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: 9

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:

- Missing data in total: 7278 values which means 3%
- Outliers in total: in 23 of 69 features which means 33%

Data cleansing and transformation - techniques description

Missing values (NaN or Unknown in our case)

Larisa approach Predictions based on logistic regression & Outliers
Afshan approach Missing value prediction

1. Predicting NaN values using Logistic Regression. As target values are categorical the problem was a classification one and I used linear regression as algorithm.

Problems encountered:

It was a little dificult to predict each column which contained NaN values at a time and at the same time to consider all the other features because some other columns may also contain NaN values. It was not possible also to build multi-class model prediction as column that contained NaN values had always diffrent data training. It not always happened to have one row that contained multiple NaN values placed in multiple columns. The problem was solved by not considering the columns that also contained NaN values and succeed to predict one at a time.

Results:
As it was expected, the columns that had dominant target classes turned to be the most frequent predicted value. In those cases, it would have been easier just to apply mode and not model training as the results are pretty similar.

Roger approach KNN Imputation
Omar approach: KNN Imputation

KNN Imputer from sklearn library was used to deal with missing values in Ntm_Speciality feature.

Ntm_Speciality had before Imputation 36 categories, among them “Unknown”. This lead to an observation that "unknown" should be distributed among 35 other specialties and perhaps consequently KNN Imputer supposes to take 35 neighbors wile dealing with missing values. Therefore, two Imputation were made over “Unknown” from Ntm_Speciality, one with 5 neighbors, other with 35, as KNN Imputer setting.

Second issue was related to features that should be taken into common array. After deliberation all features called “Concomb*” were taken as then seem to be the most relevant to Ntm_Speciality because they represent health issues. Then these features were transformed to numerical. No was represented by 0 and yes as 1. A dictionary was created for Ntm_Speciality which helped to translated this features both ways, from categorical to numerical – for KNN Imputation, and from numerical to categorical – for post imputation needs e.g. merging new dataframe with ABC Company dataset.

It was insightful that KNN Imputer transforms all data into floats and after imputation missing data might be distributed as floats while we need them to be integers to be translated into categories. There are two ways of dealing with this: rounding up or cutting decimals. Rounding up according to arithmetical rules was applied.

Full comparison of original Ntm_Speciality and after work of KNN Imputer is presented on the linked pdf.

Apart from number of neighbors unknown were distributed only over top 9 categories, as following:

Ntm_Speciality	Original DF	After KNN 5 neigh.	After KNN 35 neigh.
GENERAL PRACTITIONER	1535	1585	1538
RHEUMATOLOGY	604	752	786

Ntm_Speciality	Original DF	After KNN 5 neigh.	After KNN 35 neigh.
ENDOCRINOLOGY	458	520	574
ONCOLOGY	225	254	234
OBSTETRICS AND GYNECOLOGY	90	101	90
UROLOGY	33	36	33
ORTHOPEDIC SURGERY	30	34	30
CARDIOLOGY	22	24	22
PATHOLOGY	16	17	16

Outliers

Larisa approach
Predictions based on logistic regression & Outliers

There have been 2 ways of dealing with outliers:

1. Deleting entirely the columns that has only 2 classes and one class has >90% frequency.
2. Deleting the rows that contained outliers.

Dealing with many features

Dataset original shape: 3424 rows, 69 fetures

Suggestion which features should be considered for modelling:

- after dealing with NaN and outliers the data ready for trainig has the following shape:
  2004 rows, 52 features:
  'Gluco_Record_Prior_Ntm', 'Gluco_Record_During_Rx', 'Dexa_Freq_During_Rx', 'Dexa_During_Rx',
  'Frag_Frac_Prior_Ntm', 'Frag_Frac_During_Rx', 'Idn_Indicator', 'Injectable_Experience_During_Rx',
  'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms',
  'Comorb_Encounter_For_Immunization',
  'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx',
  'Comorb_Vitamin_D_Deficiency', 'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',
  'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx',
  'Comorb_Long_Term_Current_Drug_Therapy', 'Comorb_Dorsalgia',
  'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',
  'Comorb_Other_Disorders_Of_Bone_Density_And_Structure',
  'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',

'Comorb_Osteoporosis_without_current_pathological_fracture', 'Comorb_Personal_history_of_malignant_neoplasm', 'Comorb_Gastro_esophageal_reflux_disease', 'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations', 'Concom_Narcotics', 'Concom_Systemic_Corticosteroids_Plain', 'Concom_Anti_Depressants_And_Mood_Stabilisers', 'Concom_Fluoroquinolones', 'Concom_Cephalosporins', 'Concom_Macrolides_And_Similar_Types', 'Concom_Broad_Spectrum_Penicillins', 'Concom_Anaesthetics_General', 'Concom_Viral_Vaccines', 'Risk_Smoking_Tobacco', 'Risk_Chronic_Malnutrition_Or_Malabsorption', 'Risk_Family_History_Of_Osteoporosis', 'Risk_Vitamin_D_Insufficiency', 'Count_Of_Risks', 'Persistency_Flag_labels', 'Race_labels', 'Ethnicity_labels', 'Region_labels', 'Age_Bucket_labels', 'Ntm_Speciality_labels', 'Ntm_Specialist_Flag_labels', 'Ntm_Speciality_Bucket_labels', 'Risk_Segment_Prior_Ntm_labels', 'Tscore_Bucket_Prior_Ntm_labels', 'Risk_Segment_During_Rx_labels', 'Tscore_Bucket_During_Rx_labels', 'Change_T_Score_labels', 'Change_Risk_Segment_labels', 'Adherent_Flag_labels'

- We also tried to analise feature relatedness in order to figure out what features would have impact when training. Corelation matrix was not very realiable in this case. As an alternative, we tried PCA dimension reduction algorithm in the hope of finding the number of the features tha lead to best results when training. The results showed that all 52 features have a big influence when training.

Dataset for EDA and modelling

As part of week 9 team worked out the data set for EDA and further modelling.

EDA dataset: cleaned_data.csv - This dataset has no NaN values and no outliers. The NaN values have been replced by the reults obtained from prediction. The outliers have been elimanted as described above. This dataset contains the values in the original format, but also in numerical labeled format.

Training dataset: data_training.csv - This dataset contains only numerical values.