**Random Forest**:

Random Forest is another technique for churn analysis to reduce the variance. We determined the number of trees is 200 and the number of mtry is 2 when it has the smallest OOB error rate.

Text

Description automatically generated

The following figure shows the results of Random Forest model for the training data. The error rate of predicting No is 8.22%. The error rate of predicting Yes is 53.17%.

Text

Description automatically generated

The following plot shows how importance of each variable in the random forest model. The higher the value of Mean Decrease Accuracy or Mean Decrease Gini, the more important the variable is for the successful classification. Clearly, we can see that *Contract, tenure*, and *InternetService* also play a critical role in predicting the customer churn in the RF model. We also have some other important variables like TotalCharges, MonthlyCharhes, PeymentMethods, etc.

Table

Description automatically generated

**Evaluation of RF model (Validation data)**

The random forest model below is fairly accurate in predicting the churning or non-churning customers. The sensitivity is about 92.84% and the specificity is around 47.86% in the test data. It has accuracy of around (1439 + 268)/2110 \*100% = 80.9% in the validation subset. The prediction accuracy for random forest model is slightly larger than the decision tree model (80.9% > 80.52%).

A picture containing text, receipt

Description automatically generated

After analyzing all three model, we can compare all three models:

In logistic regression model: Sensitivity (91.35%), specificity (55.71%) and the prediction accuracy (81.90%).

In decision tree model: Sensitivity (91.23%), specificity (50.89%) and the prediction accuracy (80.52%).

In random forest: Sensitivity (92.84%), specificity (47.86%) and the prediction accuracy (80.9%).

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Sensitivity | Specificity | Prediction acccuracy |
| Logistic | 91.35% | 55.71% | 81.90% |
| Decision Tree | 91.23% | 50.89% | 80.52% |
| Random Forest | 92.84% | 47.86% | 80.9% |

Apparently, the logistic regression model outperforms the decision tree and random forest model in prediction accuracy and specificity. Although we the sensitivity rate is slight smaller than random forest, the logistic regression is still better than other two models on overall performance.

In the ROC plot below, we can see the dashed line in the diagonal to show the ROC curve of a random predictor. The random predictor is used to show whether a model is useful or not. Both curves are good since they lie above the dashed line. Logistic regression model is superior to decision tree model because at all cut-offs the true positive rate is slightly higher, and the false positive rate is slightly lower than for decision tree and random forest model. The AUC (area under the curve) for logistic regression model (85.6%) is slightly larger than the AUC for decision tree (81.3%) and random forest model (84.9%). The higher the AUC, the better the model is at correctly predicting churning or non-churning customers. The closer a ROC curve is to the upper left corner, the more efficient is the model. Hence, the logistic regression model performs better in churn analysis.

Diagram

Description automatically generated

After analyzing the decision tree model, we can compare all three models: we can see that **sensitivity** (**91.35%**) and **specificity** (**55.71%**) in logistic regression model are both relatively larger than sensitivity (91.23%) and specificity (50.89%) in decision tree model. The **prediction** **accuracy** (**81.90%**) of the logistic regression slightly exceeds the decision tree model (80.52%). Apparently, the logistic regression model outperforms the decision tree model with better values of accuracy, sensitivity, and specificity.