Group name: Cheetah

Name: Obiri Vincent

Email: vince23.ac@gmail.com

Country: Kenya

College/Company: KCA University

Specialization: Data Science

GitHub repository link:
https://github.com/Vince689/Data_Glacier_Final_Project_Week_7_to_13.git

Problem description:

Use Python to perform EDA on the datasets to provide some insights for the pharmaceutical company, and to support future modelling work.

Exploratory data analysis performed:

Seaborn was used for the plotting.

We explored the average values for each numerical variable (Count_Of_Risks and Dexa_Freq_During_Rx) amongst the Persistency flag classes. Bar charts were created to visualise this. On average, the numerical variables had higher values for the persistent class.

We created a scatterplot of Count_Of_Risks against Dexa_Freq_During_Rx. It was clear that as Count_Of_Risks increase, Dexa_Freq_During_Rx tends to decrease.

A new data frame was made, which was created by restructuring the data. We made a plot of the counts of the true labels for each of the best 6 features (found with SelectKBest in weeks 8-9, grouped by persistency flag. We created similar bar plots for other variables grouped by persistency flag, such as race, region, ethnicity, Hispanic/non-hispanic, age groups.

Outliers in original dataset, grouped by persistency flag, were investigated with box plots. There was a low number of outliers for Count_Of_Risks, and a lot of outliers for Dexa_Freq_During_Rx. Whether or not a patient had a dexa scan was more predictive of their persistency, rather than the number of dexa scans.

A correlation heatmap was created to check for multicollinearity and see the correlations between the features and the target. There wasn't

multicollinearity and the highest correlation was between Dexa_During_Rx_Y and Persistency_Flag.