Project: Data Science :: Persistency of a Drug

Week 10: Deliverables

Presented by: Sammy Suliman, Olivia Foster, and Tahsin Azad

Table of Contents:

Requirements:

Submit a pdf document and EDA ipynb file which should contain following details:

Team member's details : Group Name (give a name to your group), Name, Email, Country, College/Company, Specialization (Data Science, NLP, Data Analyst)

Problem description

Github Repo link

EDA performed on the data

Final Recommendation

Team Member Details:

Group Name: Call it Version 1.0

Member 1:

- **Name:** Sammy Suliman
- **Email:** sammysuliman@gmail.com
- **Country:** USA
- **College/Company:** UC Santa Barbara
- **Specialization:** Data Science

Member 2:

- **Name:** Olivia Foster
- **Email:** livia.n.foster@gmail.com
- **Country:** USA
- **College/Company:** NT Logistics
- **Specialization:** Data Science

Member 3:

- **Name:** Tahsin Azad
- **Email:** tahsinazad31@gmail.com
- **Country:** USA
- **College/Company:** University of California, Santa Barbara
- **Specialization:** Data Science

Problem Statement:

One of the challenges faced by Pharmaceutical companies is the persistence of a drug (that is, the extent to which a patient will act in accordance with the prescribed time interval, and dose of a medication) as the physician prescribed it. In this problem, we will automate the process of classifying factors that determine the persistence of a drug through Machine Learning and Python.

Drug persistence is a task of classifying different disorders and a patient's medical history to determine the dose and length of dose. In order to train our model, we will need to classify risk factors, medical histories, and disorders. To do this, we will be using a dataset based on over 3000 patients' records.

Project Deadlines:

Week:	**Due Date:**	**Plan:**	**Completed:**
Week 7	1/19/2023	Problem Statement, Data Collection, Data Report	Yes
Week 8	1/26/2023	Data Preprocessing	Yes
Week 9	2/2/2023	Feature Extraction	Yes
Week 10	2/9/2023	Building Model	In Progress
Week 11	2/16/2023	Model Result Evaluation	No
Week 12	2/23/2023	Flask Development and Web Application	No
Week 13	2/28/2023	Final Submission (Report + Code + Final Submission)	No

Data Intake Report:

Name: Final Project -- Data Science:: Healthcare - Persistency of a drug:: Group Project
Report date: 1/17/2023
Internship Batch: LISUM16
Version:<1.0>
Data intake by: Olivia Foster
Data intake reviewer: Tahsin Azad
Data storage location:
https://github.com/LiviaNFoster/DataGlacierFinalProject/blob/main/Healthcare_dataset.xlsx

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	xlsx
Size of the data	922 KB

Proposed Approach:

- Unique Row identified using with the “Ptid” column, a unique ID for each patient
- No duplicates of Unique ID confirmed
- Assumptions
  - No false positives or errors during any of the testing
  - No dishonest and unbiased data recorded

Data Cleaning:

We began this process last week and continued to move forward with it.

Null Values:

There were no null values.

Yes versus No:

We had over half of the data be defined as either Y or N, “Yes” or “No” respectively. We replaced Y and N with 1 and 0 respectively.

Renaming Columns:

Because many of the columns had long and difficult names or were not great descriptors, we went ahead and renamed many of them to make referencing easier and to add clarity.

Case sensitive:

Made all words lowercase to make referencing easier without having to worry about case sensitivity.

Acknowledgement of outliers:

In exploring our data, we discovered that because our data is mostly binary (such as above in the ‘Yes versus No’ section) that there were no significant outliers at this time.

‘Age_Bucket’ addressed:

In the original data, age was not presented in the same manner as in the description. Instead of producing quantifiable numbers for analysis, the values were grouped together as objects. To rectify this, each ‘bucket’ was replaced by a normal random output of values that fit the parameters of each ‘bucket.’

Data Preprocessing:

Encoding Categorical Data:

The data had certain categorical values. These values were changed into integer format, which will later be used for the model. Target Encoding, Label Encoding, One-Hot Encoding and Smoothing were observed.

Key Errors fixed:

Small key errors fixed, allowing for clean, correct encoding.

Exploratory Analysis:

Correlation:

Before building our model, we prepared several correlation maps to investigate what would have the greatest influence on our model for persistence. We found several higher level influencing factors but that reach at most a correlation 0.35.

Distribution charts:

A distribution chart for the age variable was prepared, displaying the mean and median age for the patients. This helps to visualize the data.

Box plot:

A box plot was created to look at any outliers. Since nearly all of the data is binary, only the boxplot for the number of "Risk_Counts" was observed. The boxplot proved that the "Risk_Counts" column had outliers.

Github Repo links:

General Repository:

https://github.com/LiviaNFoster/DataGlacierFinalProject.git

Week 10 Jupyter Notebook:

Please note that this file includes all of our work, where we each worked on it one at a time.

https://github.com/LiviaNFoster/DataGlacierFinalProject/blob/main/Week10code.ipynb