INTRODUCTION TO MACHINE LEARNING (NPFL054)
A template for Homework #2

Name: Jiří Klepl

School year: 2019/2020

- **Provide answers for the exercises.**
- **For each exercise, your answer cannot exceed one sheet of paper.**

1 Data analysis

If we picked 100 samples we would get on average precision 6% (6 true positives and 94 false positives – that if we did not use any predictor whatsoever and just picked at random)

1a) The table MOSHOOFD shows noticeable correlation between the type of the customer and his likelihood of purchasing a caravan. The same could be said about MOSTYPE table. In MOSHOOFD the category that is most likely to buy a caravan is 2 while 4 seems to be the least likely.

The tables can be seen in files **table-moshoofd.pdf** and **table-mostype.pdf**

1b) From the table MOSTYPE vs MOSHOOFD we can clearly see that MOSTYPE is subcategory of MOSHOOFD which is indeed true as seen in caravan.attributes.pdf.

This table can be seen in the file **table-mostype-vs-moshoofd.pdf**

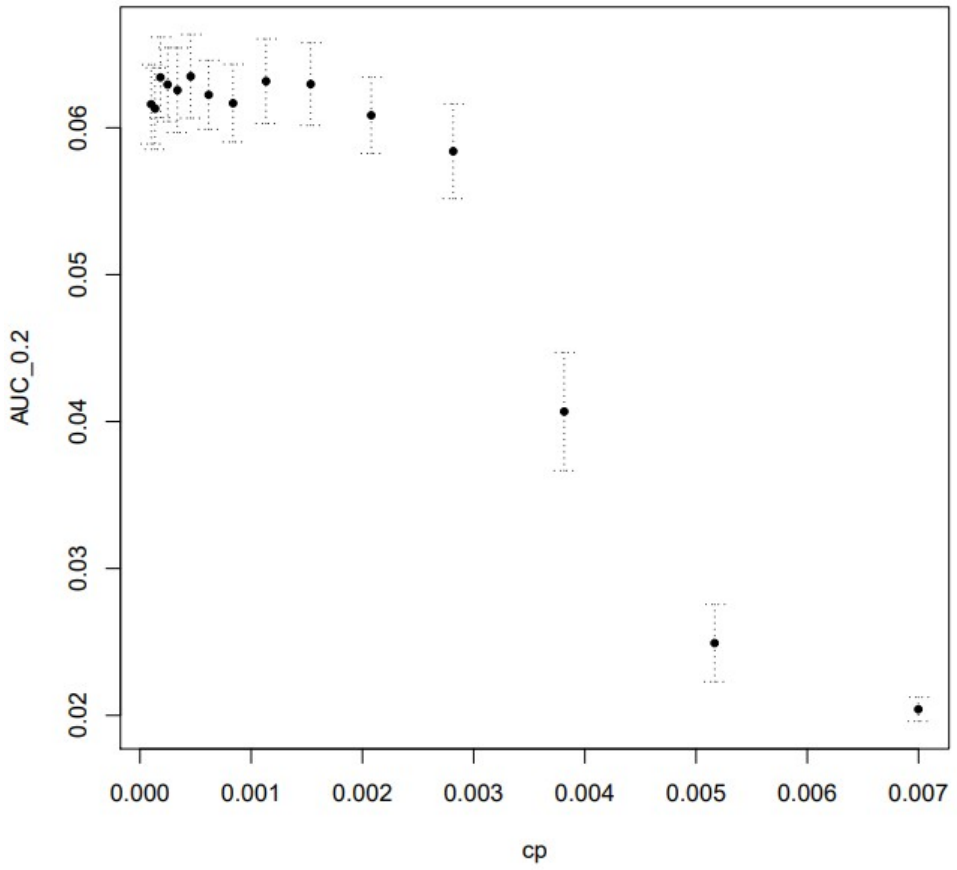
2 Model fitting, optimization, and selection

We want to achieve maximum precision by optimizing each model by maximizing $AUC_{0.2}$ as lower FPRs achieve higher precision values and FPR of 0.2 achieves at most 23% precision while higher FPRs achieve much lower values.

Each is evaluated by 10-fold cross-validation measuring $AUC_{0.2}$, and confidence intervals using standard significance level of 5%.

2a Decision Tree

Decision tree was optimized modulating it’s complexity parameter.



Here we can see that any cp below 0.0020790127 gives good results and choosing this exact number seems to be the best choice as this prevents overfitting.

2b Random Forrest

Here we optimize $AUC_{0.2}$ by modulating two values – number of trees (ntree) and feature sample size of the trees (mtry). Smaller mtry values are considered better as this leads to more variation between the trees and the predictor should be less prone to overfitting variation in data.

All experiments can be seen in **forest.pdf**.

This model doesn't give consistent results in the cross-validation and neither parameter's value seems to have any statistical significance. All confidence interval span well over half of the value of their corresponding mean $AUC_{0.2}$ and this model overall gives worse predictions according to AUC then the other two for any pair of parameter values.

2.4c Logistic regression

Here we tune two different values: lambda and alpha.

All experiments can be seen in **regression.pdf** from which we can notice that both alpha and lambda are quite significant, lambda slightly more than alpha.

For each alpha the set of desired lambdas gets thinner the higher the alpha value is and the model is more sensitive to its (lambda's) change.

The best alpha is difficult to guess as it doesn't affect AUC that much but it affect the choice of lambda. But good choice seems to be alpha = 7/9 and lambda = 0.014597743 as this seems to be the point with the lowest confidence interval out of the values values on the elbow of the plot of AUC dependent on the value of lambda.

This model seems to give the best results if tuned correctly and is better than decision tree at choosing the right threshold and decision trees tend to give more things the same probability value.

2d Evaluation of tuned models

The result of evaluation experiments can be seen in **evaluation.pdf**

The models have not so different ROC curves but the steepest on the interval we are interested in (0 to 0.2) is the logistic regression model which makes it our best choice.

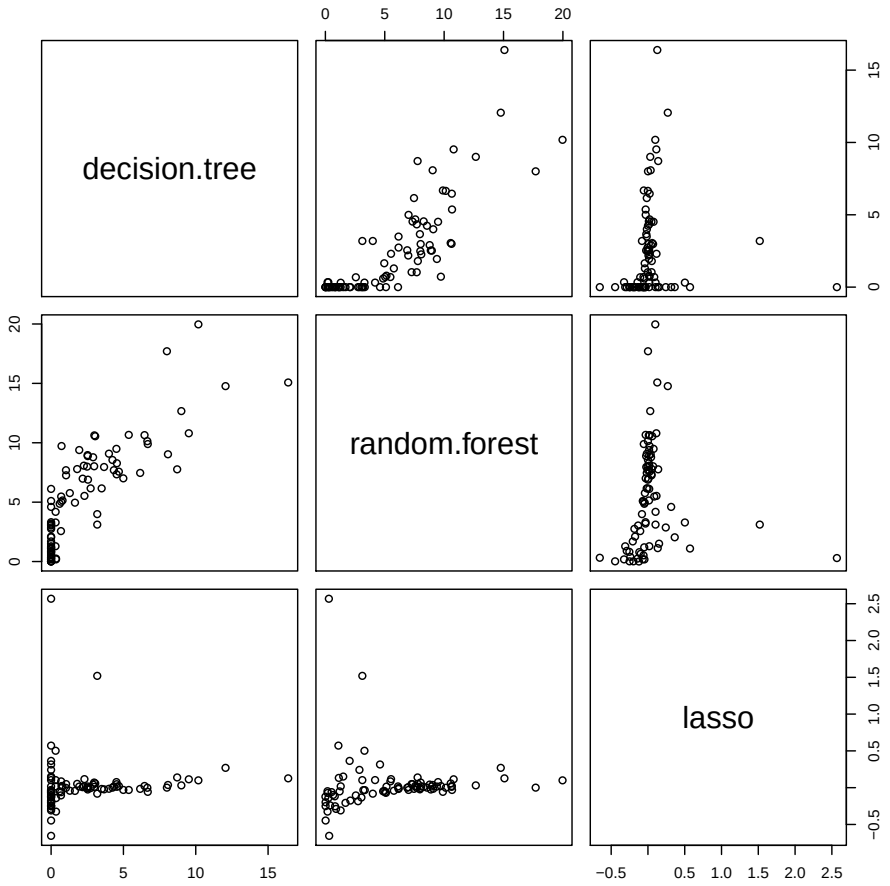
2e Choosing the ideal threshold

The ideal threshold on the training data seems to be 0.09952441, this gives us exactly 100 results of the test data classified as “Yes”.

```
correlation of feature importance values according to the models:
              decision.tree random.forest      lasso
decision.tree    1.00000000    0.82167116 0.07024048
random.forest    0.82167116    1.00000000 0.03200697
lasso            0.07024048    0.03200697 1.00000000
```

From this table we can see that lasso gives us no idea of importance of the values and that the random forest model gives similar importance to features overall as the decision tree model.

Graphically here:



4 Final prediction o the blind test

For this I have chosen the logistic regression model with the same parameters as in 2c).

It was taught on the whole Caravan set and its prediction can be found in **T.prediction** as requested by the specifics given by the teachers.