Modelling

Importing the libraries necessary for the Modelling part of the report.

Initial Data Cleaning

Now we will use the dataframe final_previous_merged. This is the dataframe that was made that consisted of the previous Financial Helath Ratings that we are utilizing to make models for predicting the Financial Health Ratings.

I am naming this dataframe df and making a very simple filtering on NA values to be equal to 0s as mentioned by the client.

Moreover upon further discussion with out group we decided to remove the following columns as well to simplify our dataset

```
#Defining the dataframe as df
df <- read_csv("final_previous_merged.csv")

#Converting NA values to 0
df[is.na(df)] <- 0

#Removing the columns in the dataframe
df<-df %>% select(-c("prev_inf_factor","prev_X.Other.Currency.to.USD","prev_financialDate"))
df
```

Random Forest Model

Now with the initial datacleaning done I am not making a Random Forest model for predicting FHR.

Random Forest models are a collection of decisions trees where each tree is trained on a random subset of data.

Our model then aggregates the predictions from the trees to give us a final prediction.

We are using random forest model as it can understand and predict our scores based on the complex relationships we have present in our fiscal data. Moreover it can handle overfitting as well as different types of data suhc as categorical and numerical as well.

Through the Random Forest we can also compute feature importance which shows which variables are the most influential in predicting our FHR scores.

For our RF(Random Forest) Model we are creating an initial train-test split. This split is done such that 80% of the data is considered for training and the rest of the 20% is considered for testing the model.

```
set.seed(222)
#Creating the partition for the split
split <- createDataPartition(df$FHR, p = 0.8, list = FALSE)

#Creating a 80-20 split for the train and test data
train_data <- df[split, ]
test_data <- df[-split, ]

#The X_train and X_test are the defined predictor variables
#The y_train and y_test are the defined target variables
```

```r
X_train <- train_data[, -which(names(train_data) == "FHR")]
y_train <- train_data$FHR

X_test <- test_data[, -which(names(test_data) == "FHR")]
y_test <- test_data$FHR

#Defining the RF model
rf_model <- randomForest(x = X_train, y = y_train, ntree = 500, importance = TRUE)

#Getting the predicitions for the RF model
y_prediction_rf <- predict(rf_model, newdata = test_data)

#Getting the metrics, in this case we are using MSE,RMSE,RSS,Total RSS, R^2

mse_rf <- (sum((y_prediction_rf - y_test)^2)/length(y_prediction_rf))
rmse_rf <- sqrt(mse_rf)

#Displaying the RMSE of the model
cat("RMSE of the RF Model: ",rmse_rf)
```

```
## RMSE of the RF Model:  9.191136
```

```r
cat("\n")
```

```r
ss_res <- sum((y_test - y_prediction_rf)^2)
ss_tot <- sum((y_test - mean(y_prediction_rf))^2)
r2 <- 1 - ss_res / ss_tot
cat("R^2 value",r2)
```

```
## R^2 value 0.8122503
```

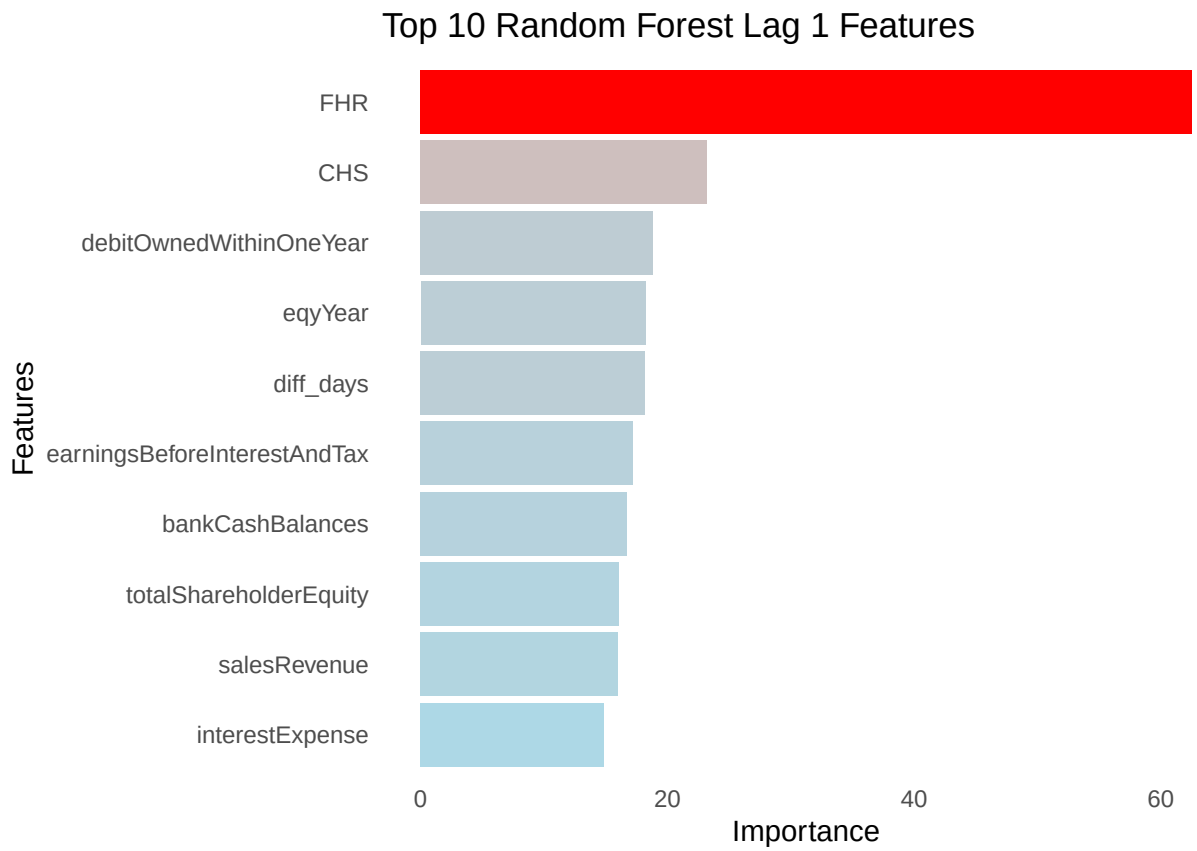
```r
#Getting the variable importance to see which variables are #important for predicting the target variab
imp <- importance(rf_model)

#Sorting the importance in descending order of importance
sorted_importance <- sorted_importance <- imp[order(-imp[, 1]), ]

#Displaying the top 10 variables that are important
sorted_df <- as.data.frame(sorted_importance) %>%
  head(10)  # Get top 10 important features

rownames(sorted_df) <- gsub("prev_", "", rownames(sorted_df))
```

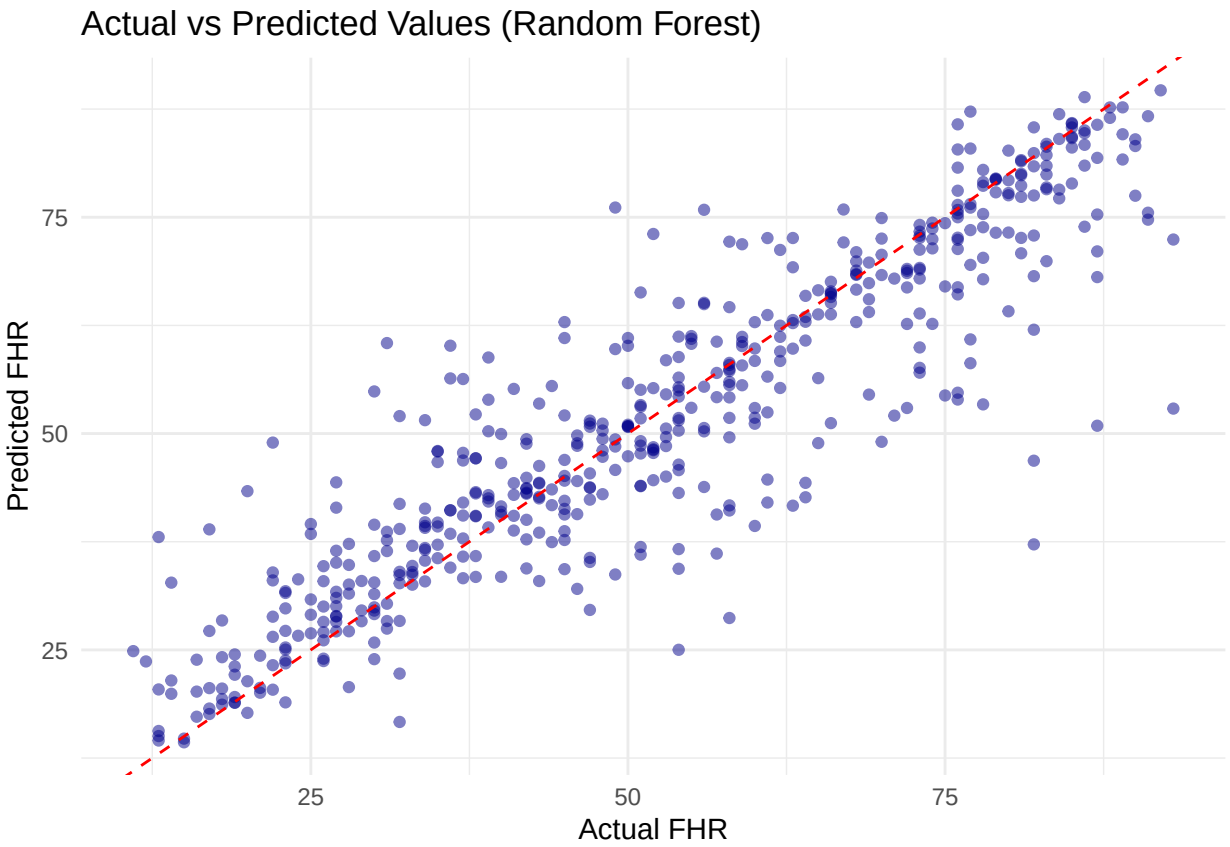
Plot for variables to show importance:



As we can see here the following are of importance in predicting the FHR score: Previous FHR Previous CHS Previous debit Owned within One year

and other fiscal factors..

Now to look at a visual representation of how well fitted the RF Model is we plot an actual versus predicted graph



From above we can see that while there are some outliers the fit described above showcases a model that has pretty good fit in correspondance with the fit line.

Now let us look at a model utilizing Ridge Regression

Ridge Regression Model

Now I am doing a similar process to the one that I did for the RF Model for the Ridge Regression Model

Ridge Regression is a type of linear regression, this is useful to see if there are any interesting linear relationships within the data given.

I am uploading the dataset and doing the usual data cleaning as mentioned in the portion for the RF Model

```
finaldf <- read_csv("final_previous_merged.csv")
finaldf_ridge<-finaldf %>% select(-c('prev_X.Other.Currency.to.USD','prev_inf_factor',
                                     'prev_financialDate'))
```

Now I am creating the train-test split as mentioned above in the RF model and getting the summary statistics to see how well the Ridge Regression Model performed

```
set.seed(222)
#Creating the partition for the split
split <- createDataPartition(finaldf_ridge$FHR, p = 0.8, list = FALSE)

#Creating a 80-20 split for the train and test data
```

```r
train_data <- finaldf_ridge[split, ]
test_data <- finaldf_ridge[-split, ]

#The X_train and X_test are the defined predictor variables
#The y_train and y_test are the defined target variables
#Since for Ridge Regression modelling we make these matrices
X_train <- as.matrix(train_data[, -which(names(train_data) == "FHR")])
y_train <- train_data$FHR
X_test <- as.matrix(test_data[, -which(names(test_data) == "FHR")])
y_test <- test_data$FHR

#Ridge Regression model
CV_ridge <- cv.glmnet(X_train,y_train,alpha = 0)

#Seeing goodness of fit
best_lambda <- CV_ridge$lambda.min

#Getting the predicitions for the Ridge Regression model
y_predicted_ridge <- predict(CV_ridge,newx = X_test, s= "lambda.min")

#Getting the metrics, in this case we are using MSE,RMSE,RSS,Total RSS, R^2
mse_ridge <- (sum((y_predicted_ridge - y_test)^2)/length(y_predicted_ridge))
rmse_ridge <- sqrt(mse_ridge)
cat("Ridge Regression model RMSE: ",rmse_ridge)
```

```
## Ridge Regression model RMSE:  10.33241
```

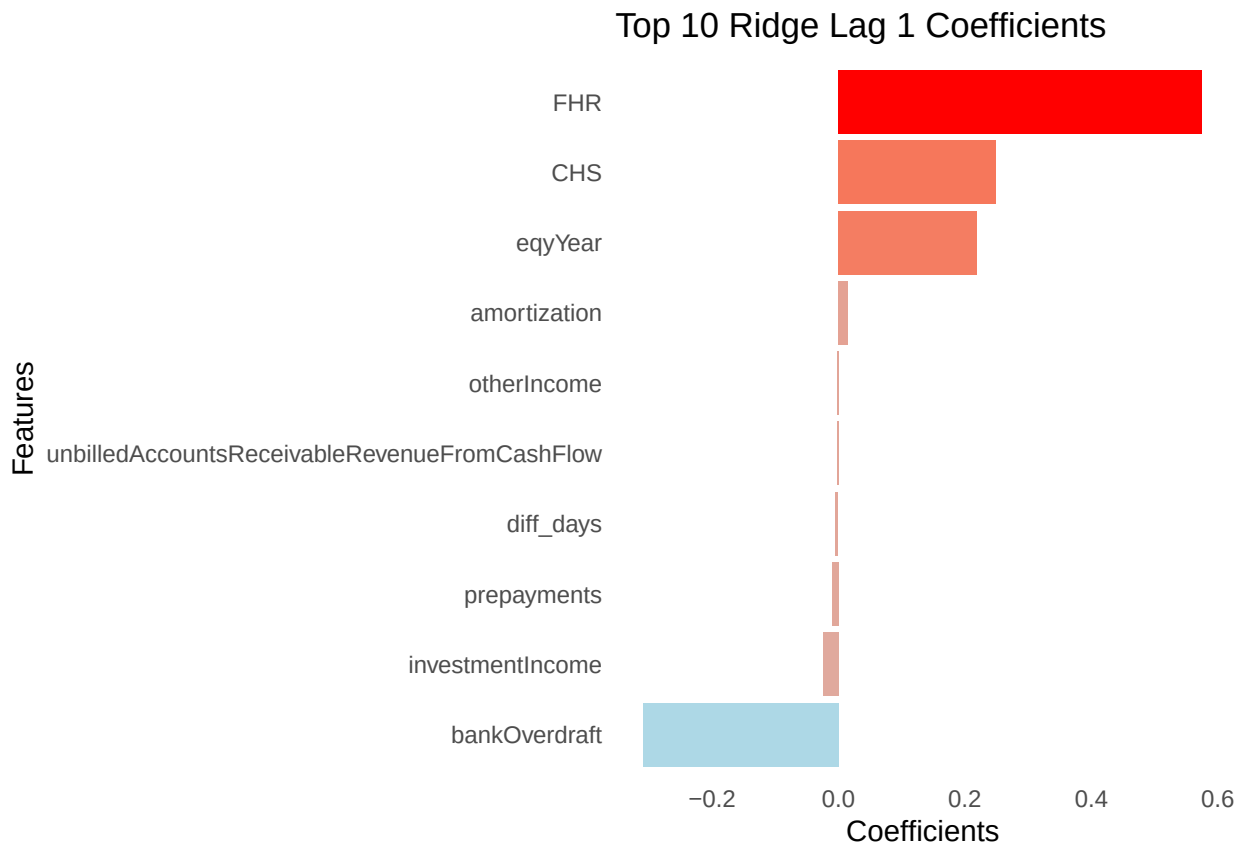
```r
cat("\n")
```

```r
#Extracting coefficients at best lambda
ridge_coef <- coef(CV_ridge, s = "lambda.min")

#Converting to a dataframe
ridge_coef_df <- as.data.frame(as.matrix(ridge_coef))
ridge_coef_df$Feature <- rownames(ridge_coef_df)  # Add feature names
colnames(ridge_coef_df) <- c("Coefficient", "Feature")
ridge_coef_df <- ridge_coef_df[-1, ]
ridge_coef_df$Abs_Coefficient <- abs(ridge_coef_df$Coefficient)
ridge_coef_df <- ridge_coef_df[order(-ridge_coef_df$Abs_Coefficient), ]

ridge_coef_df$Feature <- gsub("prev_", "", ridge_coef_df$Feature)
```

Contains a plot of the coefficients of the variables based on Ridge Regression



We can see the the Previous FHR and the Previous CHS is still some of the most important predictor variables similar to the Random Forest.

But compared to the Random Forest we can see that the RMSE value is higher at approximately 10.33. This means that the model predictions are about 10.33 units from the actual on average.

Lasso Model

Now we will look into modelling for Lasso which is a linear regression which prevents overfitting and also does variable selection. This helps again with seeing which variables can be selected to predict FHR. It is also easier to interpret in comparison to the other models such as Random Forest.

Here again I am redefining the df for the Lasso Model but referencing the same data for all the models. Datacleaning as usual is converting the NAs to 0.

Below I am also looking at which columns have character values to see if there any non-numerical columns as linear models like LASSO do not take non-numerical values for modelling

```
library(glmnet)
finaldf <- read.csv("final_previous_merged.csv")
finaldf[is.na(finaldf)] <- 0
for(col in names(finaldf)) {
  if(any(is.character(finaldf[[col]]))) {
    print(paste("Column", col, "contains character values"))
  }
}
```

Since these columns upon discussion are not necessary to extract useful insights from, we decided to remove the columns instead along with other columns.

```
finaldf_lasso<-finaldf %>% select(-c('prev_X.Other.Currency.to.USD','prev_inf_factor',
                                     'prev_financialDate'))
```

Now I am doing the test-train split for the Lasso Model using the metrics similar to the RF and Ridge Regression (80/20 split on the data for training and testing respectively). Also I am defining the Lasso Model as well.

```
set.seed(222)
#Creating the partition for the split
split <- createDataPartition(finaldf_lasso$FHR, p = 0.8, list = FALSE)

#Creating a 80-20 split for the train and test data
train_data <- finaldf_lasso[split, ]
test_data <- finaldf_lasso[-split, ]

#The X_train and X_test are the defined predictor variables
#The y_train and y_test are the defined target variables
X_train <- as.matrix(train_data[, -which(names(train_data) == "FHR")])
y_train <- train_data$FHR
X_test <- as.matrix(test_data[, -which(names(test_data) == "FHR")])
y_test <- test_data$FHR

#Defining the lasso model
CV_lasso <- cv.glmnet(X_train,y_train,alpha = 1)

#Goodness of fit
best_lambda <- CV_lasso$lambda.min

#Getting the predictions from the model
y_predicted_lasso <- predict(CV_lasso,newx = X_test, s= "lambda.min")

#Getting the metrics, in this case we are using MSE,RMSE,RSS,Total RSS, R^2
mse_lasso <- (sum((y_predicted_lasso - y_test)^2)/length(y_predicted_lasso))
rmse_lasso <- sqrt(mse_lasso)
lasso_coef <- coef(CV_lasso, s = "lambda.min")
ss_res <- sum((y_test - y_predicted_lasso)^2)
ss_tot <- sum((y_test - mean(y_predicted_lasso))^2)
r2 <- 1 - ss_res / ss_tot
cat("The Lasso Model R^2:",r2)
```

```
## The Lasso Model R^2: 0.7930211
```

```
cat("\n")
```

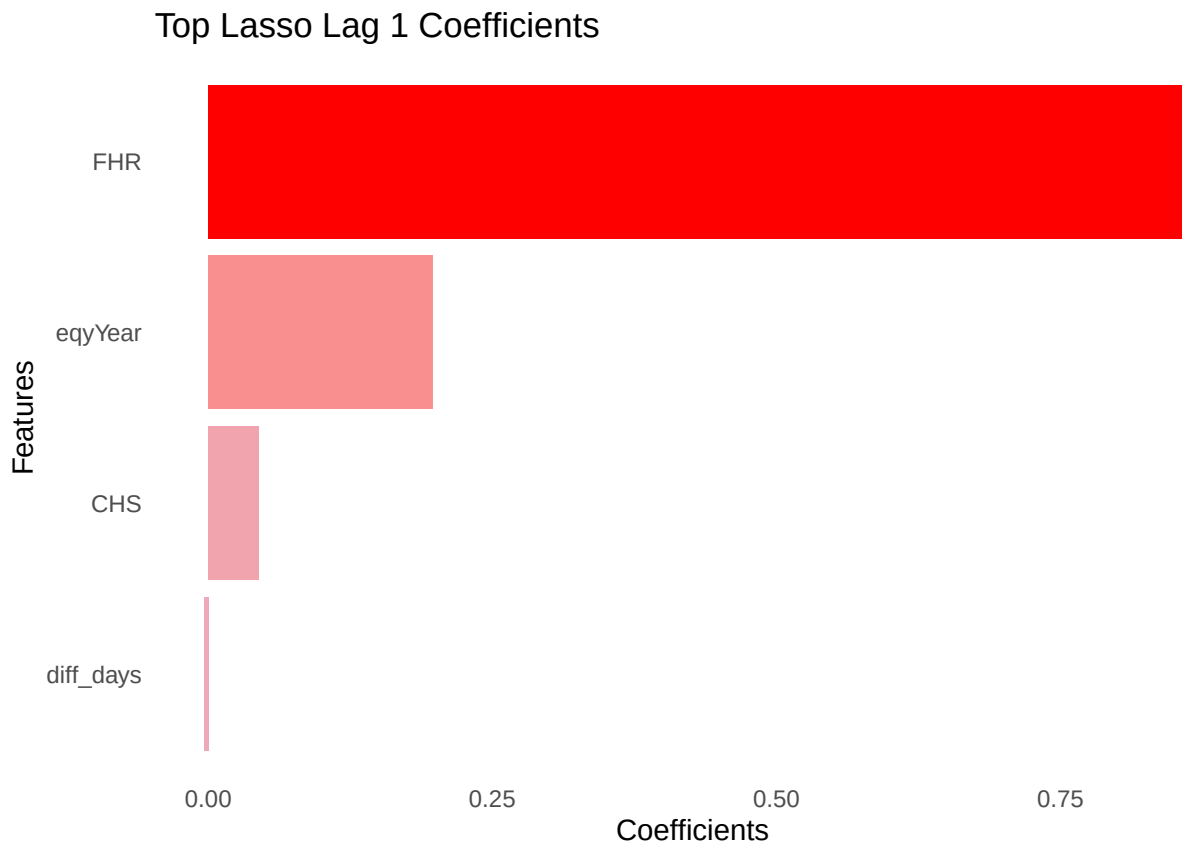
```
cat("The Lasso Model RMSE:",rmse_lasso)
```

```
## The Lasso Model RMSE: 9.652856
```

We can see that the Lasso does better that the Ridge in terms of the RMSE value which is 9.65 but not as good as the Random Forest model. Using these values we can get a sense that the Random Forest has the best RMSE.

Now looking at the importance of variables for which Lasso does variable selection on:

Plot to show the feature coefficients based on the top variables



For additional modelling I decided to explore XGBoost.

XGBOOST

This is gradient boosting model where the method creates an ensemble of decision trees where each new tree it creates it corrects the error made by the previous trees, hence minimizing errors through gradient descent. The decision trees are done sequentially where each new tree reduces the error of the previous ensemble of trees.

Now modelling using XGBoost for predicting the previous FHR

```
set.seed(222)
#Defining the data frame for the XGBoost similar to the other models and doing some data cleaning throu
finaldf <- read.csv("final_previous_merged.csv")

finaldf[is.na(finaldf)] <- 0

finaldf_xgboost<-finaldf %>% select(-c('prev_X.Other.Currency.to.USD','prev_inf_factor','prev_financiall

#Creating a 80-20 split for the train and test data
split_index <- createDataPartition(finaldf_xgboost$FHR, p = 0.8, list = FALSE)
train_data <- finaldf_xgboost[split_index, ]
test_data <- finaldf_xgboost[-split_index, ]
```

```r
#The X_train and X_test are the defined predictor variables
#The y_train and y_test are the defined target variables
X_train <- as.matrix(train_data %>% select(-FHR))
y_train <- train_data$FHR
X_test <- as.matrix(test_data %>% select(-FHR))
y_test <- test_data$FHR

#For processing need to convert it to a XGBoost DMatrix
dtrain <- xgb.DMatrix(data = X_train, label = y_train)
dtest <- xgb.DMatrix(data = X_test, label = y_test)

#Defining the hyperparameters for the XGBoost model
params <- list(
  objective = "reg:squarederror",  #For regression
  eta = 0.1,            #Learning rate
  max_depth = 6,        #Max. depth of trees
  min_child_weight = 1,#Minimum sum of instance weights for a child
  subsample = 0.8,   #Subsample ratio for training instances
  colsample_bytree = 0.8     #Subsample ratio of columns when constructing each tree
)

#Cross validation to find optimal number of boosting rounds
cv_results <- xgb.cv(
  params = params,
  data = dtrain,
  nrounds = 1000,      #Max. # of boosting rounds
  nfold = 5,           #5 fold Cross validation
  early_stopping_rounds = 50, #Early stopping is to indicat to stop if there is no
                              #improvements for 50 rounds
  verbose = 0
)

#Optimale number of rounds based on the cross validation
best_n_rounds <- which.min(cv_results$evaluation_log$test_rmse_mean)

#XGBoost model for optimal number of rounds
xgb_model <- xgb.train(
  params = params,
  data = dtrain,
  nrounds = best_n_rounds,
  watchlist = list(train = dtrain, test = dtest),
  verbose = 0
)

#Predictions from the model
y_pred_xgb <- predict(xgb_model, dtest)

#Calculating the metrics of RMSE,MSE
xbg_mse <- mean((y_pred_xgb - y_test)^2)
xgb_rmse <- sqrt(xbg_mse)

cat(paste("RMSE:", xgb_rmse))
cat("\n")
```

As we can see while this does better that the Linear models,as the data contains complex fiscal relationships that are better captured by trees and tree based-models through either bagging or boosting, we can see that in this case RF Models tend to perform better than the gradient boosting model with XGBoost.

Hence we will use the RF Model for final modelling for the future FHR score as is performed better in terms of metrics in comparison to the other models.

Now we are using the model to predict the future FHR scores.

```
#Now we are using a dataframe that have the previous FHR scores and using this we will get predictions
suppliers <- read_csv("final_future_merged.csv")
#Doing some data cleaning on the dataframe
suppliers[is.na(suppliers)] <- 0
suppliers<-suppliers %>% select(-c("prev_inf_factor","prev_X.Other.Currency.to.USD",
                                   "prev_financialDate"))

#Predicting the model for the data given
pred <- predict(rf_model, newdata = suppliers)

results <- data.frame(supplierID = suppliers$prev_Supplier.Number, predicted_FHR = pred)

#Saving it to a CSV for further analysis
write.csv(results, "predicted_FHR_supplier.csv", row.names = FALSE)
```

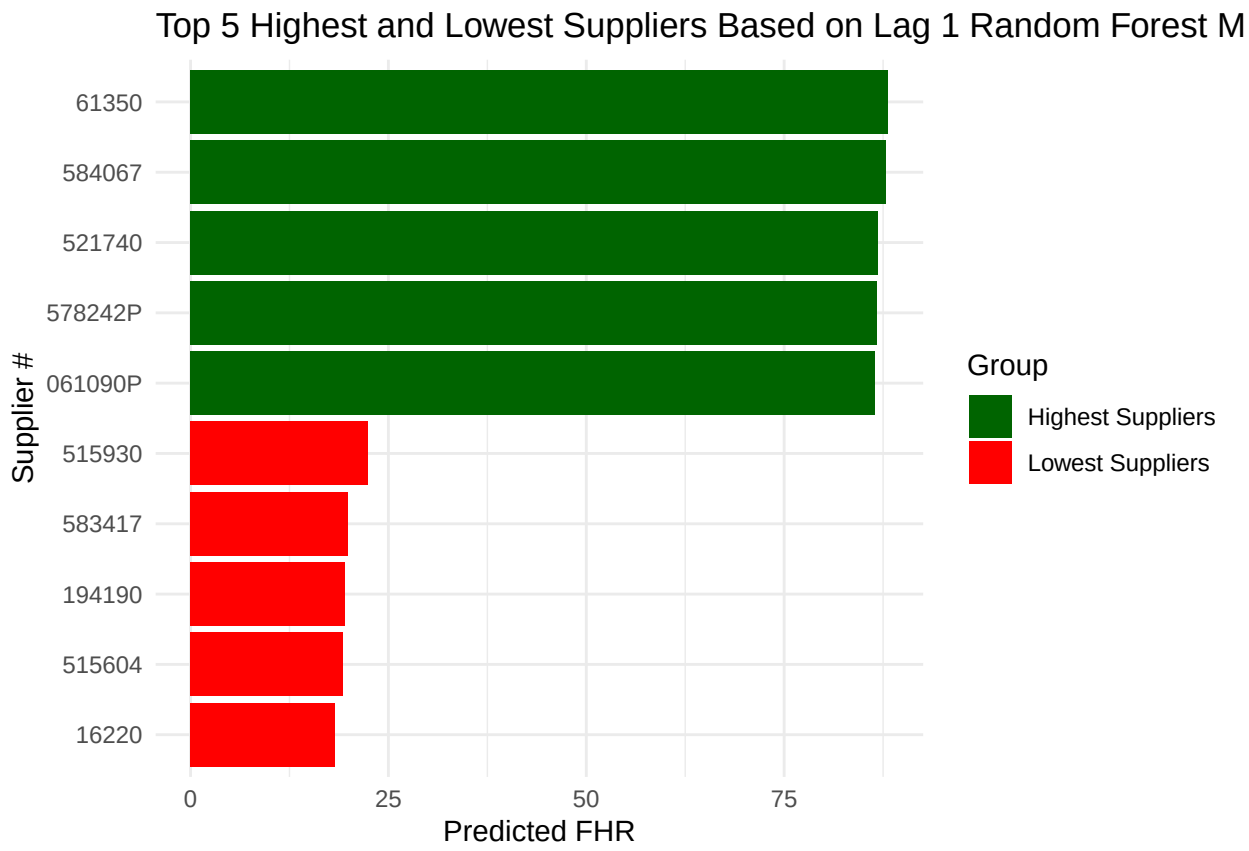
We will now utilize the predictions to do further analysis on the results gained.

```
#Getting the predictions we got from running the model in the cell before
final_predictions<-read_csv("predicted_FHR_supplier.csv")
supplier_info<-read_csv("final_merged_updated.csv")
```

Now we will look at the top 5 and bottom 5 suppliers in terms of the FHR score

Further we are seeing which countries of origin based on the currency the supplier uses.

Plotting the top and bottom currencies to see which countries perform the best or perform poorly



Now I will look into the supplier data and try to find out their country of origin

```
#Analysis on the perspective of currency on the two dataframes
latest_currency_data <- supplier_info %>%
  group_by(Supplier.Number) %>%
  arrange(desc(financialDate)) %>%
  dplyr::slice(1) %>%
  select(Supplier.Number, currency, financialDate)

result <- final_predictions %>%
  left_join(latest_currency_data,
            by = c("supplierID" = "Supplier.Number"))
```

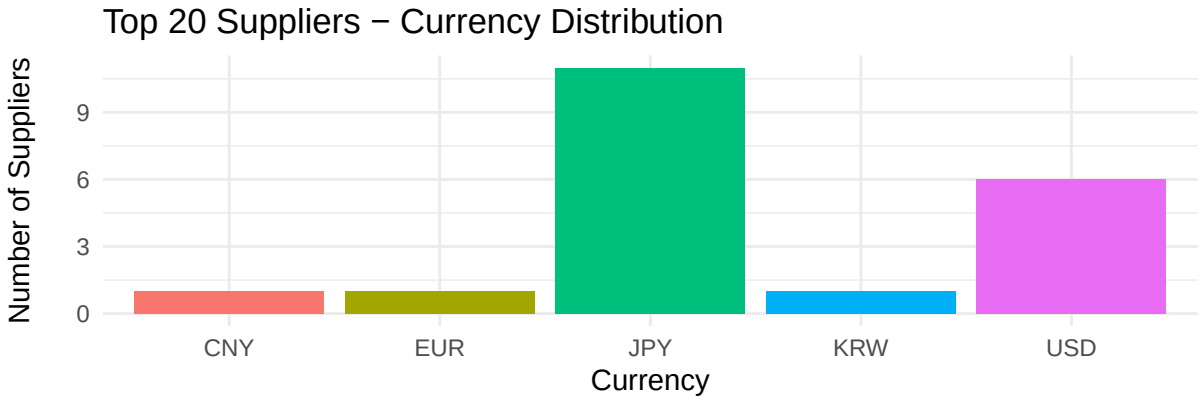
It might be interesting to see the country of origin for the suppliers to see if we see similar trends to what we saw during EDA where Canadian and Japanese suppliers perform better than the rest.

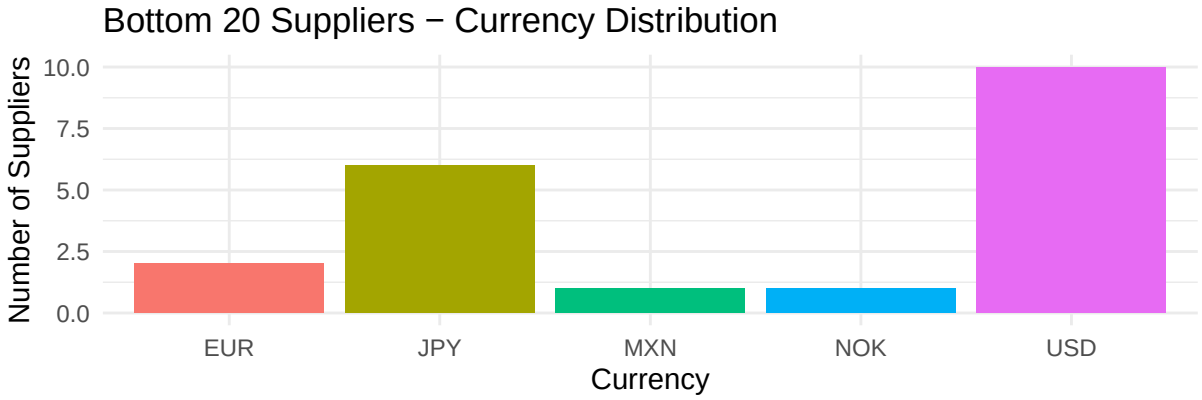
Getting the top 20 and bottom 20 suppliers based on predicted FHR scores

Further we are seeing which countries of origin based on the currency the supplier uses.

Getting their countires of origin for further analysis

Plotting the top and bottom currencies to see which countries perform the best or perform poorly





These charts show an interesting currency distribution pattern.

We can see how the suppliers with countries of origin that have the top number of suppliers show variability in terms of the predicted FHR scores. This can be identified with Japanese and US suppliers being high in number in terms of count.

But it is interesting to note that Japanese suppliers are top performing while US has some under performers.

It will be interesting to consider Japanese based suppliers and analyzing what they offer better in terms of quality for such high performance.

Moreover the performance of Japanese suppliers can also be seen as high even in our EDA process and hence we can say that in general suppliers from Japan could be advantageous to have more trade with.

It is interesting to see that one supplier from Korea is also a high performer along with a supplier from China. This could maybe correlate with the East Asian region having high performers in terms of supplier FHR scores.