Data Glacier

Your Deep Learning Partner

Final Presentation

HEALTHCARE - PERSISTENCY OF A DRUG

Group Name: DS_SS

Name: SONIYA SUNNY

Email: soniyasunny1210@gmail.com

Country: Canada

Company: Data Glacier

Specialization: Data Science

Agenda

Executive Summary
Problem Statement
Approach
Data Cleaning
EDA
EDA Summary
Models
Recommendations

Problem Statement

- To identify the persistency of a drug, a pharmaceutical company approached to develop a model based on data analysis.
- Factors that affect the persistence of drugs should be identified, along with data insights with predictive analytics, to help the company for their smooth and efficient functioning, with the help of dataset provided by the company.

Approach

- Collect data
- Analyze the data
- Detect outliers
- Combine datasets
- Curate the data
- Feature engineering
- Detect correlations
- Analyze patterns
- Provide insights
- Tools used – Microsoft Excel, Python, Microsoft Power BI

Data Cleaning

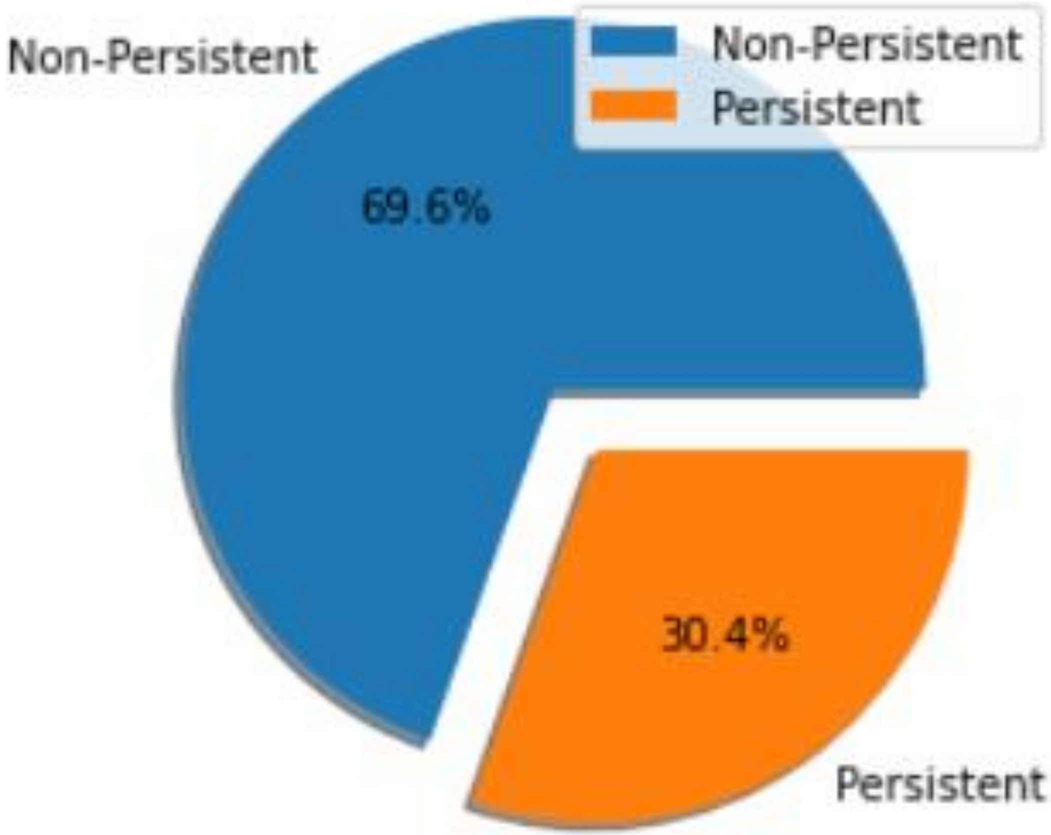
- Dataset details:
  - ➢ File name: Healthcare_dataset.xlsx
  - ➢ No. of rows = 3424
  - ➢ No. of columns = 69
  - ➢ Target variable = **‘Persistency_Flag’**
- There were no null values in the dataset.
- Some variables have large number of value counts, which is reduced by grouping same names, treating “Unknown” values, etc.
- A new dataset was created after data curation steps, with 2942 rows & 66 columns.

Exploratory Data Analysis

1. Basic Data Exploration
2. Demographic Analysis
   - Gender
   - Age
   - Race
   - Region
   - Ethnicity
3. Clinical Factors' Analysis
   - T score
   - Risk Segment
   - Glucocorticoid Recency
   - Fragility Fracture Recency
4. Disease/Treatment Factors' Analysis
   - Comorbidity
   - Risk factors
   - Concomitancy

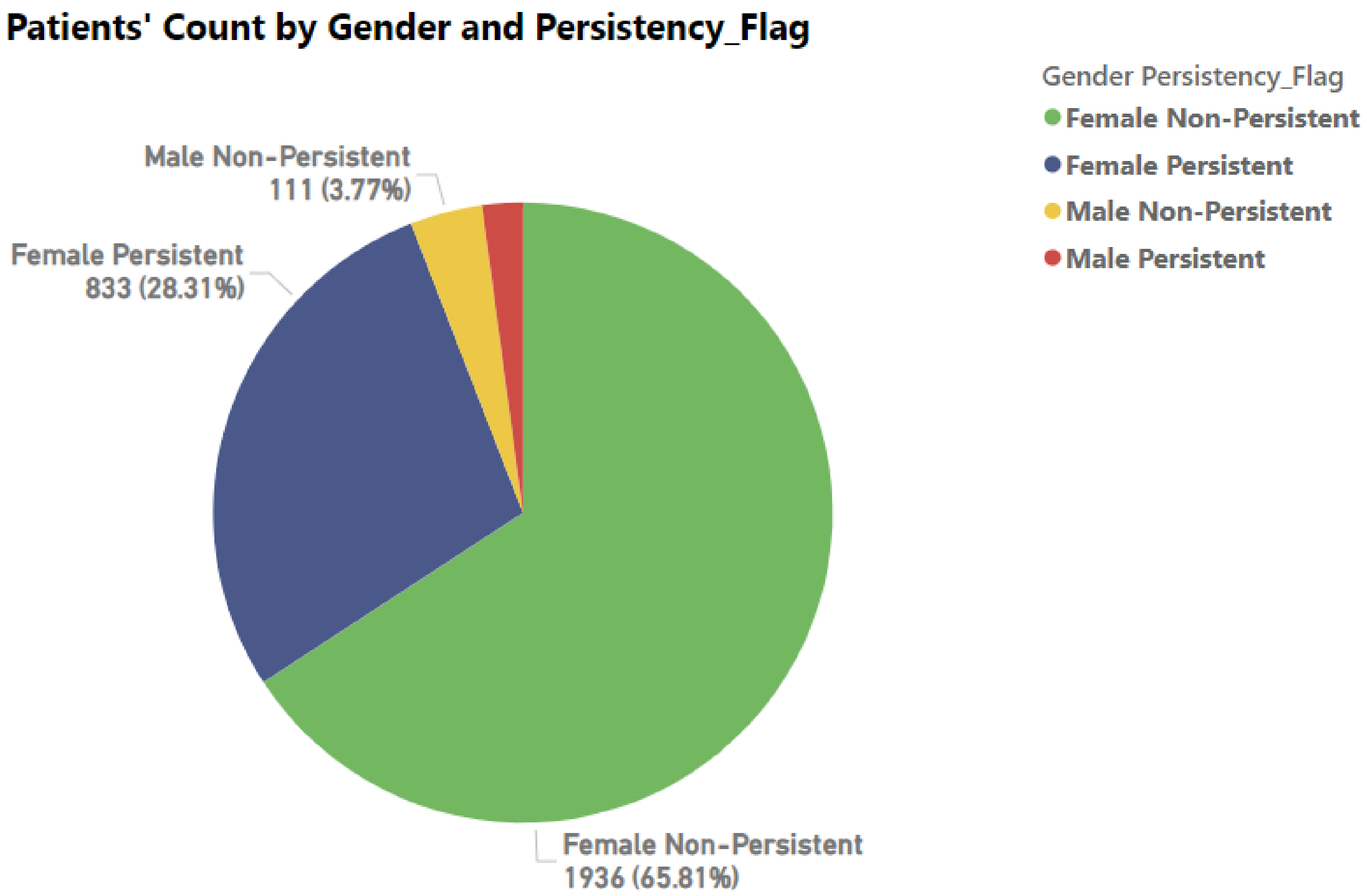
Data Glacier
Your Deep Learning Partner

Target Variable Analysis



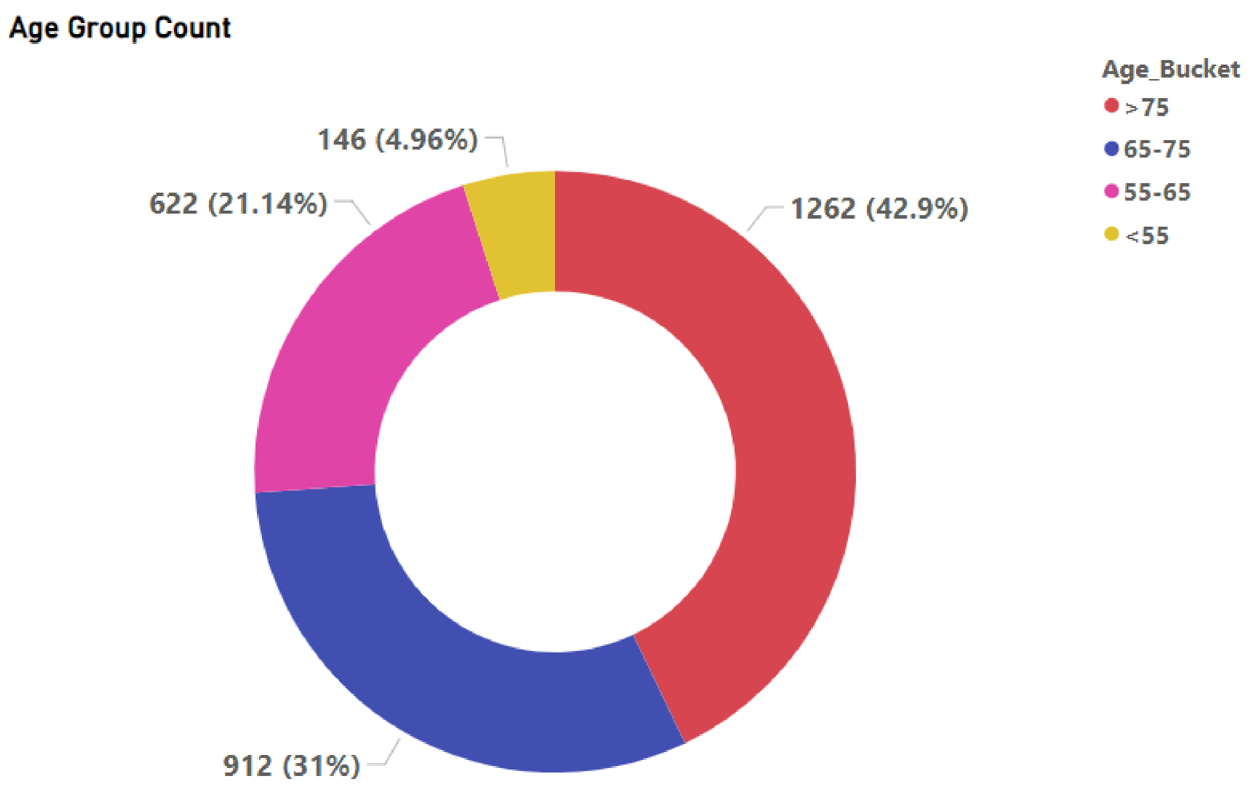
Data Glacier

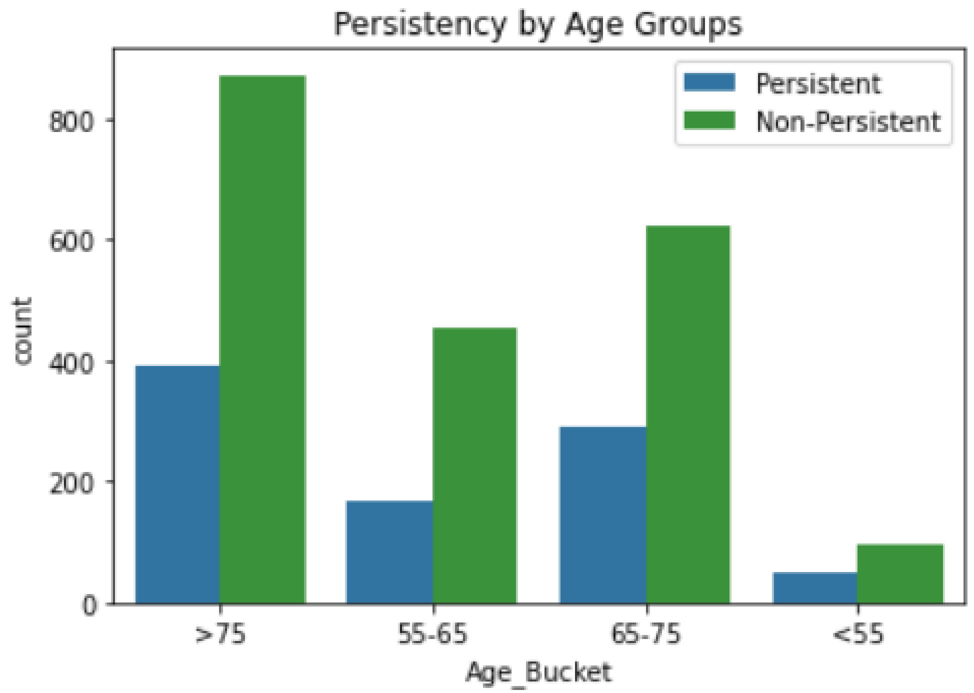
Gender Analysis



Data Glacier

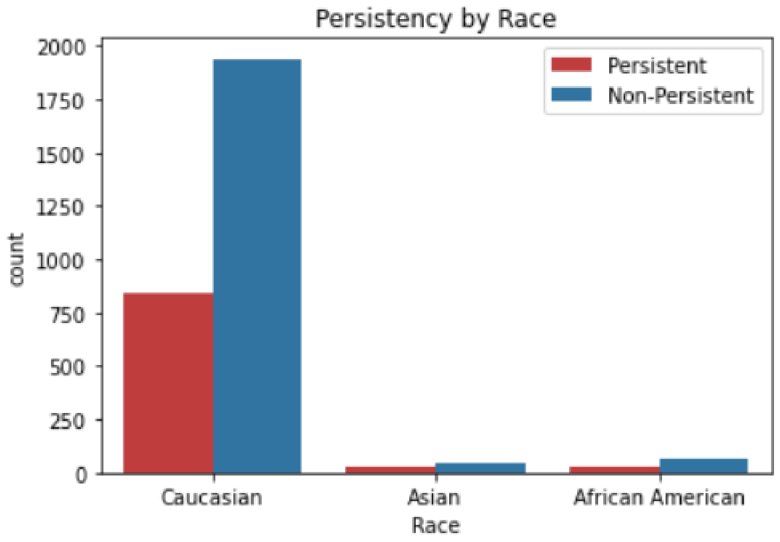
Age Group Analysis

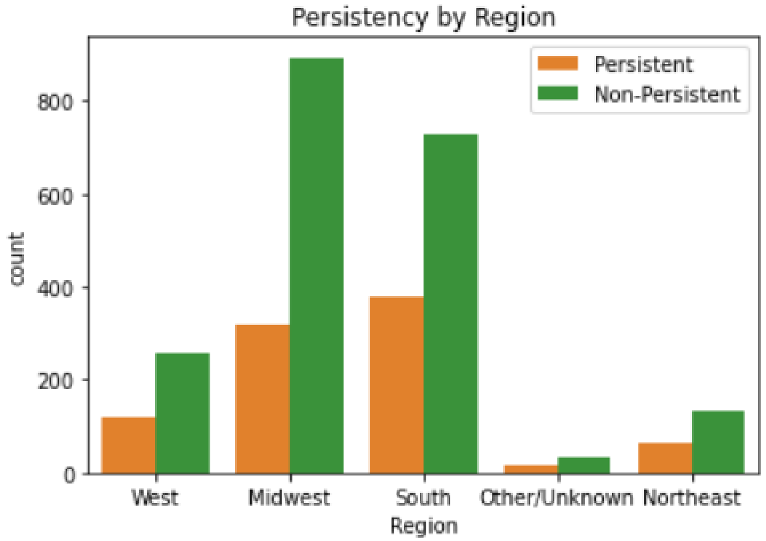


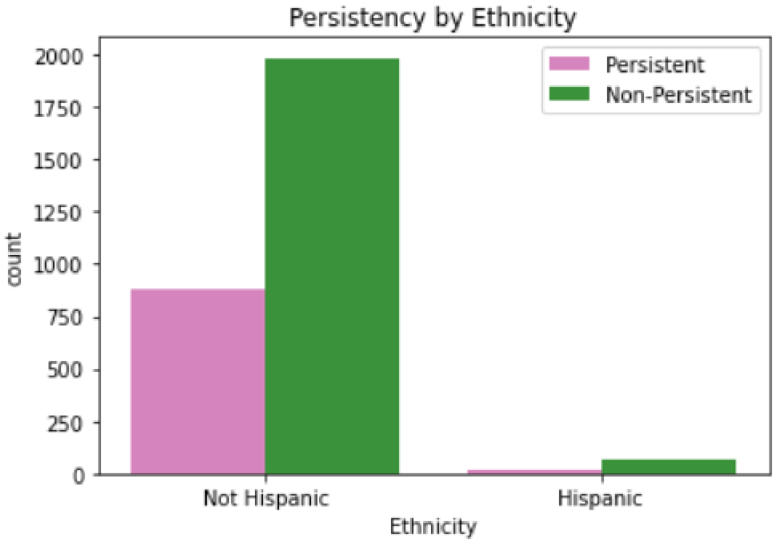


Data Glacier

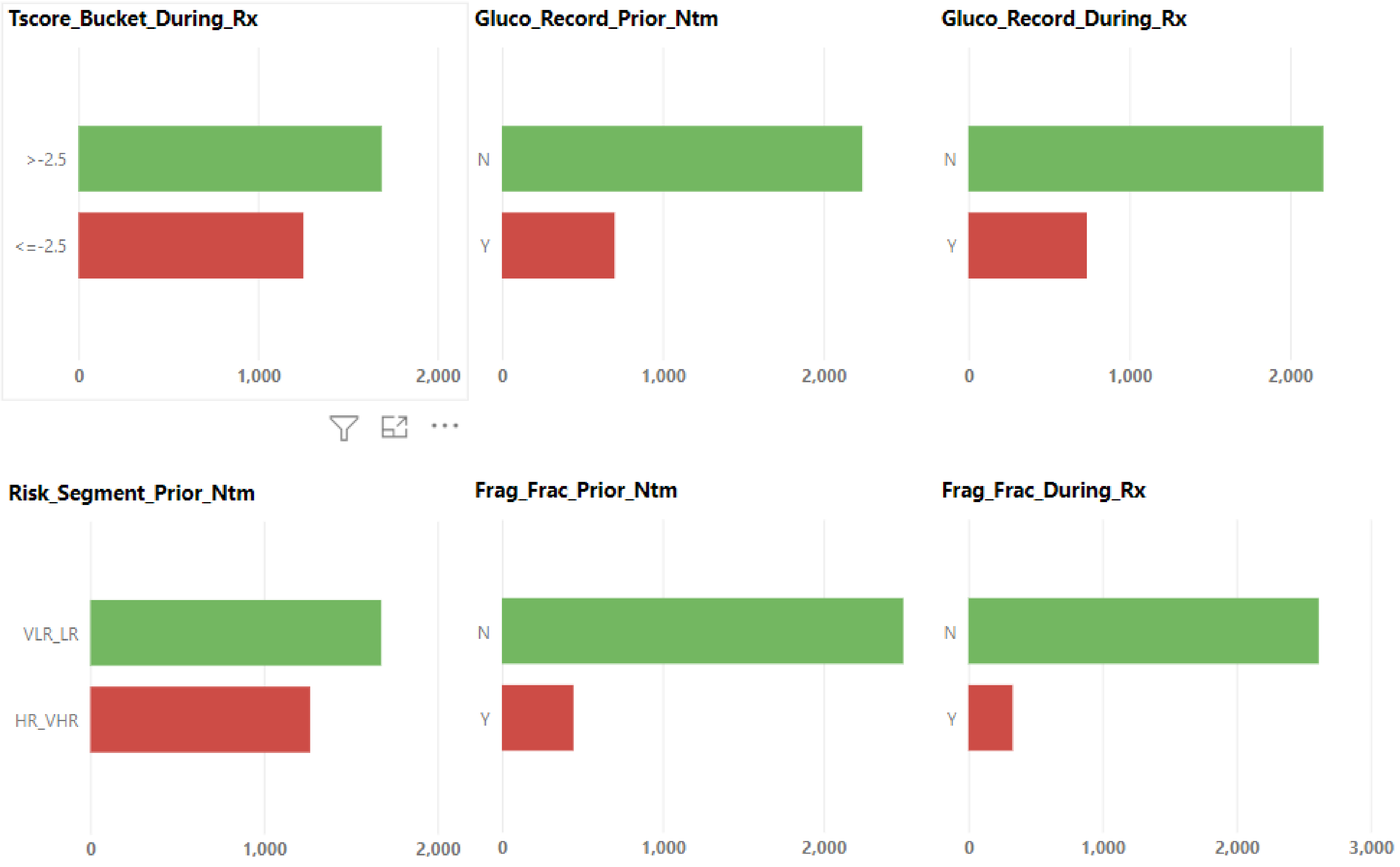
Race, Region & Ethnicity





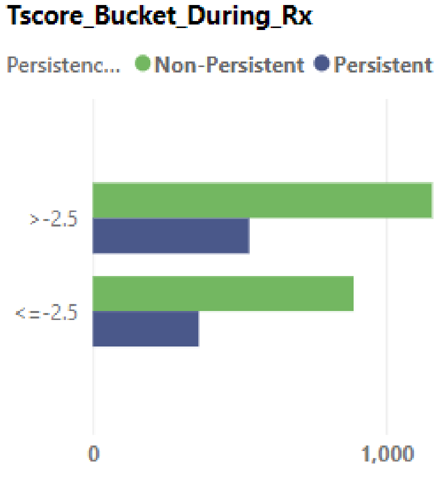


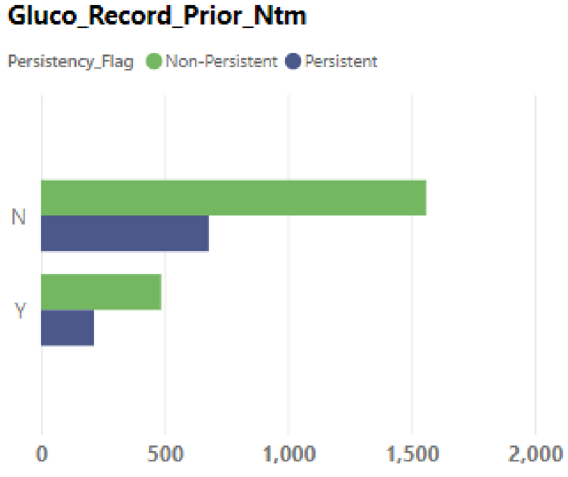
Clinical Factors

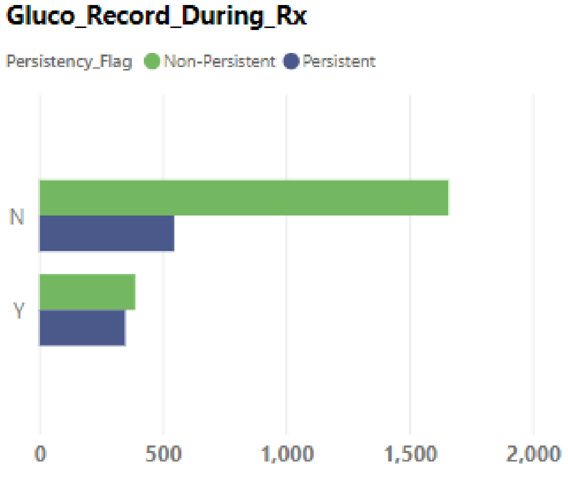


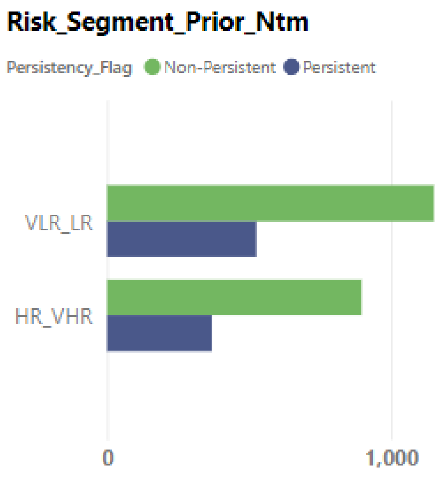
Data Glacier

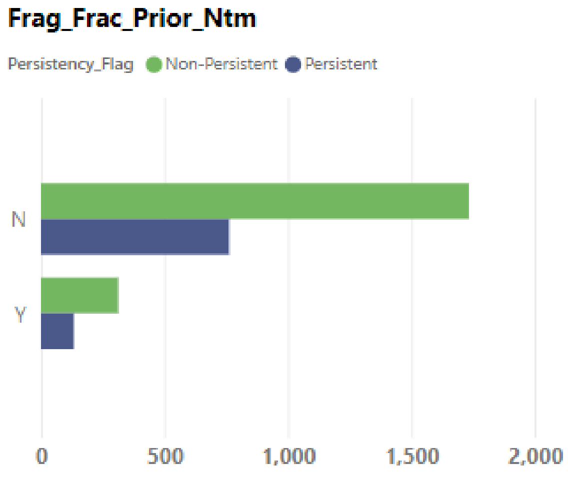
Clinical Factors - Persistency

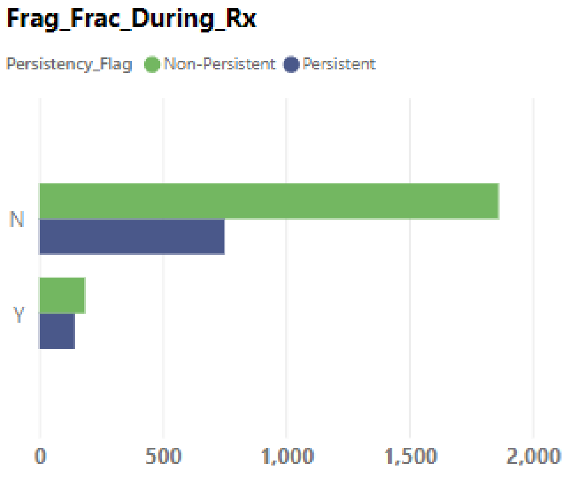












Data Glacier

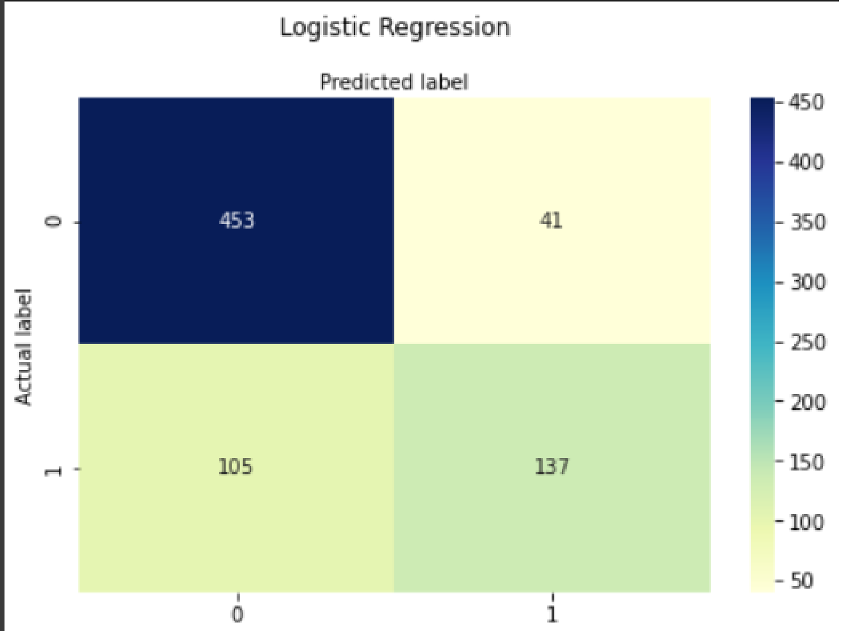
Disease/Treatment Factors

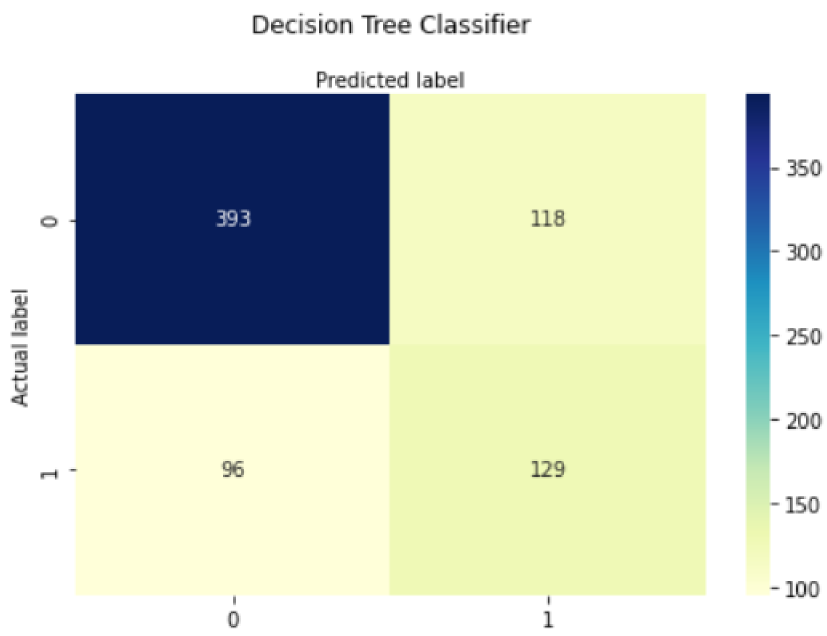
- NTM-Comorbidity: Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias has highest influence.
- NTM-Risk Factors: Risk_Vitamin_D_Insufficiency has highest influence.
- NTM-Concomitancy: Concom_Narcotics has highest influence.

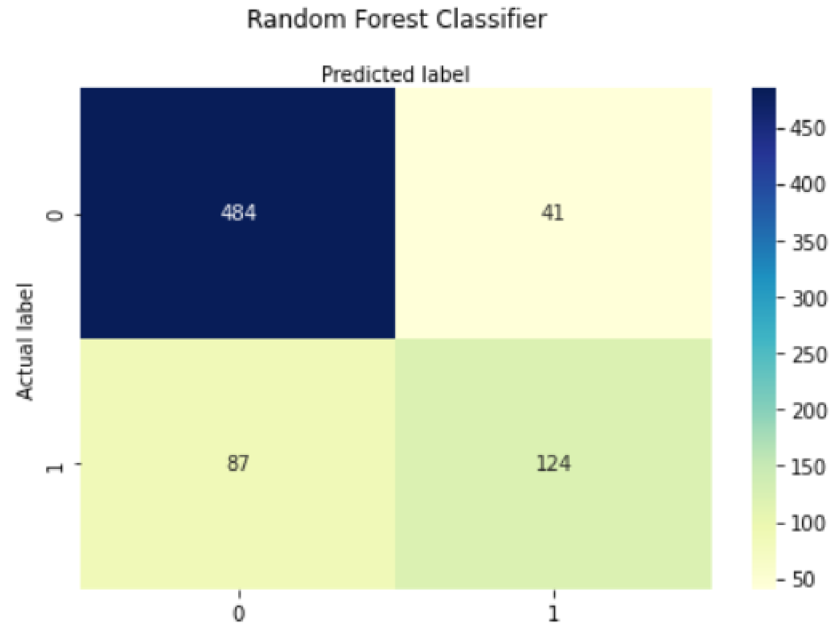
EDA Summary

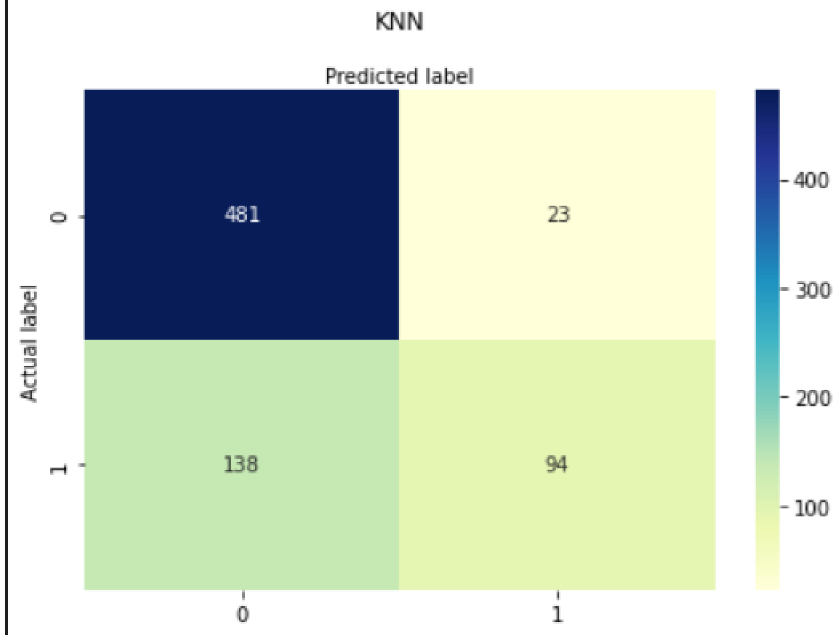
- Drug Persistency is affected by many factors
- Patients older than 65 years show more persistency
- T-square also has influence in persistency
- Disease/Treatment factors:
  - Comorbidity
  - Risk factors
  - Concomitancy

Models


Logistic Regression
Predicted label
Actual label
453
41
105
137


Decision Tree Classifier
Predicted label
Actual label
393
118
96
129


Random Forest Classifier
Predicted label
Actual label
484
41
87
124


KNN
Predicted label
Actual label
481
23
138
94

Models

Model	Accuracy Score
1. Logistic Regression	80.16%
2. Decision Tree Classifier	70.92%
3. Random Forest Classifier	82.61%
4. k-Nearest Neighbors (kNN)	78.13%

Classification Reports

1. Logistic Regression

	precision	recall	f1-score	support
Non-Persistent	0.81	0.92	0.86	494
Persistent	0.77	0.57	0.65	242
accuracy			0.80	736
macro avg	0.79	0.74	0.76	736
weighted avg	0.80	0.80	0.79	736

2. Decision Tree Classifier

	precision	recall	f1-score	support
Non-Persistent	0.80	0.77	0.79	511
Persistent	0.52	0.57	0.55	225
accuracy			0.71	736
macro avg	0.66	0.67	0.67	736
weighted avg	0.72	0.71	0.71	736

3. Random Forest Classifier

	precision	recall	f1-score	support
Non-Persistent	0.85	0.92	0.88	525
Persistent	0.75	0.59	0.66	211
accuracy			0.83	736
macro avg	0.80	0.75	0.77	736
weighted avg	0.82	0.83	0.82	736

4. KNN

	precision	recall	f1-score	support
Non-Persistent	0.78	0.95	0.86	504
Persistent	0.80	0.41	0.54	232
accuracy			0.78	736
macro avg	0.79	0.68	0.70	736
weighted avg	0.79	0.78	0.76	736

Recommendations

- Classifier algorithms tend to produce more effective results
- These models help to categorize patients based on certain factors
- Random Forest Classifier produced most accurate results
- Factors that affect comorbidity, risks and concomitancy were also found out

Thank You