STATML - Assignment 2

Ross Jackson

2023-04-03

Packages:

```
library(rpart)
library(rpart.plot)
library(ROCR)
```

Fit a logistic regression model and a classification tree using the data in data.

```
data_speed <- read.csv("data_speed_dating.csv")
n <- nrow(data_speed)
set.seed(2023)
set <- sample(1:n, 2290)
data_test <- data_speed[set,]    # for assessing the predictive performance
data <- data_speed[-set,]

#changing match column to binary values 0 and 1
data$match <- ifelse(data$match == "yes", 1, 0)

#LOGISTIC REGRESSION
#Fit log reg model
fit <- glm(match ~ ., data = data, family = "binomial")
#Predict probabilities for the test data
p_test <- predict(fit, newdata = data_test, type = "response")

#CLASSIFICATION TREE
#Fit classification tree model
tree_fit <- rpart(match ~ ., data = data, method = "class")
#Predict probabilities for the test data
tree_p_test <- predict(tree_fit, newdata = data_test, type = "prob")[, 2]
```

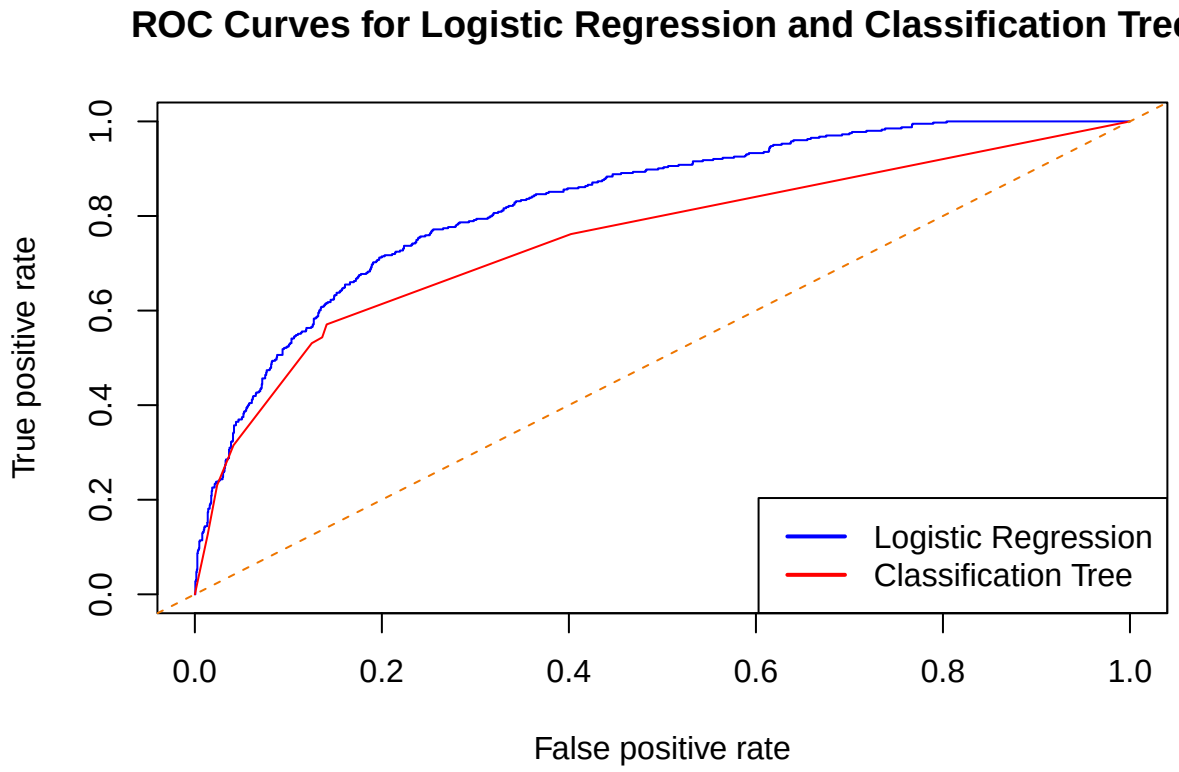
Using the data in data_test, assess appropriately, compare, and comment on the predictive performance of these two classifiers. In doing so, note that the company developing the app is primarily interested in the ability of the models to correctly identify positive matches between users.

```
#PREDICTIVE PERFORMANCE:
#Calc and plot ROC curves for models
pred_obj_lr <- prediction(p_test, data_test$match)
roc_lr <- performance(pred_obj_lr, "tpr", "fpr")

pred_obj_tree <- prediction(tree_p_test, data_test$match)
roc_tree <- performance(pred_obj_tree, "tpr", "fpr")

plot(roc_lr, col = "blue", main = "ROC Curves for Logistic Regression and Classification Tree")
plot(roc_tree, col = "red", add = TRUE)
```

```
abline(0, 1, col = "darkorange2", lty = 2) # add bisect line
legend("bottomright", legend = c("Logistic Regression", "Classification Tree"),
       col = c("blue", "red"), lwd = 2)
```




```
#calc the area under the ROC curve for models
auc_lr <- performance(pred_obj_lr, "auc")
auc_tree <- performance(pred_obj_tree, "auc")

cat("AUC for Logistic Regression:", auc_lr@y.values[[1]], "\n")
```

```
## AUC for Logistic Regression: 0.8296665
```

```
cat("AUC for Classification Tree:", auc_tree@y.values[[1]])
```

```
## AUC for Classification Tree: 0.7521457
```

As we can see from the graph above, both models achieve a decent level of classification for the data. The logistic regression performed better than the classification tree, with an AUC of 0.8296665, while the classification tree had an AUC of 0.7521457, which is relatively weaker. The higher the AUC value represents a higher level of discrimination between positive and negative cases.