

HW3

```
library(microbenchmark)
library(mlbench)
library(gbm)
library(xgboost)
library(ranger)
library(pROC)
library(ROCR)
library(randomForest)
library(rpart.plot)
library(rpart)
library(tidymodels)
library(dplyr)
library(caret)
library(randomForest)
```

About the data set

The data set contains 12 variables with 8250 observations, Also it has customer transaction and demographic related data, It holds risky and not risky customer for specific banking products, and a financial institution wants to understand its customers' payment behavior and predict their future payment performance. The "Payment_data.csv" dataset contains various customer attributes and their payment history.

Detail your task with the problem, features, and target.

```
customer = read.csv("~/customer_data.csv")
str(customer)
```

```
'data.frame':  1125 obs. of  13 variables:
 $ label : int  1 0 0 1 0 0 1 1 0 0 ...
 $ id    : int  54982665 59004779 58990862 58995168 54987320 59005995 59001917 54984789 5898...
 $ fea_1 : int  5 4 7 7 7 6 4 5 5 4 ...
 $ fea_2 : num  1246 1277 1298 1336 NA ...
 $ fea_3 : int  3 1 1 1 2 3 3 3 3 2 ...
 $ fea_4 : num  77000 113000 110000 151000 59000 56000 35000 78000 218000 35000 ...
 $ fea_5 : int  2 2 2 2 2 2 2 2 2 2 ...
 $ fea_6 : int  15 8 11 11 11 6 8 15 15 8 ...
 $ fea_7 : int  5 -1 -1 5 5 -1 9 -1 5 5 ...
 $ fea_8 : int  109 100 101 110 108 100 85 111 112 101 ...
 $ fea_9 : int  5 3 5 3 4 3 5 3 4 3 ...
 $ fea_10: int  151300 341759 72001 60084 450081 60091 60069 60030 151300 60029 ...
 $ fea_11: num  245 207 1 1 197 ...
```

Problem:

Credit Risk Production, some people dont pay their debt, so we cant give the credit those people. ## Task Task: Build a machine learning model on this dataset to predict customer payment behavior.

Features

The dataset containg 1125 observation with 13 variables, but features unnamed, so impossible to explain features. But label 0 means that the customer has low risk credit level, so we can give credit Label 1 means that customer has high risk credit level, it is risky to give credit.

```
# The data set has tons of missing value, and useless variables we need to get rid of them
customer = na.omit(customer) # It deletes missing values.
customer2 = subset(customer, select = -c(id)) # It deletes useless variable which is id
```

Train a logistic regression model, a decision tree, and a random forest model.

```
# We need to split the data set to train (80%) with test (20%)
set.seed(123) # Set the seed value for randomness control
train_index = createDataPartition(customer2$label, p = 0.8, list = FALSE) # Creating index
train = customer2[train_index, ] # Train dataset
test = customer2[-train_index, ] # Test dataset
```

Logistic regression model

```
glmmodel <- glm(label ~ ., data = train, family = "binomial")

glmmodel
```

Call: `glm(formula = label ~ ., family = "binomial", data = train)`

Coefficients:

| | | | | | |
|-------------|-----------|------------|------------|------------|------------|
| (Intercept) | fea_1 | fea_2 | fea_3 | fea_4 | fea_5 |
| -1.011e+00 | 1.680e-01 | -1.011e-04 | 1.631e-01 | -6.098e-06 | 3.397e-02 |
| | fea_6 | fea_7 | fea_8 | fea_9 | fea_10 |
| | 5.251e-04 | -2.390e-02 | -1.091e-02 | 2.515e-02 | 6.546e-07 |
| | | | | | fea_11 |
| | | | | | -1.474e-04 |

Degrees of Freedom: 780 Total (i.e. Null); 769 Residual

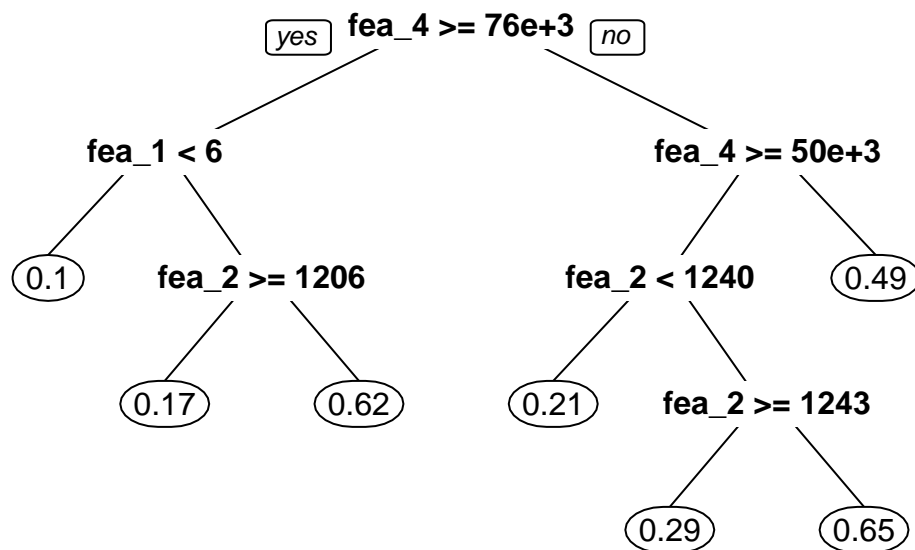
Null Deviance: 764.1

Residual Deviance: 733.5 AIC: 757.5

- In the “Coefficients” section, the estimated coefficient values for each feature (fea_1, fea_2, ..., fea_11) are listed. For example, the estimated coefficient for fea_1 is 0.1680. These coefficients represent the impact of the corresponding feature on the target variable. A positive coefficient indicates that the feature contributes to increasing the value of the target variable, while a negative coefficient indicates a decreasing effect.
- In the “Degrees of Freedom” section, it is stated that there are a total of 780 degrees of freedom and 769 residual degrees of freedom. These degrees of freedom are determined based on the total number of data points and the number of features used in the model.
- The “Null Deviance” value is given as 764.1. This represents the deviation when the model tries to explain the target variable using only the intercept term.

- The “Residual Deviance” value is provided as 733.5. This represents the deviation when all the features are included in the model. A lower “Residual Deviance” value indicates that the model better explains the data.
- Finally, the “AIC” value (Akaike Information Criterion) is given as 757.5. AIC is an information criterion that balances the model’s fit and complexity. A lower AIC value indicates a better fit of the model. ### Decision Tree

```
decmodel = rpart(label ~ ., data = train)
prp(decmodel)
```



Random Forest

```
train$label <- as.factor(train$label)
rfmodel = randomForest(label ~ ., data = train, type = "classification")
predict = predict(rfmodel, train)
rfmodel
```

Call:

```
randomForest(formula = label ~ ., data = train, type = "classification")
```

```
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 3
```

```
OOB estimate of error rate: 19.46%
Confusion matrix:
  0  1 class.error
0 615 16  0.02535658
1 136 14  0.90666667
```

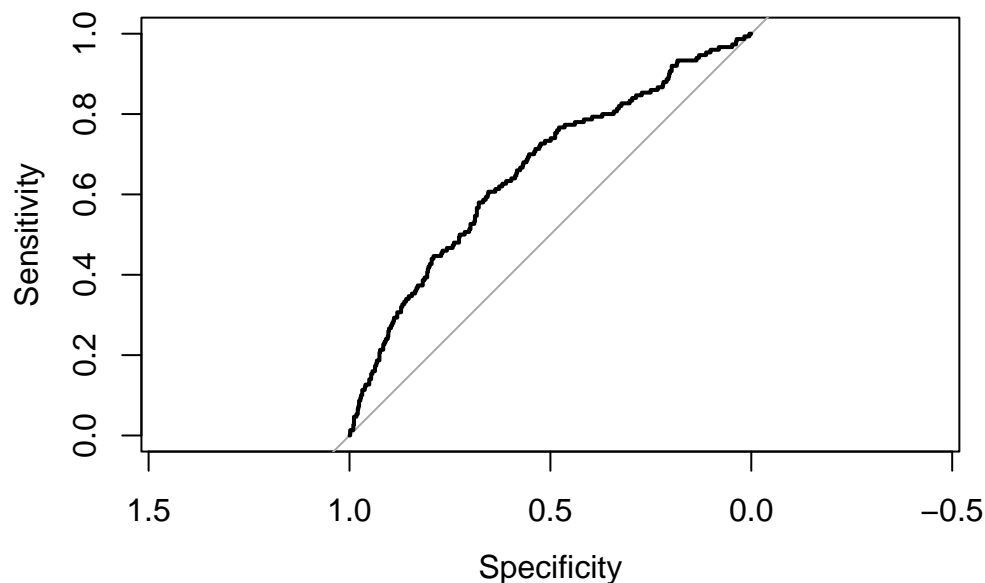
- Number of trees: The model consists of 500 trees.
- Number of variables tried at each split: 3 random variables are tested at each node split.
- The OOB estimate of the error rate for the model is 19.33%. The OOB error rate is a prediction made using the data samples that were not used during the training process and helps evaluate the overall performance of the model.
- For class 0, it can be seen that 618 examples are correctly classified, but 13 examples are misclassified (class errors). In this case, the class error rate for class 0 can be calculated as $13 \text{ errors} / 631 \text{ total examples} = 0.02060222$, resulting in an approximate class error rate of 2.06%.
- For class 1, it can be seen that 138 examples are correctly classified, but 12 examples are misclassified (class errors). In this case, the class error rate for class 1 can be calculated as $12 \text{ errors} / 150 \text{ total examples} = 0.08$, resulting in an approximate class error rate of 8%.

The ROC

I prefer the ROC, and AUC metrics to compare performance which is strong metrics to understand the performance, Also we already calculate AIC and confusion matrix performances, so there is no problem to compare metrics.

```
pred = predict(glmmodel, newdata = train)
roc = roc(response = train$label, predictor = pred)
auc = auc(roc)

plot(roc)
```



auc

Area under the curve: 0.6584

- Area under the curve is 0.6584.
- This indicates that the model has an average performance. Being Close to the upper left corner of the curve indicates that the model has a middle accuracy rate. However, The curve only shows a significant increase after the probability threshold of 0.5, indicating that the model's classification performance is not particularly high.

Compare the performance of the models with a criterion which is independent from the cutoff value.

Training bagging trees

```
set.seed(123)
trained_bt <- ranger(label ~ .,
                     data = train,
                     mtry = 11)
```

```
trained_bt
```

Ranger result

Call:

```
ranger(label ~ ., data = train, mtry = 11)
```

| | |
|----------------------------------|----------------|
| Type: | Classification |
| Number of trees: | 500 |
| Sample size: | 781 |
| Number of independent variables: | 11 |
| Mtry: | 11 |
| Target node size: | 1 |
| Variable importance mode: | none |
| Splitrule: | gini |
| OOB prediction error: | 19.21 % |

Training random forests

```
set.seed(123)
trained_rf <- ranger(label ~ .,
                     data = train)
```

```
trained_rf
```

Ranger result

Call:

```
ranger(label ~ ., data = train)
```

| | |
|----------------------------------|----------------|
| Type: | Classification |
| Number of trees: | 500 |
| Sample size: | 781 |
| Number of independent variables: | 11 |
| Mtry: | 3 |
| Target node size: | 1 |
| Variable importance mode: | none |
| Splitrule: | gini |
| OOB prediction error: | 19.08 % |

Bt's confusion matrix

```
test$label = as.factor(test$label)
preds_bt <- predict(trained_bt, test)

confusionMatrix(preds_bt$predictions,
                 test$label)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 151 | 36 |
| 1 | 6 | 2 |

Accuracy : 0.7846
95% CI : (0.7202, 0.8401)
No Information Rate : 0.8051
P-Value [Acc > NIR] : 0.7939

Kappa : 0.0206

Mcnemar's Test P-Value : 7.648e-06

Sensitivity : 0.96178
Specificity : 0.05263
Pos Pred Value : 0.80749
Neg Pred Value : 0.25000
Prevalence : 0.80513
Detection Rate : 0.77436
Detection Prevalence : 0.95897
Balanced Accuracy : 0.50721

'Positive' Class : 0

Rf's confusion matrix

```
preds_rf <- predict(trained_rf, test)

confusionMatrix(preds_rf$predictions,
                 test$label)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 152 | 37 |
| 1 | 5 | 1 |

Accuracy : 0.7846
95% CI : (0.7202, 0.8401)
No Information Rate : 0.8051
P-Value [Acc > NIR] : 0.7939

Kappa : -0.0081

McNemar's Test P-Value : 1.724e-06

Sensitivity : 0.96815
Specificity : 0.02632
Pos Pred Value : 0.80423
Neg Pred Value : 0.16667
Prevalence : 0.80513
Detection Rate : 0.77949
Detection Prevalence : 0.96923
Balanced Accuracy : 0.49723

'Positive' Class : 0

- It is seen that the test set performance of the bagging trees is better than the random forest. Random forest has worse performance especially in Accuracy.

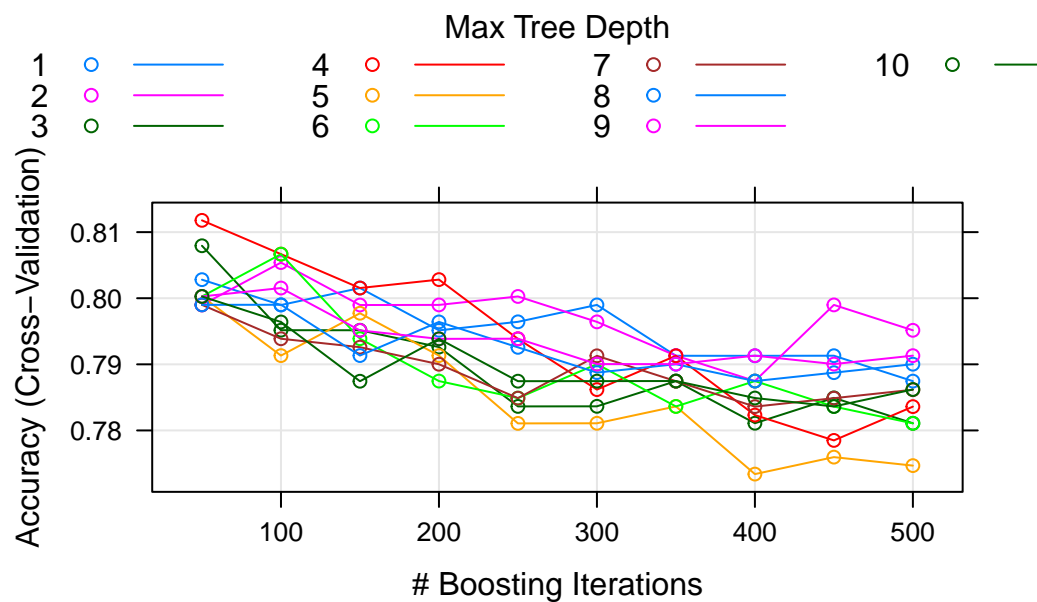
Use the tools that you learned during the courses to increase the model prediction performance.

```
set.seed(123)
gbm_fit = train(label ~ .,
  data = train,
  method = "gbm",
  trControl = trainControl(method = "cv", number = 5),
  verbose = FALSE,
  tuneLength = 10)
```

```
gbm_fit$bestTune
```

```
n.trees interaction.depth shrinkage n.minobsinnode
31      50              4      0.1          10
```

```
plot(gbm_fit)
```



Measuring the GBM performance

```
gbm_preds <- predict(gbm_fit, test)

confusionMatrix(gbm_preds,
                 test$label)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 152 | 37 |
| 1 | 5 | 1 |

Accuracy : 0.7846
95% CI : (0.7202, 0.8401)
No Information Rate : 0.8051
P-Value [Acc > NIR] : 0.7939

Kappa : -0.0081

McNemar's Test P-Value : 1.724e-06

Sensitivity : 0.96815
Specificity : 0.02632
Pos Pred Value : 0.80423
Neg Pred Value : 0.16667
Prevalence : 0.80513
Detection Rate : 0.77949
Detection Prevalence : 0.96923
Balanced Accuracy : 0.49723

'Positive' Class : 0

Training XGB model

```
set.seed(123)
xgbmfit <- train(label ~ .,
  data = train,
  method = "xgbTree",
  trControl = trainControl(method = "cv", number = 5))
```

```
[18:51:59] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:00] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:02] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:02] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:06] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:06] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:07] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:07] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:08] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:08] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:09] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:09] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:09] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:10] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:10] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:11] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:11] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:12] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:14] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:14] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:15] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:15] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:16] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:16] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:17] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:17] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:52:18] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
```

[illegible]

[illegible]

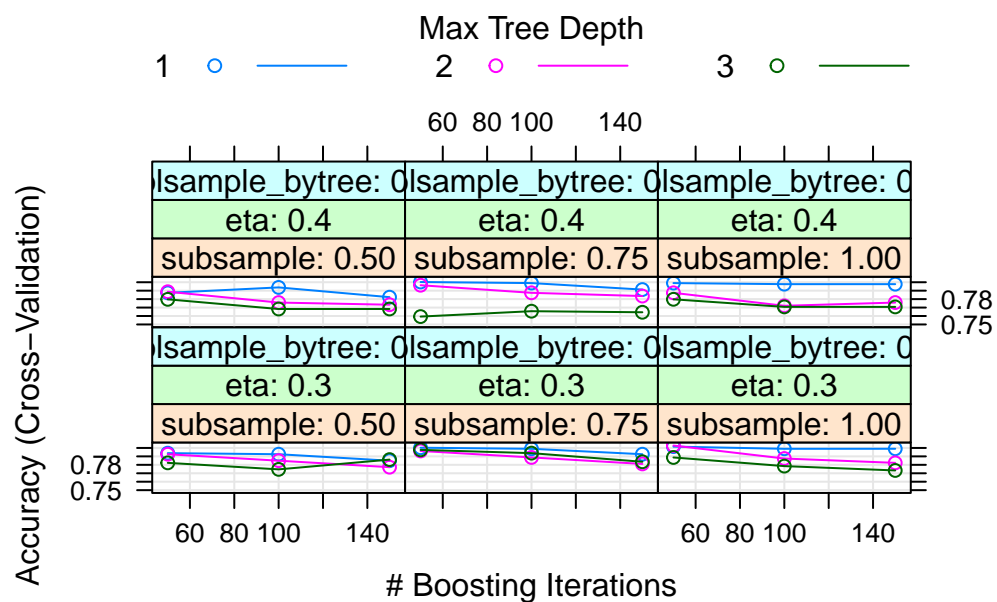
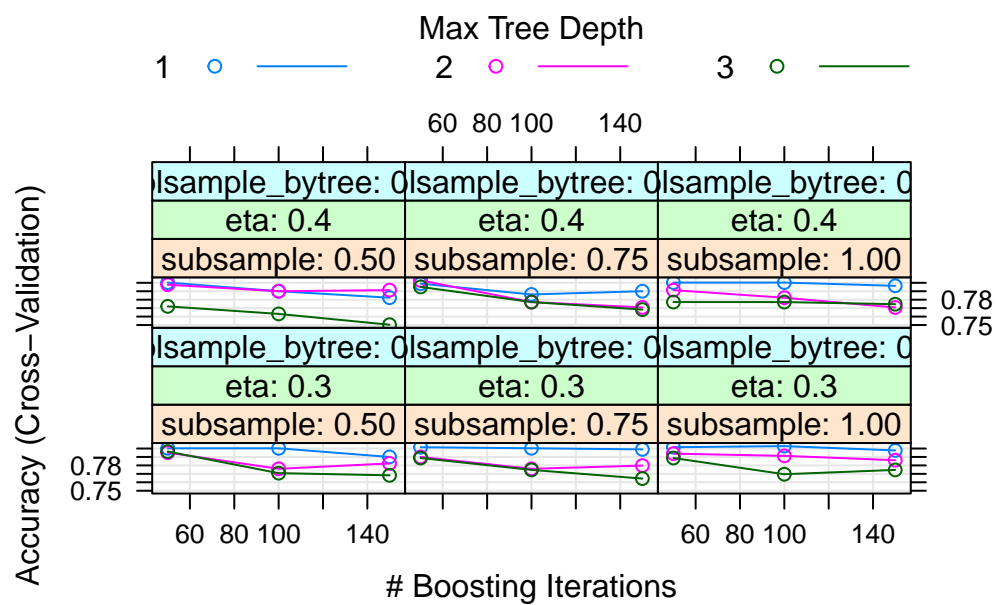
[illegible]

```
[18:55:02] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:02] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:03] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:03] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:03] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:04] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:04] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:04] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:04] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:05] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:06] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:06] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:07] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:07] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:08] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:08] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:09] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:09] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:10] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:10] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:11] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
[18:55:11] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range`
```

```
xgbmfit$bestTune
```

```
  nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
76       50        2 0.4    0                0.6                1      0.75
```

```
plot(xgbmfit)
```



Measuring the performance

```
xgb_preds <- predict(xgbmfit, test)

confusionMatrix(xgb_preds,
                 test$label)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | 0 | 1 |
| 0 | 150 | 38 |
| 1 | 7 | 0 |

Accuracy : 0.7692
95% CI : (0.7037, 0.8264)
No Information Rate : 0.8051
P-Value [Acc > NIR] : 0.9103

Kappa : -0.0645

McNemar's Test P-Value : 7.744e-06

Sensitivity : 0.9554
Specificity : 0.0000
Pos Pred Value : 0.7979
Neg Pred Value : 0.0000
Prevalence : 0.8051
Detection Rate : 0.7692
Detection Prevalence : 0.9641
Balanced Accuracy : 0.4777

'Positive' Class : 0

- Accuracy (Accuracy): It shows the rate of correct classification by the model. In this example, the accuracy value is 0.7692, which means the model has a 76.92% accuracy rate in making correct predictions.
- 95% CI (Confidence Interval): It represents the 95% confidence interval of the accuracy value. In this example, the confidence interval is given as (0.7037, 0.8264).

- No Information Rate: Simply represents the percentage of the most frequent class. In this example, the no information rate value is 0.8051.
- P-Value [Acc > NIR]: It shows the statistical significance between the accuracy and the no information rate. In this example, the p-value is 0.9103, indicating that the accuracy value is not statistically significant compared to the no information rate.
- Kappa: It is the value of the Cohen's Kappa statistic, which shows how much better the model performs compared to random classification. In this example, the kappa value is -0.0645, suggesting that the model performs worse than random guessing.

-Sensitivity (Sensitivity): It shows the rate of correctly detecting true positives (class 1). In this example, the sensitivity value is 0.9554, indicating that the model can detect true positives with a 95.54% accuracy.

-Specificity: It shows the rate of correctly detecting true negatives (class 0). In this example, the specificity value is 0.0000, meaning that the model has not correctly detected any true negatives.

-Pos Pred Value (Positive Predictive Value): It represents the probability of positive predictions being true positives. In this example, the positive predictive value is 0.7979, indicating that 79.79% of the samples predicted as positive are actually positive.

- Neg Pred Value (Negative Predictive Value): It represents the probability of negative predictions being true negatives. In this example, the negative predictive value is 0.0000, meaning that the model has not correctly predicted any negative examples.
- Prevalence: It shows the prevalence rate of the true positive class. In this example, the prevalence value is 0.8051. Detection Rate: It shows the rate of detecting true positives. In this example, the detection rate is 0.7692, meaning the model can detect 76.92% of the true positives.
- Detection Prevalence: It shows the rate of samples predicted as positive by the model. In this example, the detection prevalence is 0.9641.
- Balanced Accuracy: It is the average of sensitivity and specificity values. In this example, the balanced accuracy value is 0.4777.
- These metrics are used to evaluate the performance of the model and understand the classification results. For example, a low specificity value indicates that the model is not correctly detecting negatives, while a high sensitivity value indicates that it is correctly detecting true positives.