Problems Inclass 8_2. You can comment in this document and submit a pdf of your work. Please mark clearly all your answers and answer problems in the order provided.

1. Think through and answer the following problems to the best of your abilities.

   a) Valentine Day is approaching. A restaurant is trying to decide if to organize a singles' night or if to offer a special romantic menu. The restaurant has an established base of customers and collects demographic, income, social media and behavioral information on its customers. They decide to use the help of a data scientist to make sense of their Valentine's day menu in order to maximize sales (Valentine's days tend to be cash cows for restaurants). What algorithm would you use?

First, I would separate the customer data into two main parts: single customers and couples. Thus I would use a Linear Classifier. From there the income should analyzed to predict which demographic will yield the greatest profit. That alone however is not enough. The behavioral information can be studied to predict which group will be more likely to attend singles night or go there for a date.

   b) Describe the type of information you would collect (what features) to decide if an email is spam or non-spam and what machine learning algorithm you would use.

The type of information I would collect would be typical characteristics of spam email. For this, I would use logistic regression. Of course the first thing to check would be how many times I have already received an email from this sender and how many of those times I have manually sent it to spam. Having sent it at least once should be a red flag and a big indicator. If the From: and Reply To: addresses are different, that is another indicator. Some other include if the message body is only HTML, if message body contains remote image, or message contains only or mainly tags.

   c) Describe the type of information you would collect (what features) and from what sources to decide if to buy or sell a stock (financial investment). What machine learning algorithm can you use?

Type of information: Stock History, current state of stock, buyers and sellers of that stock, current news about the stock company. I would use Linear Regression in this case because we would be using information on a series of features (described above) that characterize the observations and it will identify the impact of each feature.

   d) How would you use Facebook to recommend certain products to people and what machine learning algorithm would you use?

Keep track of what people click on, like pictures and links, and of the things they share and like. I would use the Nearest neighbor ML Algorithm to find similarities between observations to make predictions.

2. A classification algorithm classifies emails into spam and non-spams. The following confusion matrix was returned by using the classifier on the testing set:

264	14
22	158

Consider “non-spam” = “positive” class. The matrix has the organization described in class. Calculate and interpret the following:

1) Accuracy rate
   = (TP + TN) / (TP + TN + FP + FN) = 0.92 = 92%

2) Precision

   = TP/ (TP + FP) = 0.91 = 91%

3) Recall
   = TP / (TP + FN) = 0.88 = 88%
4) F1
   = 2*Precision*(recall / (precision + recall)) = 0.89

5) Sensitivity
   = TP / (TP + FN) = 0.95 = 88%
6) Specificity
   = TN / (TN + FP) = 0.95 = 95%

7) In your opinion, is it more important to have good recall or precision?

It is more important to have good precision because this will reduce the amount of false positives passing through. In other words, you want to mark fewer spam messages because the spam messages that make it through can be sent to spam manually, and you really don’t want to send important emails to spam.