Name	MOHAMMADISTIYAK SHAIKH
Email ID	Shaikhistiyak9824@gmail.com
Country	United Kingdom
College	London Metropolitan University
Specialization	Data Science
Project	Persistency of a drug

Contents

Problem description

ABC Pharma wants to understand patient **drug persistency**—whether patients continue to take their medications as prescribed by physicians. The goal is to build a **machine learning model** that predicts persistency using patient demographics, clinical history, risk factors, and treatment behavior. Automating this process will help physicians and the pharma company improve adherence strategies and personalize patient interventions.

Business understanding

The business objective of this project is to improve drug adherence by identifying patients who are at risk of non-persistence. This initiative aims to reduce costs associated with non-adherence, enhance patient health outcomes, and enable more effective targeting of patient support programs. Key stakeholders include the pharmaceutical company ABC Pharma, the data science team, and healthcare providers. The success of the project will be measured by the development of a predictive model that demonstrates high recall and precision in identifying non-persistent patients, along with sufficient model explainability to support actionable clinical and programmatic interventions.

Project lifecycle along with deadline

Phase	Task
Week 8	EDA, cleaning, preprocessing
Week 9	Feature engineering, modeling
Week 10	Model tuning, evaluation
Week 11	Deployment
Week 12	reporting

Type of data have got for analysis

The dataset consists primarily of **categorical and binary features** collected from patients, along with one key numerical column:

- **Categorical features** include demographic information such as Gender, Race, Ethnicity, Region, Age_Bucket, and clinical details like Ntm_Speciality, Ntm_Specialist_Flag, and bucketed scores.
- **Binary features** (e.g., risk factors, comorbidities) are represented by "Y"/"N" values and indicate presence or absence of certain medical conditions or risk factors.
- **Numerical features** are minimal; a key example is Count_Of_Risks, which is an actual integer count.
- The target variable is Persistency_Flag, which is binary and classifies patients as Persistent or Non-Persistent with their medications.

Property	Value
Number of Rows	3424
Number of Columns	69
Number of Numeric Columns	2
Number of Categorical Columns	67
List of Numeric Columns	[Dexa_Freq_During_Rx, Count_Of_Risks]
List of Categorical Columns	[Ptid, Persistency_Flag, Gender, Race, Ethnici...
Missing Values (Total)	0
Missing Values (Per Column)	{'Ptid': 0, 'Persistency_Flag': 0, 'Gender': 0...

Problems in the data

Missing Values (NA):
Upon inspection, the dataset does **not contain any missing values** (nulls). This is ideal and suggests a pre-cleaned or high-quality data source.

Outliers:
While most columns are categorical or binary (where outliers do not apply), numeric features like Count_Of_Risks and Change_T_Score may contain **extreme values**.
These outliers can disproportionately influence models like logistic regression or linear models, so identifying and treating them is important.

Skewed Features:
Some categorical features may be **highly imbalanced**. For example, certain regions, age groups, or risk factors may dominate the data, potentially leading to biased model performance. Also, the target variable Persistency_Flag should be checked for balance.

Column	Skewness	Interpretation	Suggestion
Persistency_Flag	0.51	Mild right skew (more 0s)	Leave as is (if binary)
Dexa_Freq_During_Rx	6.81	Strong right skew	Transform or clip outliers
Count_Of_Risks	0.88	Moderate right skew	Optional transform

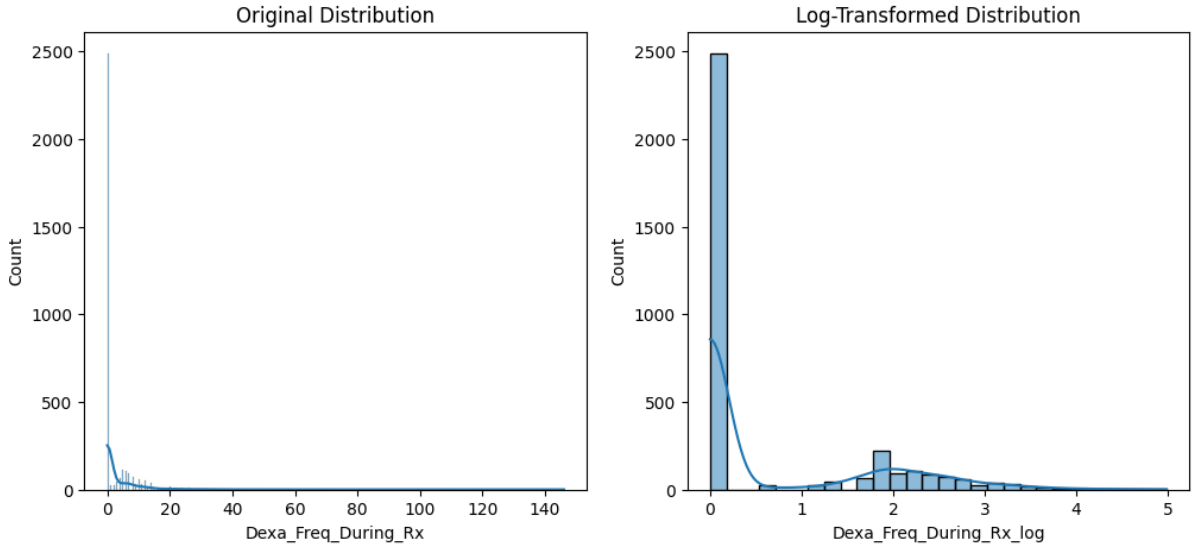


FIGURE 1:LOG TRANSFORMATION

Best Strategy: Log Transformation

The variable is **strongly right-skewed** → it can dominate distance-based models or linear relationships.

Log transformation **reduces skew**, improves **model performance**, and **preserves all data points**.

It's interpretable in the medical context (e.g., frequency on a log scale is still meaningful to analysts or physicians).

Approaches will be trying to apply on the dataset to overcome problems

Since the dataset does not have NA values, the focus shifts to cleaning and transforming the data for model readiness:

- **Text Cleaning:**
  Strip whitespaces and fix inconsistencies in text-based columns to avoid incorrect grouping or misclassification during encoding.
- **Binary Encoding of Y/N Columns:**
  All Y/N columns were converted to 1 and 0, making them suitable for modeling without introducing artificial ordinal relationships.
- **Label Encoding of Categorical Columns:**
  Non-ordinal categorical variables (e.g., Gender, Race, Region) are label encoded. This keeps the dataset compact for models that can handle categorical codes.
- **Outlier Treatment:**
  Numerical columns like Count_Of_Risks are treated using the **IQR (Interquartile Range) method**. This helps cap extreme values without completely removing data, preserving data integrity while limiting influence on model performance.

Outlier Count per Numeric Column (IQR Method):

Persistency_Flag: 0 outliers

Dexa_Freq_During_Rx: 460 outliers

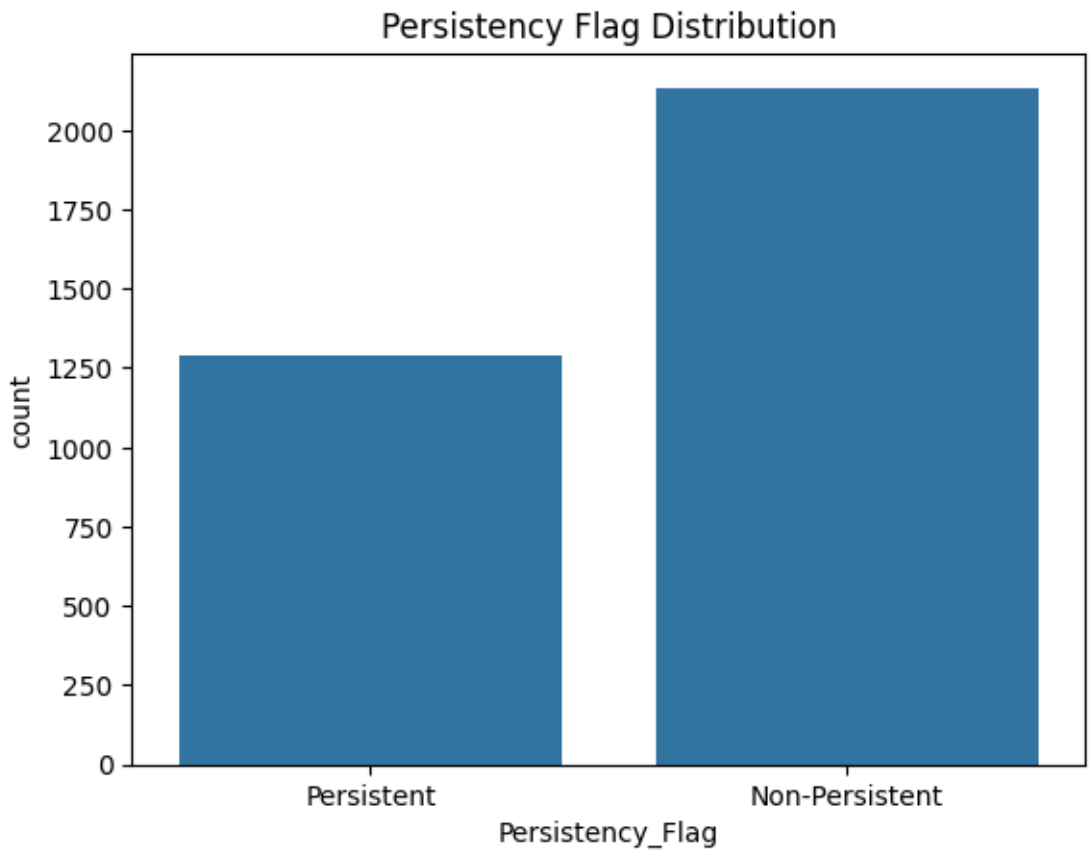
Count_Of_Risks: 8 outliers

Column	Outlier Count	Interpretation
Persistency_Flag	0	Expected — it's a binary target variable (0 or 1).
Dexa_Freq_During_Rx	460	**Significant outliers** — heavy right skew.
Count_Of_Risks	8	Minor outliers — could be genuine high-risk patients.

After the applying the log transformation on the column name Dexa_Freq_During_Rx, and relaced with the transformed column name, Dexa_Freq_During_Rx_log. Now in the dataset outlier condition mentioned below:

Column	Outlier Count
Persistency_Flag	0
Dexa_Freq_During_Rx_log	460

Class distribution check for imbalance



Target Variable Distribution:
Persistency_Flag
Non-Persistent 2135
Persistent 1289

FIGURE 2: TARGET VARIABLE DISTRIBUTION

- **Feature Scaling:**
  Applied **StandardScaler** to scale numeric features. This is important for algorithms like logistic regression, SVM, and gradient boosting that are sensitive to feature magnitudes.
- **Train-Test Split and Model Pipeline:**
  The data is split into training and test sets to prevent data leakage and allow reliable performance evaluation. A gradient boosting model is used as the final classifier, supported by explainable models like logistic regression and random forests for comparison.

After applying encoding final dataset, we got:

We performed.

1. Identify Y/N columns (binary risk factors) and convert it from "Y", "N" to 1,2
2. Apply label encoding to non-ordinal categorical columns

Ptid	Persistency_Flag	Gender	Race	…..	Count_Of_Risks	Dexa_Freq_During_Rx_log
P1	1	1	2		0	0.0
P2	0	1	1		0	0.0
P3	0	0	3		2	0.0
……					..	….
P3424	0	0			1	0.0

Week-10

EDA (Exploratory Data Analysis)

Dataset Shape

Rows: 3424, Columns: 69

Data Types and Missing Values

Column	Non-Null Count	Dtype
Ptid	3424	object
Persistency_Flag	3424	int32
Gender	3424	int32
Race	3424	int32
Ethnicity	3424	int32
Region	3424	int32

Age_Bucket	3424	int32
Ntm_Speciality	3424	int32
Ntm_Specialist_Flag	3424	int32
Ntm_Speciality_Bucket	3424	int32
Gluco_Record_Prior_Ntm	3424	int64
Gluco_Record_During_Rx	3424	int64
Dexa_During_Rx	3424	int64
Frag_Frac_Prior_Ntm	3424	int64
Frag_Frac_During_Rx	3424	int64
Risk_Segment_Prior_Ntm	3424	int32
Tscore_Bucket_Prior_Ntm	3424	int32
Risk_Segment_During_Rx	3424	int32
Tscore_Bucket_During_Rx	3424	int32
Change_T_Score	3424	int32
Change_Risk_Segment	3424	int32
Adherent_Flag	3424	int32
Idn_Indicator	3424	int64
Injectable_Experience_During_Rx	3424	int64
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	3424	int64
Comorb_Encounter_For_Immunization	3424	int64
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	3424	int64
Comorb_Vitamin_D_Deficiency	3424	int64
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	3424	int64
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	3424	int64
Comorb_Long_Term_Current_Drug_Therapy	3424	int64
Comorb_Dorsalgia	3424	int64
Comorb_Personal_History_Of_Other_Diseases_And_Conditions	3424	int64
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	3424	int64
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	3424	int64
Comorb_Osteoporosis_without_current_pathological_fracture	3424	int64
Comorb_Personal_history_of_malignant_neoplasm	3424	int64
Comorb_Gastro_esophageal_reflux_disease	3424	int64
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	3424	int64
Concom_Narcotics	3424	int64
Concom_Systemic_Corticosteroids_Plain	3424	int64
Concom_Anti_Depressants_And_Mood_Stabilisers	3424	int64
Concom_Fluoroquinolones	3424	int64
Concom_Cephalosporins	3424	int64
Concom_Macrolides_And_Similar_Types	3424	int64
Concom_Broad_Spectrum_Penicillins	3424	int64
Concom_Anaesthetics_General	3424	int64
Concom_Viral_Vaccines	3424	int64
Risk_Type_1_Insulin_Dependent_Diabetes	3424	int64
Risk_Osteogenesis_Imperfecta	3424	int64
Risk_Rheumatoid_Arthritis	3424	int64
Risk_Untreated_Chronic_Hyperthyroidism	3424	int64
Risk_Untreated_Chronic_Hypogonadism	3424	int64
Risk_Untreated_Early_Menopause	3424	int64
Risk_Patient_Parent_Fractured_Their_Hip	3424	int64
Risk_Smoking_Tobacco	3424	int64

Risk_Chronic_Malnutrition_Or_Malabsorption	3424	int64
Risk_Chronic_Liver_Disease	3424	int64
Risk_Family_History_Of_Osteoporosis	3424	int64
Risk_Low_Calcium_Intake	3424	int64
Risk_Vitamin_D_Insufficiency	3424	int64
Risk_Poor_Health_Frailty	3424	int64
Risk_Excessive_Thinness	3424	int64
Risk_Hysterectomy_Oophorectomy	3424	int64
Risk_Estrogen_Deficiency	3424	int64
Risk_Immobilization	3424	int64
Risk_Recurring_Falls	3424	int64
Count_Of_Risks	3424	int64
Dexa_Freq_During_Rx_log	3424	float64

Target Variable Distribution

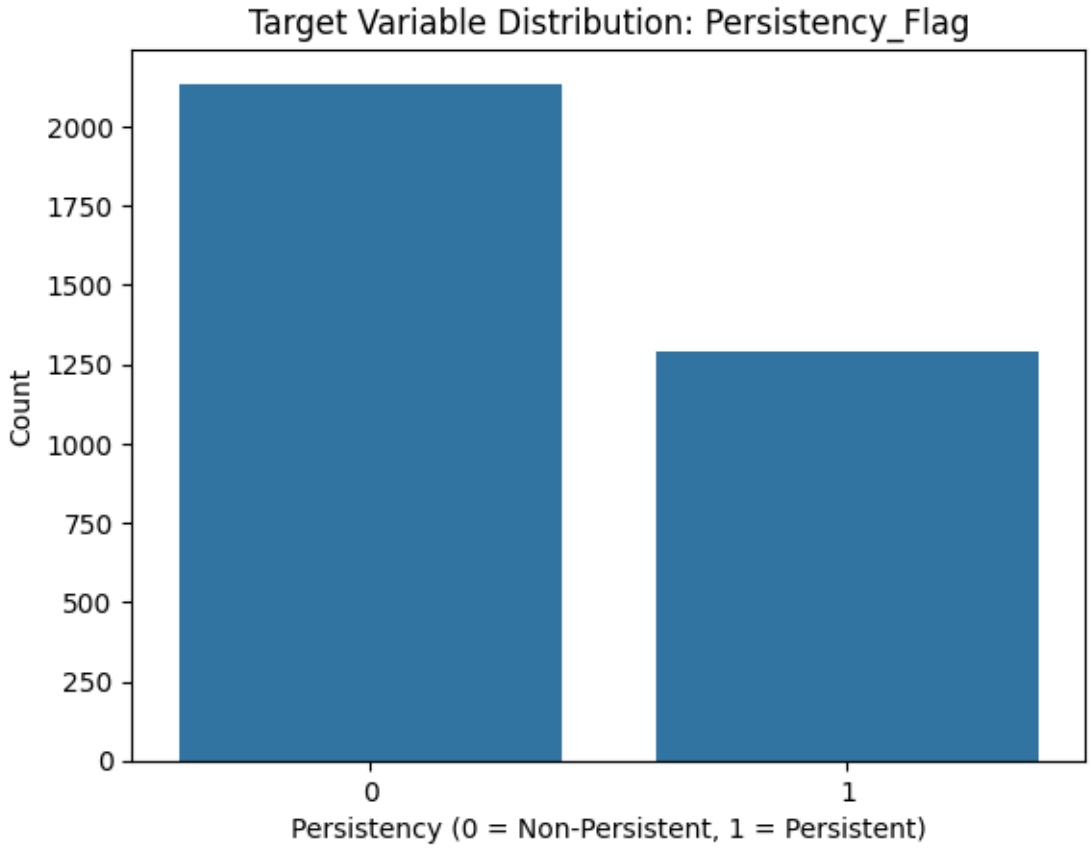


FIGURE 3:PERSISTENCY FLAG DISTRIBUTION

Persistency_Flag

0 2135

1 1289

Name: count, dtype: int64

Categorical Feature Distributions

Some examples:

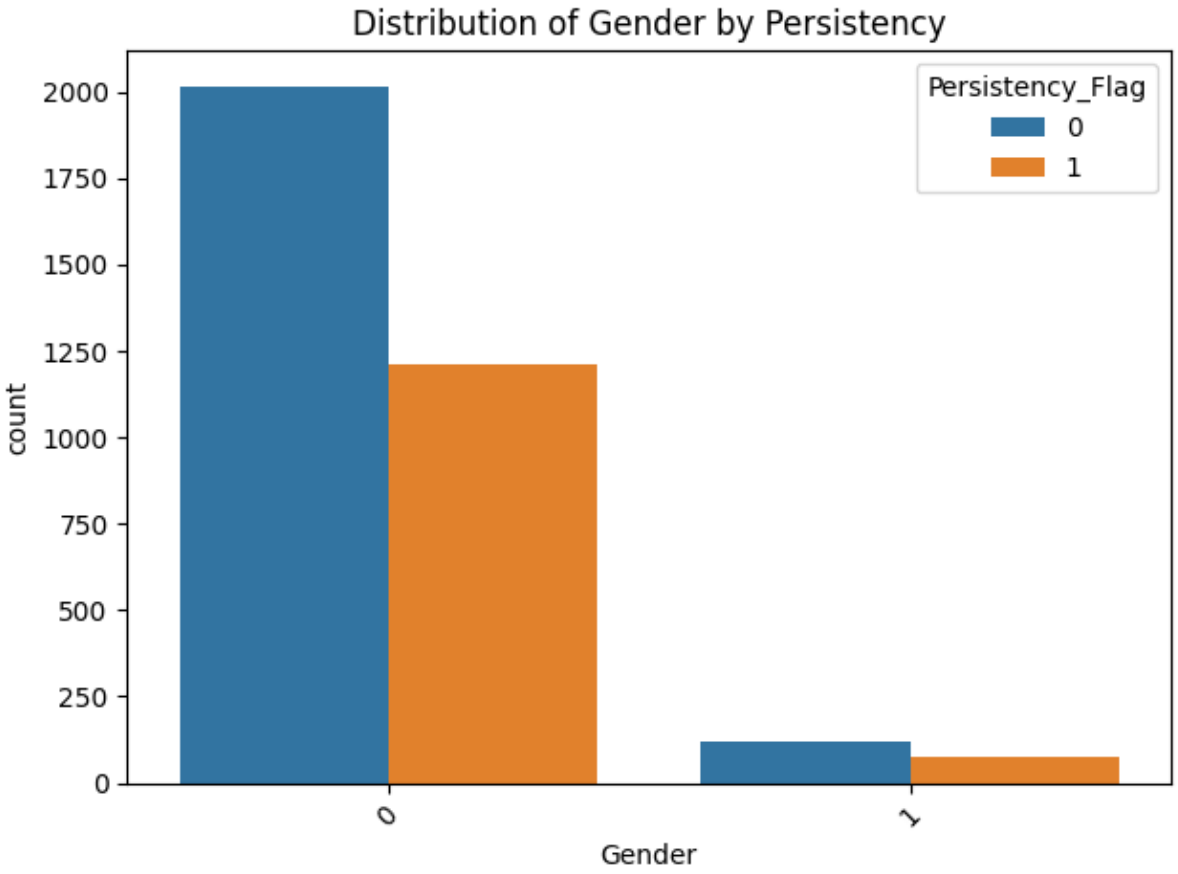


FIGURE 4: GENDER DISTRIBUTION

Numeric Feature Distributions + Outliers

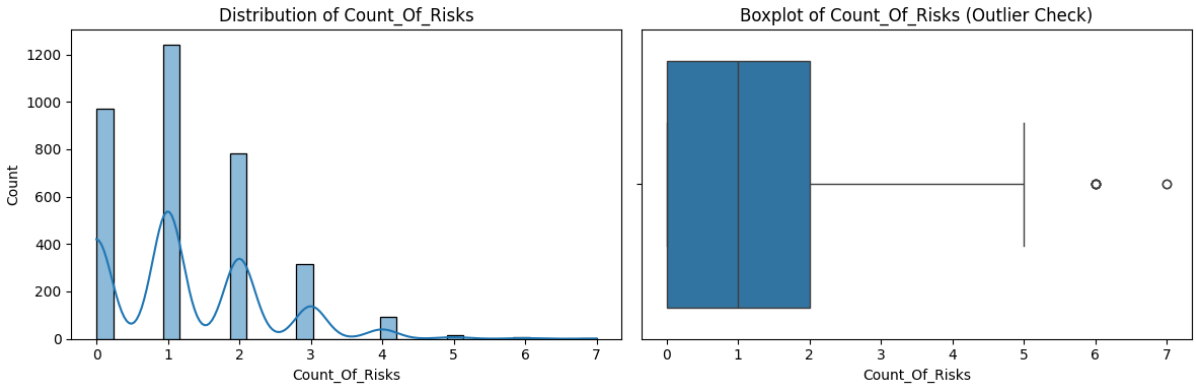


FIGURE 5: COUNT OF RISK DISTRIBUTION AND OUTLIER

Correlation Heatmap

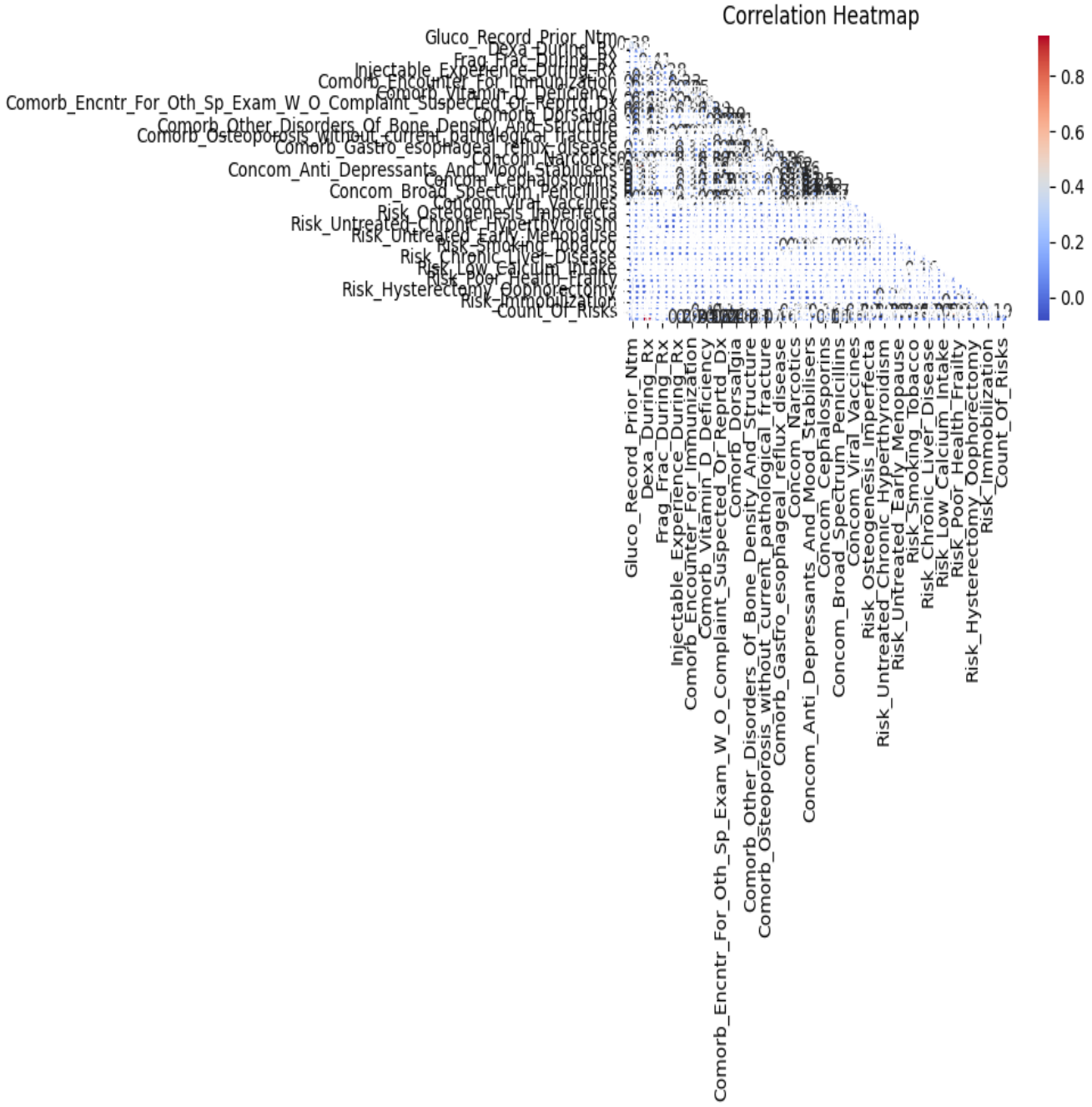


FIGURE 6: CORRELATION MATRIC OF NUMERIC VALUES

Model Selection and Comparison:

Problem Type: Classification

Target Variable: Persistency_Flag (0 = Non-Persistent, 1 = Persistent)
Business Constraint: Needs interpretability

Model Selection:

Base Model: Logistic Regression

- **Type**: Linear Model
- **Why**: Simple, interpretable, useful as a baseline
- **Pros**:
    - Coefficients show feature importance
    - Easy to explain to business stakeholders
- **Cons**:
    - May underperform on complex relationships

Ensemble Model: Random Forest Classifier

- **Why**: Combines multiple decision trees to improve accuracy
- **Pros**:
    - Handles both numeric and categorical
    - Feature importance is available
- **Cons**:
    - Less interpretable than Logistic Regression

Boosting Model: Gradient Boosting Classifier (GBM / XGBoost)

- **Why**: Builds sequential models to correct previous errors
- **Pros**:
    - Best accuracy in many structured data problems
    - SHAP values provide interpretability
- **Cons**:
    - More complex, requires explainability tools

Optional Advanced: Stacking Classifier

- Combines predictions from multiple models
- Usually improves performance
- **Not preferred if business needs interpretability**

Model	Accuracy	Interpretability	Business Fit	Final Verdict
Logistic Regression	★★★	★★★★★	✅ High	**Baseline**
Random Forest	★★★★	★★★	✅ Medium	Considerable
Gradient Boosting	★★★★★	★★★★ (with SHAP)	✅ High	**Recommended**
Stacking Classifier	★★★★★	★★	❌ Low	Use only for R&D phase

Model Comparison:

Model	Accuracy	Precision (1)	Recall (1)	F1-score (1)	Remarks
Logistic Regression	**0.81**	0.76	0.69	0.72	Interpretable (white-box model)
Random Forest	0.80	0.76	0.68	0.71	Less interpretable, robust
Gradient Boosting	**0.81**	**0.77**	**0.71**	**0.73**	Slightly better performance
Stacking Classifier	**0.81**	0.75	0.71	0.73	Complex, harder to interpret

Recommendation

Best Overall Model (Balanced View):

Gradient Boosting Classifier

- Slightly better **F1-score** and **recall** for class 1 (Persistent) — which is critical for this use case.
- Handles non-linear relationships well.
- More robust to outliers and feature interactions than Logistic Regression.

Best Interpretable Model (Business-Friendly):

Logistic Regression

- **Easiest to interpret**: coefficients directly show impact of features.
- **Useful if transparency is required**, especially in regulated industries (e.g., healthcare).
- Only slightly behind in performance (~1–2% drop in F1-score).

Logistic Regression

Why Logistic Regression

Criteria	Logistic Regression (✔ Selected)
Interpretability	High (white-box model)
Performance (Accuracy)	81%
F1-Score for Class 1	0.72
Transparency	Easy to explain to stakeholders
Speed	Fast to train and deploy
Feature Importance	Direct via coefficients

Confusion Matrix

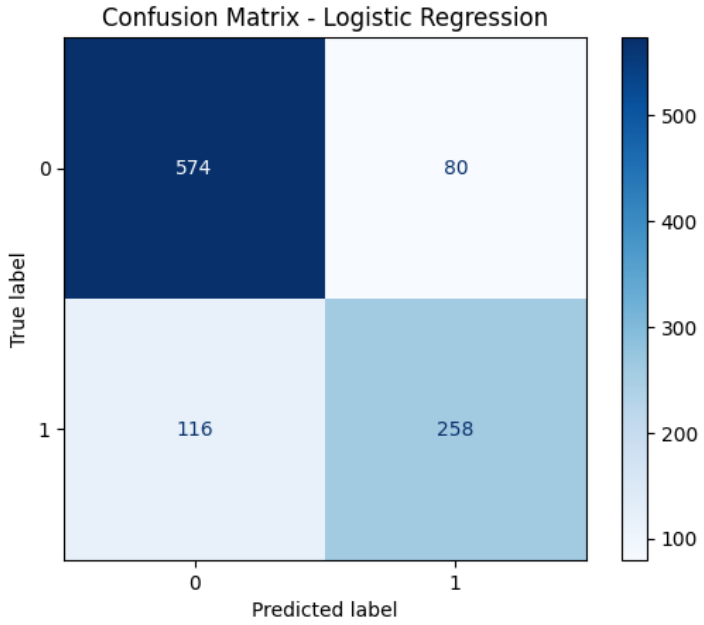


FIGURE 7: CONFUSION MATRIX

A **confusion matrix** helps evaluate the performance of a classification model by comparing **actual vs. predicted labels**.

Your Confusion Matrix:

	Predicted 0	**Predicted 1**
Actual 0	574 (TN)	80 (FP)
Actual 1	116 (FN)	258 (TP)

- **True Negative (TN = 574):** Model correctly predicted class 0.
- **False Positive (FP = 80):** Model incorrectly predicted 1 when it should be 0.
- **False Negative (FN = 116):** Model predicted 0 when it should be 1.
- **True Positive (TP = 258):** Model correctly predicted class 1.

Key Metrics:

- **Accuracy** = (TP + TN) / Total = (574 + 258) / (574 + 80 + 116 + 258) ≈ **82.7%**
- **Precision (for class 1)** = TP / (TP + FP) = 258 / (258 + 80) ≈ **76.3%**
- **Recall (for class 1)** = TP / (TP + FN) = 258 / (258 + 116) ≈ **68.9%**
- **F1 Score** = Harmonic mean of precision and recall ≈ **72.4%**

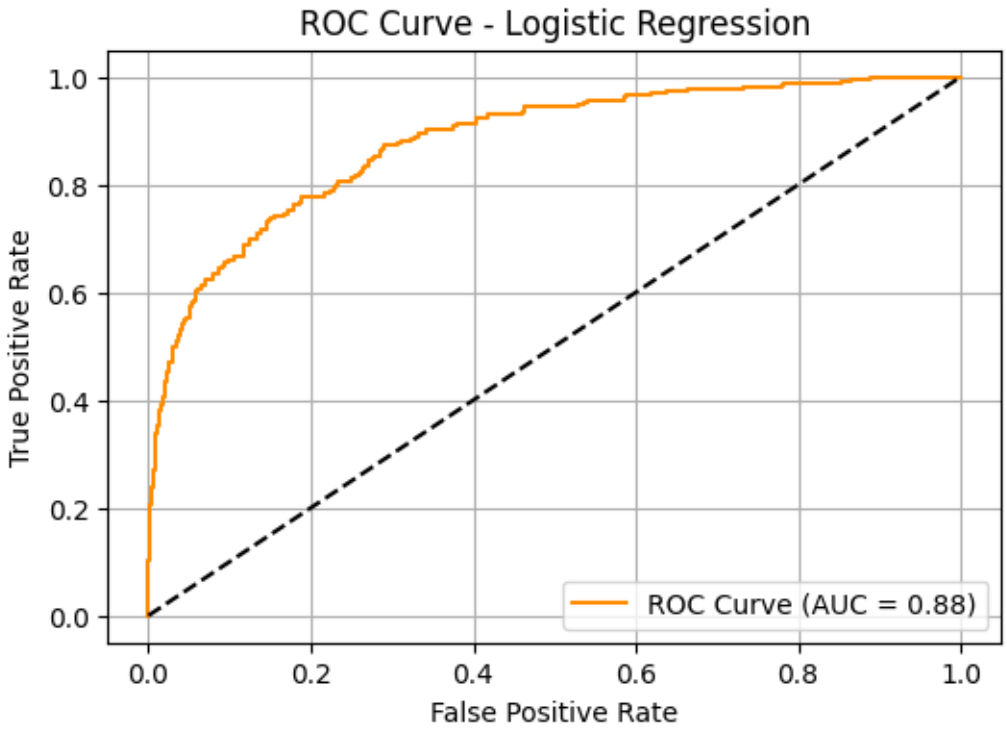


FIGURE 8: ROC CURVE

The **ROC curve** plots the **True Positive Rate (Recall)** against the **False Positive Rate (FPR)** at different classification thresholds.

What the Graph Shows:

- The **orange curve** shows how well your model separates classes.
- The **black diagonal line** is the baseline (random guessing).
- The **higher the curve above the line**, the better the model.

Key Metric:

- **AUC (Area Under Curve) = 0.88** — This is quite good. It means:
    - There's an **88% chance** the model ranks a random positive example higher than a random negative one.