INTRODUCTION TO MACHINE LEARNING (NPFL054)
Homework #3

Name: Pavel Mikuláš

School year: 3.

1. Data analysis

1a)

	1	2	3	4	5	6	7	8	9	10
Population	552.0	502.00	886.00	52	569.00	205.00	550.00	1563.00	667.0	276.00
caravan%	**8.7**	**13.15**	**6.66**	**0**	**2.64**	**1.95**	**3.64**	**5.69**	**6.3**	**1.81**

This table shows the population and relative percentages of people that purchased the insurance for each of the L2 groups based on the MOSHOOFD attribute. We can see that Driven Growers and Successful hedonists are the most likely to purchase the insurance. Career Loners doesn't seem to purchase the insurance whatsoever, but there are only 52 records of them so we can't really deduce anything from that. Farmers and cruising seniors are also not likely to purchase the insurance for some reason.

	1	2	3	4	5	6	7	8	9	10
Population	124.00	82.00	249.00	52.00	45.00	119.00	44.00	339.00	278.00	165.00
caravan%	**10.48**	**7.32**	**10.04**	**3.85**	**4.44**	**10.08**	**6.82**	**15.04**	**4.32**	**5.45**
	11	**12**	**13**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Population	153.00	111.00	179.00	5	16	9	19	3	25	15
caravan%	**5.88**	**14.41**	**7.26**	**0**	**0**	**0**	**0**	**0**	**8**	**0**
	22	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**
Population	98.00	251.00	180.00	82.00	48.00	50	25	86.00	118.00	205.00
caravan%	**4.08**	**1.59**	**2.78**	**2.44**	**2.08**	**2**	**0**	**2.33**	**3.39**	**2.93**
	32	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**
Population	141.00	810.00	182.00	214.00	225.00	132.00	339.00	328.00	71	205.00
caravan%	**5.67**	**5.68**	**4.95**	**3.74**	**7.11**	**7.58**	**6.78**	**5.79**	**0**	**2.44**

The following table shows the same realtionship for the MOSTYPE attribute. There we can deduce that Affluent young families, Middle class families and High income people are the most likely to purchase the caravan insurance. While people living in cities and cosmopolitan people never do so. Probably because they don't own caravans.

1b)

Next we analyze the relationship between the MOSTYPE and MOSHOOFD attributes. The following table shows the distribution of the L0 types between the L2 types. We can see that is it distributed really nicely. Implying we may be able to omit the L0 group and only group the customers using the L2 group so we can reduce the dimensionality of our models.

	1	2	3	4	5	6	7	8	9	10
1	124	0	0	0	0	0	0	0	0	0
2	82	0	0	0	0	0	0	0	0	0
3	249	0	0	0	0	0	0	0	0	0
4	52	0	0	0	0	0	0	0	0	0
5	45	0	0	0	0	0	0	0	0	0
6	0	119	0	0	0	0	0	0	0	0
7	0	44	0	0	0	0	0	0	0	0
8	0	339	0	0	0	0	0	0	0	0
9	0	0	278	0	0	0	0	0	0	0
10	0	0	165	0	0	0	0	0	0	0
11	0	0	153	0	0	0	0	0	0	0
12	0	0	111	0	0	0	0	0	0	0
13	0	0	179	0	0	0	0	0	0	0
15	0	0	0	5	0	0	0	0	0	0
16	0	0	0	16	0	0	0	0	0	0
17	0	0	0	9	0	0	0	0	0	0
18	0	0	0	19	0	0	0	0	0	0
19	0	0	0	3	0	0	0	0	0	0
20	0	0	0	0	25	0	0	0	0	0
21	0	0	0	0	15	0	0	0	0	0
22	0	0	0	0	98	0	0	0	0	0
23	0	0	0	0	251	0	0	0	0	0
24	0	0	0	0	180	0	0	0	0	0
25	0	0	0	0	0	82	0	0	0	0
26	0	0	0	0	0	48	0	0	0	0
27	0	0	0	0	0	50	0	0	0	0
28	0	0	0	0	0	25	0	0	0	0
29	0	0	0	0	0	0	86	0	0	0
30	0	0	0	0	0	0	118	0	0	0
31	0	0	0	0	0	0	205	0	0	0
32	0	0	0	0	0	0	141	0	0	0
33	0	0	0	0	0	0	0	810	0	0
34	0	0	0	0	0	0	0	182	0	0
35	0	0	0	0	0	0	0	214	0	0
36	0	0	0	0	0	0	0	225	0	0
37	0	0	0	0	0	0	0	132	0	0
38	0	0	0	0	0	0	0	0	339	0
39	0	0	0	0	0	0	0	0	328	0
40	0	0	0	0	0	0	0	0	0	71
41	0	0	0	0	0	0	0	0	0	205

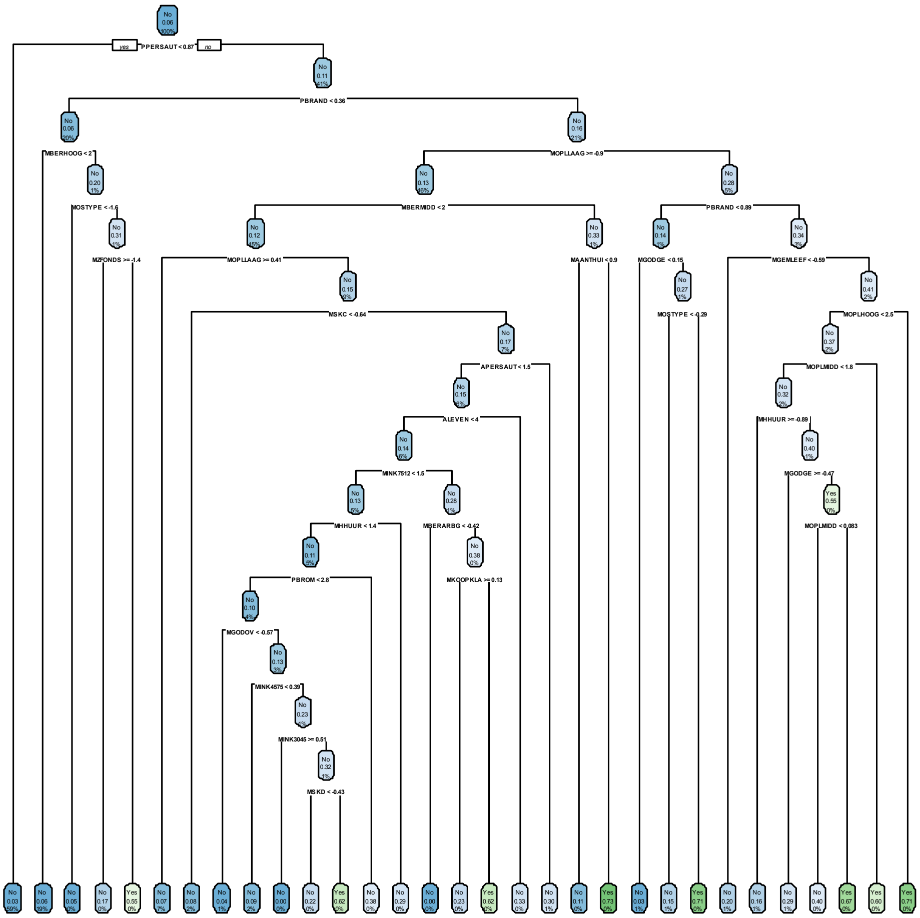
2. Model fitting, optimization and selection

2a)

Next we are gonna try to find and fit the best model for our dataset. We will be using cross-validation over training set to estimate the model performance. Since the dataset it really unbalanced(only 6% positive examples) rather than optimizing for accuracy we will be optimizing the AUC(Area Under the ROC Curve).

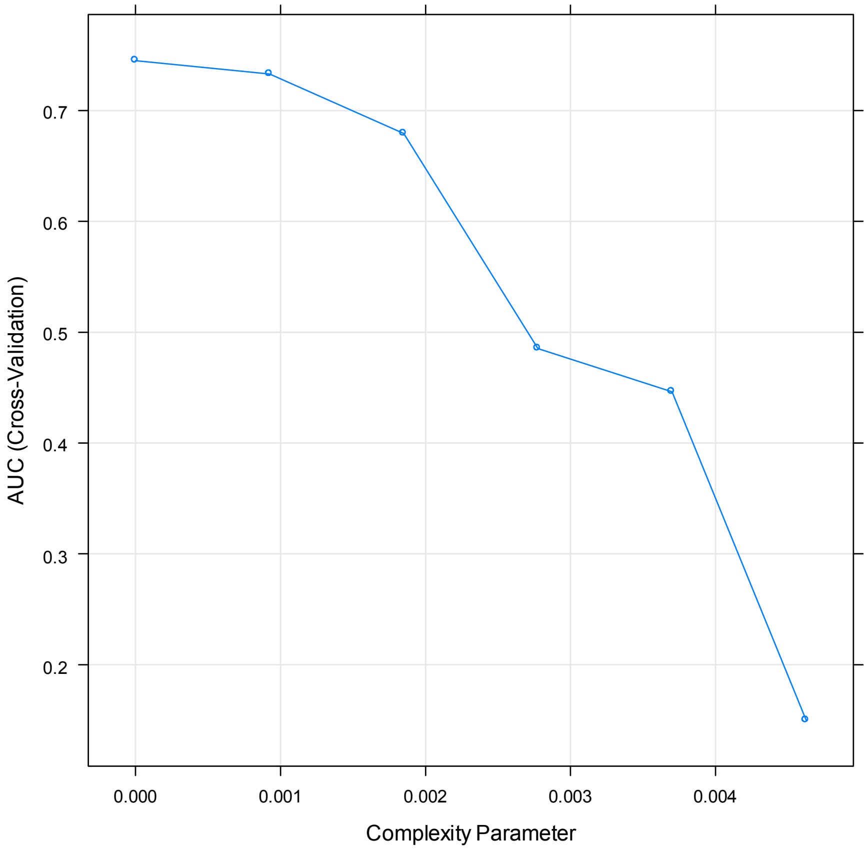
We should be dividing our folds carefully with how unbalanced our dataset is. But since I wanna be using the caret package for training and optimizing to model I can't easily to that. So I'm gonna be running the cross-validation multiple times to try and compensate for that.

Decision tree



Unfortunately the model is too large to be easily readable but we can atleast see the structure of the tree. It would certainly be better to choose some tree with lower cp within the standard deviation. But since we won't be using this algorithm any further it's not necessary.

The cp selected here is 0. So there is a high probability of the model being overfitted too.



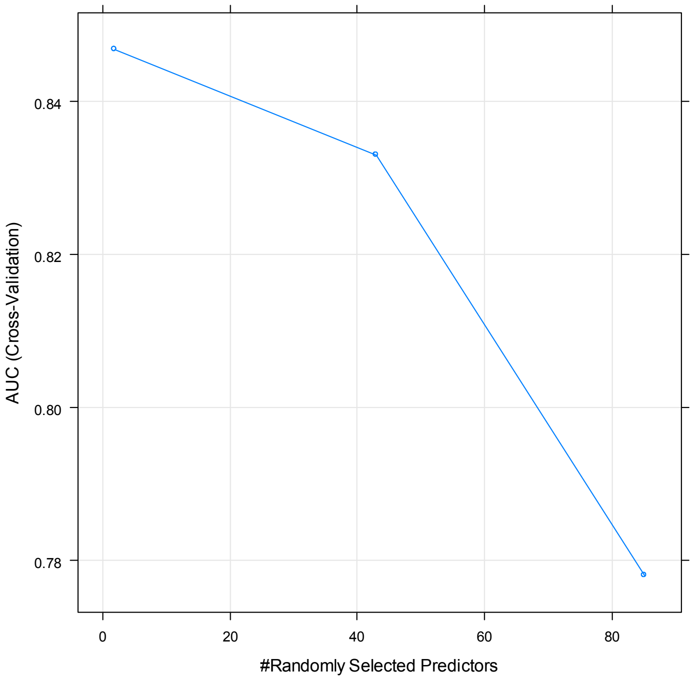
On this picture we can see how the AUC decreases with increasing values of cp. So any cp above 0.0001 would probably yield satisfying results.

2b)

Random forest

The best selected parameters from cross-validation are mtry = 2 and ntree = 500. I haven’t optimized ntree for timesaving reasons and also because the risk of overfitting by adding more trees is usually pretty low for random forests. This model has OOB estimate of 6.16% which is roughly the percentage of positive examples. So it’s probably not very good.

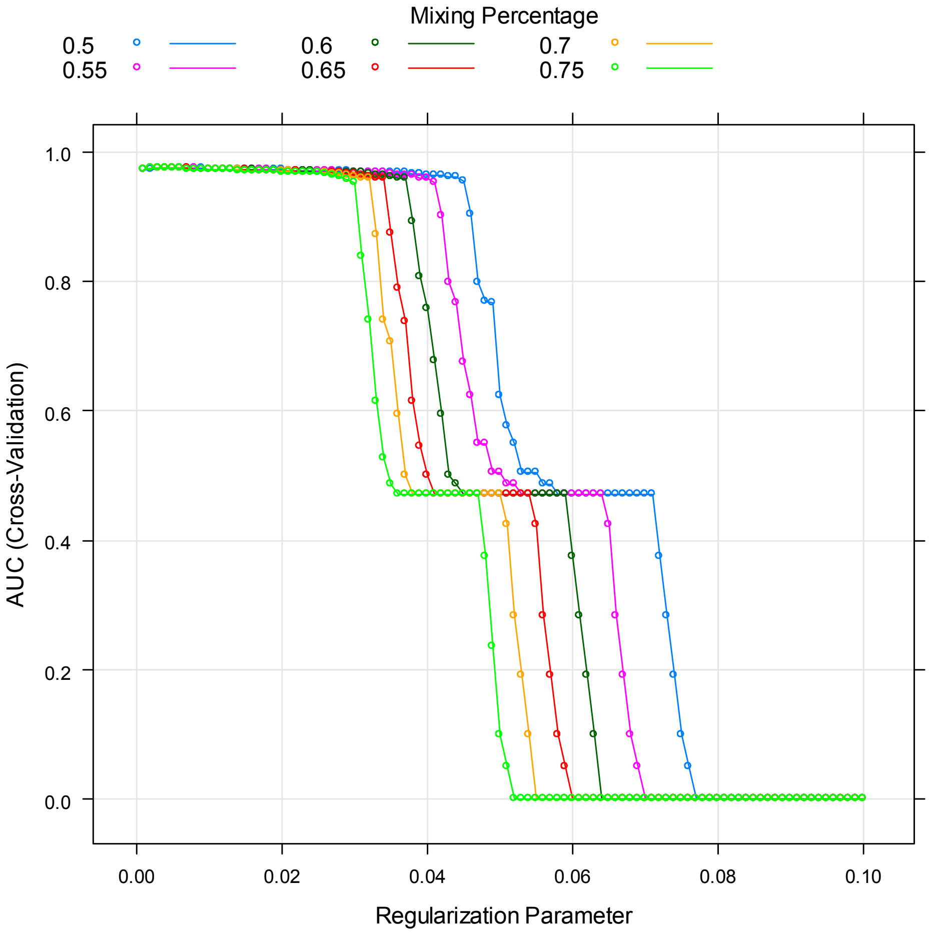
On the following picture we can see the decrease of AUC with increasing mtry parameter.



2c)

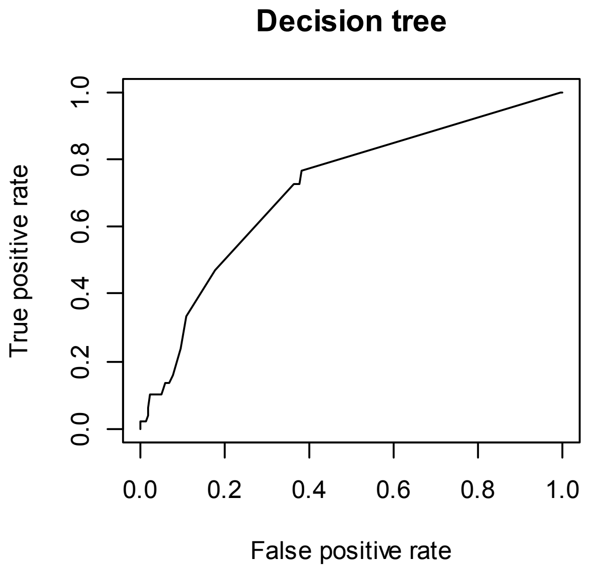
Logistic regression

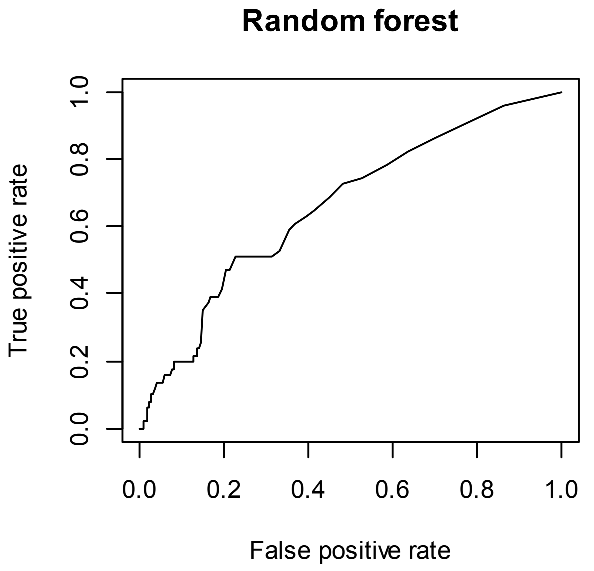
The optimal parameters found by cross-validation are alpha=0.5 and lambda = 0.006. On the following picture we can see the relationship between those two parameters.

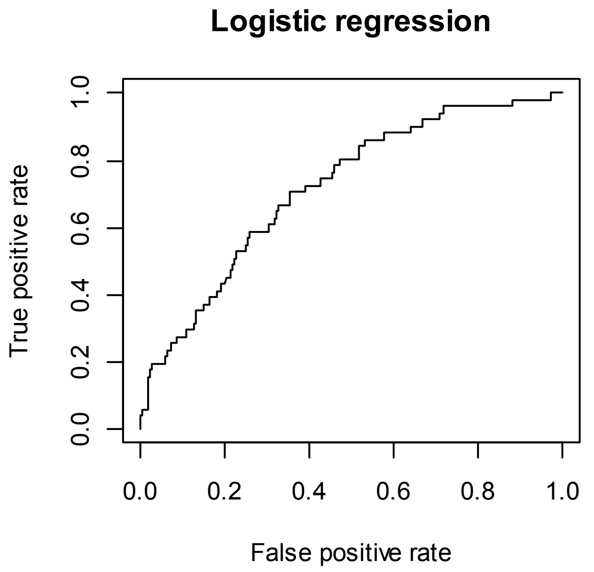


This is the model we will be working with from now on. It can surely be optimized more by doing repeated cross-validation or being smarter about it. But it will suffice.

I’ll now provide a short summary of the training results and compare the ROC curves of trained models.







2d,e)

In the pictures above we can see the ROC curves for 3 different algorithms: Decision tree, Random forest and Logistic regression. We are mostly interrested in the area below the ROC curve in the range between 0 and 0.2. The reason for that is that with a dataset so unbalanced we are gonna reach maximum recall pretty soon and from that point we are only gonna be losing precision as FPR will increase.

The AUC0.2 values are:

Decision tree	0.052
Random forest	0.042
Logistic regression	0.054

Those results are kind of underwhelming because I didn’t have enough time to train the models for longer. But they are good enough.

We can see that Logistic regression is the best algorithm for this task since it has the highest AUC. I’m now gonna provide a short summary of selected parameters for each model.

There is also no need to choose the optimal cut-off threshold. Since it will really depend on the data we are given. For instance for this model the optimal threshold for selection 100 potentional customers would be 0.131. And we would correctly predict 14 customers.

3. Feature selection and model interpretation

Using Lasso model I have selected the 13 most important features:

"MOPLHOOG", "MOPLLAAG", "MBERBOER", "MAUT1", "MINKM30", "MINKGEM", "MKOOPKLA", "PWAPART", "PPERSAUT", "PBRAND", "APERSAUT", "APLEZIER", "ABYSTAND"

Comparing those with the features used in my decision tree model:

"PPERSAUT", "MOSTYPE”, "MZFONDS", "PBRAND", "MOPLLAAG", "MSKC", "APERSAUT", "ALEVEN", "MINK7512", "PBROM" , "MGODOV" , "MINK4575" "MINK3045", "MSKD" "MHHUUR", "MBERARBG", "MKOOPKLA", "MBERMIDD", "MAANTHUI", "MGODGE", "MGEMLEEF", "MOPLHOOG", "MOPLMIDD"

We can see that the two sets are not the same. The reason for that could be that Lasso uses regression while decision tree are always trying to find the optimal split.

Comparing that with the 13 most important features from the random forest model:

MOSTYPE, MOSHOOFD, MFWEKIND, MOPLMIDD, MOPLLAAG, MBERMIDD, MBERARBG, MSKC, MHHUUR, MKOOPKLA, PPERSAUT, PBRAND, APERSAUT

We can see that they are pretty similar to the ones from decision tree model. Which further confirms my hypothesis stated above.

We could now use those reduced attribute sets to further train and improve out models. But I’m really short on time now so I’m not gonna to that either.

4. Final prediction on the blind test set

Done and sent.