Regression Assignment

Scenario Information

The Centers for Disease Control and Prevention (CDC) along with the Nation Center for Health Statistics (NICHS) make available a lot of US-centric health data including details of births and natality. Past studies conducted[1] have concluded that factors such as a mother's BMI and weight gain during pregnancy as well as ethnicity and race are correlated with infant weight at birth. Using recently released 2021 birth records, you are to see how well you can predict a baby's weight at birth (in grams). This predication can help in predicting the need for additional postnatal care for a baby with a birth weight outside of the 2500-4000 gram range.

Data Sources

Data available for use in the analysis is from 2021 US birth records for 1st time pregnancy births (approximately 1.2 of the 3.6 million births are first time). The file name is **2021NatalityFirstPregnancy.zip** and available for download at the course GitHub site **https://raw.githubusercontent.com/SueMcMetzger/MachineLearning/main/chpt4/**
A second file – **2021NatalityFirstPregnancyPredictions.csv** – is also provided. It includes 5 sample records you will use to predict birth weight.

Data fields provided are as follows:

- Baby Female°
- Baby Non-cephalic at Birth
- Birth Place Hospital°
- Birth Weight Grams (what you will predict)
- Birth Weight Recoded – (NOTE: this is highly correlated with Birth Weight Grams)
- Cigarettes°
- C-Section°
- Day of Week*
- Father Age Bracket
- Father Education
- Father Race Recode 6*
- Gestation Recode 10
- Married°
- Month
- Month Prenatal Care Began Recoded
- Mother Age
- Mother BMI
- Mother Born in US°
- Mother Education
- Mother Height
- Mother Race*

[1] https://www.cdc.gov/nchs/data/nvsr/nvsr70/nvsr70-16.pdf

- No Known Infections
- No Known Risk Factors
- No Prenatal Visits
- No Prenatal Visits Recoded (highly corelated to No Prenatal Visits)
- Plurality (number of babies)
- Time of Birth
- US Resident°
- Weight Gain Recoded

While most fields are numeric, those fields marked with * are categorical. Those marked with ° are binary. Missing data indicates the value is unknown or not collected.

Instructions

Complete the following steps. Note that all analysis is to be performed using Python.

1. Source the data provided. Because the data file is large, it is recommended you first filter the data and only use data from the month of September. Otherwise, your processing will take too long!

2. The data is relatively clean so very little cleansing is necessary, but you will want to consider highly correlated fields, what to do with non-ordinal fields, how (if any) you may wish to stratify the data, and what do with missing data values.

3. Explore the data both visually and statistically. Identify any concerns you have using this data to proceed with your analysis.

4. Prepare your data for analysis.

5. Determine a baseline by which you can evaluate your models against.

6. Create a model using the best performing linear regression algorithm. Create a model using the best performing non-linear regression algorithm. For each model, see that
    a. Cross validation is used
    b. The model is tuned using GridsearchCV or RandomizedSearchCV
    c. Each model predicts the birth weight of 5 different baby details provided in the file **2021NatalityFirstPregnancyPredictions.csv**

Submission

Complete the presentation file called AssignRegression.ppt. This format structure will require that your analysis be concise and direct (bullet points rather than paragraphs). Save this presentation as PDF to ensure it preserves all formatting and visuals accordingly. Submit to Blackboard your presentation, your Jupyter notebook file (.ipynb), as well as a PDF of your Jupyter notebook (File -> Print as PDF). See that the Jupyter notebook includes the most current output for each cell (this aids the instructor).

You will be evaluated on your ability to analyze data, implement & tune machine learning models using Python, and assessment of the model results.