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.

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. Conretely, you should train models that predict the "Diabetes_Number" column Throughout the project, structure your notebook around communicating your analysis and results to the policymaker.

2 Exploratory Data Analysis

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. In this section, you should explore the data to uncover interesting findings that can help guide your analysis.

Plot three different graphs, and explain their relevance to 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 (race, sex, age etc.) vary in the dataset as a whole, across geographies, or in relation to each other. You might use plotting techniques such as maps, histograms, box-and-whisker plots 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 Clean Data

Remove any features that should not be used in the analysis (for instance, 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

Partition the data into train, 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.

3.3 Feature Selection

Investigate whether there are any features that you should remove prior to model fitting. For example, you might investigate whether there are highly correlated features or features with low variance. You may also consider using plots and relationships you found in the EDA stage for this question.

4 Train Models

4.1 Model Description

Do the following:

- Choose 5 different machine learning techniques. See available ones in the 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 Feature Selection

Conduct feature selection using techniques specific to these models. For instance, you might use coefficient cutoffs or variable importance plots. If you used a model that does automatic feature selection, detail those results as well.

5.3 Test Set

Choose your best performing model, select out unimportant features, 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?

5.4 Implement a Cross-Validation Approach

Do the following:

- Using your preferred model, 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 Paragraphs

Discuss your findings and recommendations. 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 the 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?