Chris Donaton, Data Science specialization

Problem description

Given an array of features that describe a patient's medical history, calculate that patient persistency likelihood.

Data understanding

The data set is well constructed with 69 fields (2 numeric and 67 categorical). There is no missing data, but there are some fields with some form of 'other' indicator, such as race, religion, and medical specialty to name a few. Currently, I see no benefit to treat these as missing values.

What type of data you have got for analysis

The categorical data is mainly comprised of indicator variables (binary) with a few fields that group and segment the population into ranges, such as age range.

What are the problems in the data (number of NA values, outliers, skewed etc.)

There are some balance issues that will drive modeling decisions. Males and persistently encoded patients, those that followed the treatment regimen completely, are overrepresented by a ratio of 16:1 and 2:1 respectively.

Additionally, both numeric fields, one which counts the number of Dexa scans received during treatment and the other which counts the number of risk factors, are right skewed.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier, etc. and why?

It is never an ideal approach to remove data. For the fields with 'other' indicators, I will assess the predictive factor on selected models and consider removing the fields altogether.

For the skewed numeric columns, I will apply transformations to counteract the skewness.

Github Repo: https://github.com/rambles-tech/DG_virtual_internship/week07