Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Project Name: Healthcare – Persistency of a Drug

Group Name: Decent Healthcare Analysis

Name: Tolga Yaz

E-Mail: tolgayaz1991@gmail.com

Country: Turkey

Specialization: Data Science

29-Nov-2021

Background – Persistency of a Drug

- **One of the challenges for pharmaceutical companies is to understand the persistency of drugs as per the physician prescription. To solve this problem a pharma company wants to automate this process of identification via using analytics.**
- **Main Goal: Predicting drug persistency of patients to provide better healthcare.**
- **Main Objective: Building a classification model to predict drug persistency of patients via using provided data and data science methods.**

- **The planned project lifecycle:**

1. Problem Description, Business Understanding and Data Intake Report Preparation
2. Data Understanding and Exploratory Data Analysis
3. Eda Presentation and Proposal of Modeling Technique(s)
4. Model Selection and Model Building
5. Evaluation of Results and Final Project Report

Basics of the Examined Data and Exploratory Data Analysis Approach

- **1 data file in .xlsx format with 2 sheets were obtained:**

1. **Healthcare_dataset.xlsx**: Includes 2 sheets:

a. **Feature Description:** Includes descriptions of main features provided.

b. **Dataset :** Includes the data provided.

- **Characteristics of the data provided:**
- 3424 Observations,
- 69 Features.

- **Exploration Approach:**

1. The data was transformed into Python Pandas Dataframe to be processed.
2. The dataframe were observed carefully to identify and deal with the nulls and duplicates.
3. Analysis were done upon the dataframe file via usage of Python Libraries and Jupyter Notebooks.

Exploratory Data Analysis Phase 1 (Dealing with Missing Values)

• The data provided are mostly composed of categorical features (67 of 69 features are categorical with data type of object). And only 2 features have data type of int64.

• When the data were examined, no dominant outlier problem was seen.

• When the data were examined, it was seen that there are some values named “Unknown”.
• The “Unknown” values were changed into NaN values to easily examine them via Pandas.
• The number of NaN values and the column names they are found are as below:

1) ('Ethnicity', 91),
2) ('Ntm_Speciality', 310),
3) ('Risk_Segment_During_Rx', 1497),
4) ('Tscore_Bucket_During_Rx', 1497),
5) ('Change_T_Score', 1497),
6) ('Change_Risk_Segment', 2229)

• Since we have many columns and the data are sensitive, we dropped the columns with >200 missing values and got a new dataframe without these missing values. For the "Ethnicity" column, we used mode value to fill missing values.

Exploratory Data Analysis Phase 2 (Analysis of Relationship Between Variables)

- **Relationship between the variables was analyzed and the features which do not affect the result of the target variable were found and dropped.**

- **The 28 features that were dropped because their lack of efficiency to the target result are:**

['Gender', 'Ethnicity', 'Ntm_Specialist_Flag', 'Gluco_Record_Prior_Ntm', 'Frag_Frac_Prior_Ntm', 'Risk_Segment_Prior_Ntm', 'Tscore_Bucket_Prior_Ntm', 'Idn_Indicator', 'Injectable_Experience_During_Rx', 'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias', 'Comorb_Osteoporosis_without_current_pathological_fracture', 'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations', 'Concom_Anti_Depressants_And_Mood_Stabilisers', 'Risk_Type_1_Insulin_Dependent_Diabetes', 'Risk_Untreated_Early_Menopause', 'Risk_Patient_Parent_Fractured_Their_Hip', 'Risk_Smoking_Tobacco', 'Risk_Chronic_Malnutrition_Or_Malabsorption', 'Risk_Family_History_Of_Osteoporosis', 'Risk_Low_Calcium_Intake', 'Risk_Vitamin_D_Insufficiency', 'Risk_Poor_Health_Frailty', 'Risk_Excessive_Thinness', 'Risk_Hysterectomy_Oophorectomy', 'Risk_Estrogen_Deficiency', 'Risk_Recurring_Falls', 'Risk_Untreated_Chronic_Hyperthyroidism', 'Risk_Immobilization']

- **The shape of the data after exploratory data analysis:**

(3202, 34) with a target variable (Persistency_Flag) and 33 predictors

A Screenshot from Correlation Analysis

- **Below is a screenshot from correlation analysis (the values are pearson correlations):**

	Persistency_Flag
Persistency_Flag	1.000000
Dexa_Freq_During_Rx	0.330503
Count_Of_Risks	0.081699
Gluco_Record_During_Rx_code	0.213078
Dexa_During_Rx_code	0.495773
Frag_Frac_During_Rx_code	0.113836
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms_code	0.317499
Comorb_Encounter_For_Immunization_code	0.318606
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx_code	0.287595
Comorb_Vitamin_D_Deficiency_code	0.178660
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified_code	0.233583
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx_code	0.210992
Comorb_Long_Term_Current_Drug_Therapy_code	0.358023

Results and Model Recommendation

- **The shape of the data before exploratory data analysis: (3424, 69)**
- **The shape of the data before exploratory data analysis: (3202, 34)**

- **Some of the data were dropped because of mainly two reasons:**

1. Missed Data

2. Inefficient-to-the-Target-Result Data

- **Since the data is tabular and the target value with most of the other features are categorical, the recommended ml model is "Random Forest".**

Thanks