CS 418 Data Science

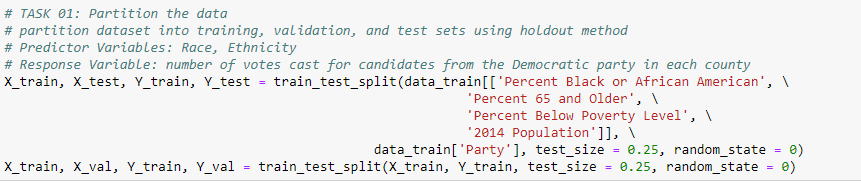
Project 01: Exploratory Data Analysis

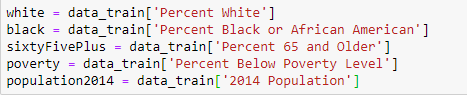
Group Members: Vani Jaiswal, Mudit Kumar, Ahmed Khan

The libraries we used for this project are:

A screenshot of a cell phone

Description generated with very high confidence

1. We partitioned the dataset into a training set and a validation set using the holdout method. We used the following predictor variables: “Percent Black or African American”, “Percent 65 and Older”, “Percent Below Poverty Level”, and “2014 Population”.



1. We standardized the training set and validation set

**A screenshot of a cell phone

Description generated with high confidence**

1. We built several simple linear regression models using different predictor variables to predict the number of votes cast for candidates from the Democratic party in each county. Among all the models, the best performing simple linear regression model was with the predictor variable “Democratic ~ population2014”. The performance of this model’s R-squared value is 0.841. We selected this variable because the county data suggest that the higher population is, the more likely the county is to be Democratic. Then we computed the summary of the evaluation metrics of the model. This shows that the model is strong.

**A screenshot of a social media post

Description generated with very high confidence**

1. We built multiple linear regression models using different combinations of predictor variables to predict the number of votes cast for candidates from the Democratic party in each county. Among all the models, we found that the best performing multiple linear regression model was with the predictor variables “Democratic ~ white”, “Democratic ~ black”, “Percent 65 and over”, “Poverty”, “Population 2014”. The performance of this model’s R-squared value is 0.843. We selected this combination of variables because it has the highest R-squared value. Then we computed the evaluation metrics.

**A screenshot of a social media post

Description generated with very high confidence**

1. We built several simple linear regression models using different predictor variables to predict the number of votes cast for candidates from the Republican party in each county. Among all the models, the best performing simple linear regression model was with the predictor variable “Republican ~ population2014”. The performance of this model’s R-squared value is 0.621. We selected this variable because the lower the population is, the more likely the county is to be Republican. Then we computed the evaluation metrics.
2. **A screenshot of a social media post

   Description generated with very high confidence**

We built multiple linear regression models using different combinations of predictor variables to predict the number of votes cast for candidates from the Republican party in each county. Among all the models, we found that the best performing multiple linear regression model was with the predictor variables “Republican ~ white”, “Republican ~ black”, “Percent 65 and over”, “Poverty”, “Population 2014”. The performance of the model’s R-squared value is 0.638. We selected this variable because this combination of variables gives us the highest R-squared value. Then we computed the evaluation metrics.

**A screenshot of a social media post

Description generated with very high confidence**

**7)**

We then built a classification model in order to classify each county as Democratic or Republican. We initially used K Neighbor classifier that resulted in a good outcome. We attempted to use linear SVM to classify each county, but after fifteen minutes of standby time we were unable to receive an output. We finally used logistic regression classifier and received the best outcome of the three models, with 75% accuracy. We have listed a confusion matrix of our data and the evaluation metrics computed:A screenshot of a social media post

Description automatically generated