Final Project (Individual)

Team Primero

Date: 25th July 2021

LISUM01

Name: Reeha Khan
Email: khanreeha22@gmail.com
Country: Pakistan
College: National University of Sciences and Technology
Specialization: Data Science

Problem description

ABC is a pharmaceutical company that wants to understand the persistency of a drug as per the physician's prescription for a patient. The company has approached an Analytics company to automate this process of identification. This Analytics company has assigned the task to Team Primero to come up with a solution to automate the persistency of a drug for the client ABC.

GitHub Repo Link

https://github.com/ReehaKhan/Project-Healthcare.git

Data Cleaning And Transformation

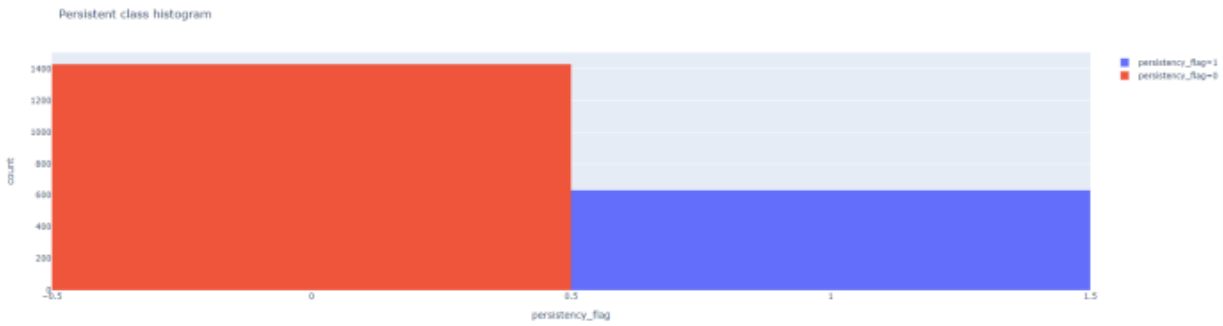
Since there are no Null values, so there is nothing to do in that regard. They are some skewness and Kurtosis in two numerical features, so we scaled their values by RobustScaler() and after that remove their outliers by calculating IQR and remove data smaller/greater than two whiskers. After removing outliers from “dexa_freq_during_rx”, we can now check the decrease in the shape of the data:

Old Shape: (3424, 69)

New Shape: (2964, 69)

All the [‘Y’, ‘N’] values have been changed to [1, 0] to train models on the data, and also we have changed the values of target feature from ['Non-Persistent', 'Persistent'] to [0, 1].

The other problem in the data was the misbalancing of the target feature:



Since misbalanced datasets makes predicting hard and interfere with the models, we can do "Up sampling" on the data. In this method, we increase the records of the minority class such that at the end count of records is same for each class.

Another thing we performed on the dataset is “one hot encoding”. We need numerical values to use classifiers. This is done by using the “get_dummies()” function from Pandas library.

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0