Schulich School of Business
York University

MMAI 5040: Business Application of AI 1 (Winter 2023)

Individual Assignment 1 (60 points) – Worth 10% of Final Grade

Instructions:

1. This assignment is to be done individually. Submit your work in two files: 1) a word or pdf file containing your answer to each question, and 2) a Jupyter notebook file containing your python code and outputs.

2. Do not leave any question unanswered in the word or pdf file. That is, do not refer the instructor/grader to outputs in your python files. Rather, make sure to present all relevant outputs and explanations as part of your answers in the word or pdf file.

3. Name the files as follows: "MMAI5040_W23_Student_Name_Assignment1"

Grading will be based:

- Accuracy, clarity and precision of conceptual arguments and strategies.
- Correctness (rather than quality) of implementation in Python
- Originality in discussions of the business implication of outputs.
- Organization and presentation

Due Mon Feb 27 @ 11.59pm. The usual penalty of 20% daily late submission applies.

Q1. Exploring Healthy Cereals [20 points] The Breakfast Cereal data (*cereals.csv*) contains nutritional information and consumer rating of 77 breakfast cereals. The consumer rating is a rating of cereal "healthiness" for consumer information (not a rating provided by consumers). Cereal information is based on a bowl of cereal rather than a serving size because most people simply fill a cereal bowl (resulting in constant volume, not weight). The data include 13 numerical variables for each cereal. (Note: Cereals.csv is a publicly available dataset. You may consult public sources online, e.g., http://lib.stat.cmu.edu/datasets/1993.expo/ for more information about the dataset and variable descriptions).

a. Explore and summarize the data as follows:

i. Which variables are numerical? Which are ordinal? Which are nominal? (1.5pts)

numerical data: Calories, Protein, Fat, Sodium, Fiber, Carbo, Sugars, Potass, Vitamins, Weight, Cups, Rating

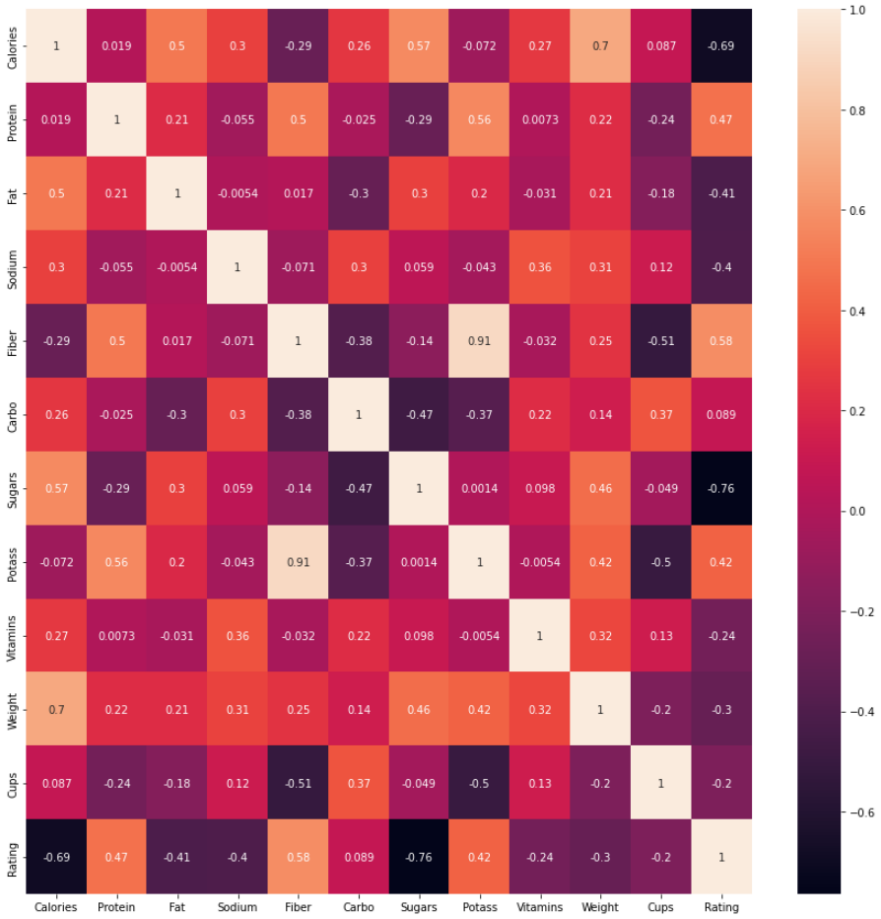
ordinal data: Shelf

nominal data: Cold, Nabisco, Quaker, Kelloggs, GeneralMills, Ralston, AHFP, Name, Manuf, Type

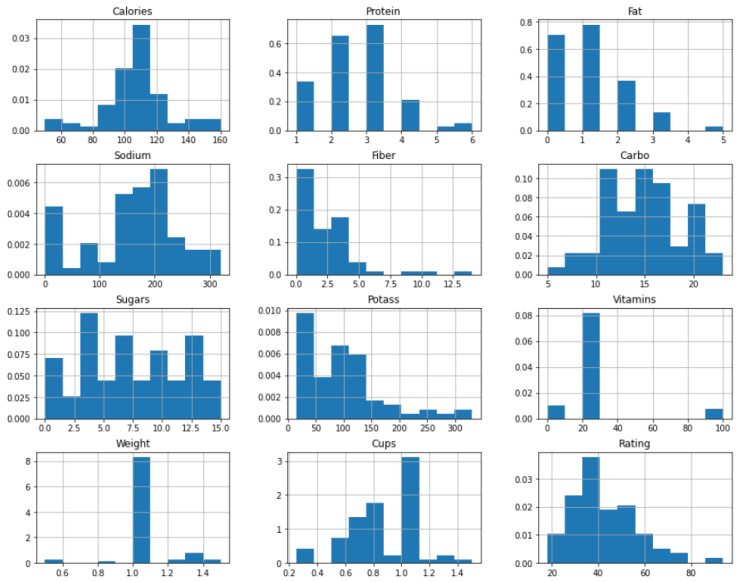
ii. Compute the mean, median, min, max, standard deviation, the correlation table, and a matrix plot for the numerical variables. (2pts).

	mean	median	min	max	standard deviation
Calories	0.114325	0.117501	0.053410	0.170911	0.021197
Protein	0.106982	0.106407	0.042563	0.255377	0.045789
Fat	0.082199	0.082199	0.000000	0.410997	0.082761
Sodium	0.103712	0.114976	0.000000	0.204402	0.052870
Fiber	0.077952	0.071657	0.000000	0.501602	0.086827
Carbo	0.112441	0.110687	0.038168	0.175573	0.029707
Sugars	0.099281	0.097771	0.000000	0.209509	0.060885
Potass	0.094580	0.086407	0.014401	0.316825	0.068049
Vitamins	0.092457	0.079556	0.000000	0.318223	0.070946
Weight	0.114998	0.111561	0.055780	0.167341	0.017115
Cups	0.111798	0.102052	0.034017	0.204104	0.032074
Rating	0.110426	0.104905	0.047022	0.244207	0.036574

	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sugars	Potass	Vitamins	Weight	Cups	Rating
Calories	1.000000	0.019066	0.498610	0.300649	-0.293413	0.257638	0.566533	-0.072063	0.265356	0.696091	0.087200	-0.689376
Protein	0.019066	1.000000	0.208431	-0.054674	0.500330	-0.025012	-0.291853	0.563706	0.007335	0.216158	-0.244469	0.470618
Fat	0.498610	0.208431	1.000000	-0.005407	0.016719	-0.300003	0.302497	0.200445	-0.031156	0.214625	-0.175892	-0.409284
Sodium	0.300649	-0.054674	-0.005407	1.000000	-0.070675	0.297687	0.058866	-0.042632	0.361477	0.308576	0.119665	-0.401295
Fiber	-0.293413	0.500330	0.016719	-0.070675	1.000000	-0.380357	-0.138760	0.911528	-0.032243	0.247226	-0.513061	0.584160
Carbo	0.257638	-0.025012	-0.300003	0.297687	-0.380357	1.000000	-0.471184	-0.365003	0.219202	0.138467	0.367460	0.088712
Sugars	0.566533	-0.291853	0.302497	0.058866	-0.138760	-0.471184	1.000000	0.001414	0.098231	0.455844	-0.048961	-0.763902
Potass	-0.072063	0.563706	0.200445	-0.042632	0.911528	-0.365003	0.001414	1.000000	-0.005427	0.419933	-0.501607	0.416009
Vitamins	0.265356	0.007335	-0.031156	0.361477	-0.032243	0.219202	0.098231	-0.005427	1.000000	0.320324	0.128405	-0.240544
Weight	0.696091	0.216158	0.214625	0.308576	0.247226	0.138467	0.455844	0.419933	0.320324	1.000000	-0.199583	-0.298124
Cups	0.087200	-0.244469	-0.175892	0.119665	-0.513061	0.367460	-0.048961	-0.501607	0.128405	-0.199583	1.000000	-0.203160
Rating	-0.689376	0.470618	-0.409284	-0.401295	0.584160	0.088712	-0.763902	0.416009	-0.240544	-0.298124	-0.203160	1.000000



iii. Plot a histogram for each of the numerical variables. Which variables have the largest variability? Which variables seem skewed? Are there any values that seem extreme? (1.5pts)

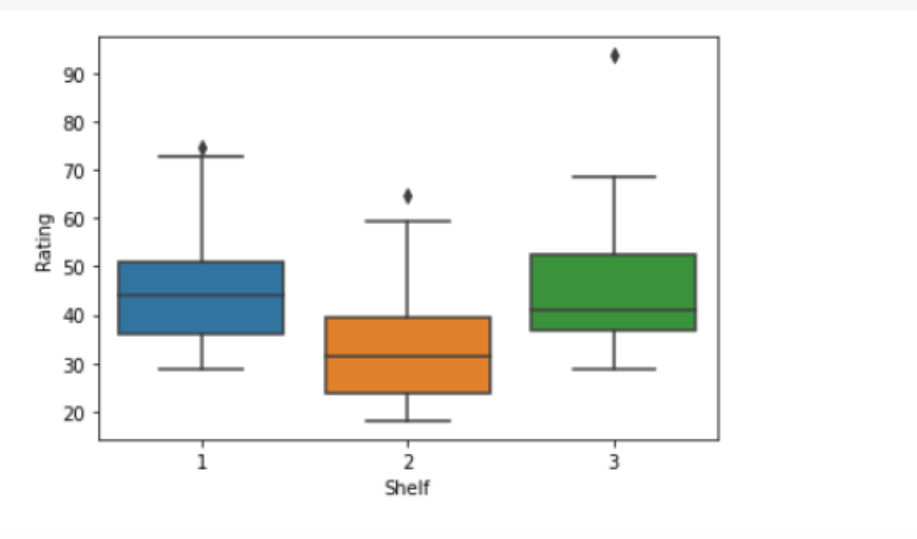


largest variability: Sugars

skewed: Protein, Fat, Potass and Rating seems right skewed; Cups, Sodium seems left skewed.

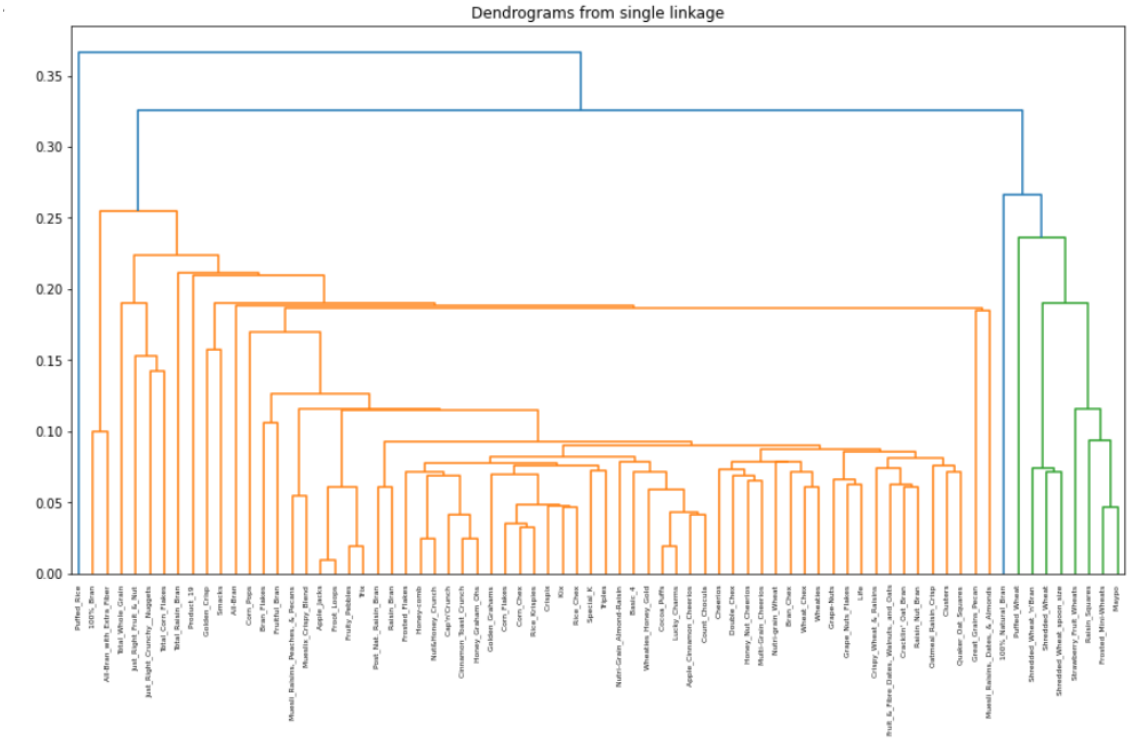
extreme: Weight

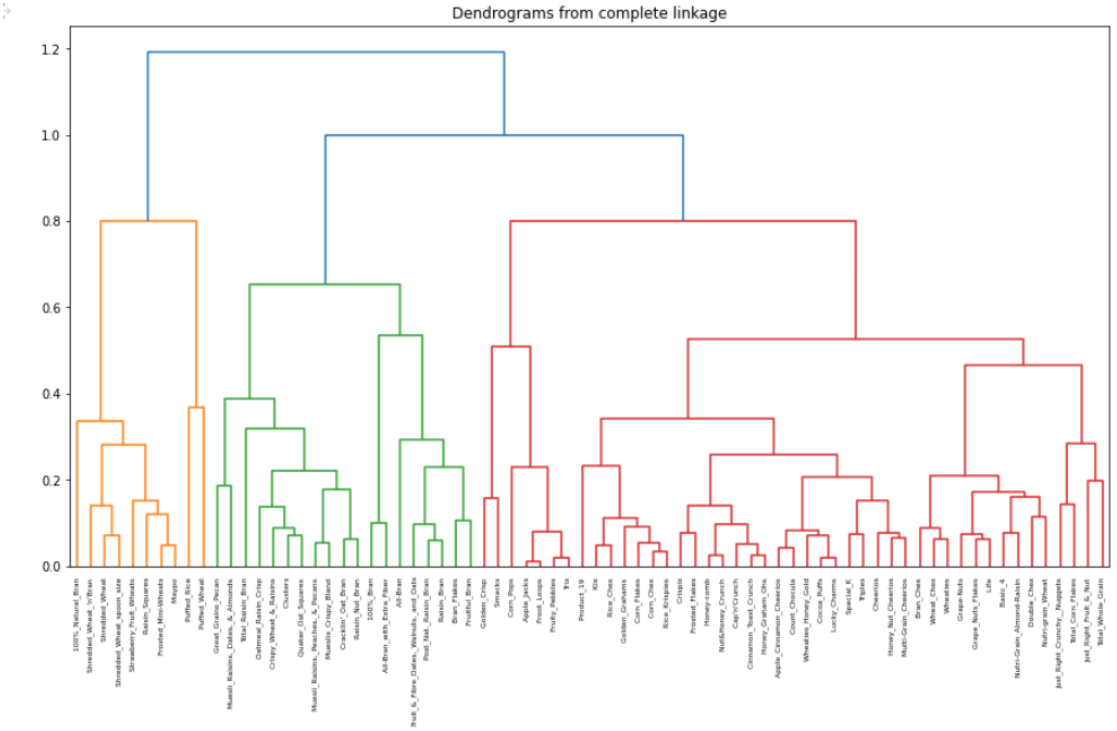
iv. Plot a side-by-side boxplot of consumer rating as a function of the shelf height. If we were to predict consumer rating from shelf height, would we need to keep all three categories of shelf height? (1.5pts)



Only need two categories of shelf height since 1 and 3 have the similar effect so pick either 1 and 2 or 2 and 3 would be enough to represent the relationship.

b. Apply hierarchical clustering to the data using Euclidean distance with normalized measurements. (Preprocess the data by removing all cereals with missing values).
   i. Compare the dendrograms from single linkage and complete linkage methods and look at cluster centroids. Comment on the structure of the clusters and their stability. (*Hint*: To obtain cluster centroids for hierarchical clustering, computer the average values of each cluster members, using *groupby( )* with the cluster centers followed by mean: `dataframe.groupby(clusterlabel).mean()`.) (5pts)





Complete Linkage Cluster Center:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.588019	0.017012	0.003364	0.018502	0.013489	0.103882	0.018120	0.618243	0.065938	0.016223	0.006041	0.006474	0.441386	0.006640	0.002534	0.002821	0.001285	0.000000	0.000000	0.00066
2	0.417845	0.011873	0.006084	0.567928	0.015435	0.045836	0.031459	0.644502	0.103392	0.010886	0.004102	0.002263	0.159885	0.003680	0.000162	0.000252	0.000982	0.001087	0.000401	0.00000
3	0.466109	0.009073	0.003580	0.788696	0.004659	0.064914	0.034486	0.249801	0.137329	0.008244	0.004344	0.003807	0.157640	0.004267	0.000000	0.000279	0.001473	0.001492	0.000413	0.00000

Single Linkage Cluster Center:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.451781	0.009904	0.004323	0.723155	0.007858	0.059250	0.033588	0.366977	0.127254	0.009029	0.004272	0.003348	0.158307	0.004092	0.000048	0.000271	0.001328	0.001372	0.00041	0.000000
2	0.584963	0.017534	0.003738	0.020558	0.014988	0.097642	0.020134	0.666419	0.073264	0.013922	0.006028	0.005825	0.407325	0.006010	0.002816	0.001767	0.001428	0.000000	0.00000	0.000733
3	0.615525	0.012310	0.000000	0.000000	0.000000	0.160036	0.000000	0.184657	0.000000	0.036931	0.006155	0.012310	0.747938	0.012310	0.000000	0.012310	0.000000	0.000000	0.00000	0.000000

Cluster structure: There are 3 clusters with a threshold of 0.3 for the single linkage cluster. The clusters' sizes and forms differ greatly from one another. There are 3 clusters for the entire linkage cluster with a threshold of 0.85. In terms of sizes and forms, the clusters are more comparable to one another than the ones in single linkage clusters.

The stability: The complete linkage approach is more stable than the single linkage cluster, as can be seen by looking at the cluster centers, where the values for the clusters are more closely spaced.

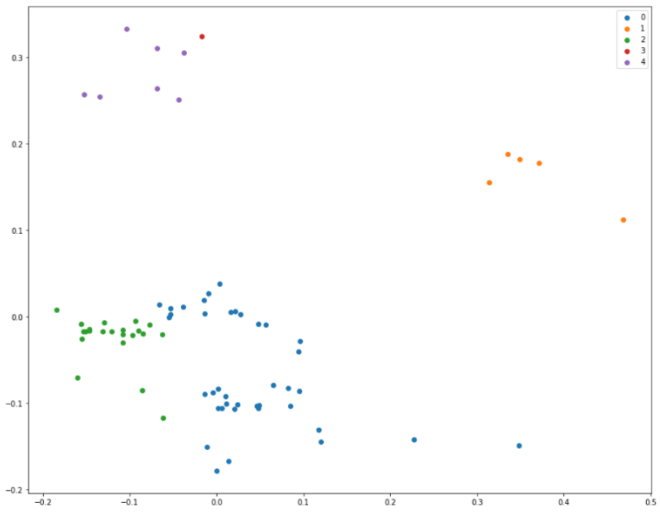
ii. Which method leads to the most meaningful clusters? How many clusters results from this method? What distance is used for this cutoff? (1.5pts)

The complete linkage method leads to more meaningful clusters. 3 clusters results from this method. The cutoff distance is 0.85.

c. Apply k-means clustering using all variables except *name* and *rating*, and with k=5.

i. Describe the characteristics of the cereals within each cluster (2pts)

	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sugars	Potass	Vitamins	Shelf	Weight	Cups	Cold	Nabisco	Quaker	Kelloggs	GeneralMills	Ralston	AHFP
0	70	4	1	130	10.0	5.0	6.0	280.0	25	3	1.0	0.33	1	1	0	0	0	0	0
1	120	3	5	15	2.0	8.0	8.0	135.0	0	3	1.0	1.00	1	0	1	0	0	0	0
2	70	4	1	260	9.0	7.0	5.0	320.0	25	3	1.0	0.33	1	0	0	1	0	0	0
3	50	4	0	140	14.0	8.0	0.0	330.0	25	3	1.0	0.50	1	0	0	1	0	0	0
5	110	2	2	180	1.5	10.5	10.0	70.0	25	1	1.0	0.75	1	0	0	0	1	0	0



First cluster: Low Calories, low Fat, high Protein

Second cluster: Low Sodium, low Fiber, low Potass, high Calories
Third cluster: High Protein, low Fat, high Sodium, high Potass
Forth cluster: Low Calories, low Sugars, high Potass, high Sodium, high Vitamins, high Protein
Fifth cluster: High Calories, low Protein, low Fiber, low Potass, high sugar

ii. Provide an appropriate name for each cluster. (2pts)

First cluster: Nutrition
Second cluster: Energetic
Third cluster: Mineral material specialist
Forth cluster: Healthy
Fifth cluster: Sweet

d. The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet.
   i. For this goal, find a cluster of "healthy cereals". (2pts)

   Forth cluster, because it is low Calories, low Sugars, at the same time, it still provides lots of energy, because it is high Potass, high Sodium, high Protein, high Vitamins.

   ii. Should the data be normalized? If not, how should they be used in the cluster analysis? (1pt)

   It's important to normalize the data. As we require a same scale for each variable, the k-means clustering won't be impacted.

Q2. **Predicting Customer Churn [25 points]** The dataset *Churn.csv* contains 15 predictors worth of information about 3333 customers, along with the target variable, Churn, an indication of whether a customer left the company (true) or not (false). Use the data to answer the following questions. (Note: Churn.csv is a publicly available dataset. You may consult public sources online, e.g., https://www.kaggle.com/datasets/blastchar/telco-customer-churn for more information about the dataset and variable descriptions).

a. A classification model has been applied to a subset of the data and has classified 88 churn customers (30 correctly so) and 952 as non-churn customers (920 correctly so). Construct the confusion matrix and calculate the following metrics: *overall error rate*, *sensitivity*, *specificity*, and *lift*. (4pts)

	Positive (Actual)	Negative (Actual)	Total
Positive (Predict)	30 (TP)	58 (FP)	88
Negative (Predict)	32 (FN)	920 (TN)	952
Total	62	978	1040

Overall error rate = (FP + FN) / (TP + TN + FN + FP) = (58 + 32) / (30 + 920 + 32 + 58) = 0.0865

Sensitivity = TP / (TP + FN) = 30 / (30 + 32) = 0.484

Specificity = TN / (TN + FP) = 920 / (920 + 58) = 0.941

Lift = (TP/(TP+FP) (TP+FN))/(TP+TN+FP+FN) = 5.72

b. Suppose that you can adjust the cutoff (threshold) for classification on a sliding scale to alter the proportion of records classified by the model as churn. Describe how moving the cutoff up or down the scale would affect:

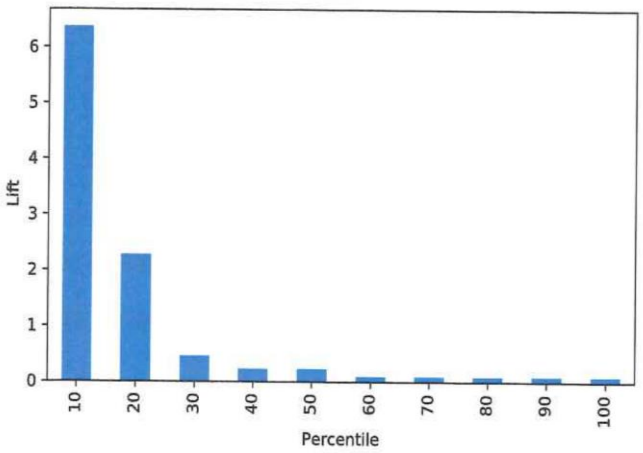
   i. the classification error rate for true churn customers. (1pt)

   Moving cutoff up will reduce the false positives, so increase true positive, thus reduce the classification error rate.
   Moving cutoff down will increase the false positives, so reduce true positive, thus increase the classification error rate.

   ii. the classification error rate for true non-churn customers. (1pt)

   Moving cutoff up will increase the false negative, so reduce true negative, thus increase the classification error rate.
   Moving cutoff down will reduce the false negative, so increase true negative, thus reduce the classification error rate.

c. Below is a decile lift chart for applying the model in part a) to classify new customers.



   i. Interpret the first and second bars from the left. (2pt)

   First bar: Taking the top 10 predictions from all the records that are applying to the model in part a) to classify new customers has six times higher accuracy rate as a random guess.

Second bar: Taking the top 20 predictions from all the records that are applying to the model in part a) to classify new customers has twice higher accuracy rate as a random guess.

ii. Explain how you might use this information in practice. (1pt)

Use this information to evaluate if it is worthy to apply the new model to the model. If the performance increment is high, then it is worth.

iii. Another analyst comments that you could improve the accuracy of the model by classifying all customer as non-churn. If you do that, what is the error rate? (1pt)

Make all the wrong prediction for actual churn.

Thus, error rate = (30 + 32) / 1040 = 0.0596

iv. Comment on the usefulness of the two model performance metrics in this case (i.e., error rate and lift). (1pt)

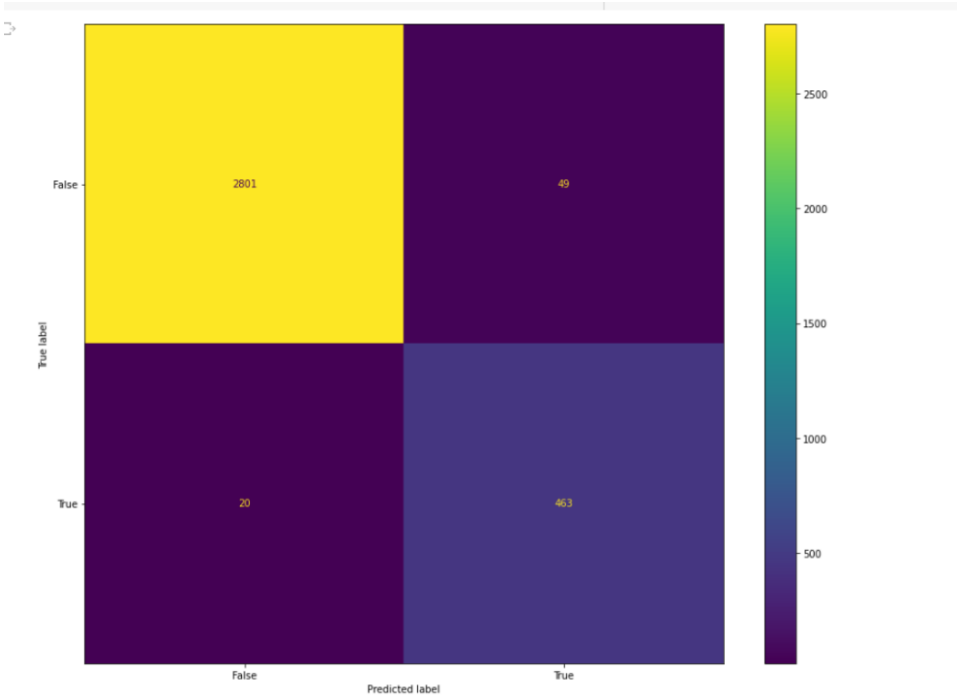
Error rate refers to a measure of the degree of prediction error of a model made with respect to the true model. Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. So, this case, lift is a better representation because for part c, although the error rate decreased, but it labeled everything as non-churn, it is actual a worse result, so lift is better in this scenario.

d. Suppose that intervening with a customer at risk of churning costs $100, and that a customer who churns represents $2000 in lost revenue. Assume that the company has 50% success rate of preventing customers from churning.

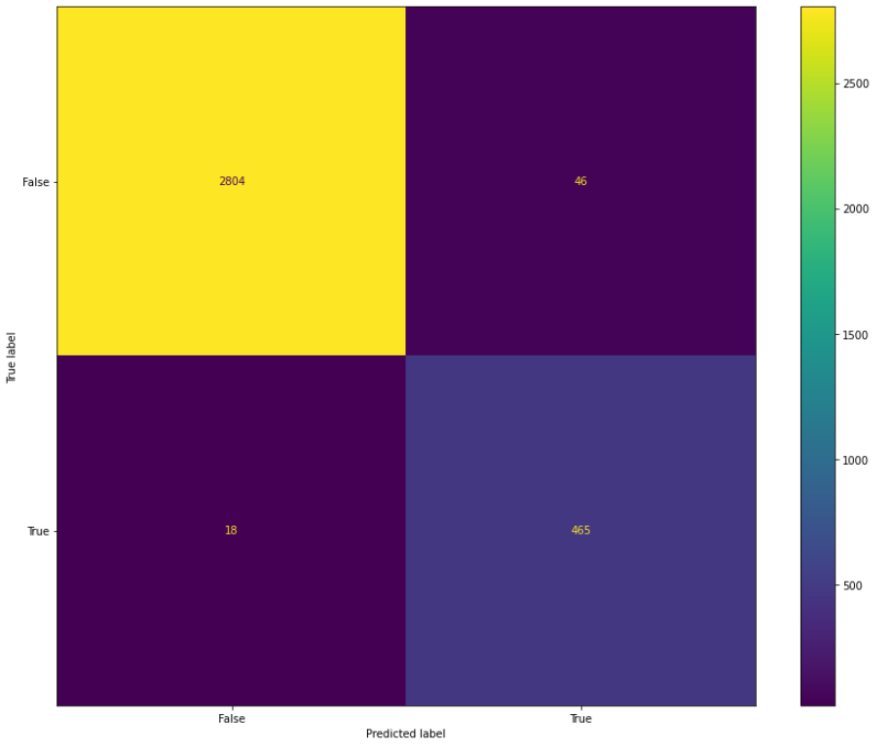
i. Construct a cost matrix for the scenario above. (4pts)

	Churn (Actual)	Non-churn (Actual)
Churn (Predicted)	0.5 * 100 + 0.5*(100+2000) = 1100	2000
Non-churn (Predicted)	100	0

ii. Develop a CART model to predict Churn without misclassification cost. (3pts)



iii. Develop a CART model to predict Churn using the cost matrix in i. (3pts)



iv. What is the increase/decrease in revenue by incorporating the cost matrix? (4pts).

Total cost without cost matrix = 2801*0+49*100+20*2000+463*1100 = 554200
Total cost with cost matrix = 2814*0+46*100+18*2000+465*1100 = 552100
Increase in revenue = 554200 – 552100 = 2100

Q3. Sales of Riding Mowers [15pts]: A riding mower manufacturer would like to identify the best sales prospects for an intensive sales campaign. The company is interested in classifying households as prospective owners or nonowners based on Income (in $1000) and Lot Size (in 1000 ft-sq). You have been hired as an analyst and given a random sample of 24 households in the file *RidingMowers.csv*.

a. Use all the data to fit a logistic regression of ownership on the two predictors. (2pts)
Equation used:
p(ownership=1) = 1 / (1 + exp(-(0.10085007 * income - 0.81168427 * lot_size)))
coefficients:

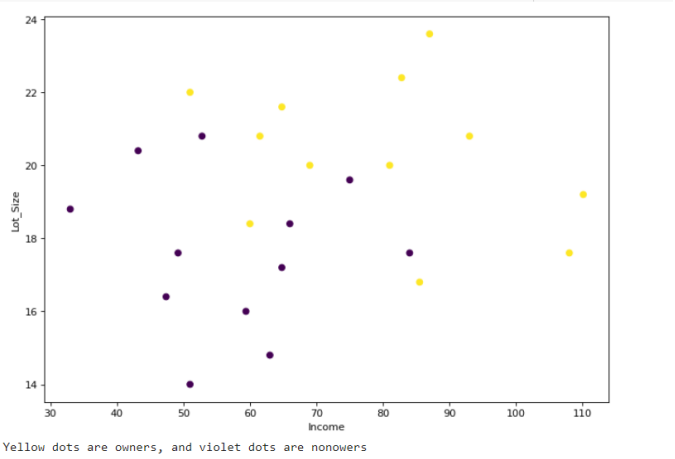
```
[ ]  model.coef_

     array([[0.10085007, 0.81168427]])
```

b. What percentage of households in the study own snow blowers? (2pts)

The percentage of households in the study own snow blowers is 50%

c. Create a scatter plot of Income vs. Lot Size using color or symbol to distinguish owners from nonowners. From the scatter plot, which class seems to have a higher average income? (2pts)



Owner class seems to have a higher average income.

d. Among nonowners, what is the percentage of households classified correctly? (2pts)

   83.3%

e. To increase the percentage of correctly classified nonowners, should the cutoff probability be increased or decreased? Explain your answer. (2pts)

   We have to decrease the cutoff probability, moving cutoff down will reduce the false negative, so increase true negative, thus reduce the classification error rate. Thus increase the percentage of correctly classified nonowners.

f. What are the odds that a household with a $60K income and a lot size of 20,000 ft-sq is an owner? Use cutoff = 0.5. (2pts)

   0.49

g. What is the minimum income that a household a with 16,000 ft-sq lot size should have before it is classified as an owner? (3pts)

   $92.6k