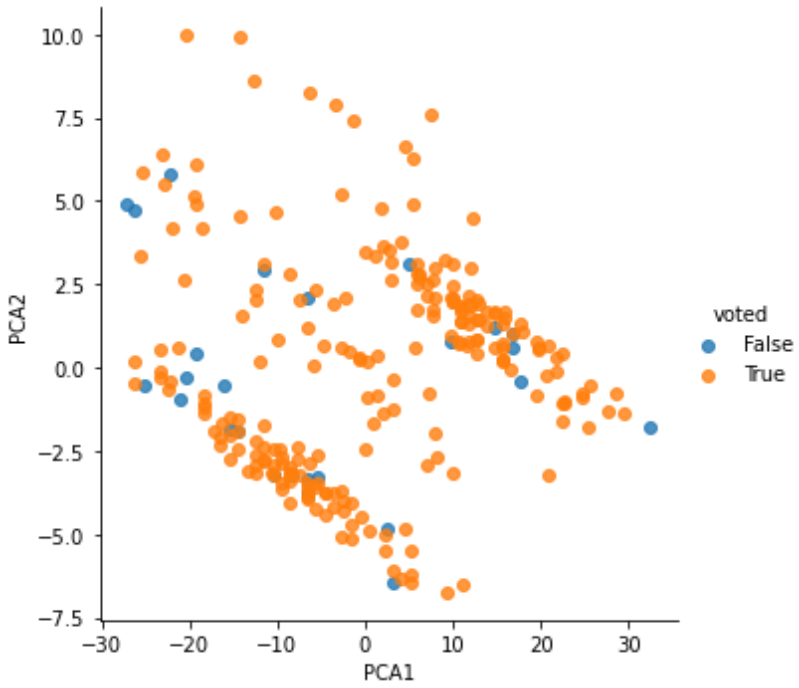
Huriye Güngör

78014

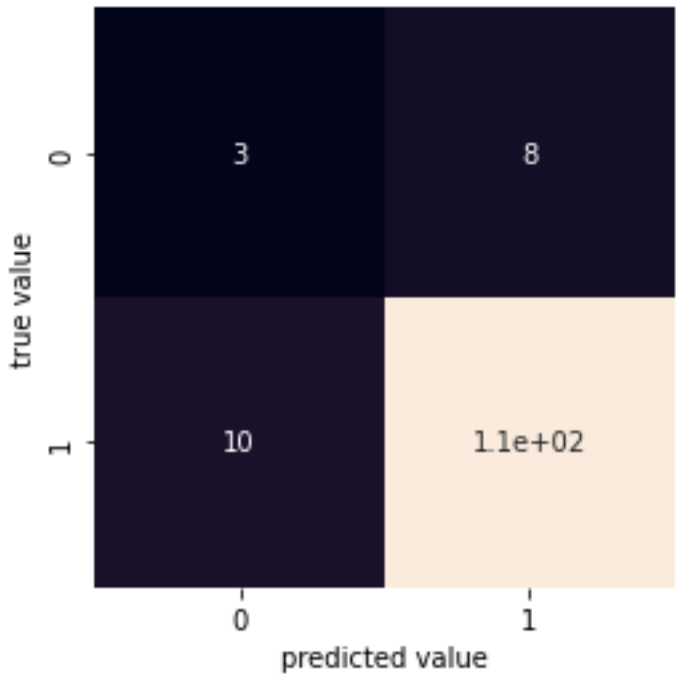
CSSM 502 Homework 3 Report

In this homework, I used machine learning algorithms to predict a person is likely to vote or not, given his/her demographic background. Firstly, I started with observing features and unique values in each column. Then, I moved on to data cleaning process. I removed all the missing and non-informative rows. Therefore, the size of my data decreased from 12451 to 257. Moreover, I dropped some of the columns in the model which have either too many unique values or just one unique value. In order to ensure the distribution of the data has not changed before and after the cleaning process, I looked for number of "True" labels in "voted" column. At first, it was 10226 out of 12451 (which is 82%), and after, it became 231 out of 257 (which is 90%). Thus, I think that it is close enough, but some of the valuable information may be lost during the preprocess.

Then, I used one hot encoding to represent categorical variables, which are not ordered. Doing so, number of columns are increased from 32 (excluding unnamed ID column) to 41 columns. After this, I move on to machine learning part. In this part, I split data into train and test parts with the same size. First of all, I tried Gaussian Naïve Bayes algorithm. However, it came up with 25% accuracy, which is far more below the required accuracy. After that, I used PCA to reduce the number of features in the model to 2 (See the figure below).



However, I did not find this graph informative; thus, I continued with K-Nearest Neighbor algorithm. I choose n=1 as a start. The accuracy score is calculated as 86% in this case, and I found it satisfying. I tried n=2 after this; however, accuracy score is decreased, and I decided to go with n=1. I created a confusion matrix to see true negative and false positive numbers (See Figure below). It seems that 110 rows labeled correctly as True, and 3 rows labeled correctly as False. However, 10 rows labeled wrong as False, and 8 rows labeled wrong as True. In False case, it is clear that number of correct estimates (which is 3) is lower than number of wrong estimates (which is 8). This result is not satisfying; however, I found it natural. Because the False label in the data is very low compared to True label, model is assuming True labels most of the time. Moreover, while data cleaning process, ratio of False labels in the data decreased from 18% to 10%, which also explains this situation.



Finally, I used cross validation methods to ensure the accuracy of the fitted model. I found their results also satisfying, and I decided KNN algorithm with n=1 is suitable to predict the future outcomes of this survey.