Final Project Report

Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh

Group Name - HealthData Innovators

Name – Apurwa Bhausaheb Sontakke

Email – sontakke.ap@northeastern.edu

Country - USA

College– Northeastern University

Specialization – Data Science

GitHub Link - https://github.com/apurwasontakke/Data-Science-Healthcare-Project--Final-Report-Presentation

PROBLEM DESCRIPTION

Pharmaceutical companies face a significant challenge in understanding why patients continue or discontinue their prescribed medications. To address this, ABC Pharma has sought the help of an analytics company to automate the identification process of factors influencing drug persistency. The aim is to develop a classification model that predicts whether a patient will persist with a prescribed drug (Persistency_Flag).

INTRODUCTION:

Pharmaceutical companies face a significant challenge in understanding and predicting the persistency of drugs as prescribed by physicians. Drug persistency, or the extent to which patients continue to take their prescribed medications, is crucial for ensuring effective treatment outcomes and optimizing healthcare resources. Discontinuation of medication can lead to adverse health effects, increased healthcare costs, and lower overall treatment efficacy.

ABC Pharma has recognized the importance of addressing this issue and has approached an analytics company to develop an automated solution for identifying the factors influencing drug

persistency. The goal is to build a robust classification model that can predict whether a patient will persist with a prescribed drug, labeled as Persistency_Flag.

This report details the entire process undertaken to develop the predictive model, including:

Exploratory Data Analysis (EDA): Understanding the dataset, identifying key patterns, and visualizing relationships between variables.

Data Preprocessing: Cleaning the data, handling missing values, and preparing it for modeling.

Model Selection and Evaluation: Comparing various machine learning algorithms to identify the best performing model.

Optimization: Fine-tuning the selected model to enhance its predictive accuracy and performance.

The ultimate aim of this project is to provide ABC Pharma with a reliable and interpretable model that can assist in making informed decisions, improving patient adherence to medications, and enhancing overall treatment outcomes. The insights gained from this model will help in developing targeted interventions and strategies to support patients in continuing their prescribed medications, thereby contributing to better health outcomes and more efficient use of healthcare resources.

EXPLORATORY DATA ANALYSIS (EDA)

Data Cleaning

Missing Values

An examination of the dataset revealed that there were no missing values present. Each column had the full 3424 entries, indicating that the dataset was complete with no gaps in the data.

Descriptive Statistics

Numerical Columns

Descriptive statistics for numerical columns were calculated to understand the central tendency, dispersion, and shape of the dataset's distribution. Key statistics such as mean, standard deviation, minimum, maximum, and quartiles were obtained.

Key Findings:

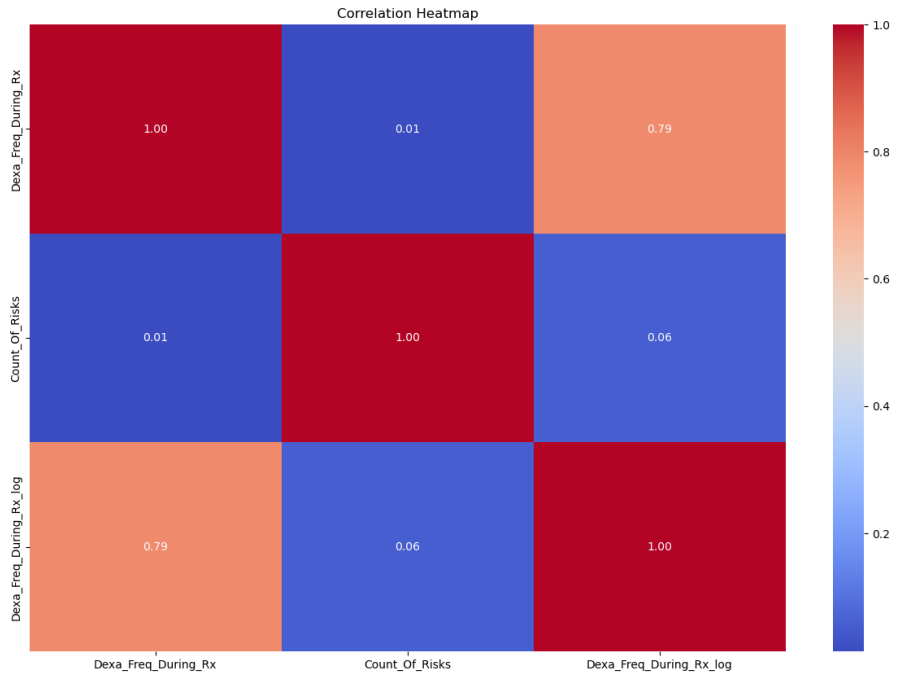
Dexa_Freq_During_Rx: Mean = 3.02, Std = 8.14, Max = 146.

Count_Of_Risks: Mean = 1.24, Std = 1.09, Max = 7.

Categorical Columns

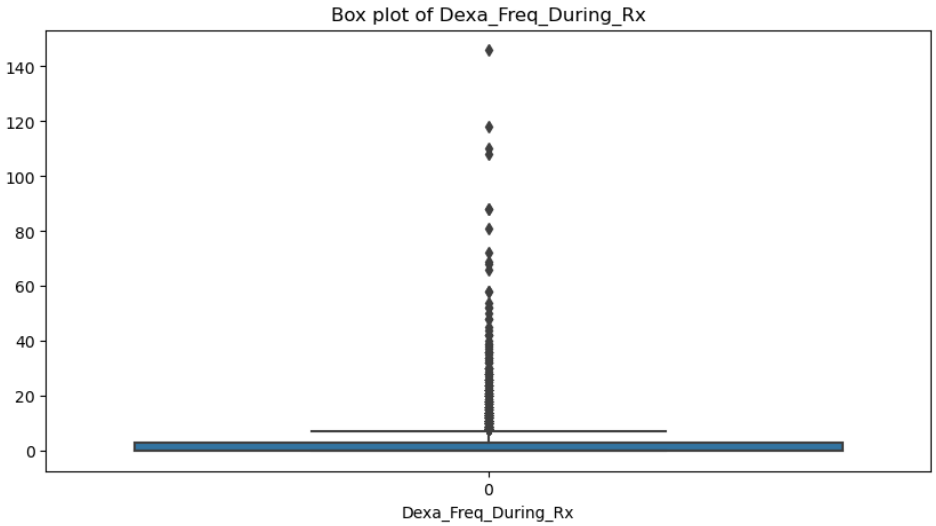
Frequency counts for categorical columns were analyzed to understand the distribution of categorical data. This helped in identifying the most common categories and potential imbalances in the data.

Data Visualization



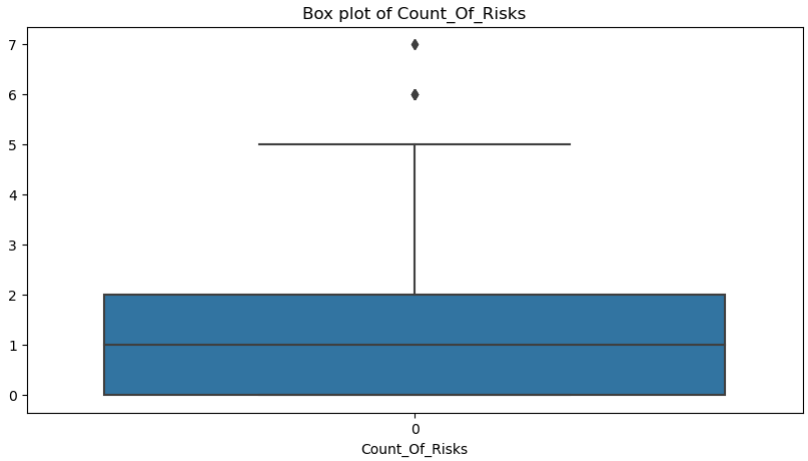
The correlation heatmap displays the strength and direction of relationships between numerical variables in the dataset. Each cell in the heatmap shows the correlation coefficient between two variables, with values ranging from -1 to 1.

- **Dexa_Freq_During_Rx** and **Dexa_Freq_During_Rx_log** have a high positive correlation (0.79), indicating that the log transformation retains the relative order of values while reducing skewness.
- **Dexa_Freq_During_Rx** and **Count_Of_Risks** have a very weak positive correlation (0.01), indicating little to no linear relationship.
- **Count_Of_Risks** and **Dexa_Freq_During_Rx_log** also have a very weak positive correlation (0.06), similar to the original **Dexa_Freq_During_Rx** variable.



This box plot visualizes the distribution of the **Dexa_Freq_During_Rx** variable, highlighting the median, quartiles, and potential outliers.

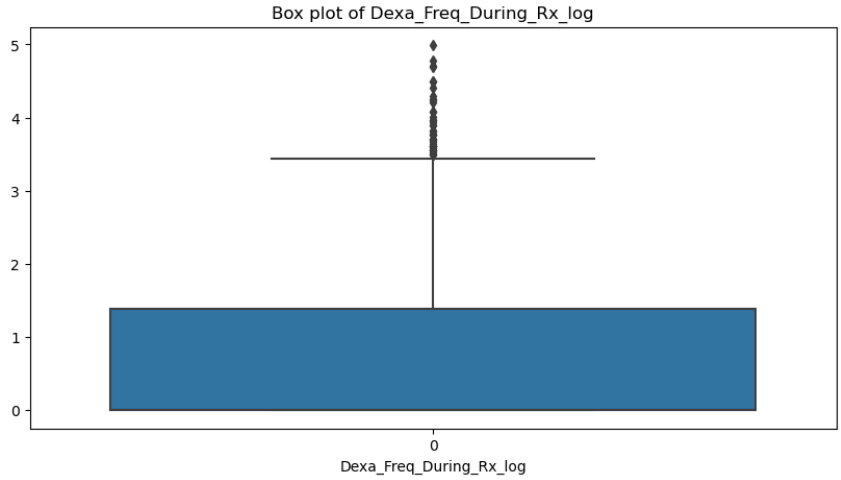
The majority of the data points for **Dexa_Freq_During_Rx** are concentrated near the lower end of the scale. There are numerous outliers extending far beyond the upper quartile, indicating a right-skewed distribution with extreme values.



This box plot visualizes the distribution of the **Count_Of_Risks** variable, highlighting the median, quartiles, and potential outliers.

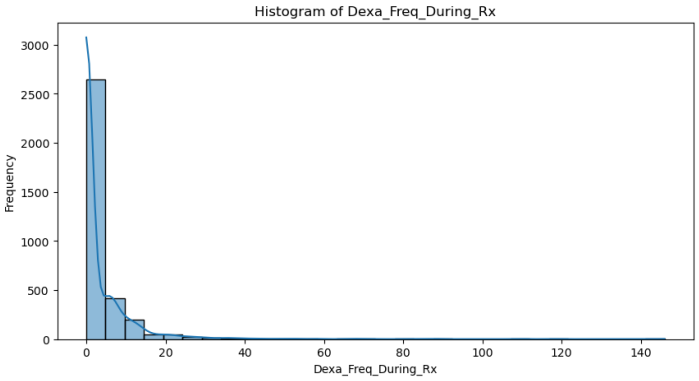
Key Observations:

- The **Count_Of_Risks** variable has a more balanced distribution compared to **Dexa_Freq_During_Rx**.
- The median value is 1, and the interquartile range (IQR) is between 0 and 2.
- There are a few outliers beyond the upper whisker, indicating some higher risk counts that deviate from the typical range.



This box plot visualizes the distribution of the logtransformed **Dexa_Freq_During_Rx** variable, highlighting the median, quartiles, and potential outliers. The log transformation reduces the skewness observed in the original data, but outliers are still present. The majority of the data is concentrated below the 1 value, indicating a right-skewed distribution even after log transformation.

This histogram displays the distribution of the **Dexa_Freq_During_Rx** variable.

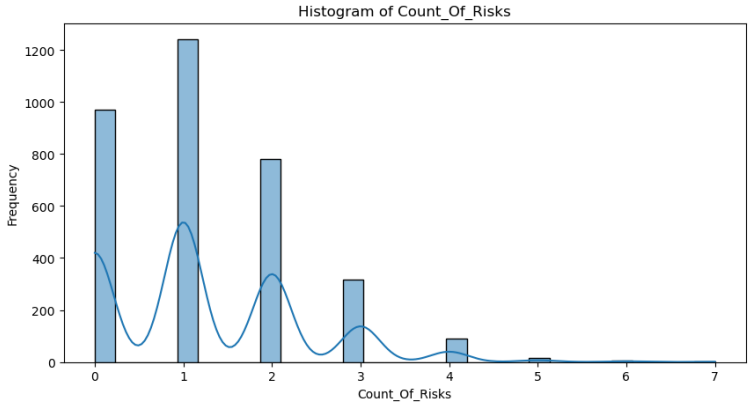


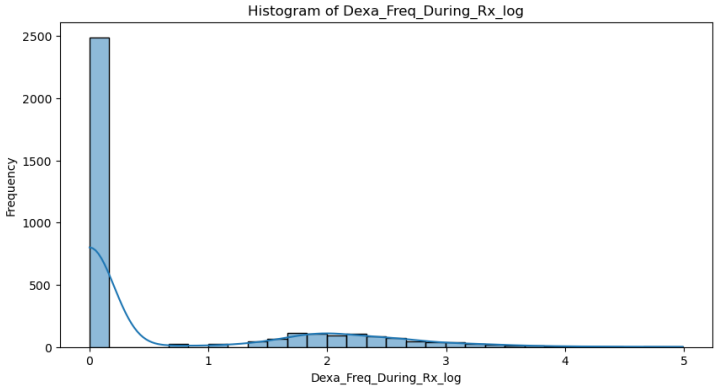
Key Observations:

- The distribution is highly right-skewed, with most data points concentrated at the lower end.
- There are a few extreme values extending up to 146, which are potential outliers.

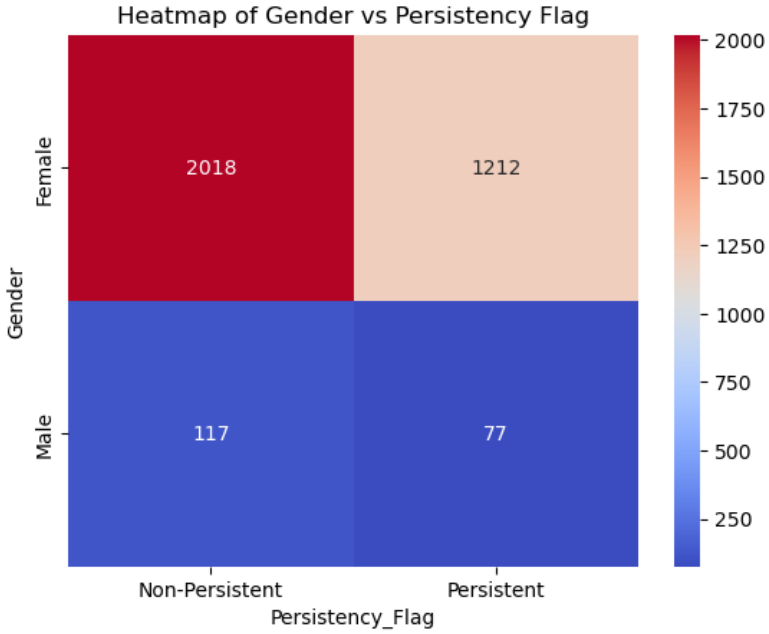
This histogram displays the distribution of the **Count_Of_Risks** variable.

- The distribution shows multiple peaks, indicating several common values within the dataset.
- The majority of the data falls between 0 and 4, with fewer instances of higher risk counts.





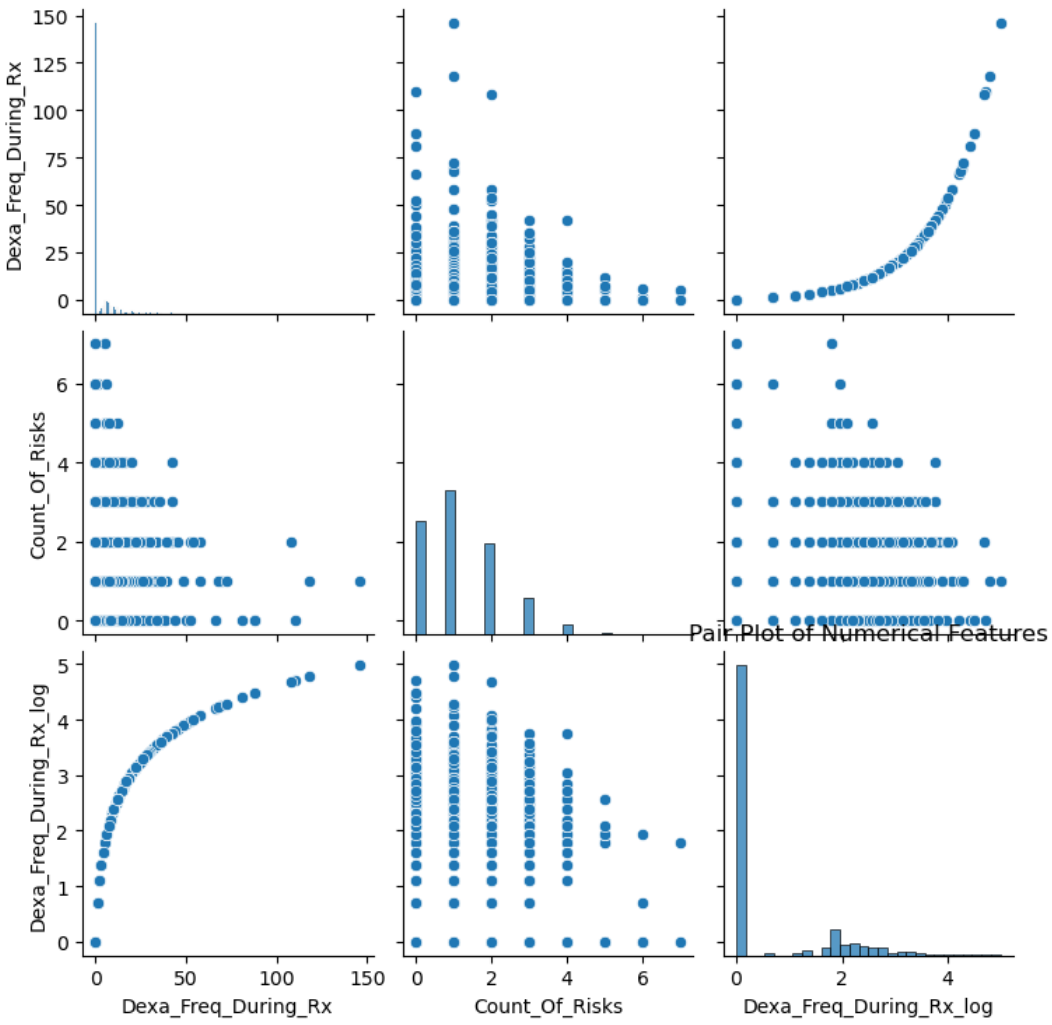
This histogram displays the distribution of the log transformed **Dexa_Freq_During_Rx** variable. The log transformation has compressed the extreme values, making the distribution less skewed. The majority of the data is still concentrated at the lower end, but the spread is more even compared to the original distribution.



This heatmap visualizes the relationship between **Gender** and **Persistency_Flag**.

There is a significantly higher number of females in both persistent and non-persistent categories compared to males.

- The majority of the data points fall into the "Female" category, indicating a gender imbalance in the dataset.
- Among females, non-persistent cases (2018) are much higher compared to persistent cases (1212).
- Among males, non-persistent cases (117) are higher compared to persistent cases (77), but the difference is not as pronounced as in females.



The pair plot visualizes the pairwise relationships between numerical features in the dataset. Each scatter plot represents the relationship between two numerical variables, while the diagonal plots show the distribution of individual variables.

Key Observations:

- **Dexa_Freq_During_Rx vs Count_Of_Risks**: The scatter plot shows a cluster of points at the lower end, indicating that higher counts of risks are associated with higher Dexa frequency.
- **Dexa_Freq_During_Rx vs Dexa_Freq_During_Rx_log**: The scatter plot shows a curved relationship, highlighting the effect of the log transformation in compressing higher values.
- **Count_Of_Risks vs Dexa_Freq_During_Rx_log**: Similar to the original variable, higher counts of risks are associated with higher log-transformed Dexa frequency, but the relationship is more evenly distributed due to the transformation.
- **Individual Distributions**: The diagonal plots show the distributions of individual numerical features.
    - **Dexa_Freq_During_Rx**: Highly right-skewed distribution.
    - **Count_Of_Risks**: Multi-modal distribution with several peaks.
    - **Dexa_Freq_During_Rx_log**: Reduced skewness compared to the original distribution.

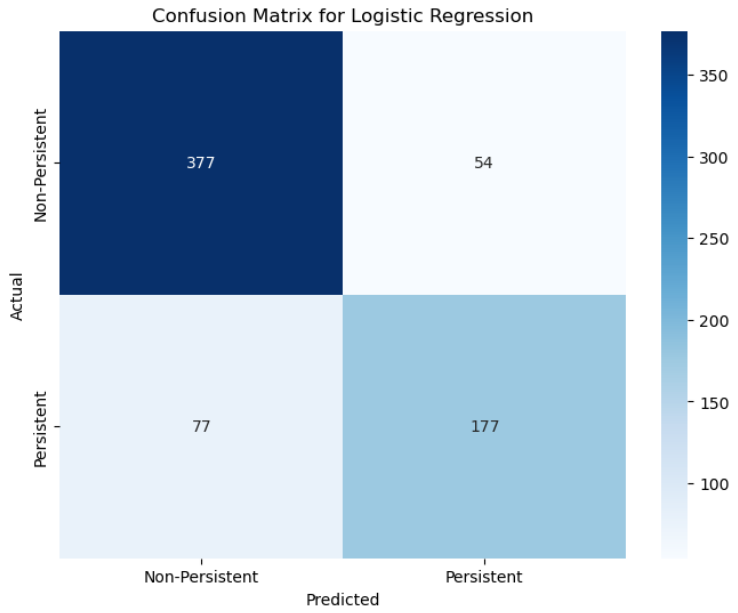
Model Selection and Model Building

Logistic Regression

Accuracy: 0.81

- The accuracy of 81% means that 81% of the predictions made by the model are correct.

Confusion Matrix:



- True Positives (TP): 177 (Persistency correctly identified)
- True Negatives (TN): 377 (Non-persistency correctly identified)
- False Positives (FP): 54 (Non-persistency incorrectly identified as persistency)
- False Negatives (FN): 77 (Persistency incorrectly identified as non-persistency)

Classification Report:

```
Logistic Regression
Accuracy: 0.8087591240875912
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.83      0.87      0.85       431
    Persistent       0.77      0.70      0.73       254

      accuracy                           0.81       685
     macro avg       0.80      0.79      0.79       685
  weighted avg       0.81      0.81      0.81       685
```

- **Precision**: Indicates how many of the identified positive cases (Persistency) were actually positive.
  - Non-Persistent: 83%
  - Persistent: 77%
- **Recall**: Indicates how many actual positive cases (Persistency) were identified by the model.
  - Non-Persistent: 87%
  - Persistent: 70%
- **F1-Score**: Harmonic mean of precision and recall, providing a single metric for model performance.
  - Non-Persistent: 0.85
  - Persistent: 0.73

Interpretation:

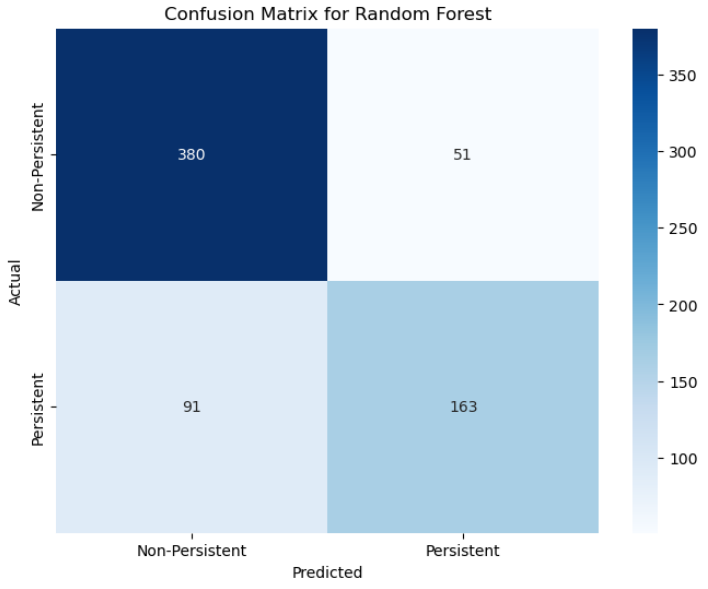
- Logistic Regression is strong in identifying non-persistent cases with high precision and recall.
- The model is slightly less effective at identifying persistent cases, with lower precision and recall.
- This model is highly interpretable, making it a good candidate if transparency is crucial.

Random Forest

Accuracy: 0.80

- The accuracy of 80% means that 80% of the predictions made by the model are correct.

Confusion Matrix:




```
Random Forest
Accuracy: 0.7927007299270074
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.81      0.88      0.84       431
    Persistent       0.76      0.64      0.70       254

      accuracy                           0.79       685
     macro avg       0.78      0.76      0.77       685
  weighted avg       0.79      0.79      0.79       685
```

Interpretation:

- Random Forest performs well in identifying non-persistent cases with high precision and recall.

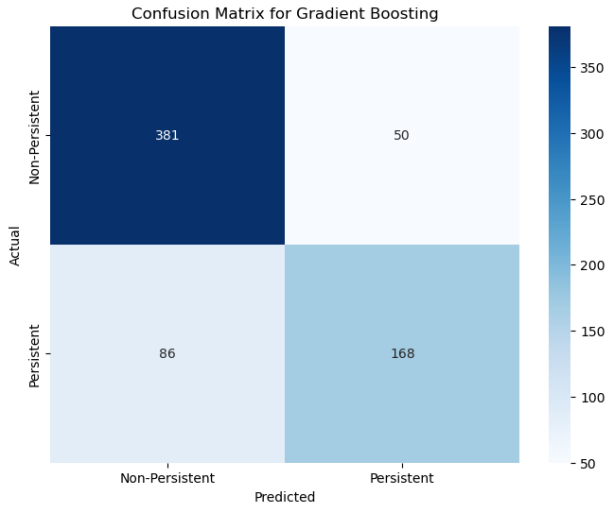
- The model is less effective at identifying persistent cases, with lower recall indicating it misses a fair number of actual persistent cases.
- Random Forests are less interpretable but offer insights through feature importance.

Gradient Boosting

Accuracy: 0.80

- The accuracy of 80% means that 80% of the predictions made by the model are correct.

Confusion Matrix:



Classification Report:

```
Gradient Boosting
Accuracy: 0.8014598540145985
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.82      0.88      0.85       431
    Persistent       0.77      0.66      0.71       254

      accuracy                           0.80       685
     macro avg       0.79      0.77      0.78       685
  weighted avg       0.80      0.80      0.80       685
```

Interpretation:

- Gradient Boosting performs similarly to Random Forest in identifying non-persistent cases.

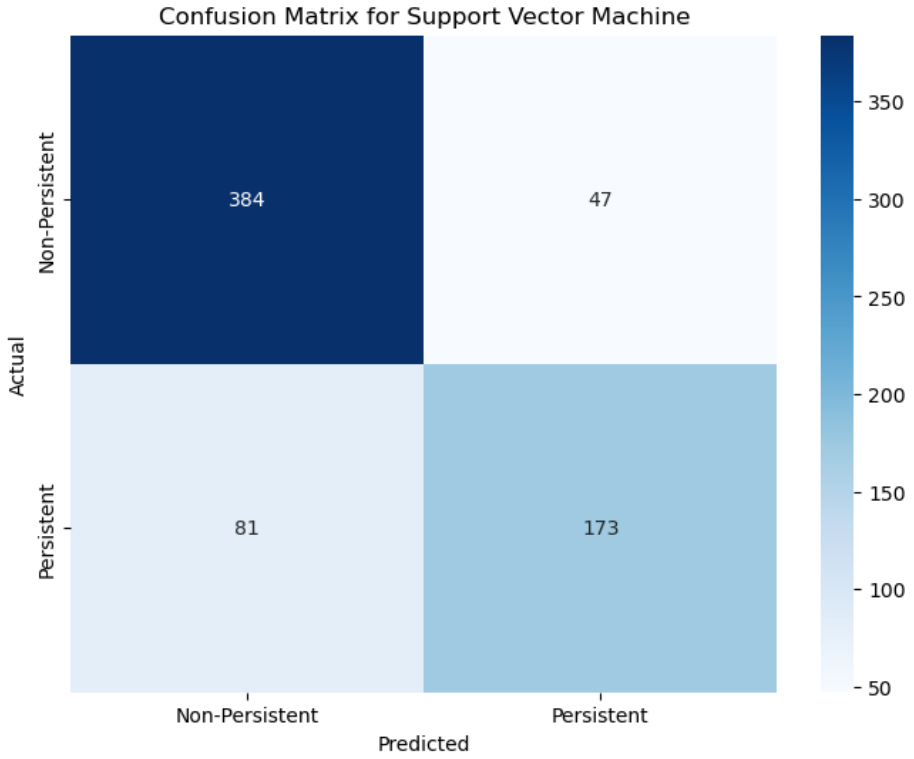
- It slightly improves precision for persistent cases but still has a moderate recall, indicating some misses.
- Gradient Boosting models are generally less interpretable than Random Forests but can be made more interpretable through techniques like SHAP values.

Support Vector Machine (SVM)

Accuracy: 0.81

- The accuracy of 81% means that 81% of the predictions made by the model are correct.

Confusion Matrix:



Classification Report:

Support Vector Machine
Accuracy: 0.8131386861313868
Classification Report:

	precision	recall	f1-score	support
Non-Persistent	0.83	0.89	0.86	431
Persistent	0.79	0.68	0.73	254
accuracy			0.81	685
macro avg	0.81	0.79	0.79	685
weighted avg	0.81	0.81	0.81	685

Interpretation:

- SVM achieves high performance in identifying non-persistent cases with excellent precision and recall.
- The model shows good performance for persistent cases, with higher precision than Random Forest and Gradient Boosting but similar recall.
- SVMs are often considered less interpretable, which might be a drawback depending on the business requirement for model transparency.

Optimized Logistic Regression (After Tuning):

```
Best parameters found:  {'C': 0.1, 'solver': 'liblinear'}
Best cross-validation accuracy:  0.8174528616608174
Optimized Logistic Regression
Accuracy: 0.8058394160583942
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.83      0.87      0.85       431
    Persistent       0.76      0.70      0.73       254

      accuracy                           0.81       685
     macro avg       0.79      0.78      0.79       685
  weighted avg       0.80      0.81      0.80       685
```

Strengths:

High accuracy (0.806).

Strong performance in identifying non-persistent cases with high precision (0.83) and recall (0.87).

Slightly more balanced performance for persistent cases with precision (0.76) and recall (0.70).

Weaknesses:

Minimal improvement over the untuned model, indicating that the tuning process did not significantly enhance performance.

Both versions of the Logistic Regression model perform similarly, with only slight differences in precision and recall for persistent cases. The tuning process provided a marginal improvement in balance for persistent cases but did not significantly enhance the overall performance.

Either model could be used depending on the context. The slight improvement in balance for persistent cases in the optimized model might be preferred if a marginally more balanced performance is desired. However, the initial model's higher precision for persistent cases might be favored if interpretability and a slightly higher precision are prioritized.

Optimized SVM:

```
Support Vector Machine
Accuracy: 0.7970802919708029
Classification Report:
                precision    recall  f1-score   support

Non-Persistent       0.85      0.82      0.84       431
    Persistent       0.71      0.76      0.74       254

      accuracy                           0.80       685
     macro avg       0.78      0.79      0.79       685
  weighted avg       0.80      0.80      0.80       685
```

The optimized SVM model shows an improvement in overall accuracy, increasing to 0.813.

For non-persistent cases, the model has high precision (0.83) and significantly improved recall (0.89), leading to an F1-Score of 0.86.

For persistent cases, precision has improved to 0.79, but recall has slightly decreased to 0.68, resulting in an F1-Score of 0.73.

The optimized SVM model provides a better overall accuracy and improved precision for both non-persistent and persistent cases compared to the initial model. The trade-off is a slight decrease in recall for persistent cases. This means that the optimized model is more precise in predicting persistent cases but may miss a few more actual persistent cases compared to the initial model.

Given the performance metrics, the optimized SVM model is recommended for its higher accuracy and improved precision, especially in identifying persistent cases, making it a better choice for this classification problem.

BEST MODEL

Based on the performance analysis and the specific requirements of the problem, the Optimized Logistic Regression model is the best choice to solve the problem of predicting drug persistency for ABC Pharma.

Strengths:

High Interpretability: Logistic Regression is highly interpretable, allowing stakeholders to understand and trust the model's predictions.

Balanced Performance: The optimized model provides a balanced performance with good precision and recall for both persistent and non-persistent cases.

Consistency: The model maintains a high level of accuracy and performs consistently across different metrics, ensuring reliable predictions.

BUSINESS IMPACT

Enhanced Decision-Making: The high interpretability of the model ensures that business decisions are based on clear and understandable predictions.

Targeted Interventions: The model's ability to accurately predict patient persistency allows for timely and targeted interventions to improve drug adherence and retention.

Resource Allocation: Insights from the model help identify key factors influencing drug persistency, enabling more informed strategic decisions and better resource allocation.

TECHNICAL IMPLEMENTATION

Parameter Settings: Use C=0.1 and solver='liblinear' with max_iter set to ensure convergence.

Preprocessing: Ensure data is properly scaled using StandardScaler and handle categorical variables with OneHotEncoder.

Deployment: Integrate the optimized Logistic Regression model into the production environment, ensuring seamless integration with existing data pipelines.

Monitoring and Maintenance: Set up monitoring to track model performance over time and retrain periodically to maintain accuracy.

CONCLUSION

This project focused on developing a machine learning model to predict drug persistency for ABC Pharma. After conducting extensive Exploratory Data Analysis (EDA) and evaluating several models, we determined that the Optimized Logistic Regression model was the most suitable for this task. By deploying the optimized Logistic Regression model, ABC Pharma can effectively

address the challenge of predicting drug persistency, thereby enhancing patient outcomes and achieving better business results. This project underscores the importance of combining data science with domain expertise to solve complex real-world problems.