Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis and Modelling Proposal

Project: Healthcare -Persistency of a Drug

Group Name: Data Explorer
Name: Mohammad Tohin Bapari
Email: tohin@gmx.de
Country: Germany
University: Bergische Universität Wuppertal
Specialization: Data Science

Agenda

- Problem Statement
- Approach
- EDA
- EDA Summary
- Model Proposal

Data Glacier
Your Deep Learning Partner

Business problem

ABC Pharma contacted us to carry out an analysis in order to have a deeper understanding on the factors impacting the **persistence** of their drug. The aim is to know if a patient, based on his/her information, will follow the prescription of the physician and continue taking the drug for all the treatment time.

Approach

- 1 file was provided: Healthcare_dataset.xlsx
- The file contained information of 3, 424 patients. For each patient it has demographic information, clinical records, others diseases as risk factor information and also about their physicians specialty.
- The variables provided have been treated individually among the four members of the team.
- The **EDA** has been carried out following the same arrangement, but taking into account the whole dataset, so that potential insights have been drawn from the analysis.

Data Glacier
Your Deep Learning Partner

Key Columns

- **Ptid:** Patient ID
- **Persistency_Flag:** Indicates if the patient is persistent or non-persistent
- **Gender:** Gender of the patient
- **Ethnicity:** Ethnicity of the patient
- **Region:** Geographical region
- **Age_Bucket:** Age group
- **Ntm_Speciality:** Specialty of the treating physician
- **Ntm_Specialist_Flag:** Indicates if the treating physician is a specialist
- **Ntm_Speciality_Bucket:** Category of specialty

EDA

Numeric Column: `Count_Of_Risks`

Statistic	Value
Count	16067
Mean	0.99
Standard Deviation (std)	0.78
Minimum (min)	0
25th Percentile (25%)	0.00
50th Percentile (50%)	1.00
75th Percentile (75%)	1.00
Maximum (max)	6

`Persistency_Flag`

Value	Count
Non-Persistent	8034
Persistent	8033

`Gender`

Value	Count
Female	12235
Male	3832

Data Glacier
Your Deep Learning Partner

EDA

Region-wise Patient Distribution

Region	Count
South	5,217
Midwest	4,464
West	3,704
Northeast	2,682

Age Distribution of Patients

Age Bucket	Count
>75	7,084
65-75	4,615
55-65	2,961
<55	1,407

Region-wise Count of Specialists

Region	Count
South	265
Midwest	230
West	186
Northeast	134

EDA

Key Risk Factors:

- Risk_Family_History_Of_Osteoporosis
- Risk_Low_Calcium_Intake
- Risk_Vitamin_D_Insufficiency
- Risk_Poor_Health_Frailty
- Risk_Excessive_Thinness

Region	Family History of Osteoporosis (%)	Low Calcium Intake (%)	Vitamin D Insufficiency (%)	Poor Health/Frailty (%)	Excessive Thinness (%)
South	15.2%	10.1%	8.7%	5.4%	4.2%
Midwest	14.8%	9.9%	8.5%	5.2%	4.1%
West	16.0%	11.3%	9.0%	5.9%	4.8%
Northeast	13.5%	9.5%	7.8%	4.7%	3.9%

EDA

Risk Factors Distribution:

Risk Factor	Percentage (%)
Family History of Osteoporosis	14.8%
Low Calcium Intake	9.9%
Vitamin D Insufficiency	8.5%
Poor Health/Frailty	5.2%
Excessive Thinness	4.1%

Age Bucket	Family History of Osteoporosis (%)	Low Calcium Intake (%)	Vitamin D Insufficiency (%)	Poor Health/Frailty (%)	Excessive Thinness (%)
>75	17.3	12.1	9.5	7.2	4.6
65-75	15.8	10.7	8.2	6.1	4.3
55-65	13.6	9.2	7.1	5.4	3.8
<55	11.9	8.3	6.3	4.5	3.2

Risk Factor	Female (%)	Male (%)
Risk_Family_History_Of_Osteoporosis	14.5	10.3
Risk_Low_Calcium_Intake	11.2	7.4
Risk_Vitamin_D_Insufficiency	9.8	5.6
Risk_Poor_Health_Frailty	6.3	4.1
Risk_Excessive_Thinness	5.2	3.5

EDA Summary

The file contained information of 3, 424 patients. For each patient it has demographic information, clinical records, others diseases as risk factor information and also about their physicians specialty.

There are some significant differences between genders (vitamin D deficiencies, screening for malignant neoplasms, Hypogonadism).

Most of the patients already hold comorbidity factors, while holding risk factors is less common.

Patients older than 65 are affected by the mentioned factors in a higher proportion.

There seem to be some remarkable differences between Asian and other races.

Variables that are recorded during the treatment like Dexa_Freq_During_Rx, Dexa_During_Rx and Gluco_Record_During_Rx have more useful information for the classification than others.

Model proposal

- **Support Vector Machines** algorithm to classify the persistence of patients (1 for positives and -1 for negatives). The whole dataset is composed of 3424 feature vectors of 83 dimensions, plus the target variable. A linear kernel has been used, obtaining an **accuracy of 83.5 %** over testing data (25 % out of the whole dataset).

- **Random Forest** algorithm for classification (1 for positives and 0 for negatives). The algorithm has 1000 estimators, max_depth of 10, obtaining an accuracy of 81% and AUC of 89% over testing data.

- **Decision Tree** algorithm (0 for Persistent and 1 for Non-persistent). The input is composed by 64 features with 3424 observations. The tree got best predictions with max depth of 1, obtaining an **accuracy of 76%** on test data.

- **Logistic Regression** algorithm for binary classification. The labels are the following: 0 for Non-Persistent and 1 for Persistent. Using GridSearchCV for optimization, the LR model uses 204 columns (after one-hot-encoding) to train. The f1_score obtain is **82%**.

Thank You

Data Glacier
Your Deep Learning Partner