CS 418: Project 2 Report

Authors: Anusha Sagi, Fatima Kahack, Lydia Tse

1. (5 pts.) Partition the merged dataset into a training set and a validation set using the holdout method or the cross-validation method. How did you partition the dataset?

Answer: We have partitioned the data into training and validation sets using hold-out method by separating 75% into train data and 25% into test data.

3. (25 pts.) Build a linear regression model to predict the number of votes cast for the Democratic party in each county. Consider multiple combinations of predictor variables. Compute evaluation metrics for the validation set and report your results. What is the best performing linear regression model? What is the performance of the model? How did you select the variables of the model?

- **Repeat this task for the number of votes cast for the Republican party in each county.**

Variable Selection Rationale

We selected the variables for these models because we are aware that political party affiliation is very closely related to ethnicity and education of individuals. Additionally, we are aware that the location of the individuals can be indicative of political lines as well. Thus, we felt that they would be the best indicators of party affiliation.

Using Regression to predict Democratic values

Model Number	Description	R^2	$Adj.R^2$	RMSE	Conclusion
1	Linear Regression using all variables	0.934	0.931	-14771.995	It has a good Regression model with high Adjusted R- squared and moderate Root Mean Square Error (RMSE).
2	LASSO Regression with all variables	0.934	0.931	14768.885	LASSO Regression is slightly better than Model 1 since we see slightly higher Adj. R-squared value and a lower RMSE.
3	Linear Regression using 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or	0.927	0.926	14592.862	This model has a lower Adj. R-squared value and lower RMSE than the previous two models. This is not a better model than the above two.

	Latino', 'Percent Foreign Born'				
4	Linear Regression with variables ‘Total Population’, ‘Percent Black, not Hispanic or Latino’, ‘Percent Less than Bachelor's Degree’	0.951	0.950	12456.893	This model is the best model as it has higher R squared value and a significantly lower RMSE than the above 3 models.

Conclusion

Since Model 4 has the highest R-squared values and significantly lower RMSE values than the all of the other models, we conclude that this is the best regression model.

Using Regression to predict Republican values

Model Number	Description	R^2	$Adj. R^2$	RMSE	Conclusion
1	Linear Regression using all variables	0.724	0.711	15962.431	It has a good regression model with moderate Adj. R- squared and moderate RMSE.
2	LASSO Regression with all variables	0.724	0.711	15962.567	LASSO Regression is similar to the above regression model which includes all the variables as we can see almost a similar Adj. R-squared and similar RMSE.
3	Linear Regression using 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'	0.670	0.665	17111.714	This model has is not the best performing model as it has low Adjusted R squared value and very high Root mean square error than the previous two models.
4	Linear Regression with variables 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born',	0.730	0.723	15749.246	This model which includes the above predictors is the best model we have tried as it has higher R squared value and a very lower Root mean square error than the above 3 models.

	'Percent Age 65 and Older', 'Percent Unemployed', 'Median Household Income', 'Percent Rural'				

Conclusion

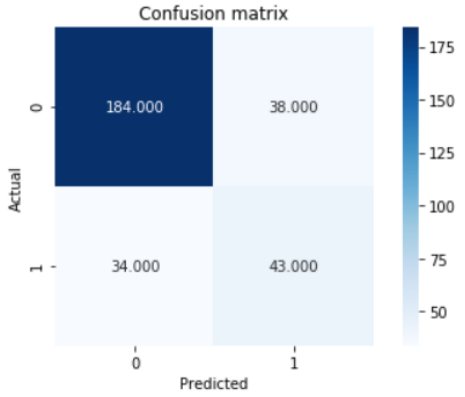
Because Model 4 has both high R-squared values and lower RMSE than the other models, we conclude that this is our best performing regression model. However, we recognize that there is still room for improvement in the model since the R-squared values are not as high as we'd like.

4. (25 pts.) Build a classification model to classify each county as Democratic or Republican. Consider at least two different classification techniques with multiple combinations of parameters and multiple combinations of variables. Compute evaluation metrics for the validation set and report your results. What is the best performing classification model? What is the performance of the model? How did you select the parameters of the model? How did you select the variables of the model?

Answer:

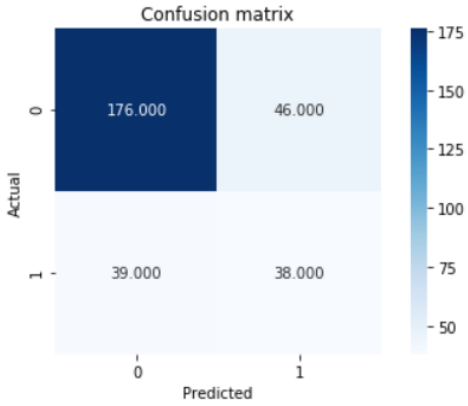
Decision Tree Classifier: To predict each county as Democratic or Republican

a) **Model 1 - Using all variables and entropy**



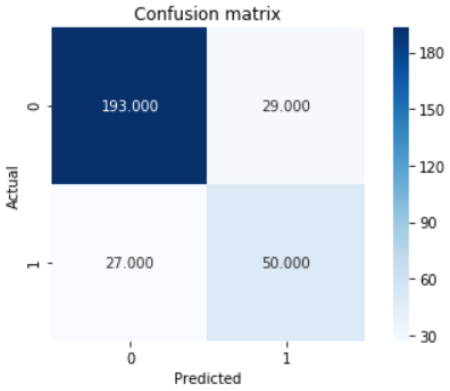
No of nodes: 197
Accuracy: 0.759
Error: 0.241
Precision: [0.8440367, 0.5308642]
Recall: [0.82882883, 0.55844156]
F1-score: [0.83636364, 0.5443038]

b) Model 2 - Using all variables and the Gini index



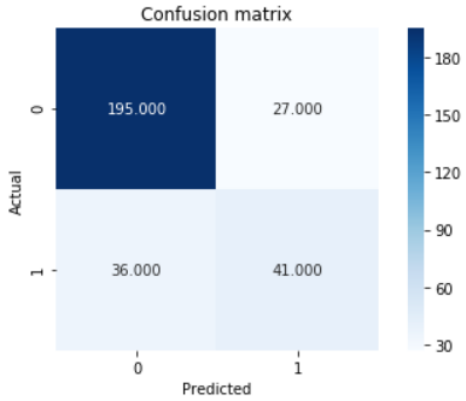
No of nodes: 231
Accuracy: 0.716
Error: 0.284
Precision: [0.81860465, 0.45238095]
Recall: [0.79279279, 0.49350649]
F1-score: ([0.80549199, 0.47204969]

c) Model 3 - Using some variable and entropy - 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Less than Bachelor's Degree'



No of nodes: 263
Accuracy: 0.813
Error: 0.187
Precision: [0.87727273, 0.63291139]
Recall: [0.86936937, 0.64935065]
F1-score: [0.87330317, 0.64102564]

d) Model 4 - Using some variable and gini - 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Less than Bachelor's Degree'



No of nodes: 253
Accuracy: 0.789
Error: 0.211
Precision: [0.84415584, 0.60294118]
Recall: [0.87837838, 0.53246753]
F1-score: [0.86092715, 0.56551724]

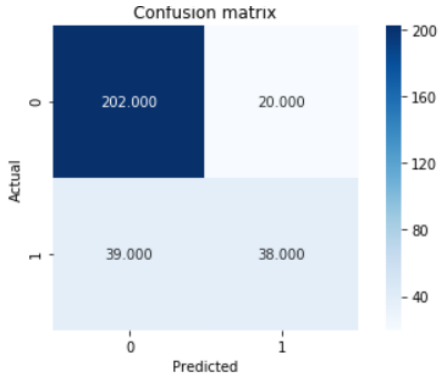
Even though this model is able to Party - '0' i.e., Republican very accurately, the confusion matrix for Party - '1' i.e., Democratic is not predicted well. Hence, this model can't be taken as the best model.

Conclusion: Model 3 is the best performing model so in classification. As, it was able to classify good number of democratic and republican parties. It has the highest F1-score combined for both republican and democratic than any other model. We used the decision tree classifier with criterion as "Entropy" and with all the above variables mentioned and with min_samples_leaf = 2.

K-Nearest Neighbors - To predict each county as democratic or republican

a) **Model 1 - Using all variables with k = 3**

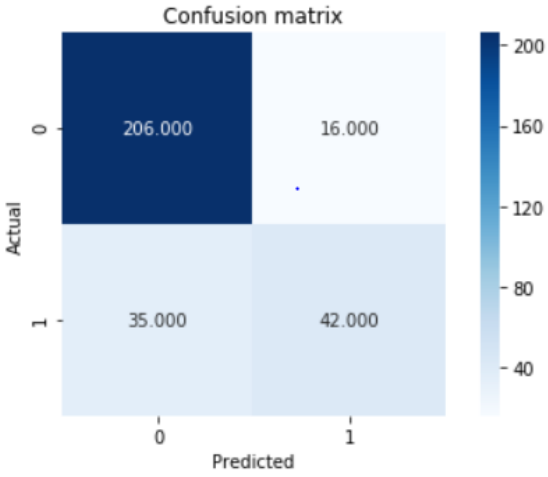
Confusion Matrix:



Accuracy: 0.793
Error: 0.207
Precision: [0.80769231, 0.69230769]
Recall: [0.94594595, 0.35064935]
F1-score: [0.87136929, 0.46551724]

Even though this model is able to Party - '0' i.e., Republican very accurately, the confusion matrix for Party - '1' i.e., Democratic is not predicted well and the f1-score for Party '1' is very low. Hence, this model can't be taken as the best model.

b) **Model 2 - Using some variables with k = 3 - 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Less than Bachelor's Degree'**

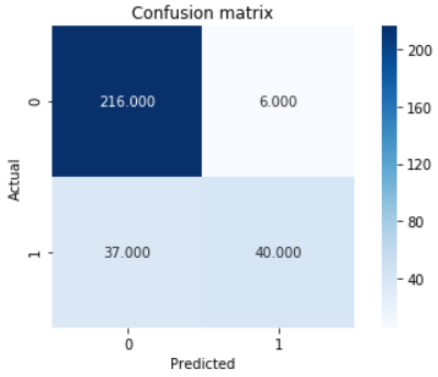


Accuracy: 0.829
Error: 0.171
Precision: [0.85477178, 0.72413793]
Recall: [0.92792793, 0.54545455]
F1-score: [0.88984881, 0.62222222]

Similar to above model, this model is able to Party - '0' i.e., Republican very accurately, the confusion matrix for Party - '1' i.e., Democratic is not predicted well and the f1-score for Party '1' is very low. But this is a better model in K-Nearest Neighbours than the previous model. But, this is not a better model than decision tree model 3. Hence, this model can't be taken as the best model.

Support Vector - To predict each county as democratic or republican

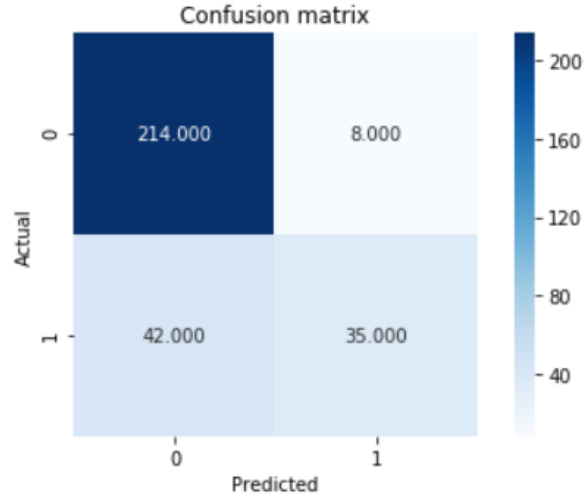
a) **Model 1 - Using all variables**



Accuracy: 0.856
Error: 0.144
Precision: [0.85375494, 0.86956522]
Recall: [0.97297297, 0.51948052]
F1-score: [0.90947368, 0.6504065]

This model has a good F1-score. But this is more biased towards prediction Party '0' than Party ' 1'. A model is a good model when it is able to detect both the observations accurately. This model is not the best performing model as it doesn't better detect the Party '1' than the decision tree classifier (model 3).

b) **Model 2 - Using some variables - 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Less than Bachelor's Degree'**



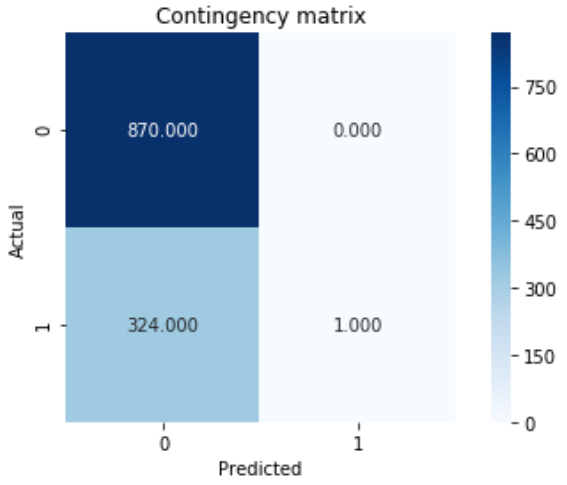
Accuracy: 0.833
Error: 0.167
Precision:[0.8359375 , 0.81395349]
Recall: [0.96396396, 0.45454545]
F1-score: [0.89539749, 0.58333333]

5. (25 pts.) Build a clustering model to cluster the counties. Consider at least two different clustering techniques with multiple combinations of parameters and multiple combinations of variables. Compute unsupervised and supervised evaluation metrics for the validation set with the party of the counties (Democratic or Republican) as the true cluster and report your results. What is the best performing clustering model? What is the performance of the model? How did you select the parameters of model? How did you select the variables of the model?

Answer:

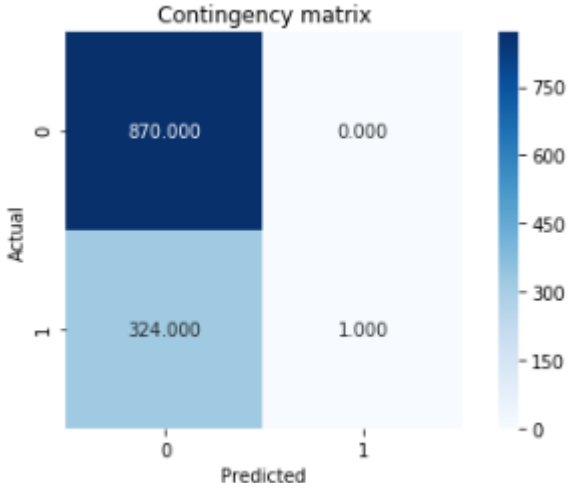
Hierarchical clustering with single linkage method

Model 1: Uses variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'



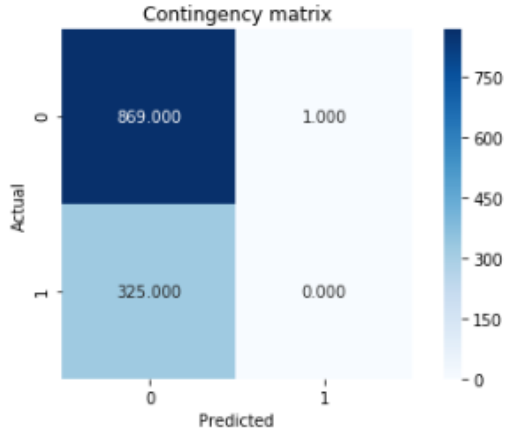
Evaluation Metrics:
Adjusted Rand Index: 0.003
Silhouette Coefficient: 0.697

Model 2: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 29 and Under', 'Percent Age 65 and Older'



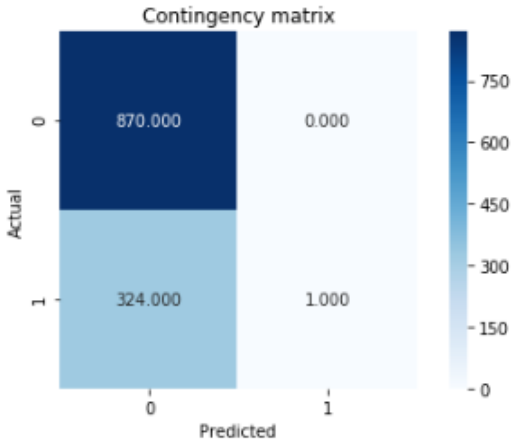
Evaluation Metrics:
Adjusted Rand Index: 0.003
Silhouette Coefficient: 0.671

Model 3: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Female'



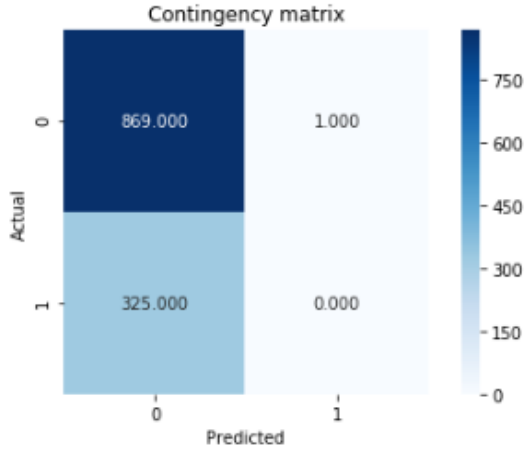
Evaluation Metrics:
Adjusted Rand Index: -0.001
Silhouette Coefficient: 0.795

Model 4: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Less than High School Degree', 'Percent Less than Bachelor's Degree'



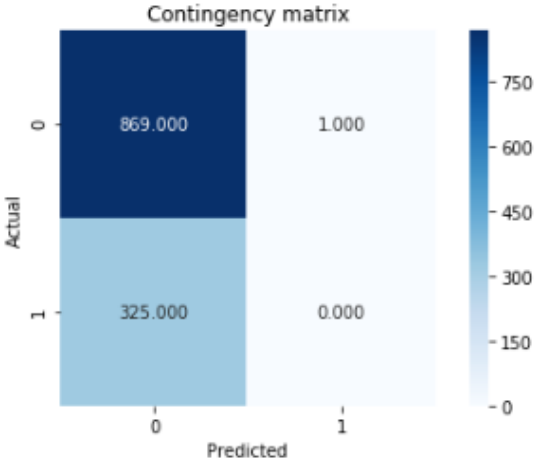
Evaluation Metrics:
Adjusted Rand Index: 0.003
Silhouette Coefficient: 0.675

Model 5: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 29 and Under', 'Percent Age 65 and Older', 'Percent Less than High School Degree', 'Percent Less than Bachelor's Degree'



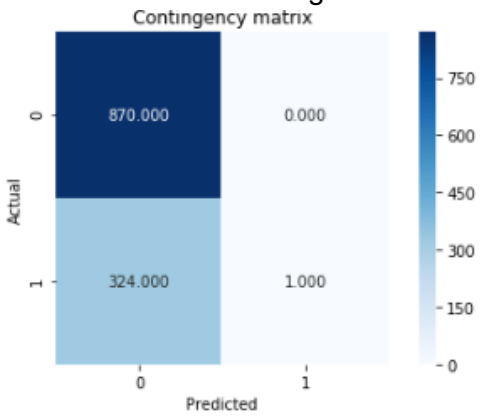
Evaluation Metrics:
Adjusted Rand Index: -0.001
Silhouette Coefficient: 0.422

Model 6: Using variables 'Percent Age 29 and Under', 'Percent Age 65 and Older', 'Percent Less than High School Degree', 'Percent Less than Bachelor's Degree'



Evaluation Metrics:
Adjusted Rand Index: -0.001
Silhouette Coefficient: 0.506

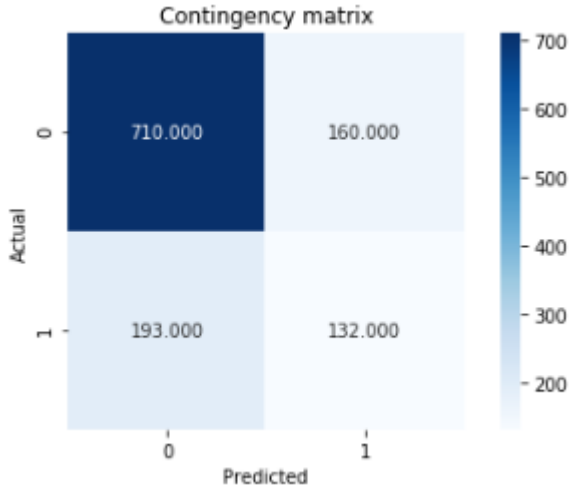
Model 7: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Less than High School Degree', 'Percent Less than Bachelor's Degree'



Evaluation Metrics:
Adjusted Rand Index: 0.003
Silhouette Coefficient: 0.585

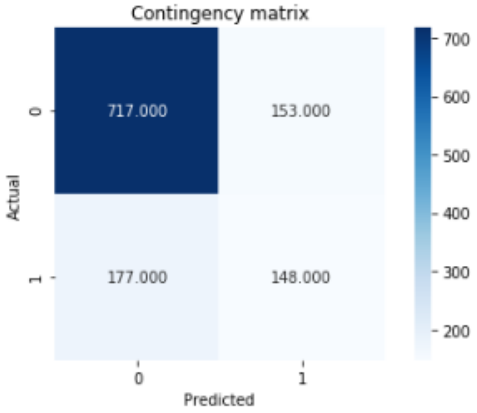
KMeans clustering with k = 2

Model 1: Uses variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'



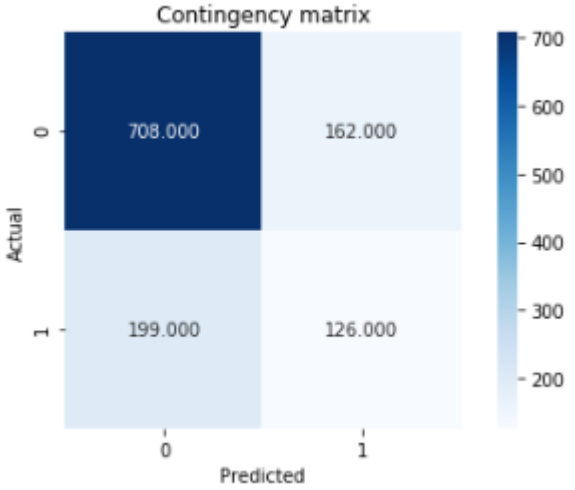
Evaluation Metrics:
Adjusted Rand Index: 0.119
Silhouette Coefficient: 0.582

Model 2: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 29 and Under', 'Percent Age 65 and Older'



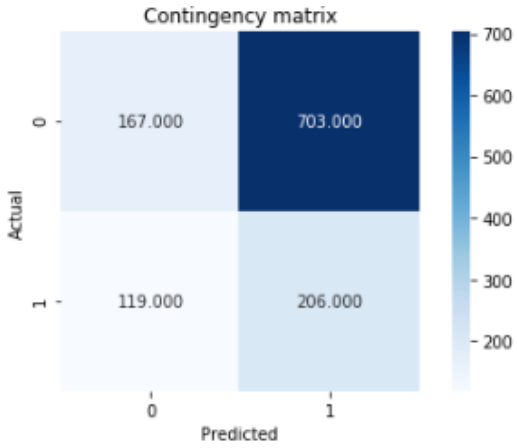
Evaluation Metrics:
Adjusted Rand Index: 0.157
Silhouette Coefficient: 0.291

Model 3: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Female'



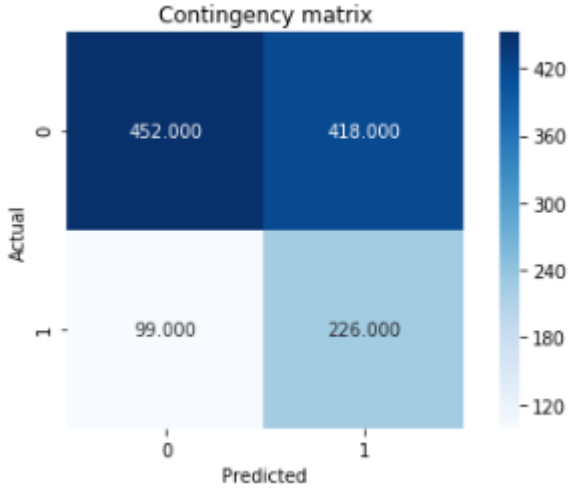
Evaluation Metrics:
Adjusted Rand Index: 0.106
Silhouette Coefficient: 0.522

Model 4: Using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Less than High School Degree', 'Percent Less than Bachelor's Degree'



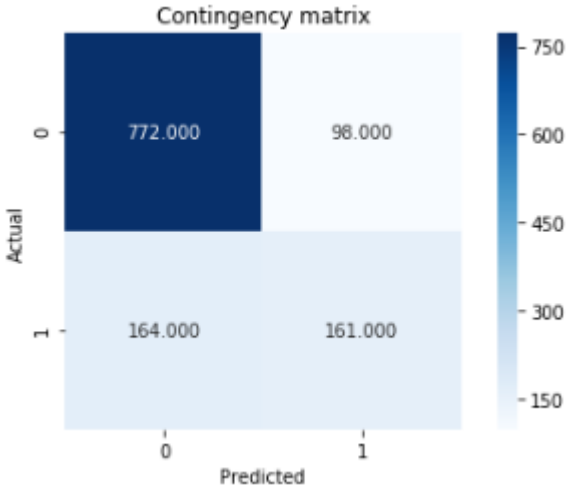
Evaluation Metrics:
Adjusted Rand Index: 0.0892
Silhouette Coefficient: 0.464

Model 5: Using variables 'Percent Age 29 and Under', 'Percent Age 65 and Older'



Evaluation Metrics:
Adjusted Rand Index: 0.016
Silhouette Coefficient: 0.462

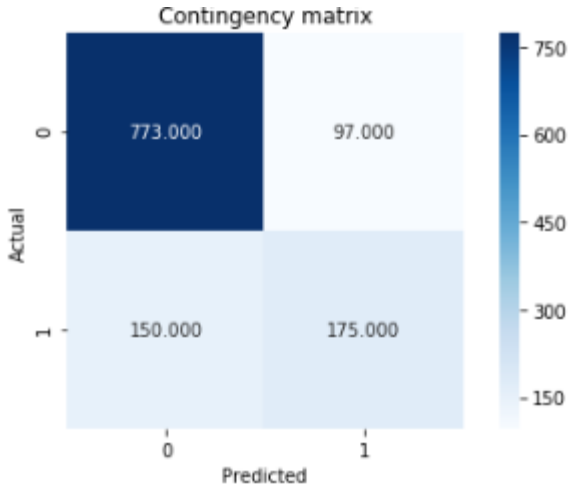
Model 6 Using variables 'Percent Black, not Hispanic or Latino', 'Percent Less than Bachelor\'s Degree'



Evaluation Metrics:
Adjusted Rand Index: 0.266
Silhouette Coefficient: 0.545

Model 7 (BEST MODEL)

Using variables 'Percent Black, not Hispanic or Latino', 'Percent Less than Bachelor\'s Degree', 'Percent Age 29 and Under', 'Percent Unemployed'



Evaluation Metrics:
Adjusted Rand Index: 0.301
Silhouette Coefficient: 0.325

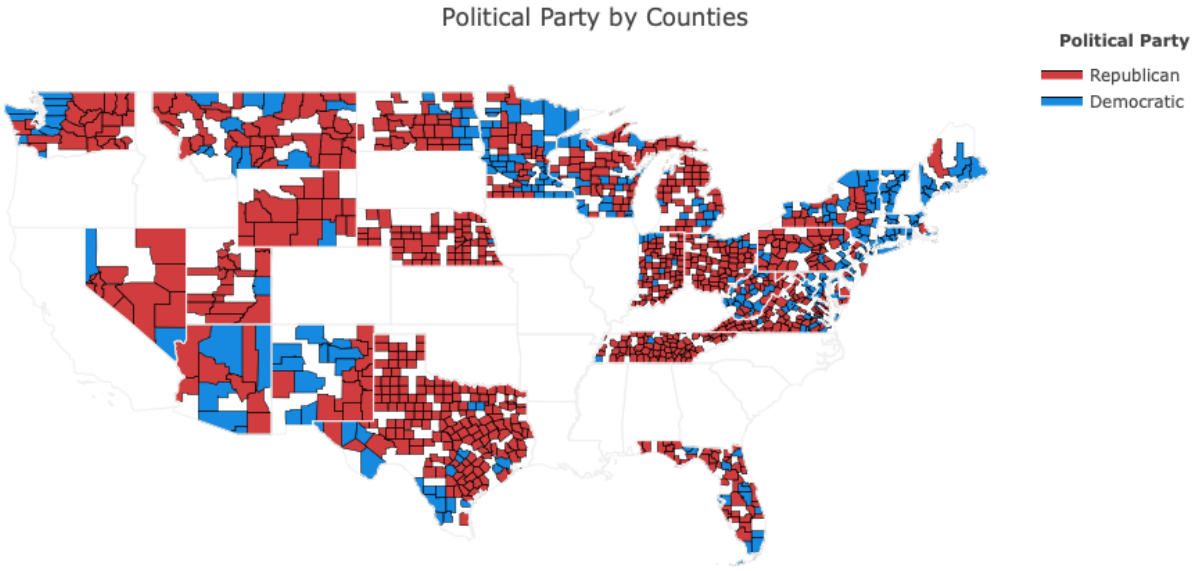
Conclusion

The model that performs the best is Model 7 of the K-mean clustering method. This model accurately predicts 948 observations of the data. The variables for this model were chosen based on conclusions from project 1. The parameters for this model are n_clusters = 2 and n_init = 10 The n_clusters parameter determines how much clusters and since we are predicting whether a county is a Democrat or Republican, we would set this parameter equal to 2. The n_init parameter is the number of iterations and 10 is a good number of iterations to run the algorithm.

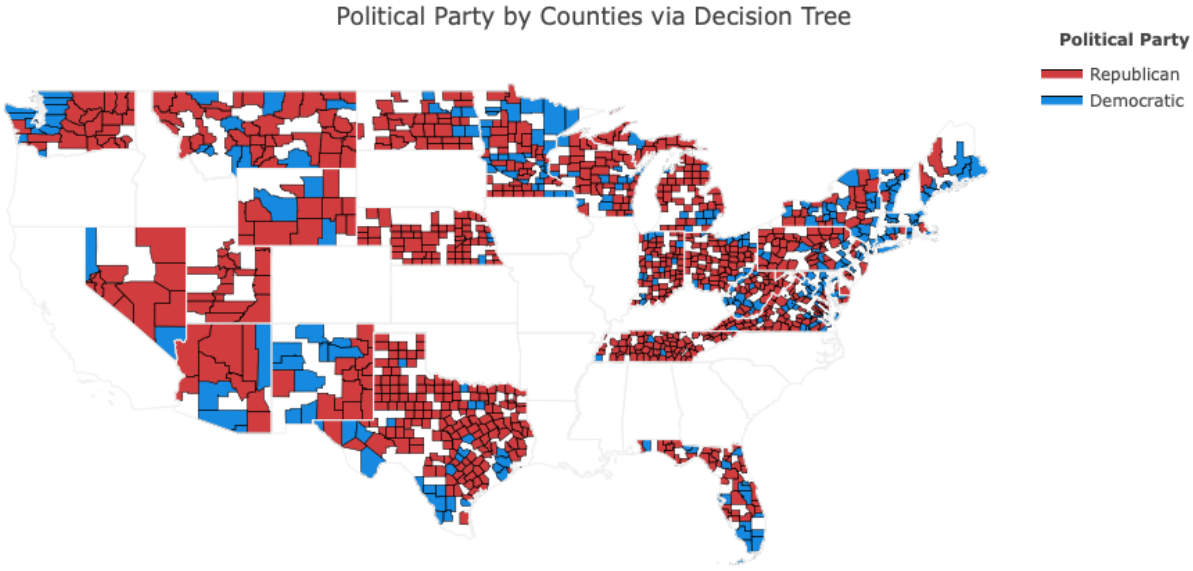
6. (10 pts.) Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Compare with the map of Democratic counties and Republican counties created in Project 01. What conclusions do you make from the plots?

Answer

Choropleth from Project 01



Choropleth from the best classifier (Model 3: Decision Tree)



Through comparing the two choropleths, we can see that while the plot generated by the decision tree correctly predicted many of the counties. However, there are still counties with incorrectly predicted parties such as Collier county in Florida which had a true party value of Republican but was classified as Democratic via the decision tree model. Likewise, Coconino county in Arizona also was incorrectly classified as being Republican despite being Democratic. This most certainly has to do with the fact that demographic data used in Model 3, ('Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Less than Bachelor's Degree'), in particular data on ethnicity and education, does not fully account for the potential political views.

7. (5 pts.) Use your best performing regression and classification models to predict the number of votes cast for the Democratic party in each county, the number of votes cast for the Republican party in each county, and the party (Democratic or Republican) of each county for the test dataset (demographics_test.csv). Save the output in a single CSV file. For the expected format of the output, see sample_output.csv.

Answer:

[621]:

	State	County	Democratic	Republican	Party
0	NV	eureka	0.000000	10279.986522	0
1	TX	zavala	0.000000	0.000000	1
2	VA	king george	21823.049764	18795.181860	0
3	OH	hamilton	183669.476767	112375.441324	1
4	TX	austin	7294.738614	6193.106586	0

This is the output file generated. To have a full look of the output file, please refer to output.xlsx. There are few negative values in the Democratic and Republican and hence these are replaced with zero values.