Machine Learning Homework 6

Eric Larsen, A02176917

April 2024

Problem 1: Reinforcement Learning

This paper is a good overview of reinforcement learning and how the methodology works. The paper goes through and talks about the reason we want to create RL methods, allowing for independent agents to interact with the world without required human input. Doing this is easier said than done, but the grounding of RL methods finds its roots in behavioral sciences. The main objective of RL methods is to optimize a reward given the states it is in. Each state consisting of what it has been programmed to know. Using this state information it then tries to achieve the best next state by learning an optimal function to transition between states. With this being said it can be achieved using a Markov Decision Process, building future state conditions only on current state knowledge.
The paper describes two ways of doing RL methods, either through a value function, or through a policy search. Where the value function method focuses on making an optimal Q function based on maximizing the Q function for the next state. Policy search methods instead look for the best policy Π with a starting policy set. This is then updated overtime using either gradient free methods, or gradient methods. Where gradient methods are the desired method instead of having to do a search of next policies using gradient free methods. The paper then goes through and talks about how you can implement these methods by either modifying the agent or the environment to include additional dependencies. Another interesting idea presented was instead of using the full state, you can break it into individual actions allowed to then allow the agent to weight the values of these actions and their expected rewards.

Problem 2: Exploratory Data Analysis

All parts of this portion of the assignment are found in the submission file Problem2_HW6.py

Part 1)

From the dataset there were missing values in the income of some of the customers listed. To impute this information I set up a dataset to do Random Forest regression to create this missing data. I used random forest because it makes sense to impute this data from other samples that look like it, i.e. data points that follow the same tree system. Taking these points requires us to make the assumption that people of similar expenses have the same income. Which is not realistic to all applications but it gives us a better vision of these people than taking a mean income or some other simple statistic.

Part 2)

Generating the distributions for Age, Income, Wine purchased, Meat purchased, and Gold Products. With these results shown in Figures 1- 5 below.

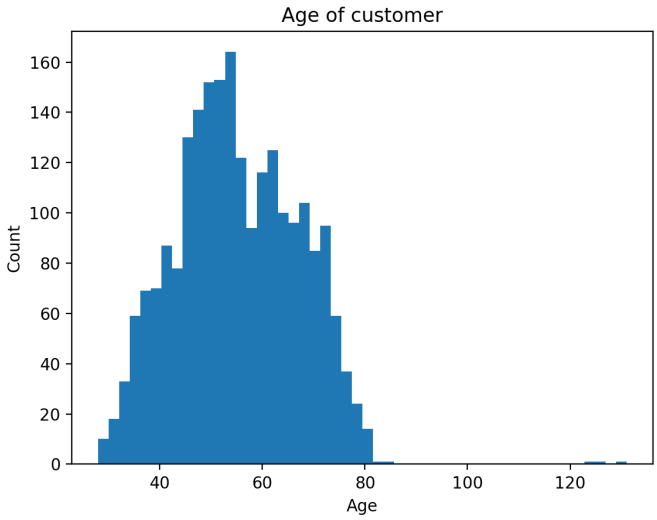


Figure 1: Age Distribution

Here it is clear to see that the histogram of ages shows most people are of middle age. It is also clear to see that one person lied about their age being over 100 years old. For the most part it is a Gaussian distribution centered round 50 years of age. The Income distribution is also clearly Gaussian.

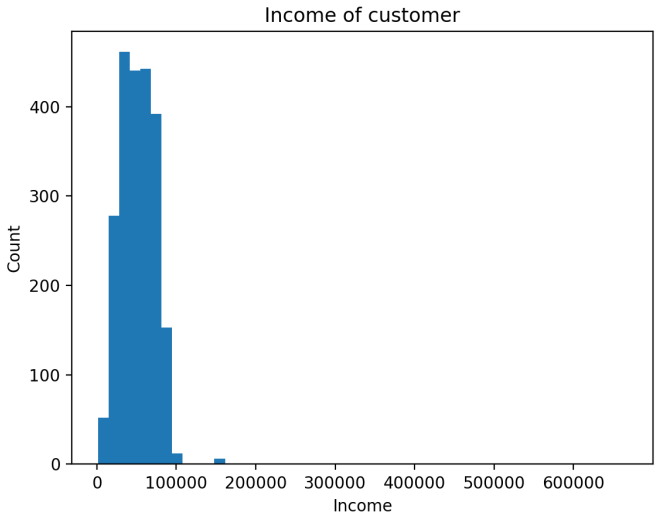


Figure 2: Income Distribution

Here is is also clear to see that one individual makes the curve harder to see. but the general trend is a Gaussian centered at 80,000 dollars and the most of them make under 100,000. Next we can see the distribution of amount of wine purchased.

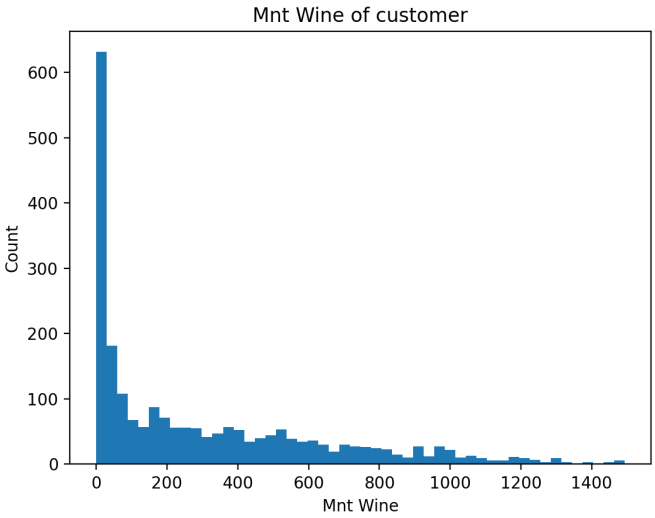


Figure 3: Wine Distribution

Here it is clear to see that most people do not spend a lot of money on wine. It looks like an exponential decay as you increase the amount of wine purchases. This trend is expected due to the link between income and amount one has to purchase more luxury time items. In contrast most people show a common trend in amount of meat products.

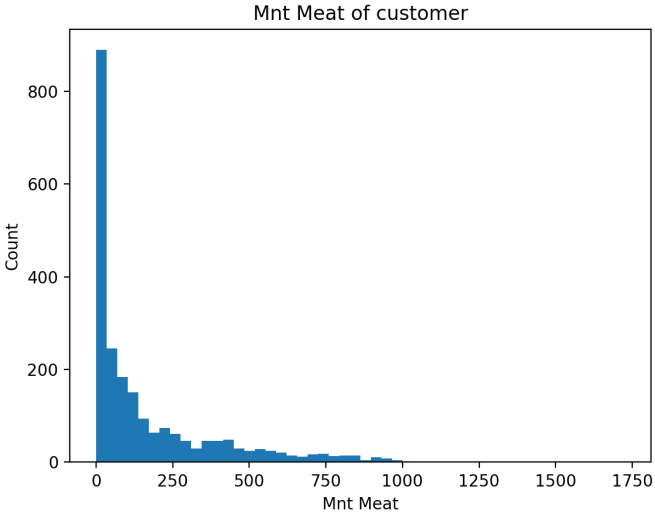


Figure 4: Amount spent on meat products

here it is clear to see in general most people spend the same amount on meat, with the general group spending less than 250 dollars a month. Finally looking at the amount spent on Gold products is shown below.

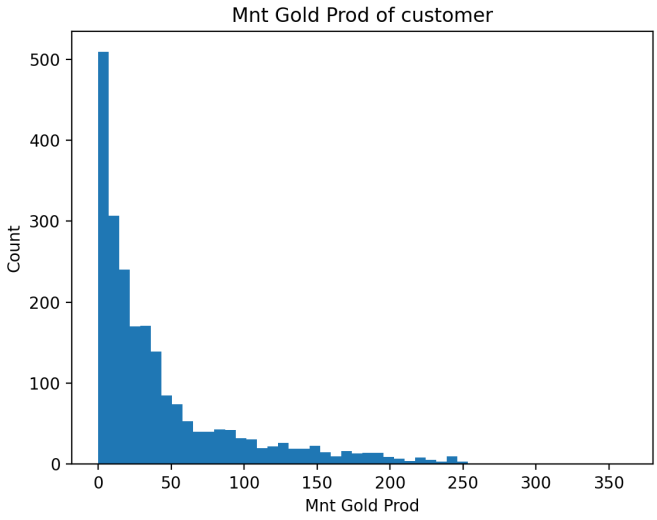


Figure 5: Amount spent on gold products

Here like the other distributions the majority of individuals spend low, and it decays quickly. Overall the trends show that most people have similar tastes and have spend reasonably low monthly on the products.

Part 3)

Here we can see the distribution of categorical values, namely Education, Relationship status, and number of kids at home.

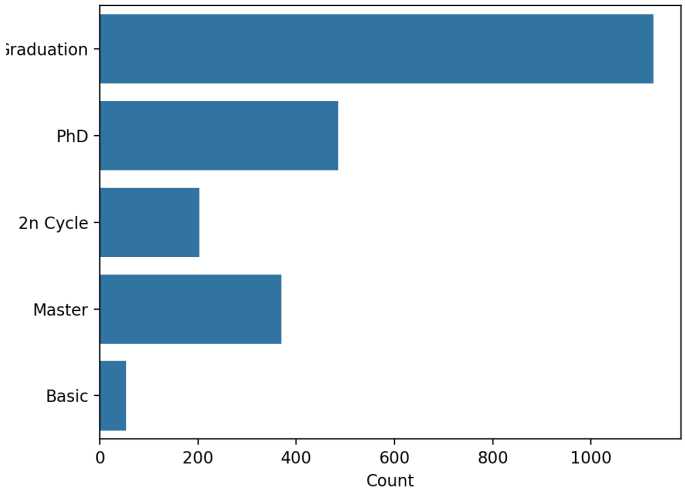


Figure 6: Education of customers

Here we have 5 different categories of education with the highest population coming from the graduation category. This seems to be a high school education. Next looking at the marital status of the customers.

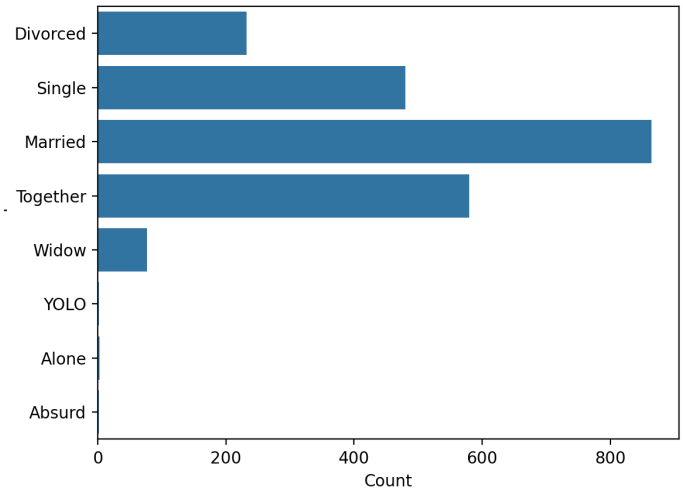


Figure 7: Marital status of customers

here from the eight status categories the majority being married, with the next set being classified as together. Here there are a couple of 'erroneous' categories that one could contribute to people being allowed to put their own responses in. Continuing with the family theme I found chose the third category to be kids at home.

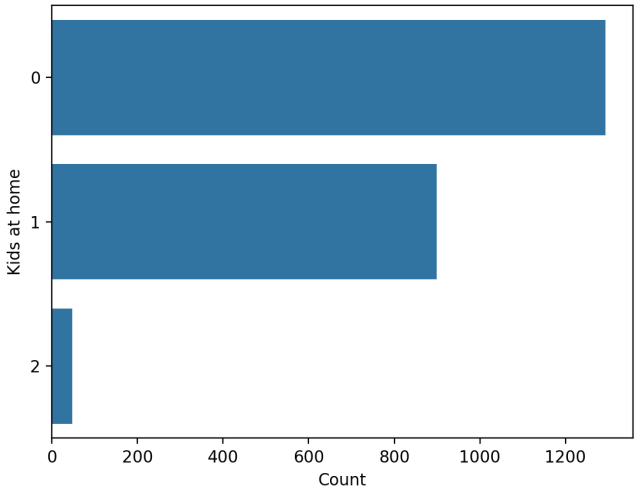


Figure 8: Bar plot of kids at home

Here it is clear to see that the majority of customers have zero kids at home. Which I didn't expect given the known number of married customers, but that could be a preconceived bias on the data.

Part 4)

Exploring the number of purchases and the amount spent on different products gives the following figures from CCA.

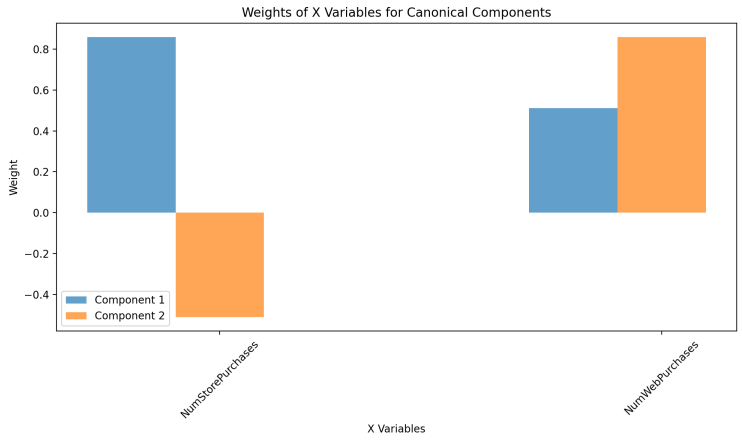


Figure 9: Number of purchases influence CCA

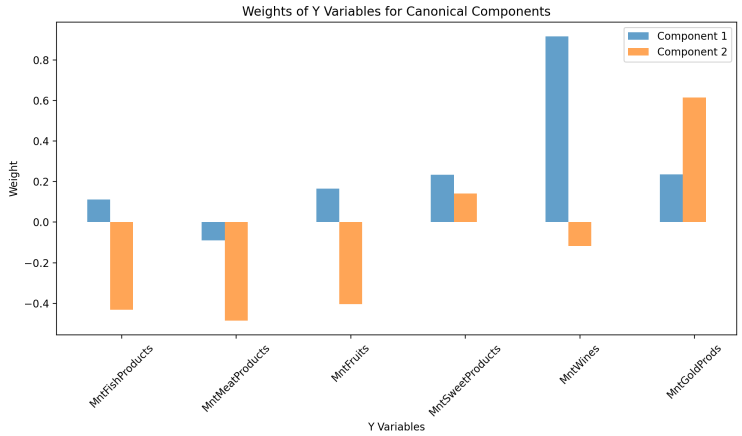


Figure 10: Products purchased CCA

Here it is clear to see that the amount of store purchases has a strong correlation to the amount of wine purchased. Where the number of web purchases has a negative connection to fish, meat, fruit purchased. But does have a strong connection to the number of gold products purchased.

Part 5)

Exploring the age of the customer and the income on different products gives the following figure using CCA.

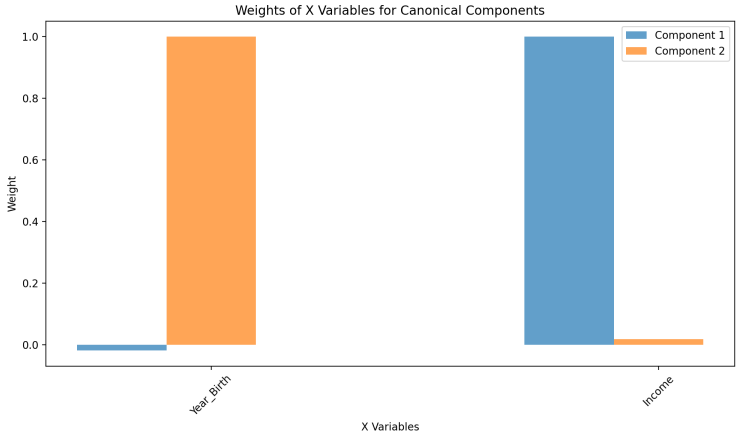


Figure 11: Demographic influence CCA

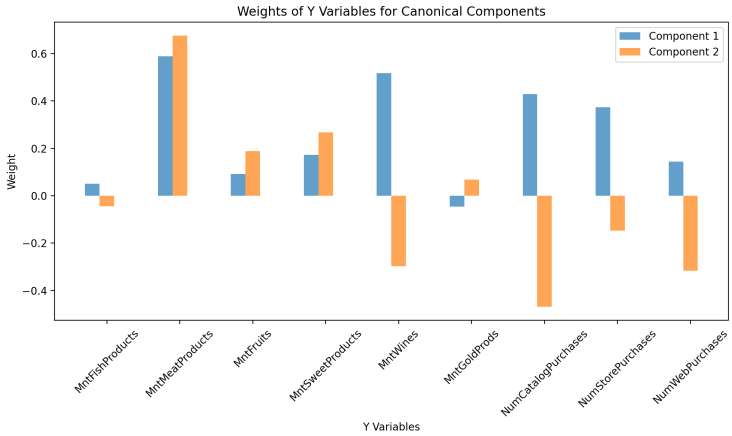
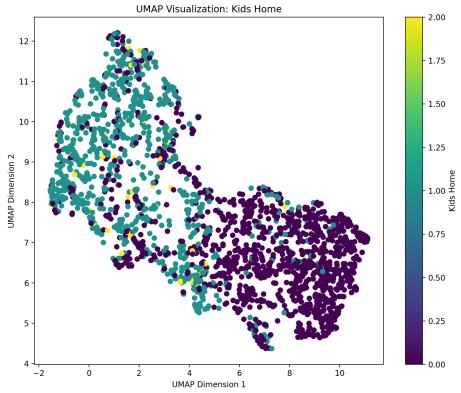


Figure 12: Products purchased CCA via demographic information

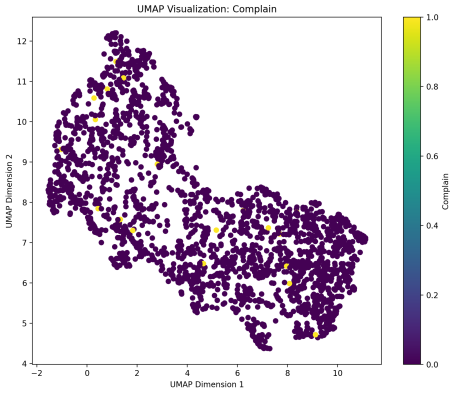
Here it is clear to see that the age of the individual has the most influence on the amount of meat they purchase, and surprisingly negative effect on the number of catalog purchases and the amount of wine purchased. Then the income is strongly connected to the amount of meat, wine, and purchases in general made by the customers.

Part 6)

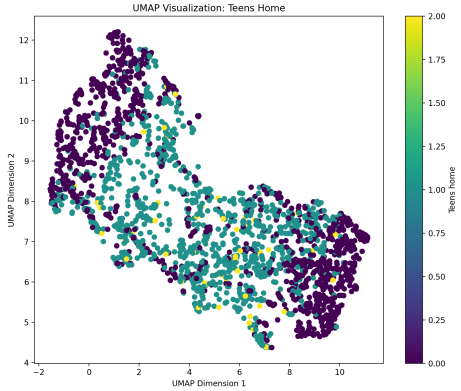
For visualizing the data I decided to use UMAP to visualize the data, and the following five figures show coloration using the five different categorical values; namely Kids at home, Complaints, Teens at home, Education, and Marital Status.



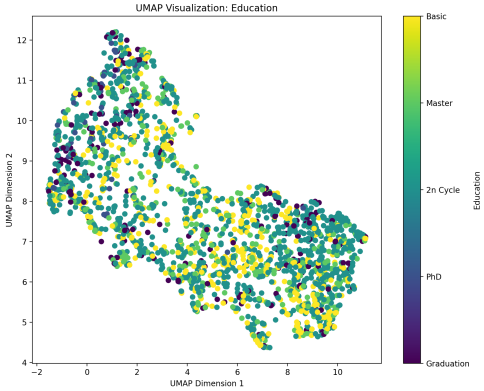
(a) UMAP kids at home



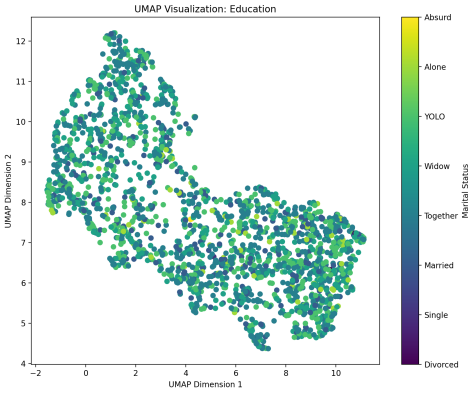
(b) UMAP complaints from customer



(c) UMAP teens at home



(d) UMAP education of customer



(e) UMAP marital status of customer

Figure 13: UMAP visualization of the dataset

From these results it is clear to see that there is not a clear separation between the different categories of the customers in the dataset. The most clear coloration comes from the kids at home, with marital status

showing the hardest to follow coloration. These issues were not unexpected due to the complexity of the data. From this it is clear to see the data is not cleanly separable.

Part 7)

Using this UMAP embedding for visualization we can then apply K means clustering to the data, where the selected K is 2 due to the number of possible selections for the desired list of if people buy the promotion or not. This visualization is shown in in the figure below

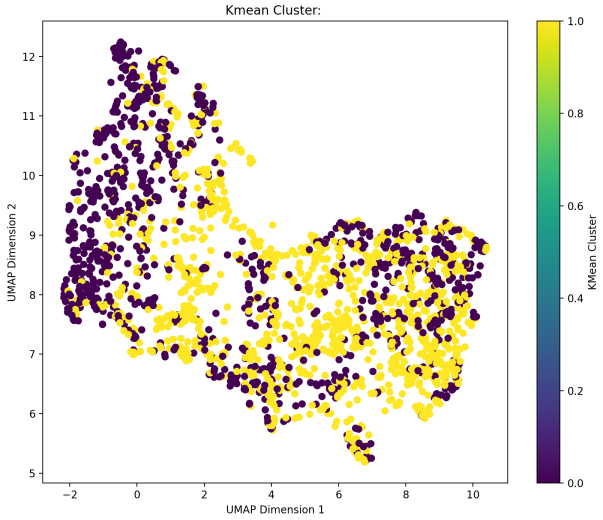


Figure 14: UMAP embedding colored by Kmeans clusters n=2

Here again we wee that there is not a clear separation of the data. Though clearly it can be seen that there is a trend with the 0 class being along the edges of the clustering visualization. While the 1 class is more centralized in the UMAP embedding. From these results I believe that we know that our data is not going to be easy to separate at least at a 2D level from what is seen from UMAP.

Part 8)

Next going through and exploring the trend in customer visits to the company website over time yields the following figure

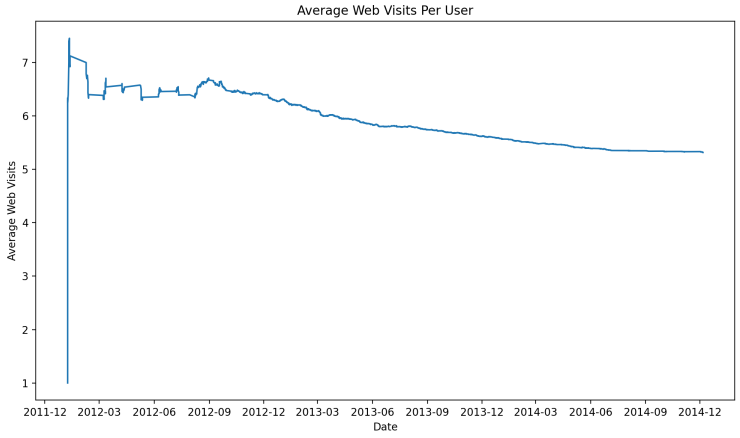


Figure 15: Average number of web visits per customer

Here you can see the general trend is consistent with a long term average of 6 uses per customer per month. This trend comes over time with the beginning of the curve being at a higher average use of the website with averaging a 6.7 uses per customer. So over time the use of the website reduces in the customer base. Looking for a more granular understanding by user add data gives the following figure not using a rolling averages over the total number of customers

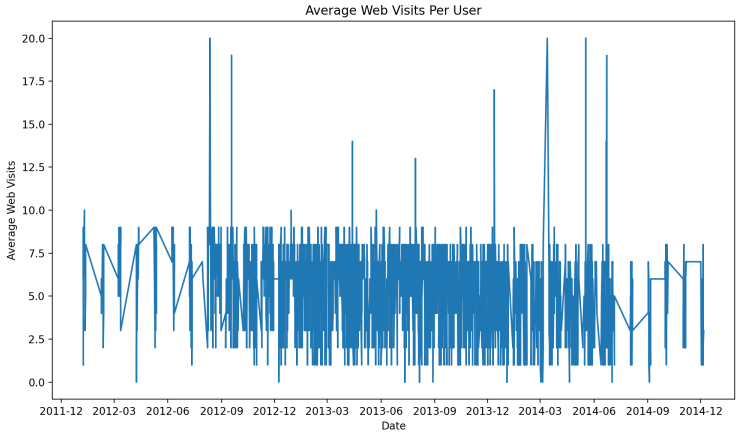


Figure 16: Granular view of web visits per customer over enrollment date

Here there is the same trend, though there are the users that have a much higher use rate of the website. Overall the general trend is an average of 6 uses per person. With this consideration looking at the customer behavior over time I just looked at the average trend per user by enrollment date because it shows the same trend as a granular view.

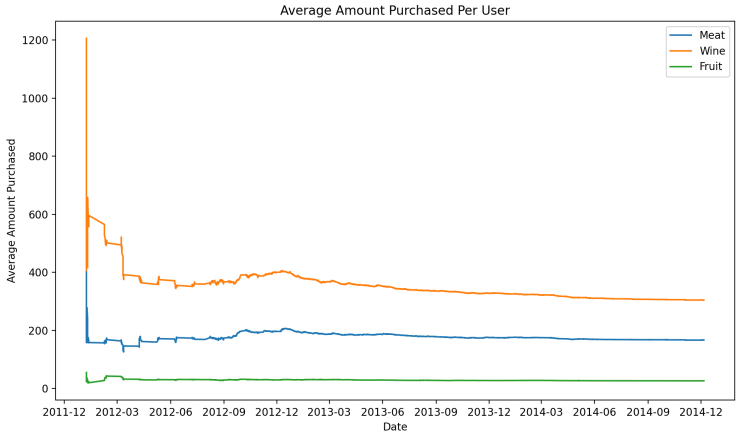


Figure 17: Purchase history of users over enrollment date

In the amount spent on commodities the fruit is consistent overtime, while the wine and meat see a peak around Dec 2012. This shows a large contribution to these products from those that enrolled around that time, they have a higher spending habit than other individuals. It is interesting to see that over time the wine amount purchased increases over the average customer.

Problem 3: Machine Learning

All parts of this portion of the assignment are found in the submission file Problem3_HW6.py

Part 1)

for this assignment because there was a class disparity between those that did not accept the promotion (class 0) and those that did accept the promotion (class 1) I set up a test case that pulled 50 of each class to see how it compared. This allowed me to see how the performances of each method compare to random guessing. If one was to pick just one class in the test set the performance would be 0.50 for accuracy. This acts as a baseline understanding for the dataset.

Part 2)

for this section the code can be found in the submitted

Part 3)

The relative performance of the modeled developed in the code are shown below in table 1.

Table 1: Performance metrics of different models

Model	Accuracy	Precision	Recall	F1 Score	Roc_auc
KNN	0.52	1.0	0.04	0.0769	0.52
LDA	0.63	0.8421	0.32	0.4638	0.63
QDA	**0.72**	0.8235	**0.56**	**0.6667**	**0.72**
LR	0.53	1.0	0.06	0.1132	0.53
SVC	0.54	1.0	0.08	0.1481	0.54
RF	0.59	0.8462	0.22	0.3492	0.59
XG	0.69	**0.9524**	0.4	0.5634	0.69

Here the performance of each method is clearly seen. While other methods were able to have an Precision of 1.0 this is most likely due to predicting the majority class as you can see poor performance in the recall value. Compared to other methods the best method for this supermarket dataset was QDA, while all methods developed did not perform as high as desired. QDA performed well across the board given all the performance parameters measured. Comparing the performance of these methods via confusion matrices shown below in Fig. 18.

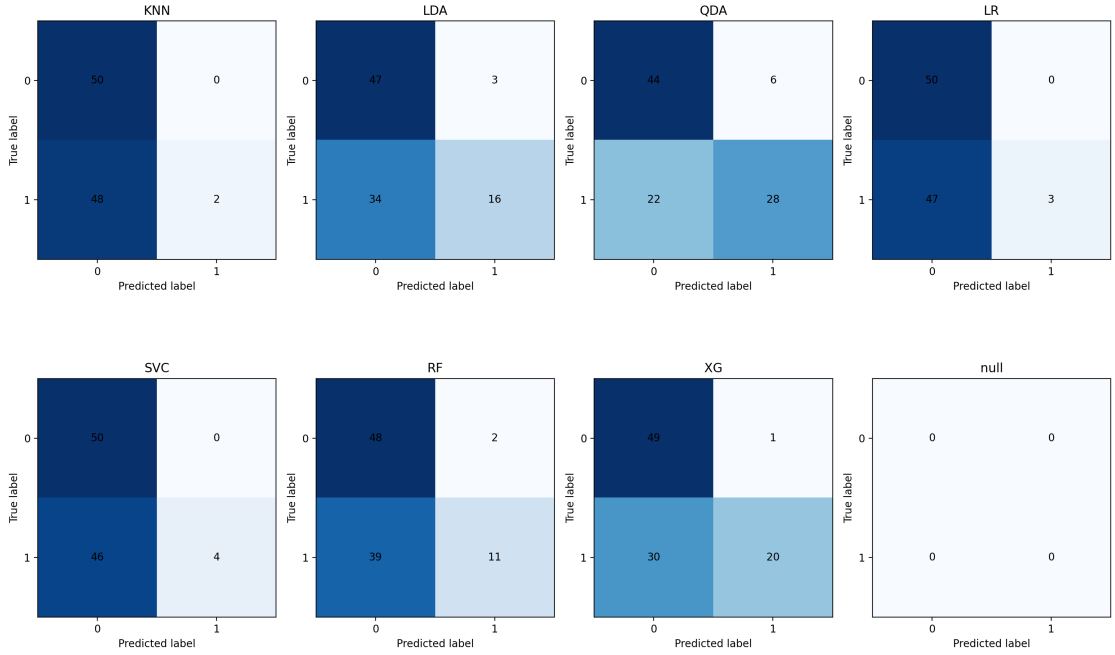


Figure 18: Confusion matrix of different methods used

Again from this result the best performing network would be QDA for this dataset. From Fig. 18 it can be seen that most methods predict that no one will purchase the promotion no matter what was given to it, from this it is clear to see that one needs to use a method like QDA or XGBoost to get a better prediction.

Part 4)

Looking into the variable importance I used the developed XGBoost method. The realative feature importance is shown below in Fig. 19.

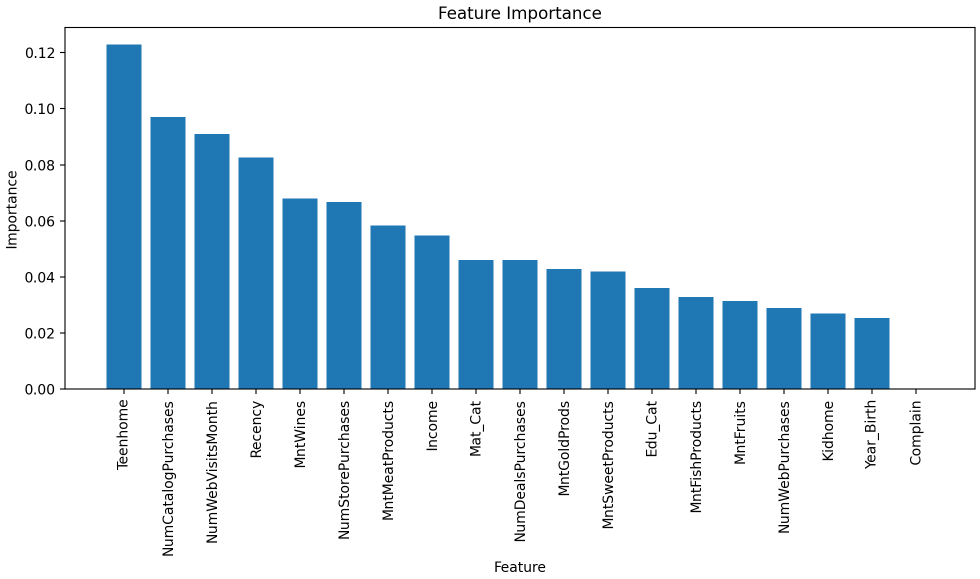


Figure 19: Feature importance from developed XGBoost method

These results are surprising to me because these are not the categories I would have made an informed prediction on personally. The most important feature is the number of teens at home, with the next being the number of catalog purchases made. I would have expected income to be a major contributing factor to the overall decision. One of the lease contributing features, at least from this XGBoost method is the age of the individual, and whether they complained in the last two years. The features with the highest importance seem to come together, at least make logical sense that they are connected. Each of the features would contribute to how much one uses the store and therefore the amount they have access to the promotion. If you are using the store more, either by visiting the store, using the catalog, or by using the website you are more likely to get the most out of the promotion. If someone uses it more they are likely to want to get the most out of their store experience.