Given the amount of categorical variables with many values we have in the data (especially from our governmental sources), it would require extensive binning of the data to arrive at stable results. For these reasons, the logistic regression model is not the recommended model for this application. Validation performance metrics include: RSquare of .7380, misclassification rate of 9.25%, and AUC of 0.9521.

### **CART Analysis**

CART analysis here refers to a classification tree. Using this method, parcels can be classified as dangerous or not. The results of the CART analysis were more compelling than logistic regression. One advantage seen over logistic is the ability to bin data itself, meaning this method will be less intensive in the pre-processing of data than the logistic regression. Despite better performance than the logistic regression, other models show even better performance than the classification tree. Validation performance metrics include: R Square of .7933, misclassification rate of 7.8%, and AUC of 0.9578.

### **Bootstrap Forest**

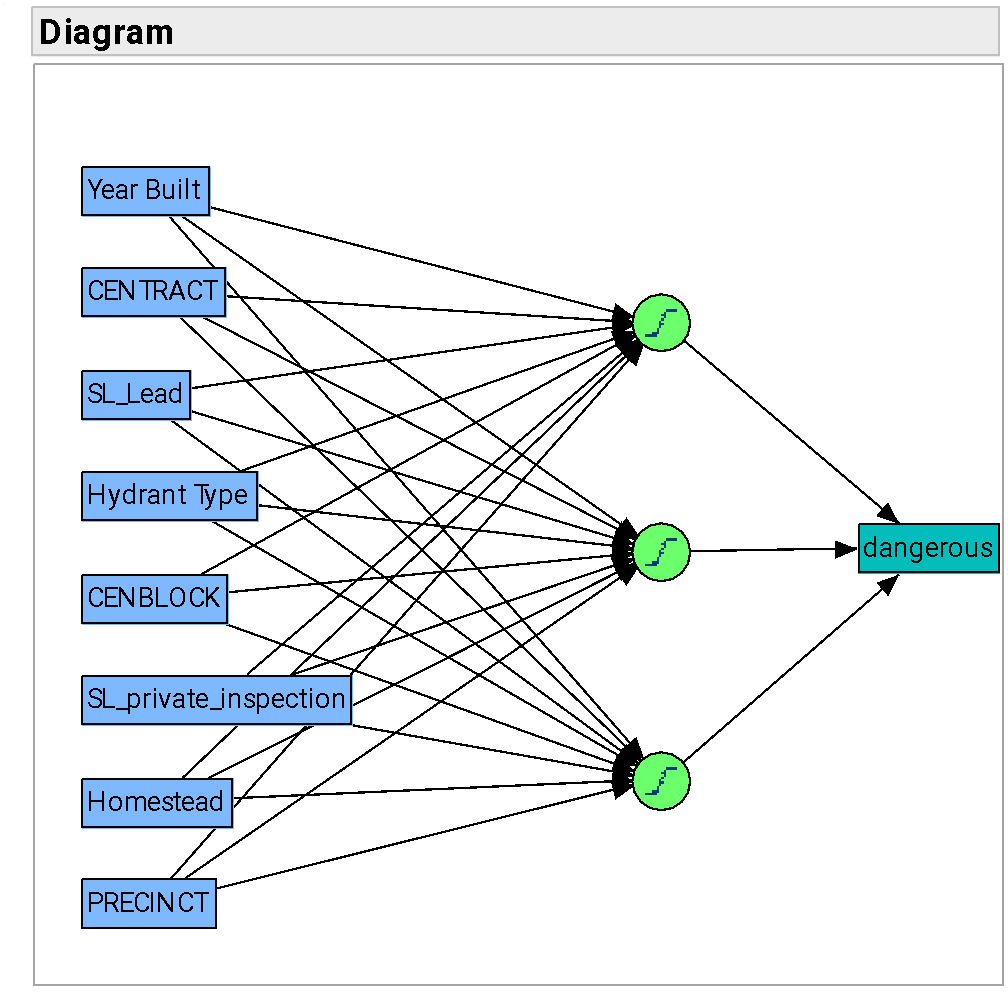
The bootstrap forest method combines CART analysis with ensembling. This method of ensembling essentially produces many models based on different samples from the data set and combines their results. This bootstrap forest model produced our best results. Interestingly, this was the method with the highest validation AUC but also is a method that helps avoid overfitting, indicating its performance will likely stay strong as it is applied to test data. Validation performance metrics include: R Square of .8182, misclassification rate of 7.46%, and AUC of 0.9684.

### **Boosted Tree**

A boosted tree combines the methods of CART analysis with a form of ensembling. It can be a more powerful model than others as it works to improve upon areas where the model makes prediction errors. This was one of the stronger methods produced. Validation performance metrics include: R Square of .7981, misclassification rate of 7.89%, and AUC of 0.9617.

### **Neural Net**

A neural network is a method that can be used for classification. In JMP, the neural net takes inputs and optimizes algorithm values that minimize errors. To determine the inputs, we took the top variables that proved important in our tree models as those models were the best performing. The neural net performed better than logistic regression and CART analysis but could not beat the ensembled tree methods. Validation performance metrics include: R Square of .7903, misclassification rate of 7.97%, and AUC of 0.9597.



Inputs into the neural network, hidden layer, and output layer shown.

## **COMPARING MODELS**

### **Ensembles**

Knowing that an ensemble, or a combination of, models often results in more accurate predictions, one such model is presented. Some of the best performing models were the methods which use forms of ensembling in their algorithms: bootstrap forest and boosted tree. Their performance speaks to the power of ensemble and mark them as the models to try in an ensemble model combining the two.

The ensemble method used was to average the propensities. The ensemble of these two models resulted in performance between the bootstrap forest and boosted tree. Validation performance metrics include: R Square of .6825, misclassification rate of 7.67%, and AUC of 0.9670.

## **CONCLUSION**

Given the vast number of predictors we could include in the models, we used various methods to see which emerge as important throughout our analysis. This includes using stepwise for logistic regression and using methods like trees which have a built-in selection. In general, year built (whether continuous or binned into groups) was important. We also saw variables related to location emerge in many models such as precinct, census tract, and census block. This means, in areas where we find dangerous levels of lead, we’re likely to find more in the area. Perhaps unsurprisingly, variables related to historic records (SL\_Lead) or actual inspections (SL\_private\_inspection) for lead appeared in several models. These variables indicate that the information about lead location that we do have, although imperfect and incomplete, is useful to our analysis. Finally, hydrant type surfaced as a useful predictor.

As for the models, tree methods performed well overall. The best performance, if judged on area under the ROC curve, is the bootstrap forest model. For context, perfect performance on this metric is equal to 1. For validation data, the AUC of the ROC for the bootstrap forest is .9684. Additionally, this type of ensembled tree is a machine learning method that helps to reduce overfitting to data. This means, it should continue to perform well when used on new data once it is deployed for use.

## **RECOMMENDATIONS**

Given an extensive review of appropriate methods, the recommendation is to deploy the bootstrap forest model. This model should help the city make more accurate predictions regarding lead. With better predictions, budget utilization will be improved. Most importantly, acting upon these predictions where there is most likely to contain lead will help to save the citizens of Flint from further harm.

## **APPENDIX**

### **Variables**

The variable “Condition” was created from Housing Condition 2012, Housing Condition 2014, and Commercial Condition 2013. This variable was created by concatenating all three variables. From there, using recording, we used the most recent or the non-missing for housing conditions. Then for commercial conditions, that would be used if available. Only one record contained a value for all three columns where we used the mode. This left me with only a handful of missing values.

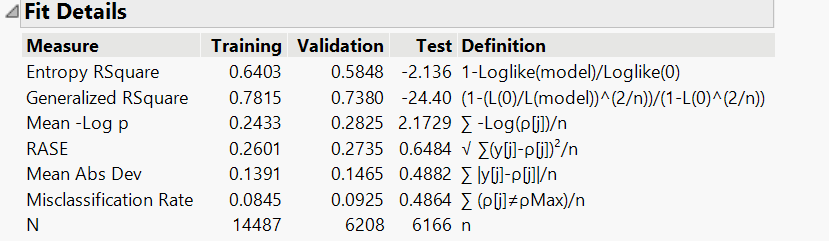
The variable “Building Value” is simply the sum of Residential Building Value and Commercial Building Value

### **Technical details of Models**

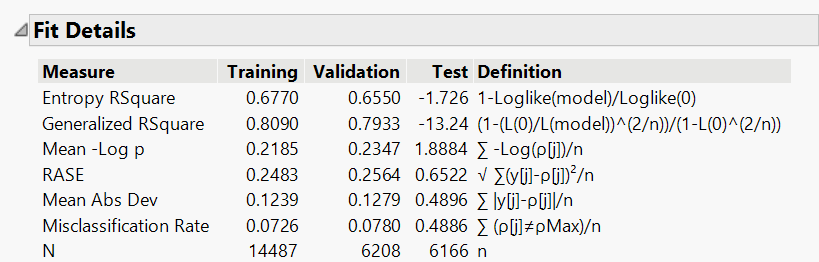
### 

### **8.2.1. Fit details**

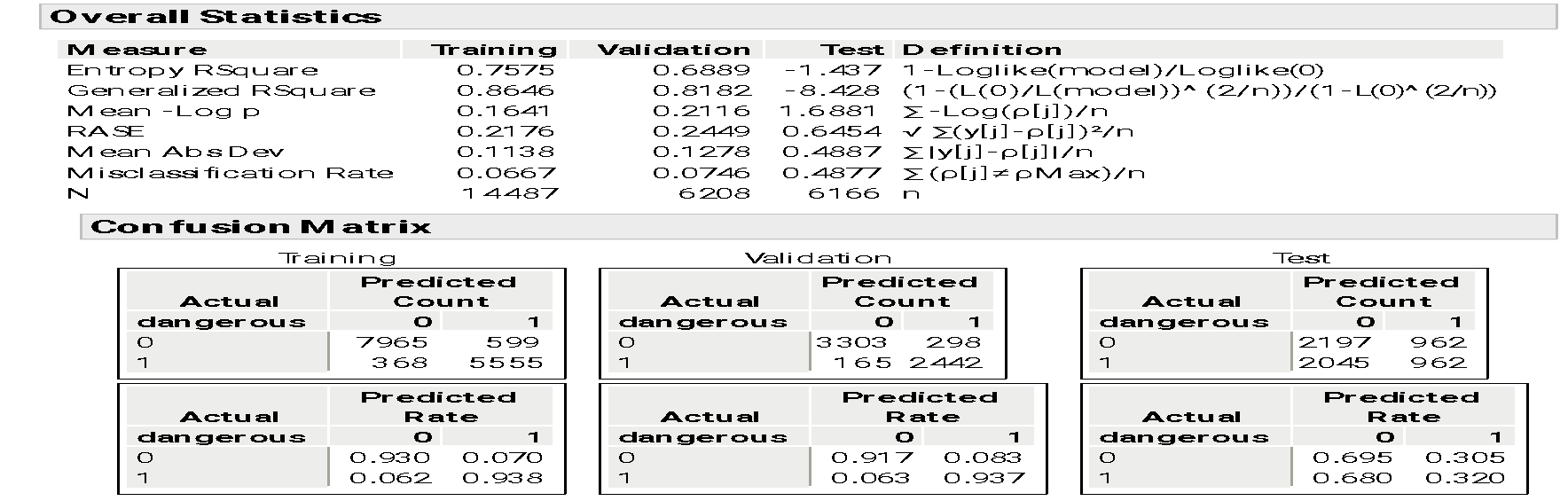
Logistic regression:



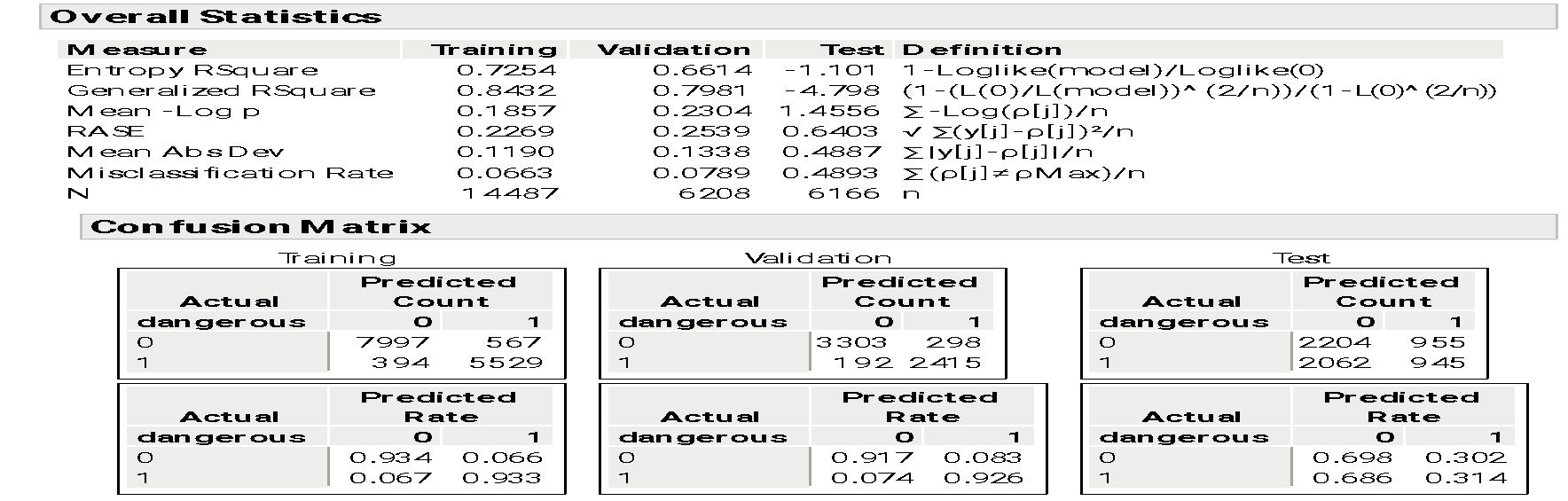
CART Analysis:



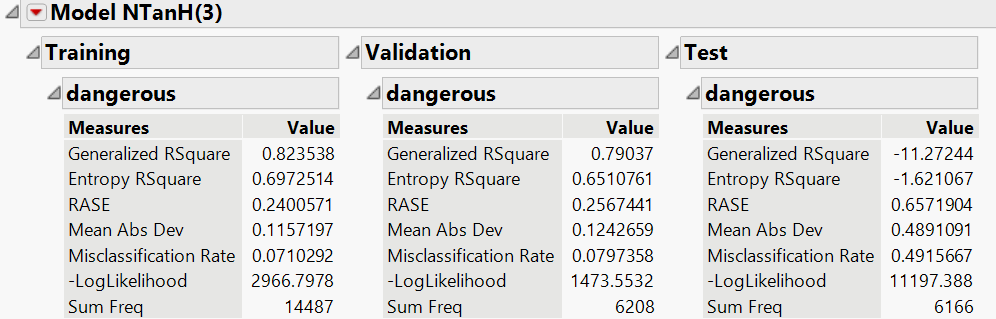
Bootstrap forest:



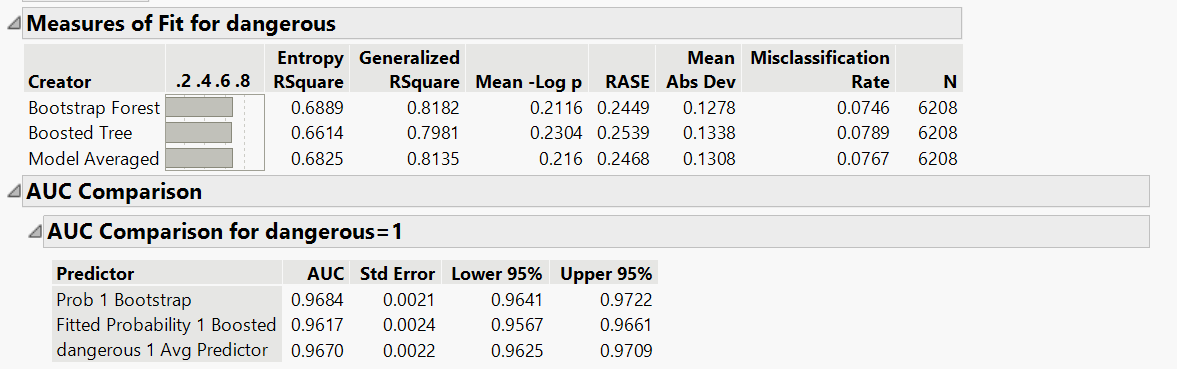
Boosted tree:



Neural network:

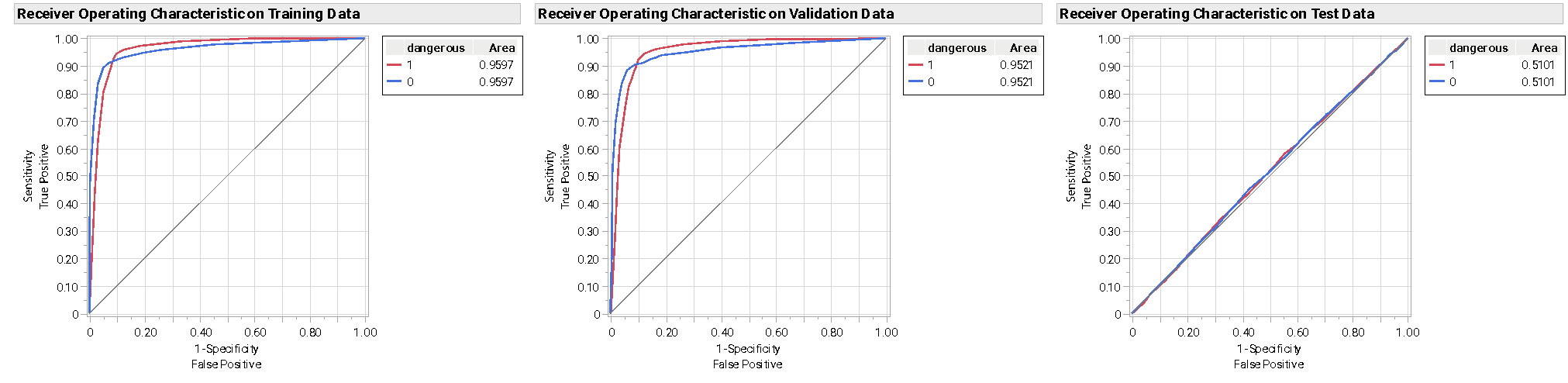


Ensemble of bootstrap forest and boosted tree:

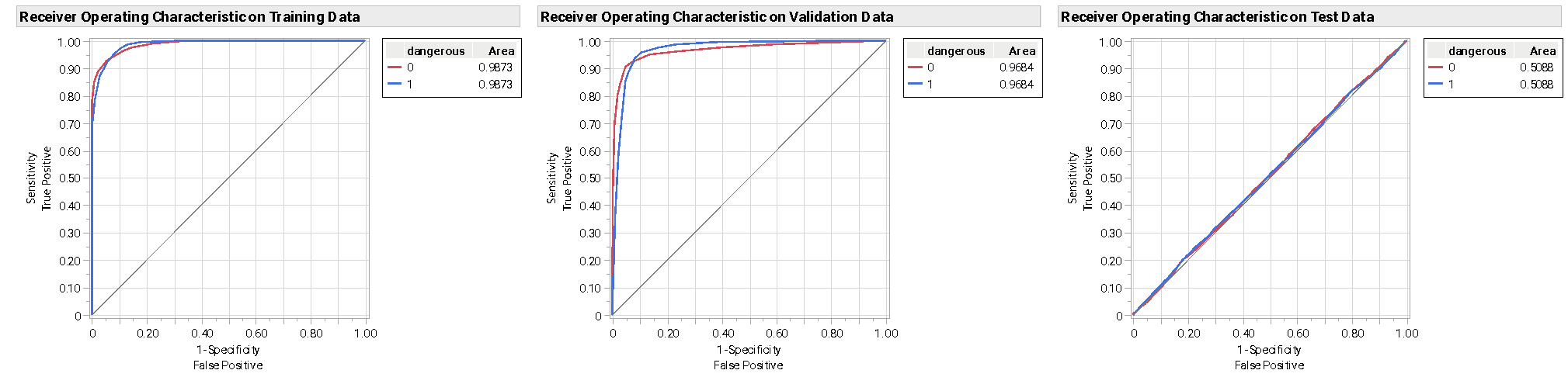


### **ROC curve**

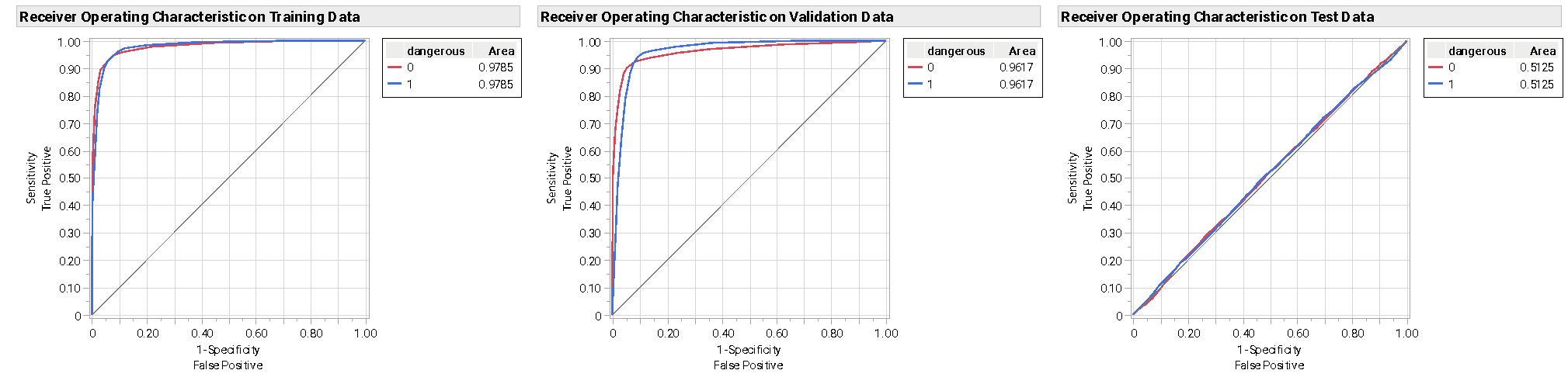
Logistic regression:



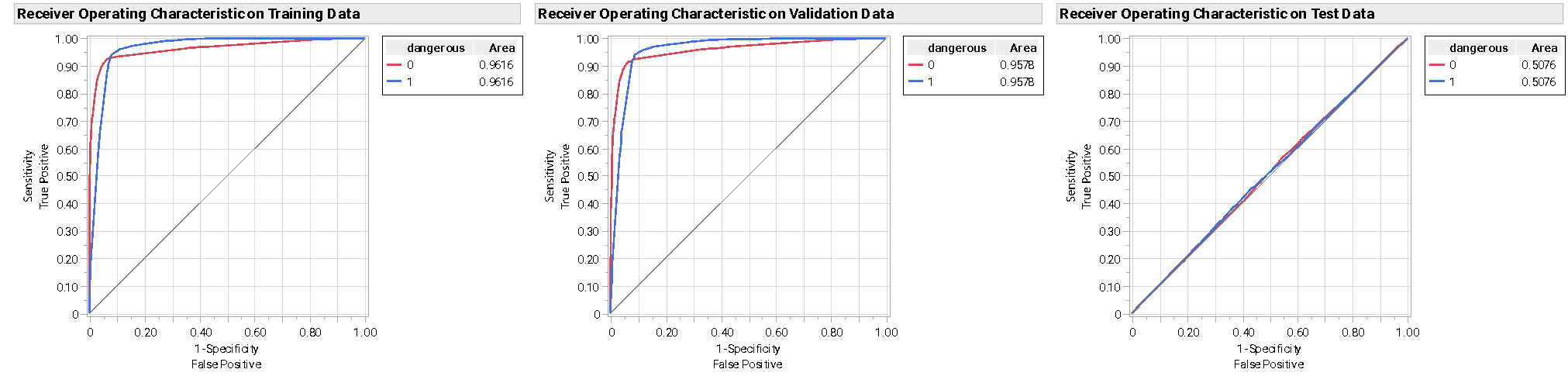
Bootstrap forest:



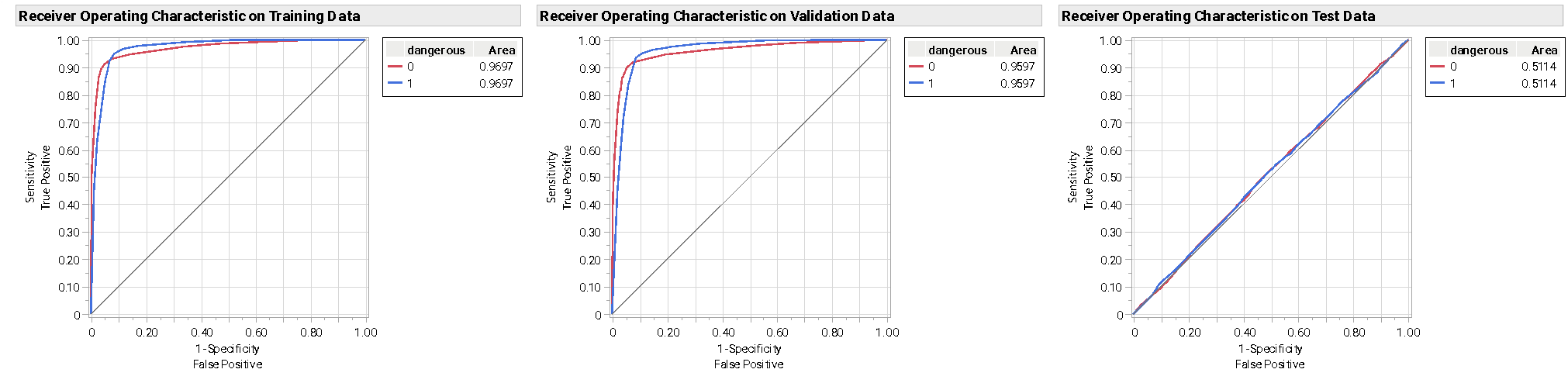
Boosted tree:



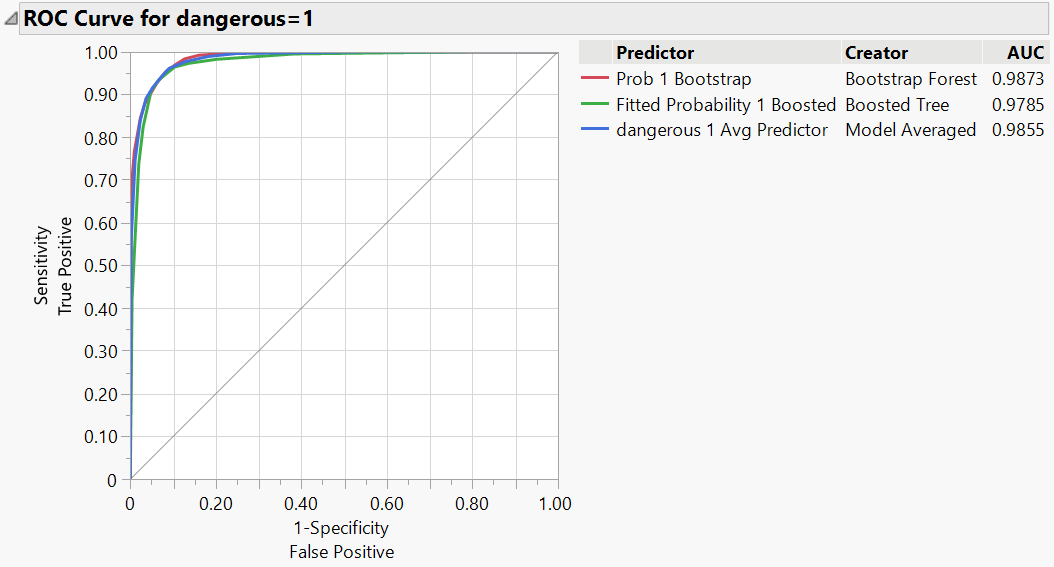
CART Analysis:



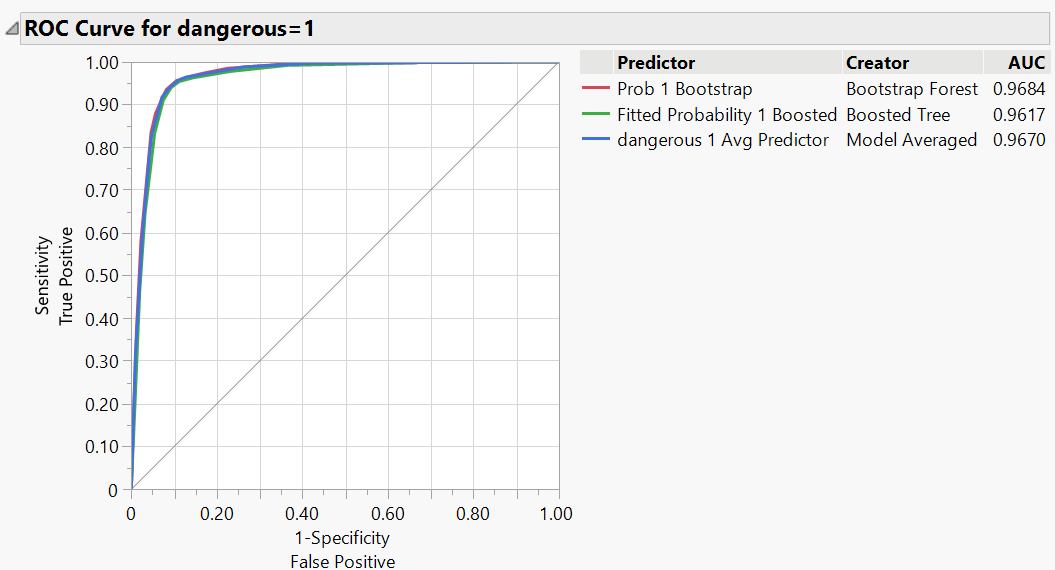
Neural network:



Ensemble of bootstrap forest and boosted tree:



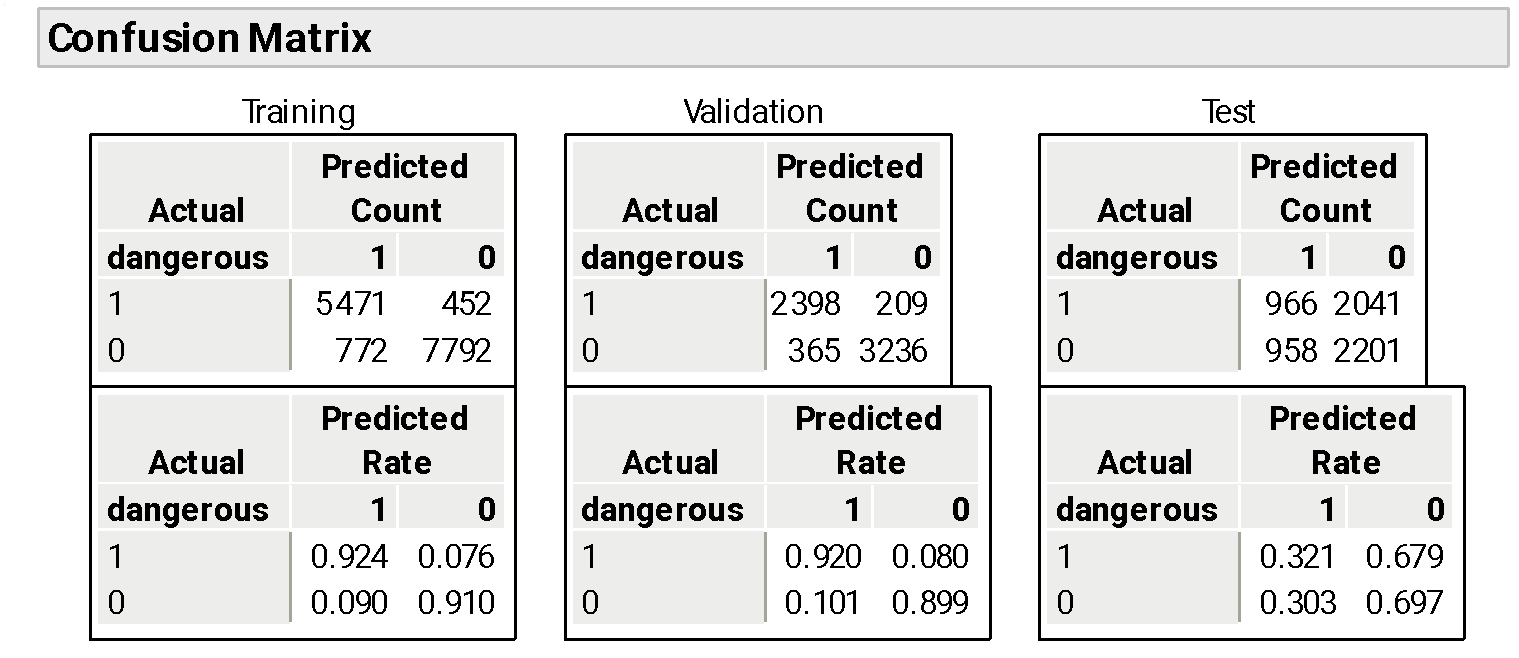
ROC Curve for training with ensembled model in blue



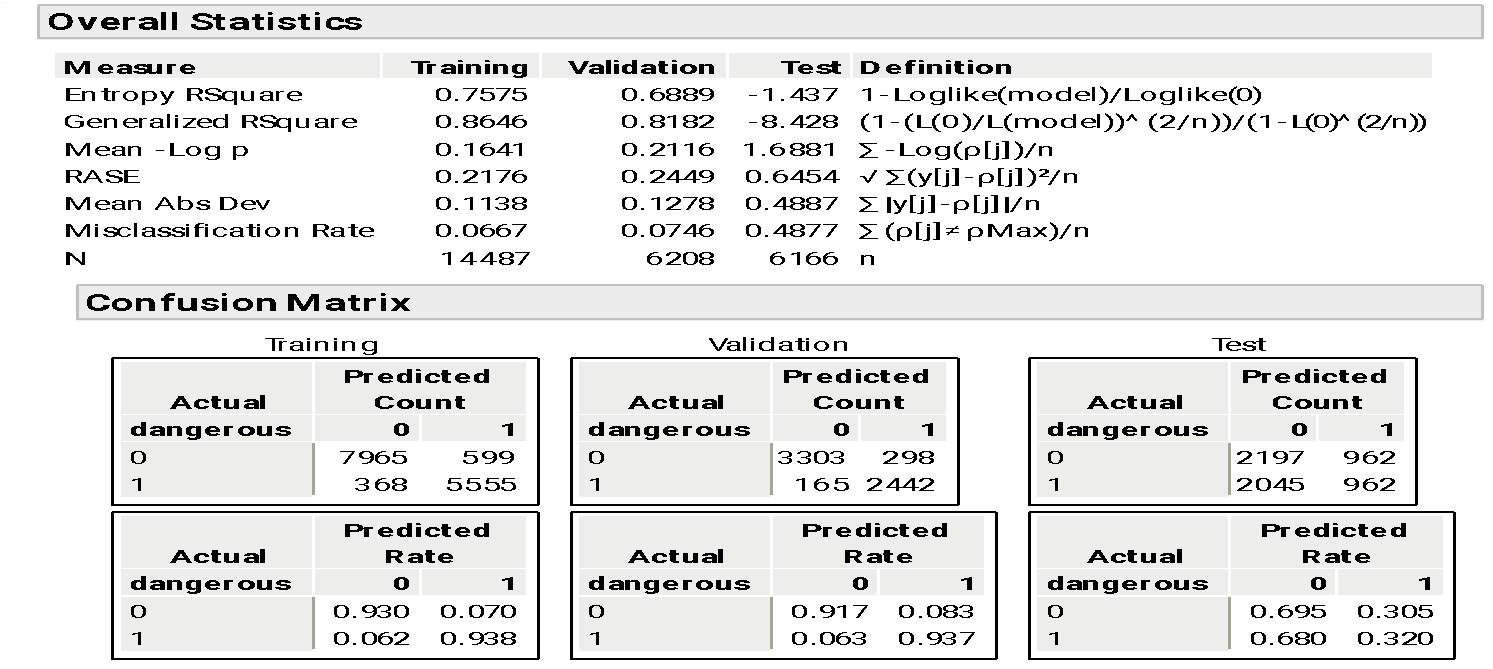
ROC Curve for validation with ensembled model in blue

### **Confusion matrix**

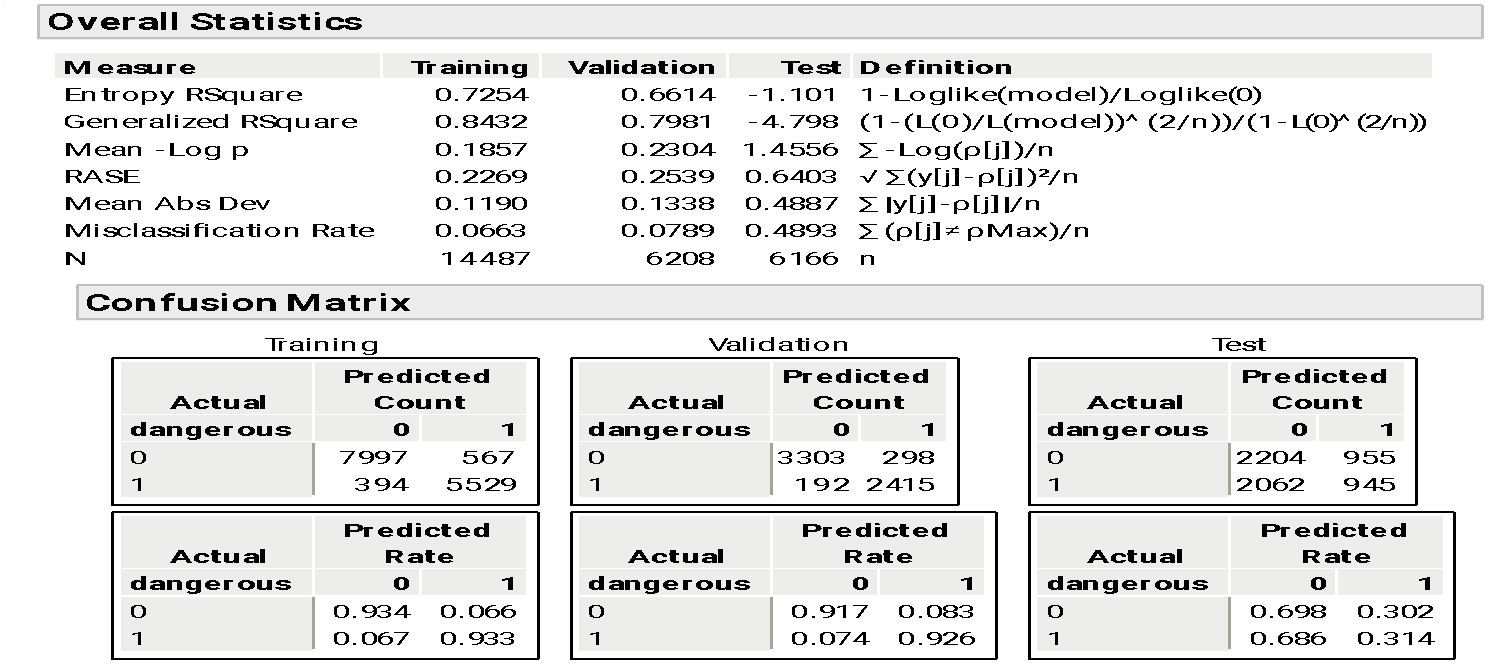
Logistic regression:



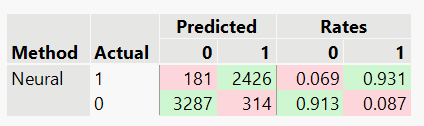
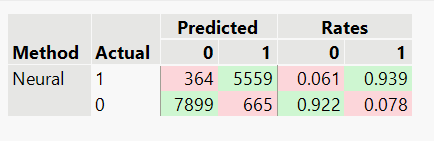
Bootstrap forest:



Boosted tree:

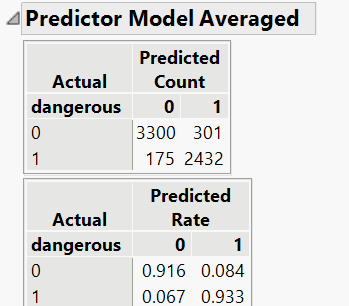
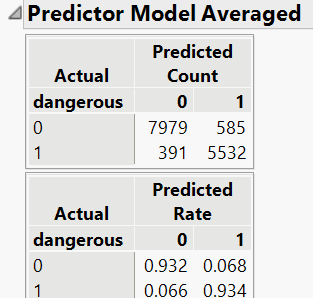


Neural network:



training validation

Ensemble:



training validation

## **INSTRUCTIONS FOR SCORING**

Step 1: Replace the values of the dangerous column for the 6,166 test set records into the existing dataset, starting from Row 20698 where the Test Data section is located.

Step 2: First, run the script for “Bootstrap Forest for Dangerous” and the ROC curve for the test set should now display accurate AUC. This is the model we suggest as the best model overall. Can see if this prediction holds for the test data.

Step 3: One can run the scripts for all the models (Fit Nominal Logistic, Boosted Tree of Dangerous, Decision tree of Dangerous 2, Neural of Dangerous) if needed to verify the best model amongst.

Step 3: And then for the Ensemble, to use this model with test data, must rerun “model comparison for BoostedTree BootstrapForest” script, select average the models in model comparison drop down, new predictor columns will appear, and rerun model comparison with the new averaged predictors.

This allows verifying each model in addition to one we predict will be the best overall.