Begüm Arslan

CSSM502: HOMEWORK3

In this machine learning project, “cses4_cut.csv” file that contained subset of CSES Wave Four data set was used. Initially, data was analyzed and found out that it included some categorical variables. Thus, the categorical variables were one-hot encoded. The dataset also did not contain any null values, thus there wasn’t any need left for imputing. Then, the dataset was divided into X and y, which the target variable included if the respondent voted in the last presidential election or not. Then, sklearn’s train-test-split method was applied which the test size was given as 30% of the whole dataset.

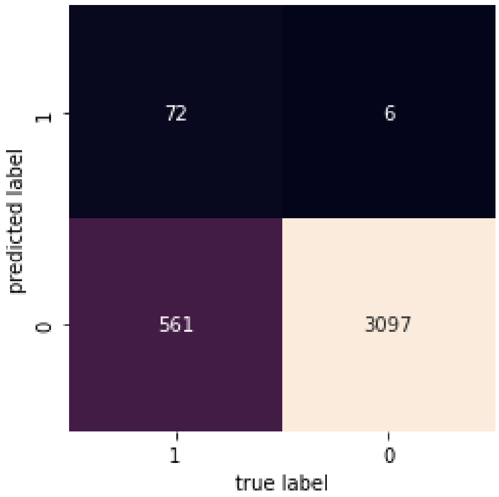
- **Model without Scaling and Feature Selection**

At first, the raw data which was one-hot encoded was used with Random Forest Classifier model. In order to hypertune the parameters, GridSearchCV was used which the parameters of n_estimators, max_features, max_depth and criterion was searched for the model in order to find the best fit. The best parameter of GridSearchCV for this part came as follows:

```
{'criterion': 'entropy',
     'max_depth': 8,
 'max_features': 'auto',
   'n_estimators': 200}
```

Thus, the outputs of GridSearchCV was used while modelling with Random Forest Classifier and Accuracy score came as: 0.848. The resulting confusion matrix is as follows:

In order to see if the model overfit or not, cross validation was used and the average score was taken which resulted as 0.839. This result shows that the model gives consistent results.



As it can be seen, the model classified 72 people who voted and 3097 people who did not voted correctly and it misclassified 567 points.

- **Feature Selection**

SelectKBest algorithm was used in order to get the most important features to take into account and reduce number of features of dataset. First 10 features with highest score is given below:

```
[('D2021', 23616.010551455885),
 ('D2022', 9679.99694947531),
 ('D2023', 7728.432886280269),
 ('age', 4588.480589050823),
 ('D2004_3', 567.9176851522903),
 ('D2010_5', 488.1860541958613),
 ('D2014_8', 225.182953263611),
 ('D2013_8', 199.21572514062333),
 ('D2025_8', 181.93018300427522),
 ('D2003_99', 142.86322557527365)]
```

New dataset was created as X_selected with the 10 highest scored features.

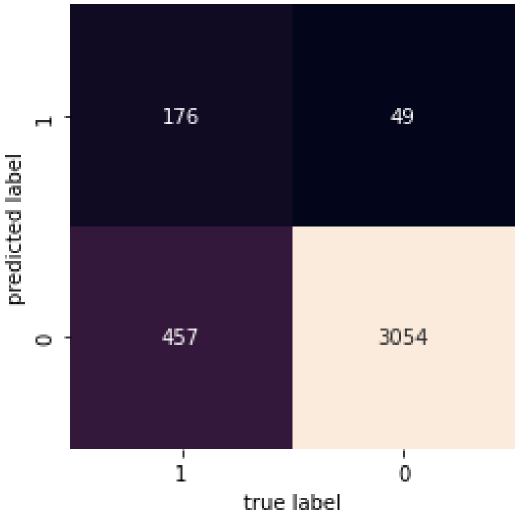
- **Modeling with Feature Selection, Without Scaling**

X_selected was used as input data. Two models of Random Forest Classifier and KNeighborsClassifier was used. Random Forest Classifier was hypertuned with GridSearchCV and the best parameters are as follows:

```
{'criterion': 'gini',
 'max_depth': 7,
 'max_features': 'auto',
 'n_estimators': 200}
```

The training data was fit into hypertuned model and fit into test data. The accuracy came out as: 0.865.

Resulting confusion matrix is as follows:



As it can be seen, the model classified 176 people who voted and 3054 people who did not voted correctly and it misclassified 506 points.

In order to see if the model overfit or not, cross validation was used and the average score was taken which resulted as 0.858. This result shows that the model gives consistent results.

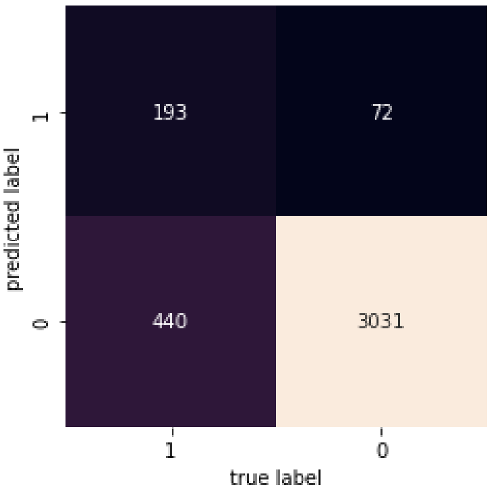
As a second model, KNeighborsClassifier was hypertuned with parameters of leaf_size, n_neighbors and p. The best parameters resulted as follows:

```
{'algorithm': 'auto',
 'leaf_size': 31,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': None,
 'n_neighbors': 27,
 'p': 1,
 'weights': 'uniform'}
```

The hypertuned model was fit with X_train and y_train and test data was used to predict new unseen points. The accuracy of the model resulted as 0.863.

Cross validation was applied and gave the output of 0.857.

The resulting confusion matrix is as follows:



As it can be seen, the model classified 193 people who voted and 3031 people who did not voted correctly and it misclassified 512 points.

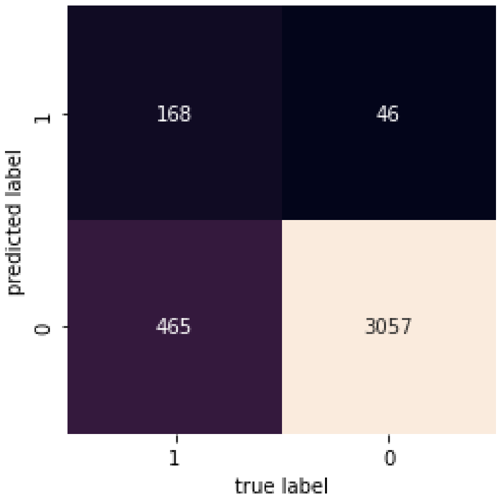
- **Modelling with Feature Selection and Scaling**

A pipeline was created in order to scale and classify the data. Standard Scalar was used in order to normalize the data. Two models of Random Forest Classifier and K Neighbors Classifier were analyzed to see which model gave the best results at our case. GridSearchCV was used to find the best parameters of the models. As a result, Random Forest Classifier gave the best results with the following parameters:

```
Pipeline(steps=[('scaler', StandardScaler()),
                ('classifier', RandomForestClassifier(max_depth=5))])
```

A new pipeline with the given best parameters were fitted into the model and gave accuracy of 0.863. Cross validation was used in order to see if the model overfit and if it can give consistent results with the other folds of the data. Cross validation resulted with 0.856 accuracy of mean which means that model gives consistent results.

The resulting confusion matrix is as follows:



As it can be seen, the model classified 168 people who voted and 3057 people who did not voted correctly and it misclassified 511 points.

As a result, it can be seen that Random Forest can deal with nonscaled data, thus scaling is not necessary for Random Forest. Also, Feature selection increased the accuracy of the model which shows that feature selection is an important part of modelling.