Using classifiers for mail promotions. Part II. Business analisys

Lab 2.2

Lab consists of two parts: classification and business analysis

- Part I. Data mining: build the classifier and use it for the prediction of potential responders
- Part II. Business analytics: how to design the most profitable campaign

Plan

Part I. Data Mining. Classification with WEKA.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Part II. Business analysis

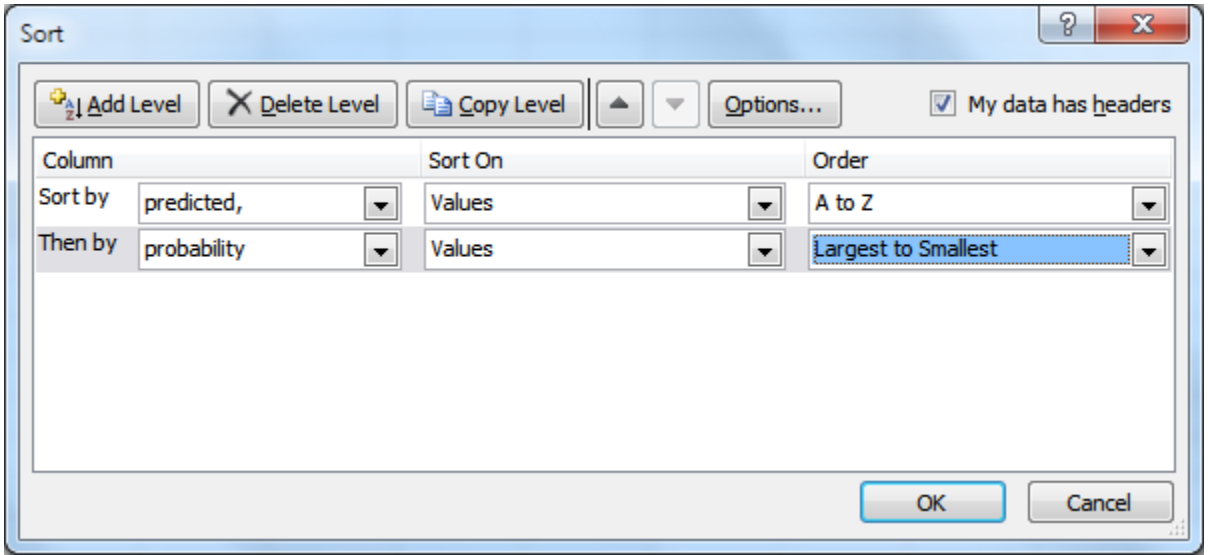
1. Generate Lift chart(s)
2. Cost-benefit analysis
3. Recommendations

Part II. Business analysis

- Here we want to calculate our future possible revenues – what can we get by applying our learned classification model

Ranking predictions

- Copy results to a new Sheet: Lift charts
- Leave only columns inst#, actual, predicted, probability:
- Select data and sort:
  - First yes, then no, and within each group in descending order of probabilities (of buying PEP)



Lift charts

- Lift – the ratio of the expected ‘Yes’ responses using the top of predicted list to the number of ‘Yes’ responses from the same number of random customers
- Lift measures the change in the concentration of a target class when the model is applied to the original dataset

Part I. Data Mining.

Part II. Business analysis

1. Generate Lift chart(s)
2. Cost-benefit analysis
3. Recommendations

Insert count of actual positive responses

Copy

inst#,	actual,	predicted,	probability	count resp
77	1:YES	1:YES	0.118	1
122	1:YES	1:YES	0.118	1
133	2:NO	1:YES	0.118	0

=IF(C2=$K$1,1,0)

1:YES K1

Add column *rank*: 1,2,3…

This is just to have a reference to a string constant (1:YES), used in the formula

This formula means: if the value in column 'actual,' equals '1:YES' then put 1, otherwise put 0.
$K$1 refers to a cell where we put our constant

Drag – expand to the rest of the cells in this column

Insert cumulative sum of actual positive responses

inst#,	actual,	predicted,	probability	count resp	cum sum
77	1:YES	1:YES	0.118	1	=1
122	1:YES	1:YES	0.118	1	=SUM($F$2:F3)
133	2:NO	1:YES	0.118	0	
18	1:YES	1:YES	0.083	1	
21	1:YES	1:YES	0.083	1	

Starting from the second row: add up 'count resp' counts starting from $F$2 up to the current row

Total 83 positive responses out of 180 mailings

Count cumulative percent of positive responses

rank	inst#,	actual,	predicted,	probability	count resp	cum sum	cum percent
1	77	1:YES	1:YES	0.118	1	1	1.204819
2	122	1:YES	1:YES	0.118	1	2	2.409639
3	133	2:NO	1:YES	0.118	0	2	2.409639
4	18	1:YES	1:YES	0.083	1	3	3.614458
5	21	1:YES	1:YES	0.083	1	4	4.819277
6	57	1:YES	1:YES	0.083	1	5	6.024096

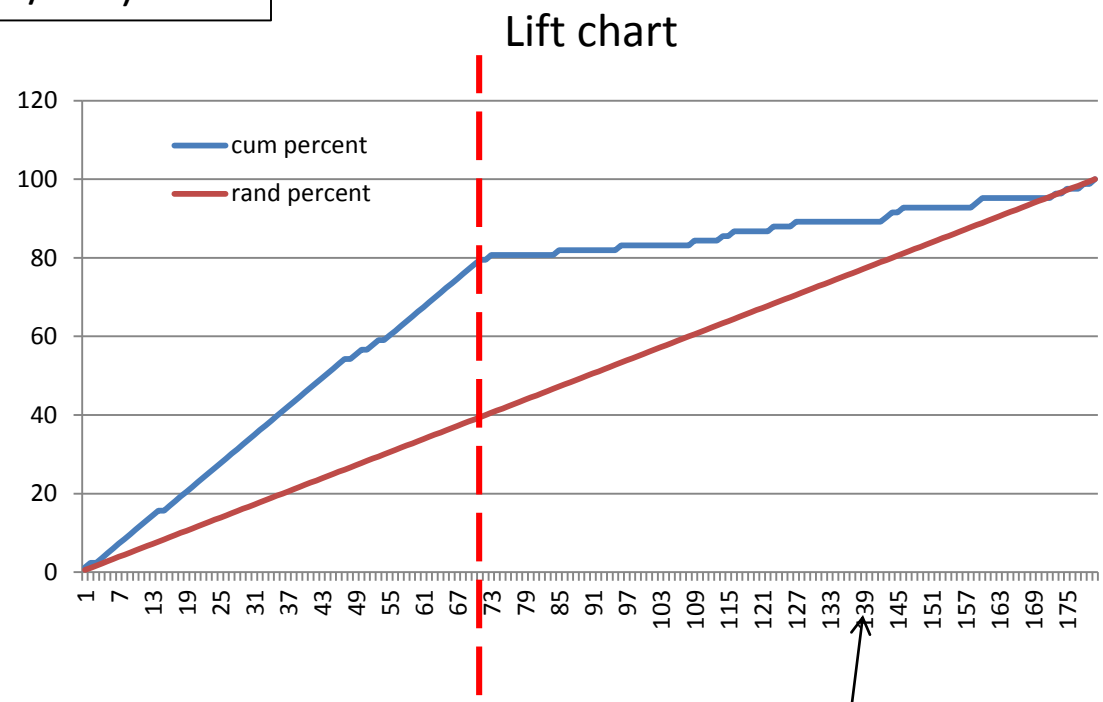
=(G2/83)*100

Total 83 positive responses out of 180

Add column for rand percent and generate lift chart

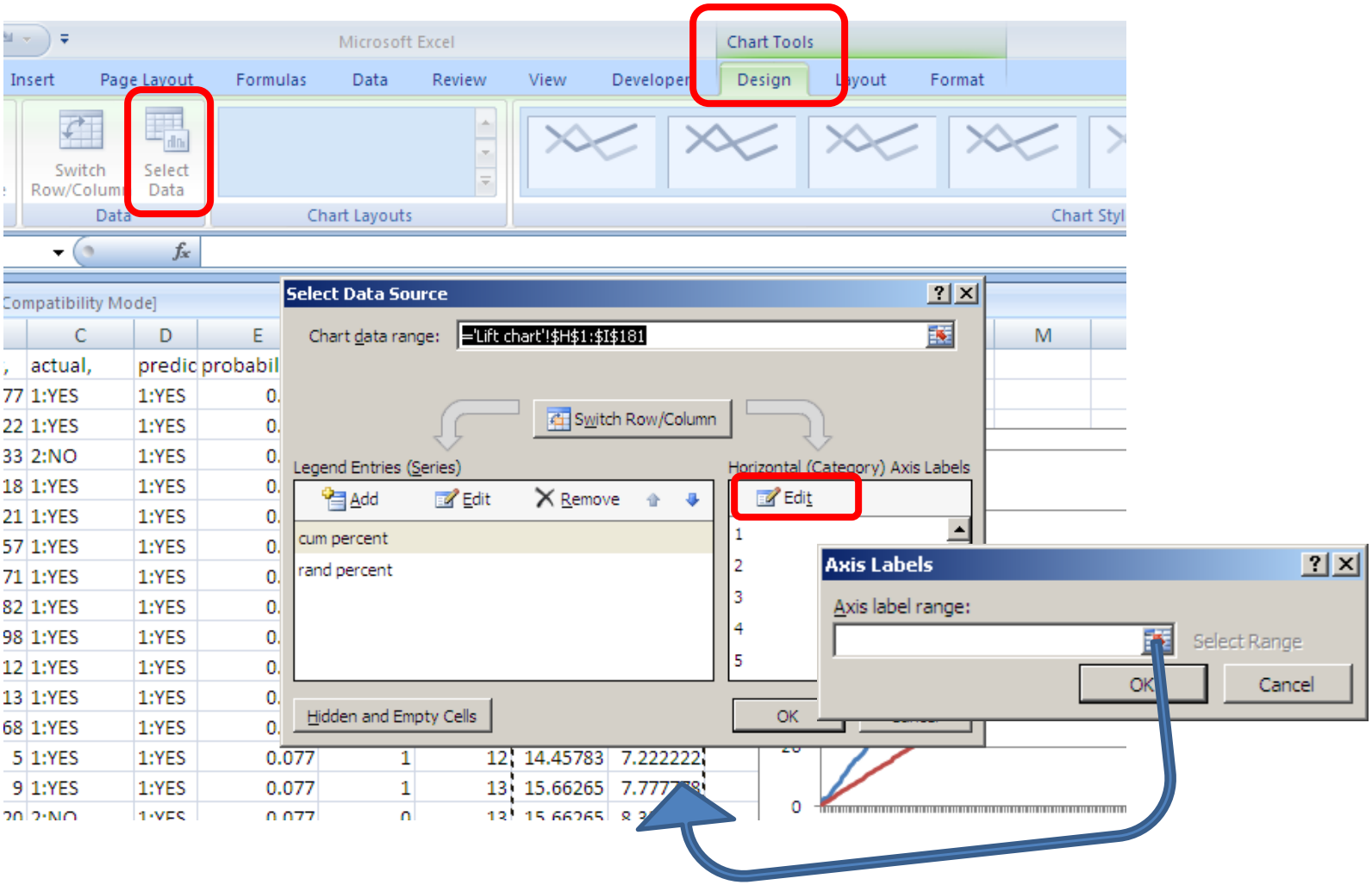
- Add column ‘rand percent’
- Create chart for ‘cum percent’ and ‘rand percent’

=(A2/180)*100



We need to re-label the default scale in the horizontal axis: to the percent of targeted customers (i.e. the same as the ‘rand percent’)

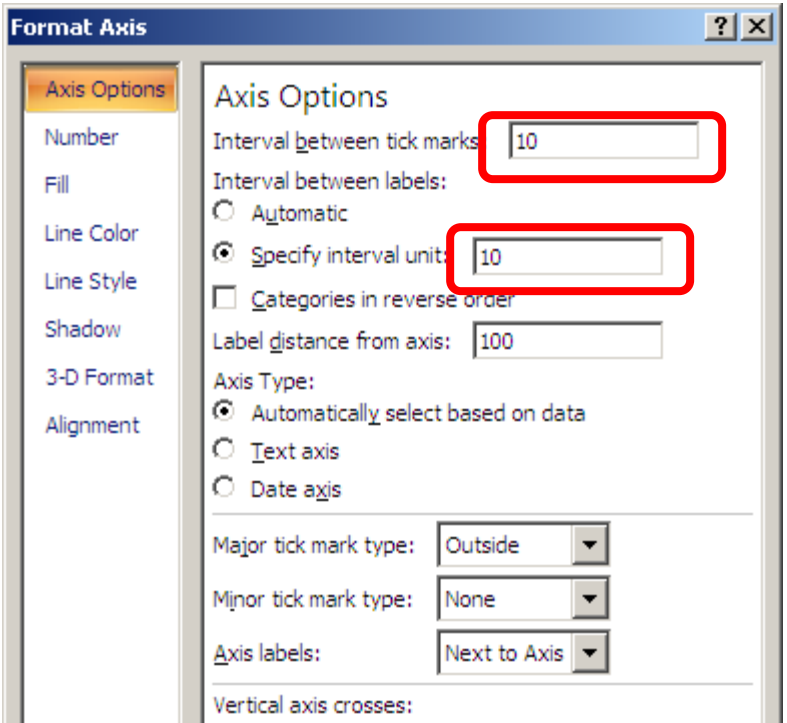
Lift chart: X-axis

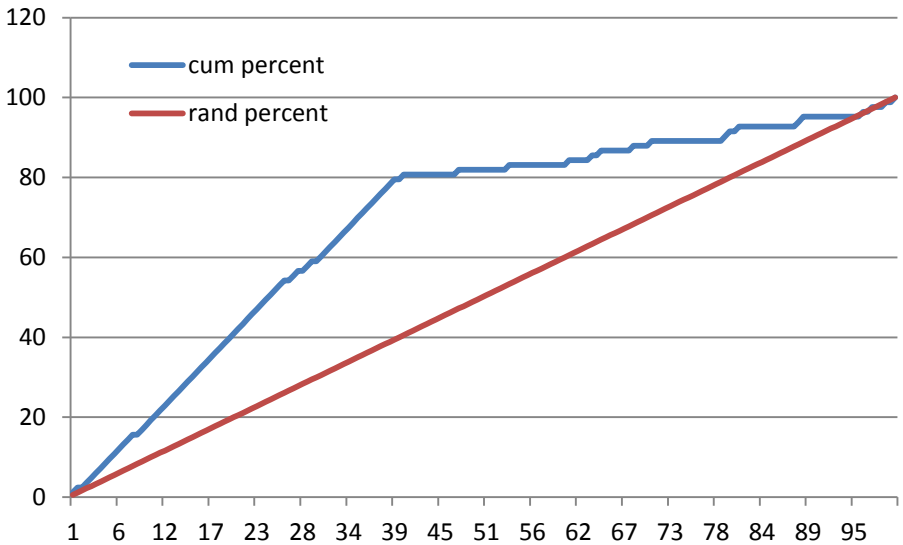


Select column ‘rand percent’ to label X –axis (without the column title)

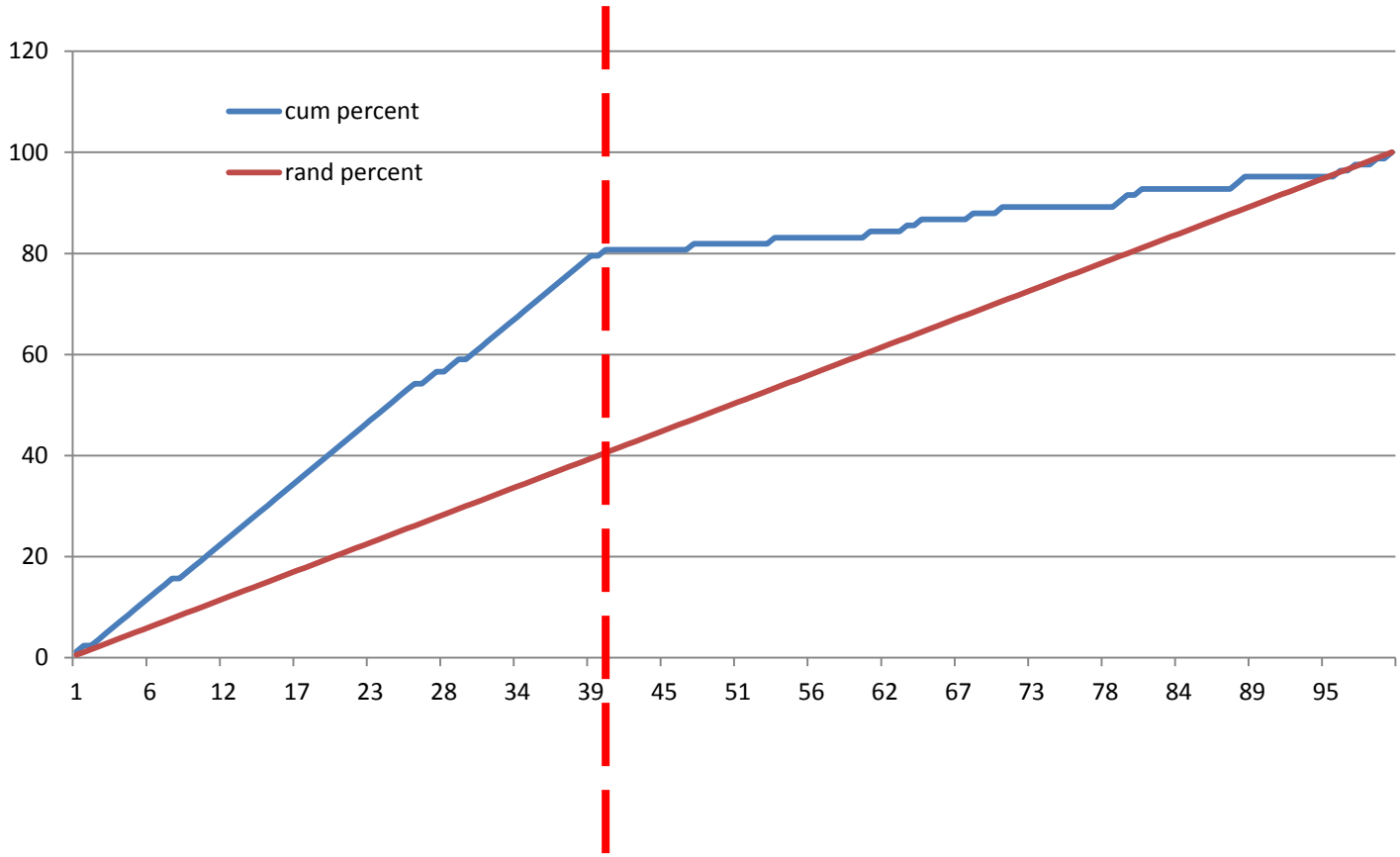
X-axis labels: percent of mailings out of total 180 customers

- Format values in 'rand percent' column: to have 0 decimal points
- You can also format axis to make interval unit 10%





The resulting Lift chart



By sending mails to only 40% of the customers from the top of the ranked list, we may cover 80% of potential responders (80% out of 83 = 66 customers). Note that we cannot cover all 100% of customers unless we send letters to everybody. Our model does not give 100% correct answers

Optimal number of letters

- How to choose the optimal number of letters to send?

- This depends on the:
  - Cost of each mail: for example, let it be $5
  - Benefit from each accepted PEP: let it be $10

- These unrealistic numbers are chosen because of the small size of the validation dataset – to demonstrate the concept of maximum profit
- In real life, the cost of mailing is several cents, and the benefit is hundreds and thousands dollars, but the datasets of mailings contain millions of records

Part I. Data Mining.

Part II. Business analysis

1. Generate Lift chart(s)
2. Cost-benefit analysis
3. Recommendations

Cost-benefit analysis

- Copy data to a new Sheet “cost analysis”
- Insert column rand_sum: =(A2/180)*83

$O$1=5

$P$1=10

Unit costs for mailing and for revenues

- Insert column cost: I2=A2*$O$1
- Insert column benefit: J2=G2*$P$1
- Insert column expected revenue: =J2-I2
- Expand all columns to entire columns

Results

					cum sum	rand sum	cost	benefit	exp revenue
					1	0.461111	5	10	5
					2	0.922222	10	20	10
							...		
70	161	1:YES	1:YES	0	1	65	32.27778	350	650	300
71	174	1:YES	1:YES	0	1	66	32.73889	355	660	305
72	176	2:NO	1:YES	0	0	66	33.2	360	660	300
73	180	1:YES	1:YES	0	1	67	33.66111	365	670	305
74	13	2:NO	2:NO	1	0	67	34.12222	370	670	300
75	32	2:NO	2:NO	1	0	67	34.58333	375	670	295
76	34	2:NO	2:NO	1	0	67	35.04444	380	670	290
77	35	2:NO	2:NO	1	0	67	35.50556	385	670	285
78	63	2:NO	2:NO	1	0	67	35.96667	390	670	280

The optimal number of letters

- Maximizes the revenue:

  71-73 letters ~ $300 profit

Note: this profit cannot be achieved by random mailing

End of task