CS 418: Introduction to Data Science
Project 02: Regression, Classification, and Clustering
Fall 2020

Instructions

This assignment is due Wednesday, November 18, at 11:59PM (Central Time).

For this assignment, you must work in teams of three students. Each member of the team must be assigned one of three roles (*project manager*, *scribe*, or *timekeeper*) and everyone must switch roles in every project.

Deliverables for this assignment (see *Deliverables* section below) must be submitted on *Blackboard* by the *project manager*. Only one submission per team is required. Additionally, every member of the team must submit a self- and peer-evaluation form.

Late submissions will be accepted within 0-12 hours after the deadline with a 5-point penalty and within 12-24 hours after the deadline with a 20-point penalty. No late submissions will be accepted more than 24 hours after the deadline.

Offering or receiving any kind of unauthorized or unacknowledged assistance in this assignment is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Project Description

Given the merged dataset created in Project 01 (*merged_train.csv*), perform the following tasks:

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?***
2. **(5 pts.) Standardize the training set and the validation set.**
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.**
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?

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?*
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?*
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*.

Hint: Use your conclusions from Project 01 as a starting point to select the variables of the models.

Deliverables

Submit a compressed (zipped) folder on Blackboard containing the following files:

- README text file with the name, NetID, and UIN of the members of the team, as well as the role (*project manager*, *scribe*, or *timekeeper*) and contribution of each member to the assignment. Also include all necessary instructions to run your code.
- Minutes of all team meetings (saved as PDF files). Every team is required to hold at least two online meetings: one at the beginning of the project to discuss the tasks and one at the end of the project to share and discuss the results.
- Jupyter notebook (saved as both a PDF file and ipynb file) with your code and output for all the tasks in the project description.
- Output (saved as a single CSV file) of your best performing regression and classification models on the test dataset (*demographics_test.csv*). For the expected format of the output, see *sample_output.csv*.
- Report (3-5 pages, saved as a PDF file) with your answers to all the questions in the project description. Also include all corresponding results and plots. You cannot submit the PDF of your Jupyter notebook as your report.