INTL550: ADVANCED DATA ANALYSIS IN PYTHON
SERHAN TÜRKAN
HOMEWORK 3

Introduction

The purpose of this homework is to review and practice fundamental machine learning concepts. The idea is to build a predictive model of whether a respondent likely voted in their last presidential election. For this purpose, I used "cses4_cut.csv" file which containing a subset of the CSES Wave Four data set.

Trying multiple approaches

I tested different classifiers and regressors to see their behavior without any pre-processing or dimensionality-reduction operation.
Results are as follows:

Model	***Accuracy***
Random Forest	86.64%
K-Nearest Neighbors	84.47%
Linear Discriminant Analysis	83.75%
Logistic Regression	83.26%
Support Vector Machine	82.47%
Decision Tree	78.20%
Quadratic Discriminant Analysis	69.86%
Bayes	69.34%

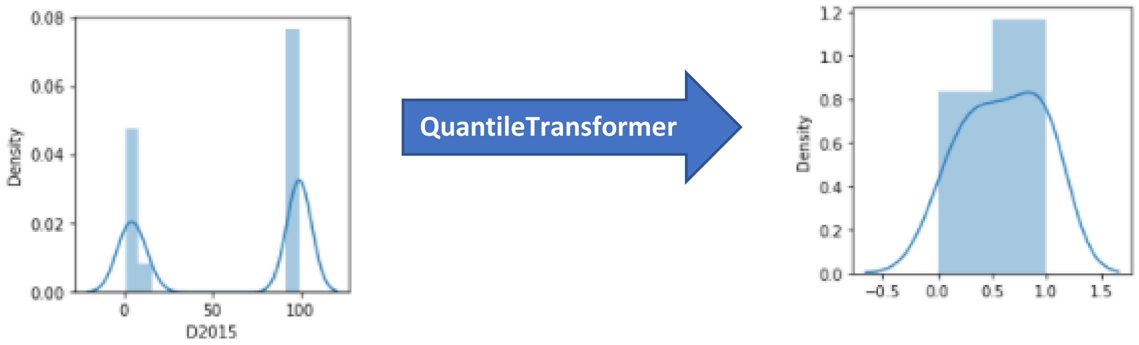
Dimensionality-reduction with feature selection

Feature selection is a technique where we choose those features in our data that contribute most to the target variable. In other words, we choose the best predictors for the target variable. With feature selection we can reduce overfitting, improve accuracy, and reduce training time. For this purpose I used **sklearn.feature_selection.SelectKBest** function and took 12 features with the highest score which are:
D2011, D2015, D2016, D2021, D2022, D2023, D2026, D2027, D2028, D2029, D2030, age

Pre-processing

There are some unwanted data in data set like:
7.REFUSED
8.VOLUNTEERED: DON'T KNOW
9.MISSING

These data disrupt the distribution of data. I used quantile transformer method **(sklearn.preprocessing.QuantileTransformer**) to solve this problem. This method transforms the features to follow a uniform or a normal distribution. Therefore, for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of outliers.



Classifiers with dimensionality-reduction and pre-processing

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

Model	***Accuracy***
Random Forest	86.03%
Support Vector Machine	84.99%
Linear Discriminant Analysis	83.54%
Logistic Regression	83.52%
K-Nearest Neighbors	83.40%
Quadratic Discriminant Analysis	78.51%
Decision Tree	78.35%
Bayes	77.45%

Optimizing the model and its hyperparameters

I took the top 5 classifiers and regressors and looped them until I found the best hyperparameters.
Results are as follows:

Model	***Accuracy***
Random Forest	86.04%
Support Vector Machine	85.65%
K-Nearest Neighbors	84.23%
Linear Discriminant Analysis	83.54%
Logistic Regression	83.54%

Best results were achieved with these parameters:

Random Forest Classifier	:n_estimators=1000, criterion='entropy'
Linear Discriminant Analysis	: solver= 'svd'
Logistic Regression	: penalty='none'
K-Nearest Neighbors	:n_neighbors= 9
Support Vector Machine	: C=5, kernel='precomputed2'