

Week 10: EDA and recommendations

Team Members:

| S/N | Name | Email | Country | College | Specialization |
| --- | --- | --- | --- | --- | --- |
| 1 | Tomisin Abimbola Adeniyi | tomisin\_adeniyi11@yahoo.com | Nigeria |  | Data Science |
| 2 | Fabio Pontecchiani | pontecchianifabio@gmail.com | Belgium | University of Sheffield | Data Science |
| 3 | Bilikis Omolara Alayo | berlykis@gmail.com | United Kingdom |  | Data Science |



1. Missing Values

The missing values present in the data were addressed by using the imputation methods. Mean imputation and model-based imputation were applied separately on the dataset. Since all the missing values were of categorical types, model imputation could be effective here because the method is simple and robust as it replaces missing values with the most frequent category in the features.

2. Skewness

From the two methods used to handle skewness, log transformation appears to have a much lower skewness when applied on the numerical features.

Log transformation skewness:

Dexa\_Freq\_During\_Rx 1.405860

Count\_Of\_Risks -0.091583

Square root transformation skewness:

Dexa\_Freq\_During\_Rx 1.992495

Count\_Of\_Risks -0.327599

Hence, log transformation method will be used to handle the skewness in the numerical features.

3. Class Imbalance

The imbalance discovered in the target variable (`Persistency\_Flag`) will be handled by using the Synthetic oversampling techniques (SMOTE) as it creates synthetic samples that are typical of the minority class, which improves the model's capacity to generalise to previously unknown data. Also, it has the ability to reduce overfitting caused by merely replicating minority class data.

4. Feature Selection

The three methods of feature selection selected different numbers of features based on the logic behind each individual method. However, some of these features were commonly selected by all the three methods.

Therefore, Race, Ethnicity and few of the 'Risk' categories will be dropped as they were not selected by these methods.

It is also recommended to exclude demographic features like 'Race', 'Ethnicity', and 'Region' from the machine learning model as they are unlikely to provide significant contributions. Similarly, considering the large proportion of missing values relative to the total number of rows (3423), it is advisable to drop 'Risk\_Segment\_During\_Rx', 'Tscore\_Bucket\_During\_Rx', 'Change\_T\_Score', and 'Change\_Risk\_Segment' to maintain the integrity and effectiveness of the model.

Few key points from EDA:

* Most of the physicians are not specialists.
* Most of the patients are mapped to IDN
* Most of the patients are female (around 15 times more females)
* Majority the patients are not Hispanic (20 times more not hispanic)
* Most of the patients are from Midwest and South regions
* Majority of the patients are Caucasians
* Majority of the patients during the therapy are above 75 years old

The correlation diagram has to be re-run as it has not been satisfactory in selecting the right features for the model.

5. Encoding

Encoding of non-numerical categorical data has been performed, to convert the risk factors into 1-0 instead of yes/no. All the risk factors are more oriented on the ‘No’, while the one with more balanced answers is the “Vitamin D insufficiency”. Which could play an important role in the Drug Persistency determination.