MATH 4780 PROJECT PROPOSAL : LIFE EXPECTANCY

11/30/23

Nicky, Ethan and Andrew

In this project, we plan to use the “Life Expectancy (WHO) fixed” dataset found on Kaggle. Within this dataset, we are given 21 variables with 179 countries that already appear to potentially have underlying correlations and multicollinearity which we could have a discussion and exploration opportunities for the presentation. This dataset is guaranteed to spark discussion and get the audience’s attention as it is interesting and provocative. We plan to explain the data to the class followed by any cleaning and management that was necessary.

Following this, we will illustrate the exploratory analysis we performed to get comfortable with the data while also displaying possible correlations to further analyze. Depending on our findings in the exploration, we will continue by deciding on the best strategy and model in order to demonstrate and achieve our desired outcome.

We will try to create a reasonable multivariate model that deals with any issues in the data may have, testing along the way different variations of feature selections and also any transformation we see fit. Along with this given that we have 21 variables, perhaps reducing the amount of regressors that have little to no prediction/explanatory power on our response variable “life expectancy”.

Here is our link to our dataset:
https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

USED FOR OUR REFERENCE:

PROPOSAL DUE 12/1 11:59PM

- JIST OF PROJECT:
    - PRESENTATION 14-16 MIN LONG
    - FOLLOWED BY 1-3 MIN Q/A

- Slides, Code, Data
    - CODE NEEDS TO BE ABLE TO REPRODUCE THE OUTPUTS WE DISPLAY ON SLIDES
- CONTENT
    - DATA VISUALIZATION
        - PLOTS, EXPLORING RELATIONSHIPS BETWEEN VARIABLES. TO CHOOSE THE REGRESSION MODEL, HOW DID WE CHOOSE THE REGRESSORS
    - WHY DID WE CHOOSE THIS MODEL FOR ANALYZING DATASET
    - INCLUDE REGRESSION DIAGNOSTICS:
        - MODEL ADEQUACY, RESIDUAL DIAGNOSTICS, LEVERAGE AND INFLUENCE DIAGNOSTICS, AND COLLINEARITY DIAGNOSTICS
            - EXPLAIN VIOLATIONS OF ASSUMPTIONS AND COLLINEARITY IF THEY EXIST (HOW DID WE FIX THEM)
    - DEMONSTRATE THE CORRECTNESS OF THE MODEL
        - INFERENCE (ESTIMATION AND TESTING) OR PREDICTION METHODS.
            - IF ISSUES REMAIN
                - EXPLAIN WHY THEY CANNOT BE FIXED
                - IMPROVING ANALYSIS
- CAVEAT:
    - WE CANNOT USED ANY DATASET THAT WE HAVE SEEN IN CLASS

- WHAT WE NEED TO DO:
    - FIND A INTERESTING DATASET:
        - DO THE ABOVE ON THE DATASET

- Dataset Sources:

Life Expectancy(fixed):

https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

Aircraft Wildlife Strikes:

https://www.kaggle.com/datasets/dianaddx/aircraft-wildlife-strikes-1990-2023

TidyTuesday

DATASETS FROM TIDY:

GPT detectors in R:

https://github.com/simonpcouch/detectors/

Kaggle

DATASETS FROM KAGGLE:

Awesome Public Datasets

DATASETS FROM AWESOME PUBLIC DATASETS:

Harvard Dataverse

DATASETS FROM HARVARD DATAVERSE:

UCI Machine Learning Repository

DATASETS FROM UCI ML REPOSITORY:

CREDIT APPROVAL DATASET:

https://archive.ics.uci.edu/dataset/27/credit+approval

DEFAULT OF CREDIT CARD CLIENTS:

https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

FiveThirtyEight

DATASETS FROm THIRTY EIGHT :