Group Project Deliverables Week 7

Project Name: Healthcare - Persistency of a drug
Report date: 19.09.2023
Internship Batch: LISUM24

- **Team member's details**

Group: HealthData Collective
Group members:

Name	email	country	College/Company	Specialization
Selena	selena.elensto@gmail.com	Macedonia	-/-	Data Science
Anusha Asim	anushaasim21@yahoo.com	UAE	Regent Middle East	Data Science
Sukurat	salamsukurat@gmail.com	United Kingdom	Teesside University	Data Science
Khushboo	Khushboomasih193@gmail.com	United Kingdom	University of Aberdeen	Data Science

- **Problem description**

The problem to be solved is related to pharmaceutical companies' need to understand the persistence of drug usage among patients based on physician prescriptions. Specifically, the challenge is to automate the process of identifying whether a patient is persistent in following their prescribed therapy regimen. The primary problem is to build a classification model using machine learning to determine the persistency of patients based on various features and factors.

- **Business understanding**

The business understanding is about using data-driven insights and machine learning to tackle the problem of medication persistency and improve patient adherence to prescribed therapies, which can ultimately lead to better patient outcomes and reduced healthcare costs.
This problem revolves around the pharmaceutical industry's interest in ensuring that patients adhere to their prescribed medications and therapies. Poor medication persistency can lead to health complications and increased healthcare costs. Therefore, pharmaceutic company aims to gain insights into the factors that influence a patient's persistency in following their prescribed therapies.
Key points of business understanding include:

- **Data Sources**: Data for this analysis is gathered from various sources, and includs patient demographics, physician information, clinical factors, provider attributes, disease/treatment factors, and adherence information.
- **Target Variable**: The primary outcome of interest is the "Persistency_Flag," which indicates whether a patient was persistent or not in adhering to their prescribed therapy.

- **Factors Impacting Persistency**: The analysis aims to identify factors that impact persistency, such as age, race, region, ethnicity, gender, physician specialty, clinical factors like T-Score and risk segment, use of multiple risk factors, Dexa scan frequency and recency, fragility fractures, glucocorticoid usage, comorbidities, concomitant drug usage, and adherence to therapies.

- **Machine Learning Approach**: The solution involves building a classification model that uses these features to predict whether a patient is likely to be persistent in following their prescribed therapy. This model will help the pharmaceutic company to identify at-risk patients and potentially develop targeted interventions to improve persistency rates.

- **Project lifecycle along with deadline**

WEEK #	TASKS:
WEEK 8	- Data understanding (types of data, issues like missing values, outliers, skewness) - Approaches planned to handle data issues (e.g., NA imputation, outlier detection and handling)
WEEK 9	- Data cleansing and transformation steps performed - Application of at least 2 techniques for data cleansing (e.g., handling NA values, outlier treatment)
WEEK 10	- EDA performed on the Data - Final recommendations based on EDA findings
WEEK 11	- Prepare an EDA presentation for business users - Include a technical user slide with recommended models for the dataset
WEEK 12	- Select base model and explore models from different families (Linear, Ensemble, Boosting, etc.) - Submit a dashboard - Ensure selected models align with business requirements - Merge code from individual Team members into a single Repo
WEEK 13	- Provide the link to the code and report - Hold a Team discussion to Select the best solution that meets project requirements - Prepare a PowerPoint presentation - Merge code from individual Team members into a single Repo

Data Intake Report

Project Name: Healthcare - Persistency of a drug
Report date: 19.09.2023
Internship Batch: LISUM24
Version:<1.0>
Group: HealthData Collective
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/venusflytrapfairy/HealthDataCollective-Group

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.csv
Size of the data	898 KB

Proposed Approach:

1. **Data Profiling and Understanding:**
   - Getting know the 4 datasets and the columns within them.
   - Checking for total number of observations and number of unique values in each column.

2. **Data Cleaning:**
   - Checking for missing values, converting data types, and formatting data:

3. **Deduplication Criteria, Approach, and Implementation:**
   - Identifying the columns that are relevant for deduplication:
   - Hashing: selected due to the length of the dataset. It calculates hash values for records based on selected columns and compares hash values.
   - Implementation: on the cleaned dataset, considering selected columns.
   - Describe the Results and Action taken during the deduplication.

4. **EDA analysis:**
   - Perform Exploratory Data Analysis.

5. **ML modeling:**
   - Explore ML various models.
   - Select a ML model with the best evaluation values and that fit business requirements.

6. **Findings and Solutions:**
   - Prepare dashboard and presentation explaining the findings and the solutions provided.