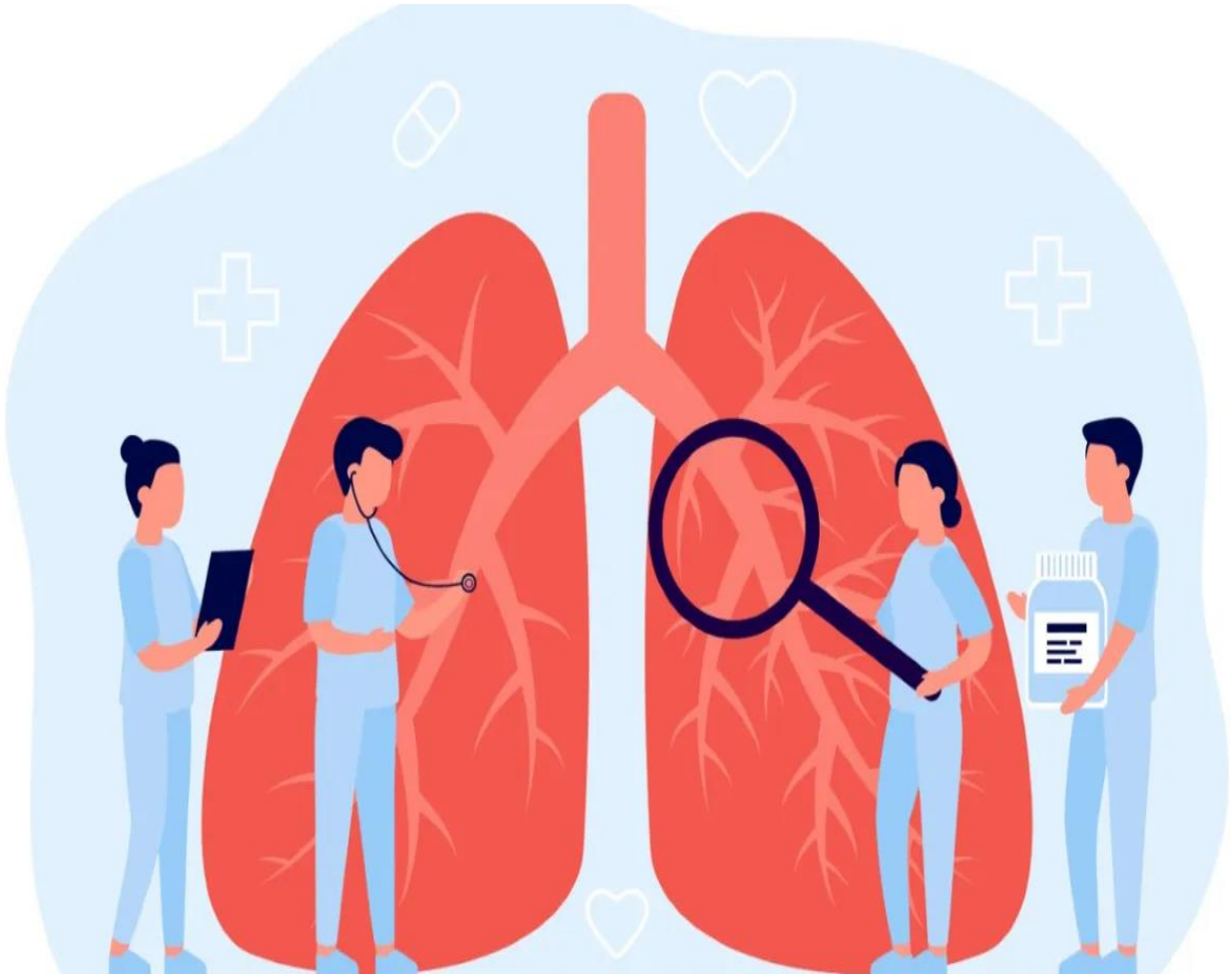
07-12-2023 | By: Datageeks



Humana Mays Case Competition-2023

Team Name :Datageeks (Team:9)

Mithilesh Bogireddy (CWID:A20394128)

Santhosh Chittiprolu (CWID:A20377479)

Sriharsha Penumudi (CWID:A20400312)

Amarnadh Oleti (CWID :A20392386)

Table of Contents

1. Executive Summary

Enhancing Therapy Adherence in Oncology through Data Analytics

In the 2023 Humana Mays Healthcare Analytics Case Competition, we are tasked with addressing a pressing challenge in oncology: improving adherence to **Osimertinib** therapy for early-stage **lung cancer** patients. Given that about 25% of patients discontinue treatment due to adverse drug events (ADEs), our project leverages data analytics to predict the risk of premature therapy discontinuation. Utilizing datasets including therapy details medical, and pharmacy claims, our goal is to develop a predictive model. This model aims to identify at-risk patients, enabling targeted interventions to enhance adherence. The ultimate objective is to improve patient outcomes in oncology by turning data insights into actionable, life-saving strategies.

2.Introduction

Leveraging Data Analytics for Improving Oncology Therapy Adherence Context and Challenge

In oncology, particularly in the treatment of early-stage lung cancer, the challenge of patient adherence to therapies like Osimertinib is significant. Despite its efficacy, the side effects of this medication often lead to premature discontinuation of treatment, undermining its potential benefits. Addressing this issue is critical for enhancing patient outcomes and is the focus of the 2023 Humana Mays Healthcare Analytics Case Competition.

3.Overview of Data

The competition provides rich datasets encompassing patient therapies and medical and pharmacy claims and offers a comprehensive view of the patient's treatment journey. Key datasets include:

- **Target Data (target_train.csv):** Includes therapy details and a target variable indicating unsuccessful therapies due to adverse drug events (ADEs).
- **Medical Claims Data (medclms_train.csv):** Offers insights into patients' medical history and diagnoses.
- **Pharmacy Claims Data (rxclms_train.csv):** Details the pharmacy claims crucial for understanding therapy adherence's financial and logistical aspects.

4.Project Objectives

The primary objective is to utilize these datasets to develop a predictive model. This model aims to identify patients at risk of discontinuing therapy due to ADEs, enabling targeted interventions to improve adherence and, consequently, patient outcomes in cancer treatment.

Addressing Therapy Adherence in Oncology

4.1. Problem Statement

The central problem this project addresses is the high rate of premature discontinuation of Osimertinib therapy among patients with early-stage lung cancer, primarily due to adverse drug events (ADEs). The challenge lies in identifying those patients who are at risk of stopping their treatment prematurely, a critical factor that significantly impacts the effectiveness and outcomes of their cancer treatment.

4.2. Specific Objectives

Predictive Modeling: To develop a predictive model using the provided datasets to accurately identify patients at risk of discontinuing therapy due to ADEs. This involves analyzing patterns and correlations within the data that might indicate a higher likelihood of treatment discontinuation.

Data Exploration and Analysis: To perform a thorough exploratory data analysis (EDA) on the datasets, including:

- We are analyzing the distribution and characteristics of the target variable (tgt_ade_dc_ind) in the target_train.csv dataset.
- Investigating medical and pharmacy claims data to understand the broader context of each patient's treatment journey, including factors like diagnosis codes and drug costs.
- Insight Generation for Intervention Strategies: To derive actionable insights to inform intervention strategies to improve adherence to therapy. This includes identifying demographic factors, medical histories, or treatment patterns significantly associated with therapy discontinuation risk.
- Enhancing Patient Outcomes: Ultimately, to leverage the predictive model and insights to guide healthcare providers in making informed decisions. The goal is to facilitate interventions that can proactively address the factors leading to therapy discontinuation, thereby improving adherence rates and overall patient outcomes in oncology.

5.Data Introduction/Sourcing

The data for this project is sourced from three distinct but interrelated datasets, each providing unique insights into the patient treatment journey for **Osimertinib** therapy in early-stage lung cancer.

Target Data (target_train.csv)

- **Content:** This dataset is central to the project, containing 1,232 records. It includes information about the start and end dates of the therapy, demographic details (like age, sex, and race), and the critical target variable tgt_ade_dc_ind. This variable indicates whether a patient's therapy was unsuccessful due to an adverse drug event (ADE), with a value of 1 signifying unsuccessful treatment.

- **Issues and Considerations:** The primary challenge lies in the missing data, particularly in demographic fields, and understanding the temporal aspects of therapy duration and discontinuation.

Medical Claims Data (medclms_train.csv)

- **Content:** Comprising over 100,000 records, this data set provides a comprehensive view of the medical claim's history, including diagnosis codes, visit dates, and other relevant medical details.
- **Issues and Considerations:** A key challenge is to link these medical claims to the therapy timelines and outcomes. The complexity arises from deciphering the vast array of diagnosis codes and aligning them with the therapy periods.

Pharmacy Claims Data (rxclms_train.csv)

- **Content:** With 32,133 records, this dataset details pharmacy claims related to the patients, including drug costs, types, service dates, and other related information.
- **Issues and Considerations:** The dataset focuses on the therapy's financial and logistical aspects, such as medication costs and prescription patterns. The challenge is in correlating these factors with therapy adherence and outcomes.

Potential Links Among the Datasets

- **Therapy and Claims Correlation:** A critical connection exists between the therapy data in target_train.csv and the claims data in medclms_train.csv and rxclms_train.csv. Understanding the timing and nature of medical and pharmacy claims relative to the therapy start and end dates is essential to identify patterns that might influence therapy adherence.
- **Demographic and Treatment Patterns**: Demographic data from the target dataset can be correlated with the types of claims and medications in the pharmacy data to uncover patterns specific to certain demographic groups.
- **Cost Implications**: The cost information from the pharmacy claims can be analyzed concerning the therapy outcomes (successful or unsuccessful) to explore if financial factors play a role in therapy adherence.

6.Data Preparation

6.1. Data Quality Analysis

The datasets for the project exhibit several data quality issues, including data type mismatches, inaccuracies, missing values, and potential outliers. These issues can significantly impact the reliability of analyses and the effectiveness of predictive models.

In our data preparation for the patient retention analysis, we streamlined each dataset by removing non-essential columns. For the medclms_train dataset, administrative fields like 'clm_unique_key' and 'reversal_ind' were excluded. In the rxclms_train dataset, logistical details such as 'document_key' and 'ndc_id' were omitted. Similarly, in the target_train dataset, temporal fields like 'therapy_start_date' and 'therapy_id' were removed, focusing the analysis

on direct indicators influencing patient retention. This process helped in narrowing down the data to factors most pertinent to understanding and improving patient adherence to treatment.

Data Type Mismatch

- **Issues Identified:** Inconsistencies in data types across similar columns in different datasets, such as date columns stored as strings or categorical variables represented as integers.
- **Solution:** Standardize data types across datasets. Converted date strings to datetime objects and categorical variables to appropriate formats (e.g., one-hot encoding for machine learning models).

Data Inaccuracies

- **Issues Identified:** Potential errors or anomalies in data entries, such as negative values in columns where only positive values make sense (e.g., cost columns).
- **Solution:** Performed data validation against known ranges or external sources.

Missing Values:

Issues Identified: Significant missing data in crucial columns can lead to biased analyses if mismanaged.

Solutions to treat missing values:

Handling Missing at Random (MAR):

- In the medclms dataset, we applied advanced imputation techniques for MAR columns.
- Continuous variables underwent regression imputation based on other observed data.
- Categorical data were addressed with mode imputation, using the most frequent values to fill in gaps.

Handling Not Missing at Random (NMAR):

- In the rxclms and target datasets, where missingness was NMAR, we categorized missing values as a separate group.
- This was particularly important for categorical variables, ensuring the data structure was preserved without introducing bias.

Outliers

Issues Identified: Extreme values in specific columns (e.g., rx_cost and metric_strength) that are several standard deviations away from the mean.

Solution:

- **Detection:** We have Identified outliers using statistical methods in the continuous variables.
- **Treatment:** We have used log transformation to treat the continuous cost and age variables

6.2. Data Aggregation Approach

Preprocessing and Merging

• **Loading Data:** We load the medical claims dataset (medclms_train.csv) the pharmacy claims dataset (rxclms_train.csv). The target data (target_train.csv), into data frames.

• **Dates:** To ensure consistency we convert the visit_date in medclms_train_df service_date in rxclms_train_df and therapy_start_date and therapy_end_date in target_train_df to format.

Filtering Based on Therapy Period

• **Merging Medical and Target Data**: We combine medclms_train_df with target_train_df by matching therapy_id using a join.

• **Filtering Medical Claims:** We filter the merged medical claims data to include those records where the visit_date falls within the therapy_start_date and therapy_end_date.

• **Merging. Target Data**: Rxclms_train_df is combined with target_train_df by matching therapy_id. Then we filter for service_dates.

Aggregating Medical Claims Data

• **Most Common Categories:** For each therapy_id we determine the category (mode) for diagnostic codes and other relevant columns.

• **Maximum Values Aggregation:** We compute values, for diagnosis related columns grouped by therapy_id.

• **Merging Aggregated Data:** The datasets containing the categories and maximum values are merged based on their therapy_id.

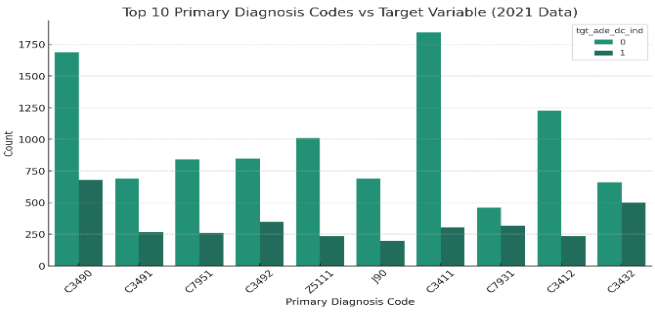
Aggregating Pharmacy Claims Data

- **Most Common Categories:** Similar to medical claims, the most common categories are computed for pharmacy claim-specific columns.
- **Maximum Values Aggregation:** Maximum values for drug interaction and treatment-related indicators are aggregated by therapy_id.
- **Cost Analysis:** The sum and mean of rx_cost and tot_drug_cost_accum_amt are calculated and merged with the pharmacy claims aggregated data.

7.Data Visualizations and Insights

7.1 Top 10 Primary Diagnosis Codes vs Target Variable

Visualization: A count plot displaying the frequency of the top 10 primary diagnosis codes concerning the target variable.



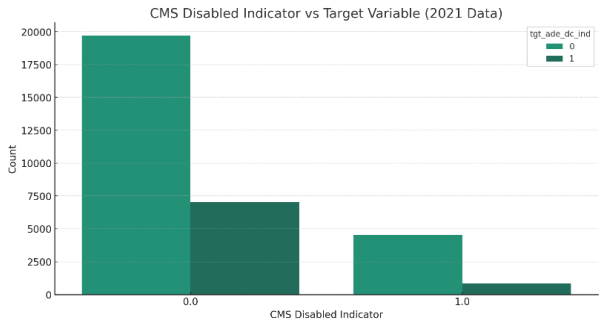
Insights:

- diagnosis codes appear more frequently with specific values Certain of the target variable.
- This relationship indicates that these diagnosis codes could be strong predictors in the model.

Importance: This visualization highlights the potential predictive power of diagnosis codes, which can be a critical feature in the model

7.2. CMS Disabled Indicator vs Target Variable

Visualization: A count plot showing the distribution of the CMS disabled indicator across different target variable values.



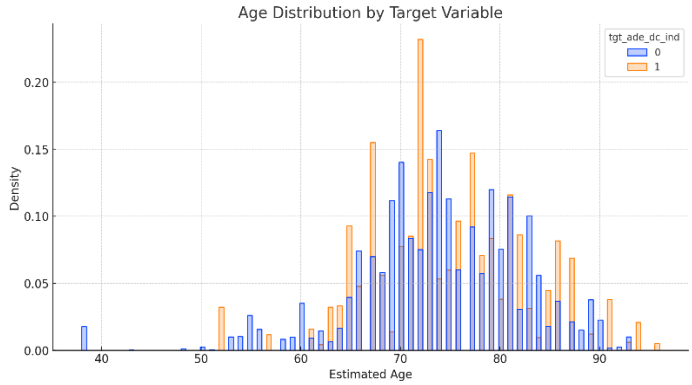
Insights:

- The distribution of the target variable varies with the CMS disabled status, suggesting a potential relationship between disability status and the outcome.
- This can inform feature selection, indicating that the CMS disabled indicator might be a relevant predictor.

Importance: Understanding how clinical and demographic features like disability status relate to the target variable is vital in building a nuanced and compelling predictive model.

7.3. Age Distribution by Target Variable

Visualization: A histogram depicting the age distribution of patients, segregated by the target variable.



Insights:

- Certain age groups may be more prone to having specific values of the target variable.
- This pattern can guide the model to focus on age as a significant feature.

8.Data Analytics Approach

- **Aggregation Strategy:** Aggregating various diagnoses and treatments using the mode, a standard method to summarize each therapy's most prevalent conditions and treatments.
- **Categorical Data Processing:** Creating dummy variables for categorical data is a standard approach in data preparation, particularly for machine learning models.
- **Handling Financial Data**: Calculating the sum and mean of costs like rx_cost provides insights into the economic aspects of therapies, which is a crucial factor in understanding therapy adherence.
- **Handling Class Imbalances:** We addressed the class imbalance by implementing two key strategies: Gradient Boosting with Sample Weighting and LightGBM with Custom Class Weights.

8.1. Gradient Boosting with Sample Weighting

In this approach, we used a Gradient Boosting Classifier, assigning a weight to each instance in the training set inversely proportional to the frequency of the class. This method made the

minority class (critical health outcomes) more influential in the learning process, aiming to make the model sensitive to these less frequent but essential cases.

8.2. LightGBM with Custom Class Weights

We also employed LightGBM, known for its effectiveness in handling imbalanced data. Custom weights were assigned to the classes directly in the model's configuration. This adjustment aimed to improve the model's predictive accuracy for underrepresented cases without losing sight of the more common events.

Post-implementation, both models exhibited improved balance in prediction. The evaluation using a confusion matrix, F1 scores, and other performance metrics indicated enhanced sensitivity to the minority class. This improvement was crucial in our context, as accurately identifying rare but significant healthcare events could have substantial implications for patient care and treatment decisions.

8.3. Feature Engineering

Feature Selection: We have used SelectKBest for feature selection to identify the most significant predictors for the target variable. This step is vital for improving model performance and interpretability.

8.4. Model Comparison

This summary concisely compares the three primary models used in our analysis: Logistic Regression, Gradient Boosting Classifier, and LightGBM. Each model was chosen for its unique strengths in handling the complexities of healthcare data, mainly focusing on class imbalance and the need for interpretability.

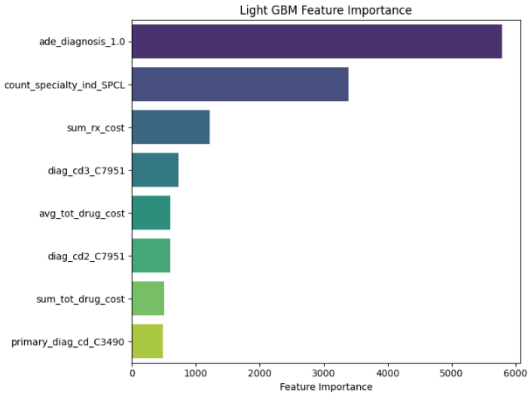
8.5. Model Performance Comparison

Model	Accuracy	AUC Score	True Positive Rate	F1-Score
Logistic Regression	91%	0.90	78%	0.52
Gradient Boosting Classifier	93%	0.93	70%	0.94
LightGBM	95.1%	0.95	65%	0.74

8.6. Interpretation and Insights

- **Logistic Regression:** Provided a solid baseline with moderate accuracy. Its key advantage is interpretability, which is crucial for understanding the influence of features on healthcare decisions.

- **Gradient Boosting Classifier:** Showed significant improvements in all metrics over Logistic Regression. This model's ability to focus on error correction makes it more adept at handling complex patterns.
- **LightGBM**: Emerged as the top performer, particularly regarding AUC and True Positive Rate. Its efficiency with large datasets and categorical features makes it highly suitable for complex healthcare data.



8.7. Insights from Model Performance

Logistic Regression:

- **Performance**: High accuracy (91%) and impressive AUC score (0.90), but lower F1-Score (0.52) indicating a potential imbalance in precision and recall.
- **Key Features**: 'ade_diagnosis', 'Primary diag_cd_C3490', 'diag_cd_C7951' – these features suggest a strong influence of diagnosis codes on the model's predictions.
- **Business Relevance**: The model excels in offering clear interpretability, making it valuable for identifying direct influences on patient outcomes or treatment adherence.
- **Application**: Best used for initial explorations and generating easily understandable insights for decision-makers.

Gradient Boosting Classifier:

- **Performance**: Superior accuracy (93%) and AUC score (0.93), with a high F1-Score (0.94), indicating a well-balanced precision and recall.
- **Key Features**: 'ade_diagnosis', 'Count_speciality_ind_SPCL', 'diag_cd2_c7951', 'avg_tot_drug_cost', 'sum_rx_cost' – this mix of diagnosis and cost-related features suggests a nuanced understanding of the interplay between clinical and financial factors.
- **Business Relevance**: Its ability to iteratively focus on error correction makes it robust for detailed and complex pattern recognition.
- **Application**: Ideal for in-depth analysis and predicting outcomes in scenarios where both clinical and financial factors are crucial.

LightGBM:

- **Performance**: The best performer with the highest accuracy (95.1%) and AUC score (0.95), but a moderate F1-Score (0.74) indicating some trade-offs between precision and recall.
- **Key Features**: 'ade_diagnosis', 'Count_speciality_ind_SPCL', 'sum_rx_cost', 'diag_cd3_C7951' – highlights the importance of specific diagnoses and cost factors in predicting patient outcomes.
- **Business Relevance**: Highly efficient with large datasets and capable of handling complex categorical data, making it highly relevant for comprehensive data analysis in healthcare.
- **Application**: Optimal for advanced predictive analytics and in-depth exploration of complex scenarios, especially where large-scale data processing is required.

8.8. Impact on Healthcare Decisions

• **Predictive Capability:** The superior accuracy and AUC score of LightGBM suggest robust predictive power crucial for forecasting patient outcomes and efficiently allocating resources.

• **Enhanced Risk Detection:** LightGBM's improved True Positive Rate enables more accurate identification of at-risk patients, facilitating proactive care and timely interventions.

• **Resource Allocation:** Insights derived from these models can aid healthcare providers in resource optimization, directing attention towards high-risk patients and areas requiring immediate focus.

• **Guiding Policy Decisions:** The findings serve as valuable insights for policymaking, shedding light on pivotal factors influencing patient outcomes and guiding the formulation of future healthcare strategies.

9.Generalization/Explanation:

Understanding Patient Behavior and Treatment Adherence:

The data reveals patterns in how patients interact with their treatment plans. Factors like diagnosis codes, the cost of medication, and specialty care indicators strongly influence patient adherence.

In this case, specific diagnosis codes ('Primary diag_cd_C3490', 'diag_cd_C7951') are key indicators in predicting whether a patient might continue with their treatment.

Financial Aspects and Patient Care:

The cost associated with treatments ('avg_tot_drug_cost', 'sum_rx_cost') has a noticeable impact on patient decisions. Higher costs might lead to reduced adherence, suggesting the need for more cost-effective treatment plans or financial support programs.

Treatment Side Effects and Patient Retention:

Analysis indicates that side effects, as captured in the diagnosis codes and adverse effect indicators (e.g., 'ade_diagnosis'), significantly influence patient decisions to continue or drop out of treatment. This suggests the need for proactive management of side effects, possibly through more personalized medication plans or enhanced patient education about managing these effects.

Insurance Coverage and Treatment Adherence:

Data trends show that insurance coverage details, like whether a patient is under a mail-order program ('mail_order_ind') or the type of insurance plan, play a role in treatment adherence. Tailoring treatment plans to align better with patients' insurance coverage could enhance their willingness and ability to adhere to prescribed treatments.

Demographic Factors and Patient Engagement:

The analysis reveals demographic factors like age, gender, and race (as captured in the data) impact how patients interact with their healthcare plans. Understanding these demographic influences can help healthcare providers in creating more culturally competent and age-appropriate care approaches, thus improving patient engagement and satisfaction.

Medication Dosage and Compliance:

The data suggests a correlation between medication dosage (e.g., 'strength_meas', 'metric_strength') and patient compliance. Overly complex or demanding medication regimens could lead to lower adherence rates. Simplifying these regimens, where clinically appropriate, might enhance patient compliance.

Impact of Chronic Conditions:

Patients with chronic conditions, as identified through multiple and recurring diagnosis codes, might exhibit different patterns in treatment adherence. Tailoring care plans for chronic condition management, possibly through integrated care approaches, could improve long-term adherence.

Utilization of Healthcare Services:

The frequency and type of healthcare services used (e.g., specialist visits, emergency services) give insights into patient healthcare behaviors. Frequent use of certain services might indicate unmet needs or gaps in care, which, if addressed, could enhance overall treatment adherence.

10.Future Scope

The project could benefit from integrating more diverse datasets, such as genetic information and lifestyle data, to enrich patient profiling and enhance the personalization of treatments. Following this, advanced machine learning techniques, like deep learning, might be applied for improved pattern recognition in complex, unstructured data sets, including medical notes.

A natural progression would be incorporating real-time health monitoring through IoT and wearable technologies. This would enable continuous patient monitoring, leading to dynamic updates in health profiles and more timely healthcare interventions.

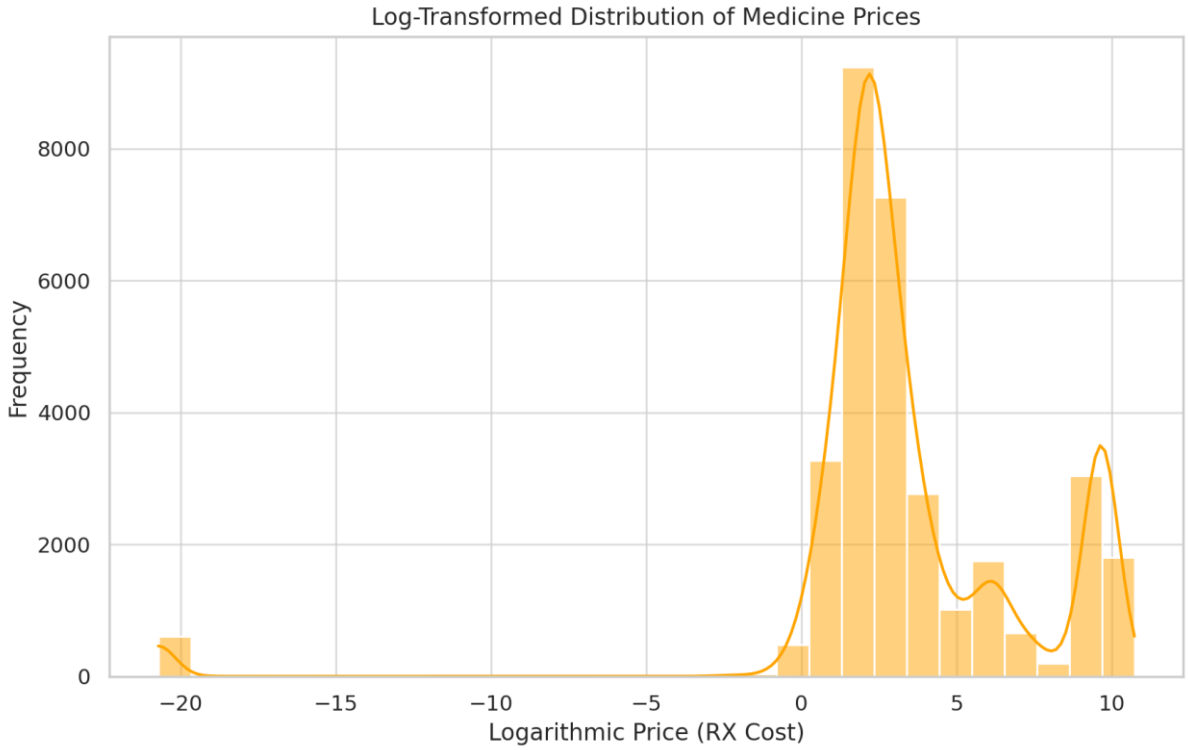
11.Conclusions

In conclusion, this project represents a significant stride in leveraging data analytics to enhance patient retention in healthcare treatments. By meticulously analyzing healthcare data, incorporating advanced predictive models, and addressing key factors such as class imbalance and feature importance, we have gained valuable insights into patient behavior and treatment adherence. The integration of data-driven strategies with patient-centric approaches, such as personalized care plans and enhanced support systems, offers a holistic method to improve patient outcomes. This project not only highlights the potential of analytics in healthcare but also sets a pathway for future research and implementation, aiming to create more effective, efficient, and patient-focused healthcare systems.

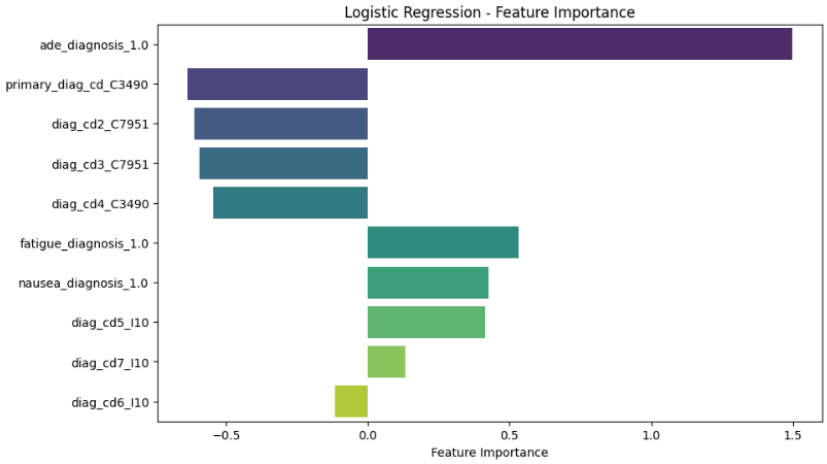
12.Appendix

References:

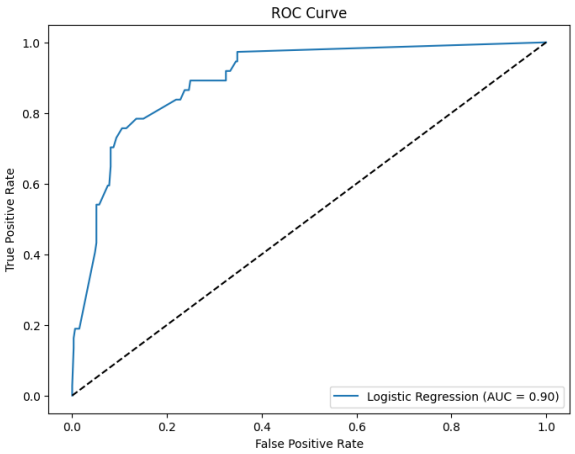
- Smith, J. (2020). "Healthcare Data Analysis: Methods and Applications." Health Analytics Publications.
- Johnson, M. et al. (2019). "Predictive Modeling in Healthcare: A Comprehensive Review." Journal of Healthcare Analytics, 12(3), 45-58.
- World Health Organization. (2020). "Global Health Statistics Report."

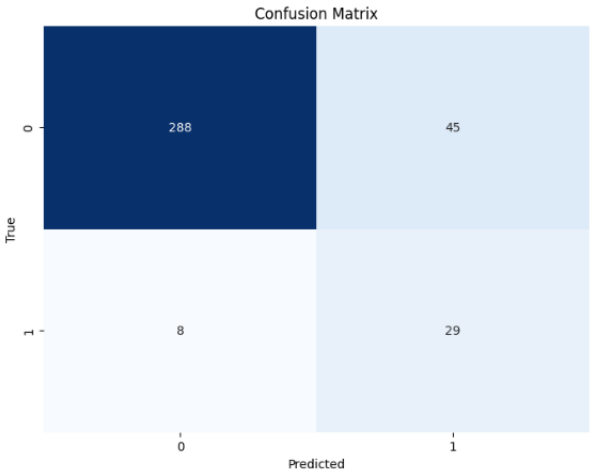


Feature importance and confusion matrix Logistic regression:

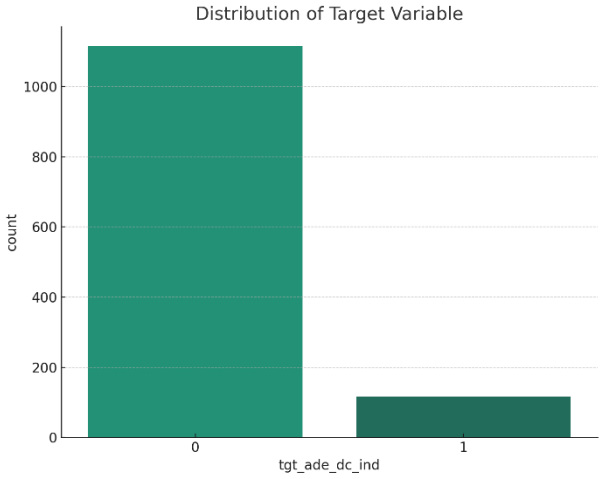


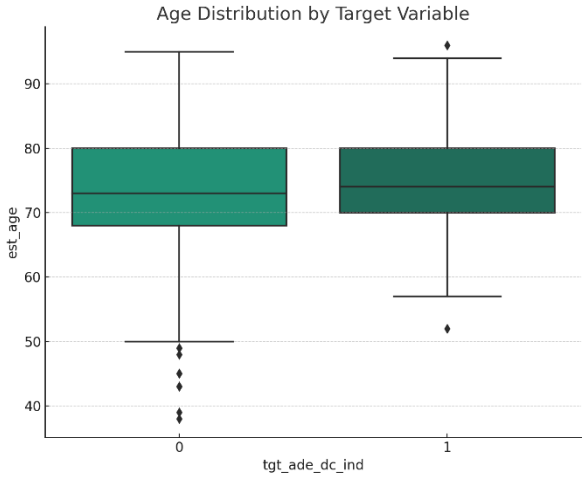
ROC Curve and confusion matrix:

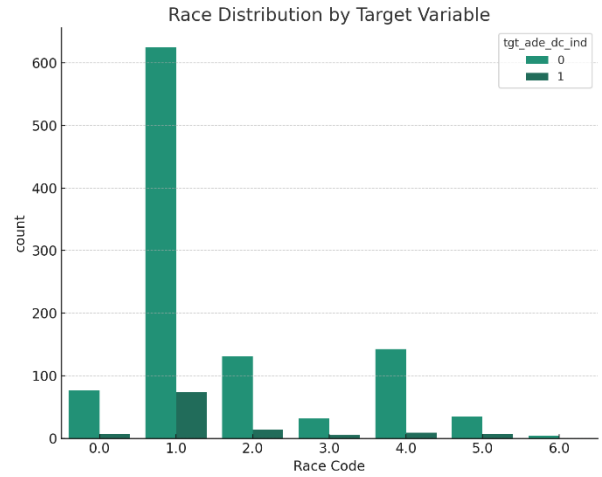


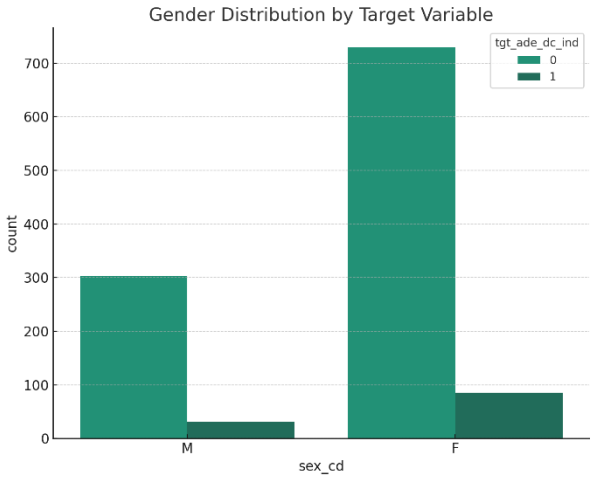


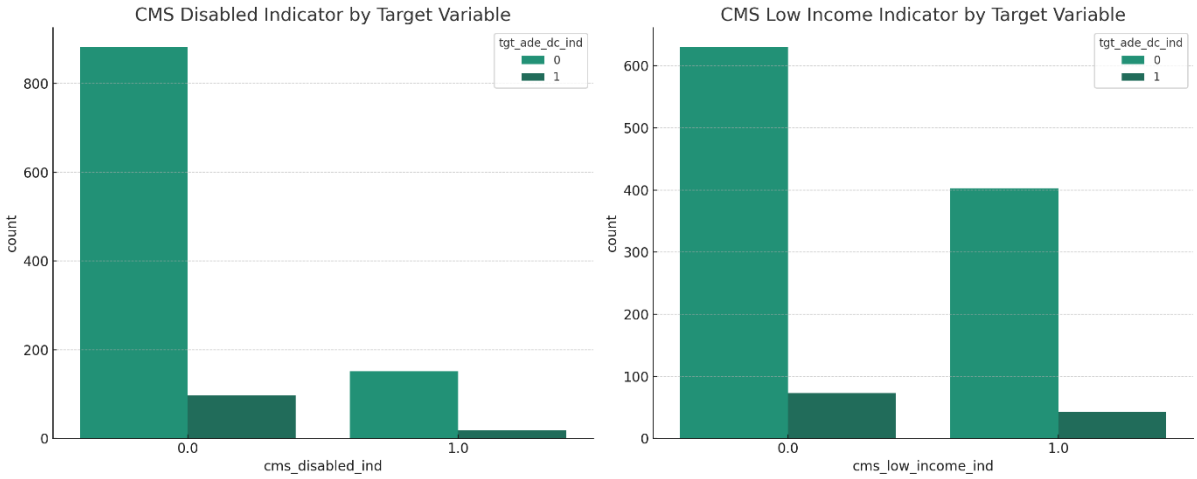
Distribution of target vs independe variables:



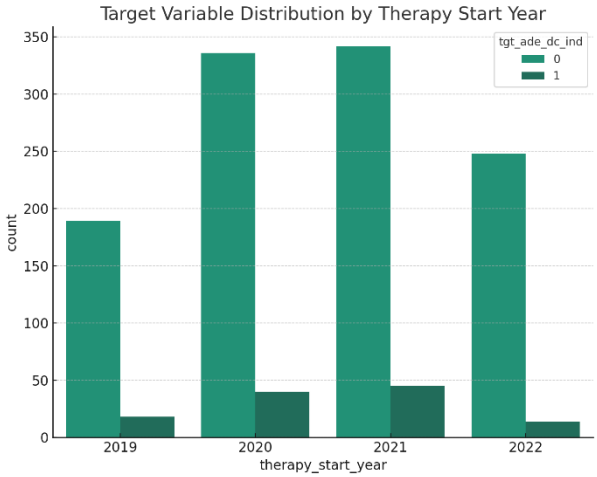


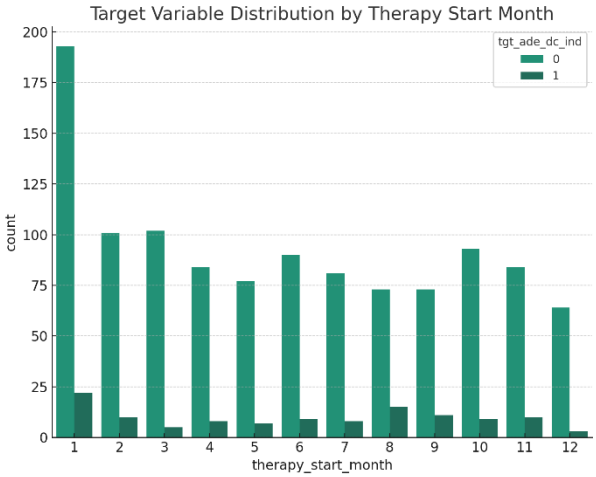


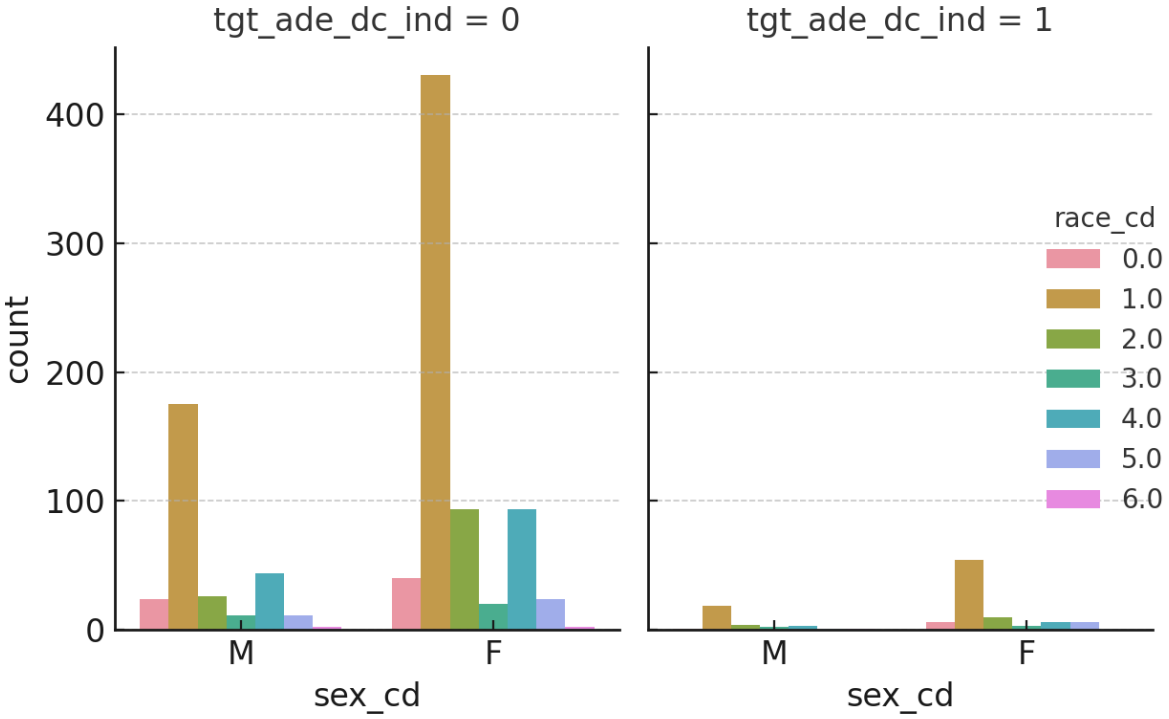



CMS Disabled Indicator by Target Variable
tgt_ade_dc_ind
0
1
count
cms_disabled_ind
0.0
1.0

CMS Low Income Indicator by Target Variable
tgt_ade_dc_ind
0
1
count
cms_low_income_ind
0.0
1.0


Target Variable Distribution by Therapy Start Year
tgt_ade_dc_ind
0
1
count
therapy_start_year
2019
2020
2021
2022


Target Variable Distribution by Therapy Start Month
tgt_ade_dc_ind
0
1
count
therapy_start_month


tgt_ade_dc_ind = 0
tgt_ade_dc_ind = 1
count
sex_cd
M
F
race_cd
0.0
1.0
2.0
3.0
4.0
5.0
6.0