HW3 Report
Arif Akkan

I used “cses4_cut.csv” for a machine learning project. My aim is to predict a citizen will vote in the presidential election based on her characteristics.

I have used a subset of the data which consists of following variables:
Ordered variables : D2020 Household income, d2003 education, d2023 religiosity, d2029 residence status.
Categorical Variables: D2005 union membership D2013 employment type d2004 marital status

All variables are labeled by numbers. I have used StandardScaler from “preprocessing” set of sklearn module for the ordered variables. For the categorical ones, I have used One-Hot Encode method to make data suitable for analysis.

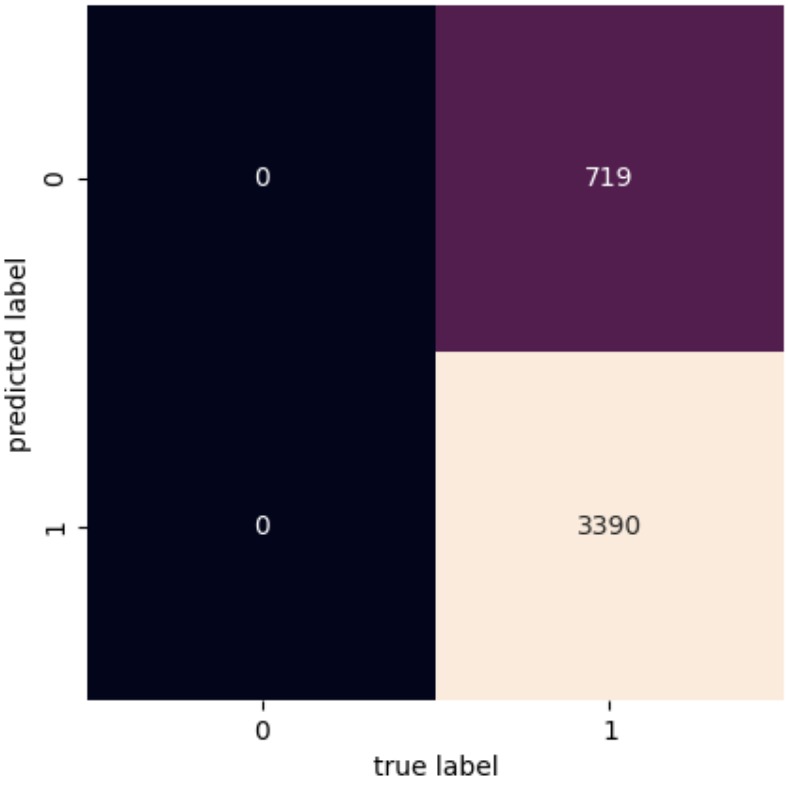
As part of the preprocessing, I have used SelectKBest function with f_classif criterion, to use the statistically best variables for our purposes and decrease the number of variables for efficiency.

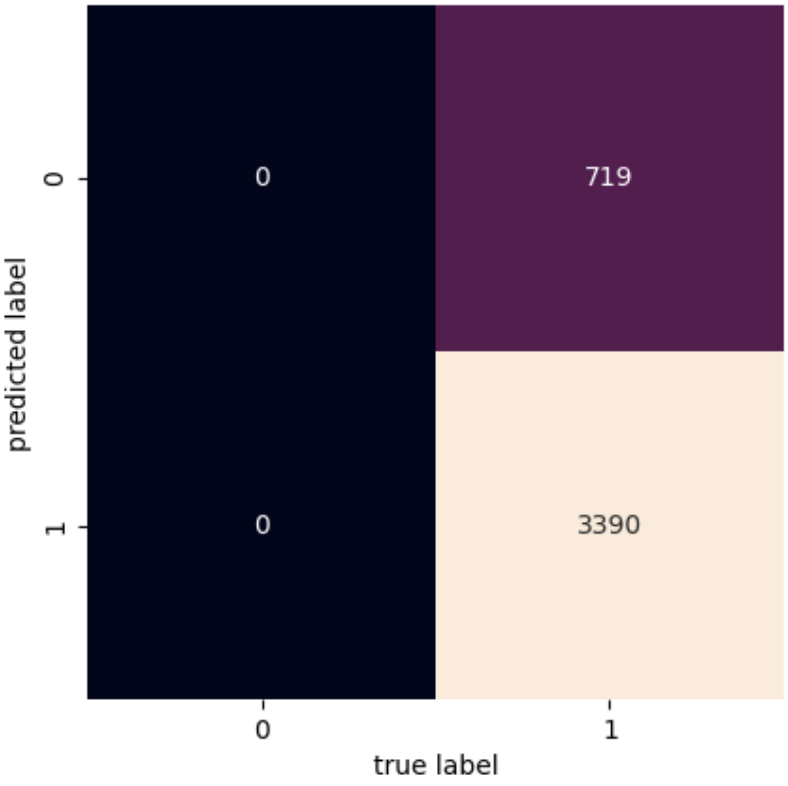
As part of the model selection, I have created cross validation scores of different machine learning models: Decision Tree, KNeighbor, Support Vector Machine, Gaussian Naïve Bayes and Logistic Regression. According to the cross-validation scores, Support Vector Machine and Logistic Regression are the most efficient models.

| Cross Validation Scores (5-fold) Follow as: | |
|---|---|
| KNeigbors | 0.8046019026600169 |
| Decision Tree: | 0.8198272777823279 |
| Support Vector Machine | 0.8206665344776306 |
| Gaussian Naive Bayes | 0.7796684641501896 |
| Logistic Regression | 0.8206665344776306 |

I have trained the models with the data. I have used GridSearchCV function to identify the best model conditions. At that point, even though my results identified the best settings, my accuracy scores based on the real data and predictions have not increased.

I have produced confusion matrices for the models. They both gave the same plot mysteriously. It shows that %17 of results are false negatives.


0
719
0
3390
predicted label
0
1
0
1
true label


0
719
0
3390
predicted label
0
1
0
1
true label