Assignment 7

Applied Machine Learning

Continuing to the previous machine learning problem, let's get back to the pre-processed dataset `Suicide Rates Overview 1985 to 2016` file. We would like to have a machine learning model to predict the suicide rate '`suicides/100k pop`'.

1. [20 pts] Use your previous pre-processed dataset, keep the variables as one-hot encoded and develop a multiple linear regression model. Use your model to predict the target variable for the people with age 20, male, and generation X. What is the MAE error of this prediction? How many regression coefficients are there?

2. [30 pts] Now use the original sex, age and generation variables in numerical form and develop a new model. Use your model to predict the target value for the people with age 20, male, and generation X. What is the MAE error of this prediction? How many line coefficients are there? (Note that for this step you have to think of a way of encoding the original nominal age feature and generation feature into numerical features.)

3. [10 pts] Did you note any change in these two model performances?

4. [10 pts] What is the prediction for age 33, male and generation Alpha (i.e. the generation after generation Z)?

5. [10 pts] Give one advantage when using regression (as opposed to classification with nominal features) in terms of independent variables.

6. [10 pts] Give one advantage when using regular numerical values rather than one-hot encoding for regression.

7. [10 pts] Now that you developed both a classifier (previously) and a regression model for the problem in this assignment, which method do you suggest to your machine learning model customer? Classifier or regression? Why?