Some information about the data

Name	Email	Country	College/Company	Specialisation
Kelvin Mpofu	mpofukelvintafadzwa@gmail.com	South Africa	n/a	Data science
Purity Nyagweth	purityeverniter@gmail.com	Kenya	n/a	Data Science
Reshma Jayapalan	reshma.jayapalan@gmail.com	UAE	n/a	Data Science
Hanouf Hazza	hanouf.haz@gmail.com	Saudi Arabia	n/a	Data Science

Problem Description

ABC pharma company has a challenge in understanding drug persistency as per physician prescription and to solve this problem it wants to automate the process of identification.

Objective is to build a classification model for drug persistency identification.

This will automate the process of identifying drug persistency for ABC pharma company thus helping the company to understand drug persistency as per physician prescription.

Data Understanding

What type of data is given the analysis - .xlsx type of data

Data shape – 3424 by 69

Variable types – categorical variables are 67 and the numerical variables are 2

The data set has many unknown values specifically in the features 'Race', 'Ethnicity','Region','Ntm_Speciality','Risk_Segment_During_Rx', 'Tscore_Bucket_During_Rx','Change_T_Score','Change_Risk_Segment'. We consider the missing values to be features which can be used by the machine learning algorithm so we did not remove them. There are outliers in the feature sets Dexa_Freq_During_Rx. The count of risks column is positively skewed.

What problems are with the data:

- **Number of NA values –** 0
- **Number of duplicates -** 0
- **Outliers –** There are outliers in the numerical columns; Dex_Freq_During_Rx and Count_Of_Risk
- **Skewed data –** The numerical columns have skewed data.
- **Class imbalance in the target variable –** There's a slight imbalance

Solutions to the problems

1. **Outliers**
    - Dropping the outliers by using the Inter-Quartile-Range method
2. **Skewed data**
    - Dropping the outliers (worked perfectly).
    - Log transformation of the numerical features
    - Applying the min-max to the numerical features

3. **Class Imbalance**
    - Oversample the minority class
    - Downsample the majority class

Vocabularies

1. **Drug persistency –** The act of continuing the treatment for the prescribed duration.
2. **Adherence –** The extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen.
3. **IDN -** a network of healthcare providers and facilities within a specific geographic region that offers a full range of healthcare services. An IDN is often designed to offer a full spectrum of care inclusive of primary care physicians, specialists, general acute care (i.e. inpatient services), and home health services.
4. **NTM –** Non-Tuberculosis Mycobacteria
5. **Rx –** Medical prescription
6. **DEXA Scan –** Also known as bone density scan. Used to measure calcium and other minerals in the bone.
7. **Fragility fracture –** a fracture resulting from a fall.
8. **Glucocorticoid –** Are steroid hormones used for the treatment of inflammation, autoimmune diseases and cancer.
9. **Injectable drugs –** Drugs injected to users
10. **Comorbidity –** condition of having two or more diseases at the same time.
11. **Concomitancy –** Existing or occurring together. Concomitant medication are other prescription drugs, over the counter drugs or dietary supplements that are taken other the drug under investigation.
12. **T score –** Measures how much a bone density is higher or lower than that of a healthy 30-year-old adult.