Ali-Just Team Members

Name: Alireza Samadifardheris

Email: [alirezasamadii71@gmail.com](mailto:alirezasamadii71@gmail.com)

Country: Rome, Italy

College: Sapienza University of Rome, Computer Engineering

Specialization: Data Science

Name: Justin Lee

Email: [justindavinlee@gmail.com](mailto:justindavinlee@gmail.com)

Country: Ontario, Canada

College: Wilfrid Laurier University, Data Science Concentration Big Data

Specialization: Data Science

Problem Description

One of the challenges for all Pharmaceutical companies is to understand the persistence of drugs as per the physician’s prescription. the persistency of a drug may be defined as “the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen.” Medication persistence refers to the act of continuing the treatment for the prescribed duration.

Data Understanding

The data was retrieved from <https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view> where there are 69 features, and 3423 entries. The dataset was originally in an .xlsx format, which includes a unique row id (Patient ID), a target variable (Persistency\_Flag), and many features ranging from demographics, provider attributes, clinical factors and disease/treatment factors which may be used to train and test the machine learning models. The unique row id (Patient ID) will be dropped since using a unique row id as a feature in our models will result in overfitting the data. Most features are of the ‘object’ dtype, which are categorical, as well as strings. These features will have to be converted to numeric values or dummy variables. We have also decided to change one of the names of the features, due the format having problematic symbols. In terms of problems within the data, there are no missing values within the dataset, although the dataset was skewed and heavily unbalanced. To solve the issue of skewness, we will remove skewness from features with low correlation, as removing skew from data with higher correlation will have a higher impact and could be problematic. To deal with the problem of the unbalanced data, we will perform sampling techniques such as: RandomUnderSampler and SMOTE (Synthetic Minority Oversampling Technique).

Project Lifecycle Along with Deadline

**Start Date:** November 17th, 2022

**Exploratory Data Analysis:** November 25th, 2022

**Data Cleaning and Transformation:** December 1st, 2022

**Recommendations:** December 8th, 2022

**Presentation:** December 15th, 2022

**Model Selection/ Building:** December 22nd, 2022

**Final Project Report and Code:** December 30th, 2022

Data Intake Report

Name: Healthcare Industry

Report date: November 26th, 2022

Internship Batch: LISUM14

Version: 0.1

Data intake by: Alireza Samadifardheris and Justin Lee

Data intake reviewer:

Data storage location: https://docs.google.com/spreadsheets/d/1P\_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/edit#gid=2047360270

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 3424 |
| **Total number of files** |  |
| **Total number of features** | 69 |
| **Base format of the file** | .xlsx |
| **Size of the data** | 900 KB |

**Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.**

**Please do not forget to remove this section while converting the file into pdf.**