Group Name: Ignotus

Team Member Details:

- Corey Hamren, hamhams86@gmail.com, USA, Self-teaching, Data Science
- Motamen MohammedAhmed, Motamen.salih@hotmail.com, UAE, Zayed University, Data Science
- Francis Kim, fkim39@gmail.com, Fkim39@gmail.com, USA, University of Maryland, Data Science
- Inna Soltsman-Groysman, innasol90@gmail.com, USA, Data Science

Problem Description:

There are different diseases that affect people around the world. Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Business Understanding:

There are different diseases that affect people around the world. Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Project Life-cycle and Deadline:

Task	Beginning Date	Due Date
Problem understanding (week7)	09/27/2021	10/04/2021
Data understanding (week8)	10/04/2021	10/11/2021
Data cleansing and transformation done on the data. (week9)	10/11/2021	10/18/2021
EDA performed on the data. (week10)	10/18/2021	10/25/2021
EDA Presentation and proposed modeling technique. (week11)	10/25/2021	11/01/2021
Model Selection and Model Building. (week12)	11/01/2021	11/08/2021
Final Project Report. (week13)	11/08/2021	11/15/2021

Data Intake Report:

Name: Medicare Group Project
Report date: 09/29/2021
Internship Batch: LISUM03
Version: 1.0
Data intake by: Ignotus
Data intake reviewer: Data Glacier
Data storage location: https://github.com/TheHamhams/dg_group_project

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	0.88 MB

Proposed Approach:

In terms of dedup validation, we will used Excel's duplicate identification feature to identify and remove duplicate addresses (if any). In addition, null values can be identified in either Excel or Python and removed. However, we found the column value: "Unknown" to be useful and not the same of a null value. The reason being is that column value "Unknown" is an important count for analyzing the distribution count of data we do have Vs. the unknown values. Dataset will also be cleaned through the lowering of column titles in Python.

GitHub Repository Link:

- https://github.com/TheHamhams/dg_group_project