Project title: Healthcare - Persistency of a drug
Group name: DG_team_project_PL-RO-KSA-EGY
Github repo: https://github.com/Omar-Safwat/HealthCare_project
Week: 8

Team members

Name	Specialization	Country	Email
Ms. Larisa Popa	Data Science	Romania	Larisapopa4@gmail.com
Ms. Afshan Hashmi	Data Science	Kingdom of Saudi Arabia	afshanhashmi786@gmail.com
Mr. Omar Safwat	Data Science	Egypt	omarksafwat@gmail.com
Mr. Roger Burek-Bors	Data Science	Poland	roger.burek-bors@hotmail.com

Problem description

A machine learning model cannot be built without sufficient data. Quality data is fundamental to any data science engagement. To gain actionable insights, the appropriate data must be sourced and cleansed. There are two key stages of Data Understanding: a Data Assessment and Data Exploration. Our team provides Data Understanding insights within week 9 assignment.

The Pharmaceutical company provided dataset called “Healthcare_dataset” in xlsx format consisting of:

- Basic description of features (Feature Description tab)
- Data (Dataset tab) where we found:
  - 69 features
  - 3424 data entries

Data understanding

Key dataset characteristic are as following:

- 69 features
- 3424 data entries
- 99% features are provided as categorical data and we need to turn them into numerical before feeding into ML model (e.g. “Yes”/”No” as 1/0, NTM_Speciality as dictionary), only one feature has numerical values ("Count_Of_Risks")
- We can replace all “Yes”/”No” as 1/0 and for each other column that does not contain numerical values ('Persistency_Flag', 'Race','Ethnicity','Region','Age_Bucket','Ntm_Speciality','Ntm_Specialist_Flag','Ntm_Speciality_Bucket','Risk_Segment_Prior_Ntm', 'Tscore_Bucket_Prior_Ntm', 'Risk_Segment_During_Rx', 'Tscore_Bucket_During_Rx', 'Change_T_Score', 'Change_Risk_Segment', 'Adherent_Flag'). We can make a different column containing the corresponding numerical label obtained using LabelEncoder
- Data is labelled and “Persistency_Flag” feature will serve as the label for ML model
- Missing values are present as "Unknown" or "Other/Unknown" in following features: "Race", "Ethnicity", "Region", "NTM_Speciality", "Risk_Segment_During_Rx","Tscore_Bucket_During_Rx", "Change_T_Score", "Change_Risk_Segment".
- Outliers are present in the following features:
  - Gender: predominant females, the rest outliers
  - Race: predominant Caucasian, the rest outliers
  - Ethnicity: predonimamnt Not hispanic, the rest outliers
  - Region: cathegory other outliers
  - Ntm_specialty: predonimant categories: general_preactitioner, unknown, endocrinology, rheumatology, oncology, obstretics and gynecology; the rest outliers
  - Dexa_freq_during_rx : predominant : 0; the rest outliers
  - Adherent_flag: predominant adherent; the rest outliers
  - Risk_Type_1_Insulin_Dependent_Diabetes: predominant False(0); True(1) outlier
  - Risk_Osteogenesis_Imperfecta: predominant False; True outlier
  - Risk_Rheumatoid_Arthritisitis: predominant False; True outlier
  - Risk_Untreated_Chronic_Hyperthyroidism: predominant False; True Outlier
  - Risk_Untreated_Chronic_Hypogonadism: predominant False; True Outlier
  - Risk_Untreated_Early_Menopause: predominant False; True Outlier
  - Risk_Patient_Parent_Fractured_Their_Hip: predominant False; True Outlier
  - Risk_Chronic_Liver_Disease: predominant False; True Outlier
  - Risk_Low_Calcium_Intake :predominant False; True Outlier
  - Risk_Poor_Health_Frailty: predominant False; True Outlier
  - Risk_Excessive_Thinness: predominant False; True Outlier
  - Risk_Hysterectomy_Oophorectomy: predominant False; True Outlier
  - Risk_Estrogen_Deficiency: predominant False; True Outlier
  - Risk_Immobilization : predominant False; True Outlier
  - Risk_Recurring_Falls: predominant False; True Outlier
  - Count_of_risks: 5,6,7 outliers

Missing values (NaN or Unknown in our case)

Why NaN values are problematic?

- we cannot train data
- data is not informative

Dealing with NaN values:

- delete them
- predict them (using regression technique) or impute them (using KNN technique)
- relace them with the category that has the highest ocuurence (mode)

We may distinguish some cases:

1. NaN values occurence is higher than the other column categories. Example 3 classes, NaN values occurence is the highest and the other 2 categories occurencies have almost similar dirtibution:
   - Relacing the NaN values with the category that has the highest ocuurence (mode), we influence learning on only one category. ---> Not a good idea
   - We may delete them ----> we can lose data, so it is not recommended
   - We may predict them ---> try and compare with mode data result; if the result are similar it is recommended to used mode as training is time consuming
2. NaN values occurence is very lower than the other column categories. -> in this case it is definitely suitable to use mode

Outliers

Why the outliers are a problem?

They do not contribute to model learning. The values are irrelevat and the model will only learn the dominant cathegories. Large data slows down the training time, and for no reason.

Dealing with outliers:

1. If for example there are only two categorical values in a feature and one of the features has very few values that can be considered outliers, it would be a good idea to not consider at all those columns when training and delete columns entirely. Because dominant data will determine learning only one category, it will not influence learning, but it contributes to speeding training time. It this case, we may delete the following columns:
   'Gender','Risk_Type_1_Insulin_Dependent_Diabetes','Risk_Osteogenesis_Imperfecta','Risk_Rheumatoid_Arthritis','Risk_Untreated_Chronic_Hyperthyroidism','Risk_Untreated_Chronic_Hypogonadism','Risk

2. Delete rows that contain outliers.

Dealing with many features

69 features is a large number. And usually, not all data is relevent to training. We must find the features that matters. After deleting possible columns that contain only one category, we can still have a large number of features. One solution for a clean and relevant data, and also for a faster training would be to apply dimension reduction and see how relevant is each feature when training.

Skewed data

Not present in our dataset.