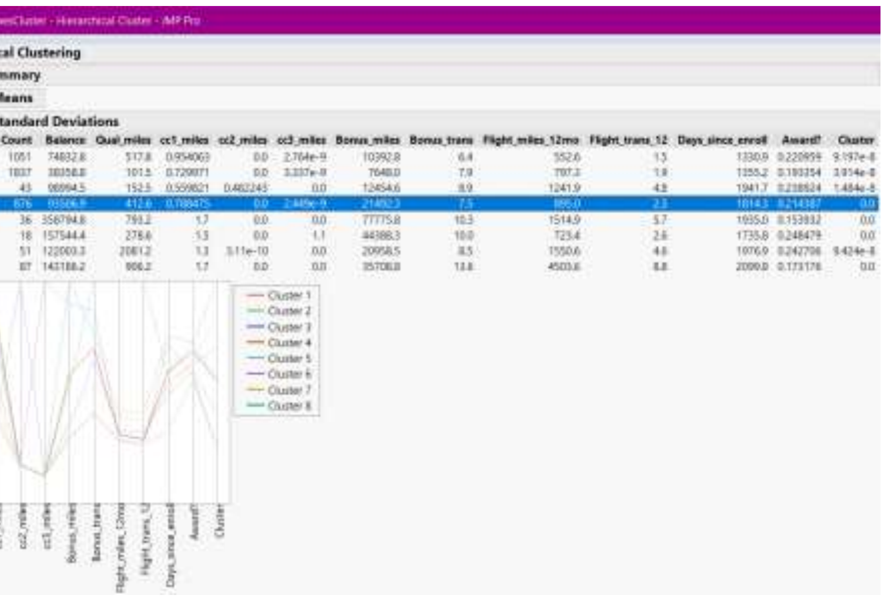
Sebastián Pastor

Data Mining – Assignment 04

04/30/2020

Question 01: Clustering


Standard Deviations

Count	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?	Cluster
1051	74832.8	517.8	0.954063	0.0	2.764e-9	10392.8	6.4	552.6	1.5	1330.9	0.220959	9.197e-8
1837	38358.8	101.5	0.729871	0.0	3.337e-8	7648.0	7.9	787.3	1.8	1355.2	0.193354	3.914e-8
43	98994.5	152.5	0.559821	0.482245	0.0	12454.6	8.9	1241.9	4.8	1941.7	0.238924	1.484e-8
676	93506.9	412.6	0.788475	0.0	2.449e-9	21492.3	7.5	895.0	2.3	1814.3	0.214387	0.0
36	358794.8	793.2	1.7	0.0	0.0	77775.8	10.3	1514.9	5.7	1935.0	0.153932	0.0
18	157544.4	278.6	1.5	0.0	1.1	44388.3	10.0	723.4	2.6	1735.8	0.248479	0.0
51	122003.3	2081.2	1.3	3.11e-10	0.0	20958.5	8.5	1550.6	4.6	1976.9	0.242798	9.424e-8
87	143188.2	906.2	1.7	0.0	0.0	35708.8	13.8	4503.6	8.8	2090.0	0.173176	0.0

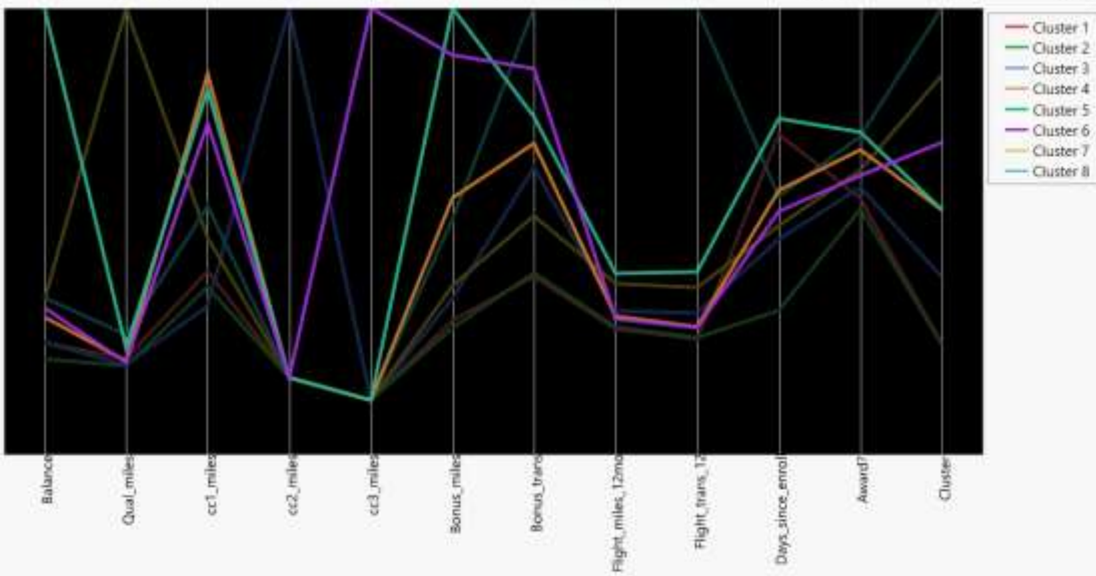
a. Apply hierarchical clustering and Ward's method. Make sure to standardize the data. Use the dendrogram and the scree plot, along with practical considerations, to identify the ''best'' number of clusters. How many clusters would you select? Why?

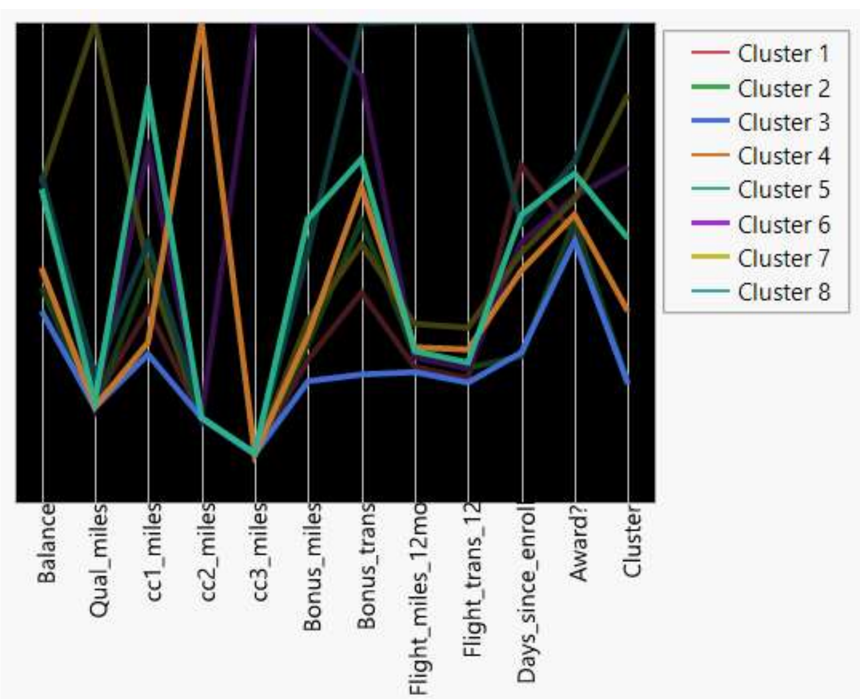
Between 4-6 clusters seems to be the best amount in this model given the clusters standard deviations.

b. What would happen if the data were not standardized?
Given that the features come in different scales, had the data not been standardized, the cluster distances would have been skewed significantly towards features such as balance and days since enrollment, which have higher frequencies given their particular scales.

c. Explore the clusters to try to characterize them. Try to give each cluster a label. i. Compare the cluster centroids (select Cluster Summary) and click on the lines to characterize the different clusters. ii. Save the clusters to the data table and use graphical tools and the Column Switcher.

Cluster	Count	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?	Cluster
1	1051	74832.8	517.8	0.954063	0.0	2.764e-9	10392.8	6.4	552.6	1.5	1330.9	0.220959	9.197e-8
2	1837	38358.8	101.5	0.729971	0.0	3.337e-9	7648.0	7.9	797.3	1.9	1355.2	0.193354	3.914e-8
3	43	98994.5	152.5	0.559821	0.482243	0.0	12454.6	8.9	1241.9	4.8	1941.7	0.238924	1.484e-8
4	876	93506.9	412.6	0.788475	0.0	2.449e-9	21492.3	7.5	895.0	2.5	1814.3	0.214307	0.0
5	36	358794.8	793.2	1.7	0.0	0.0	77775.8	10.3	1514.9	5.7	1935.0	0.153932	0.0
6	18	157544.4	278.6	1.5	0.0	1.1	44388.3	10.0	723.4	2.6	1735.8	0.248479	0.0
7	51	122003.3	2081.2	1.3	3.11e-10	0.0	20958.5	8.5	1550.6	4.6	1976.9	0.242706	9.424e-8
8	87	143186.2	906.2	1.7	0.0	0.0	35708.8	13.6	4503.6	8.8	2099.8	0.173176	0.0

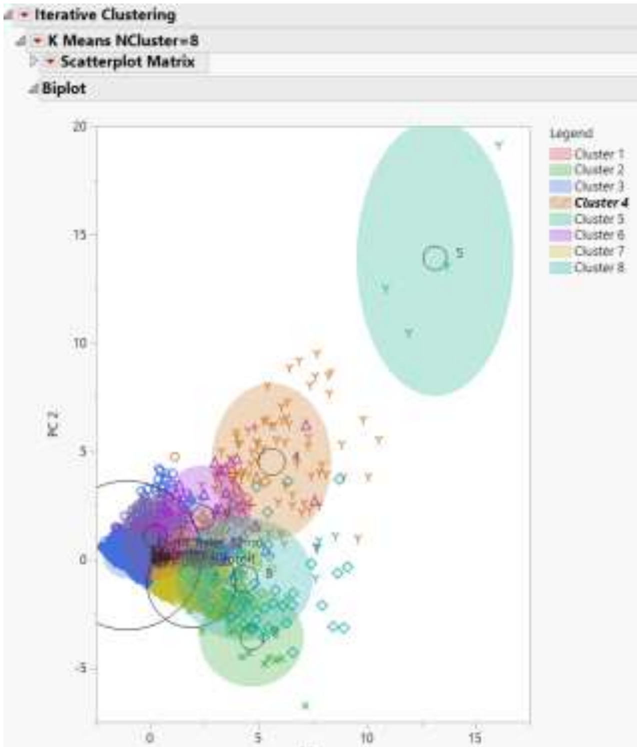




d. To check the stability of the clusters, hide and exclude a random 5% of the data (you can partition using a random seed of 4279), and repeat the analysis. Does the same picture emerge?

Hierarchically, the data looks identical. However, the clusters are not as defined as previously. Cluster 6 suffers the most, but clusters 5 and 4 remain constant. This tells us that perhaps cluster 5 is best.

e. Use k -means clustering with the number of clusters that you found above. How do these results compare to the results from hierarchical clustering? Use the built-in graphical tools to characterize the clusters.



Delta for clusters 4 and 5 visualized based on K-Means. We can see that results are very similar to those of the original hierarchical cluster we ran.

Hierarchical:

Cluster Means

Cluster	Count	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?
1	1051	68948	126.4	0.568982	-2.4e-16	3.42e-16	8222.0	8.0	213.0	0.658421	3913	0.514173
2	1837	39890	9.9	0.374524	-3e-16	4.86e-16	6064.9	8.3	257.6	0.731628	2655	0.467092
3	43	68877	23.3	0.139535	1.3	-6.9e-18	14689.8	17.5	582.6	2.209302	3969	0.561279
4	876	113589	87.4	3.0	-2.2e-16	2.74e-16	44234.4	19.7	475.4	1.398886	4878	0.723188
5	36	656459	298.1	2.8	3.47e-18	-6.9e-18	100048.4	22.0	1347.4	4.638889	6204	0.799638
6	18	129951	65.7	2.4	-5.2e-18	2.72	86259.9	26.2	422.2	1.333333	4489	0.611821
7	51	143288	5715.0	0.980392	6.94e-18	-1e-17	18424.2	13.3	1141.6	3.705882	4223	0.635495
8	87	147640	509.0	1.4	5.2e-18	-1.2e-17	37982.9	31.5	6747.1	20	4764	0.781133

K-Means:

Cluster Means

Cluster	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?
1	68876.5814	23.255814	1.13953488	2.34883721	1	14689.8372	17.5348837	582.627907	2.20930233	3968.93023	0.39534884
2	138061.4	78.8	3.46666667	1	4.06666667	93927.8667	28.0666667	506.666667	1.6	4613.86667	0.53333333
3	47779.7359	49.1756757	1.34623813	1	1.00036523	5734.54091	7.64828342	232.297663	0.6844412	3779.3382	0.22388605
4	130878.753	493.741573	2.25842697	1	1	33430.8989	29.3258427	6171.94382	18.4044944	4744.39326	0.83146067
5	131999.5	347	2.5	1	1	65634.25	69.25	19960	49.25	2200.25	1
6	124780.189	5712.33962	2	1	1	18436.2453	12.4716981	1003.15094	3.05660377	3993.84906	0.52830189
7	97384.5876	69.9721649	3.92061856	1	1.00206186	41901.2309	19.4175258	382.445361	1.14226804	4844.26186	0.68659794
8	519834.149	270.471264	3.7816092	1	1	68520.5057	21.8045977	1353.98851	4.54022989	6215.62069	0.81609195

Cluster Standard Deviations

f. **Which clusters would you target for offers, and what types of offers would you target to customers in that cluster?**

Cluster Summary

Cluster	Count
1	43
2	15
3	2738
4	89
5	4
6	53
7	970
8	87

Step	Criterion
16	0

The K-Means cluster pictured to the left offers a smaller count, therefore I would use the hierarchical clusters to target offers.
Cluster Number 4 has a high amount of earnings when using their flight rewards credit card. However, they have a low level of flight miles and a low level of flight transactions. So offers in those areas might appease to them to engage more in order to earn awards.

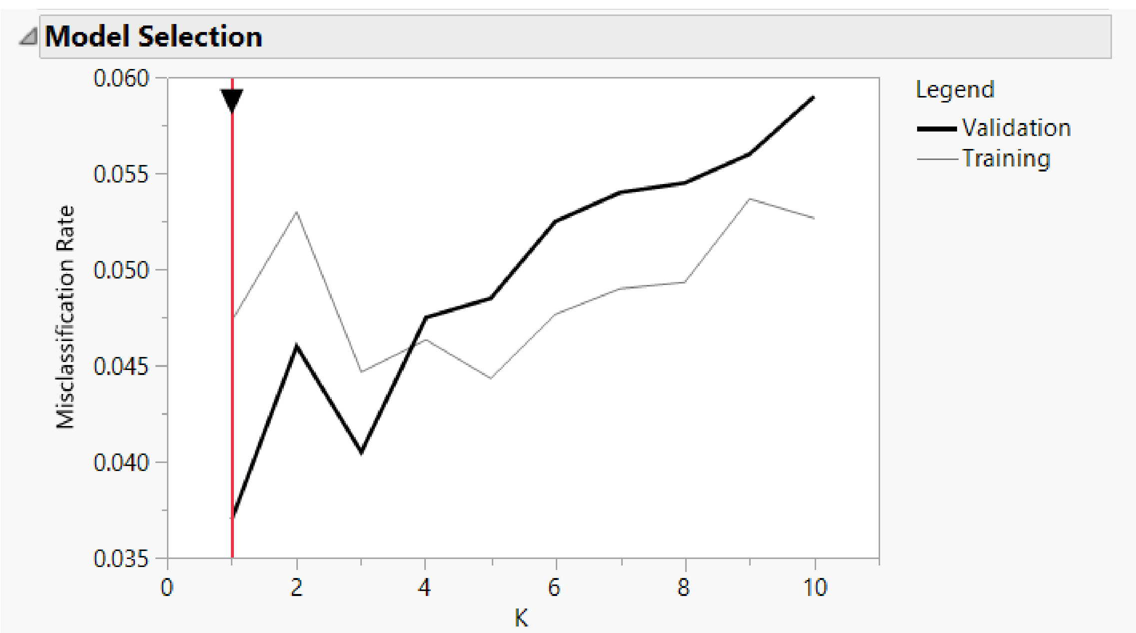
Question 2: K – Nearest Neighbors

Partition the data into training (60%) and validation (40%) sets using a random seed of 4279. a. Consider the following customer: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education = 2, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit Card = 1. Perform a k-NN classification with all predictors except ID and ZIP code using k = 10. How would this customer be classified? (Note: This analysis may take a few minutes.)

This customer would be classified as not receiving the offer.

b. What is a choice of k that balances between overfitting and ignoring the predictor information?

For the model created, the balance seems to lie at 1.



c. Show the classification matrix for the validation data that results from using the best k.

Confusion Matrix for Best K=1

Training

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	173	109
No	33	2685

Validation

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	148	50
No	24	1778

d. Consider the following customer: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education = 2, Mortgage = 0, Securities Account = 0, CDAccount = 0, Online = 1 and Credit Card = 1. Classify the customer using the best k.

This person would not qualify for a loan.

e. Repartition the data, this time into training, validation, and test sets (50%, 30%, 20%). Apply the k – NN method with the k chosen above. Compare the classification matrix of the test set with that of the training and validation sets. Comment on the differences and their reason.

Confusion Matrix for Best K=1

Training

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	173	109
No	33	2685

Validation

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	148	50
No	24	1778

Confusion Matrix for Best K=1

Training

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	167	74
No	29	2231

Validation

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	105	49
No	17	1328

Test

Actual Personal Loan	Predicted Count Yes	Predicted Count No
Yes	52	33
No	13	902

The test column gave us less true positives, which is what we are looking for. Given that test has less true positives and more false negatives, we get a weaker matrix.

Question 3: Naïve Bayes Classifier

a. Check to make sure that the variables are coded as Nominal, and that none of the variables has the Value Labels column property (remove this column property if needed).

b. Create a summary of the data using Tabulate, with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the cells should convey the count (how many records are in that cell).

		Online	
Personal Loan	CreditCard	0	1
1	0	128	209
	1	61	82
0	0	1300	1893
	1	527	800

c. Consider the task of classifying a customer that owns a bank credit card and is actively using online banking services. Looking at the tabulation, what is the probability that this customer will accept the loan offer? (This is the probability of loan acceptance (Loan = 1) conditional on having a bank credit card (CC = 1) and being an active user of online banking services (Online = 1)).

82/(209+82+128+61) = 82/480 = (0.17 * 480/5000) / (882/5000) = **92.3%**

d. Create two tabular summaries of the data. One will have Loan (rows) as a function of Online (columns) and the other will have Loan (rows) as a function of CC .

	Online	
Personal Loan	0	1
1	189	291
0	1827	2693

||

	CreditCard	
Personal Loan	0	1
1	337	143
0	3193	1327

e. Compute the following quantities [P ($A|B$) means "the probability of A given B"]: i. P (CC = 1|$Loan$ = 1) (the proportion of credit card holders among the loan acceptors) ii. P ($Online$ = 1|$Loan$ = 1) iii. P ($Loan$ = 1) (the proportion of loan acceptors) iv. P (CC = 1|$Loan$ = 0) v. P ($Online$ = 1|$Loan$ = 0) vi. P ($Loan$ = 0)

- 143/(337+143) = 29.8%

- 291/480 = 60.6%

- 480/5000 = 9.6%

- 2693/4520 = 59.6%

- 4520/5000 = 90.4%

f. Use the quantities computed above to compute the naive Bayes probability P ($Loan$ = 1|CC = 1, $Online$ = 1).

(480/5000 * .298 * .606) / (882/5000) = 98%

g. Compare this value with the one obtained from the tabulation in (b). Which is a more accurate estimate of P ($Loan$ = 1|CC = 1, $Online$ = 1)?

The Naïve Bayes is slightly less accurate since it is a generalized calculation with no precise values.

h. Which of the entries in this table are needed for computing P ($Loan$ = 1|CC =1, $Online$ = 1)? In JMP, use Naive Bayes to compute the probability that P ($Loan$ = 1|CC = 1, $Online$ = 1). Compare this to the number you obtained in (e).

Naive Bayes

Fit Details

Personal Loan

Training

Count	Misclassification Rate	Misclassifications
3000	0.09400	282

Validation

Count	Misclassification Rate	Misclassifications
2000	0.09900	198

Confusion Matrix

Training

Actual Personal Loan	Predicted Count 1	0
1	0	282
0	0	2718

Validation

Actual Personal Loan	Predicted Count 1	0
1	0	198
0	0	1802

ROC Curve for Personal Loan = 1

Prediction Profiler