SESSION 13: Decision Tree Based Models Assignment 2

5. Problem Statement

1. Use the given link below:

https://archive.ics.uci.edu/ml/machine-learning-databases/00304/

Problem- prediction of the number of comments in the upcoming 24 hours on those blogs, the train data was generated from different base times that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, the you should use the provided, temporally disjoint train and test splits to ensure that the evaluation is fair.

a. Create a linear regression model to predict the number of comments in the next 24 hours (relative to base time).

b. Fine tune the model and represent important features Visualize the dataset and make inferences from that.

c. Interpret the summary of the linear model.

d. Report the test accuracy vs. the training accuracy

```
# A. Create a linear regression model to predict the number of comments in the next 24 hours
# (relative to basetime)

TARGET <- lm(target~., data = train)
# step <- stepAIC(TARGET, direction = "both")

final_model <- lm(target ~ checkin + talking + d5 + d6 + d7 + d8 + d9 + d10 + d11 +
d12 + d13 + d16 + d17 + d19 + d20 + d21 + d22 + d23 + d24 +
```

```
cc1 + cc2 + cc3 + cc4 + basetime + postshre + Hhrs + wed +
thu + fri + basemon + basewed, data = train)
summary(final_model)

# B. Fine tune the model and represent important features

final_model <- lm(target ~ talking + d5 + d7 + d8 + d10 + d11 +
d12 + d13 + d16 + d17 + d19 + d20 + d22 + d23 +
cc1 + cc2 + cc3 + cc4 + basetime + postshre + Hhrs, data = train)
summary(final_model)


prediction <- predict(final_model, test)
predicted <- data.frame(cbind(actuals = test$target, prediction = prediction))
predicted$prediction <- ifelse(prediction<0, 0, round(prediction,0))
cor(predicted)
View(predicted)

# C. Interpret the summary of the linear model

# Residual error is distributed between -346.83 to 1271.33
# P-value of the model is less than alpha (0.05), hence we can accept the model
# 32.46% variability is represented by the model
#-------------------------------------------------------------------------------
# D- report the test accuracy vs. the training accuracy

# test accuracy
round(accuracy(predicted$prediction,predicted$actuals),3)

prediction <- predict(final_model, test)
predicted <- data.frame(cbind(actuals = test$target, prediction = prediction))
predicted$prediction <- ifelse(prediction<0, 0, round(prediction,0))

min_max_accuracy <- mean(apply(predicted, 1, min) / apply(predicted, 1, max))
min_max_accuracy

# training accuracy
```

```
round(accuracy(predicted$prediction,predicted$actuals),3)

prediction <- predict(final_model, train)
predicted <- data.frame(cbind(actuals = train$target, prediction = prediction))
predicted$prediction <- ifelse(prediction<0, 0, round(prediction, 0))
min_max_accuracy <- mean(apply(predicted, 1, min) / apply(predicted, 1, max))
min_max_accuracy
```