CS 418: Introduction to Data Science
Project 02: Regression, Classification, and Clustering
Fall 2018

Instructions

This assignment is due Wednesday, November 28, at 11:59PM (Central Time).

For this assignment, you must work in groups of 2-3 students.

Deliverables for this assignment (see *Deliverables* section below) must be submitted on *Blackboard*. Only 1 submission per group is required.

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 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, perform the following tasks:

1. **Partition the 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. **Standardize the training set and the validation set.**
3. **Build a simple linear regression model (one predictor variable) to predict the number of votes cast for candidates from the Democratic party in each county. Consider multiple predictor variables. Compute evaluation metrics and report your results. *What is the best performing simple linear regression model? What is the performance of the model? How did you select the variable of the model?***
4. **Build a multiple linear regression model (more than one predictor variable) to predict the number of votes cast for candidates from the Democratic party in each county. Consider multiple combinations of predictor variables. Compute evaluation metrics and report your results. *What is the best performing multiple linear regression model? What is the performance of the model? How did you select the variables of the model?***
5. **Repeat task 3 for the number of votes cast for candidates from the Republican party in each county.**
6. **Repeat task 4 for the number of votes cast for candidates from the Republican party in each county.**

7. **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 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?***
8. **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 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?***
9. **(*BONUS*) Create a map of the counties in each cluster using Python's Plotly library (plot.ly/python/county-choropleth/). Compare with the map of Democratic counties and Republican counties created in Project 01.**

***Hint*: You can 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 group, as well as the contribution of each member to the assignment. Also include all necessary instructions to run your code.**
- ~~**CSV file with merged dataset from Project 01 used for training.**~~
- **Jupyter notebook (saved as a ipynb file) with your code for all the tasks in the project description.**
- ~~**Jupyter notebook (saved as a ipynb file) with your code for training and testing your best performing model for each task. Use *project_02.ipynb* (posted on Blackboard) as a template.**~~
- **Output of each task (saved as a single CSV file) obtained using your best performing model 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 a PDF of your Jupyter notebook as your report.**