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 cleansing and transformation done on the data.

There were several groups of missing data in the data set. I decided to drop all rows with NaNs if fields that totaled less than 5% of the overall data.

For some of the missing data, I was able to drop the entire column due to high correlation to other features, such as ntm_specialist_flag and ntm_specialist_bucket. Since we already have an ntm_speciality field, we do not need these other aggregated features.

For the missing ntm_speciality features, I imputed data to match the proportion of specialties in the original dataset to avoid over weighting the mode, general practitioner.

Finally, there is roughly a third of the data set that is missing values for the same rows in the following features: change_t_score, risk_segment_during_rx, and tscore_bucket_during_rx. This represents too much of the data set to execute an arbitrary imputation. I conducted a two sample T-test to confirm that NaNs for these fields are not equally distributed in both persistent and not persistent groups.

I will continue to look for fields that are highly correlated with these fields. If I cannot find an adequate replacement, I will have to drop these rows.

I was able to use log transformations to reduce the effect of outliers in the two numeric fields. Square root and min-max method did not produce adequate results.

Github Repo: https://github.com/rambles-tech/DG_virtual_internship/week09