DSC 433/533 – Class 14 – Nearest neighbors example for Direct Mail case

This class exercise is based on the **Direct Mail Fundraising** Case (see separate document on the handouts page of the course website) and the Excel dataset **Donor.xls** (available on the data page of the course website).

1. The **Donor.xls** has already been partitioned into Training (1872), Validation (1248), and Test (2000) samples. Do a k-Nearest Neighbors analysis with k=1, using variables zip_1, numchld, numprom, maxramnt, lastgift, totmonths, and avggift to classify whether a former donor made a donation in response to the current campaign (target_b=1).
   - Select XLMiner > Classification > k-Nearest Neighbors.
   - Move zip_1, numchld, numprom, maxramnt, lastgift, totmonths, and avggift to the “Input variables” box and target_b to the “Output variable” box.
   - Hit “Next.”
   - Check “Normalize input data” and make sure “Number of nearest neighbors (k)” is 1 and “Scoring option” is selected to be “Score on specified value of k as above.”
   - Select “Summary report” for “Score training data.”
   - Select “Detailed scoring,” “Summary report,” and “Lift charts” for “Score validation data.”
   - De-select “Summary report” for “Score test data.”

   The estimated probability of donating for each individual is based on the proportion of nearest neighbors who donated. You can find these estimated probabilities on the “KNNC_ValidScore1” worksheet in the column headed “Prob. for 1 (success)” next to the column headed “Actual Class” which shows whether the individual donated or not.
   - Select and copy these two columns (from E7 to F1255), and paste them into the O and P columns.
   - Select Data > Sort to sort the O and P columns in descending order using “Prob. for 1 (success)” in column P to sort on.

   You should discover that 650 individuals in the validation sample have an estimated donation probability of 1, of which 342 actually made a donation. Similarly, 5 individuals in the validation sample have an estimated donation probability of 0.67, of which 2 actually made a donation, and so on.

2. We can use this information to construct a Lift Chart for *donations* for the Validation sample. Use the following table typed into Excel (or just copy and paste it):

	0	342	344	346	347	350	352	621
Cumulative solicitations sent	0	650	655	659	671	680	684	1248
k-NN analysis with k=1	0	34.9	35.1	35.3	35.4	35.7	35.9	63.3
Simple random sample	0	33.0	33.2	33.4	34.1	34.5	34.7	63.3

   The numbers in the top row of the table represent the cumulative number of donations when solicitations are sent to 0/650/655/659/671/680/684/1248 individuals. The third row consists of re-weighted results for the k-NN analysis with k=1 above (since actual response rate is 0.051 and the oversampled training/validation response rate is 0.50, we need to multiply the numbers of donations by 0.051/0.50, e.g., 342*0.051/0.50 = 34.9). **Note that you do NOT have to do any re-weighting for homework 7 based on the Charles Book Club case since there is no over-sampling in that dataset.** The final row consists of expected results when sending out solicitations at random. For example, we would expect (621*(0.051/0.50)/1248)*650 = 33.0 donations if sending out 650 solicitations at random.

   Then highlight the whole table, start the “Chart Wizard” by clicking on the chart icon (), select “XY (Scatter)” for the Chart type, and select “Scatter with data points connected by lines” for the Chart sub-type.

3. We can also construct a Lift Chart for *profit* for the Validation sample. From the case description, it costs $0.68 to send a solicitation to a former donor and, out of those that respond, the average donation is $13.00. Use the following table typed into Excel (or just copy and paste it):

Cumulative solicitations sent	0	650	655	659	671	680	684	1248
k-NN analysis with k=1	0	11.5	10.7	10.7	3.8	1.7	1.6	-25.2
Simple random sample	0	-13.1	-13.2	-13.3	-13.5	-13.7	-13.8	-25.2

The numbers in the table represent the total profit (in $) when solicitations are sent to 0/650/655/659/671/680/684/1248 individuals for the k-NN analysis with k=1 above and the expected results when sending out solicitations at random. For example, we would expect profit of 342*(0.051/0.50)*13 – 650*0.68 = $11.5 if we only send out solicitations to the first 650 individuals in the k-NN analysis with k=1. **Note that you do NOT have to do any re-weighting for homework 7 based on the Charles Book Club case since there is no over-sampling in that dataset.** Similarly, we would expect profit of (621*(0.051/0.50)/1248)*650*13 – 650*0.68 = –$13.1 if sending out 650 offers at random. If we send out offers to all 1248 individuals we would expect our profit to be 621*(0.051/0.50)*13 – 1248*0.68 = –$25.2.

Then highlight the whole table, start the “Chart Wizard” by clicking on the chart icon, select “XY (Scatter)” for the Chart type, and select “Scatter with data points connected by lines” for the Chart sub-type.

4. Can we do better using a k-Nearest Neighbors analysis with k=4, using the same 7 variables and instructions from part 1?

Donations:

	0	61	251	254	258	258	447	449	452	453	455	457	580	582	621
Cumulative solicitations sent	0	112	433	439	447	448	855	859	868	873	877	882	1165	1168	1248
k-NN analysis with k=4	0	6.2	25.6	25.9	26.3	26.3	45.6	45.8	46.1	46.2	46.4	46.6	59.2	59.4	63.3
Simple random sample	0	5.7	22.0	22.3	22.7	22.7	43.4	43.6	44.1	44.3	44.5	44.8	59.1	59.3	63.3

Profit:

Cumulative solicitations sent	0	112	433	439	447	448	855	859	868	873	877	882	1165	1168	1248
k-NN analysis with k=4	0	4.7	38.4	38.3	38.1	37.5	11.3	11.3	9.1	7.0	7.0	6.2	-23.1	-22.5	-25.2
Simple random sample	0	-2.3	-8.7	-8.9	-9.0	-9.0	-17.3	-17.3	-17.5	-17.6	-17.7	-17.8	-23.5	-23.6	-25.2

5. Can we do better using a k-Nearest Neighbors analysis with k=10, using the same 7 variables and instructions from part 1?

Donations:

	0	1	23	68	68	68	151	153	155	158	159	302	303	303	305	305
Cumulative solicitations sent	0	3	32	101	102	103	257	259	264	269	270	526	527	528	530	531
k-NN analysis with k=10	0	0.1	2.3	6.9	6.9	6.9	15.4	15.6	15.8	16.1	16.2	30.8	30.9	30.9	31.1	31.1
Simple random sample	0	0.2	1.6	5.1	5.2	5.2	13.0	13.1	13.4	13.7	13.7	26.7	26.7	26.8	26.9	27.0

	440	441	443	445	525	527	527	528	530	530	581	582	608	608	621	621
Cumulative solicitations sent	788	789	792	797	1011	1015	1017	1020	1025	1027	1163	1164	1224	1225	1246	1248
k-NN analysis with k=10	44.9	45.0	45.2	45.4	53.6	53.8	53.8	53.9	54.1	54.1	59.3	59.4	62.0	62.0	63.3	63.3
Simple random sample	40.0	40.0	40.2	40.5	51.3	51.5	51.6	51.8	52.0	52.1	59.0	59.1	62.1	62.2	63.2	63.3

Profit:

Cumulative solicitations sent	0	3	32	101	102	103	257	259	264	269	270	526	527	528	530	531
k-NN analysis with k=10	0	-0.7	8.7	21.5	20.8	20.1	25.5	26.8	26.0	26.6	27.2	42.8	43.4	42.7	44.0	43.4
Simple random sample	0	-0.1	-0.6	-2.0	-2.1	-2.1	-5.2	-5.2	-5.3	-5.4	-5.5	-10.6	-10.6	-10.7	-10.7	-10.7

Cumulative solicitations sent	788	789	792	797	1011	1015	1017	1020	1025	1027	1163	1164	1224	1225	1246	1248
k-NN analysis with k=10	47.6	48.2	48.9	48.1	8.7	8.6	7.2	6.5	5.8	4.4	-20.4	-19.8	-26.1	-26.8	-23.8	-25.2
Simple random sample	-15.9	-15.9	-16.0	-16.1	-20.4	-20.5	-20.5	-20.6	-20.7	-20.7	-23.5	-23.5	-24.7	-24.7	-25.2	-25.2