Data Glacier

Your Deep Learning Partner

Healthcare - Persistency of a Drug

Name: Amrin Shaikh
Email: amrin02@gmail.com
Country: United Kingdom
Specialization: Data Science

16-Dec-2024

Problem Description

Problem Overview:

Medication adherence is crucial in ensuring the effectiveness of drug treatment and preventing unnecessary hospitalizations. This project aims to automate the identification of drug persistency, which will help healthcare providers understand patient adherence patterns.

Objective:

- Develop a classification model to predict medication persistency.
- Provide insights into factors influencing patient adherence, such as demographics, risk indicators, and clinical features.
- Improve patient adherence through the identification of critical influencing factors.

Data Overview

Dataset Summary:

- **Size:** 3424 rows × 69 columns
- **Data Type:** Mix of numerical and categorical features
  - **Numerical Columns:** Dexa_Freq_During_Rx, Count_Of_Risks
  - **Categorical Columns:** Gender, Race, Ethnicity, Age_Bucket, Region, and binary risk/comorbidity indicators

Key Columns:

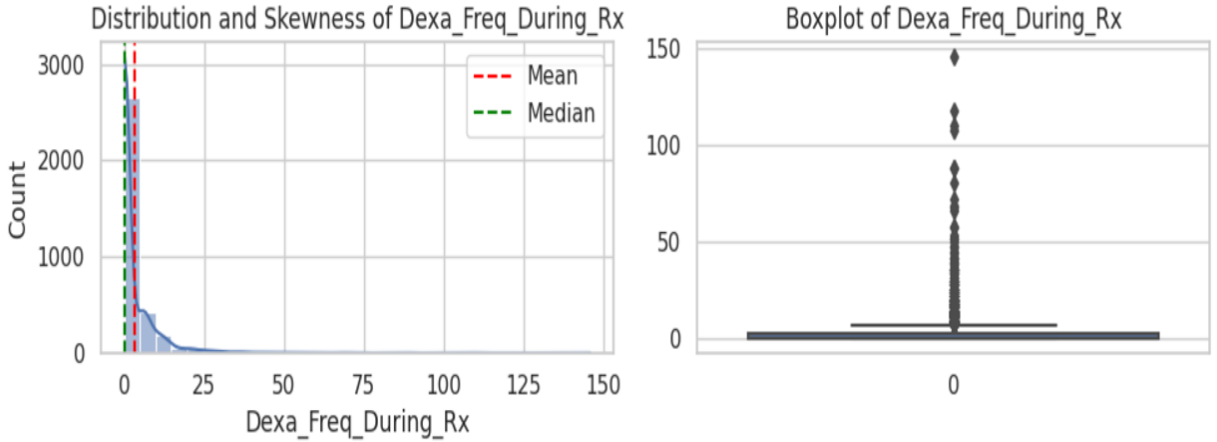
- **Persistency_Flag:** Target variable indicating patient adherence (Yes/No)
- **Dexa_Freq_During_Rx:** Frequency of DEXA scans during treatment
- **Count_Of_Risks:** Number of associated health risks
- **Demographics:** Age_Bucket, Gender, Race, Ethnicity, Region
- **Risk and Comorbidity Indicators:** Various binary health conditions

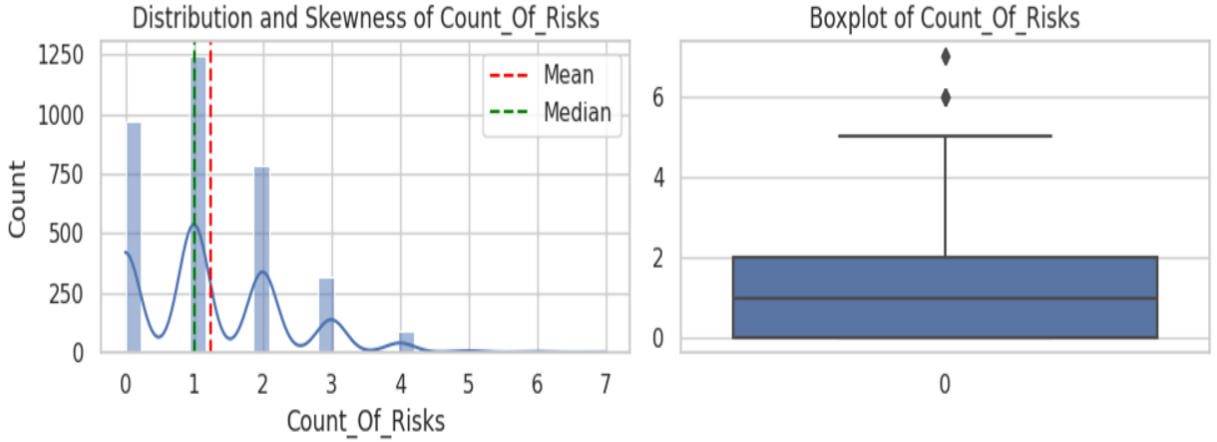
Data Cleaning & Preprocessing

- **Null Values**: No missing data, reducing preprocessing complexity.

Data Cleaning & Preprocessing

Outlier Detection:





	attributes	min	max	range	mean	median	std	skew	kurtosis
0	Dexa_Freq_During_Rx	0	146	146	3.016063	0.0	8.136545	6.805747	74.647502
1	Count_Of_Risks	0	7	7	1.239486	1.0	1.094914	0.879405	0.897420

1. IQR (Interquartile Range) Method:

- **Dexa_Freq_During_Rx:** 460 outliers identified.
- **Count_Of_Risks**: 8 outliers identified.
- Outliers detected using the IQR method were extreme data points falling outside the range of Q1 - 1.5IQR or Q3 + 1.5IQR.

2. Isolation Forest Method:

- Total Outliers: 172 records identified as outliers across all numerical columns.
- This method flagged fewer outliers compared to the IQR method for Dexa_Freq_During_Rx, which had 460 outliers.

3. Descriptive Statistics:

- **Dexa_Freq_During_Rx:** High skewness (6.81) with values ranging from 0 to 146.
- **Count_Of_Risks:** Mild skewness (0.88) with values ranging from 0 to 7.

Data Cleaning & Preprocessing

Handling Outlier & Skewness:

Handling Outliers:

- **Z-score Method:** Identifies data points far from the mean (Z > 3 or Z < -3) and marks them as outliers.
- **IQR Method:** Data points outside the IQR thresholds (Q1 - 1.5IQR or Q3 + 1.5IQR) were treated as outliers.

Handling Approach:

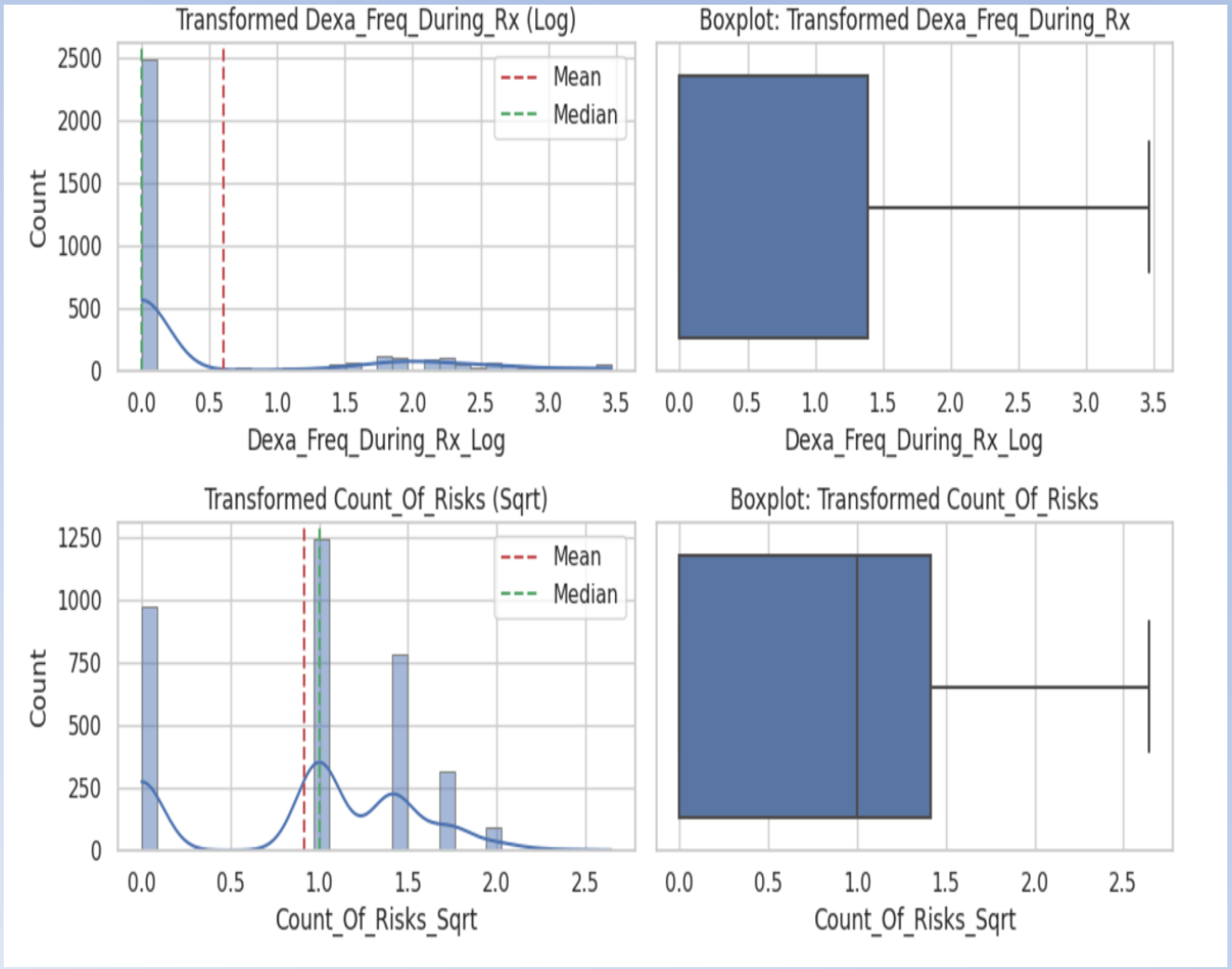
- For critical features, outliers were capped at specific thresholds.
- For non-critical features, outliers were removed to reduce distortion in the model.

Handling Skewness:

- **Log Transformation:** Applied to Dexa_Freq_During_Rx to reduce positive skewness.
- Resulted in a decrease in skewness from 6.81 to 1.33 and removed 42 outliers.

Square Root Transformation:

- Applied to Count_Of_Risks to reduce mild skewness.
- Resulted in a skewness of -0.33, indicating a more normal distribution with no extreme outliers.



Normalizing / Transforming Data

1. Transformations Applied:

Log Transformation:

- Applied to Dexa_Freq_During_Rx to reduce extreme positive skewness.
- Helped normalize the data, but skewness remained moderate after transformation.

Square Root Transformation:

- Applied to Count_Of_Risks to reduce mild skewness.
- Resulted in a more symmetric distribution with reduced skewness.

Box-Cox Transformation:

- Applied to Dexa_Freq_During_Rx for further reduction of skewness.
- Skewness After Box-Cox: 1.02 (still moderate).
- Box-Cox was chosen for its ability to handle skewed data when other transformations were not enough.

2. Skewness Assessment:

Threshold for Symmetry:

- Skewness between -0.5 and 0.5: considered approximately symmetric.
- Skewness of 1.02 indicates moderate skewness, which may be tolerable for tree-based models like Random Forest or Gradient Boosting.
- For models sensitive to skewed data (e.g., Linear Regression or SVM), further transformation may be needed.

3. Further Transformations

Winsorization:

- Purpose: Caps outliers contributing to skewness.
- Skewness After Winsorization: 1.22 (slightly higher than Box-Cox).
- Despite multiple transformations, the Dexa_Freq_During_Rx_Log feature still retains moderate skewness (1.02).

4. Model Considerations:

- Since the model is robust to skewed distributions (tree-based models), the untransformed Dexa_Freq_During_Rx will be used for analysis despite moderate skewness.

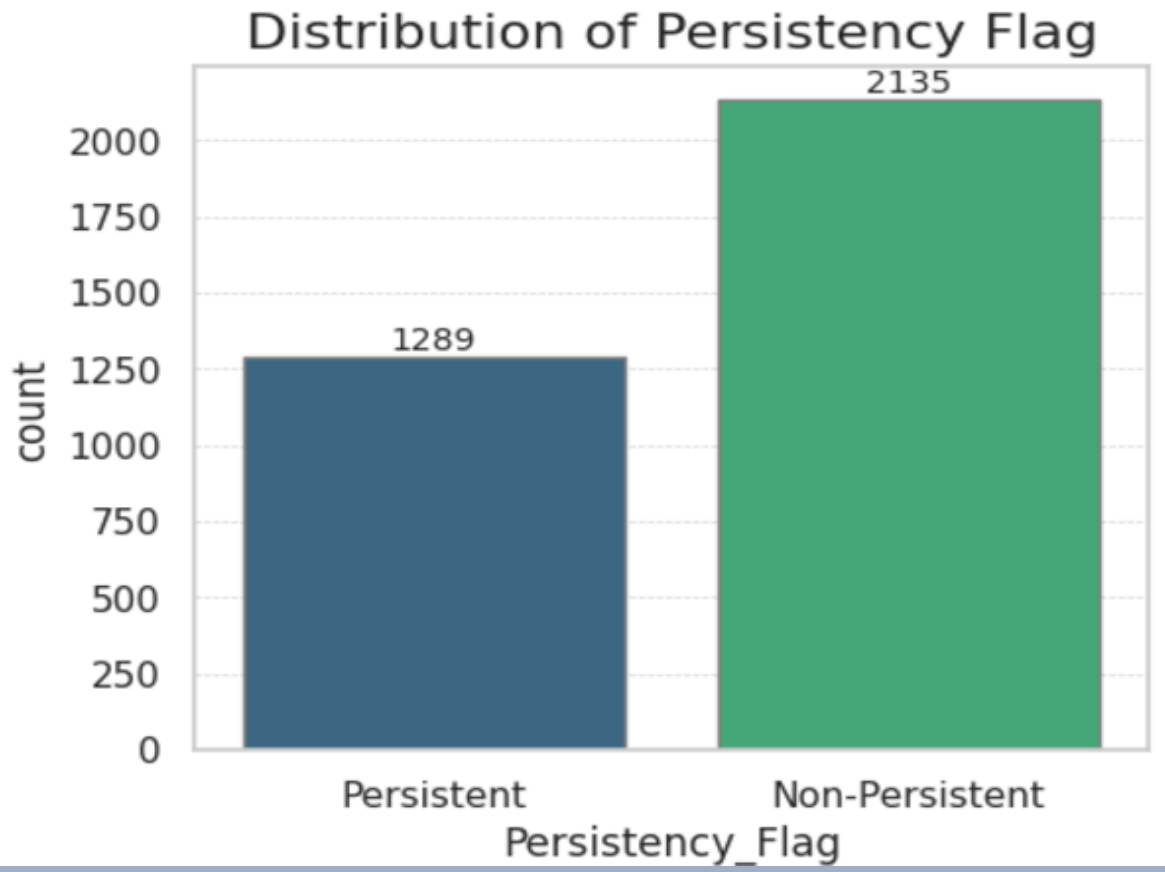
Distribution of Key Features

Distribution of Persistency Flag:

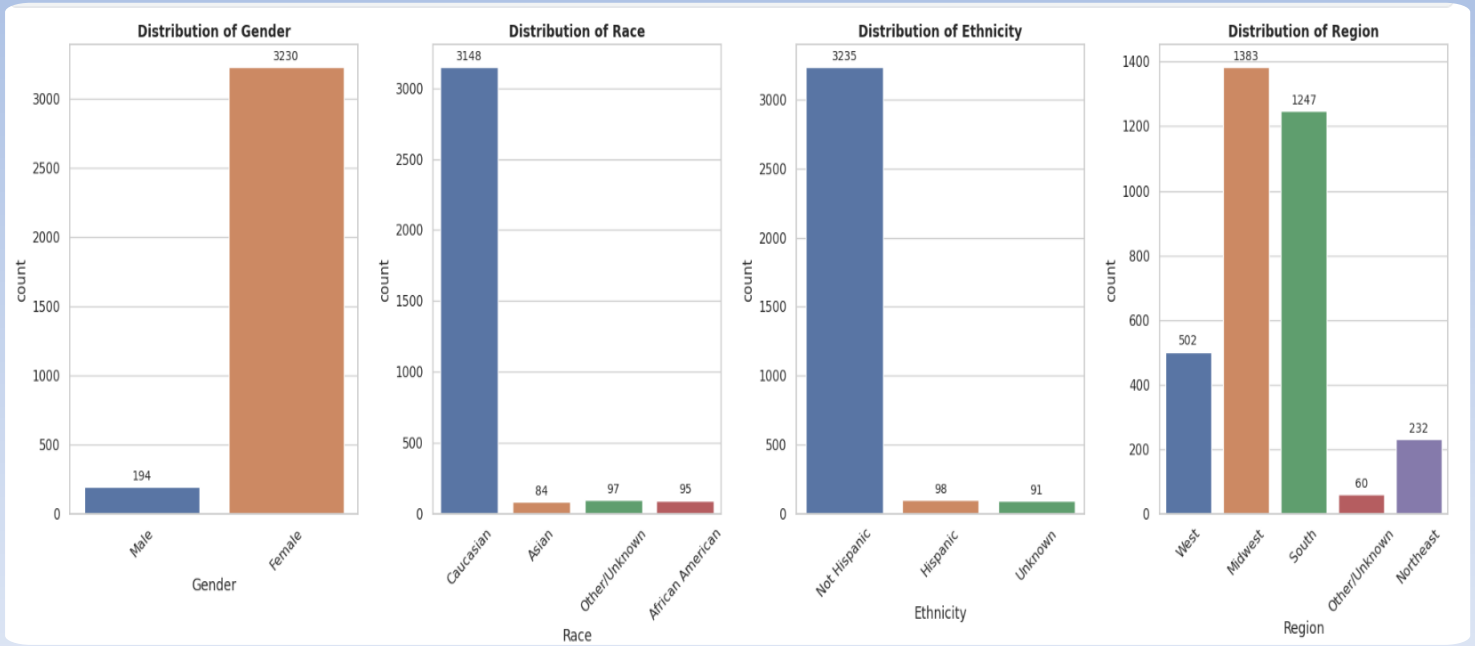
Non-Persistent instances (2,135) outnumber Persistent instances (1,289).

Hypothesis:

The dataset shows a higher prevalence of non-adherence, indicating potential issues in long-term medication adherence or irregularity in persistency.



Distribution of Key Features



The charts show the distributions of Gender, Race, Ethnicity, and Region in the dataset. Key observations are:

- **Gender Distribution:** Predominantly Female (3,230) compared to Male (194), with more Non-Persistent cases across both genders.

- **Hypothesis:** Gender-specific healthcare-seeking behaviour may explain this disparity. Women may face challenges maintaining adherence due to caregiving responsibilities, medication side effects, or systemic healthcare access barriers.

- **Race Distribution:** Caucasians (3,148) represent the majority, followed by smaller groups: Asian (84), Other/Unknown (97), and African American (95). Race distribution is heavily skewed towards Caucasians, which may limit insights into other racial groups.

- **Hypothesis:** Racial disparities in healthcare delivery and medication adherence may influence persistency, with systemic inequities and cultural perceptions impacting adherence rates.

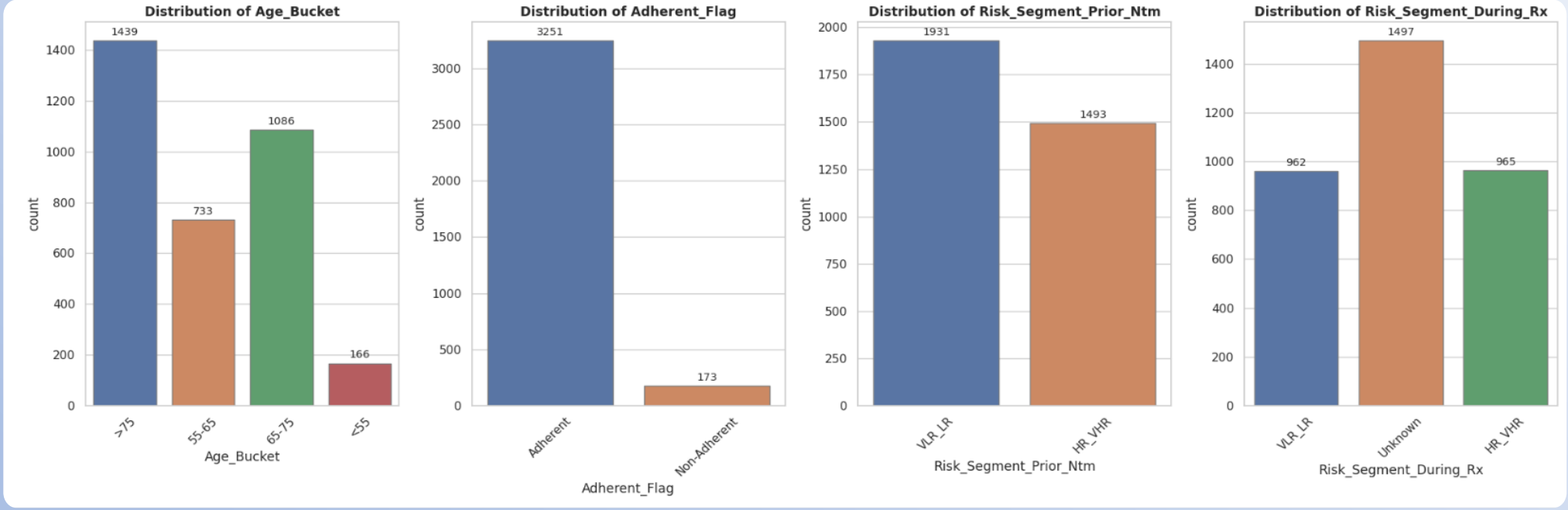
- **Ethnicity Distribution:** Majority are Non-Hispanic (3,235), with few Hispanic (98) and Unknown (91).

- **Hypothesis:** Ethnic factors like cultural, social, or access-related challenges could influence adherence. Non-Hispanic individuals may face barriers to continued adherence due to healthcare accessibility and socioeconomic factors.

- **Region Distribution:** The largest groups are from the Midwest (1383) and South (1247), followed by West (502), Northeast (232), and Other/Unknown (60). Regional trends are more likely to reflect data from the Midwest and South, given their larger representation.

- **Hypothesis:** Regional healthcare access, socioeconomic differences, and regional healthcare policies may affect persistency.

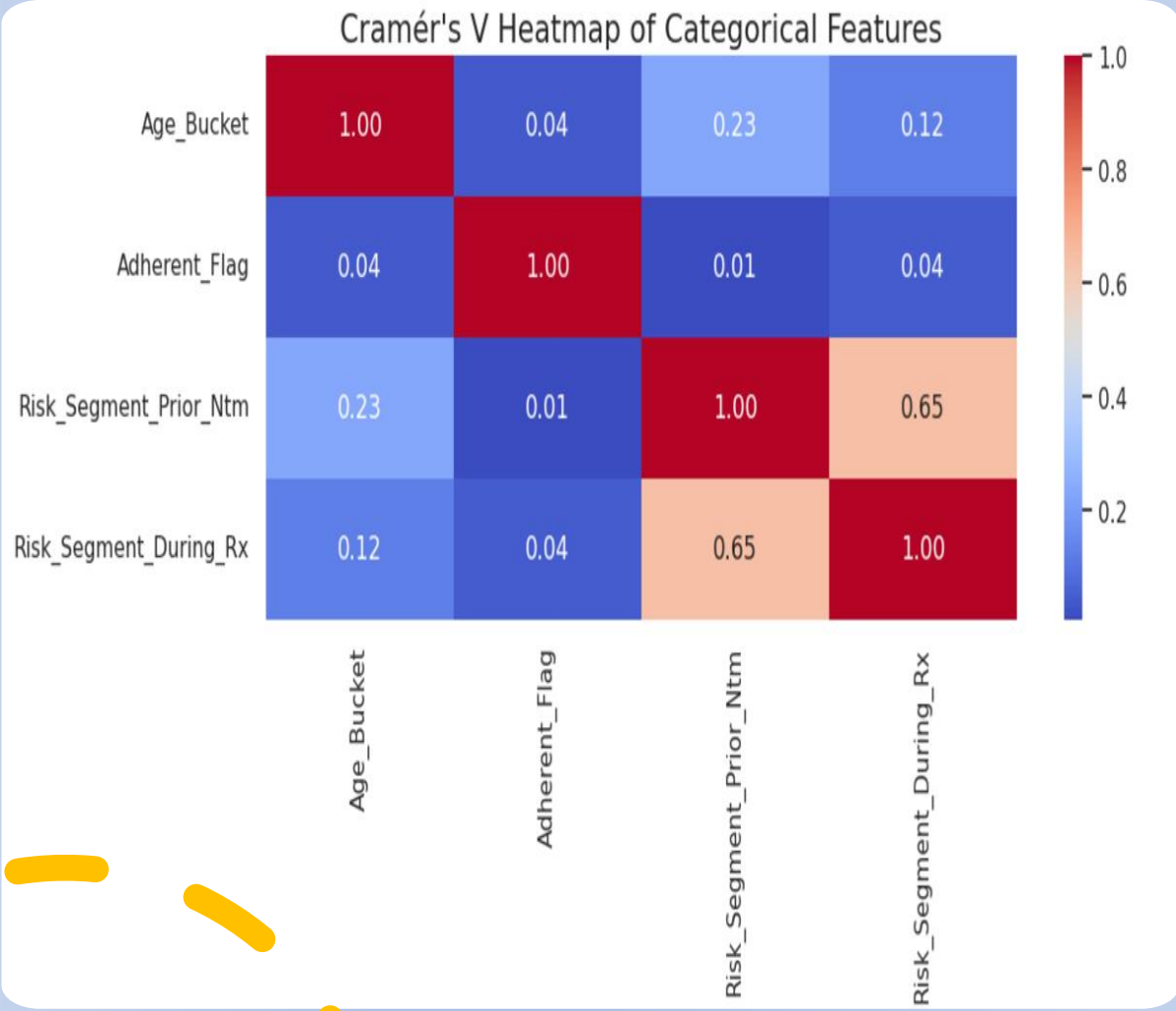
Distribution of Key Features



The charts show the distributions of Gender, Race, Ethnicity, and Region in the dataset. Key observations are:

- **Age Distribution and Adherence:** Most patients are older adults (>75 years), followed by those aged 65-75 years. Very few are under 55.
- **Hypothesis:** Age plays a role in adherence and risk segments, with older adults likely dominating trends, In other hand Adherence is showing high, but specific factors may influence non-adherence
- **Risk Segments (Prior NTM vs. During Rx):** Prior risk segments show a higher number of Very Low Risk - Low Risk (VLR-LR) individuals compared to High Risk - Very High Risk (HR-VHR). During treatment, Unknown risk dominates, with VLR-LR and HR-VHR being nearly equal.
- **Hypothesis:** Prior risk levels reflect successful interventions, but unclear risk categorization during treatment points to potential issues in data quality or risk recording.

Correlation & Relationships



Heatmap of Categorical Features (Cramér's V)

- Visualize the association between categorical variables using Cramér's V to identify strong relationships.

Age Bucket vs. Other Variables:

- **Observation:** The correlation between Age Group and other features like Adherence Flag (0.04) and Risk Segments (0.23 for prior risk, 0.12 for during treatment) is weak.
- **Hypothesis**: Age has minimal influence on adherence and risk levels, suggesting that age-related factors are not strongly driving persistency outcomes in this dataset.

Adherence Flag Correlation:

- **Observation:** The correlation of the Adherence Flag with other variables is very low (max 0.04).
- **Hypothesis:** Adherence appears largely independent of other demographic and risk-related variables, meaning factors influencing adherence might be more complex or individualized.

Risk Segment (Prior NTM vs. During Rx):

- **Observation:** There is a strong relationship between Prior NTM Risk and During Rx Risk (correlation coefficient: 0.65).
- **Hypothesis:** Prior and during-treatment risk segments are closely linked, suggesting that individuals classified in a higher risk group prior to treatment are likely to maintain similar risk profiles during treatment. This trend indicates continuity in risk classification, potentially driven by underlying health factors or treatment effectiveness.

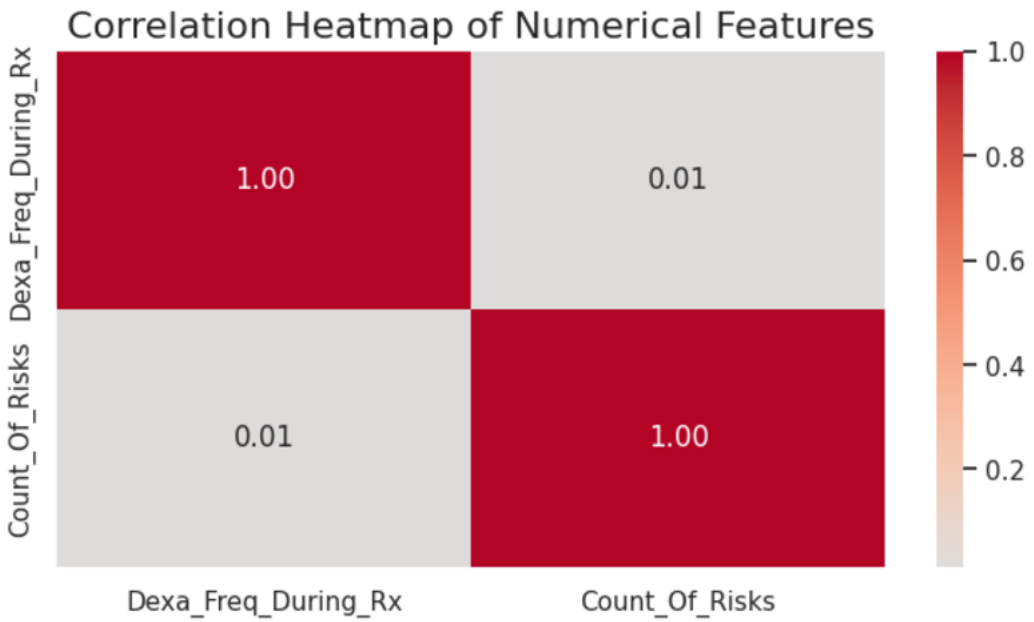
Correlation & Relationships

- **Correlation between "Dexa_Freq_During_Rx" and "Count_Of_Risks“**

The correlation coefficient is 0.01, indicating a very weak relationship between the frequency of DEXA scans during treatment and the count of risks.

- **Hypothesis:**

There is no significant correlation between these two variables. This suggests that changes in the frequency of DEXA scans do not have a substantial impact on the number of risks observed.



Correlation Analysis: Glucose Monitoring & Risk Segments

1. Positive Correlation Between Gluco_Record_Prior_Ntm & Gluco_Record_During_Rx (0.38):

There is a moderate positive correlation between glucose records prior to and during treatment.

Hypothesis: Patients who had glucose monitoring prior to treatment are likely to continue monitoring during treatment.

2. Weak Negative Correlations with Risk Segments:

- **Gluco_Record_During_Rx vs Risk_Segment_During_Rx (-0.11):** Indicates a slight negative correlation—as glucose records increase during treatment, the risk segment decreases slightly.

  Hypothesis: Improved glucose monitoring during treatment may contribute to better risk management (lower risk).

- **Gluco_Record_Prior_Ntm vs Risk_Segment_During_Rx (-0.04):** This weak relationship suggests that prior glucose records have minimal influence on the risk segment during treatment.

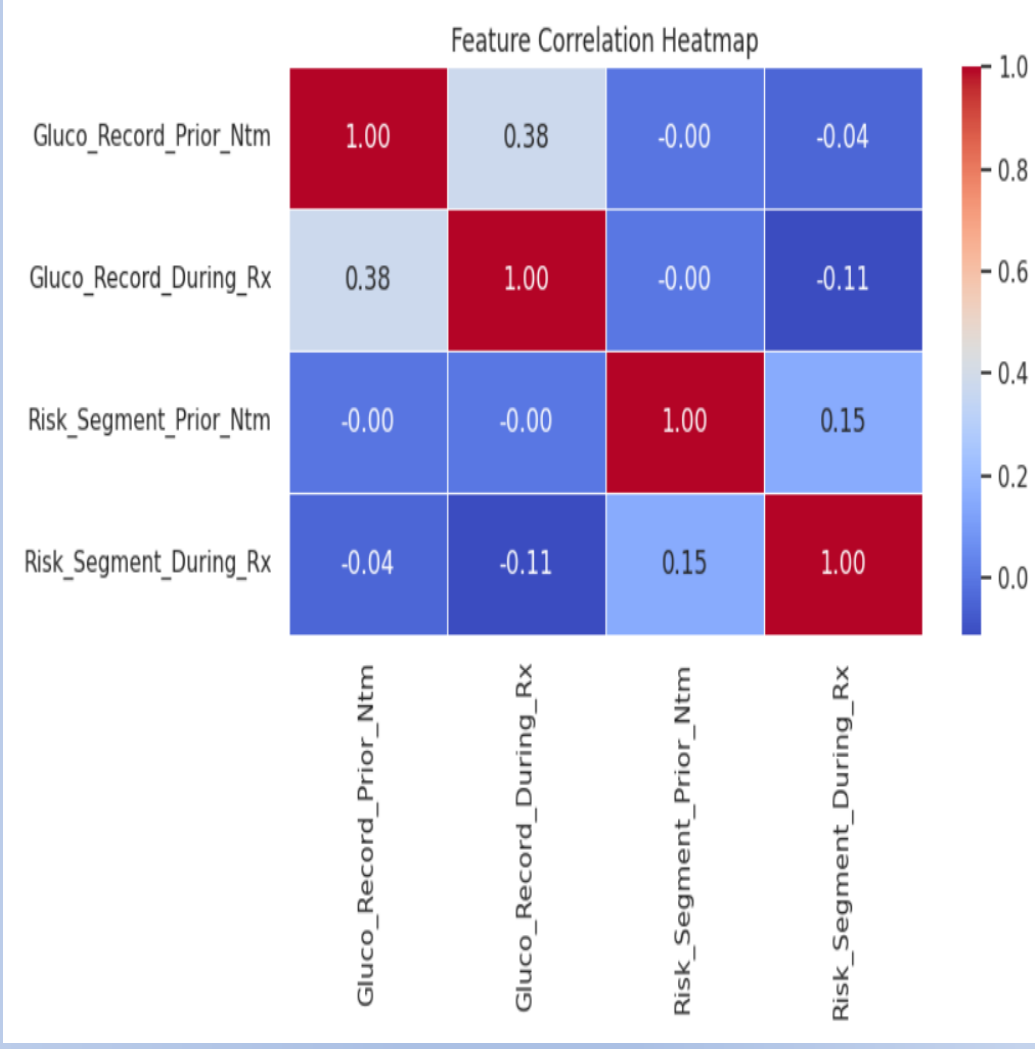
3. Positive Correlation Between Risk_Segment_Prior_Ntm & Risk_Segment_During_Rx (0.15):

There is a small positive correlation between the risk segments prior to and during treatment.

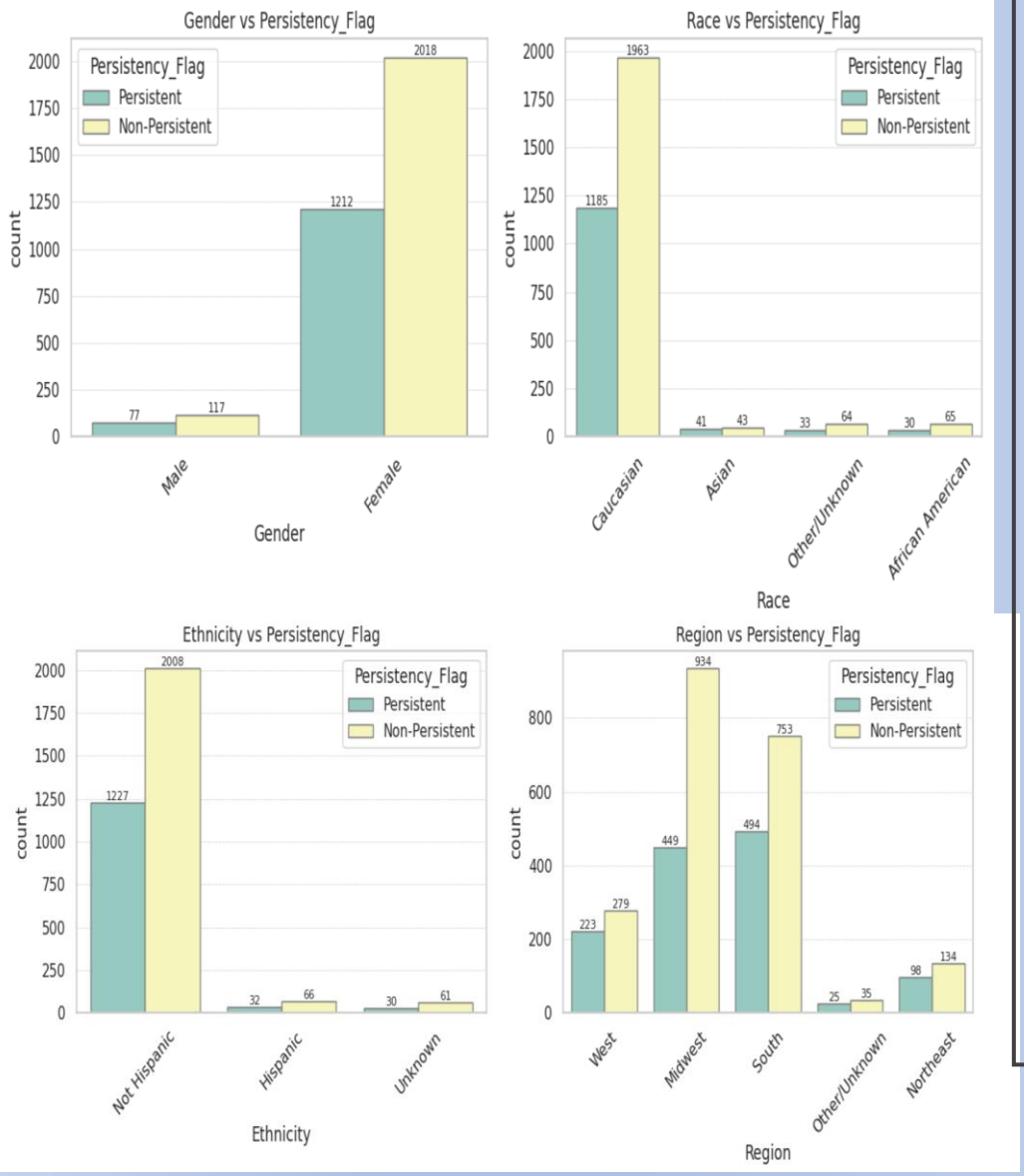
Hypothesis: Patients with a higher risk before treatment may continue to show elevated risk during treatment, indicating some persistence of risk level over time.

4. Near-Zero Correlations:

- The correlations between Gluco_Record_Prior_Ntm and risk segments are almost zero (close to -0.00).
- This indicates no linear relationship between glucose monitoring before treatment and the risk levels.



Demographic Insights on Persistency



1. Gender vs Persistency_Flag

Females dominate the dataset with 1,212 Persistent and 2,018 Non-Persistent cases, while Males show a smaller gap: 77 Persistent and 117 Non-Persistent.

Hypothesis: Gender differences in healthcare-seeking behaviour may explain this disparity. Women are generally more likely to seek healthcare but face challenges in maintaining adherence due to caregiving roles, medication side effects, or systemic barriers in access to care.

2. Race vs Persistency_Flag

The majority of cases belong to the Caucasian group, with 1,185 Persistent and 1,963 Non-Persistent cases. Smaller groups like Asian, African American, and Other/Unknown also show a higher incidence of Non-Persistent cases.

Hypothesis: Racial disparities in healthcare delivery may impact medication persistency. Systemic inequities, cultural perceptions of adherence, and trust in the healthcare system could influence adherence rates across different racial groups.

3. Ethnicity vs Persistency_Flag

Among Non-Hispanic individuals, Non-Persistent cases (2,008) significantly outnumber Persistent cases (1,227). Fewer cases are present in Hispanic and Unknown ethnicity groups, but the trend of Non-Persistent cases dominating remains.

Hypothesis: Ethnicity may affect adherence due to cultural, social, or access-related challenges. Non-Hispanic individuals may initiate treatment but face difficulties in maintaining adherence, possibly due to accessibility or socioeconomic barriers.

4. Region vs Persistency_Flag

In all regions (West, Midwest, South, Northeast, Other/Unknown), Non-Persistent cases dominate. The Midwest and South regions show the highest counts, with a significant gap favouring Non-Persistent cases.

Hypothesis: Geographic location may influence persistency, with disparities in healthcare access, regional healthcare policies, or socioeconomic factors. Rural areas, particularly in the Midwest and South, might face more challenges in ensuring consistent medication adherence.

Key Insights for Medication Adherence

❑ Clinical Features Impacting Adherence

Across multiple comorbidities, Non-Persistent cases dominate, indicating challenges in adherence related to specific clinical conditions.

1. Key Insights Malignant Neoplasms (Screening): Patients undergoing screening for malignant neoplasms show higher adherence (843 Persistent) compared to those without (446 Persistent).

Implication: Regular screenings might encourage better follow-ups and adherence.

2. Immunization Encounters: Patients with immunization encounters demonstrate better adherence (829 Persistent) compared to those without (460 Persistent).

Implication: Preventive care engagements boost persistency through proactive healthcare visits.

3. General Exam Without Complaint: Individuals with general exams exhibit higher persistency (744 Persistent) compared to those without (545 Persistent).

Insight: Routine health check-ups promote adherence to treatment regimens.

4. Vitamin D Deficiency: Persistency is fairly balanced among individuals with deficiency (545 Persistent) versus without (744 Persistent).

Insight: Vitamin deficiencies may not significantly influence adherence but require attention for overall care.

6. Other Joint Disorders: Non-Persistent cases dominate when no joint disorders are reported (1688 cases), but patients with disorders show relatively better persistence (552 Persistent).

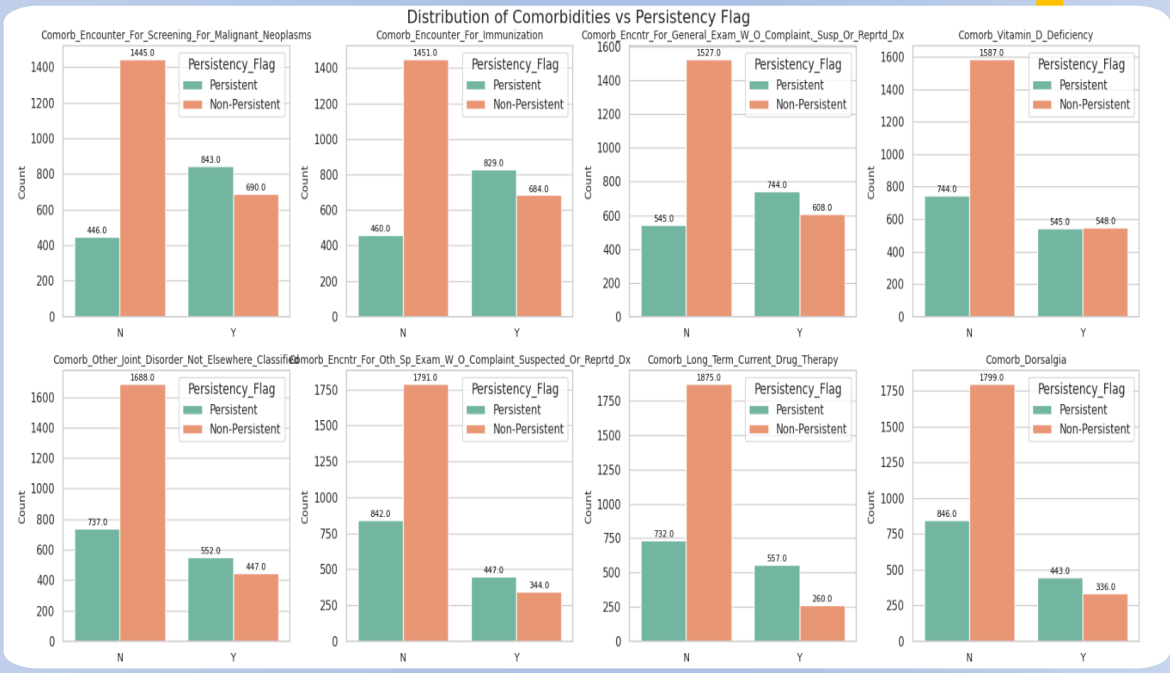
Hypothesis: Pain management or chronic joint conditions might encourage medication adherence.

7. Long-Term Drug Therapy: Adherence drops significantly for those undergoing long-term drug therapy (557 Persistent vs 260 Persistent).

Implication: Longer medication durations correlate with reduced persistency, highlighting the need for interventions to maintain adherence.

8. Dorsalgia (Back Pain): Patients with back pain (Y) display lower adherence (443 Persistent) compared to those without (846 Persistent).

Insight: Chronic pain conditions might deter adherence due to side effects or inadequate pain management



Key Insights for Medication Adherence

❑ Clinical Features Impacting Adherence

Across multiple comorbidities, Non-Persistent cases dominate, indicating challenges in adherence related to specific clinical conditions.

1. Personal History of Other Diseases and Conditions: Patients with a history of other conditions show higher adherence (400 Persistent vs 277 Non-Persistent).

Insight: Increased health awareness drives better adherence.

2. Bone Density Disorders: Patients with bone density disorders have nearly double the persistency (342 Persistent vs 176 Non-Persistent).

Insight: Chronic nature of the condition motivates adherence.

3. Disorders of Lipoprotein Metabolism: Persistence is balanced for these disorders (800 Persistent vs 805 Non-Persistent).

Insight: Regular monitoring may encourage adherence, though results vary.

4.Osteoporosis: Adherence levels are balanced among patients with osteoporosis (448 Persistent vs 469 Non-Persistent).

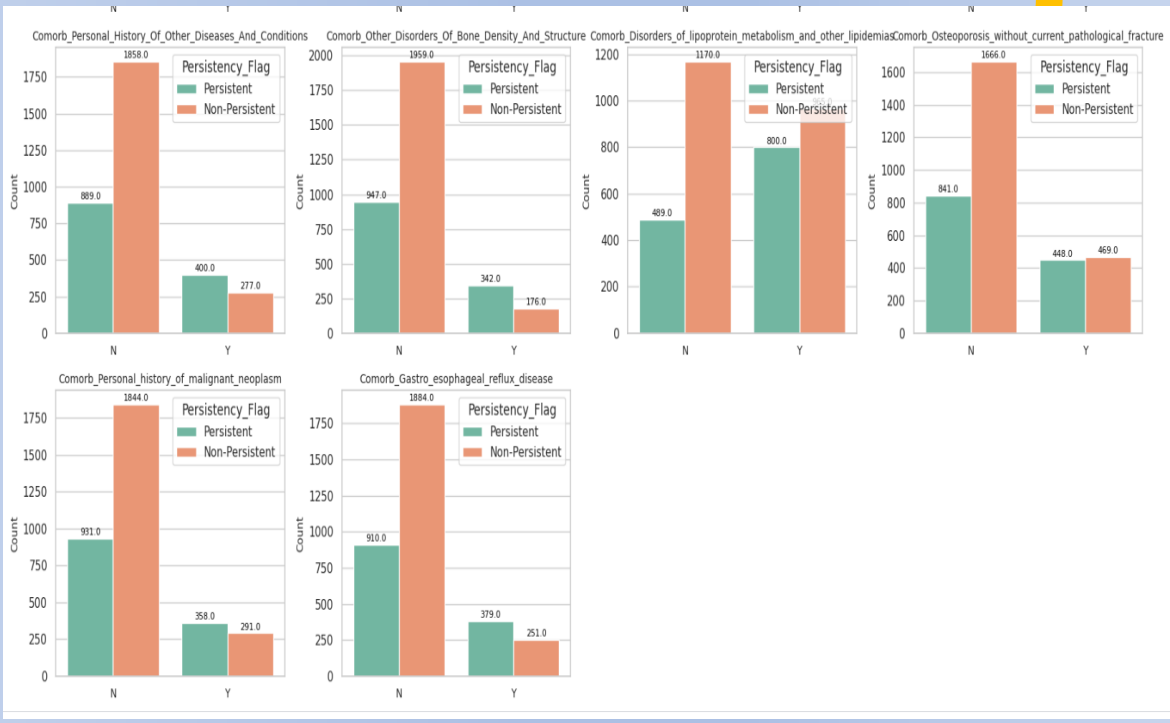
Insight: Perceived disease severity and side effects influence adherence.

5. Malignant Neoplasms: Patients with a history of cancer show improved adherence (358 Persistent vs 291 Non-Persistent).

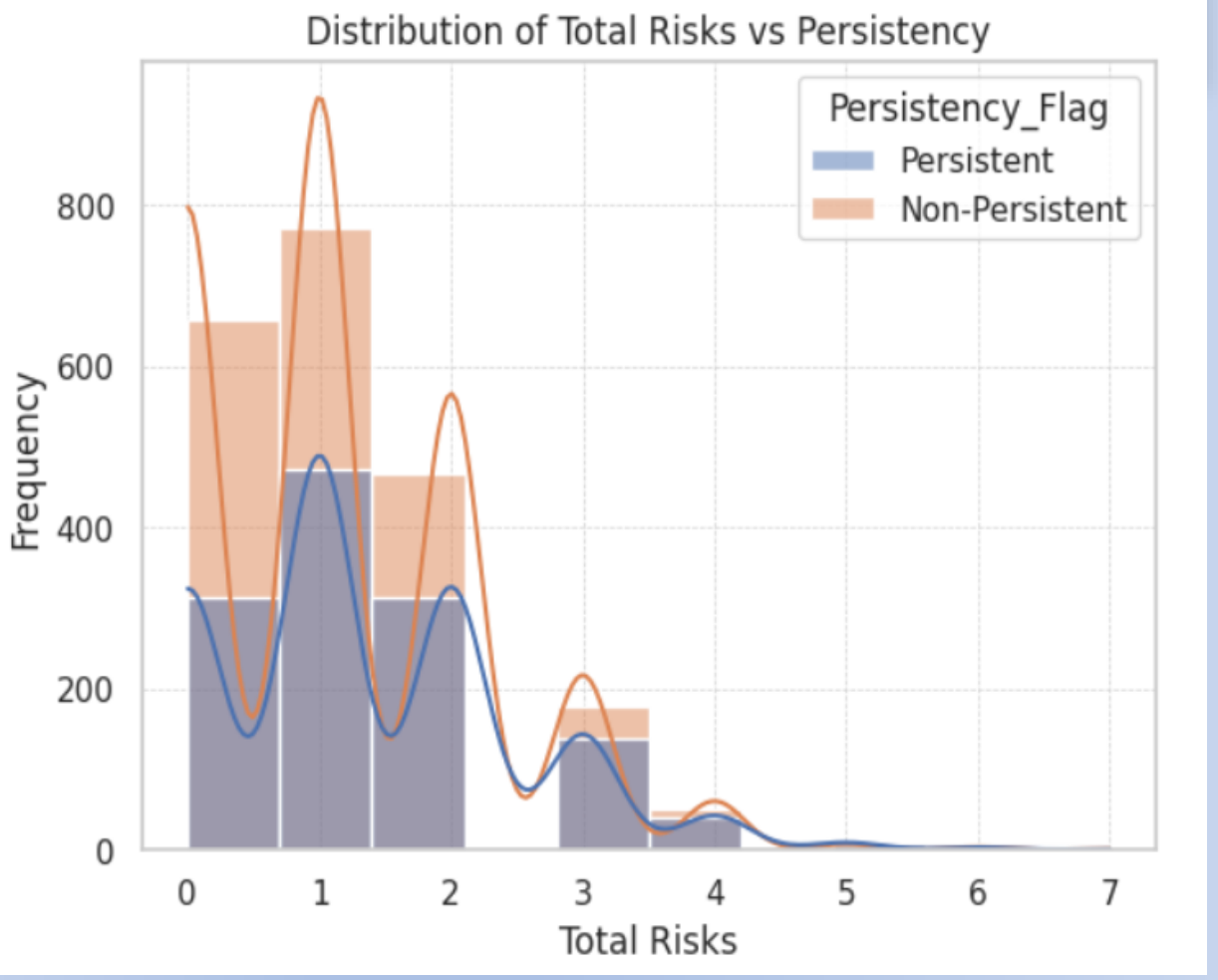
Insight: Higher health awareness in survivors boosts adherence.

6. GERD (Gastroesophageal Reflux Disease): Patients with GERD have better adherence (379 Persistent vs 251 Non-Persistent).

Insight: Continuous symptom management drives persistency.



Total Risks and Medication Adherence



Key Observations:

- Most patients, regardless of persistency, have Total Risks between 0 and 2.
- At lower risk levels (0–1), the frequency of Non-Persistent patients is noticeably higher compared to Persistent patients.
- Persistent patients' distribution decreases sharply after risk level 2, while Non-Persistent patients have a longer tail extending to higher risk levels (3–7).

Hypotheses for Analysis

Low Total Risks (0–1):

- Patients with lower risk levels are more likely to adhere to treatment due to fewer complications or barriers (e.g., side effects, complexity of the treatment regimen).
- However, the higher frequency of Non-Persistent patients at these levels suggests that some other factors, such as behavioral or socioeconomic barriers, might impact adherence.

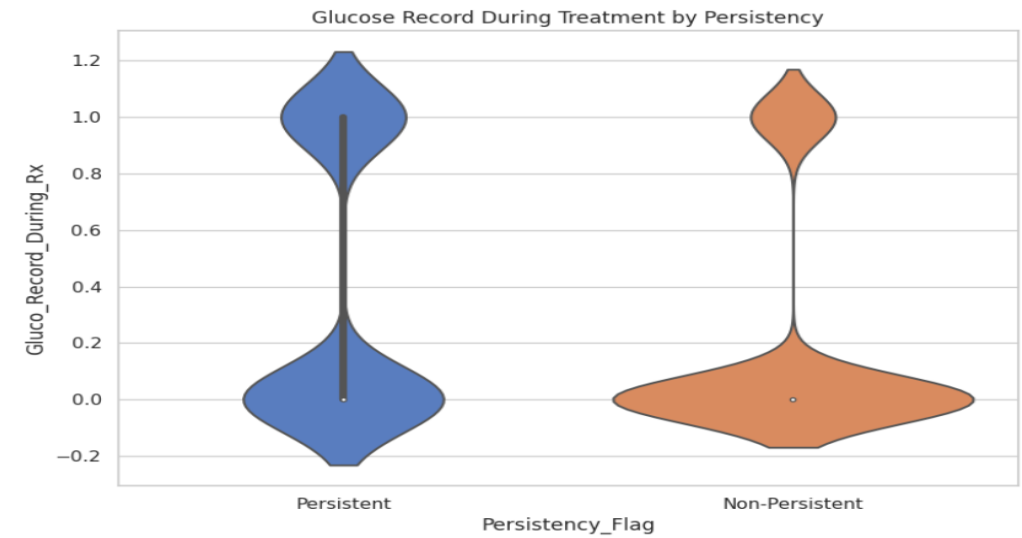
Moderate Total Risks (2–3):

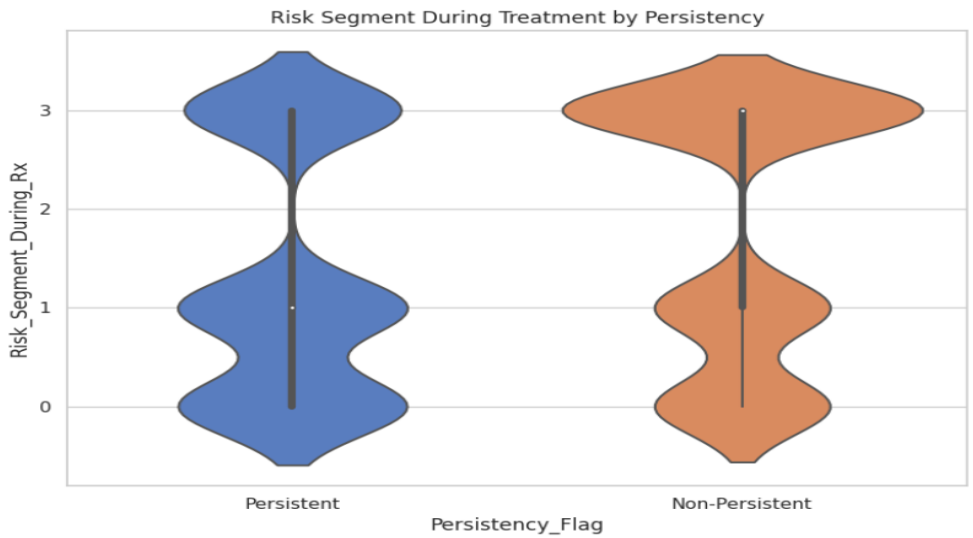
- As risks increase, persistent patients become less frequent, possibly due to increasing treatment complexity, higher side effects, or diminishing confidence in the effectiveness of the treatment.

High Total Risks (4–7):

- Patients with very high risks are predominantly Non-Persistent, which could be due to the overwhelming severity of their conditions, inability to manage complex regimens, or lack of follow-up support.

Glucose Monitoring & Risk Segments





Glucose Record During Treatment by Persistency

- For Persistent patients (left violin plot), glucose records are concentrated at higher values, suggesting better or more frequent glucose monitoring during treatment.
- For Non-Persistent patients (right violin plot), glucose records show a lower concentration overall, indicating less frequent monitoring.
- There is a clear difference in distribution, with Persistent patients exhibiting higher and more diverse glucose records, while Non-Persistent patients have a narrow, low-range distribution.

Hypothesis:

- Patients with better glucose record monitoring during treatment may exhibit better persistency (continuation with treatment). Glucose Monitoring & Risk Segments
- Lack of frequent glucose monitoring might contribute to lower treatment persistency.

Risk Segment During Treatment by Persistency

- For Persistent patients, the risk segment is bimodal (two peaks). The peaks occur at:
  - Low risk (0): A significant proportion of Persistent patients are in the low-risk category.
  - Moderate to higher risk (2–3): Some Persistent patients fall into higher-risk categories as well.
- For Non-Persistent patients, the distribution also shows bimodality, but the higher risk segment (3) appears to dominate. The lower risk group (0) is less concentrated than in Persistent patients.

Hypothesis:

- Patients in lower risk categories during treatment may exhibit better persistency because they might experience fewer health complications or issues.
- Patients in higher risk segments may drop out of treatment (Non-Persistent) due to difficulty managing their condition.

Technical Slide for Feature Selection

Summary:

Feature selection ensures that the most relevant variables are fed into the machine learning models, reducing noise and improving performance.

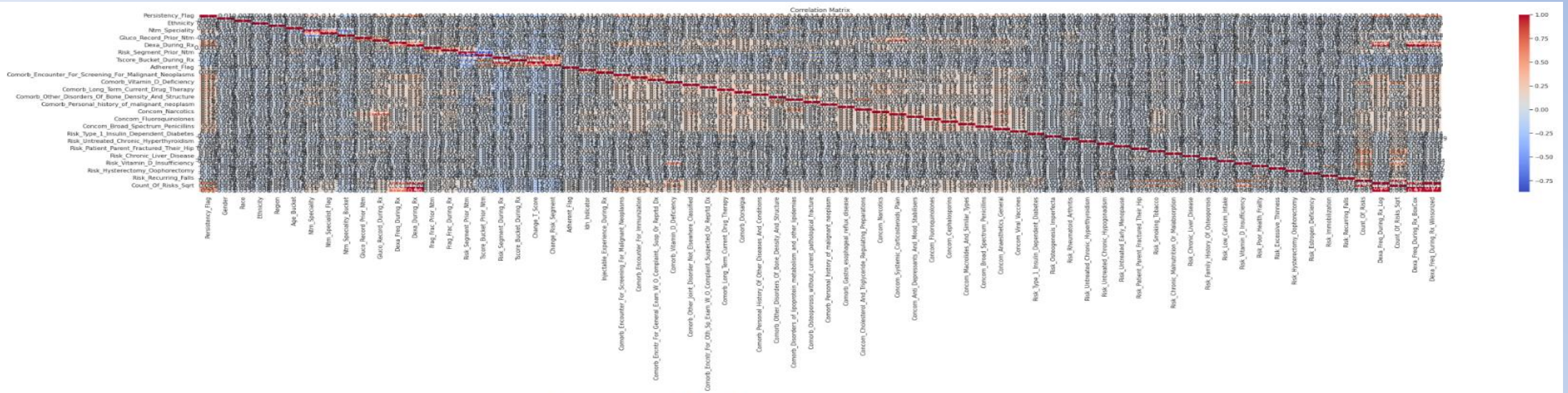
Feature Selection Techniques Used:

- **Correlation Analysis:** Removed highly correlated features (e.g., >0.85 threshold).
- **Mutual Information:** Selected features most informative for predicting "Persistency".
- **Recursive Feature Elimination (RFE):** Selected top features iteratively.
- **Feature Importance**: Leveraged Random Forest feature importance scores.

Technical Slide for Feature Selection

- **Filter Methods:** Filter methods evaluate the statistical relationship between features and the target variable. Features that are weakly correlated or irrelevant are removed before training the model.

- **Common Techniques**
- Correlation Matrix (for numerical features)
- Chi-Square Test (for categorical features)
- Variance Threshold (removes features with low variance)



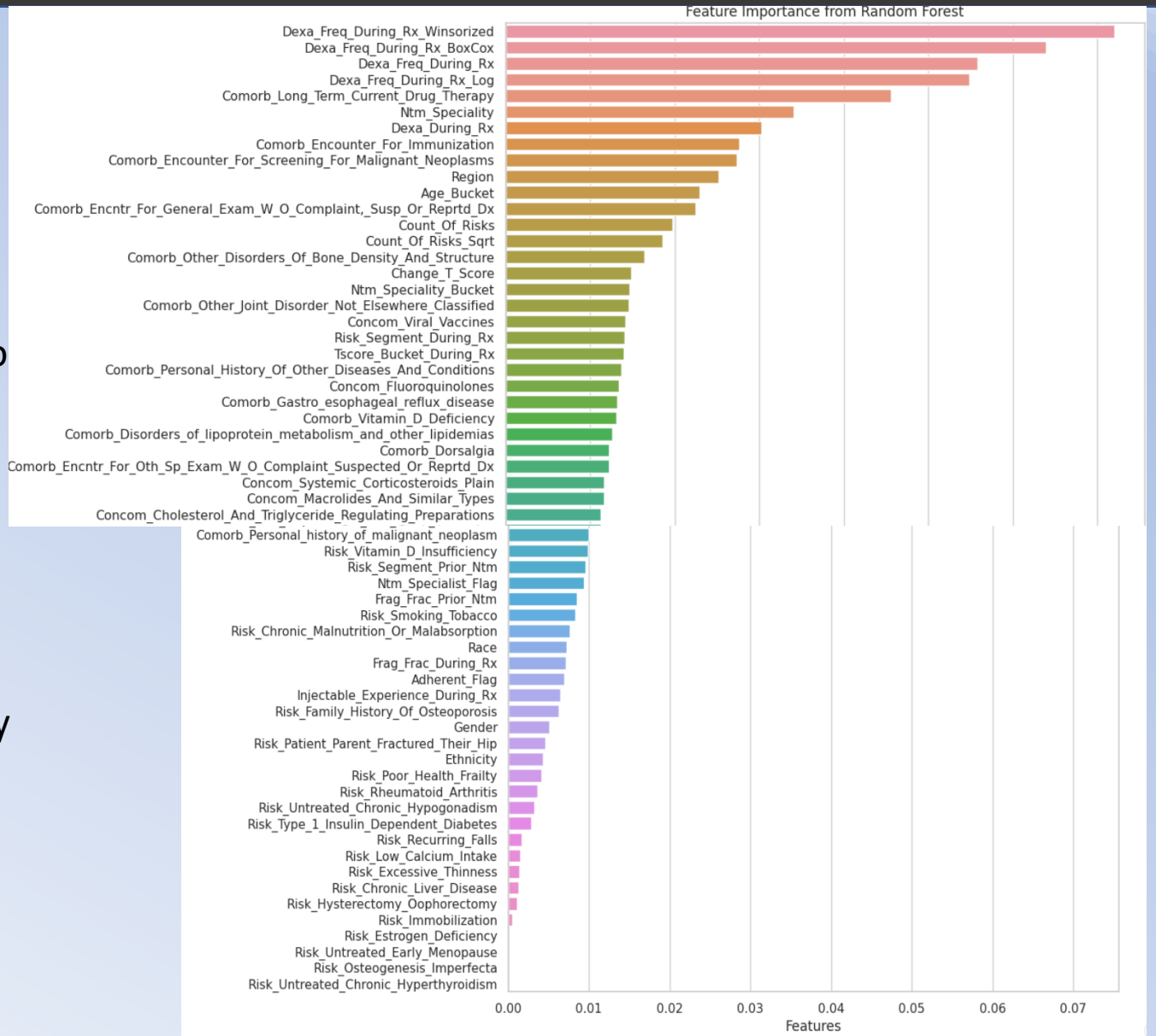

```
Highly correlated features to drop: {'Concom_Systemic_Corticosteroids_Plain', 'Dexa_Freq_During_Rx_BoxCox', 'Change_T_Score', 'Count_Of_Risks_Sqrt', 'Dexa_Freq_During_R
x_Winsorized', 'Dexa_Freq_During_Rx_Log', 'Tscore_Bucket_Prior_Ntm'}
```

Feature Importance from Random Forest

❑ **Embedded Methods:**

- Feature Importance from Tree-Based Models:

- Random Forest assigns importance scores to features during training.
- Random Forest calculates how much each feature contributes to reducing errors (feature importance).
- Visualizing feature importance helps identify the most relevant features.



Model Recommendations for Persistency Prediction

Why Classification Models?

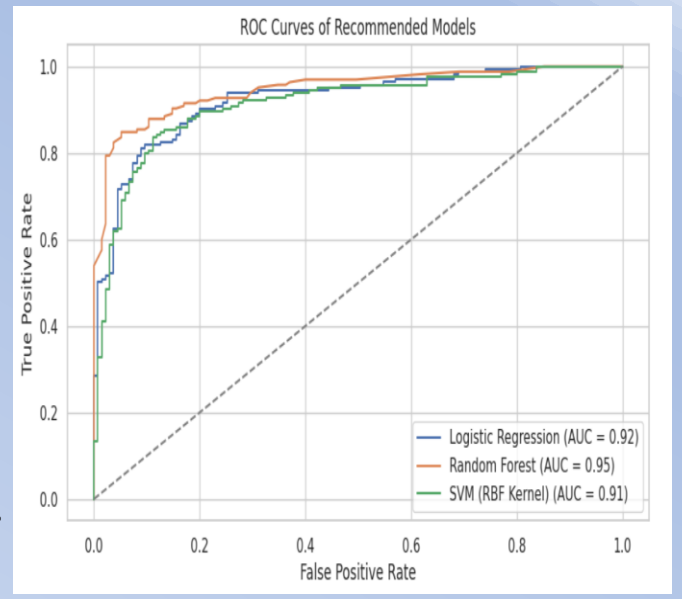
- Persistency Prediction is a binary classification problem: Persistent (1) , Non-Persistent (0)
- A classification model determines the probability of a patient being persistent or non-persistent based on input features (e.g., glucose records, risk segments, demographics, etc.).

Recommended Models:

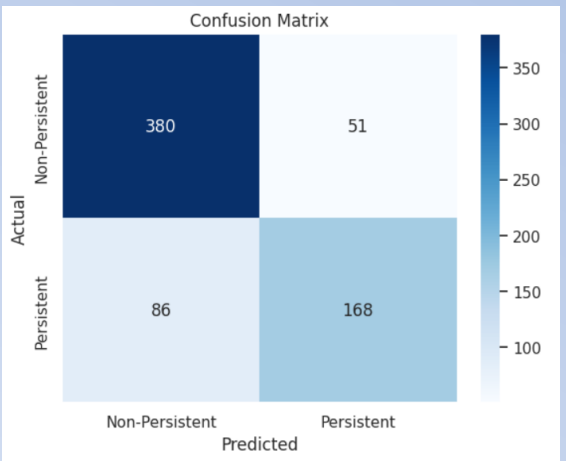
Model	Type	Why It's Recommended
Logistic Regression	Linear Model	Simple, interpretable, and effective.
Random Forest	Ensemble (Bagging)	Handles non-linearity and feature ranking.
Gradient Boosting (XGBoost, LightGBM)	Ensemble (Boosting)	High accuracy; manages imbalanced data.
Support Vector Machines	Non-Linear Model	Effective for complex decision boundaries.
Neural Networks	Deep Learning	Can capture intricate relationships.

Evaluation Metrics to Assess Classification Models

Evaluating a model is critical to understanding its performance. For binary classification tasks like Persistency Prediction, the following metrics are widely used:



- **Accuracy:** Indicates the overall correctness of predictions.
- **Precision:** Reflects how many of the predicted Persistent patients were actually persistent.
- **Recall:** Shows how well the model captured all actual Persistent patients.
- **F1-Score:** Balances Precision and Recall, especially useful if data is imbalanced.
- **ROC-AUC Score:** Evaluates the model's ability to differentiate between classes.
- **Why ROC-AUC is Key:** ROC-AUC provides a threshold-independent evaluation for binary classification.
    - High ROC-AUC indicates good class separation.
    - Closer to 1.0: Excellent performance
    - 0.5: Random predictions.



Best Practices for Model:

- Evaluation Always consider multiple metrics for a complete picture.
- Use Confusion Matrices to identify the types of errors made (e.g., FP, FN).
- For imbalanced datasets, focus on Precision, Recall, F1-Score, and ROC-AUC rather than Accuracy.

Business Requirements and Model Evaluation

❑ **Business Needs:**

- High interpretability for stakeholder understanding.
- Competitive accuracy for reliable predictions.

❑ **Model Evaluation:**

- Logistic Regression: Best for interpretability, highest AUC (0.87).
- Random Forest: Slightly better accuracy (80.29%), less interpretability.
- XGBoost and SVM: Strong but black-box; lower interpretability.
- Stacking Model: Combines strengths of multiple models, improves robustness but reduces interpretability.

Model	Accuracy (%)	AUC	Interpretability	Notes
Logistic Regression	79.56	0.87	High	Best for interpretability.
Random Forest	80.29	0.85	Medium	Higher accuracy, less clear.
XGBoost	79.12	0.85	Low	Black-box model.
SVM	78.39	0.86	Low	Not suitable.
Stacking Model	79.64	0.86	Medium	Balances performance.

Why Stacking Model?

❑ Improved Robustness:

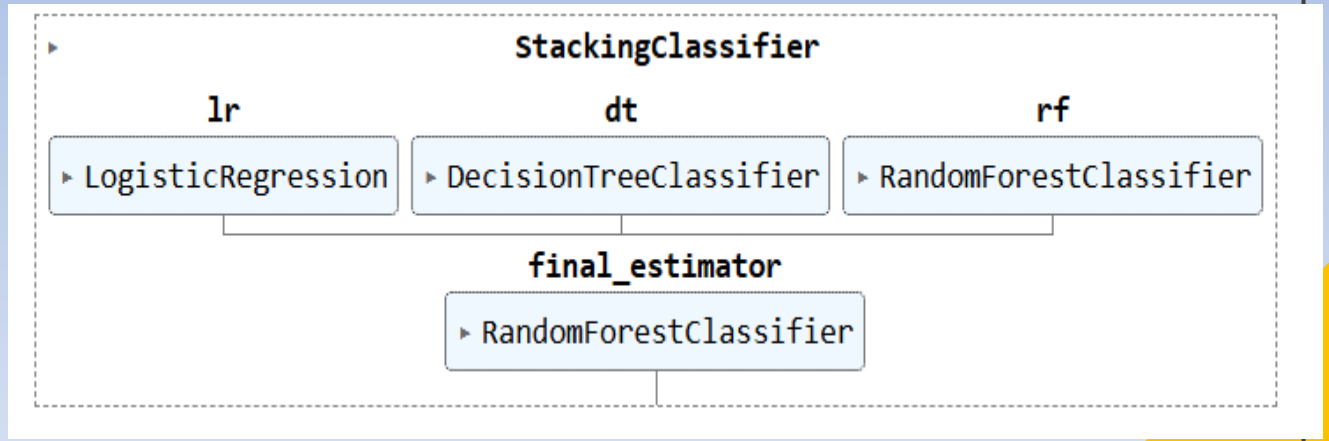
- Combines Logistic Regression and Decision Tree for interpretability.
- Uses Random Forest as the final estimator to enhance accuracy.

❑ Balanced Trade-off:

- Preserves interpretability from base models.
- Slight accuracy boost compared to standalone models.

❑ Business Impact:

- Provides actionable insights with reliable predictions.
- Addresses stakeholder needs for explainable results.



Model Optimization and Improvement

❑ Model Optimization

Hyperparameter Tuning:

- Fine-tune model parameters to improve performance.
- Use GridSearchCV or RandomizedSearchCV for systematic search of optimal hyperparameters.
- Focus on key parameters such as max_depth, n_estimators, learning_rate, and subsample.

Ensemble Methods:

- Combine multiple models to create an ensemble (e.g., VotingClassifier, Stacking).Leverage strengths of different models for improved generalization.

Cross-Validation:

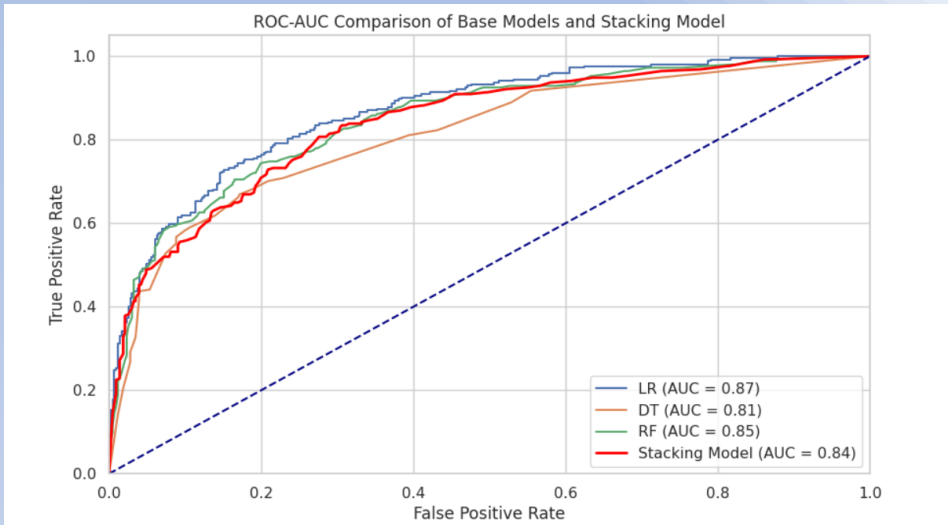
- Use K-Fold Cross-Validation to validate model performance on multiple subsets of the data.
- Ensures the model's robustness and avoids overfitting.

❑ Handling Class Imbalance

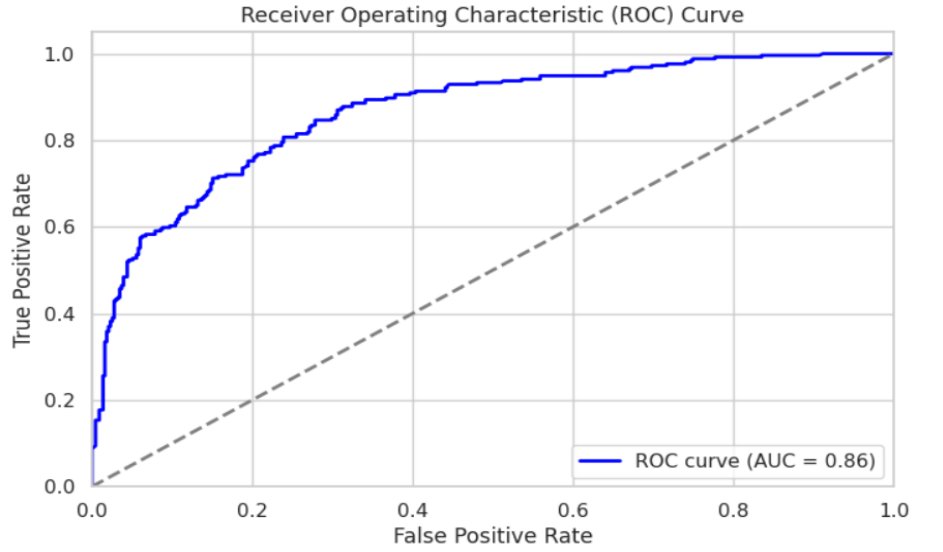
- **Oversampling:** Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes.
- **Class Weight Adjustment:** Adjust weights for the minority class in the model's loss function to reduce bias toward the majority class.

Model Implementation and Results Overview

- **Base Models:** Logistic Regression (high interpretability) and Decision Tree (rule-based clarity).
- **Final Estimator**: Random Forest tuned using GridSearchCV for improved performance.
- **Class Imbalance Solution:**
  - Applied SMOTE to balance dataset.
  - Enhanced recall for minority class (Class 1).
- **Hyperparameter Tuning:**
  - **Optimal Parameters:**
    - Max Depth: 5
    - Number of Trees: 100
  - Achieved ROC-AUC: 0.8814 (highest among all models).
- **Predicted Probabilities Example:** Class 0: 0.24, Class 1: 0.76 → Prediction: Class 1.
- **Misclassification Rate**: 22% (144/685 samples).
- **ROC-AUC Comparison:**
  - **Standalone Models:** 0.84–0.87.
  - **Stacking Model:** 0.8814 (best performance).
- **Conclusion:** The stacking model balances interpretability and performance, meeting business requirements with robust predictions and improved recall for critical classes.




```
Fitting 5 folds for each of 9 candidates, totalling 45 fits
Best Parameters: {'final_estimator__max_depth': 5, 'final_estimator__n_estimators': 100}
Best ROC-AUC Score: 0.8814629506016356
```



Preprocessing and Model Dependency in Deployment

- **Preprocessing in Model Training:**
- Preprocessing is crucial for preparing data before training (e.g., scaling, encoding, handling missing values).
- Without proper preprocessing, predictions may be inaccurate or fail entirely.
- **Example:** If trained with scaled data using StandardScaler, raw unscaled data in production leads to unreliable predictions.

- **Issue with Deployment Without Preprocessing:**
- If only the model is saved and not the preprocessing pipeline :
  - When deployed, the app will not automatically perform preprocessing (scaling, encoding, etc.).
  - **Result:** Requires manual preprocessing steps in the app, or the model will fail to provide correct predictions.

Model Deployment

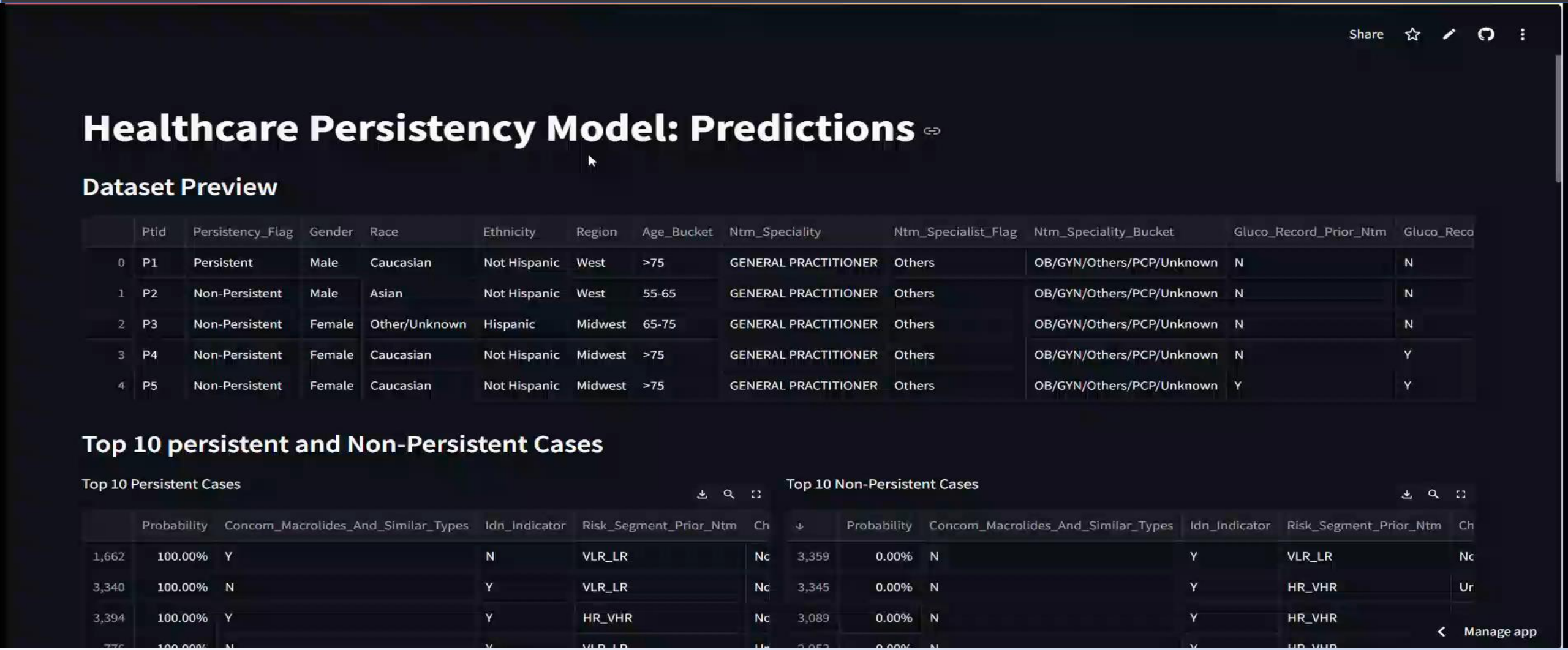
❑ Deployment Preparation

- **Model Packaging:** Save the best-performing model using Pickle or Joblib for future use or deployment.
- **API Integration:** Prepare model as a REST API using Flask or FastAPI for easy integration with other systems.
- **Cloud Deployment:** Use cloud platforms like AWS, GCP, or Azure for model deployment and scaling.

❑ Continuous Monitoring and Updates

- **Model Retraining:** Periodically retrain the model as new data is collected to keep the predictions accurate and up-to-date.
- **Model Drift Detection:** Monitor for performance degradation over time and identify potential causes for retraining.

Model Demonstration:



	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	Gluco_Record_Prior_Ntm	Gluco_Reco
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	N
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N	Y
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	Y	Y

	Probability	Concom_Macrolides_And_Similar_Types	Idn_Indicator	Risk_Segment_Prior_Ntm	Ch
1,662	100.00%	Y	N	VLR_LR	Nc
3,340	100.00%	N	Y	VLR_LR	Nc
3,394	100.00%	Y	Y	HR_VHR	Nc

	Probability	Concom_Macrolides_And_Similar_Types	Idn_Indicator	Risk_Segment_Prior_Ntm	Ch
3,359	0.00%	N	Y	VLR_LR	Nc
3,345	0.00%	N	Y	HR_VHR	Ur
3,089	0.00%	N	Y	HR_VHR	

Thank You!