Project 2
Regression for Prediction Problems
Sociology 273L: Computational Social Science

1 Introduction

In this project, you will learn how to develop machine learning models to predict diabetes rates in U.S. counties. Using outcome data from the Centers for Disease Control and Prevention (CDC) and the U.S. Census Bureau, you will explore the dataset, build several models, evaluate their performance, and choose the "best" model. Throughout the project, you will make extensive use of scikit-learn, pandas, and plotting packages (for example, matplotlib or seaborn). Be sure to get started early and ask questions frequently.

All work should be done in a juypter notebook and pushed to GitHub as a subfolder ("Project 2") in the "Computational Social Science Projects" repo you created for the first assignment. Be sure the final submission is in the main branch and submit a link to your repo on bCourse. Only the jupyter notebook will be graded, so be sure all of your answers are in the notebook–toggle between markdown and code cells. Be sure everything runs properly, since I will grade you assignment by pulling down your assignment from the main GitHub branch and running the notebook again.

Imagine that a policymaker tasked with piloting a new diabetes prevention program is consulting you. They want to know which counties to prioritize for the pilot program, and task you with building a model that will predict the counties with the highest rates of diagnosed diabetes. Concretely, you should train models that predict the "Diabetes_Number" column. Throughout the project, try to structure your notebook around communicating your analysis and results to the policymaker. Interpret and explain the code you write as you go.

The data have already been significantly cleaned and preprocessed to be used for analysis. If you are interested in looking at the raw data files and the code used to merge and clean them, see the GitHub repo here. Still, you will have to perform some checks on the data, recode variables, and drop duplicate columns.

2 Exploratory Data Analysis

In this section, you should explore the data to uncover interesting findings that can help guide your analysis. Make at least two figures (feel free to make more) and explain their relevance to the scientific problem. The goal here is to uncover interesting patterns in the data, learn more about the scope of the problem, and communicate these findings to your audience in clear ways. For instance, you might consider looking at how different attributes (e.g., race, sex, age etc.) vary in the data set as a whole, across geographies, or in relation to each other. You might use plotting techniques such as maps, histograms, box-and-whisker plots, or scatterplots, among others. Be as creative and thorough as possible with these explorations, and write your explanations as if you were communicating these findings to a policymaker or academic with some exposure to and interest in the question.

3 Prepare to Fit Models

3.1 Finalize Data

Remove any features that should not be used in the analysis (e.g, county name), transform categorical features so they can be used in a machine learning pipeline, and conduct any other steps necessary to prepare the data for fitting models.

3.2 Partition Data, Feature Selection, and Standardization

Investigate whether there are any features that you should remove prior to model fitting. You may also consider using plots and relationships you found in the EDA stage for this question. Be sure to justify your logic. Then, partition the data into training, validation, and test sets. Explain your choice of how much data to include in each set, and the tradeoffs involved with differing sizes in each set. Also, describe the purpose of each set. If you are using LASSO or Ridge, don't forget to standardize some of your features. Be sure to explain where in the workflow we should do this and why.

4 Train Models

Note that if you use Lasso, you will likely need to specify a very low penalty (e.g., 0.001) because of convergence problems.

4.1 Model Description

Do the following:

- Choose 3 different machine learning techniques from any of the 5 we have covered so far: KNN, OLS, Ridge, LASSO, and Elastic Net) Here's a link to the documentation: scikit-learn documentation.
- Detail the basic logic and assumptions underlying each model, its pros/cons, and why it is a plausible choice for this problem.

4.2 Train Models

Train each model in the training set, and be sure to tune hyperparameters if appropriate. Report any relevant summary statistics from the training set, including how well each model fits the training data.

5 Validate and Refine Models

5.1 Predict on the Validation Set

Using each of the models you trained, predict outcomes in the validation set. Evaluate how well each model did.

5.2 Test Set

Choose your best performing model, select out unimportant feature, retrain the model, and then predict on the test set. Evaluate your performance on this test set. What is the advantage of using both validation and test sets in the social sciences and public policy?

5.3 Implement a Cross-Validation Approach

Do the following:

- Using your preferred model of the three, use a k-fold cross-validation approach to refit the model.
- Describe the tradeoffs involved with the choice of k.
- Evaluate the results. How did cross-validation do compared to the train/validation/test split?

6 Discussion Questions

6.1 What is bias-variance tradeoff? Why is it relevant to machine learning problems like this one?

6.2 Define overfitting, and why it matters for machine learning. How can we address it?

6.3 Discuss your analysis in 2-3 short paragraphs

Discuss your findings and recommendations. For example, which counties or regions would you prioritize for the pilot program? Would your answers change based on whether you want to take into account certain features such as race, gender, or age composition in the county? How confident would you be deploying this sort of model in a real-world application – why or why not?