INTRODUCTION TO MACHINE LEARNING
(NPFL054)
Homework #3

Name: Lafay Gareth

School year: 2019/2020

Task 1 – Data analysis

What would be your precision if you select 100 examples by chance?

If I selected 100 examples by chance it would have around 6.40% chance of being a purchase.

1.a

Type MOSHOOFD:

level	totalPeople	percentage
1	461	8.242950
2	419	12.410501
3	738	7.046070
4	45	0.000000
5	457	2.625821
6	171	1.754386
7	437	4.347826
8	1304	5.521472
9	561	6.951872
10	229	1.310044

Type MOSTYPE

level	totalPeople	percentage
1	106	9.433962
2	71	8.450704
3	210	9.047619
4	35	5.714286
5	39	2.564103
6	106	10.377358
7	37	5.405405
8	276	14.130435
9	235	4.680851
10	138	5.797101
11	126	7.142857
12	93	15.053763
13	146	6.849315
14	0	0.000000
15	4	0.000000
16	14	0.000000
17	9	0.000000
18	16	0.000000
19	2	0.000000
20	19	10.526316
21	12	0.000000
22	78	2.564103
23	206	1.941748
24	142	2.816901
25	66	1.515152
26	41	2.439024
27	43	2.325581
28	21	0.000000
29	68	2.941176
30	95	4.210526
31	166	3.614458
32	108	6.481481
33	677	5.317578
34	154	5.194805
35	175	4.000000
36	194	6.185567
37	104	8.653846
38	281	7.117438
39	280	6.785714
40	60	0.000000
41	169	1.775148

Task 1 – Data analysis

1.b

The two tables show a similar percentage of buyers, by calculating conditional entropy we can find out more about which type is the most useful.

Entropy of MOSHOOFD = 0.32

Entropy of MOSTYPE =0.31

They are too similar to decide which is the most useful, but MOSHOOFD has less data sparsity so it should be chosen. It has less data sparsity because it has as many examples but less levels.

Task 2 – Model fitting, optimization, and selection

2.a

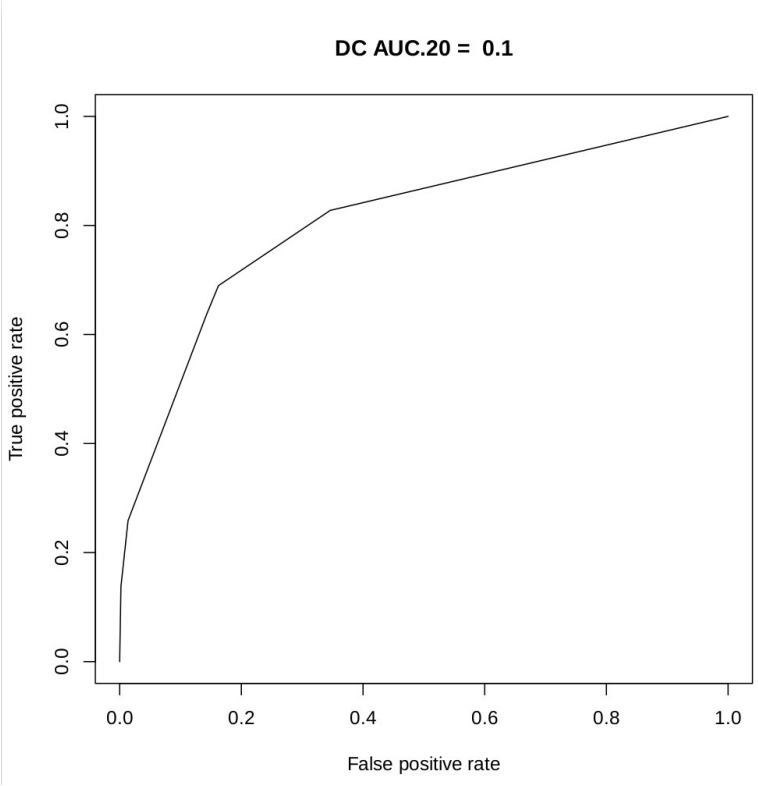
cp	mean	sd	low.interval	high.interval
0.00100	0.0548	0.01527380	0.04387378	0.06572622
0.00250	0.0518	0.01615756	0.04024158	0.06335842
0.00175	0.0536	0.01517454	0.04274479	0.06445521
0.00200	0.0510	0.01579029	0.03970431	0.06229569

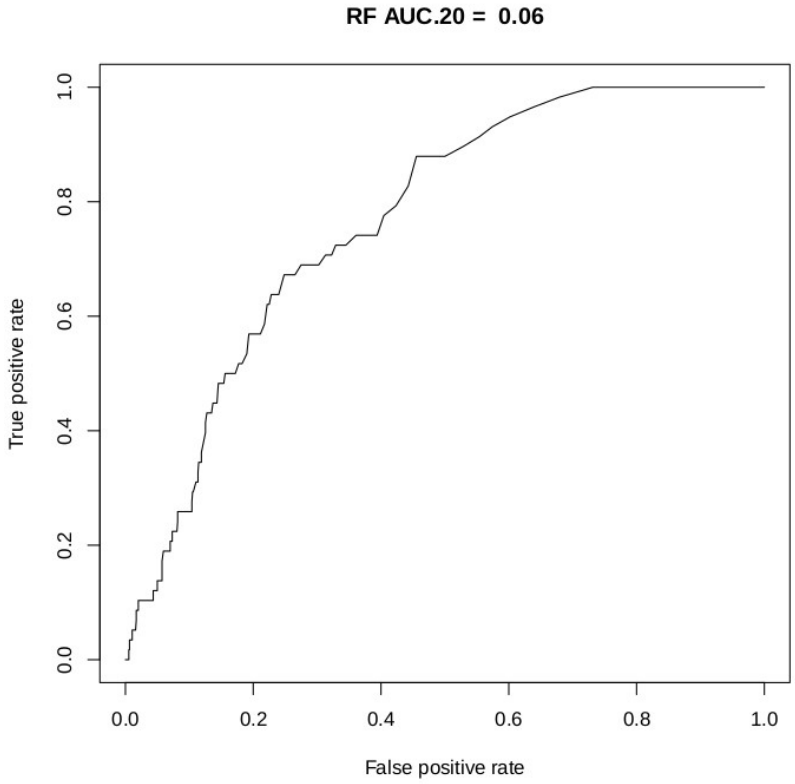
2.b

ntree	mean	sd	low.interval	high.interval
10	0.0415	0.009891073	0.03442435	0.04857565
150	0.0489	0.010375720	0.04147766	0.05632234
500	0.0491	0.010354280	0.04169299	0.05650701
600	0.0494	0.009754771	0.04242186	0.05637814
700	0.0490	0.010121484	0.04175953	0.05624047
800	0.0491	0.009711963	0.04215248	0.05604752

Task 2 – Model fitting, optimization, and selection

2.d





Task 3 – Model interpretation and feature selection

We can see that for the Decision Tree model only 5 features are really important, including MOSTYPE but not MOSHOOFD. Whereas for the random forest the difference in auc is so small that 5 features can suffice.