D209: Data Mining 1

Performance Assessment

Task 2

Logan Rosemeyer

Western Governor's University

D209: Data Mining 1

Dr. Eric Straw

January 15, 2023

Contents

A1: Proposal of Question

Is it possible to predict churn based on the services that the customer has using decision trees so the organization can reduce the number of churned customers by catching which services lead to churn?

A2: Defined Goal

The goal of this data analysis is to see if we can accurately classify whether a customer will churn based on the services they have. If we can do that, we can make a plan to lower churn rate.

B1: Explanation of Prediction Method

A decision tree model uses multiple algorithms to decide to split a node into multiple sub-nodes. Decision trees start with the root of the tree and compare the values of the root attribute with the record's attribute. It then follows the branch corresponding to the value and moves to the next node. (Chauhan, 2022)

B2: Summary of Method Assumption

One assumption of decision trees is that all the training data is passed to the root node. As we move down the tree, not all of the data will pass through all nodes, but all data needs to pass through the root node. (Teja, 2022)

B3: Packages and Libraries List

The first package I uploaded was the dplyr package, which allowed me to use the select function to narrow down the columns I wanted to use to run my model. I also uploaded the caret package to see a confusion matrix of my predictions vs. response variable. To split the data, I uploaded the caTools package. To run my model, I used the rpart package, and to visualize my decision tree, I used the rpart.plot package.

C1: Data Preprocessing

One data preprocessing goal I had was to make sure there were no duplicates or null values. Duplicates would make the data inaccurate, while null values would make the model inaccurate or possibly not even allow the model to run.

C2: Data Set Variables

All variables that I am using are categorical. They are churn, portable modem, tablet, internet service provider, phone service, multiple lines, online security add-on, online backup add-on, device protection add-on, tech support add-on, streaming TV, streaming movies, and paperless billing.

C3: Steps for Analysis

The first thing I did to prepare my data set was to load the csv into R, import the necessary libraries, and look at the head and structure of my data.

```
churn <- read.csv("D://MS DA//D209 Task 2//churn_clean.csv")
```

```
library(dplyr)
library(caTools)

library(caret)

library(rpart)

library(rpart.plot)


head(churn)

str(churn)
```

The next step was to create a data frame with the variables I wanted to use to answer my question.

```
services <- churn %>% select(20, 27:38)
```

Next I checked for any duplicates or null values. If I had continuous variables, I would have also checked for outliers.

```
sum(duplicated(churn))

colSums(is.na(churn))
```

C4: Cleaned Data Set

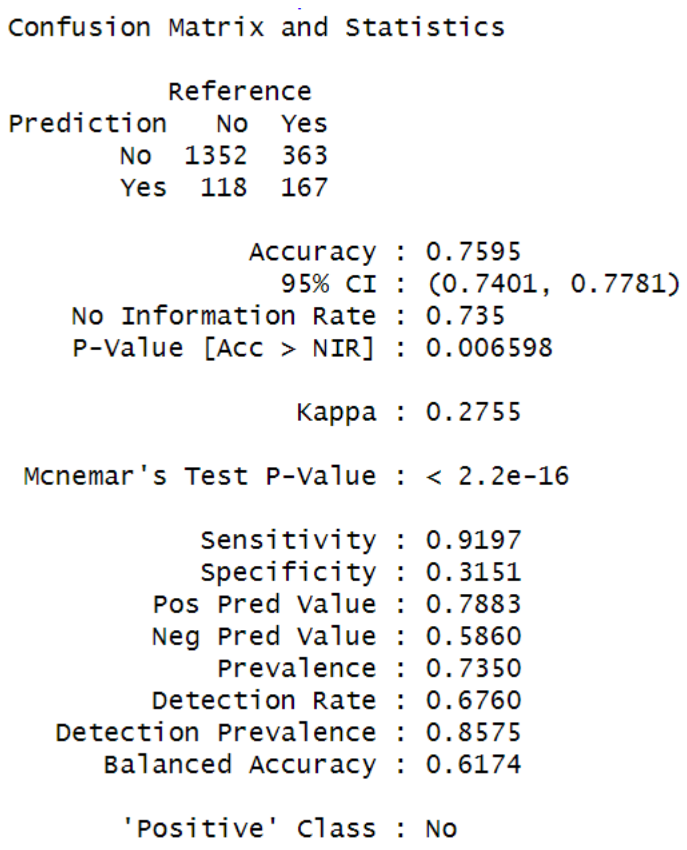
Cleaned data set has been uploaded in the submission.

D1: Splitting the Data

The training and test data sets are uploaded in the submission. 80% of the data is training, 20% is test.

D2: Output and Intermediate Calculations

I used a confusion matrix to analyze my model. As you can see, there were 481 incorrect predictions and the model was 76% accurate on the test data.



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1352  363
       Yes  118  167
                                          
               Accuracy : 0.7595          
                 95% CI : (0.7401, 0.7781)
    No Information Rate : 0.735           
    P-Value [Acc > NIR] : 0.006598        
                                          
                  Kappa : 0.2755          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.9197          
            Specificity : 0.3151          
         Pos Pred Value : 0.7883          
         Neg Pred Value : 0.5860          
             Prevalence : 0.7350          
         Detection Rate : 0.6760          
   Detection Prevalence : 0.8575          
      Balanced Accuracy : 0.6174          
                                          
       'Positive' Class : No              
```

D3: Code Execution

```
tree <- rpart(Churn ~ ., data=train)

rpart.plot(tree)

p <- predict(tree, train, type = "class")

p_test <- predict(tree, test, type = "class")

confusionMatrix(p, as.factor(train$Churn))

confusionMatrix(p_test, as.factor(test$Churn))
```

E1: Accuracy and MSE

The accuracy of my model ended up being 76%. Out of 2000 observations, there were 481 misclassifications. I also calculated the MSE of my model, which was 0.24. The smaller the MSE the better, so I would say this model is fairly accurate, but can definitely be improved upon.

E2: Results and Implications

Overall, my model ended up being fairly accurate, with an accuracy of 76% and MSE of 0.24. One place my model could be more accurate is there were too many no predictions when the reference was actually a yes. The accuracy in that scenario was not very accurate. This model could definitely help the organization reduce their churn rate by predicting which customers will churn based on the add-ons and services they have in their package.

E3: Limitation

One limitation of this model is that the churn variable is not uniform. There are many more non-churned customers than churn customers, which can result in too many non-churn predictions. This is likely why There were so many non-churn predictions for customers that actually did churn.

E4: Course of Action

I believe this model would be a good start for this organization to predict churn rate. It is more accurate than if we have no information, which is a good place to start. I think it would be possible to experiment with the model to make it even more accurate. I would recommend the organization use this model as it continues to increase the data collected, while periodically trying to optimize the model with the new information that the organization collects.

F: Panopto Demonstration

Video is included in the submission and can be found using the link below.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c9cf6f74-a690-4361-a4ab-af9300375d1e

G: Sources of Third-Party Code

confusionMatrix: Create a confusion matrix. RDocumentation. Retrieved January 22, 2023. https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix

rpart: Recursive Partitioning and Regression Trees. RDocumentation. Retrieved January 22, 2023. https://www.rdocumentation.org/packages/rpart/versions/4.1.19/topics/rpart

H: Sources

Chauhan, Nagesh Singh. *Decision Tree Algorithm, Explained.* KDnuggets. February 9, 2022. Retrieved January 22, 2023. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Teja, Kamsu Sasi. *Decision Tree Algorithm.* Scaler Topics. May 2, 2022. Retrieved January 22, 2023. https://www.scaler.com/topics/decision-tree-algorithm/