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

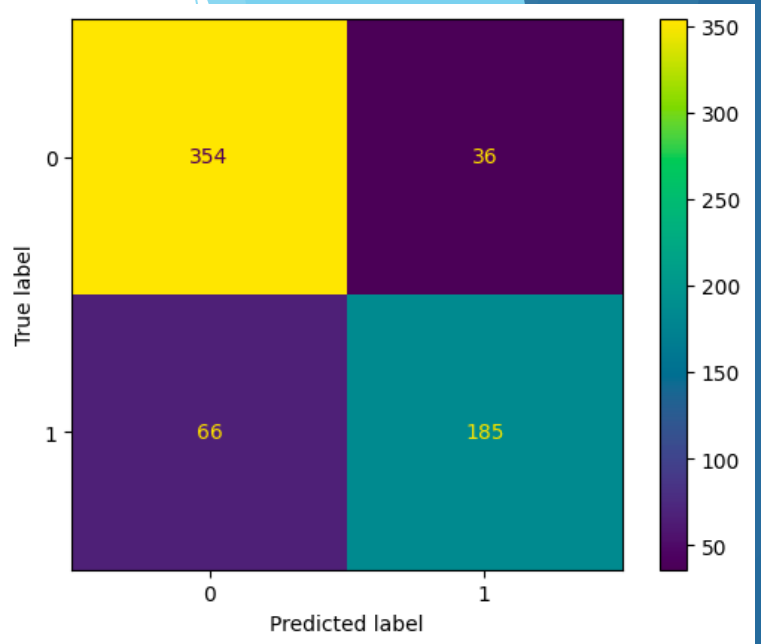
- Dexa_Freq_During_Rx: 3,424 observations; investigates DEXA scan frequency, with notable outliers
- Count_Of_Risks: Detected outliers with 6 and 7 risks counted
- Outliers in Dexa_Freq_During_Rx above the upper IQR limit: 460 observations
- Data Analysis Approach:
  - Assess if zero values represent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

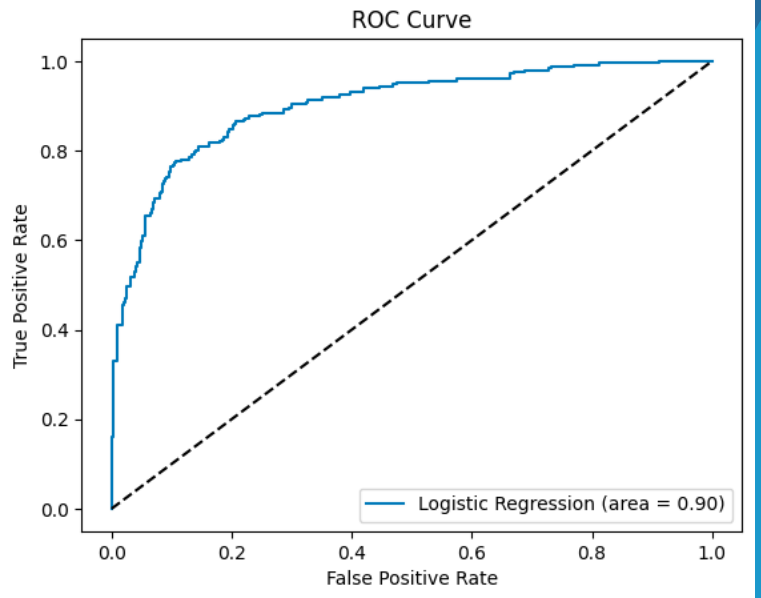
Model Selection

- **3 Models were chosen:**
  - **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.

Logistic Regression

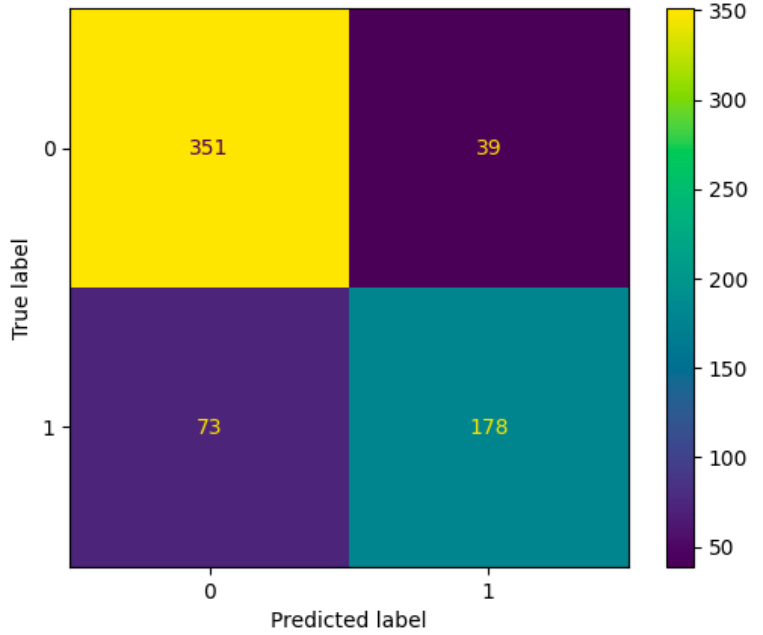
- **Performance Metrics in Percentages:**
  - **Logistic Regression Accuracy:** 84.09%
  - **Logistic Regression Precision:** 83.71%
  - **Logistic Regression Recall:** 73.71%
  - **Logistic Regression F1 Score:** 78.39%
- **Best Hyperparameters and Accuracies:**
  - **Best Hyperparameters:**
    - Regularization Strength ('C'): 0.1
    - Penalty: L2 (Ridge regression)
  - **Best Training Accuracy:** 82.44%
  - **Best Validation Accuracy:** 80.75%
  - **Best Test Accuracy (using best parameters):** 84.09%
- Logistic Regression ROC-AUC Score: 90.15%

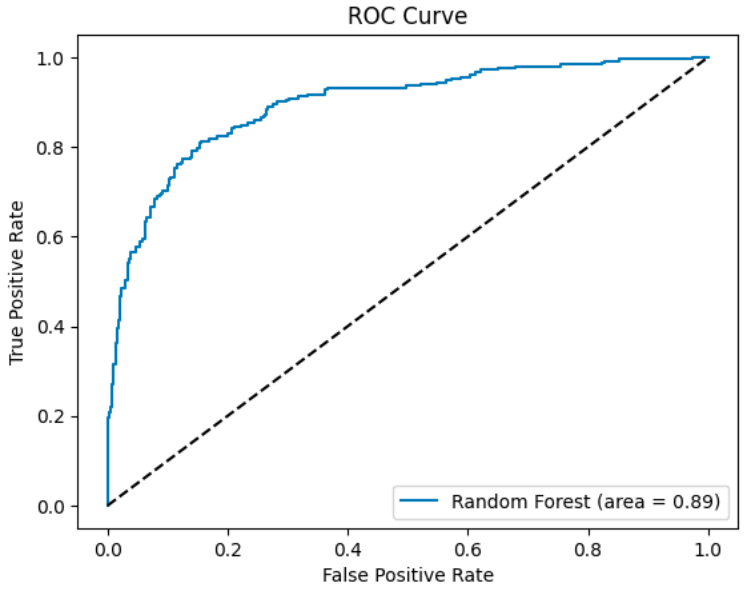




Random Forest Classification

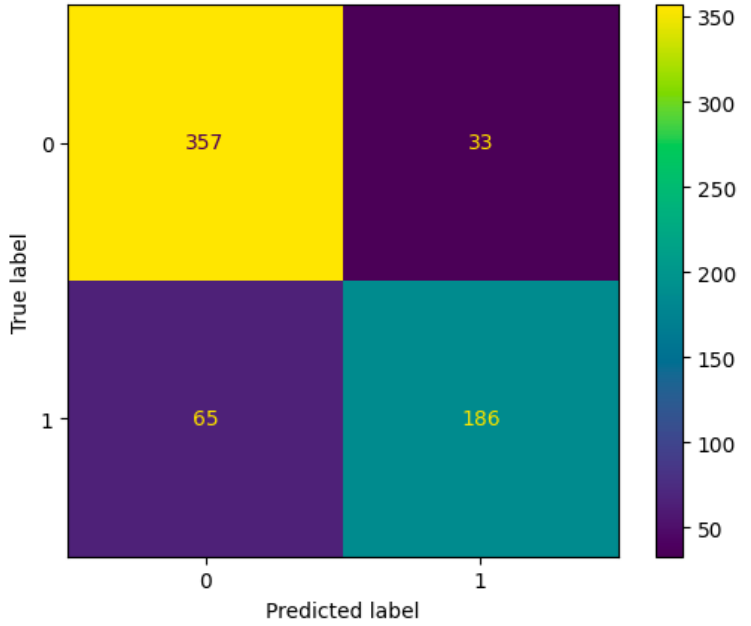
- **Performance Metrics in Percentages:**
  - **Random Forest Accuracy:** 82.53%
  - **Random Forest Precision:** 82.03%
  - **Random Forest Recall:** 70.92%
  - **Random Forest F1 Score:** 76.07%
- **Best Hyperparameters and Accuracies:**
  - **Best Hyperparameters:**
    - Maximum Depth of Trees ('max_depth'): 10
    - Minimum Number of Samples Required to Split a Node ('min_samples_split'): 2
    - Number of Trees in the Forest ('n_estimators'): 100
  - **Best Training Accuracy:** 94.46%
  - **Best Validation Accuracy:** 81.26%
  - **Best Test Accuracy (using best parameters):** 82.53%
- Random Forest ROC-AUC Score: 89.42%

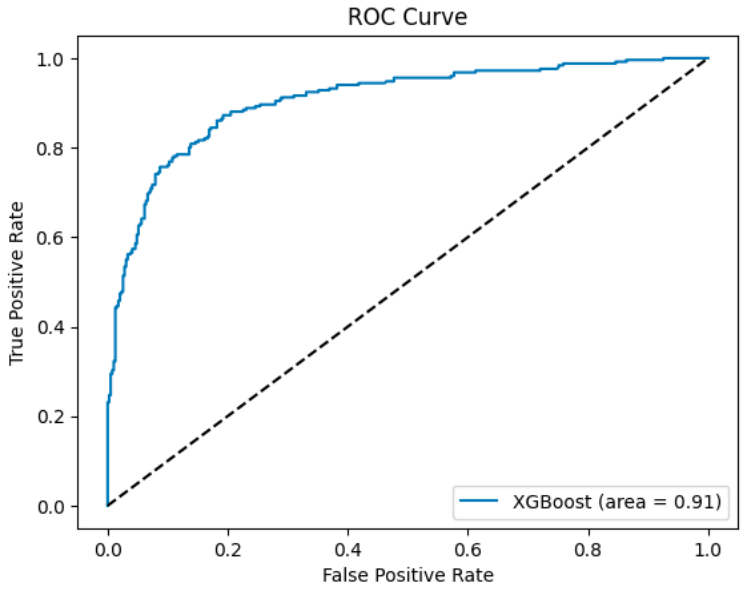




XGBoost

- **Performance Metrics in Percentages:**
  - **XGBoost Accuracy:** 84.71%
  - **XGBoost Precision:** 84.93%
  - **XGBoost Recall:** 74.10%
  - **XGBoost F1 Score:** 79.15%
- **Best Hyperparameters and Accuracies:**
  - **Best Hyperparameters:**
    - Learning Rate: 0.1
    - Maximum Depth of Trees ('max_depth'): 3
    - Number of Trees ('n_estimators'): 100
  - **Best Training Accuracy:** 86.70%
  - **Best Validation Accuracy:** 81.18%
  - **Best Test Accuracy (using best parameters):** 84.71%
- Random Forest ROC-AUC Score: 90.6%





Model Conclusions

Performance Overview:

- **Accuracy:**
  - Highest Accuracy: **XGBoost** at **84.71%**
  - Ensures overall correctness across the dataset.
- **Precision:**
  - Highest Precision: **XGBoost** at **84.93%**
  - Reflects the model's strength in reducing false positives.
- **Recall:**
  - Highest Recall: **Logistic Regression** at **73.71%**
  - Important for cases where missing a positive is costly.
- **F1 Score:**
  - Highest F1 Score: **XGBoost** at **79.15%**
  - Balances precision and recall, valuable for uneven class distribution.
- **ROC-AUC Score:**
  - Highest ROC-AUC: **XGBoost** at **90.66%**
  - Indicates the model's ability to discriminate between classes.

- **Best Model Selection:**
  - **According to Accuracy, Precision, F1 Score, and ROC-AUC Score:** XGBoost stands out as the **best model**.
  - **According to Recall:** Logistic Regression takes the lead but is outperformed by XGBoost on other metrics.
- **Conclusion:** Considering all key performance metrics, including the highest ROC-AUC score, XGBoost emerges as the superior model, providing the most reliable and balanced performance for our dataset.