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?

      Since they must do one or the other, they don't care about probabilities—they only need to know whether most of their customers are single or attached. They need a classifying algorithm. Since there appear to be several features, a Naïve Bayes classifier would probably work best.

   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.

      A decision tree along these lines might work: 1) Is the email from somebody in the contacts list? If no, 2) Is the email from the corporate domain name? If no, 3) Does the email come from list of known spamming addresses? If no, 4) Does the email contain words typical of spam (which can be discovered doing word counts on a training set of spam emails). If no, not spam.

   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?

      Historical data from stock prices tells nothing about future prices, apart from some range about the standard deviation of the various changes. Also, evidence of other investors or media pundits provide no information. You would need historical information on how firms in the same industry have fared under various changes in the economy—liquidity from the FOMC, interest rates, weather. You would have to build a model that uses each of these as a feature and model changes based on likely scenarios—sudden inflation, war in a country that supplies raw materials, etc. You would then use logistic regression to find probabilities of rises and falls.

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

You would use a clustering algorithm where the dimensions are products. You would find the k nearest neighbors, and find the products that are most common among them.

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/all = (264+14)/(264+14+22+158) 0.88535

2) Precision

TP/ (TP+FP) = 0.923077

2) Recall

TP / (TP+FN) = 0.625592

3) F1

2 * precision * recall/(precision + recall)

2 * 0.923077 * 0.625592/(* 0.923077 +0.625592)

0.745762

4) Sensitivity

TP / (TP+FN) = 0.625592

5) Specificity

TN/(TN+FP) = 0.388889

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

It depends on your needs. With Google, precision is important, because it's more important to have a good, relevant result than to recover every single relevant documents. With an academic search engine, however, where you're really trying to recover all the articles about a certain topic, recall is more important than precision. Having to ignore the non-relevant articles is preferable to missing something important.