Lab 2

Ayesha Gamage-(ayega981)/Muditha Cherangani(mudch175)

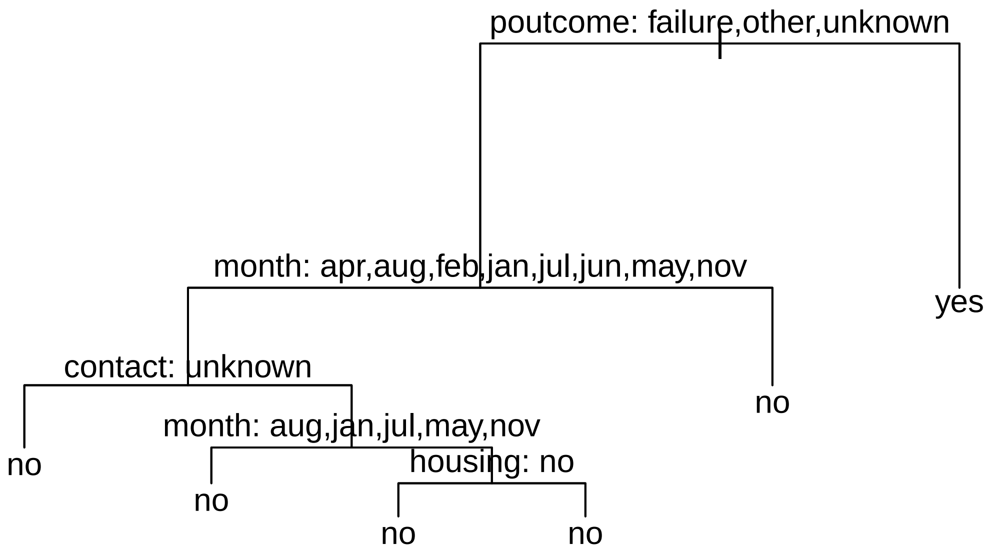
2023-12-05

Assignment 2. Decision trees and logistic regression for bank marketing

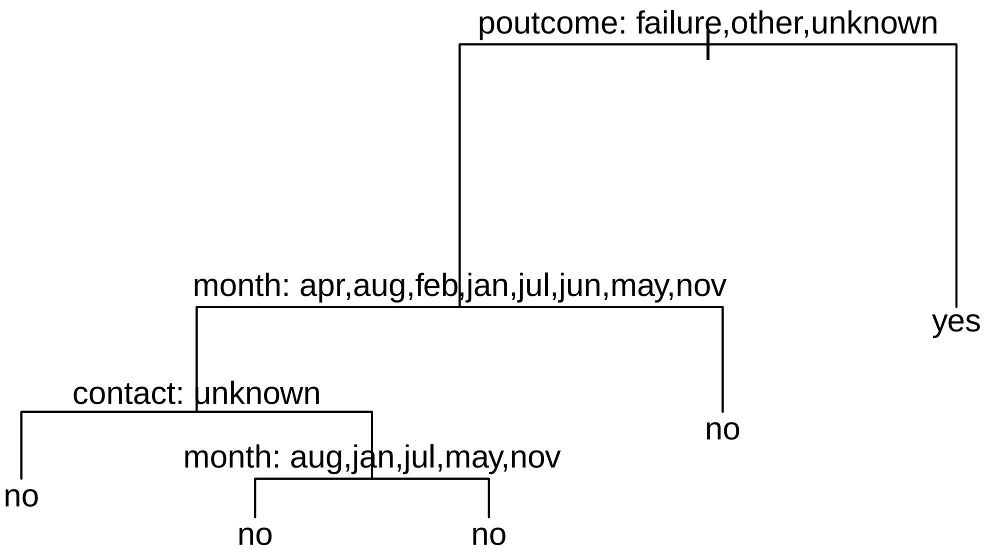
```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

a. Decision Tree with default settings


poutcome: failure,other,unknown
month: apr,aug,feb,jan,jul,jun,may,nov
yes
contact: unknown
no
month: aug,jan,jul,may,nov
housing: no
no
no
no
no

b. Decision Tree with smallest allowed node size equal to 7000.


poutcome: failure,other,unknown
month: apr,aug,feb,jan,jul,jun,may,nov
yes
contact: unknown
no
month: aug,jan,jul,may,nov
no
no
no

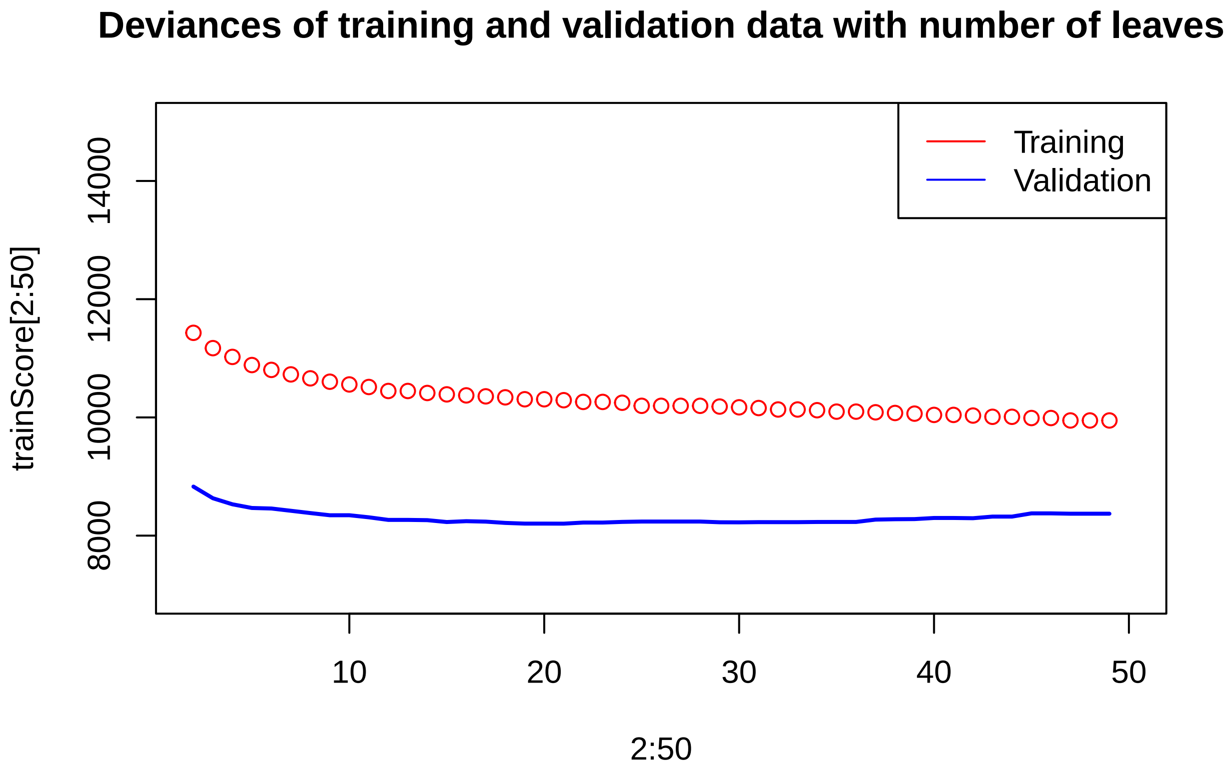
c. Decision trees minimum deviance to 0.0005.

3.Using training and validation sets to choose the optimal tree depth

Table 1: Misclassification Rate

	Train_error	Valid_error
Default	0.1048441	0.1092679
Nodesize_7000	0.1048441	0.1092679
Mindev_0.0005	0.0940058	0.1119221

According to the misclassification rates above ,the Default and Nodesize_7000 models can be regarded as the best ones. In these two models, the Valid_errors are small than the Mindev_0.0005 model and Mindev_0.0005 model's Valid_errors larger.



Examine the plot, Deviance are decrease when numbers of leaves increases.But after about 21 it has little increase. This is because as the number of leaves increase, the tree model becomes more and more complex. In here we can take optimal leave when Deviance is minimum.

```
## Optimal number of leaves 21
```

4.Confusion matrix

```
##             Reference
## Prediction    no   yes
##         no  11822   157
##         yes  1309   276
```

```
##
## Accuracy  0.8919198
```

```
## F1 score  0.9416169
```

Accuracy of the optimal model is about 0.9.This model has good predictability.

```
##
## Classification tree:
## tree::tree(formula = as.factor(y) ~ ., data = test, control = tree.control(nobs = n,
##     mincut = optimal_leaves))
## Variables actually used in tree construction:
```

```
## [1] "poutcome" "month"    "contact"  "housing"
## Number of terminal nodes:  6
## Residual mean deviance:  0.6096 = 8265 / 13560
## Misclassification error rate: 0.1081 = 1466 / 13564
```

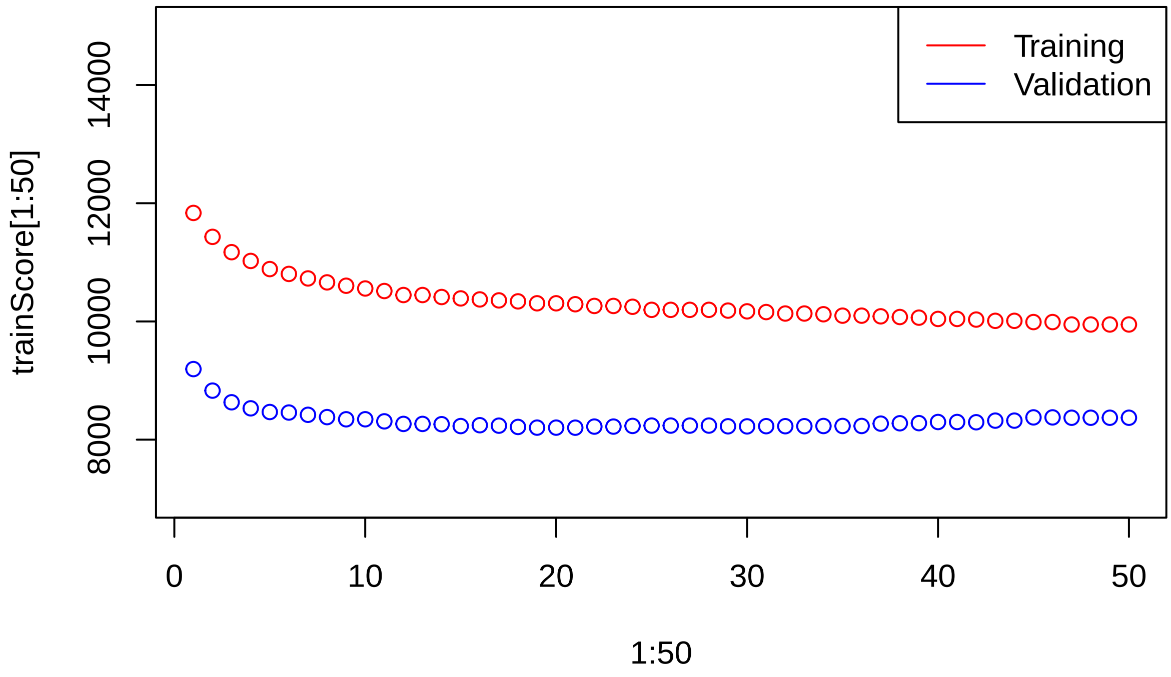
According to this summery “poutcome” ,“month”, “contact” and “housing” variables are the most important.

5.A decision tree classification

```
##           Reference
## Prediction    no   yes
##        no  11979     0
##        yes  1585     0
```

```
##
## Accuracy  0.8831466
```

```
## F1 score  0.9379478
```



```
## Optimal leave is,  21
```

4.

```
##           Reference
## Prediction    no   yes
##        no  11822  1309
##        yes   157   276
```

```
##  Accuracy
## 0.8919198
```

```
##        F1
## 0.9416169
```

5.

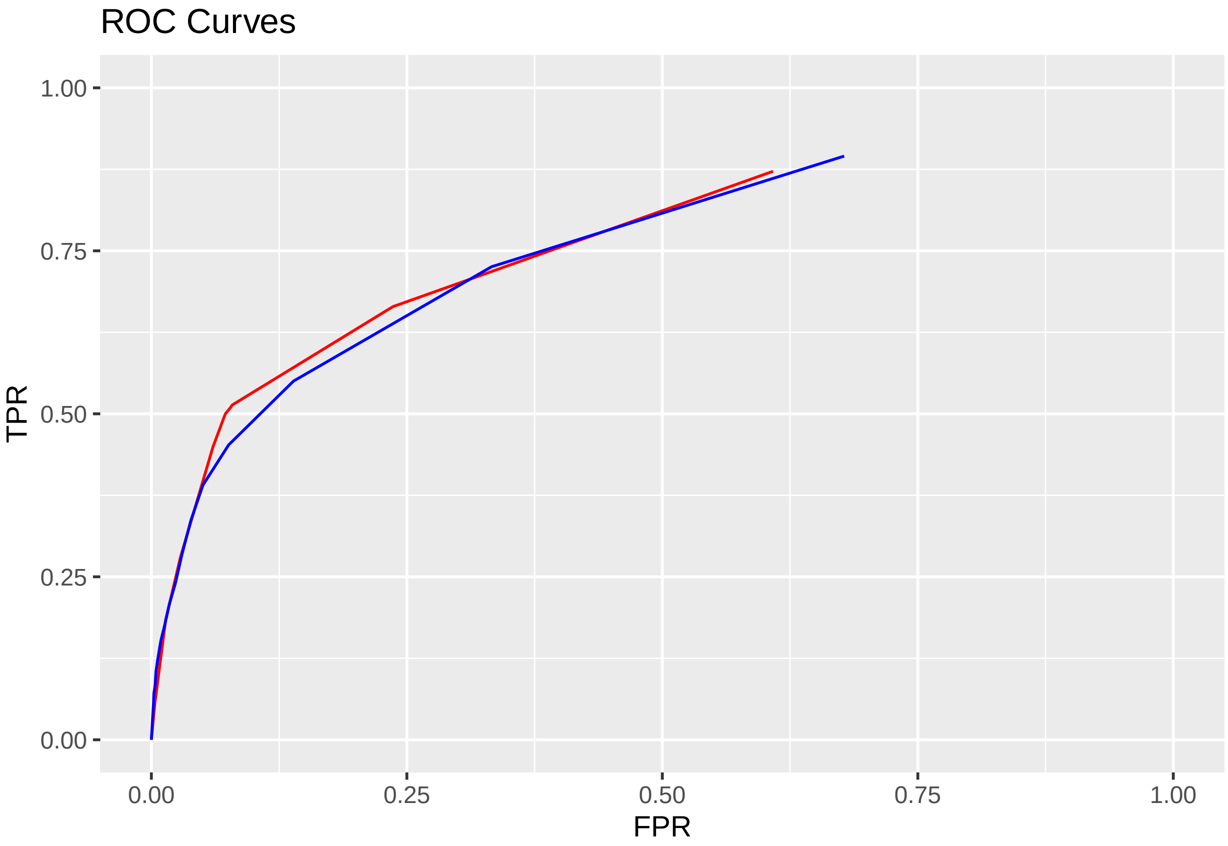
```
##           Reference
## Prediction    no   yes
##        no  11189   861
##        yes   790   724
```

```
##  Accuracy
## 0.8919198
```

```
##        F1
## 0.9416169
```

According to the data above we can see that F1 scores and Accuracy approximately equal.But in the confusion matrix, increased the predicted value of yes.We can conclude this model is batter.

6.Optimal tree and a Logistic regression



When comparing two plots for the models, we can find that the AUC of the tree model(red curve)is larger than that of glm(blue curve) Here we can conclude tree model is better.