Data Science Project

Project: Healthcare - Persistency of a Drug

Week 7 Deliverables

Team Name: Team Healthy Bones

Member: Jeeyeon Shon

E-mail: shon.jeeyeon@gmail.com

College/Company: Data Glacier

Batch Code: LISUM 10

Country: United States of America

Specialization: Data Science

Submission Date: July 19, 2022

Submitted to: Data Glacier

Github Link: https://github.com/shonjeeyeon/DG_Week_7

Table of Contents

Problem Description

A model will be established and deployed to automate identifying persistency of a certain pharmaceutical product.

Records of 3,424 patients who take the medication will be used for analysis, and correlation between medication persistency and other factors such as patient demographics, provider attributes, clinical factors, and disease factors will be investigated. Finally, an optimal model to predict persistency based on above features will be selected and developed.

Business Understanding

Medication persistence refers to a patient's continued action of taking medications for the duration instructed by the prescriber. Therefore, persistency of a drug is a critical factor which contributes to industry profits as well as patient outcomes.

Using data in real cases, a machine learning model can learn relationships between target variables (persistence) and dependent variables, such as patient related variables, prescriber attributes, and indicators of disease severity (e.g.: DEXA scans, t-scores of the bone, and history of bone fractures) at the beginning and during the therapy. Eventually, the model will be able to predict whether the patient will be persistent in medication therapy, given the values of dependent variables.

Therefore, the prediction model will benefit the company by providing a detailed picture of persistence of the medication product, which is valuable for marketing of the product or developing of a new product.

Project Lifecycle and Deadlines

Weeks	Tasks	Due Dates
Week 7	• Choose the topic • Acquire the dataset • Plan weekly tasks	Jul 19, 2022
Week 8	• Inspect the dataset for issues: ○ Missing values ○ Skewness ○ Outliers	Jul 26, 2022
Week 9	• Clean the data using at least 2 techniques ○ Impute null values ○ Resolve skewness/outlier issues	Aug 02, 2022
Week 10	• Perform EDA • Provide final recommendations	Aug 09, 2022
Week 11	• Present the EDA • Include model recommendation at the end of the presentation	Aug 16, 2022
Week 12	• Explore one model from each family: ○ Linear Models ○ Ensembles ○ Boosting ○ Other models, such as stacking (Optional)	Aug 23, 2022
Week 13	• Decide the best solution • Deliver a Powerpoint presentation	Aug 30, 2022

Data Intake Report

Name: Healthcare - Persistency of Drug
Report date: July 18, 2022
Internship Batch: LISUM 10
Version:<1.0>
Data intake by: Jeeyeon Shon
Data intake reviewer:
Data storage location:
https://drive.google.com/file/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/view

Tabular data details: Healthcare_dataset.xlsx

Total number of observations	3,424
Total number of files	1
Total number of features	69
Base format of the file	.xslx
Size of the data	898 KB

Proposed Approach:

- Remove the "Feature Description" sheet and save the data to .csv
- Ptid can be used to identify and remove duplicate observations
- Patient identifiable information had been removed already
- Most of the features are categorical; will need encoding to enable ML processing

Github Repo Link

https://github.com/shonjeeyeon/DG_Week_7