Data Science

Prediction of 2018 US Senate Office election result

Name	Net Id
Snehal Tikare	stikar2
Cesar Hernandez	cherna83
Ubemio Romero	uromer2

Task 1:

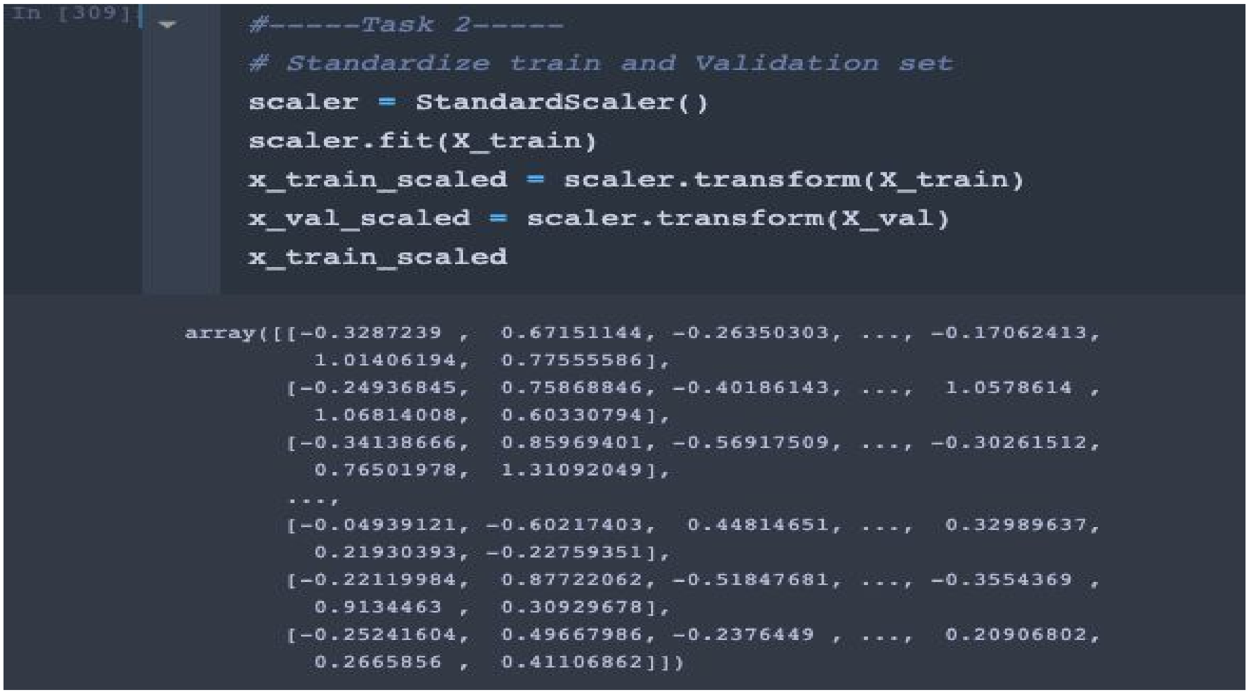
How did you partition the dataset?

- We partitioned the data using the Holdout Method. We set our training set to be 75% while our validation set equal to 25%. Our “merged_train” data consist of 1195 observation.After partitioning the data we get X_train, Y_train to be at 896 observations. Likewise, X_val, Y_val to be at 299 observations.

Task 2:

Standardize training set and validation set

- Rescaling our attributes to normalize the values



```
In [309]:
#-----Task 2-----
# Standardize train and Validation set
scaler = StandardScaler()
scaler.fit(X_train)
x_train_scaled = scaler.transform(X_train)
x_val_scaled = scaler.transform(X_val)
x_train_scaled
```

```
array([[-0.3287239 ,  0.67151144, -0.26350303, ..., -0.17062413,
         1.01406194,  0.77555586],
       [-0.24936845,  0.75868846, -0.40186143, ...,  1.0578614 ,
         1.06814008,  0.60330794],
       [-0.34138666,  0.85969401, -0.56917509, ..., -0.30261512,
         0.76501978,  1.31092049],
       ...,
       [-0.04939121, -0.60217403,  0.44814651, ...,  0.32989637,
         0.21930393, -0.22759351],
       [-0.22119984,  0.87722062, -0.51847681, ..., -0.3554369 ,
         0.9134463 ,  0.30929678],
       [-0.25241604,  0.49667986, -0.2376449 , ...,  0.20906802,
         0.2665856 ,  0.41106862]])
```

Task 3:

What is the best performing linear regression model?

The Best performing linear model for predicting number of democratic votes is the one that has the following predictor variables: Total Population, Population that is Black, Population with less than a bachelor's degree, and the unemployed population.

What is the performance of the model?

- For Democratic R square value-0.9506 and Adjusted R square -0.9503
- For Republican the R-squared: 0.7323 and Adjusted R-Square 0.7293
- As we can see, the Democratic model represents a better fit model based on our model score and R_squared. Values are closer to the value of one.
- We were able to determine the that total population would be a default predictor based on a linear regression single test. We then ended up with 8 combination and form those combinations we started to drop values based on democratic parties intuition. We also dropped predictors based on democratic influences and confirmed by performing multiple linear regression tests.

How did you select the variables of the model?

- First we made a linear regression test using “Total Population”. This proved to have a well corresponding value and was able to confirm that would be a strong variable to use for our model.
- Based on our analysis in project 1, we chose an initial set of attributes for both democratic and republican votes.Then, using Linear regression, we added and dropped attributes according to the adjusted R square value.We kept the attribute that increased the R-Square and dropped the one's, that reduced the model's performance
- We dropped and added attributes that influence the Democratic party by public knowledge. After having a strong result, we then started to make combinations of different variables to ensure it was the maximum value possible. After making sure we had the best possible combination we verify by dropping variables to guarantee a robust performance. We were able to get optimal results by having [“Total Population”, “Percent Black not Hispanic or Latino”, “Percent Unemployment”, “Percent Less than Bachelor’s Degree”]. Thus verified.
- At first glance we also used Lasso Regression a technique that would allow us to have an idea of what we should be aiming for. However Lasso does not guarantee a good result for this case. We’re dealing with a large data set and since we’ve reduced our data to what matters we know that every attribute can be held accountable. Lasso can potentially ignore some attributes by selecting some of the attributes and not all of them. In the case of linear regression we are evaluating and the more attributes the better predictions.

  -The same method was followed for selecting attributes for predicting Republican votes

Task 4:

What is the best performing classification model?

The best performing classification model is SVM. We tested four types of models. Decision Tree, Naive Bayes Classifiers, SVM, and K-nearest Neighbours with different parameters and attributes set. SVM had a better accuracy, F1 score and a less percent error.

What is the performance of the model?

- Accuracy -0.8562
- F1 Score -0.8439

How did you select the parameters of the model?

- We chose the parameters by considering the ones that increase the accuracy and f1 score on a selected set of attributes

How did you elect the variables of the model?

- Using the analysis from Project 1, we summarized the below observations.
- Attributes 'White population', ' Less than Bachelor's degree','Age 65 and over' favored Republican
- Attributes 'Black population','Foreign Born','Age under 29' favored Democrats
- Attributes 'Median Household income', 'Unemployment','Hispanic or Latino','Percent Rural' do not favor any party
- So choosing these attributes to classify the labels with higher accuracy and f1 score And selecting other attributes that increase the accuracy and f1 score
- We followed the same intuition we used in the regression model to choose the attributes.

Task 5:

What is the best performing clustering model?

- We decided that the best performing cluster model is K means clustering. After considering the other models like DBSCAN but ultimately it came to our performance and after plotting it in the map it reinforced the idea that K means works great for this scenario of votes.

What is the performance of the model?

- The performance on the model is that the Random index 0.19751656022671712 and we get a silhouette coefficient of 0.30700290833697047

How did you select the parameters of the model?

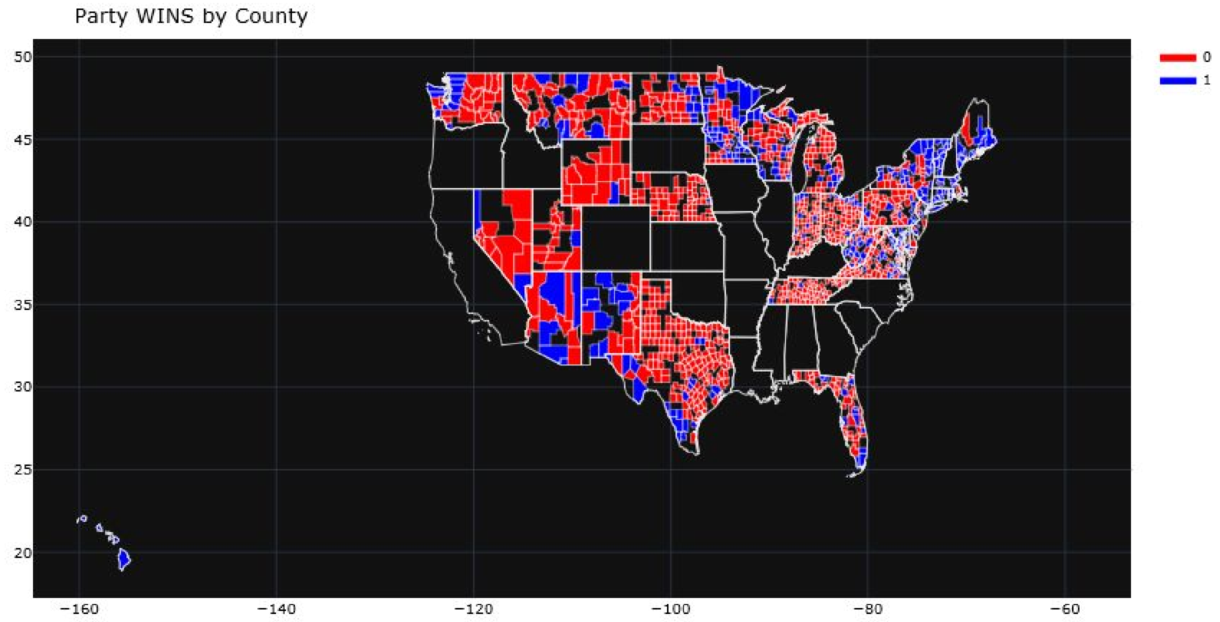
-

How did you select the variables of the model?

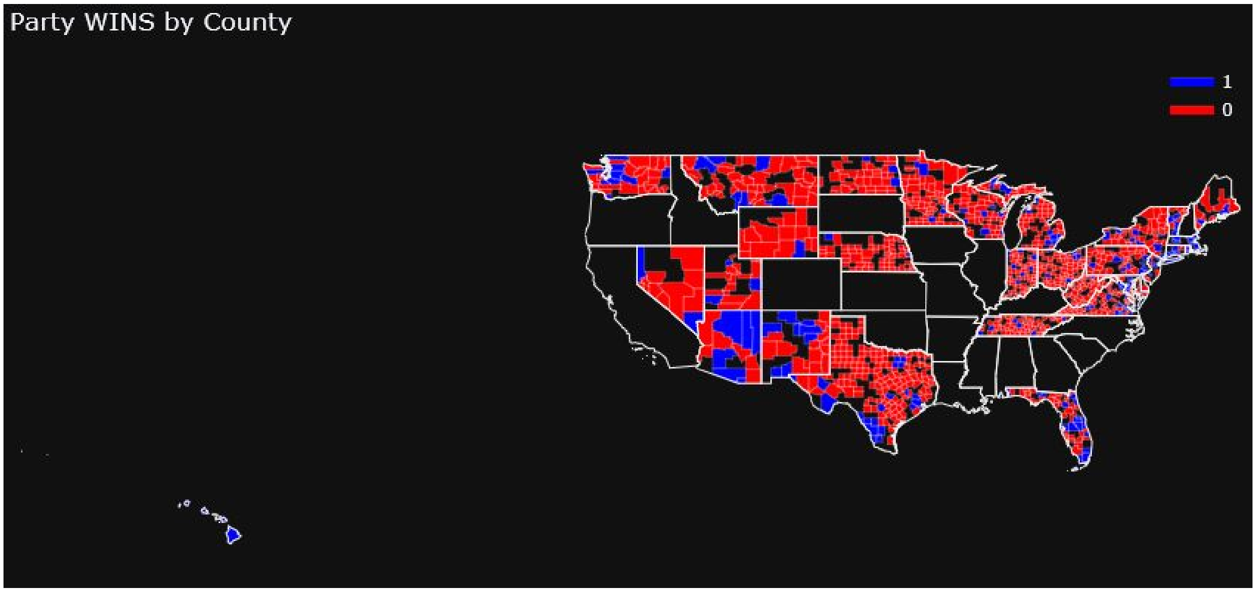
We decided to use the whole set of variables as whenever we tried using some the performance went down and this made sense since we are predicting for both parties whenever we put some variables that favored one party it shifted to that side and vice versa like in the case of the predictor of the black population the predictor works great in the case of democrats but not in republicans and the same goes with the case of rural as it had a greater influence on making them republicans, so in all by using all the predictors we were able to remain unbiased to a party.

Task 6:

Map of Democratic counties and Republican counties created in Project 01



Map of Democratic counties and Republican counties created in Project 02



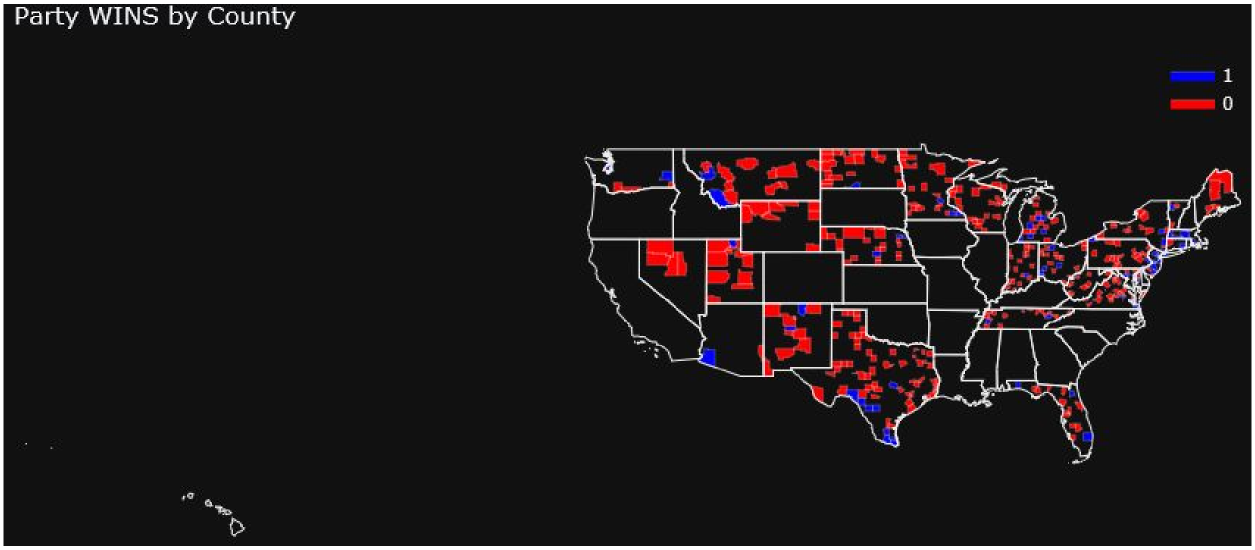
We see some difference in few counties where the built model has predicted different results.

What conclusions do you make from the plots?

- Based on these plots we can make the conclusion that based on project 1s map that the surrounding counties can play a major role as to what the county we are predicting for will vote a certain way as we can see in cases like Wyoming and Nevada the surrounding counties had an influence as to what the counties will vote either on those grounds or on literal grounds because depending on adjacent counties the land, ideas, and values will be similar to those around it so as to why we can see that most of the time the predicted counties vote leanings matched to what county's surround them from project 1's map.

Task 7:

Plot for predicted result of the county



Output saved in file output.csv file