Bank Customer Churn Predictions

A random forest model, a decision tree model, and a gradient boosted tree model were used in comparison of model accuracy. The priorities of an accurate model are as follows:

1. Precision was given highest priority
2. Recall was given second priority
3. Accuracy was given third priority

All performance statistics, 10-fold cross validation, ROC curves, and a correlation matrix all contributed to the overall decision of best model.

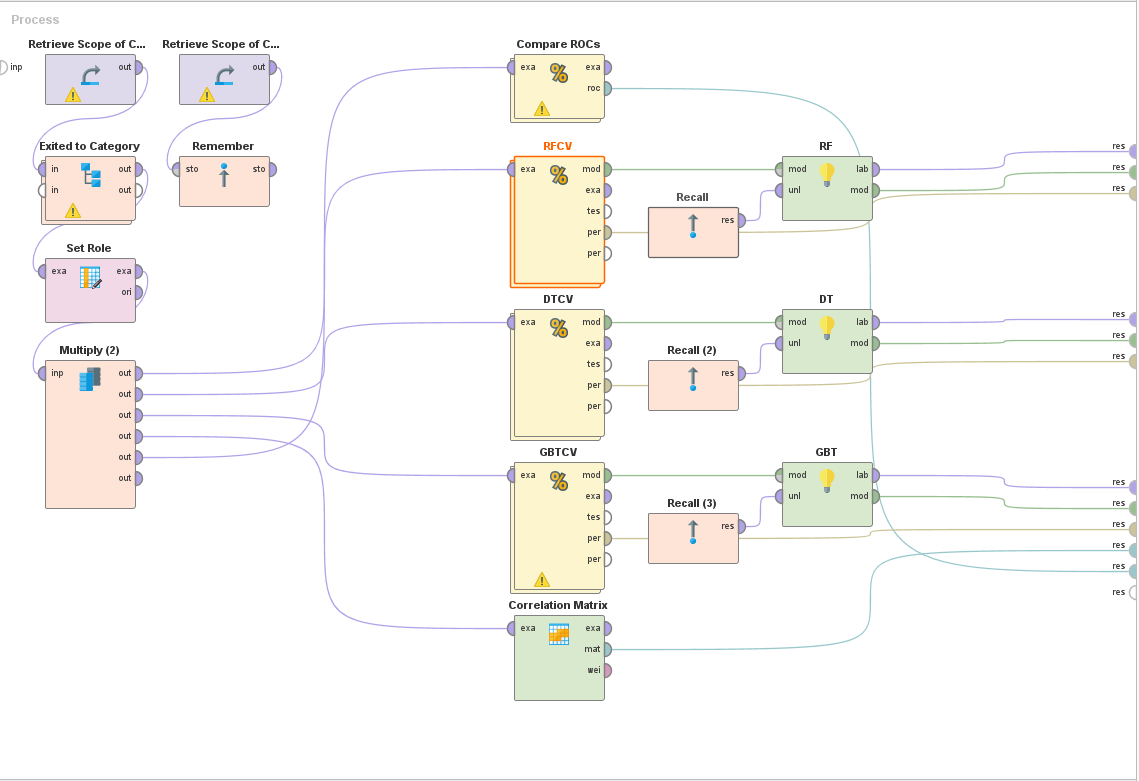
# **Data Preprocessing**

Various level of preprocessing was used to “play” around with results. It was assessed that there were no missing values and surname was found to be a useless variable. According to the outlier detection model, there were about 90 or 91 various outliers.

Ultimately the tree models won bout because of the tradeoff that existed. For relatively small amount of preprocessing a very accurate model was achieve compared to other models such as logistic regression or naïve bayes. With that said the only change made was to convert the variable exited into a category with 0 meaning no churn and 1 meaning churn everything else was left as is.

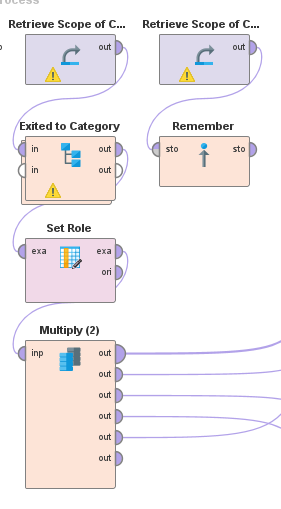
# **RapidMiner Process**

The process looks a bit convoluted but let us dive into it a bit more

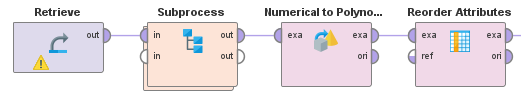


## **Data Retrieval**

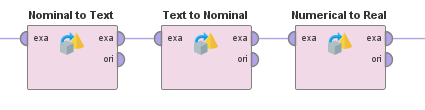
First, we retrieve the production data and store it for later use.



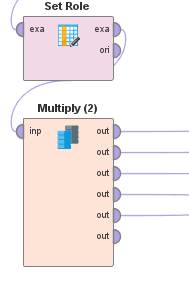
Second, the training set was retrieved and fed into a subprocess that changes exited to a category of 0 and 1, again 0 meaning no churn and 1 meaning churn was retrieved and fed into a subprocess that changes exited to a category of 0 and 1, again 0 meaning no churn and 1 meaning churn.



The subprocess within the exited to category simply changes all variables with a type of value.

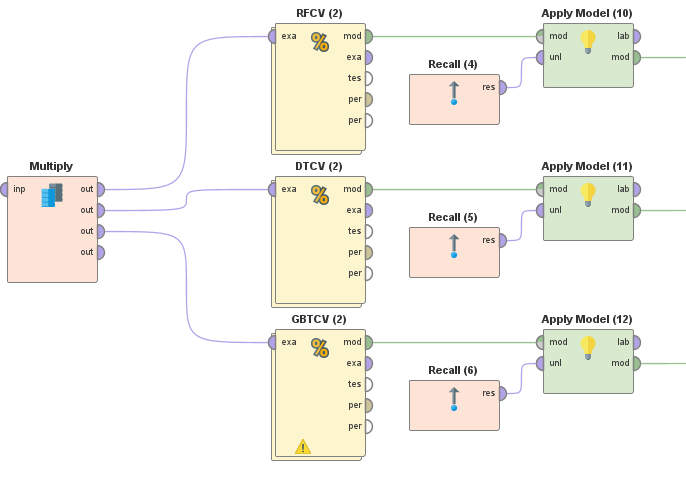


After the role of “Exited” is set to label to be able to predict churn. Then the example set is multiplied to be able to connect to various other operators.

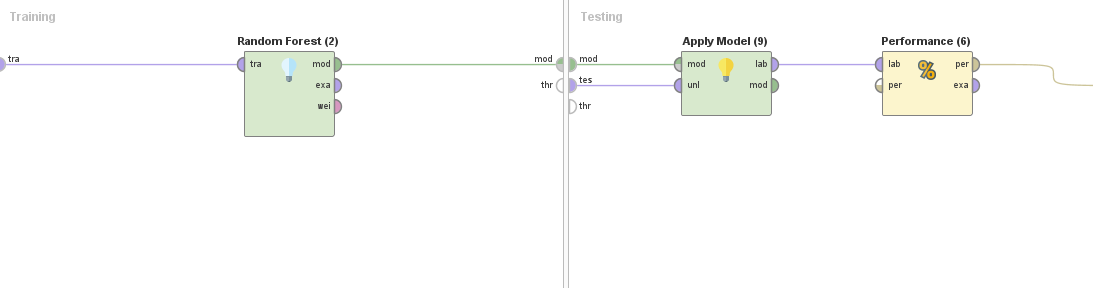


## **Modeling**

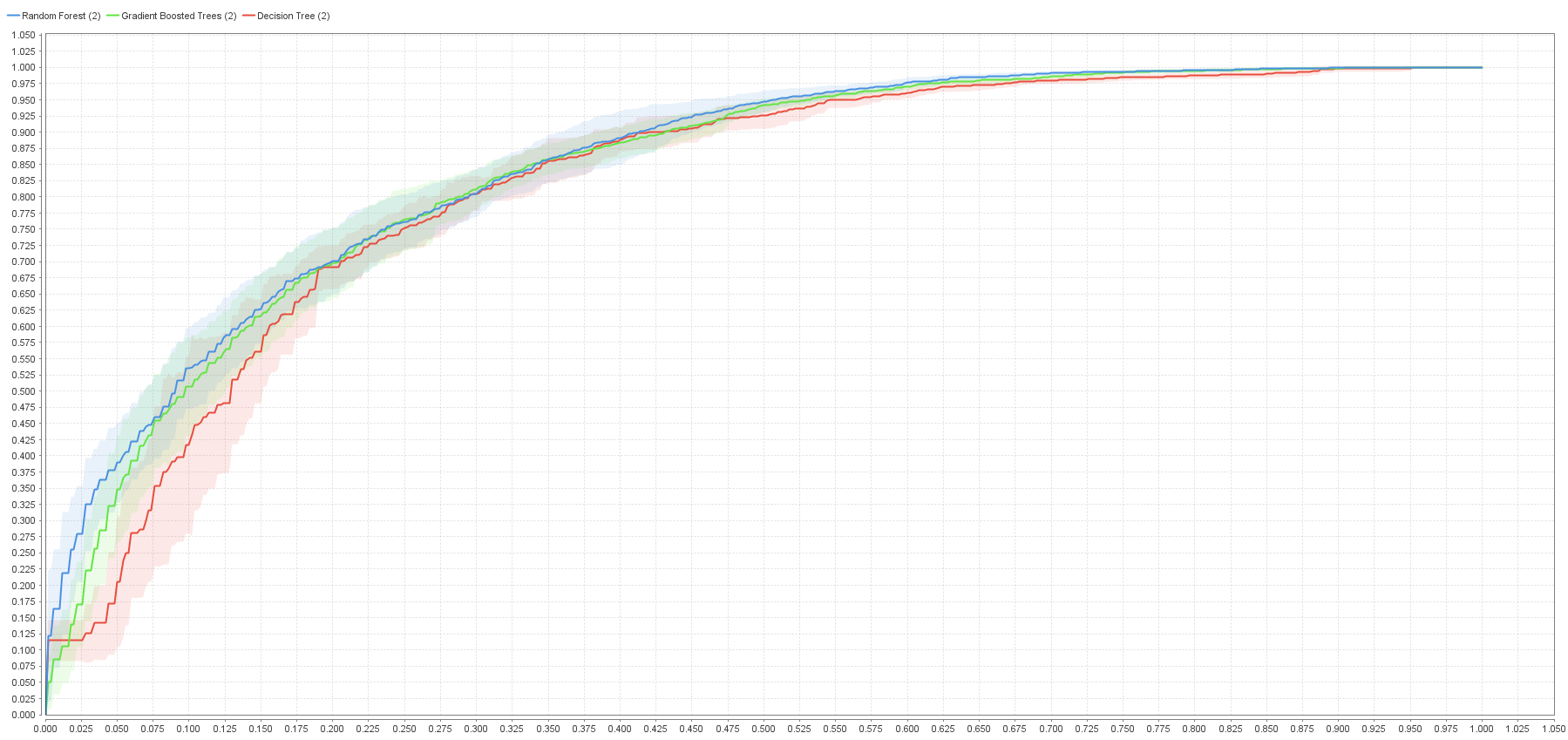
For the modeling part of the process, there are some different steps. First step is to compare the ROC curves. The “Compare ROC” subprocess has the same models used for prediction, but these are just used for modeling and ROC. The training set that was multiplied is fed into the compare ROC operator and multiplied again to be connected to each of the three models’ cross-validation process.



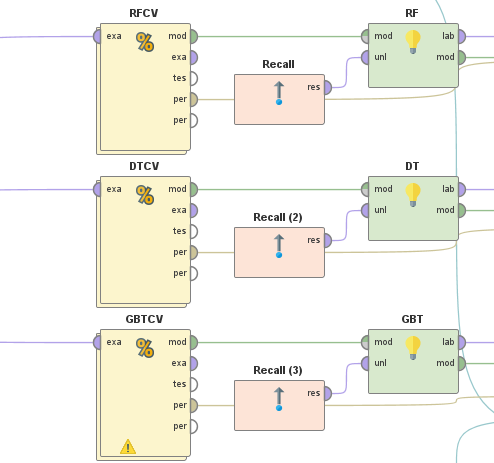
Each of the cross-validation steps looks the same albeit a different model. RFCV is the random forest, DTCV is the decision tree, GBTCV is the gradient boosted tree. Each of these has a training step and testing step. Withing the testing step, there is an apply model and a performance step.



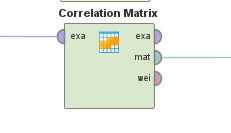
After each model is trained and tested and cross-validated, the trained model is then fed into another “apply model” step where the stored unlabeled production data is used for predictions. The “parent” operator, compare ROC, returns and ROC curve for each model prediction capabilities on the production dataset.

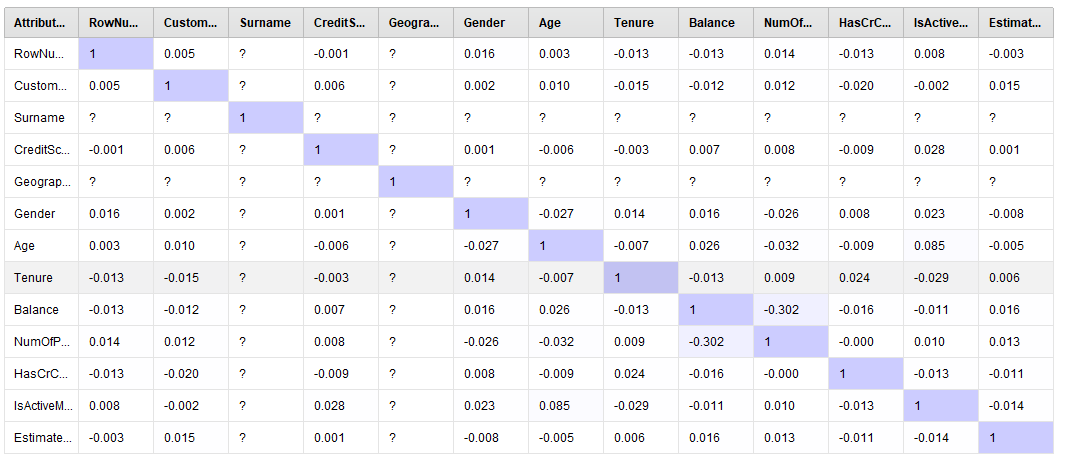


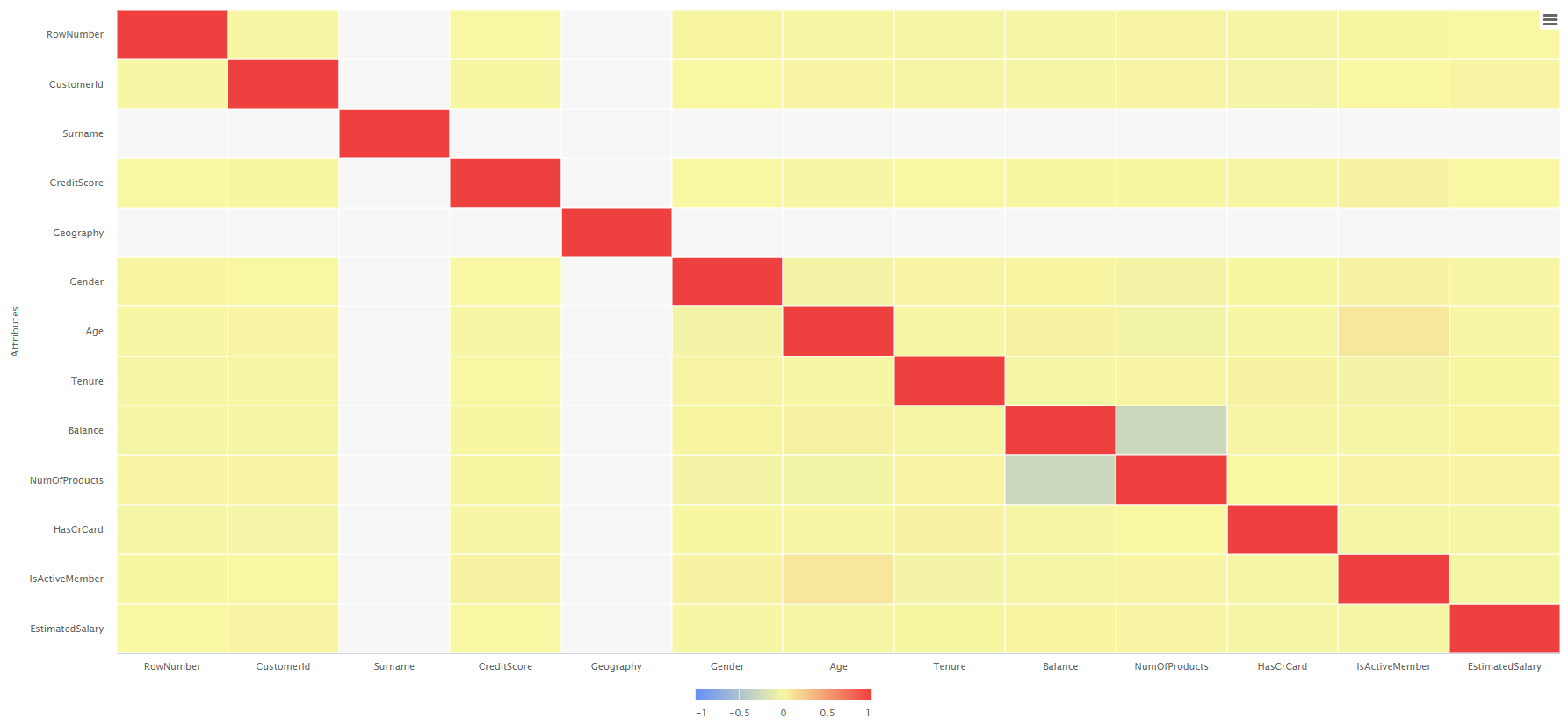
The same process is taken for predicting the label for the production dataset only this time it returns all the predictions, precision, recall, accuracy, and AUC for each model.



The last step is the correlation matrix that returns, none other than, a correlation matrix.







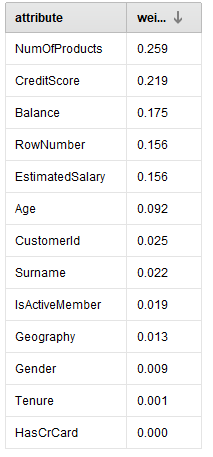
# **Predictions**

Moving forward only the gradient boosted tree model will be discussed. The model produced was accurate and well performing as shown from the metrics below

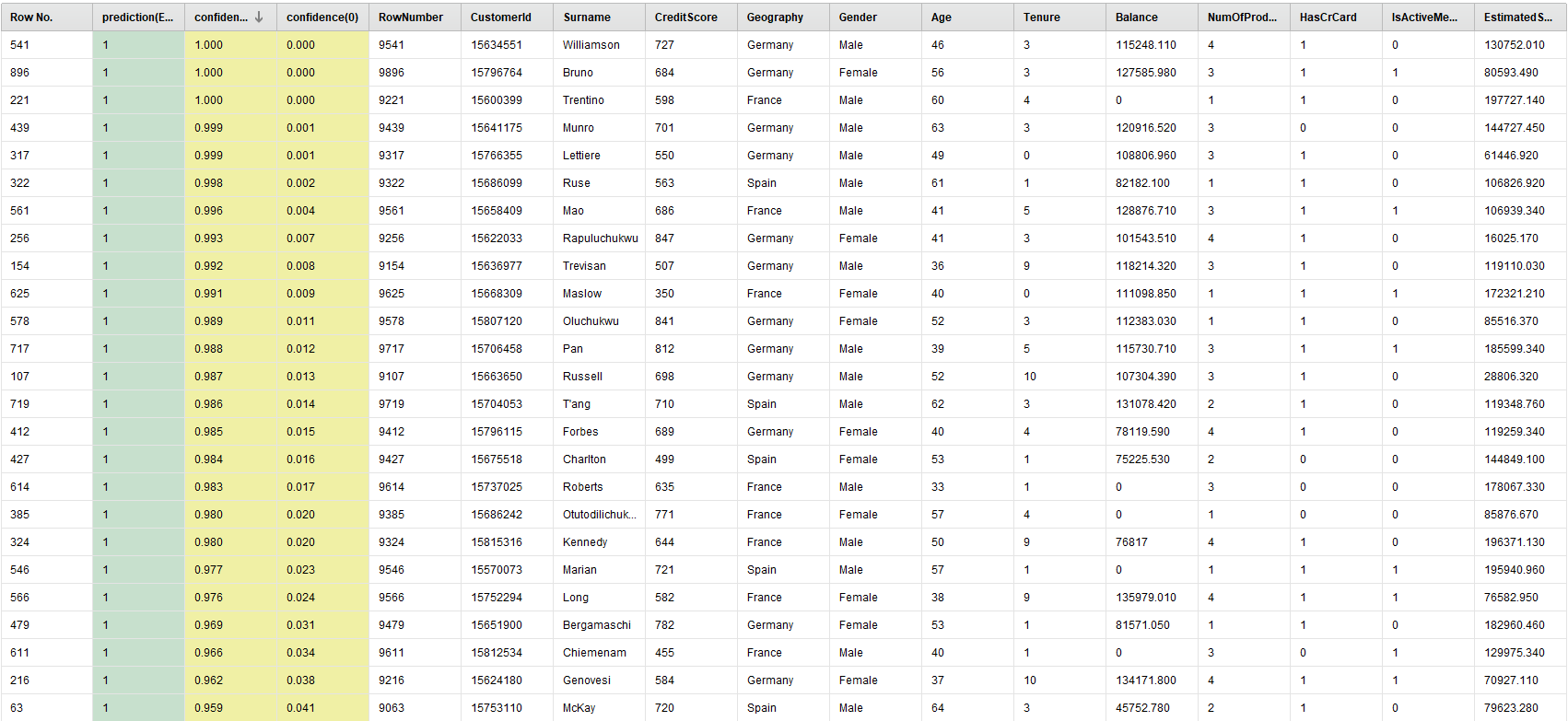
1. Precision – 89.44%
2. Recall – 88.91%
3. Accuracy – 82.86%
4. AUC – 0.836



The gradient boosted tree assigned “NumOfProducts”, “CreditScore”, and “Balance” as the three most important variables.



The following is the top 25 predictions for churn based on the model’s confidence that the customer will churn.



Based on the above information the bank should “profile” people first based on how many products they have, second by how high their credit score is, and thirdly by how high each person’s balance is.