Assignment 7

Applied Machine Learning

Continuing with 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. [10 pts] Use your previous pre-processed dataset, keep the variables as one-hot encoded, and develop a multiple linear regression model. How many regression coefficients does this model have?

2. [10 pts] Use this model to predict the target variable for people with age 20, male, and generation X. Report this prediction. What is the MAE error of this prediction?

3. [20 pts] Now go back to the original sex, age, and generation variables in their original numerical form (i.e. prior to the one-hot encoding) and build a new model. I.e., feature engineer the original nominal age and generation features into truly numerical features.) How many line coefficients are there?

4. [10 pts] Use this new Q3. model to predict the target value for the people with age 20, male, and generation X. Report the prediction. What is the MAE error of this prediction?

5. [10 pts] Did you note any change in these two model performances?

6. [10 pts] Use your Q3. model to predict the target value for age 33, male, and generation Alpha (i.e. the generation after generation Z); report the prediction.

7. [10 pts] Give one advantage when using regression (as opposed to classification with nominal features) in terms of independent variables.

8. [10 pts] Give one advantage when using regular numerical values rather than one-hot encoding for regression.

9. [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?