Rounak Saha
20/4/2025

1. Task 2: Model Development & Evaluation

1.1 Algorithm Selection

Based on the binary classification nature of churn prediction and the need for interpretability, we recommend using **Random Forest** or **Gradient Boosting Machines**:

- **Random Forest**: Robust to overfitting, interpretable via feature importances, handles nonlinear relationships.

- **Gradient Boosting (e.g., XGBoost / LightGBM)**: Often achieves higher accuracy, can be tuned for class imbalance

1.2 Model Training

```
# Xgboost Model

set.seed(220)

# Split into training and testing sets

split <- sample.split(data$ChurnStatus, SplitRatio = 0.8)

train <- subset(data, split == TRUE)

test <- subset(data, split == FALSE)

# Make column names safe

names(train) <- make.names(names(train))
names(test) <- make.names(names(test))

train$ChurnStatus <- factor(ifelse(train$ChurnStatus == "1", "yes", "no"))
test$ChurnStatus <- factor(ifelse(test$ChurnStatus == "1", "yes", "no"))

# Apply both over- and under-sampling

both_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "under")$data

control <- trainControl(method="cv", number=10, classProbs = TRUE, summaryFunction =
twoClassSummary)
```

 # Prediction with no warning printed in console

xgb_model <- train(ChurnStatus~., data=both_sampled, method="xgbTree", metric="ROC", trControl=control, verbose = 0)
xgb_model_pred <- predict(xgb_model, test, type = "prob")
xgb_model_pred_class <- as.factor(ifelse(xgb_model_pred[, "yes"] > 0.51, "yes", "no"))

confusionMatrix(xgb_model_pred_class, test$ChurnStatus, positive = "yes")

 # Extract raw xgb.Booster from caret model

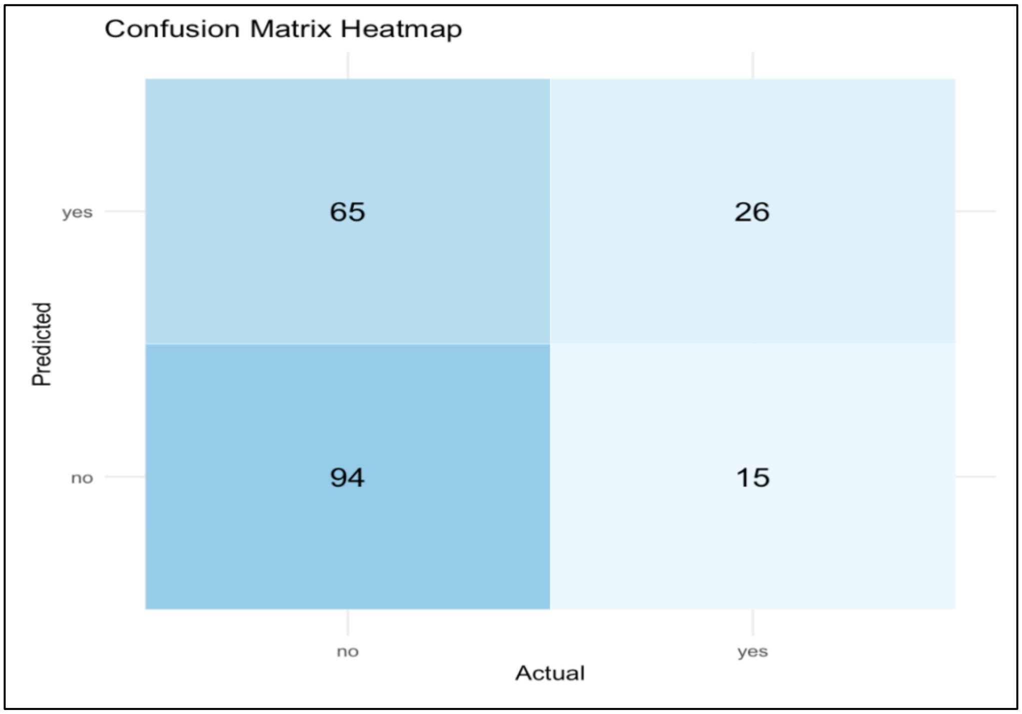
booster_model <- xgb_model$finalModel
feature_importance <- xgb.importance(model = booster_model)
print(feature_importance)
xgb.plot.importance(feature_importance, xlab = "Relative importance (Gain)",col= "maroon", top_n = 15)

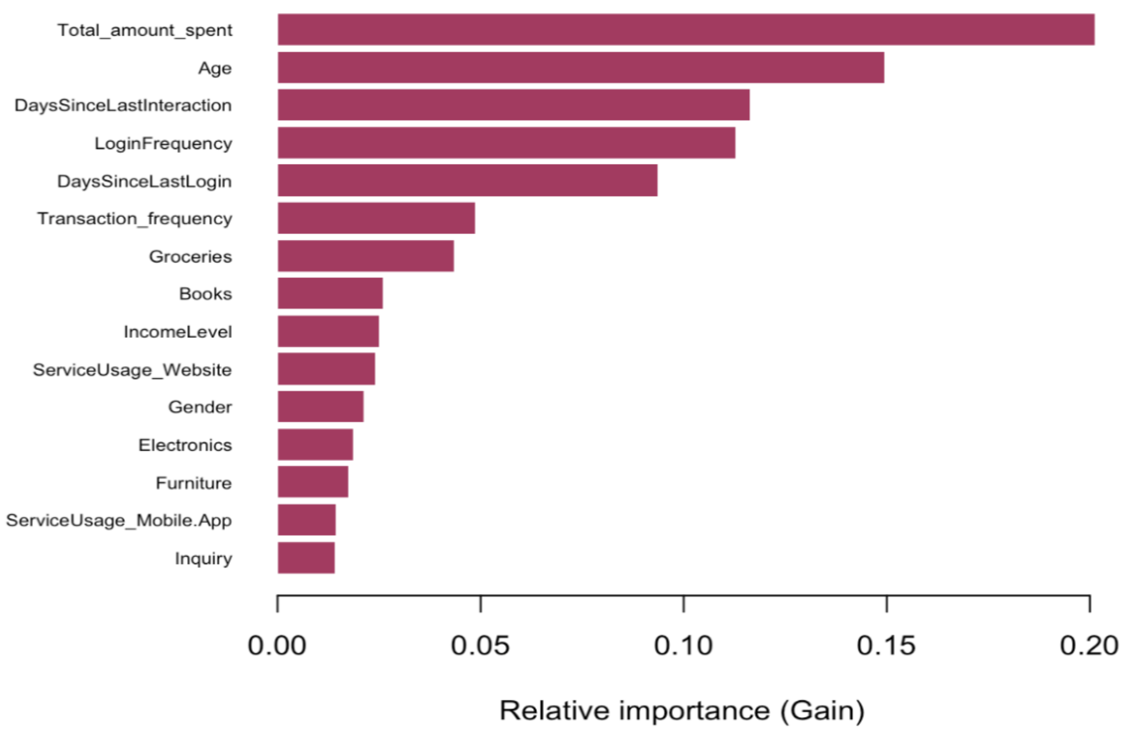
1.3 Performance Metrics & Feature Importance

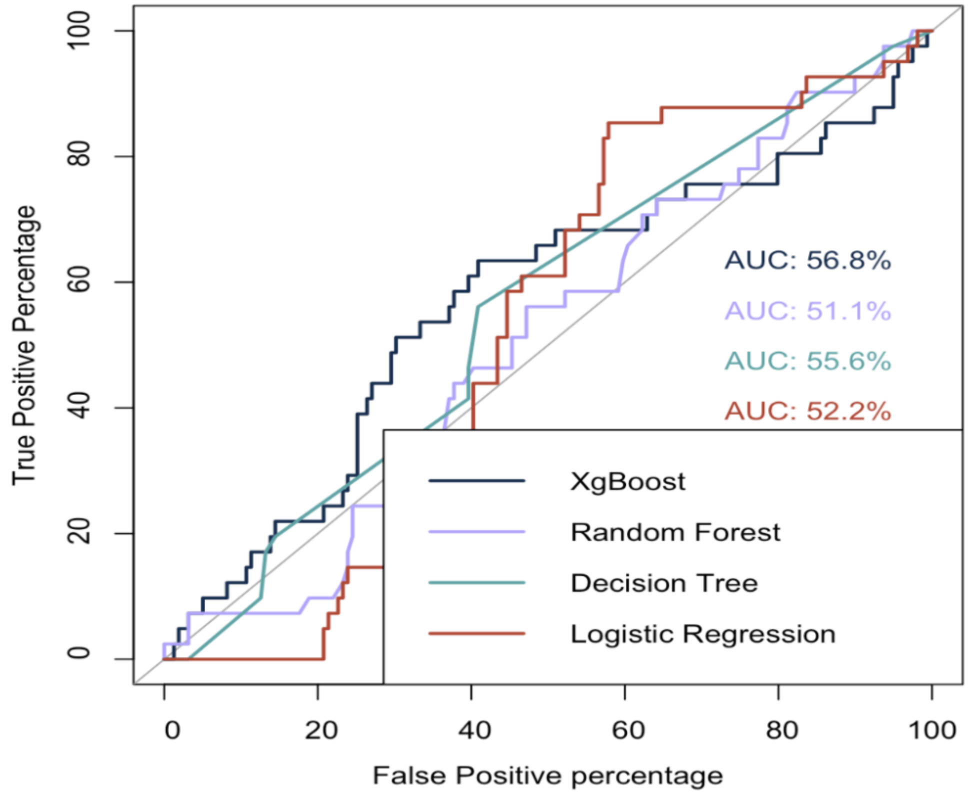
- **ROC-AUC**: Measures discrimination ability across thresholds.
- **Sensitivity (Recall / True Positive Rate)**: Focus on identifying churners (positive class) accurately.
- **Specificity (True Negative Rate)**: Specificity (also called the True Negative Rate) is the proportion of actual negative cases that are correctly identified by the model.

- **Confusion Matrix**: Visualise true positives, false positives, etc.

```
Confusion Matrix and Statistics

          Reference
Prediction no yes
       no  94  15
       yes 65  26

               Accuracy : 0.6
                 95% CI : (0.5285, 0.6685)
    No Information Rate : 0.795
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1551

 Mcnemar's Test P-Value : 4.293e-08

            Sensitivity : 0.6341
            Specificity : 0.5912
         Pos Pred Value : 0.2857
         Neg Pred Value : 0.8624
             Prevalence : 0.2050
         Detection Rate : 0.1300
   Detection Prevalence : 0.4550
      Balanced Accuracy : 0.6127

       'Positive' Class : yes
```


Confusion Matrix Heatmap
Predicted
yes
no
65
26
94
15
no
yes
Actual


Total_amount_spent
Age
DaysSinceLastInteraction
LoginFrequency
DaysSinceLastLogin
Transaction_frequency
Groceries
Books
IncomeLevel
ServiceUsage_Website
Gender
Electronics
Furniture
ServiceUsage_Mobile.App
Inquiry
0.00
0.05
0.10
0.15
0.20
Relative importance (Gain)



1.4. Recommendations for Business Application

- **Real-time Scoring**: Integrate the model into CRM to score customers on churn risk at each login.
- **Targeted Retention Campaigns**: Use risk scores to trigger personalized offers or outreach for high-risk segments.
- **Feature Monitoring**: Continuously monitor key drivers (e.g., login frequency drops) and recalibrate model as behaviors evolve.

1.5 Future Improvements

- **Ensemble Approaches**: Combine multiple algorithms (e.g., RF + GBM) for robust predictions.
- **Additional Data Sources**: Incorporate customer support interactions, sentiment analysis, or external macroeconomic indicators.
- **Explainability Tools**: Use SHAP or LIME to provide transparency into individual predictions for stakeholders.