Problem Understanding:

Business Context:

- The pharmaceutical company faces the challenge of understanding a patient's adherence to the prescribed drugs by the physicians. Understanding drug persistency, Whether the patient continues to take medicine as prescribed, can have significant implications for patient health outcomes and company's business performance.

- By analyzing the factors that influence persistency, the ABC company can make targeted strategies to improve adherence rates, leading to better health outcomes and increased revenues.

Aim:

- The aim of the project is to build a classification model to predict the persistency of a drug based on various factors related to prescription of the physicians, and patient characteristics. The model aims to identify the key factors that influence the patient to adhere to their prescribed medicine regimen.

Key terms:

- Persistency - It refers to continuity of taking prescribed medicines for a given period of time as directed by the physician.
- Persistency_Flag - If the patient is being persistent over taking medicines, it's mentioned as 1, otherwise 0.

Stakeholders:

- ABC company
- Physician
- Patient

Data Understanding:

- The dataset contains information on

1. Patient demographic data such as Age, race, region, ethnicity and gender
2. Provider data - Information about prescribing physician's speciality

3. Clinical Data - includes T-Score, risk segments, DEXA scan frequencies, fragility fractures, glucocorticoid usage, changes in T-Scores and risk segments
4. Disease and Treatment Data - Information on injectable drug usage, risk factors,comorbidities,concomitant drugs
5. Adherence data - Information on adherence to therapy.

- To identify potential issues in the data such as missing values, outliers, skewed distributions, we need to conduct an exploratory data analysis.

1. First, we proceed with finding missing values by using methods like isnull() or isna() in pandas. Then calculate the percentage of missing values in each column. Columns with a high percentage of missing values need to be dropped or imputed with appropriate values. This could be a mean, median, mode or using advanced imputation techniques.Evaluate the importance of each column before deciding to drop or impute.
2. Using box-plots or histogram to visually inspect for outliers in numerical data. We can also use Z-score and IQR methods to identify the outliers.
3. Identify Skewed features and apply transformations to normalize them.
4. Check for imbalance in the target variable and apply appropriate technique to address it.
5. Identify highly correlated features to avoid multicollinearity.