Berra Karayel

0054477

David Carlson

CSSM 502 Third Assignment

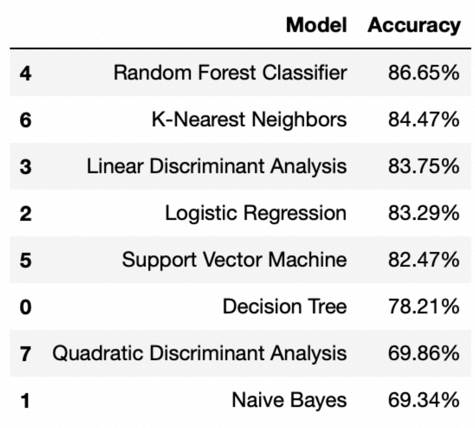
Introduction

Data Set: In this analysis, I have used “cses4_cut.csv” data set which is the subset of the CSES Wave Four data set.

Purpose of the analysis: I have create a predictive model to be able to understand the likelihood of respondents to vote in their last presidential election.

Classifiers Without Reduction and Without Pre-Processing

Without pre-processing and dimensionality-reduction operations, I have tested different classifiers and regressors to see voting behavior of respondents. Here is my results:



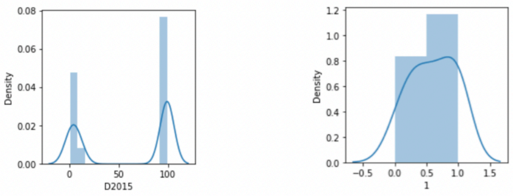
	Model	Accuracy
4	Random Forest Classifier	86.65%
6	K-Nearest Neighbors	84.47%
3	Linear Discriminant Analysis	83.75%
2	Logistic Regression	83.29%
5	Support Vector Machine	82.47%
0	Decision Tree	78.21%
7	Quadratic Discriminant Analysis	69.86%
1	Naive Bayes	69.34%

Feature Selection

In this part, I have chosen the best predictors for my target variable to reduce overfitting and training time and improve accuracy. I have selected 12 features with highest k scores by using *sklearn.feature_selection.SelectKBest* which are: D2011, D2015, D2016, D2021, D2022, D2023, D2026, D2027, D2028, D2029, D2030, and age.

Pre-Processing

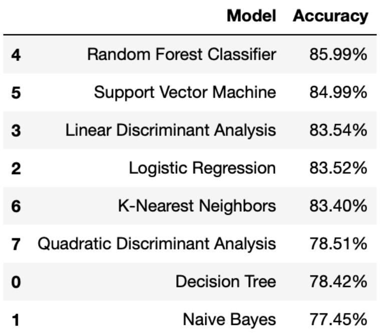
I have transformed the new data set with 12 highest features in Gaussian form, and eliminated unwanted data which disrupt the distribution of my data. To be able to do this, I have used *quantile* *transformer method* which transforms the feature to be able to follow normal distribution or uniform. This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



After quantile transformer

Classifiers with Dimensionality-Reduction and Pre-Processing

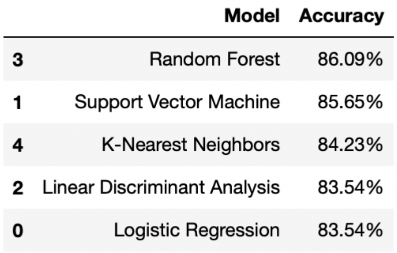
After pre-processing and feature selection, I have retrained the models. Here is my results:



	Model	Accuracy
4	Random Forest Classifier	85.99%
5	Support Vector Machine	84.99%
3	Linear Discriminant Analysis	83.54%
2	Logistic Regression	83.52%
6	K-Nearest Neighbors	83.40%
7	Quadratic Discriminant Analysis	78.51%
0	Decision Tree	78.42%
1	Naive Bayes	77.45%

Optimizing the Model and Its Hyperparameters

I have chosen the top 5 highest classifiers and regressors based on their k scores. I have looped them until I have found the best hyperparameters. Here is my results:



	Model	Accuracy
3	Random Forest	86.09%
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:

Random Forest Classifier: Best score is 0.8609207708779444 with estimator 200, criterion gini

Support Vector Machine: Best score is 0.8565310492505354 with c:5, kernel:precomputed2

Linear Discriminant Analysis: Best score is 0.835438972162741 with solver:svd

Logistic Regression: Best score is 0.8353854389721628 with penalty none

K-Nearest Neighbors: Best score is 0.8423447537473233 with number of neighbors: 9