Tutorial Business Analytics

Homework 8 – Solution

Exercise 8.5

Design a stratified 5-fold cross-validation for the below-mentioned table.

Nr.	Attribute1	Attribute2	Attribute3	Attribute4	Class
1	4.9	3.1	1.5	0.1	1
2	5.0	3.2	1.2	0.2	0
3	5.5	3.5	1.3	0.2	0
4	4.9	3.1	1.5	0.1	1
5	4.4	3.0	1.3	0.2	1
6	5.1	3.4	1.5	0.2	0
7	5.0	3.5	1.3	0.3	1
8	4.5	2.3	1.3	0.3	1
9	4.4	3.2	1.3	0.2	0
10	5.0	3.5	1.6	0.6	0
11	5.1	3.8	1.9	0.4	0
12	4.8	3.0	1.4	0.3	1
13	5.1	3.8	1.6	0.2	0
14	4.6	3.2	1.4	0.2	1
15	5.3	3.7	1.5	0.2	0
16	5.0	3.3	1.4	0.2	0
17	7.0	3.2	4.7	1.4	1
18	6.4	3.2	4.5	1.5	0
19	6.9	3.1	4.9	1.5	1
20	5.5	2.3	4.0	1.3	1

Solution 8.5

The proportion of the two types of class labels is [10, 10] = 1:1 and the size of the data set is 20. In 5-fold cross-validation, the original sample is randomly partitioned into 5 equal size subsamples. Every subsample consists of 20/5 = 4 instances. In stratified k-fold cross-validation, the folds are selected so that each fold contains roughly the same proportions of the two types of class labels. Regarding a sample with 4 instances per subsample and a proportion of 1:1 of the two types of class labels every subsample has to contain 2 instances of the class 0 and two instances of the class 1.

Example partition:
(Note: This is only one of many possible solutions!)

- P1 = {1,2,3,4}
- P2 = {5,6,7,9}
- P3 = {8,10,11,12}
- P4 = {13,14,15,17}
- P5 = {16,18,19,20}

Cross-validation:

Step	Training	Test
1	$P2 \cup P3 \cup P4 \cup P5$	P1
2	$P1 \cup P3 \cup P4 \cup P5$	P2
3	$P1 \cup P2 \cup P4 \cup P5$	P3
4	$P1 \cup P2 \cup P3 \cup P5$	P4
5	$P1 \cup P2 \cup P3 \cup P4$	P5

Note: The attributes from above-mentioned table are not necessary for the solution.

Exercise 8.6

You want to bet on soccer matches and try to predict match results. In order to improve your forecasts, you decide to use your knowledge on data mining and construct a decision tree. The table below compares the real outcome and your predicted outcome of 15 matches.

Calculate the *Accuracy*, the *True Positive Rate*, the *False Positive Rate* and the *True Negative Rate* for your decision tree based on your predictions.

True Class	Predicted Class
1	1
0	1
1	1
1	1
1	0
0	0
0	1
1	0
0	0
0	0
0	0
1	1
1	1
1	0
0	0

Solution 8.6

True Class	Predicted Class	Event
1	1	TP
0	1	FP
1	1	TP
1	1	TP
1	0	FN
0	0	TN
0	1	FP
1	0	FN
0	0	TN
0	0	TN
0	0	TN
1	1	TP
1	1	TP
1	0	FN
0	0	TN

Number of occurrences:

- True Positives (TP): 5
- False Negatives (FN): 3
- True Negatives (TN): 5
- False Positives (FP): 2

Measures:

- True Positive Rate (Recall)= TP/(TP+FN) = 5/(5+3) = 5/8 = 0.625
- False Positive Rate (False Alarm Rate) = FP/(FP+TN) = 2/(2+5) = 2/7 = 0.286
- True Negative Rate (Specificity) = TN/(TN+FP) = 5/(5+2) = 5/7 = 0.714
- Accuracy = (TN+TP)/(TP+FN+TN+FP) = (5+5)/(5+3+5+2) = 10/15 = 0.667

Exercise 8.7

Use a paired t-Test to determine whether the classifiers 1 and 2 are significantly *better* than the baseline classifier 0 at significance level 5%.

Classifier 0	Classifier 1	Classifier 2
0.71	0.69	0.76
0.69	0.67	0.79
0.83	0.74	0.78
0.72	0.63	0.73
0.91	0.89	0.92
0.74	0.70	0.78
0.72	0.78	0.83
0.83	0.86	0.75
0.67	0.66	0.71
0.79	0.81	0.85
0.82	0.78	0.83
0.72	0.72	0.87
0.71	0.78	0.86
0.84	0.85	0.77
0.80	0.78	0.82

Solution 8.7

Classifier 0	Classifier 1	C_0-C_1	Classifier 2	C_0-C_2
0.71	0.69	0.02	0.76	-0.05
0.69	0.67	0.02	0.79	-0.10
0.83	0.74	0.09	0.78	0.05
0.72	0.63	0.09	0.73	-0.01
0.91	0.89	0.02	0.92	-0.01
0.74	0.70	0.04	0.78	-0.04
0.72	0.78	-0.06	0.83	-0.11
0.83	0.86	-0.03	0.75	0.08
0.67	0.66	0.01	0.71	-0.04
0.79	0.81	-0.02	0.85	-0.06
0.82	0.78	0.04	0.83	-0.01
0.72	0.72	0.00	0.87	-0.15
0.71	0.78	-0.07	0.86	-0.15
0.84	0.85	-0.01	0.77	0.07
0.80	0.78	0.02	0.82	-0.02

Alternative hypotheses:

1. H_1: Classifier 1 is significantly better than the baseline Classifier 0
2. H_1: Classifier 2 is significantly better than the baseline Classifier 0

In order to conduct a left-tailed test we subtract the classifier from the baseline. If we did it the other way around, we would conduct a right-tailed test.
If the tested classifier is better than the baseline, their difference will be < 0 (left-tailed test).

The critical value can be retrieved from the t-table:

- degree of freedom = k – 1 = 15 – 1 = 14
- α = 5%
- t_c = 1.761 (according to the table)

We will reject the null hypothesis if:

- $t \leq -t_c$

Test 1: Baseline vs. classifier 1:

- Average difference: 0.0107
- Standard deviation: 0.0457
- t = 0.9031
- $t > -t_c \rightarrow$ not significant

Test 2: Baseline vs. classifier 2:

- Average difference: - 0.0367
- Standard deviation: 0.0711
- t = -1.9979
- $t \leq -t_c \rightarrow$ significant

Interpretation:
Test 1: Do not reject H_0. Classifier 1 is not significantly better than the baseline classifier at α = 0.05.
Test 2: Reject H_0. At α = 0.05, classifier 2 is significantly better than the baseline classifier at α = 0.05.

Exercise 8.8

a) Construct a Gain-Curve (5% steps) and the corresponding Lift-Curve based on the evaluation results mentioned below.

b) Construct a ROC-Curve for every possible cutoff-value.

Probability of class 1	True class
0.998	1
0.991	1
0.990	1
0.942	0
0.896	1
0.752	1
0.639	1
0.633	0
0.612	1
0.584	0
0.554	0
0.514	1
0.448	0
0.364	1
0.324	0
0.316	1
0.278	0
0.182	0
0.160	0
0.110	0

Solution 8.8

a) Gain-Curve and Lift-Curve

In order to construct the Gain-Curve, we sort all instances by their score (in this case: the probability of their true class being positive). Starting from the highest score, different portions of the instances are classified as positive. As a result, some instances are classified wrong. The true positive rate (tp/(tp+fn)) has to be calculated for every portion.

According to the exercise description all possible 5% steps have to be evaluated in order to construct the diagram.

The values for 25% and 50% of the instances are exemplarily calculated below.

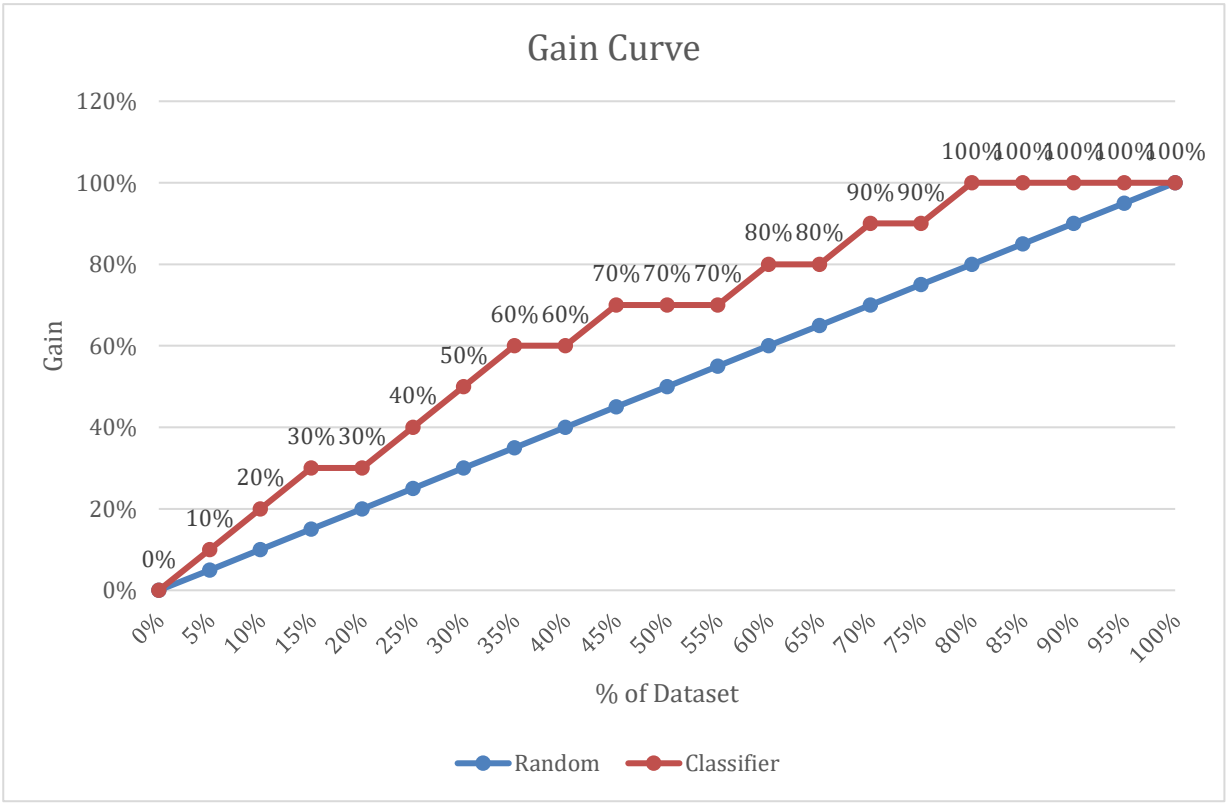
For x = 25%,

5 out of 20 instances are predicted to be positive (25%). 4 out of 10 (whole data set) positive instances are predicted correctly. Hence, the x-value is 25% and the y-value is 40%.

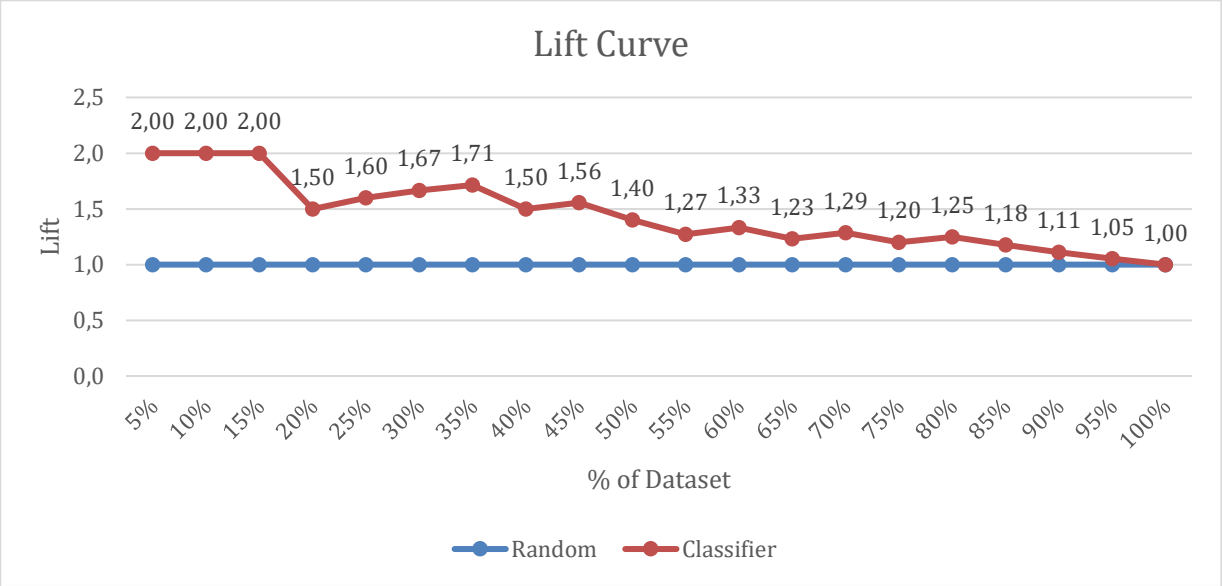
Score	True Class	Prediction	Event	Positives	Instances
0.998	1	1	TP		
0.991	1	1	TP		
0.990	1	1	TP	4	5
0.942	0	1	FP		
0.896	1	1	TP		
0.752	1	0	FN		
0.639	1	0	FN		
0.633	0	0	TN		
0.612	1	0	FN		
0.584	0	0	TN		
0.554	0	0	TN		
0.514	1	0	FN		
0.448	0	0	TN	6	15
0.364	1	0	FN		
0.324	0	0	TN		
0.316	1	0	FN		
0.278	0	0	TN		
0.182	0	0	TN		
0.160	0	0	TN		
0.110	0	0	TN		

For x = 50%,
10 out of 20 instances are predicted to be positive (50%). 7 out of 10 (whole data set) positive instances are predicted correctly. Hence, the x-value is 50% and the y-value is 70%.

Score	True Class	Prediction	Event	Positives	Instances
0.998	1	1	TP	7	10
0.991	1	1	TP		
0.990	1	1	TP		
0.942	0	1	FP		
0.896	1	1	TP		
0.752	1	1	TP		
0.639	1	1	TP		
0.633	0	1	FP		
0.612	1	1	TP		
0.584	0	1	FP		
0.554	0	0	TN	3	10
0.514	1	0	FN		
0.448	0	0	TN		
0.364	1	0	FN		
0.324	0	0	TN		
0.316	1	0	FN		
0.278	0	0	TN		
0.182	0	0	TN		
0.160	0	0	TN		
0.110	0	0	TN		



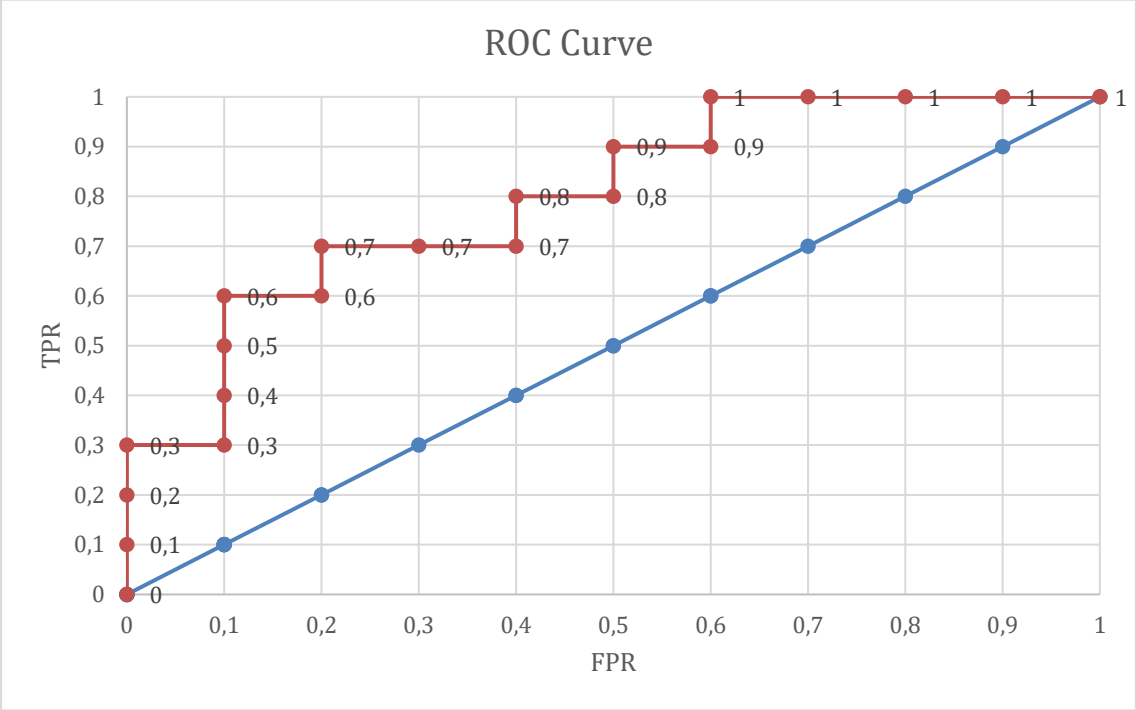
To get the corresponding Lift-Curve, we divide the y-value by the x-value for every point of the Gain-Curve.



b) The ROC-Curve is similar to the Gain-Curve, but the x-axis shows the false positive rate instead of the percentage of instances classified positive. The following example illustrates the calculation based on a cutoff value (threshold) of 0.824, e.g. every instance with a greater score is predicted as positive.

Score	True Class	Prediction	Event	Positives	Negatives
0.998	1	+	TP	4	1
0.991	1	+	TP		
0.990	1	+	TP		
0.942	0	+	FP		
0.896	1	+	TP		
0.752	1	-	FN	6	9
0.639	1	-	FN		
0.633	0	-	TN		
0.612	1	-	FN		
0.584	0	-	TN		
0.554	0	-	TN		
0.514	1	-	FN		
0.448	0	-	TN		
0.364	1	-	FN		
0.324	0	-	TN		
0.316	1	-	FN		
0.278	0	-	TN		
0.182	0	-	TN		
0.160	0	-	TN		
0.110	0	-	TN		

At threshold of 0.824, the TPR = 4/10 and FPR = 1/10. So, we draw a point at FPR (x-axis) = 0.1, and TPR (y-axis) = 0.4.



Annex

t-table

df	α = 0.1	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807