Group details

Group Name: TBC

Team member:

Name	Email	Country	College	Specialization
Yuchi Chen	ychen306@ur.rochester.edu	United States	University of Rochester	Data Science
Bolutife Akinlawon	bolu.akinlawon@gmail.com	United Kingdom	UWE, Bristol	Data Science
Alexis Collier	colliera75@gmail.com	United States	Emory University Bootcamp	Data Science
Han-Fu Lin	hanfu.lin@mail.utoronto.ca	Canada	University of Totonto	Data science

Github Repo link: [Healthcare Persistency of a drug]

Problem description

A pharmaceutical company conducts a large number of clinical trials in order to study the durability of a new drug. These trials record a large number of different attributes of experimental subjects and the results of the experiment by means of control variables. The company wants to use the data to understand what properties affect the drug's durability.

Business understanding

Problem➜ Model ➜ Solution

Problem:

According to the indicators in the data set and their results, we will analyze the influence of these indicators on the drug duration.

Model/Solution:

For the numeric variable part, we plan to apply a linear regression model to predict a new attribute.

For the binary variable part, we plan to use a classification model to divide the data into several clusters.

The final result will be the outcome of these two new attributes.

Project life cycle along with deadline

Task	Details	Start Date	End Date
Make team	Make a team and discuss the timeline	11/12/22	11/19/22
Preprocess	• Load the data into a database and perform data cleaning and transformation(if necessary). • A data understanding report will be produced in this step.	11/19/22	12/02/22
EDA	• Perform the EDA using Python and Tableau. • Output EDA Presentation and proposed modeling technique. • Present a dashboard for showing results if available.	12/02/22	12/09/22
Model Building	Select a base model and then explore 1 model of each family if its classification problem then 1 model for Linear models, 1- Model for Ensemble, 1-Model for boosting and other models	12/09/22	12/16/22
Final Result	Upload final version	12/16/22	12/23/22

Data Intake Report

Name: Data Science:: Healthcare - Persistency of a drug

Report date: 11/18/2022

Internship Batch: LISUM14

Version: 1.0

Data intake by: DataGlacier

Data intake reviewer: The whole team

Data storage location: DataSet

Tabular data details:

Total number of observations	3424
Total number of files	1
Total number of features	68
Base format of the file	.xlsx
Size of the data	899 KB

Proposed Approach:

- Read file, do data quality check and data cleaning using pandas, including missing values and mismatch data
- May build the dashboard for EDA using Tableau

- Assumption: Id column is not needed and thus will be removed in the course of analysis