CSSM502

ADVANCED DATA ANALYSIS WITH PYTHON

BATUHAN HÖKELEK

HOMEWORK 3

Introduction:

In this project, I built a predictive model to understand the likelihood of a respondent to vote in the last presidential election. I used "cses4_cut.csv" file which includes a subset of the CSES Wave Four data set.

Different classifiers and regressors were tested without pre-processing and dimensionality reduction.

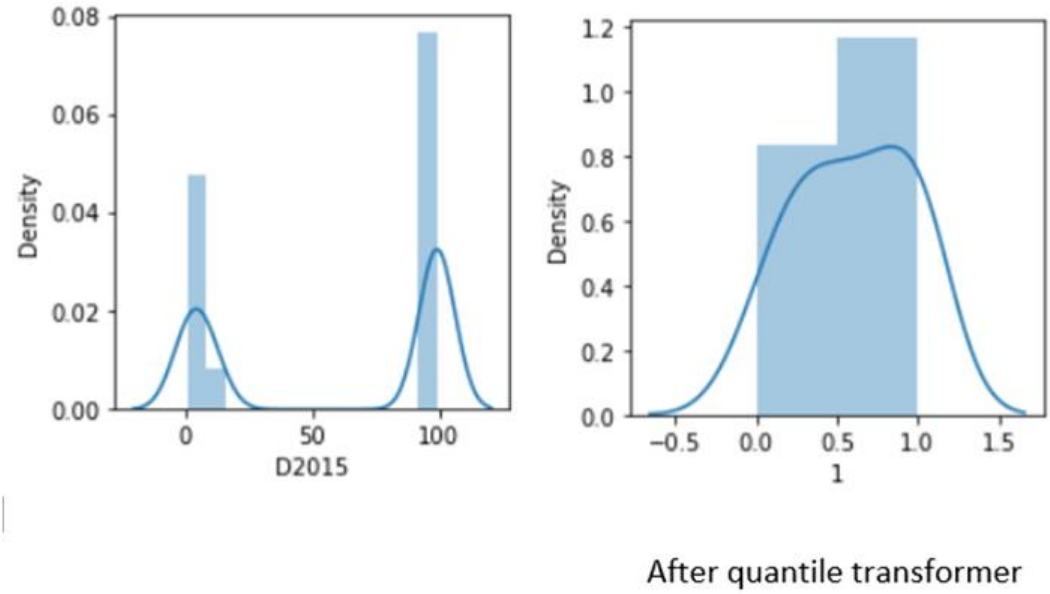
The following results were yielded.

	Model	Accuracy
6	Random Forest	86.64%
1	K-Nearest Neighbors	84.47%
2	Linear Discriminant Analysis	83.75%
0	Logistic Regression	83.26%
4	Support Vector Machine	82.47%
3	Decision Tree	78.19%
5	Quadratic Discriminant Analysis	69.86%
7	Bayes	69.34%

Feature Selection

This part involved selecting the features that best predict my target variable to reduce overfitting, training time and to improve accuracy. The highest 12 features were selected using sklearn.feature_selection.SelectKBest function and took 12 features with the highest k scores which are:
D2011, D2015, D2016, D2021, D2022, D2023, D2026, D2027, D2028, D2029, D2030 and age

I have transformed the new data set with 12 highest features in Gaussian form, and eliminated unwanted data which disrupt the distribution of my data. To be able to do this, I have used quantile transformer method which transforms the feature to be able to follow normal distribution or uniform. This method is also useful to remove outliers and spread out the most frequent values.



After quantile transformer

Classifiers with Dimensionality Reduction and Pre-processing

After pre-processing and feature selection, I re-trained the models.
Results are as follows:

	Model	Accuracy
6	Random Forest	85.94%
4	Support Vector Machine	84.99%
2	Linear Discriminant Analysis	83.54%
0	Logistic Regression	83.52%
1	K-Nearest Neighbors	83.40%
5	Quadratic Discriminant Analysis	78.51%
3	Decision Tree	78.16%
7	Bayes	77.45%

Optimizing the model and its hyperparameters

I have selected the top 5 highest classifiers and regressors according to their k scores I have looped them until I have found the best hyperparameters. My results are as follows:

	Model	Accuracy
3	Random Forest	86.11%
1	Support Vector Machine	85.65%
4	K-Nearest Neighbors	84.23%
2	Linear Discriminant Analysis	83.54%
0	Logistic Regression	83.54%

Best results yielded with these parameters:

```
Best score is: 0.8610813704496788 with estimator: 1000 criterion: entropy
Best score is: 0.8565310492505354 with c: 5 kernel: precomputed2
Best score is: 0.835438972162741 with solver: svd
Best score is: 0.8353854389721628 with penalty none
Best score is: 0.8423447537473233 with number of neighbors: 9
```