Group Name: Gamma
Name: Matthew Iversen, Jonathan Charles Allen
Email: matt.w.iversen@gmail.com, jonathancharlesallen@live.co.uk
Country: USA, UK
College/Company: N/A
Specialization: Data Science
Github Repo link: data-glacier-internship/week-7

Problem description: This project revolves around understanding the persistence of a drug based on physician prescriptions. The objective is to automate predictions of a persistency flag using a classification machine learning model. After this analysis, we will identify the most important factors in predicting persistence.

Business understanding: ABC Pharma Company aims to address a key challenge in the pharmaceutical sector: understanding drug persistency as per physician prescriptions. They seek to automate the identification of factors affecting drug adherence. The project involves creating a machine learning model to predict the "Persistency_Flag," indicating whether patients will likely stick to their prescribed medication regimens. The model will analyze patient demographics, clinical factors, and treatment histories to uncover insights into what drives medication persistence, ultimately improving patient health outcomes and medication effectiveness. Aside from health benefits, predicting persistence will help reduce healthcare costs due to fewer follow-up problems.

Project Lifecycle

Week 7: Deliverables

- Due Date: 11/19/2023
- Submit a PDF document containing the following details:
    - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
    - Problem description
    - Business understanding
    - Project lifecycle along with deadlines
    - Data Intake report
    - Github Repo link

Week 8: Deliverables

- Due Date: 11/26/2023
- Submit a PDF document containing the following details:
    - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
    - Problem description
    - Data understanding
    - What type of data do you have for analysis
    - What are the problems in the data (number of NA values, outliers, skewed, etc.)
    - What approaches are you applying to your data set to overcome problems like NA value, outlier, etc., and why?
    - Github Repo link

Week 9: Deliverables

- Due Date: 12/2/2023
- Submit a PDF document and ipynb notebook containing the following details:
    - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
    - Problem description
    - Github Repo link
    - Data cleansing and what transformations are done on the data.
    - Try at least 2 techniques to clean the data (for NA values: mean/median/mode/Model-based approach to handle NA value/WOE; like this, try different techniques to identify and handle outliers as well).
    - Try different featurization techniques for NLP and clean the data using regex and Python.

- Each member should code and review peers' work. (Review comments should be present in the GitHub repo)
- Each team member should work on a different data cleansing approach.

Week 10: Deliverables

- Due Date: 12/9/2023
- Submit a PDF document and EDA ipynb file containing the following details:
  - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
  - Problem description
  - Github Repo link
  - EDA performed on the data
  - Final Recommendation

Week 11: EDA Presentation and proposed modeling technique

- Due Date: 12/16/2023
- Submit a website URL containing the following details:
  - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
  - Problem description
  - Github Repo link
  - EDA presentation for business users
  - The last slide of EDA should be dedicated to technical users which should contain recommended models for this data set.

Week 12: Model Selection and Model Building/Dashboard

- Due Date: 12/23/2023
- Submit a website URL containing the following details:
  - Team member's details: Group Name, Name, Email, Country, College/Company, Specialization
  - Problem description
  - Github Repo link
  - Select your base model and then explore 1 model of each family (if it's a classification problem, then 1 model for Linear models, 1 model for Ensemble, 1 model for boosting, and other models if you have time, like stacking)
  - Please make sure the selected model fits your business requirements. For example, If your business does not want a black-box model, select only those models that can be used to explain the prediction.
  - Interns of the Data analysis Project should submit a dashboard this week.

Week 13: Final Project Report and Code

- Due Date: 12/30/2023

- Submit a text entry box containing the following details:
    - Provide the link to your code and report.
    - As it was a group assignment, go for a call with your team to discuss each member's solution and select the best solution per the requirement.
    - PowerPoint presentation is a must.

Data Intake Report

Name: Persistency of a Drug
Report date: 11/19/2023
Internship Batch: LISUM 26
Version: 1.0
Data intake by: Matthew Iversen, Jonathan Charles Allen
Data intake reviewer:
Data storage location: Healthcare_dataset.xlsx - Google Drive

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	0.88 MB

Proposed Approach:

- Using Pandas:
  - Utilize the Pandas library to identify duplicates based on specific columns.
- Using Python Sets:
  - Convert columns to sets and compare them to identify duplicates.
- Using Grouping and Aggregation:
  - Group data by relevant columns and count occurrences to find duplicates.
- Using Fuzzy Matching (Fuzzy Deduplication):
  - Implement fuzzy matching techniques to find similar but not identical records.
- Assumptions for Data Quality Analysis:
  - Assuming that missing values in certain columns are handled appropriately (e.g., through imputation, removal).
  - Assuming that outliers in the data are addressed or treated according to the chosen approach in the future week.
  - Assuming that the data provided is accurate and representative of the problem domain.
  - Assuming that the data preprocessing steps, such as cleaning and featurization, are performed effectively to prepare the data for analysis.