Persistency of a drug - Data Science Project

Virtual Internship – Yusuf Shamsi

Background/Introduction

- One of the challenges for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.
- ABC Pharma aims to automate the drug persistency identification process to improve treatment efficacy and patient care. They want to develop a machine learning model to classify patients based on adherence to prescribed medication, using the "Persistency_Flag". This effort seeks insights into factors affecting drug persistency for targeted patient adherence interventions.
- The model will analyse Clinical Factors—like T-Scores, changes in T-Score, risk segments, DEXA scans, fragility fractures, and glucocorticoid usage—providing insights into the patient's health status. Disease/Treatment Factors will also be considered, including treatment history, risk factors, comorbidities, and concomitant drug usage, offering a comprehensive view of the patient's medical history and treatment regimen. Together, these factors aim to predict patient persistency, enabling targeted support for medication adherence.

Data Background

- Healthcare_dataset – 2nd Sheet
- 70 Features (including 1 derived features)
- Total data points: 3,424
- **General Observations**
  - Predominantly categorical data with significant "unknown" entries
  - Boolean variables are prevalent, along with categories containing 'unknown' data
  - Essential to integrate domain knowledge to discern data recording errors from valid rarities
- Target Variable was ‘Persistency_Flag’ – values were mapped as 'Persistent': 1, 'Non-Persistent’: 0
- Ptid – was the index (Patient id number)

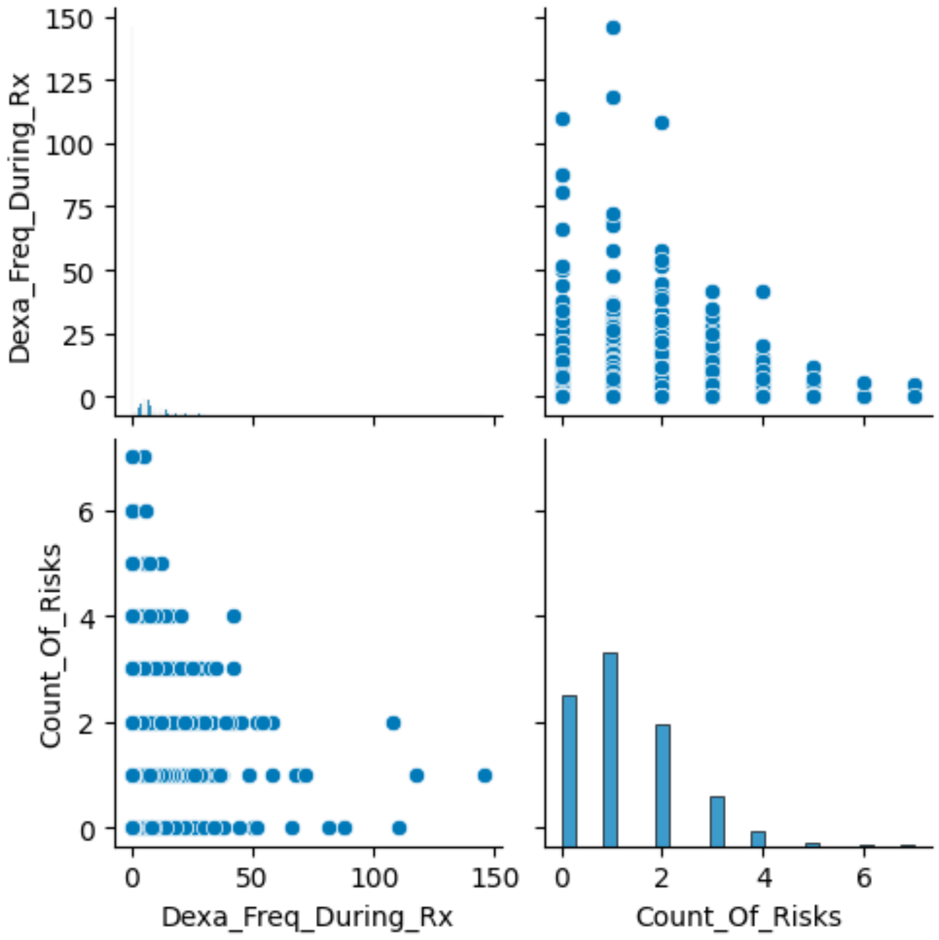
Data Pre-processing - Categorical Data Challenges

- Ntm_Speciality: 310 observations labeled as unknown
- Ntm_Speciality_Bucket: Unintuitive structure, mixing unknowns with specific categories
- Ethnicity: 91 unknown values; limited to Hispanic and Non-Hispanic categories
- Race: 97 entries of "Other/Unknown"
- Region: 60 entries of "Other/Unknown"
- Risk_Segment_During_Rx: 1,497 unknown values; majority unspecified
- Tscore_Bucket_During_Rx: 1,497 unknown values; majority unspecified
- Change_T_Score: 1,497 unknown values; majority unspecified
- Change_Risk_Segment: 2,229 unknown values; majority unspecified
- Strategies for Addressing Issues:
  - Restructure Ntm_Speciality_Bucket, creating an 'Other' category for infrequent specialties
  - Remove unknown values under 5%; retain others to avoid information loss
  - Conduct sensitivity analysis to evaluate the impact of outliers on model performance

Data Pre-processing – Numerical Data

- Dexa_Freq_During_Rx: 3,424 observations; investigates DEXA scan frequency, with notable outliers
- Count_Of_Risks: Detected outliers with 6 and 7 risks counted
- Outliers in Dexa_Freq_During_Rx above the upper IQR limit: 460 observations
- Data Analysis Approach:
  - Assess if zero values represent missing data or actual measurements
  - Retain outliers, assuming zeros are prevalent and high values may be significant but rare

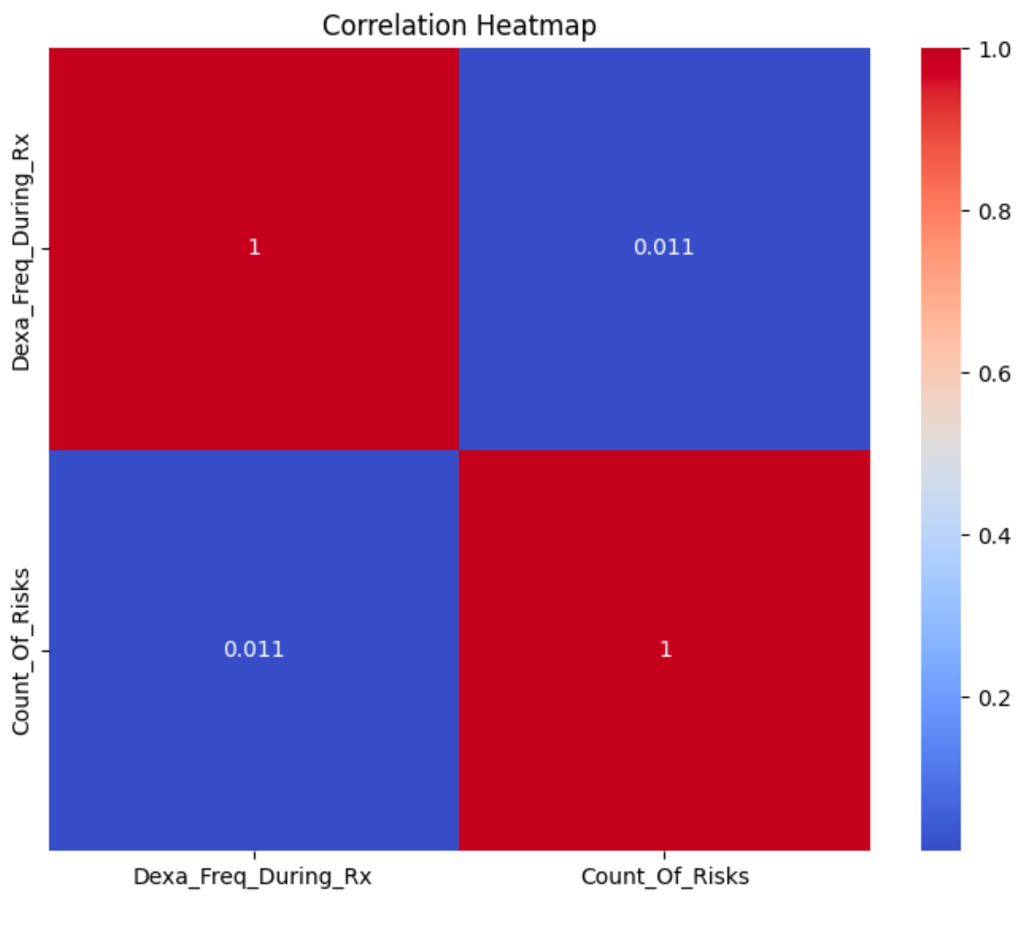
Association Between Dexa Frequency and Number of Risks



The scatter plot suggests a weak to no clear correlation between the frequency of Dexa during prescription and the count of risks. Data points are heavily clustered at lower values with few outliers, indicating that most observations have a low Dexa frequency irrespective of the number of risks.

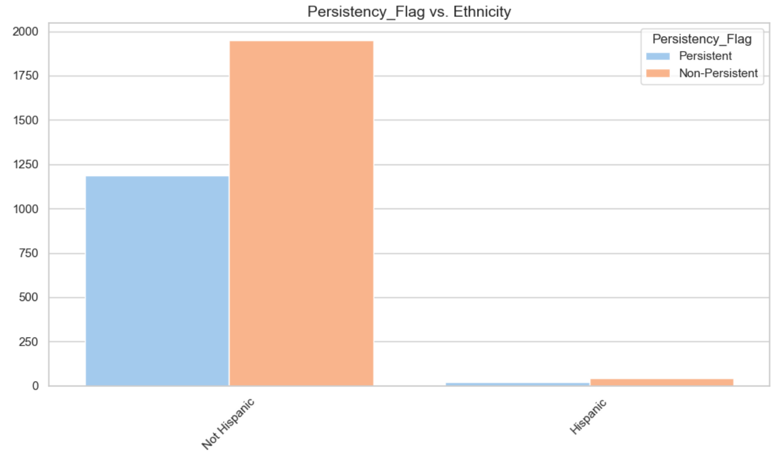
Correlation Between Dexa Frequency and Risk Count

- The correlation heatmap shows a very weak positive correlation (0.011) between Dexa frequency and risk count. This implies that there is no significant linear relationship between these two variables.

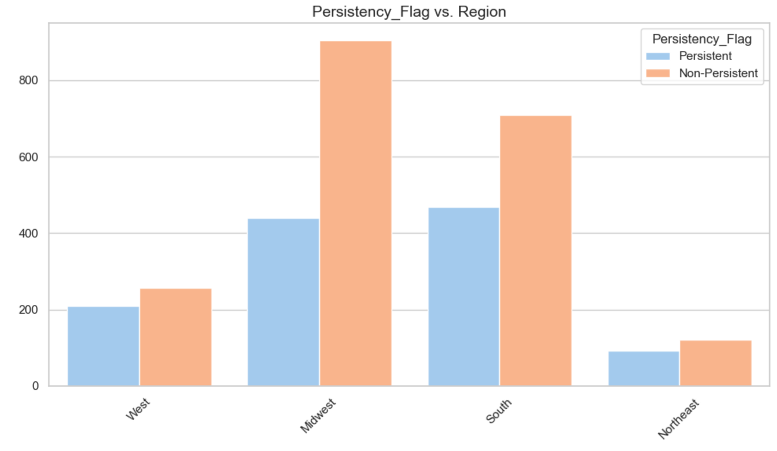


Impact of Ethnicity, Region, and Other Factors on Treatment Persistency

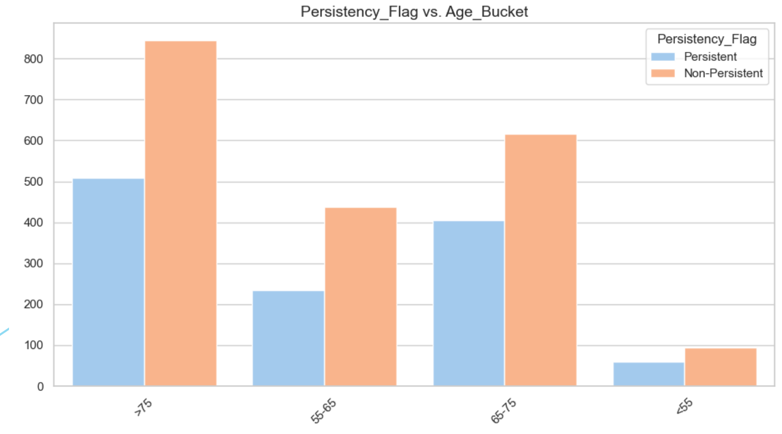
The series of bar charts would be can to examine the impact of various factors like ethnicity, race, region, and age on treatment persistency. They show how each subgroup contributes to the overall persistency and non-persistency rates. Trends in these charts may suggest which demographics or characteristics are more likely to be associated with higher or lower treatment adherence.



Persistency_Flag vs. Ethnicity



Persistency_Flag vs. Region



Persistency_Flag vs. Age_Bucket

Statistical Association Between Persistency_Flag and Other Categorical Variables – First 10 Values Shown

This table presents the results of the Chi-squared tests for independence, analyzing the association between various categorical variables and the Persistency_Flag feature.

The analysis considered a p-value threshold of ≤0.05 to denote statistical significance. All listed variables meet this criterion, which suggests that they have a statistically significant association with treatment persistency.

The variables are sorted in ascending order of their p-values, with the lowest values at the top. This order prioritizes the variables according to the strength of their association, with the strongest predictors appearing first.

Comparison against Persistency_Flag	P_Value
Persistency_Flag	0.000000e+00
Dexa_During_Rx	1.130748e-172
Comorb_Long_Term_Current_Drug_Therapy	7.012163e-91
Comorb_Encounter_For_Immunization	2.250979e-72
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	6.952994e-72
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	2.779927e-59
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	5.088962e-46
Concom_Systemic_Corticosteroids_Plain	3.034717e-43
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	1.184227e-39
Concom_Anaesthetics_General	1.346314e-37
…	…

Model Selection

- **3 Models will be chosen for Persistency_Flag prediction :**
  - **Logistic Regression (Baseline Measurement):**
    - Chosen for its simplicity and efficiency in binary classification tasks.
    - Provides clear probabilistic interpretation of model outputs.
    - Useful for understanding the influence of individual features on the outcome.
  - **Random Forest Classification:**
    - Selected for its robustness and ability to handle non-linear data.
    - Excels in handling large datasets with higher dimensionality.
    - Offers insights into feature importance, enhancing model interpretability.
  - **XGBoost:**
    - Opted for its high performance and speed in training.
    - Known for delivering state-of-the-art results on many machine learning challenges.
    - Incorporates built-in regularization, reducing overfitting and improving overall model performance.