Mehmet Kutay Keleş

64755

QMBU 472 HW-3 Report

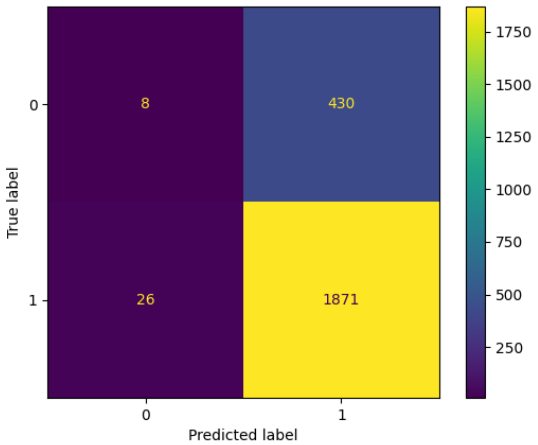
Variable Selection and Feature Engineering

Firstly, all the variables presented in "csesCodebook.txt" are examined. After investigating their meanings from "https://cses.org/wp-content/uploads/2019/03/cses4_Questionnaire.txt", variables D2002, D2004, D2005, D2006, D2007, D2008, D2009, D2012, D2013, D2014, D2017, D2018, D2019, D2031, D2003, D2010, D2015, D2021, D2022, D2023, D2020, D2024, D2025 are chosen to work with. Then, for each variable, Missing values are replaced with the most frequent one. Volunteered: Refused and Volunteered: Don't Know values are kept as it is, as I thought they might have significant effect on the output of the classifier. After that, data is split into train set and test set before training any algorithms.

Example Model of PCA and Gaussian Naive Bayes Classifier

Secondly, an example model with PCA and Naïve Bayes Classifier is trained. For the PCA, the number of components is set to 10. For the Naïve Bayes Classifier, the distribution is set as Gaussian distribution. Then again, a train-test split is done. After training the algorithm and evaluating the result, the following accuracy score and confusion matrix is obtained:

- *The accuracy score of the example model is 0.8047109207708779*

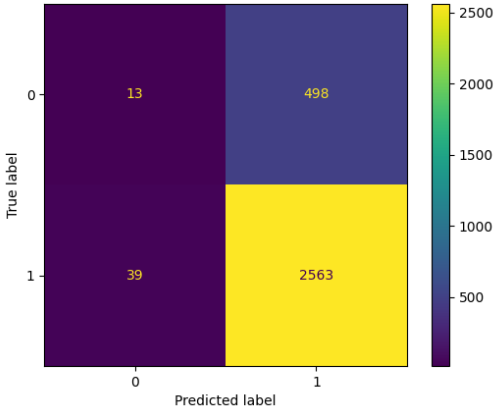


Then, in order to further analyze the accuracy of the sample model further, a 7-fold cross validation test is performed. The result is as follows:

- *7-fold Cross Validation Score List:*
  [0.81409295, 0.79985007, 0.80284858, 0.80809595, 0.8125937, 0.80209895, 0.81109445]
  Average: 0.8072392375240952

Finally, the model is tested on the initial hold-out test set and the results were as follows:

- *Accuracy score on test data: 0.8274975907484742*



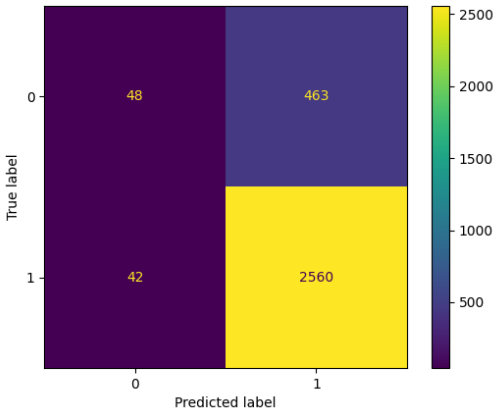
Finding the Optimal Model with PCA and Naive Bayes Classifier

Thirdly, a GridSearchCV is performed on a pipeline with PCA and Naïve Bayes Classifier for tuning the parameters in order to find the best possible optimal model. For the PCA, the range for number of components is set from 2 to 40. For the Naïve Bayes Classifier, the distribution is set to alter between Gaussian distribution and Bernoulli distribution. After running GridSearchCV algorithm, the optimal model is found to be PCA with 17 components and Naïve Bayes Classifier with Bernoulli distribution. Then, in order to analyze the accuracy of the optimal model , a 7-fold cross validation test is performed. The result is as follows:

- *7-fold Cross Validation Score List:*
  [0.81634183, 0.82158921, 0.82158921, 0.81409295, 0.81484258, 0.82083958, 0.82308846]
  Average: 0.8189119725851359

Finally, the optimal model is tested on the initial hold-out test set and the results were as follows:

- *Accuracy score on test data: 0.8371345968519114*



Clearly, both looking at the 7-fold cross validation scores and the accuracy score on test data, the performance is significantly improved. However, before deciding to finalize the model, I wanted to train and test another model with a random forest classifier.

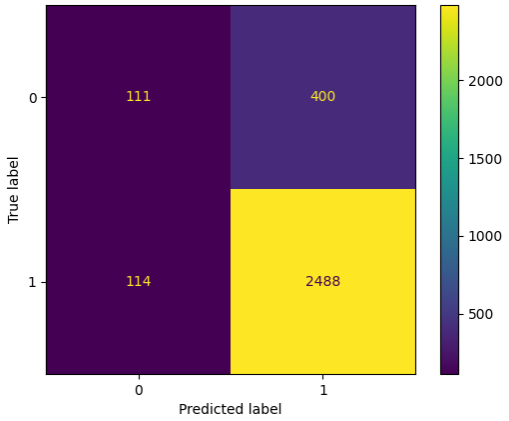
Training a Random Forest Model for Comparison

Lastly, a model with Random Forest Classifier is trained. Then, in order to analyze the accuracy of the model , a 7-fold cross validation test is performed. The result is as follows:

- *7-fold Cross Validation Score List:*
  [0.82158921, 0.81934033, 0.80509745, 0.81409295, 0.82983508, 0.81184408, 0.81409295]
  Average: 0.8165560077104306

Then, the optimal model is tested on the initial hold-out test set and the results were as follows:

- *Accuracy score on test data: 0.8200256986829425*



After examining all the results that are obtained, I have finalized my model to be PCA with 17 components and Naïve Bayes Classifier with Bernoulli distribution as the optimal and best solution.