Group name: Lone Wolf

Name: Nathan Adam

Email: nathan.adam@myport.ac.uk

Country: England

College/Company: Appen

Specialization: Data Science

GitHub repository link: **https://github.com/N-A-ML/Data_Glacier_Final_Project_Week_7_to_13**

Problem description:

To understand the healthcare dataset (persistency of a drug) and devise a plan regarding how we will clean, transform and handle the data.

Data understanding and type of data for analysis:

To start with, the data available is structured, with 3424 rows and 69 columns. Persistency_Flag is the target variable, and indicates whether or not the patient is persistent with treatment. There are only 2 numerical variables: Dexa_Freq_During_Rx and Count_Of_Risks, the others are categorical or have values in the form of strings. The variables Age_Bucket and Ptid will be converted to numbers the 'replace' command. The categorical variables are to be converted to binary variables with pd.get_dummies so that they can be used for modelling later on. The numerical variables are not normally distributed.

Problems in the data:

1)The numerical variables are not scaled

2) There are outliers present in these variables: Dexa_Freq_During_Rx , Count_Of_Risks. In principle this isn't generally problematic, but we found that the data seemed to be of a higher quality when outliers were removed.

3) There are a very high number of columns / features

Some of the classes for the categorical variables are imbalanced (most notably the Gender variable), although this is not necessarily problematic. We see that some of the classes are imbalanced / skewed. For example there are only 194 males and 3230 females in the dataset. The vast majority of the patients are caucasian (3148 compared to 97, 95, 84), but that might be expected for the country the data was collected in. The regions are reasonably balanced. We see that the age ranges are mostly for older patients (~ 3300 55+ patients and only 166 patients aged <55. With regards to the specialist type, some specialisms are almost non-existent (e.g., radiology, podiatry, opthamology and some others only have 1 entry in the dataset). For most variables, there is a realistic-looking distribution for the classes. Some classes may have poor representation due to the circumstances / environment under which the data was collected, which is not a cause for concern.

There are no NA, NaN, or missing values in the dataset, and no duplicate values.

Approaches to overcome problems:

1) We can use MinMaxScaler to scale the numerical variables. This will be useful for future analyses, especially since the data is not normally distributed
2) It could be useful to remove outliers (values which exceed 3 standard deviations of the mean) to reduce the variance in the data.
3) We can use SelectKBest to choose the k best features, and just use those for future analysis.