HUMANA-MAYS
ANALYTICS CASE COMPETITION
2023

Data-Driven Care: Predicting and Improving Osimertinib Adherence

Executive Summary

Problem Statement and Opportunity

As per the World Health Organization, an alarming statistic reveals that only about half, or approximately 50 percent of individuals with chronic diseases adhere to their prescribed medication regimens. (Sabate, E) The impact of non-adherence extends far beyond individual health, encompassing substantial avoidable healthcare costs in the United States, estimated to range between $100 billion to $300 billion annually. (Iuga and McGuire) The case addresses the persistent challenge of therapy discontinuation due to manageable side effects associated with Osimertinib, a potentially life-saving medication for early-stage lung cancer patients with a specific genetic mutation (EGFR). The opportunity lies in leveraging data analytics to predict therapy discontinuation, allowing for targeted interventions to enhance medication adherence and ultimately improve patient survival rates and quality of life.

Modeling Target

The primary objective of this challenge is to develop a predictive model identifying therapies at risk of premature discontinuation, specifically due to reported adverse drug events (ADEs). In order to construct an accurate and actionable model we preprocessed and harmonized our data which involved data cleaning and imputation, feature engineering, feature selection, principal component analysis. We then performed model selection and validation and fine-tuned our model using GridSearch. Our final Extreme Gradient Boosting model achieved an AUC ROC score of 0.8247 and a Fairness Disparity score 0.9788 on the holdout dataset. Our model provided various insights which we translated into business recommendations for Humana to implement.

Recommendations

We were able to identify the primary drivers behind members dropping out of the Osimertinib therapy on the basis of which we developed data-driven personalized solutions in order to improve the adherence rate for this therapy. Our solutions include In-Application Tracking and Alerting Notifications, 1:1 Mentoring, Establishing Support Groups, Periodic Follow Ups, Incentivized Premium Plans and Personalized Communication Channels. These actionable solutions will help Humana drive the adherence rate for the Osimertinib therapy and help their members live longer fulfilling lives.

Table Of Content

Executive Summary **2**
Table Of Content **3**
1. Introduction **5**
1.1 Background 5
1.2 The Humana-Mays Analytics Case Competition 6
1.2.1 The Business Issue 6
1.2.3 Key Performance Indicators 7
2. Preliminary Research **10**
3. Data Analysis **12**
3.1 Dataset Description 12
3.2 External Data 15
3.3 Exploratory Data Analysis 17
4. Data Preprocessing **23**
4.1 Medical Claims Data Transformation 23
4.2 Pharmacy Claims Data Transformation 25
4.3 Merged Data Transformation 26
4.3.1 Missing Data Imputation 26
4.3.2 One Hot Encoding 26
4.3.3 Checking and Correcting for Skewness 27
4.3.4 Outlier Detection using Principal Component Analysis (PCA) 28
4.3.5 Feature Selection using Correlation Plot/Matrix 29
4.3.6 Scaling of Data 30
4.3.7 Dataset Resampling 30
5. Predictive Model Development **32**
5.1 Hyperparameters Tuning 32
5.2 Final Model Construction 33
5.3 Key Performance Indicators Analysis 35
5.4 Relationship among features 38
5.5 Data Biases 39
6. Results and Findings **40**
7. Recommendations **41**
7.1 Interpretation of Results, Classification of patients based on Likelihood Score 41
7.2 Prioritizing Early Intervention and Patient Awareness in Osimertinib Treatment 42
7.3 Recommended Business Strategies 43
7.3.1 1:1 Mentoring 44
7.3.2 Support Groups 45
7.3.3 Incentivized Premium Plans 46
7.3.4 Educational Webinars 47
7.3.5 ALS Emergency Care 48

7.3.6 Periodic Follow-Up..... 49
7.3.7 In app tracking and alerts..... 50
8. Potential Impact On Business..... 51
9. Cost Benefit Analysis..... 53
10. Scope for Improvement..... 54
11. References..... 57

1. Introduction

1.1 Background

a. An overview on Lung Cancer

Reports show that every 2.5 minutes, somebody in the United States is diagnosed with lung cancer and the daily deaths due to this disease are as high as 365. ("State of Lung Cancer | Key Findings") The disease is primarily associated with long-term exposure to harmful substances such as tobacco smoke, environmental pollutants, and certain occupational hazards.

Lung cancer can be broadly categorized into two main types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC is the most common form, comprising about 85% of all lung cancer cases. SCLC is less common but tends to grow and spread rapidly.

The development of lung cancer is a complex interplay of genetic, environmental, and lifestyle factors. Smoking is the leading cause of lung cancer, accounting for the majority of cases. However, non-smokers can also develop lung cancer due to exposure to secondhand smoke, radon, asbestos, air pollution, or a family history of the disease. Early-stage lung cancer often presents minimal or no symptoms, making it difficult to diagnose in its initial stages.

b. Key Statistics on Lung Cancer ("Lung Cancer Statistics | How Common is Lung Cancer?")

- Lung cancer (both small cell and non-small cell) is the second most common cancer in both men and women in the United States.
- The American Cancer Society's estimates for lung cancer in the US for 2023 are:
  About 238,340 new cases of lung cancer (117,550 in men and 120,790 in women)
  About 127,070 deaths from lung cancer (67,160 in men and 59,910 in women)
- Lung cancer mainly occurs in older people. Most people diagnosed with lung cancer are 65 or older; a very small number of people diagnosed are younger than 45. The average age of people when diagnosed is about 70.
- Lung cancer is by far the leading cause of cancer death in the US, accounting for about 1 in 5 of all cancer deaths. Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined.
- On a positive note, the number of new lung cancer cases continues to decrease, partly because more people are quitting smoking (or not starting). The number of deaths from lung cancer continues to drop as well, due to fewer people smoking and advances in early detection and treatment.

c. Treatment of Non-Small Cell Lung Cancer

The treatment options for non-small cell lung cancer (NSCLC) are based mainly on the stage (extent) of the cancer, but other factors, such as a person's overall health and lung function, as well as certain traits of the cancer itself, are also important. In many cases, more than one type of treatment is used.

d. Targeted Therapy for Non-Small Cell Lung Cancer

As researchers have learned more about the changes in non-small cell lung cancer (NSCLC) cells that help them grow, they have developed drugs to specifically target these changes. Targeted drugs work differently from standard chemotherapy drugs. They sometimes work when chemo drugs don't, and they often have different side effects. At this time, targeted drugs are most often used for advanced lung cancers, either along with chemo or by themselves.

1.2 The Humana-Mays Analytics Case Competition

1.2.1 The Business Issue

a. Osimertinib: A Targeted Therapy for Lung Cancer

Osimertinib is a targeted therapy medication that has shown to be highly effective in treating patients with early-stage non-small cell lung cancer (NSCLC) with a specific targetable mutation known as EGFR. Studies have shown that osimertinib can double the likelihood of survival compared to standard chemotherapy, and patients who adhere to the prescribed regimen are 80% less likely to experience a recurrence of cancer. ("Treatment for Early-Stage EGFR+ NSCLC | TAGRISSO® (osimertinib)")

However, osimertinib can also cause a variety of side effects, including nausea, fatigue, pain, high blood glucose, and constipation. These side effects can make it difficult for patients to tolerate the medication, and many people choose to discontinue treatment altogether. ("About TAGRISSO® (osimertinib) for Early-Stage EGFR+ NSCLC")

In a recent study conducted by Humana, approximately one quarter of members undergoing osimertinib treatment experienced side effects and discontinued therapy within the initial six months. This highlights the critical need to identify and support patients who are at risk of discontinuing treatment.

b. The Problem with Medical Adherence Among Patients with Chronic Diseases

As per the World Health Organization, an alarming statistic reveals that only about half, or approximately 50%, of individuals with chronic diseases adhere to their prescribed medication regimens. (Sabate, E) This figure is deeply concerning when considering the potentially fatal

implications of chronic diseases and the lives at risk. Non-adherence to prescribed medications not only poses a significant threat to an individual's health and personal well-being but also exerts a substantial burden on the healthcare system as a whole.

In fact, the impact of non-adherence extends far beyond individual health, encompassing substantial avoidable healthcare costs in the United States, estimated to range between $100 billion to $300 billion annually. (Iuga and McGuire) These costs are attributed to preventable health complications stemming from patients not adhering to their medication plans. Additionally, non-adherence is associated with a tragically high loss of tens of thousands of lives each year, underlining the urgent need for effective strategies to improve medication adherence and promote better health outcomes.

c. Using Data and Analytics to Improve Medication Adherence

Humana has presented a challenge to construct a robust prediction model to anticipate the likelihood of Humana members discontinuing osimertinib therapy following a side effect occurrence. This predictive tool is instrumental in allowing Humana to proactively target these members early in their therapy journey, offering tailored support and guidance.

By identifying and intervening early, Humana aims to enhance members' ability to complete the osimertinib therapy successfully, ultimately improving their chances of surviving lung cancer and leading fulfilling lives. This model signifies a significant step forward in personalized healthcare, demonstrating Humana's commitment to providing the highest level of care and support to its members facing this critical health challenge.

1.2.3 Key Performance Indicators

Our main focus is to build a robust and accurate predictive model, find insights based on the data and research and provide actionable recommendations to solve the challenge of discontinuing Osimertinib therapies.

Our Key Performance Indicators for aligned to with our insights and recommendations are:

1. Patient Outcome Score (POS):

Importance: This metric will concentrate on the percentage of people achieving a positive outcome like disease free periods or prolonged survival.

$$\text{POS} = \frac{\text{Total Positive Patient Outcomes} \times 100}{\text{Total Osimertinib Patient Population}}$$

Business Significance: This will help key stakeholders realize the effectiveness of Osimertinib towards generating positive outcomes. Stakeholders can leverage this to tailor strategies to influence better experience and outcomes.

2. Patient Compliance Score (PCS):
Importance: This metric will provide a high level overview of the number of patients complying with and following the Osimertinib treatment plan as proposed.

$$PCS = \frac{\text{Total Compliant Patients}}{\text{Total Osimertinib Patient Population}} \times 100$$

Business Significance: This will help key stakeholders realize what proportion of the patient population follows the prescribed plan. It will aid them in identifying compliance patterns allowing them for targeted interventions to enhance patient engagement and minimize discontinuity.

3. Post-Progression Resilience Ratio (PPRR):
Importance: This metric emphasizes the patient's resilience following disease progression, offering a clear vision into continued therapy post condition worsening.

$$PPRR = \frac{\text{Patients Adherence Post Progression}}{\text{Total Patients with Disease Progression}} \times 100$$

Business Significance: This will guide oncologists and care teams to enhance and tailor post progression interventions, thereby improving survival and quality of life for patients.

4. Discontinuity Alert Score (DPS):
Importance: A score based on predictive modeling on potential discontinuity of a particular patient from Osimertinib treatment.
Business Significance: This is a proactive approach to patient management to prevent discontinuity by reducing the likelihood of adverse events and optimizing patient care.

5. Therapy Harmony Score (THS):
Importance: This metric will provide a compatibility index of Osimertinib with any ongoing prescribed medications for the patients.

Business Significance: This will not only address concerns regarding potential drug complications but also augment therapeutic harmony. This will also help in enhancing patient safety by proactively managing complications which in turn improves patient satisfaction.

6. Therapy Continuity Index (TCI):

Importance: This metric provides a holistic measure involving Time-to-Discontinuity, Post-Progression Survival, and Patient Adherence and portrays likelihood of the patient continuing therapy.

$$TCI = \frac{\text{Actual Treatment Duration} \times 100}{\text{Proposed Treatment Duration}}$$

Business Significance: Such a comprehensive measure will guide stakeholders to execute strategies that influence long term outcomes with an aim for sustained patient engagement and better therapy continuity.

2. Preliminary Research

Osimertinib, a third-generation epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor (TKI), was approved by the FDA in 2018 due to its demonstrated superiority over its predecessors in certain aspects. After a number of trials, it came out as a promising therapeutic option for patients with advanced non-small cell lung cancer (NSCLC) harboring EGFR-activating mutations. Two trials that particularly showed the superiority and effectiveness of Osimertinib were FLAURA and ADAURA.

The FLAURA trial was a phase III, double-blind trial that looked at first-line osimertinib in patients with advanced NSCLC and EGFR mutations. It tested whether osimertinib was more effective as an initial treatment than the earlier-generation EGFR inhibitors (gefitinib or erlotinib). ("Osimertinib in Advanced Lung Cancer with EGFR Mutations")

The ADAURA trial is a randomized, double-blind, placebo-controlled, global phase III trial. It tested the efficacy and safety of osimertinib (Tagrisso) versus placebo in patients with completely resected stage IB-IIIA non-small-cell lung cancer (NSCLC). ("Osimertinib After Surgery Significantly Improves Survival in Patients With Resected EGFR-Mutated Non-Small Cell Lung Cancer").

2.1 Key findings from the FLAURA trial:

Metric	**Osimertinib group**	**Gefitinib/Erlotinib group**
Median Overall Survival	38.6 months	31.8 months
Treatment Continuation Post 3 Years	28%	9%
Discontinuation Due To Side Effects	15%	18%

2.2 Key findings from the ADAURA trial: (Nyberg)

Metric	**Osimertinib group**	**Placebo group**
Median Disease-Free Survival	65.8 months	21.9 months
4-Year Rates	70%	29%
Completion of Planned 3-Year Treatment	66%	41%

Metric	Osimertinib group	Placebo group
Subsequent Anticancer Treatment	22%	54%

2.3 Challenges:

Although osimertinib offers considerable benefits in treating non-small cell lung cancer (NSCLC) with EGFR mutations, it does come with some side effects. These commonly include issues like diarrhea, paronychia (a nail disorder), dry skin, and itching. In more severe cases (grade 3 or above), patients may experience serious problems such as persistent diarrhea, mouth sores (stomatitis), pneumonia, cardiac issues, QT prolongation and lung-related issues like inflammation (pneumonitis). ("Deconstructing ADAURA. It is Not Yet Time to Forgo Platinum-based Adjuvant Chemotherapy in Resected Early Stage (IB-IIIA) EGFR-mutant NSCLC")

These studies gave us a better understanding into the effectiveness of osimertinib, its benefits over traditional treatments and potential side effects. Armed with this knowledge, we could delve deep into our analysis in a more focussed direction and figure out the factors influencing patient adherence, exploring innovative strategies to mitigate discontinuity, and ultimately aiming to enhance the overall patient experience with osimertinib therapy.

3. Data Analysis

3.1 Dataset Description

Humana provided data centered around the Osimertinib therapy for its members which was split into training and holdout data. The training and holdout datasets contained the exact same information with the exception of the target data. There were three primary data groups:

1. Target Data: This comprised the target_train (1232 records) and target_holdout (420 records) datasets. It contains sensitive information about Humana members such as their age, sex, race, etc. along with their Osimertinib therapy information such as therapy start date and end date. The only difference between the target_train and target_holdout datasets is that the training dataset contains a column 'tgt_ade_dc_ind' indicating if the member discontinued the therapy prematurely. This column is obviously not present in the holdout data. These datasets are on a member-therapy granularity, i.e, each record is unique on member ID and therapy ID.
2. Medical Claim Data: This comprised the medclms_train (100159 records) and medclms_holdout (23232 records) datasets. This contains simplified information about the medical claims made by a member 90 days prior to their Osimertinib therapy and through the end of the therapy. This data includes visit and process dates, diagnosis codes and indicators for diagnosis codes of interest. It is unique on the medical claim ID.
3. Pharmacy Claim Data: This comprised the rxclms_train (32133 records) and rxclms_holdout (6670 records) datasets. It contains simplified information about all pharmacy claims for an individual during the time 90 days before their Osimertinib therapy and through the end of therapy. This data includes service and process dates, drug identifier codes (NDC) and indicators for drug codes of interest. It is unique on the pharmacy claim ID.

Dataset	Column Name	Description	Data Type
target dataset	id	Person Identifier - unique for a member	Unique Identifier (String)
	therapy_start_date	The date of the member's first fill of Tagrisso.	Datetime
	therapy_end_date	The date the member runs out of their supply of tagrisso. OR six months after therapy_start_date. Only available in the training data	Datetime
	tgt_ade_dc_ind	An indicator for whether this person meets the target criteria of reporting an ADE and discontinuing therapy before 6 months. Only available in training data	Categorical Nominal (Integer)

	race_cd	a numeric indicator for race	Categorical Nominal (Integer)
	est_age	The member's estimated age	Quantitative Discrete (Integer)
	sex_cd	Indicates the member's sex	Categorical Nominal (Integer)
	cms_disabled_ind	indicates if the member is classified as disabled by CMS	Categorical Nominal (Integer)
	cms_low_income_ind	indicates if the member receives low income subsidies from CMS	Categorical Nominal (Integer)
all datasets	therapy_id	therapy identifier - concatenation of sdr_person_id, drug name, and therapy number	Unique Identifier (String)
medical claims dataset	medclm_key	indicator key for a medical claim	Unique Identifier (String)
	primary_diag_cd	The primary diagnosis code for this claim in the ICD-10 format. Lookup available online.	Categorical Nominal (String)
	visit_date	The date of the medical visit	Datetime
	diag_cd#	non-primary diagnosis codes for a medical claim. Each claim has space for up to 8 non-primary diagnosis codes in the ICD-10 format. Lookup available online.	Categorical Nominal (String)
	pot	place of treatment for this claim	Categorical Nominal (String)
	util_cat	Combination of admit_type and pot for use in creating utilization categories	Categorical Nominal (String)
	heids_pot	Uses Healthcare Effectiveness Data and Information Set Place of Treatment (HEDIS) ValueSets to label various place of treatment descriptions	Categorical Nominal (String)
	ade_diagnosis	Indicates if the diagnosis codes in this claim report an adverse drug event (ADE)	Categorical Nominal (Integer)
	seizure_diagnosis	Indicates if the diagnosis codes in this claim report seizures	Categorical Nominal (Integer)
	pain_diagnosis	Indicates if the diagnosis codes in this claim report pain	Categorical Nominal (Integer)
	fatigue_diagnosis	Indicates if the diagnosis codes in this claim report fatigue	Categorical Nominal (Integer)

	nausea_diagnosis	Indicates if the diagnosis codes in this claim report nausea	Categorical Nominal (Integer)
	hyperglycemia_diagnosis	Indicates if the diagnosis codes in this claim report hyperglycemia	Categorical Nominal (Integer)
	constipation_diagnosis	Indicates if the diagnosis codes in this claim report constipation	Categorical Nominal (Integer)
	diarrhea_diagnosis	Indicates if the diagnosis codes in this claim report diarrhea	Categorical Nominal (Integer)
pharmacy claims dataset	document_key	unique identifier for a prescription claim document	Unique Identifier (String)
	ndc_id	National Drug Code Identifier: a national/FDA identifier for a specific drug.	Categorical Nominal (String)
	service_date	Date of a prescription fill	Datetime
	pay_day_supply_cnt	The number of days supply of a drug	Categorical Nominal (Integer)
	rx_cost	The cost of the prescription	Quantitative Discrete (Integer)
	tot_drug_cost_accum_amt	The cumulative cost amount for a member year-to-date	Quantitative Discrete (Integer)
	mail_order_ind	Indicates whether this prescription was filled with the mail-order pharmacy	Categorical Nominal (Integer)
	generic_ind	indicates whether this drug is branded or generic	Categorical Nominal (Integer)
	maint_ind	indicates whether this drug is a maintenance or non maintenance drug	Categorical Nominal (Integer)
	gpi_drug_class_desc	Generic Product Identifier drug class description	Categorical Nominal (String)
	gpi_drug_group_desc	Generic Product Identifier drug group description	Categorical Nominal (String)
	hum_drug_class_desc	Humana Drug Class Description	Categorical Nominal (String)
	strength_meas	the unit of measure for the drug filled in this claim	Categorical Nominal (String)

	metric_strength	The metric strength of the drug filled in this claim	Categorical Nominal (Integer)
	specialty_ind	Indicates whether this claim is for a specialty drug	Categorical Nominal (String)
	ddi_ind	Indicates if this claim is for a drug with a know interaction with Tagrisso	Categorical Nominal (Integer)
	anticoag_ind	Indicates if this claim is for an anticoagulant	Categorical Nominal (Integer)
	diarrhea_treat_ind	indicates if this claim is for a drug used to treat diarrhea	Categorical Nominal (Integer)
	nausea_treat_ind	indicates if this claim is for a drug used to treat nausea	Categorical Nominal (Integer)
	seizure_treat_ind	indicates if this claim is for a drug used to treat seizures	Categorical Nominal (Integer)

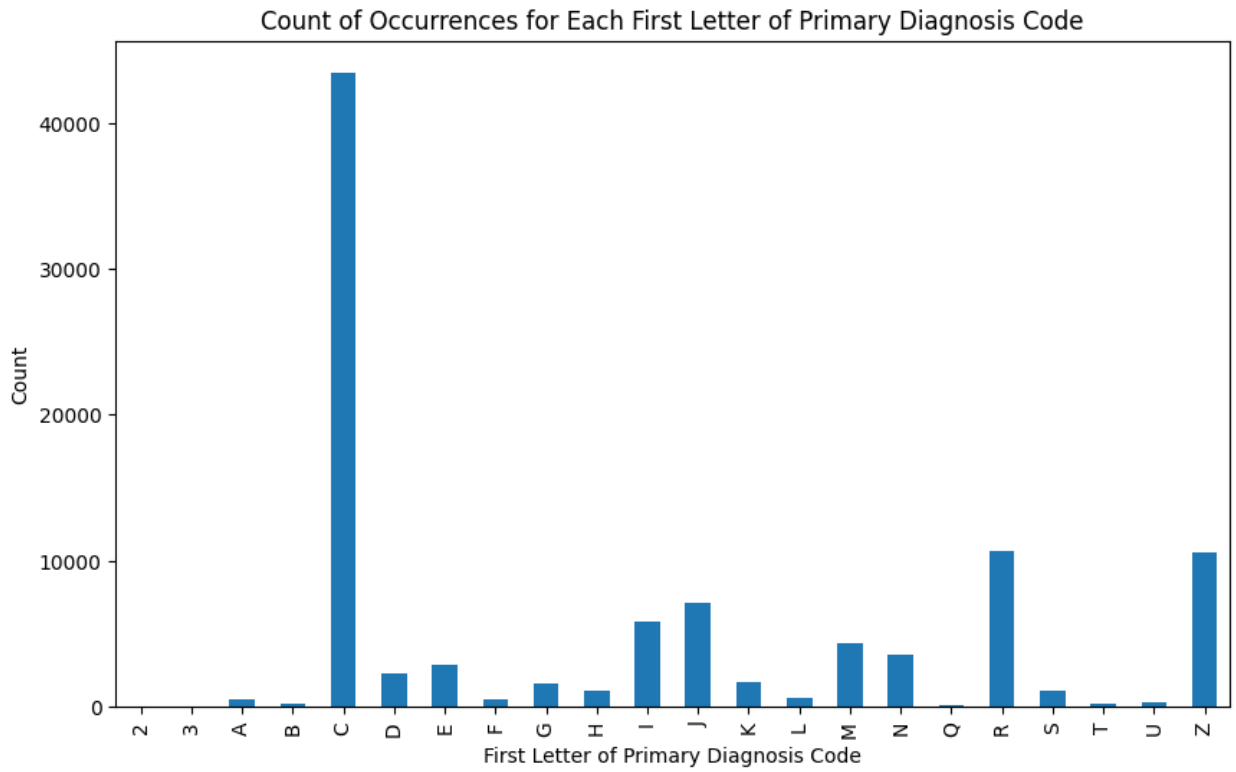
3.2 External Data

We used the CMS ICD-10 Primary Diagnosis Codes List data to analyze the different diagnosis codes reported in the data. (“ICD Code Lists”).

CODE	SHORT DESCRIPTION (VALID ICD-10 FY2023)	LONG DESCRIPTION (VALID ICD-10 FY2023)
A000	Cholera due to Vibrio cholerae 01, biovar cholerae	Cholera due to Vibrio cholerae 01, biovar cholerae
A001	Cholera due to Vibrio cholerae 01, biovar eltor	Cholera due to Vibrio cholerae 01, biovar eltor
A009	Cholera, unspecified	Cholera, unspecified
A0100	Typhoid fever, unspecified	Typhoid fever, unspecified
A0101	Typhoid meningitis	Typhoid meningitis
A0102	Typhoid fever with heart involvement	Typhoid fever with heart involvement
A0103	Typhoid pneumonia	Typhoid pneumonia
A0104	Typhoid arthritis	Typhoid arthritis
A0105	Typhoid osteomyelitis	Typhoid osteomyelitis
A0109	Typhoid fever with other complications	Typhoid fever with other complications
A011	Paratyphoid fever A	Paratyphoid fever A
A012	Paratyphoid fever B	Paratyphoid fever B
A013	Paratyphoid fever C	Paratyphoid fever C
A014	Paratyphoid fever, unspecified	Paratyphoid fever, unspecified
A020	Salmonella enteritis	Salmonella enteritis

We performed preliminary analysis to see what type of diagnosis codes occurred the most in our data and tried to see if there was any correlation between people being diagnosed with specific diseases and leaving the treatment.

We found high occurrences of diagnosis codes starting with C which corresponded to Cancer related issues which made sense as our target focus group is members having Lung Cancer.



We then went on to explore which were the top 10 diagnosis codes found in people who were leaving the Osimertinib therapy. The findings of Pneumonia, Shortness of breath and abnormal findings in the lungs are coherent with the side effects given on the Tagrisso website which are regarded as serious side effects which may cause people to leave the drug.

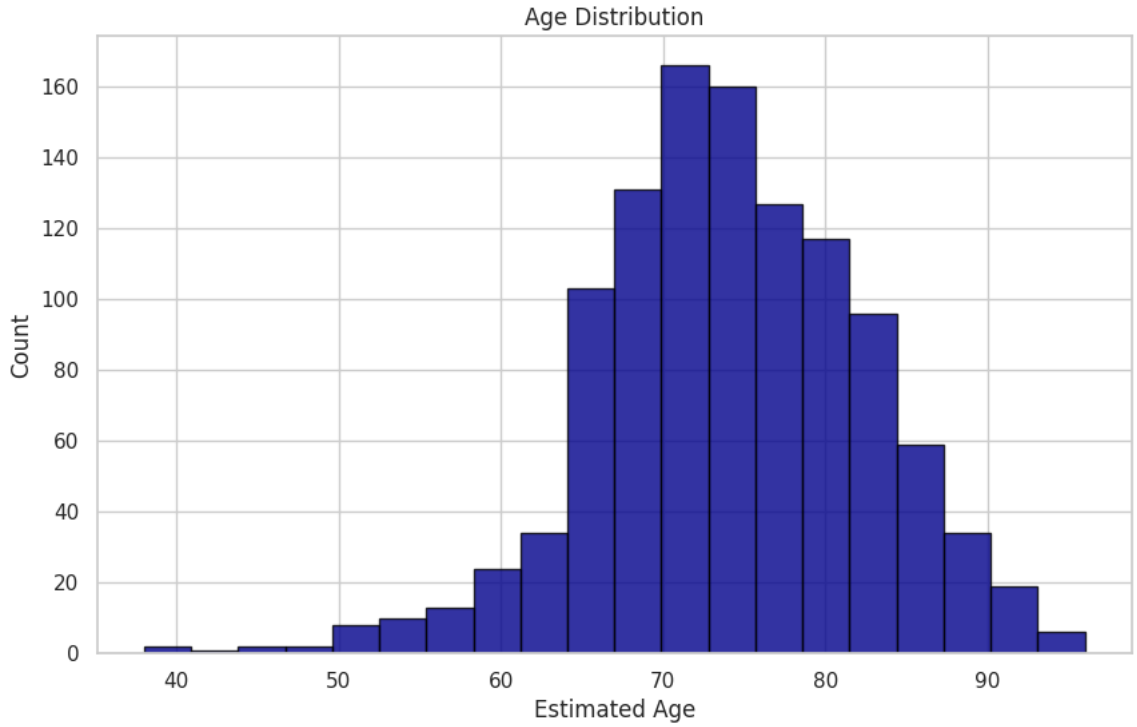
I10	Essential (primary) hypertension
I2699	Other pulmonary embolism without acute cor pulmonale
J189	Pneumonia, unspecified organism
J90	Pleural effusion, not elsewhere classified
J9601	Acute respiratory failure with hypoxia
N186	End stage renal disease
R0602	Shortness of breath
R531	Weakness
R918	Other nonspecific abnormal finding of lung field
U071	COVID-19

3.3 Exploratory Data Analysis

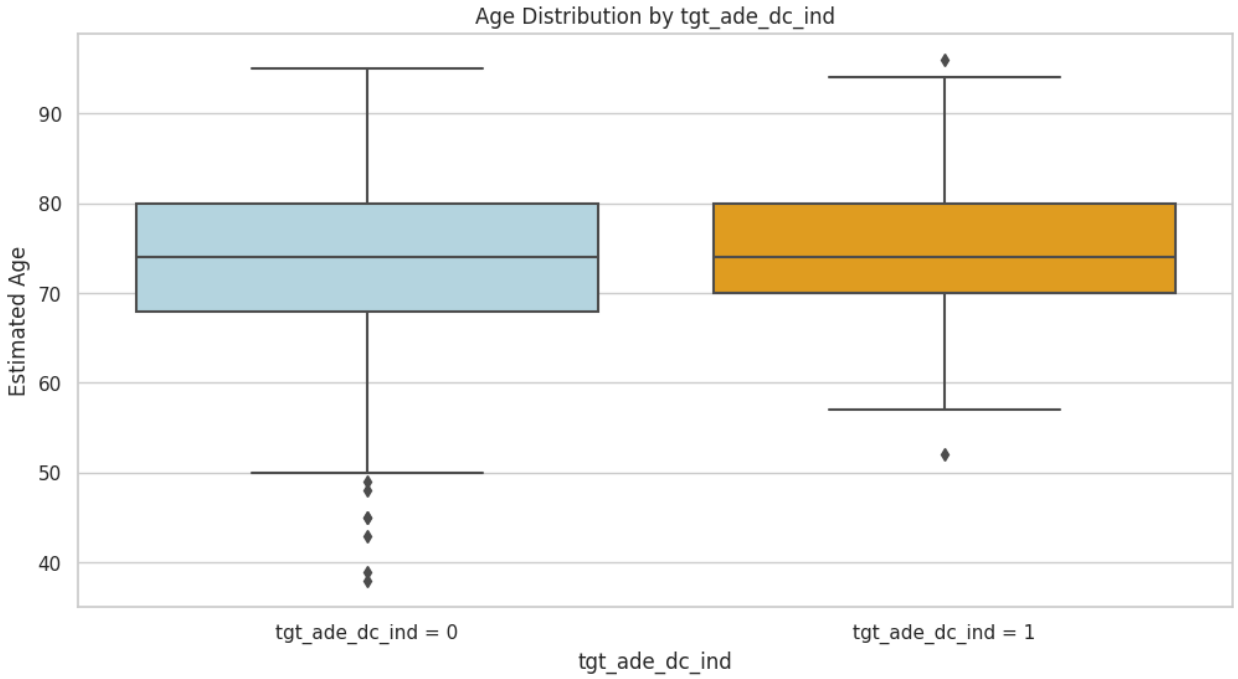
Prior to the data cleaning and model preparation phases, we conducted an in-depth exploration of the training dataset. Through a combination of exploratory data analysis visualizations and descriptive statistics methods, we aimed to gain a comprehensive understanding of the individuals represented in the dataset. This analysis allowed us to discern patterns, potential biases, or anomalies in the data, providing valuable insights into the demographics and characteristics of the dataset's members.

a. Age Distribution:

The analysis revealed that the average age of individuals undergoing Osimertinib therapy is approximately 74 years old. Furthermore, the age distribution exhibits a skewness towards older ages, indicating a concentration of individuals in the higher age range. A fractional percentage of the members were below the age of 45. This aligns with the findings provided by the American Cancer Society which states that lung cancer mainly occurs in older people. Most people diagnosed with lung cancer are 65 or older; a very small number of people diagnosed are younger than 45. The average age of people when diagnosed is about 70. ("Lung Cancer Statistics | How Common is Lung Cancer?")

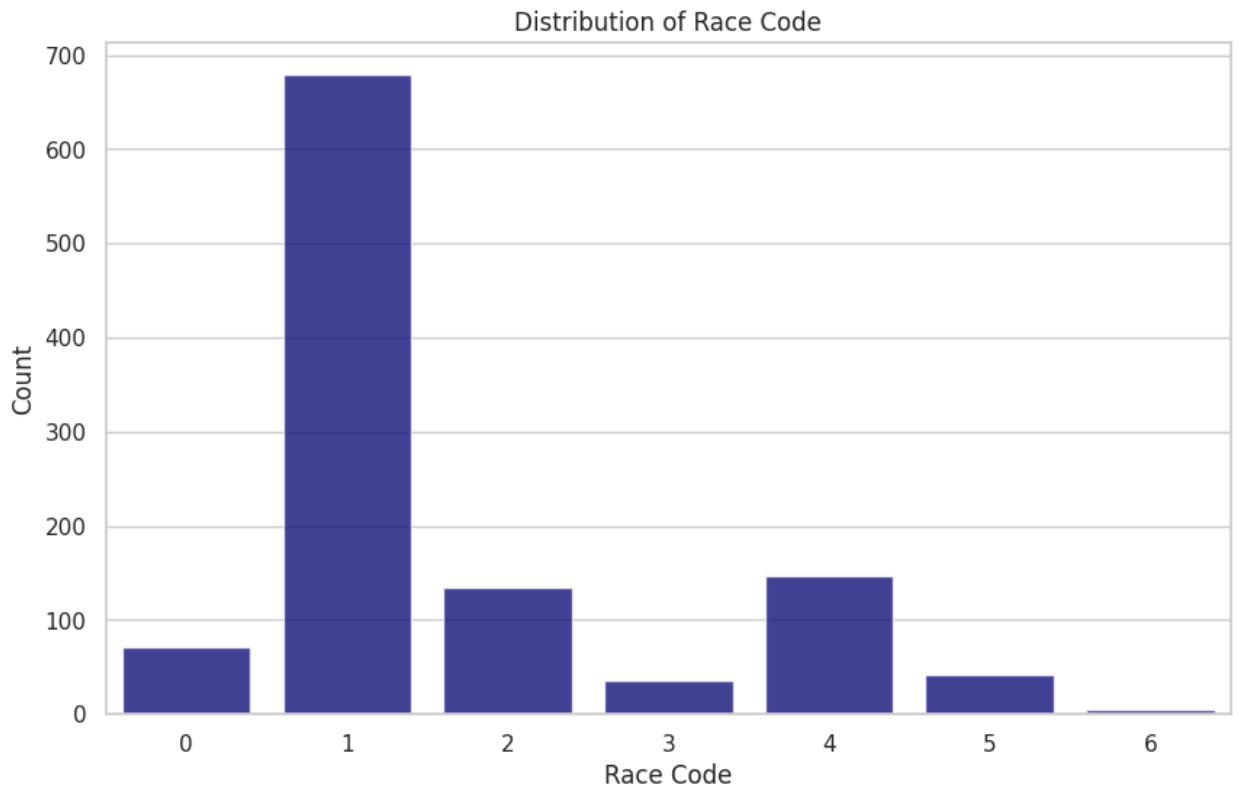


We also wanted to investigate the age distribution of people who left the therapy versus people who continued the therapy. Both had a similar mean but for people dropping out of the therapy, there are more people above the mean age than below indicating people dropping out tending to be of higher ages.

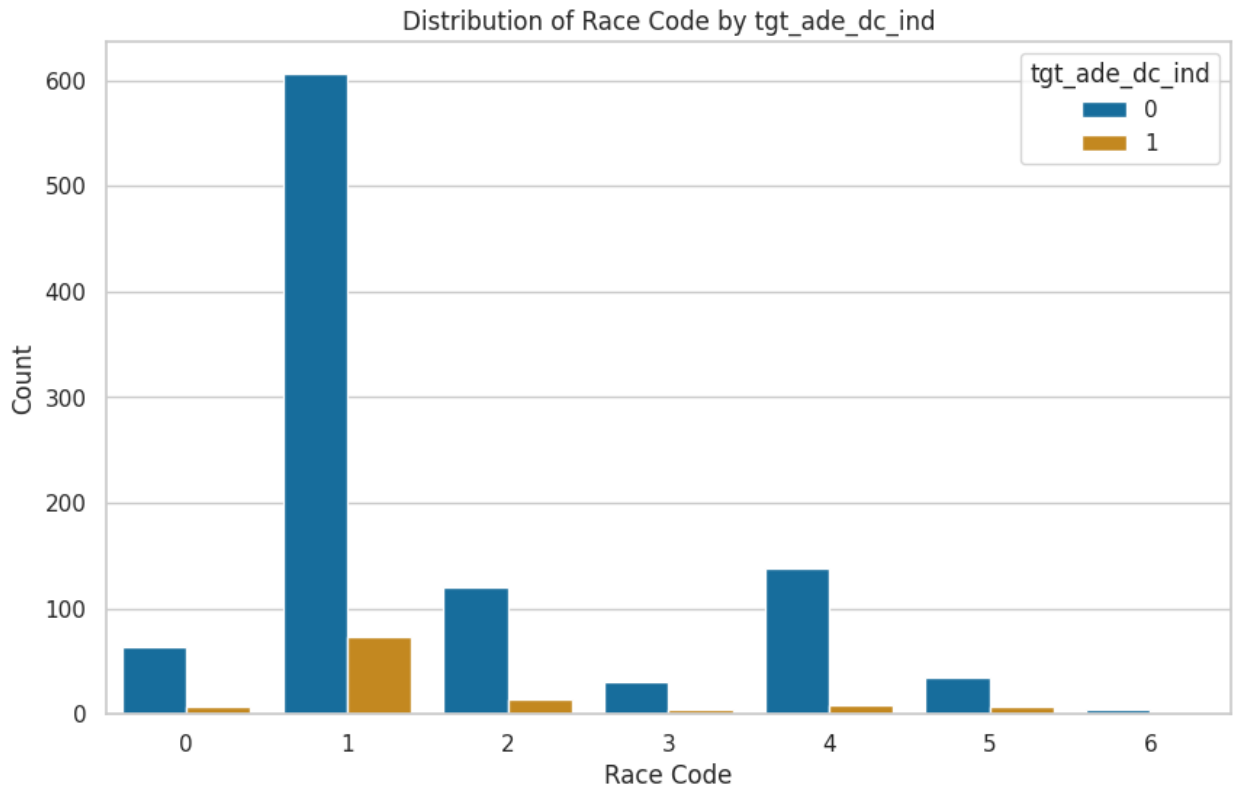


b. **Race Distribution:** We found a high percentage of white population in the dataset with about 700 white people out of the 1232 members. Below is the key for the race code and description mapping:

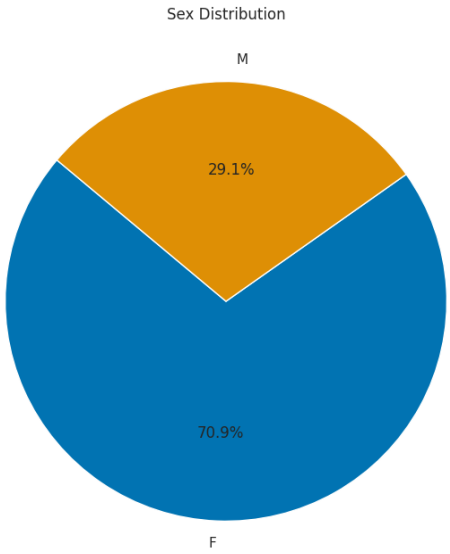
RACE CODE	DESCRIPTION
0	Unknown
1	White
2	Black
3	Other
4	Asian
5	Hispanic
6	Native American



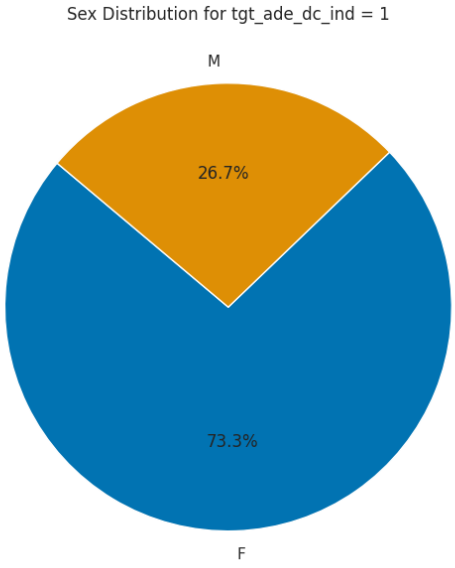
No strong observations were found regarding the race of people dropping out of the therapy. It is in proportion to the distribution of the total members in the data.

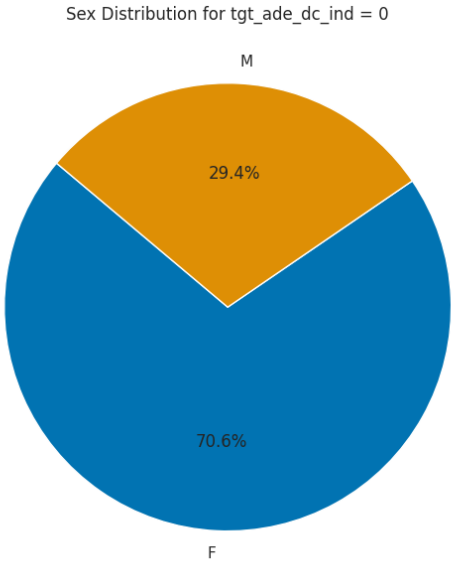


c. **Sex Distribution**: We found the number of females in the training data to be disproportionately higher than the number of males. This is different from the observed distribution provided by The American Cancer Society which is almost 50-50. ("Lung Cancer Statistics | How Common is Lung Cancer?" 2023)

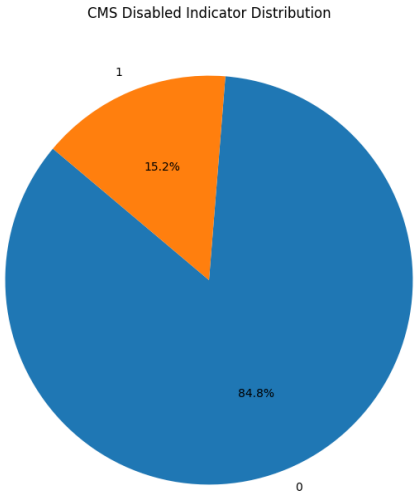


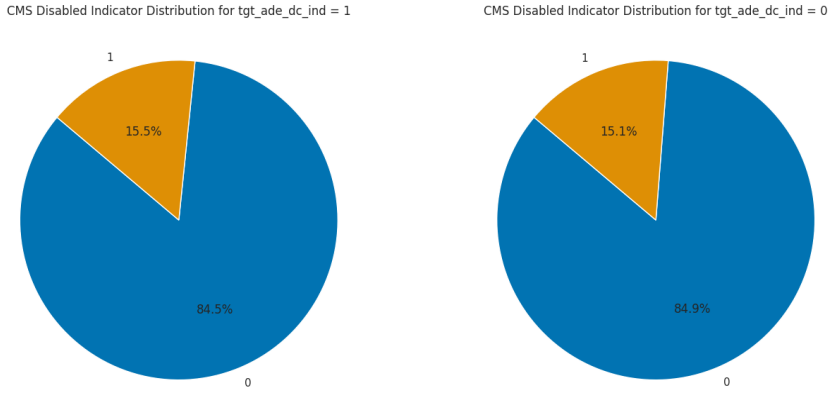
The distributions of members dropping out versus continuing therapy were similar with respect to the sex of the member.



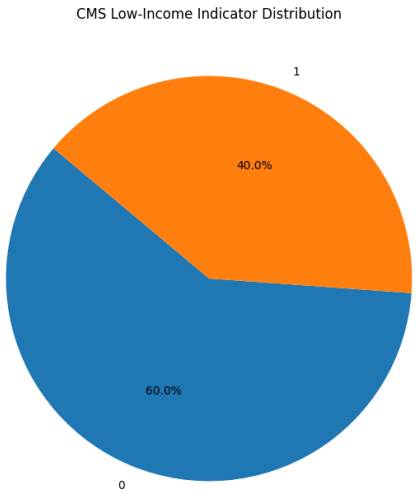


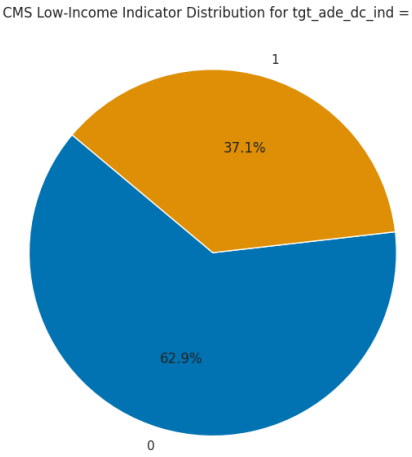
d. **Disability Distribution:** No strong correlation was observed with respect to the disability indication of a member.

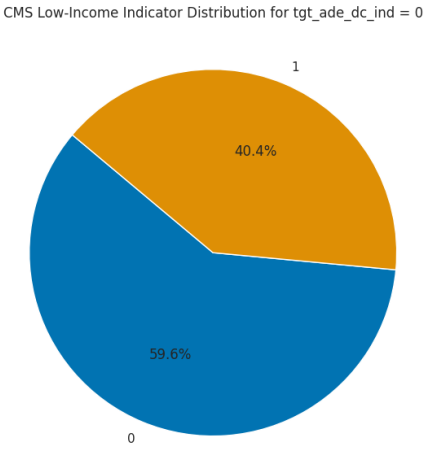




e. **Low Income Distribution:** No strong correlation was observed with respect to the disability indication of a member.

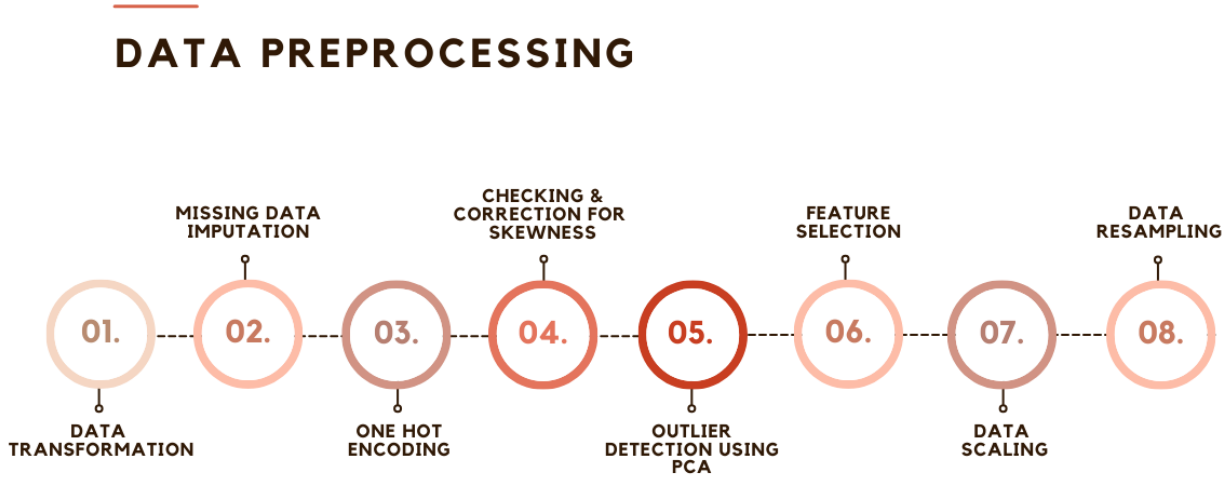



CMS Low-Income Indicator Distribution for tgt_ade_dc_ind = 1
1
37.1%
62.9%
0


CMS Low-Income Indicator Distribution for tgt_ade_dc_ind = 0
1
40.4%
59.6%
0

4. Data Preprocessing

The following steps were performed in order to clean and pre-process the data before training different models on the training dataset. This step includes imputation of missing values, outlier detection and management, feature engineering, feature selection, and many more steps in an attempt to arrive at a clean dataset which can be used to train machine learning models.



4.1 Medical Claims Data Transformation

We noticed that the medical claims dataset had many duplicates if we excluded the medical claim key field. This meant that there were multiple records of the same patient receiving the same treatment on the same day which does not add any value to the dataset. Hence, these rows were removed from the dataset before performing any exploratory data analysis.

A study on the various diagnosis codes reported in the medical claims data revealed that those patients with a primary diagnosis code that corresponds to shortness of breath and abnormal findings in the lungs were more likely to stop their treatment for Osemertinib due to an adverse drug event. Hence, in order to bring this information into the feature matrix, a new column was created which recorded whether a patient had any of their diagnosis codes belonging to these conditions.

The medical claims dataset has a few fields that, in context of this dataset, do not have any relation to the target variable. For example, the field claim type differentiates a pharmacy claim and a medical claim, this information is already evident with the table names and hence does not

have any impact on the model. The next step in the data preparation process involved removing these fields as listed below -

diag_cd2	diag_cd7	process_date
diag_cd3	hedis_pot	reversal_ind
diag_cd4	clm_type	pot
diag_cd5	diag_cd8	util_cat
diag_cd6	diag_cd9	

To ensure that the target train dataset, which contains one row for each therapy_id, can be effectively joined with the medical claims and pharmacy claims tables, which have multiple rows for each therapy_id, it was determined that we should aggregate the medical claims data at a therapy_id level while preserving all relevant features. This approach minimizes the risk of expanding the number of rows and ensures that no information is lost during the aggregation process. The following table describes each of the aggregations performed in the medical claims table.

Field Name	**Type of Aggregation**	**New Column Name**	**Rationale**
primary_diag_cd	nunique (Number of unique values)	Number_Of_Medical_Issues	To keep track of how many different diagnoses a patient has received.
visit_date	nunique (Number of unique values)	Number_Of_Visits	To count the number of visits a patient has had.
ade_diagnosis	mean	ADE_intensity	This is the number of times a patient has experienced this particular side effect divided by the total number of visits. This is a measure of how many times the patient has experienced any of these side-effects.
seizure_diagnosis	mean	Seizure_Intensity	
pain_diagnosis	mean	Pain_Intensity	
fatigue_diagnosis	mean	Fatigue_Intensity	
nausea_diagnosis	mean	Nausea_Intensity	
hyperglycemia_diagnosis	mean	Hyperglycemia_Intensity	
constipation_diagnosis	mean	Constipation_Intensity	
diarrhea_diagnosis	mean	Diarrhea_Intensity	

This data is now ready to be joined with the target train data.

4.2 Pharmacy Claims Data Transformation

A similar approach was applied to the pharmacy claims dataset as was done with the medical claims dataset. However, there were some differences in the treatment of certain fields, namely specialty_ind, generic_ind, and maint_ind. These fields consisted of binary categorical values, and they were transformed by pivoting the data. Specifically, these fields were split into two new fields, each representing the count of a particular binary value for each therapy_id. For example, maint_ind was divided into two new fields, MAINT and NONMAINT, each containing the count of their respective values for each therapy_id. This transformation closely resembles the concept of one-hot encoding or dummy variables, which we will discuss later.

The rest of the data was transformed in a similar way to what was done with the medical claims data. The following table describes how each field was aggregated and the rationale behind each decision (most of which are self-explanatory) -

Field Name	**Type of Aggregation**	**New Column Name**	**Rationale**
service_date	nunique	Number_of_Rx_Visits	To count the number of Rx visits made by the patient
ndc_id	nunique	Drug_type_Count	This gives us a distinct count of drug types
ddi_ind	sum	Number_Of_known_interactions	Number of times a patient has taken a drug that is known to have interactions with Osimertinib
anticoag_ind	mean	Mean_Of_AntiCoag	Number of times a patient has taken a drug for these symptoms is divided by the total number of visits to the pharmacy.
diarrhea_treat_ind	mean	Diarrhea_Rx_Mean	
nausea_treat_ind	mean	Nausea_Rx_Mean	
seizure_treat_ind	mean	Seizure_Rx_Mean	
rx_cost	sum	Total_Rx_Cost	Total cost at the pharmacy for a given therapy_id.

Now that both tables have been prepared, they can be merged with the target train table.

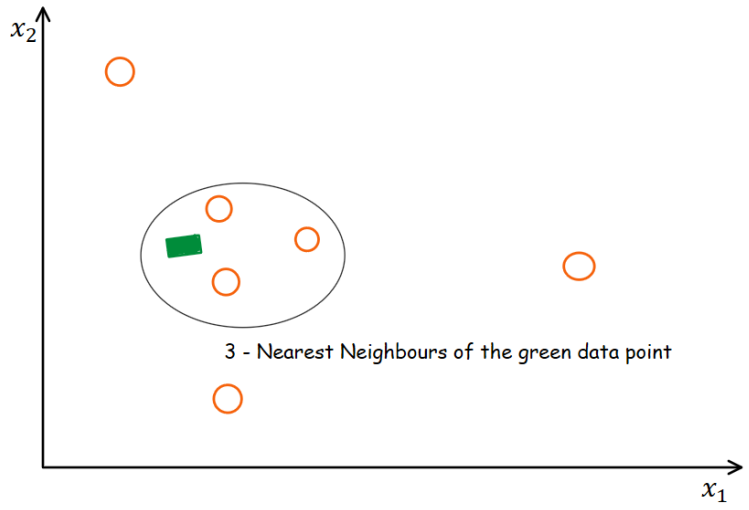
4.3 Merged Data Transformation

The following steps were performed on the merged data one after the other as part of pre-processing.

4.3.1 Missing Data Imputation

After merging the data, we noticed that the medical claims data was unavailable for 500+ patients. Further, there were a few rows in the merged table that had missing values which stemmed from the raw data provided. The easiest way to deal with this is to remove rows which have data missing from any of its features. However, this potentially could lead to a loss of important data. Hence, we decided to use one of the many options available to fill in missing data.

When it comes to handling missing data, there are numerous strategies at our disposal. After thoughtful deliberation, we settled on utilizing the kNN Median Imputation method as our preferred approach.This method leverages k-nearest neighbor models created using all other features except for those with missing values and fills in the missing values with the median of the k -nearest neighbors. For our model, we used k=5 and computed the median.



4.3.2 One Hot Encoding

For the next step, quantitative features were separated from the categorical features. All categorical features were found to be nominal.. Most machine learning models require all fields to be numerical in nature. Hence, it is a common practice to encode categorical features with numerical values.

When dealing with nominal variables, a widely adopted method for managing categorical data is one-hot encoding. Here, we create new dummy variables which take boolean values of 0 and 1

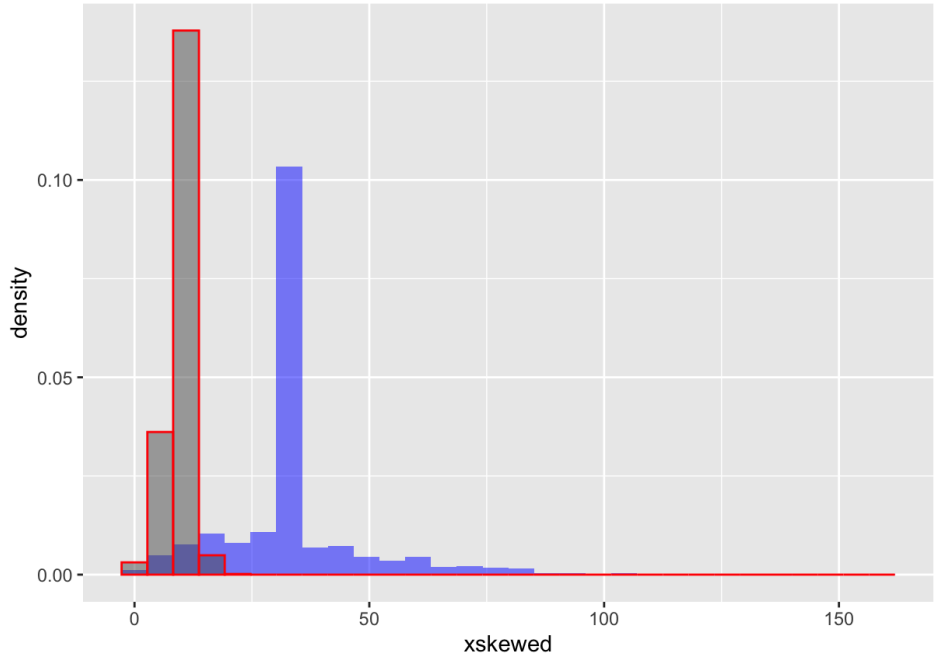
and the combination of these variables leads to one particular value in the original categorical variable. The following are the list of categorical variables -

- Race
- Sex
- CMS_Disabled_Ind
- CMS_Low_Income_Ind

4.3.3 Checking and Correcting for Skewness

All quantitative features were then checked for skewness by computing the skewness factor. A threshold value of 2 was selected and any features with a skewness factor above 2 or below -2 were corrected using the Yeo Johnson transformation. The following list of fields were corrected for skewness. The graph below the table shows the density distribution for the field - Total number of medical claims, before(Blue) and after(Red) correcting for skewness.

Number of Visits	Fatigue_Intensity	Number_Of_known_interactions
ADE_Intensity	Nausea_Intensity	Diarrhea_Rx_Count
Seizure Intensity	Hyperglycemia_Intensity	GENERIC
Pain_Intensity	Constipation_Intensity	BRANDED

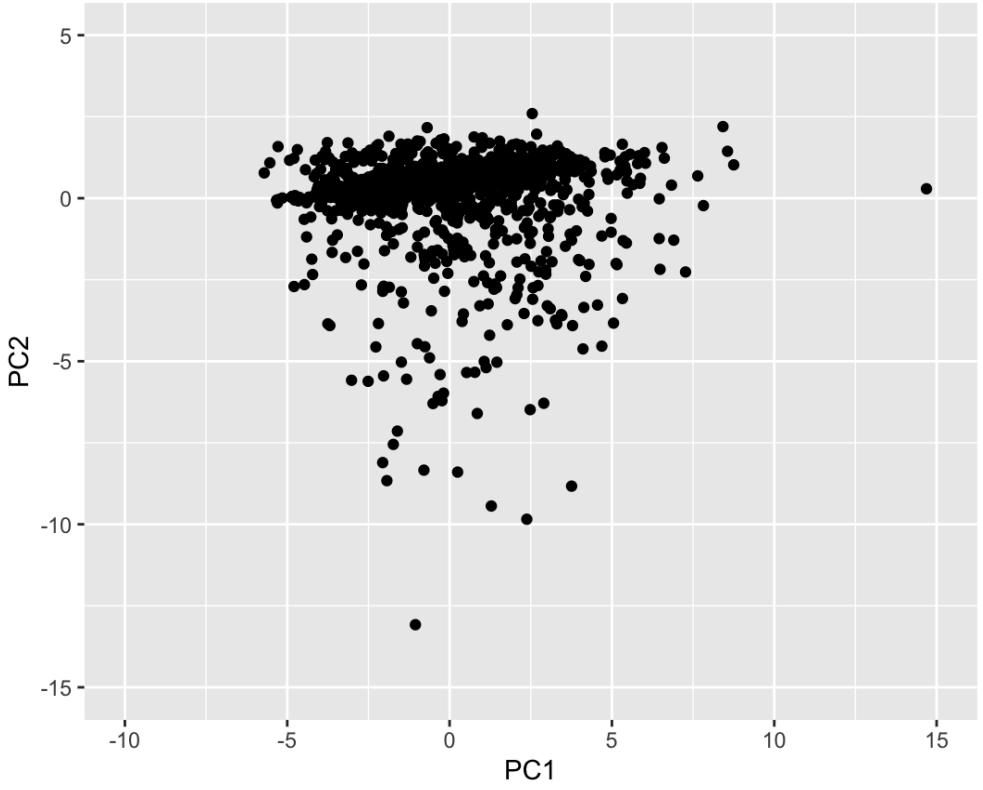


4.3.4 Outlier Detection using Principal Component Analysis (PCA)

Outlier detection can provide early warning signals for abnormal conditions, allowing experts to identify and address issues before they escalate. Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction while preserving relevant information. Due to its sensitivity, it can also be used to detect outliers in multivariate datasets (Ridley and Guide).

Principal component analysis was performed for the quantitative variables in our dataset and PC1(principal component 1) was plotted against PC2(principal component 2) to detect any outliers as shown in the image below.

From the plot, it can be seen that there are clearly two outliers in the dataset(PC1 value of 15 and the second one with PC2 value of about -13). Upon further investigation, it was noticed that one of these patients has an abnormally high ADE intensity value of 0.73 (which means that 73% of their visits involved a side-effect) which is way off from the mean value(0.034). Similarly, for the other outlier, the number of Rx visits(92) is much greater than the median(17).

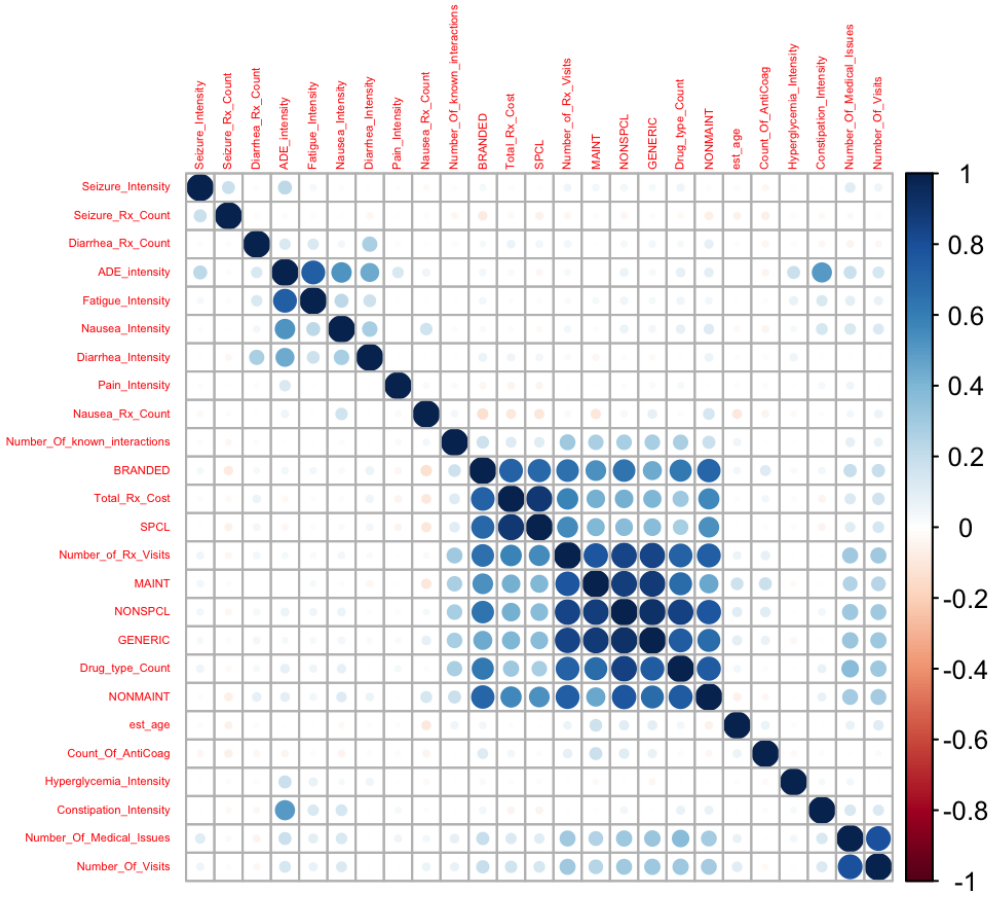


These outliers were then removed from the training dataset. Once the outliers were removed, some of the pre-processing steps were performed again to eliminate the effect of these outliers in the data (such as imputation). Nevertheless, given that we used the kNN median imputation technique, it is unlikely that these outliers had a substantial impact on the imputation process.

4.3.5 Feature Selection using Correlation Plot/Matrix

Feature selection using correlation matrix is the process of removing features that are highly correlated with each other. This means that these features have the same effect on the dependent variable and hence the presence of both features does more harm than good while training the model. Numerous strategies are available to address this issue, with the two most prevalent options being either removing one of the columns from the feature matrix or deriving a principal component for highly correlated features.

Given below is a plot of the correlation values of each of the quantitative features. Features with a correlation value of greater than 0.8 were filtered out and only one of them was retained in the final feature matrix.



The following fields were removed from the feature matrix as a result of feature selection -

- NONSPCL
- Number of Rx Visits
- GENERIC

4.3.6 Scaling of Data

Data scaling is a crucial preprocessing step for numerical features. Several machine learning algorithms often necessitate data scaling to yield optimal results. Scaling ensures that all features are given equal importance before training of the model takes place. It makes sure that higher quantitative values of certain features do not outweigh other features that have inherently lower values.

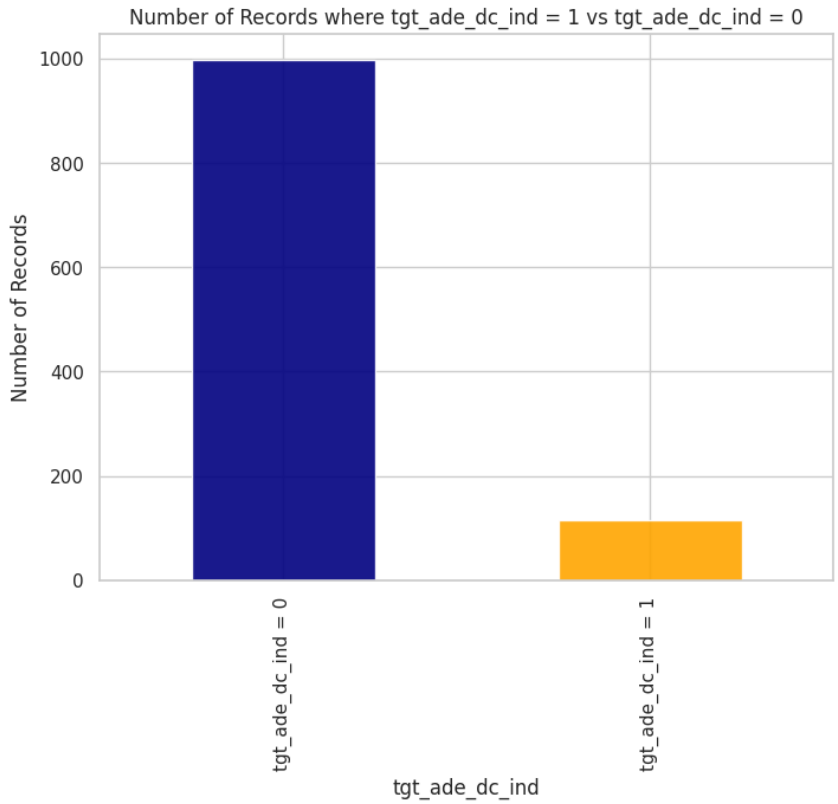
We used the MinMax scaler in order to scale all the features. The MinMax scaler is a technique used to standardize features in such a way that the minimum value of each feature is set to zero, and the maximum value is set to one. This scaler effectively compresses the data into a predefined range, typically from 0 to 1. It accomplishes this by rescaling the feature values, ensuring that they fall within the specified range, all while preserving the original data distribution's shape.

The formula used to scale data using MinMaxScaler is given below -

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4.3.7 Dataset Resampling

Humana declared that in the 2020-2021 period, approximately 24% (419 out of 1765) of the therapies were discontinued. However, in the training data provided to us, we observed that there were only 116 members out of 1232 who discontinued the therapy. This is approximately 9% which is extremely low. We wanted our model to be more accurate at limiting the number of false negatives rather than being accurate at limiting false positives. Hence, we decided to resample the data by oversampling the records of people who discontinued the therapy. We performed oversampling using the RandomOverSampler method from the imblearn.over_sampling package in Python. We brought the final ratio of the number of people who discontinued the therapy to the total number of members to 0.30.


Number of Records where tgt_ade_dc_ind = 1 vs tgt_ade_dc_ind = 0
Number of Records
1000
800
600
400
200
0
tgt_ade_dc_ind = 0
tgt_ade_dc_ind = 1
tgt_ade_dc_ind

5. Predictive Model Development

To solve the classification problem, we analyzed it as a supervised learning classification statement. By keeping several factors in mind while choosing the model like the size of the dataset, the prediction variables and interpretability of the features, we came up with the four algorithms: Logistic Regression, Decision Tree, Random Forest, Light GBM and Gradient Boosting.

We wanted to analyze the performance of each of the models and then choose the final model.

Logistic Regression is favored when the relationship between features and the target variable is linear, making it apt for binary classification and probability estimation. Decision Trees excel in capturing non-linear relationships and are prized for their interpretability, providing clear insights into feature importance and decision-making logic. Random Forest, an ensemble method of multiple decision trees is good in handling complex datasets by reducing overfitting and adeptly managing missing values. Its ability to aggregate predictions from multiple trees enhances accuracy and robustness. On the other hand, Gradient Boosting, another ensemble technique, constructs trees sequentially, correcting errors of previous trees. This sequential learning approach results in powerful predictive models, making it ideal for tasks demanding high accuracy. LightGBM is particularly useful when dealing with large and high-dimensional datasets, as it efficiently handles big data and can quickly train accurate models even with millions of data points and numerous features.

We used a 80-20 test-train split for training the model. Then, we performed cross-validation with 80% training data and 20% testing data across 5 models. The k was set as 5 to perform the cross-validation and reduce the bias and variance. The evaluation metric we used was AUC, which is Area Under the Receiver Operating Curve (ROC) which shows how well the model can distinguish between the classes. We also calculated the Accuracy that reflects the measure of how many predictions it got right out of the total number of predictions made. (Brownlee)

For each model, we implemented several steps to compare the performances of these models. We did GridSearch to analyze which hyperparameters are the best for the data and the model. After obtaining the best hyperparameters of each model and doing cross validation on each model, we compared the AUC and accuracy of these models to find the best fit.

5.1 Hyperparameters Tuning

With the new engineered features along with our original features, we performed hyperparameter tuning on our models. We used the approach called Gridsearch from scikit-learn ‘s

GridSearchCV that performs an exhaustive search over the listed hyperparameters to obtain the best combination. We implemented it for 4 hyperparameters for 3 different values, that means the algorithm will iterate over 4 x 3 = 12 fits of the model. We tuned the following hyperparameters for the model:

- N_estimators: 50; It determines the number of decision trees in the random forest. A higher value can lead to a more robust model, but it also increases computation time.
- Max_depth: 5 ; This sets the maximum depth of each decision tree in the random forest. Deeper trees can capture complex patterns, but they might overfit, so it's crucial to find an optimal depth.
- Min_samples_split:2; It represents the minimum number of samples required to split an internal node. Higher values prevent overfitting by ensuring a node has enough samples to split.
- Min_samples_leaf: 1; This parameter sets the minimum number of samples required to be at a leaf node. It prevents creating nodes that only fit a small number of samples, promoting a more generalized model.
- Random_state: 42; This parameter in machine learning functions allows us to set a seed value for the random number generator. Setting a specific random_state ensures reproducibility, as the same random processes will be generated every time you run the code.

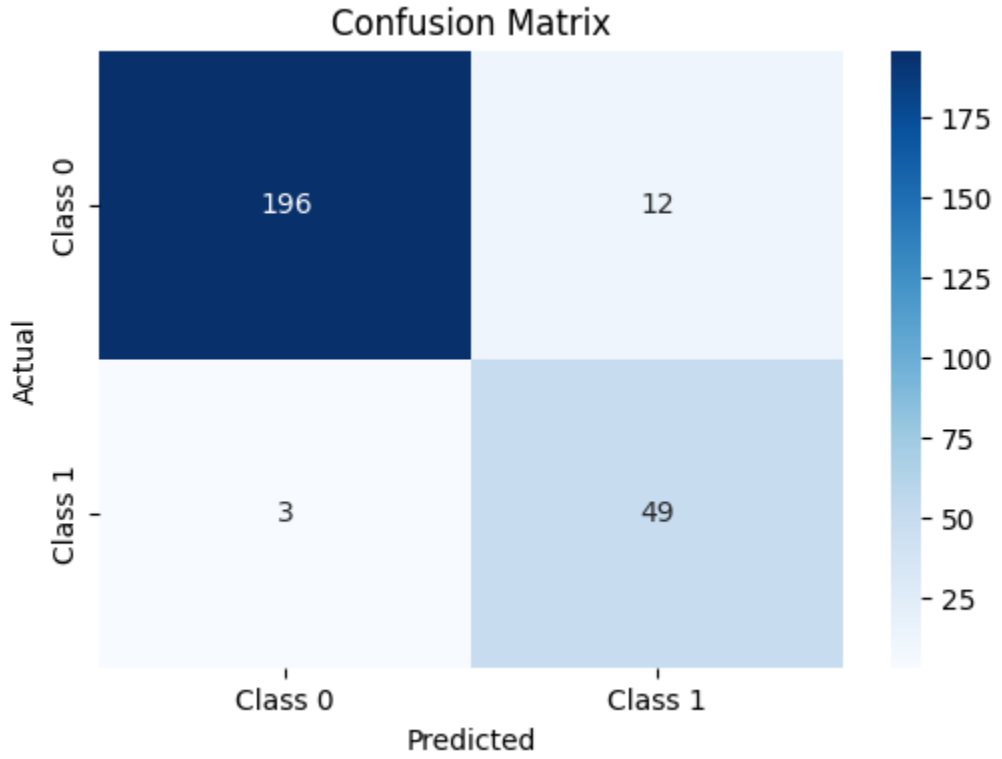
5.2 Final Model Construction

We obtained the best parameters for each of the three models we wanted to evaluate and then compared them.

Models	**Accuracy**	**ROC AUC Score**
Logistic Regression	0.82	0.88
Decision Tree	0.90	0.728
Random Forest	0.911	0.968
Light GBM	0.928	0.937
XGBoost	0.942	0.978

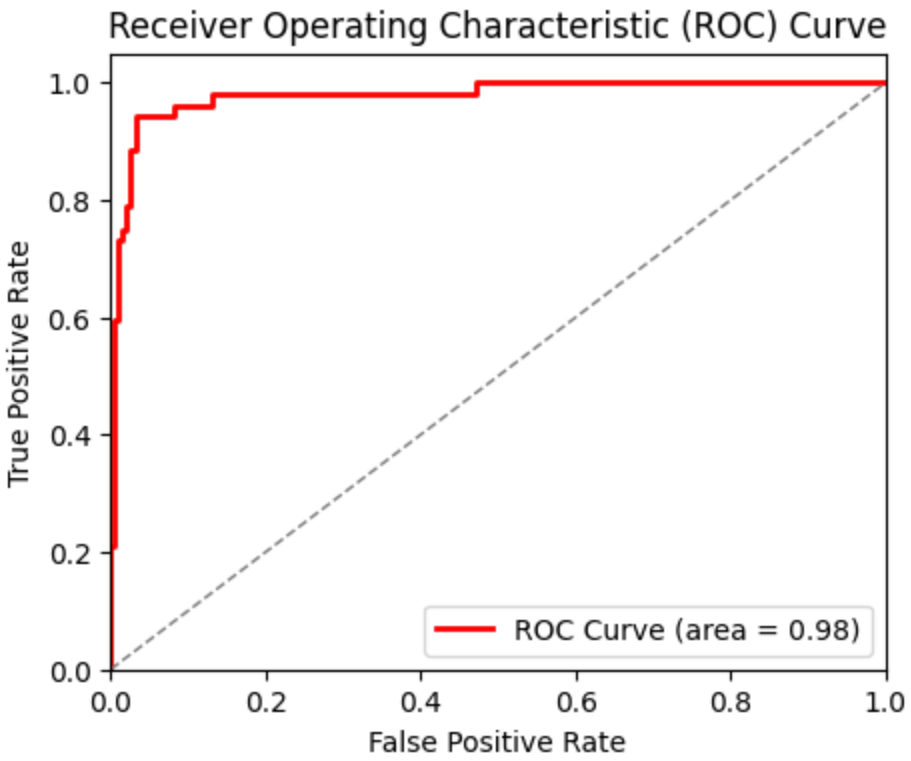
Based on the AUC scores of various models. We decided to go ahead with XGBoost which showed 0.942 AUC ROC score. We also created the confusion matrix for our model that

determined the true positives and the false negatives to better analyze our model. We kept the threshold to be 0.5 for creating the confusion matrix.



Based on the confusion matrix, we can compute recall, sensitivity, and accuracy:

- Sensitivity: Sensitivity is a measure of a model's ability to detect true positive cases. It is obtained by calculating True positives by the total positives. For our model, it comes to be 49/(12+49) = 80.32%
- Specificity: Specificity is a measure of a model's ability to detect true negative cases. It is obtained by calculating true negatives by the total number of negatives. For our model, it is 196/(196+3) = 98.49%
- Accuracy: Accuracy is obtained by calculating the true positives and true negatives by the total values. This can be calculated by (196+49)/(196+3+12+49) = 94.23%



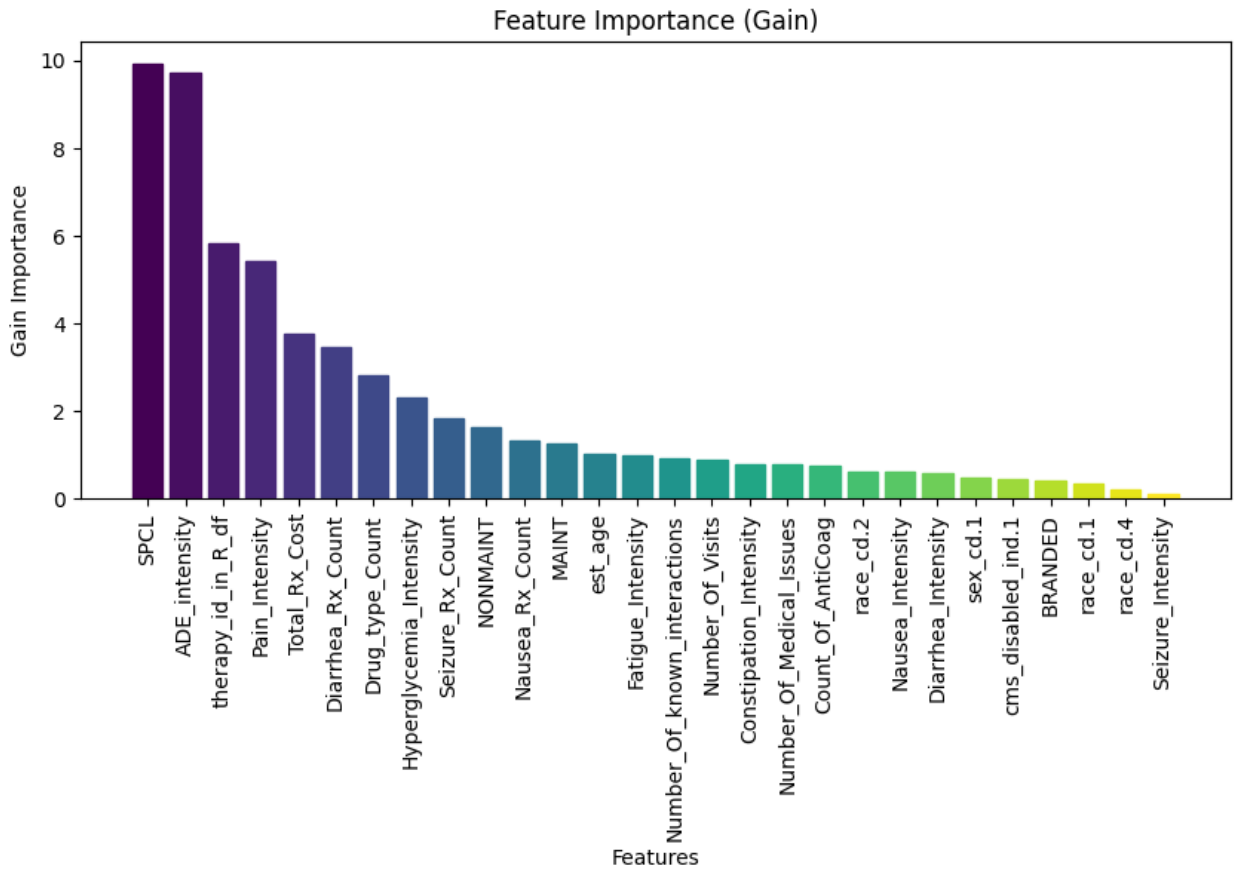
By analyzing the ROC AUC Curve , we can observe that the model has optimal compromise between specificity and sensitivity. The top left point denotes a perfect classifier and our model is very close to that classification.

5.3 Key Performance Indicators Analysis

Feature Importance

To identify the top key performance indicators, we identified the features that were important to the model using Feature importance analysis on XGBoost Gain importance and SHAP values.

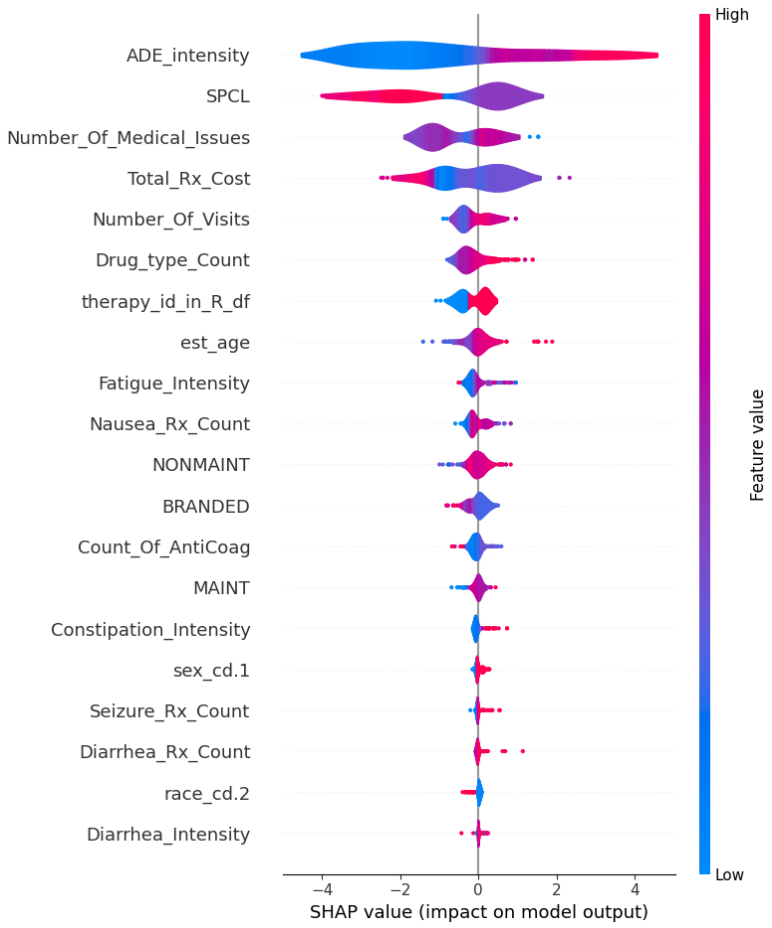
- **Gain Importance**



We used a built-in XGBoost model to obtain the important features and the above graph shows 28 features that are obtained by calculating the gain factor. It calculates the contribution of each feature to the model's predictions by measuring the average gain (or improvement in accuracy) of the decision trees where the feature is used. Features with higher "gain" values are considered more important as they lead to a larger improvement in prediction accuracy when included in the model's decision-making process.

- **SHAP Values**

In our feature importance analysis, SHAP helps us to visualize the important features in our model by quantifying the contribution of each feature into our model. Positive SHAP values indicate a feature's positive impact on the prediction, while negative values suggest a negative impact. SHAP values can also show how features interact with each other. It helps in understanding not only the individual feature importance but also how combinations of features affect predictions. (O'Sullivan)



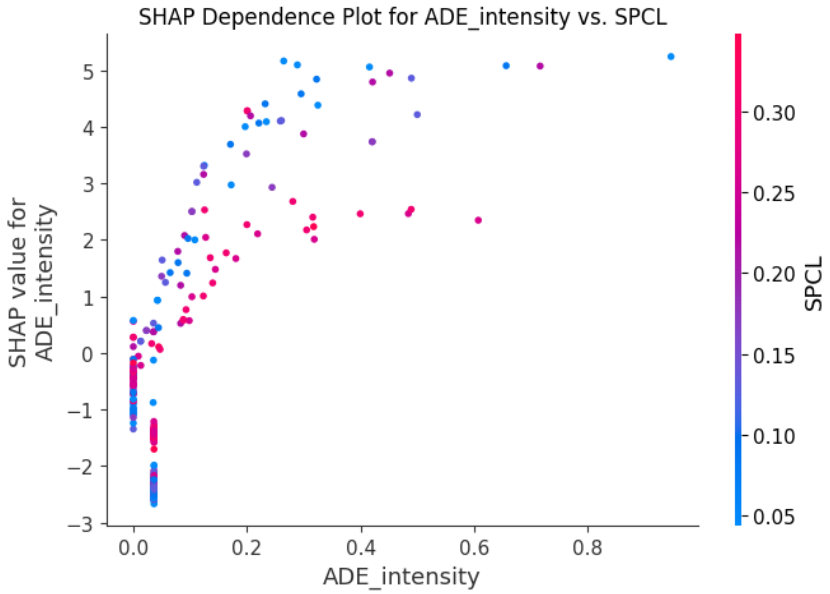
As seen from the gain feature importance plot and the SHAP values plot, we have identified the features that stand out and that are most important in our XGBoost Model. And comparing the aspects of all variables, the features can be categorized as the following:

1. Number of times a patient is diagnosed with an ADE: ADE_intensity is an important feature as per both the plots as it denotes the number of occurrences of ADE=1 for a patient. This shows patients having higher ADE_intensity are more likely to discontinue the therapy.
2. Speciality Drug Indicator: SPCL is also an important feature from the plots and it is negatively correlated with our model. As per insights from the data, we observed that SPCL was highly related to chemotherapy drugs like Osimertinib and Tarceva in the rx_claims dataset.
3. Number_of_Medical_Issues: This feature is of very high value It represents patients having a higher number of primary diagnosis incidents. We considered that the number of medical issues a patient has is correlated to the patient discontinuing the therapy. This inference underscores the vital role played by the complexity of a patient's medical condition in influencing their treatment adherence decisions.

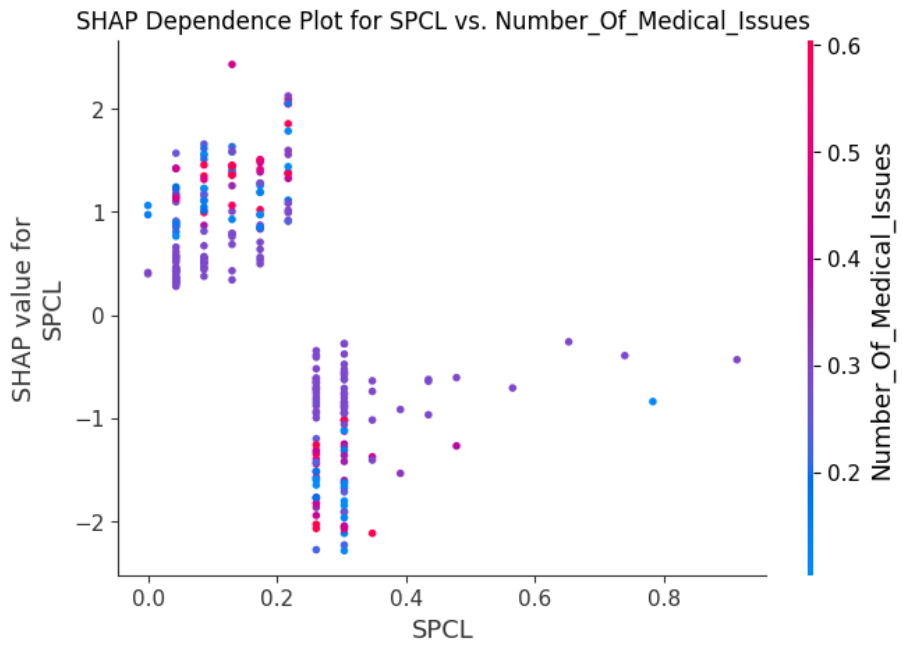
4. Total_Rx_Cost: The presence of this feature in both plots suggests a notable correlation between the total pharmacy drug costs and therapy discontinuation. This correlation is intuitively logical: higher drug costs may compel individuals to halt therapy due to financial constraints. The trend underscores a significant aspect of healthcare decision-making, highlighting the impact of economic factors on treatment adherence.
5. Pain_Intensity: The feature Pain_Intensity indicates that a patient has gone through pain through ADEs. It has high feature importance in the Gain importance which is directly related to the patient discontinuing the therapy before 6 months(Target ADE=1).

5.4 Relationship among features

As we saw from the SHAP values plot, we obtained the features that are most important to our classification model. To further obtain the relationship between features, we obtained the SHAP dependence plot.



- This plot shows that a patient having a higher count on ADE intensity is more likely to discontinue Osimertinib therapy because of more encounters with ADE. This shows positive correlation. Also, a patient having a higher count on SPCL would have built more tolerance to ADE and hence the ADE intensity count will be less. Therefore they are less likely to discontinue and it has a negative correlation.



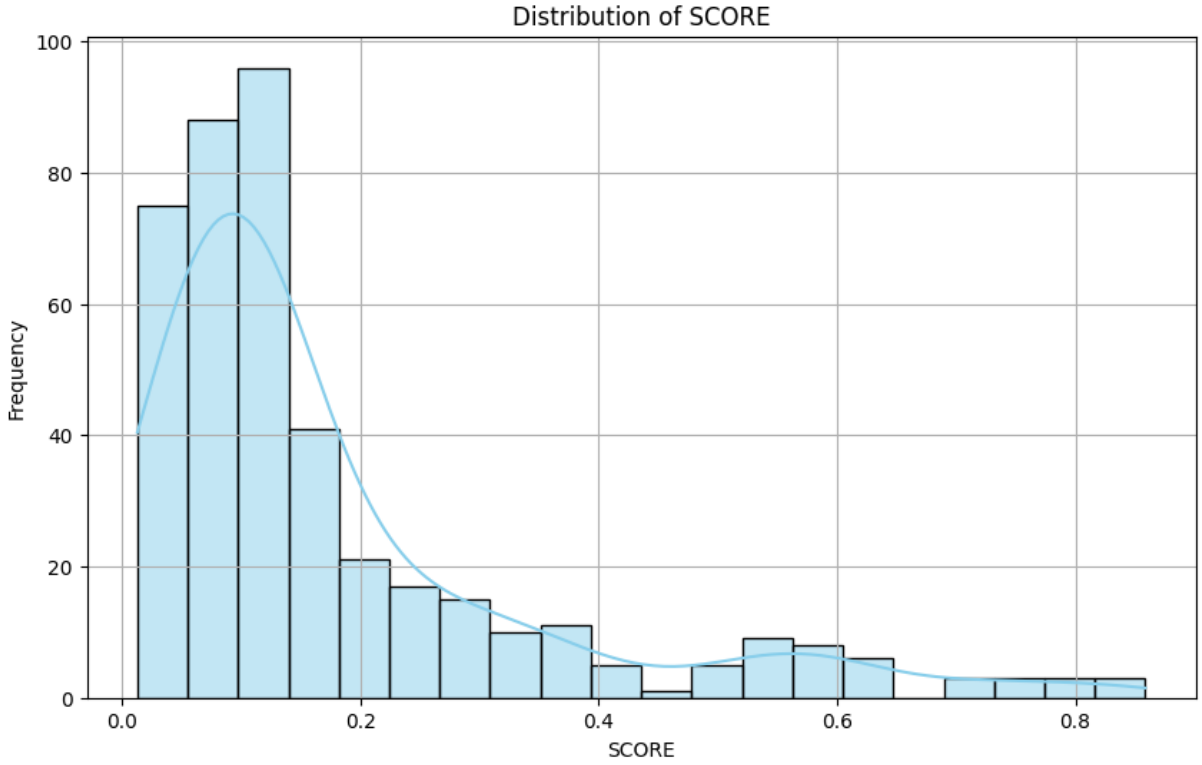
- This plot shows that a patient having a higher number of medical issues is more probable to discontinue Osimertinib therapy. This is because they are more likely to find the therapy burdensome than befitting from it. Therefore there is a positive correlation. Also, a patient having a higher count on SPCL would have built more tolerance and hence the number of medical issue counts will be less. Therefore they are less likely to discontinue and it has a negative correlation.

5.5 Data Biases

While performing exploratory analysis on the data we came across a certain bias in the data. We found that in terms of one of the protected attributes, race, there were a disparately higher number of white people in the dataset. We were not sure if this was a data bias but we would suggest Humana to look closely into this. We also found a disproportionate amount of females as compared to males in the data. However, the competition was judging the fairness criteria based on the disparity score which was based on the true positive rate.

6. Results and Findings

We used our XGBoost model for inference on the holdout dataset which gave an ROC-AUC score of 0.8247 and a Disparity score of 0.9788. We can observe the distribution of the likelihood variable that denotes the probability of members dropping out of the Osimertinib therapy after experiencing a side effect.



The key insights that we generated from our model were:

1. The key features for determining the probability of a dropping out of therapy are ADE intensity, number of medical issues, pharmacy drug cost, and specialty drug indicator.
2. Members are highly likely to leave the Osimertinib therapy during the initial 2 months of the therapy and that likelihood progressively lowers as the duration of the therapy increases.
3. The overall driver for members dropping out of the Osimertinib therapy is a combination of Adverse Drug Event factors, cost factors and member attributes.

7. Recommendations

7.1 Interpretation of Results, Classification of patients based on Likelihood Score

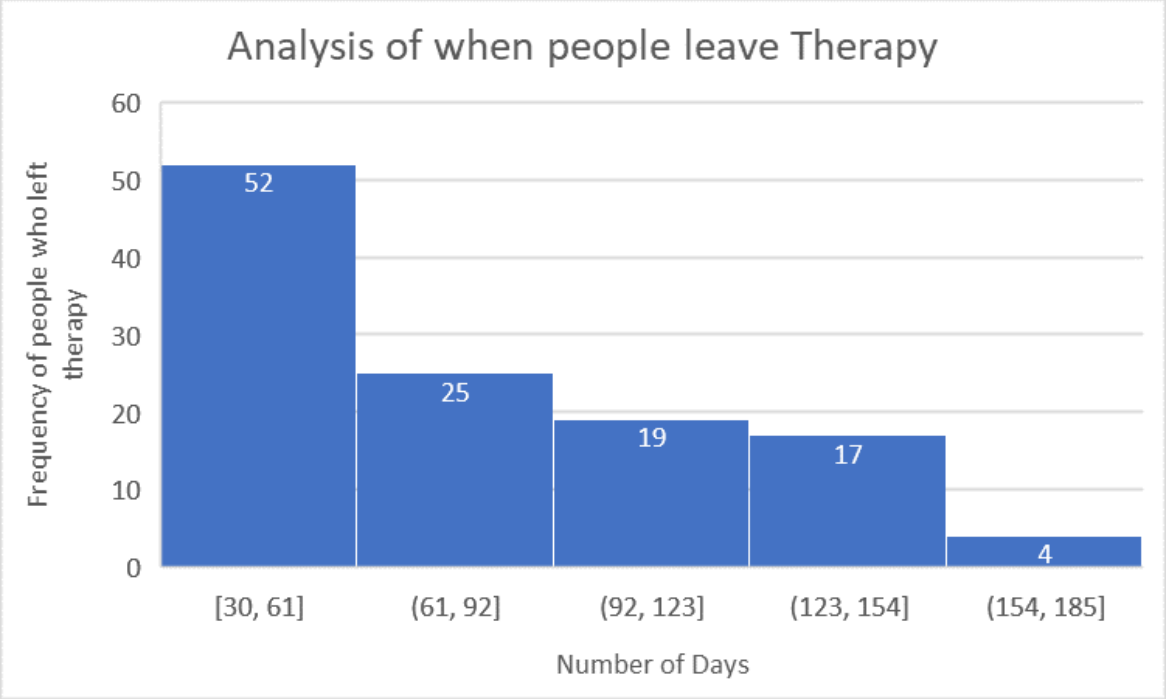
Now that we have trained our model and have used it to make predictions, it is now time to interpret the results and make data-driven decisions that enable the overarching goal of saving lives. The main objective of this report is to come up with ways to ensure that patients do not discontinue their Osimertinib therapy. Everything we have done up to this point is related to identifying those patients who are more likely to discontinue treatment.

Our strategy is based on differentiating patients according to the likelihood score that we arrived at using our classification model. The likelihood score is a measure of the probability of a patient leaving Osimertinib treatment due to side effects. Based on this likelihood score, we decided to split the patients into three different groups, High risk, medium risk and low risk patients. These buckets were determined by splitting the score in the holdout dataset based on quantiles as seen in the screenshot below -

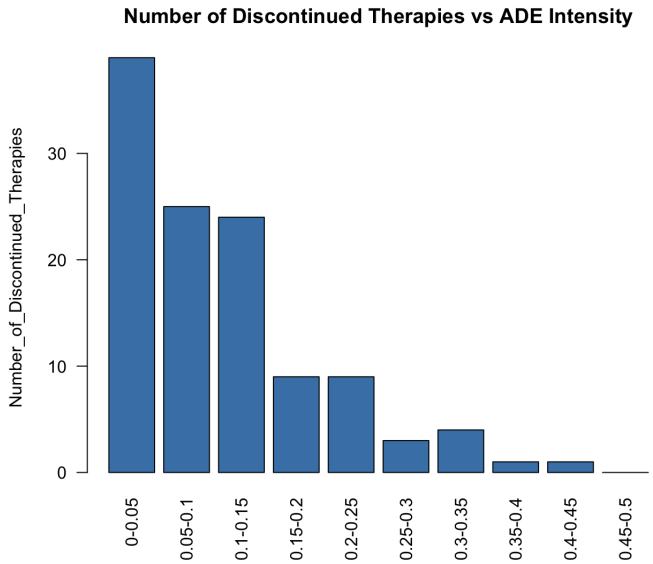
```
        0%        30%        90%       100%
0.01306203 0.07876982 0.42397575 0.85857892
```

Based on the values above, we determined that those with a likelihood score of greater than 0.42 shall be categorized as High risk patients and those with a score lesser than 0.0787 shall fall under low risk. Anyone in between are considered to be at a medium risk to discontinue therapy.

7.2 Prioritizing Early Intervention and Patient Awareness in Osimertinib Treatment



In the above chart, we can see that maximum people leave the therapy within the first two months and the dropping rate decreases as time progresses.



From the above plot, it is evident that with an increase in ADE intensity (which is a measure of the number of times a patient experienced side effects), they are less likely to leave treatment.

We can infer from both plots that patients are more likely to discontinue treatment early in their journey of the treatment since they tend to discontinue at the first instance of experiencing a side effect.

Our recommendation is that medical professionals emphasize the importance of closely monitoring patients, particularly when they first exhibit side effects. Prompt intervention at the earliest onset can greatly enhance patient adherence to the treatment regimen. Equally vital is the need to ensure patients are comprehensively informed about potential side effects. Emphasizing the significance of persisting through initial challenges, in the broader context of lung cancer treatment, can make a substantial difference.

In addressing the challenge of prioritizing early intervention, several strategies can be considered, especially after pinpointing those more susceptible to side effects. Hence, our strategy is split on the basis of what stage a particular patient is in their 6 month therapy. We have categorized our patients based on two primary criteria: the progression of their treatment plan and their risk level for discontinuing therapy. By intersecting these two dimensions, we've identified nine distinct patient groups. Each group has its own tailored strategy to ensure optimal care and adherence to treatment. This classification of patient groups ensures a more personalized, proactive and structured approach to patient support in an attempt to reduce therapy discontinuation.

7.3 Recommended Business Strategies

Below is a list of strategies that we have come up with in order to help improve adherence to Osimertinib -

1. 1:1 Mentor
2. Support Groups
3. Educational webinars
4. Periodic follow Up
5. ALS Emergency Care
6. Incentivized Premium Plans
7. In app tracking and alerts

The strategies are then mapped to each group of patients depending on whether they belong to high/medium/low risk groups and how far along they are in their treatment plan.

STRATEGIES	HIGH			MEDIUM			LOW		
	1-2 Months	3-4 Months	5-6 Months	1-2 Months	3-4 Months	5-6 Months	1-2 Months	3-4 Months	5-6 Months
1:1 Mentoring	✓	✓		✓			✓		
Support Groups	✓	✓	✓	✓	✓	✓	✓	✓	✓
Educational Webinars	✓	✓	✓	✓	✓	✓	✓	✓	✓
Periodic Follow Up	✓	✓	✓	✓	✓				
ALS Emergency Care	✓	✓	✓						
Incentivized Premium Plans	✓	✓	✓	✓	✓	✓	✓	✓	✓
In-app Tracking & Alerts	✓	✓	✓	✓	✓	✓	✓	✓	✓

7.3.1 1:1 Mentoring

Strategy Rationale

The rationale behind implementing a one on one coaching solution for Humana's members experiencing side effects from the Osimertinib therapy is grounded in a proactive and personalized approach to improve medical adherence and member outcomes. We've seen that members are highly likely to discontinue treatment in the early stages of their therapy, hence there is a knowledge and communication barrier which is causing them to not realize that their side effects can be managed.

Strategy Implementation Plan

Humana can follow this step by step approach to implement the one and one coaching solution for its members:

1. Coaching Set-Up: Using our model, we can estimate the number of members who would need one-on-one coaching and determine the number of coaches needed. Members can leverage the MyHumana app to set up a feature of either a virtual session or a physical session with the coach.
2. Coach Training: We can train the coaches on Osimertinib therapy specifics, common side effects, empathetic listening and communication skills.

3. Patient Segmentation and Identification: Utilize the predictive model to identify and segment patients at high risk of discontinuation due to manageable side effects. Prioritize patients based on risk levels to ensure efficient resource allocation for coaching.
4. Coaching Sessions and Support: Conduct regular one-on-one coaching sessions with identified patients, addressing their concerns, explaining side effects, and providing strategies for effective management. Encourage open dialogue, active listening, and empathetic responses during coaching sessions.

Cost Analysis

The number of patients that are given 1:1 can be seen from the matrix ,i.e. People that fall under 'HIGH' probability of discontinuing. From the current holdout data , we observed that 10% people fall in the HIGH category. To obtain a number, we are taking the 2021 Humana's data that shows out of 1765 people, 24% tend to discontinue. Therefore, 176 patients will fall in this category of HIGH zone. We suggest that Humana arranges 1:1 monthly counseling with these patients for 6 months, it will be beneficial for the patients and also cost-efficient for Humana. Nextly, the cost per session taken by a wellness coach is about $75 per session. By multiplying 75*176*6 we obtained a cost of $13,200.

75$ per session [10] * 6 * 176 = $79,200

[10] (Gambill)

7.3.2 Support Groups

Strategy Rationale

Support groups can offer valuable psychological support to patients with lung cancer undergoing osimertinib therapy. Connecting patients with others facing similar challenges fosters a sense of community, reduces feelings of isolation, and provides emotional reassurance. Sharing experiences, coping strategies, and successes in a supportive environment can enhance patients' mental and emotional well-being.

Strategy Implementation Plan

Humana can take the following steps to implement this strategy to create and foster support groups. For Humana, facilitating support groups aligns with the company's commitment to holistic patient care, potentially improving treatment adherence and overall health outcomes.

1. Needs Assessment: Patients can take an assessment based on their psychological status and their feelings, preferences and availability. Assessments can be used to identify common challenges faced by the patients undergoing osimertinib therapy.

2. Develop community groups within Humana: Humana can create several community groups that can include patients who have undergone the non small-cell lung cancer targeted therapy and have continued it.
3. Online and Offline support forums: Both types of support forums can be created where patients can interact with each other. Online support sessions can be useful for patients that can have comfort from their homes and offline can be useful if they want personal human interactions.
4. Professional facilitators: They can keep guest-sessions from oncologists and nutritionists to keep sessions that are beneficial to the community members and where they can ask any questions that they have.
5. Regular meet-ups: Support groups can hold regular meetings to interact with each other and share stories so they feel supported and emotionally secure.

Cost Analysis

As the support groups will be formed by the people who are members of Humana, those wont have additional costs. The facilitator can create community groups and leaders can be assigned who drive the sessions or meet-ups. For occasional guest sessions by cancer specialists and psychologists, there is a cost of $500 per session. (Clark) . If guest sessions are held monthly, then annual costs would be $500 * 12= $6000.

7.3.3 Incentivized Premium Plans

Strategy Rationale:

This is an incentive based program that will help Humana foster a better relationship with their patients as well as enhance their adherence rate. By offering a discount on premiums contingent on the duration of therapy continuation (2 months, 4 months, and 6 months), patients will be more committed to the therapy and this will in turn also improve health outcomes. This will aid patients to face potential difficulties and side effects during the course of their therapy in return for tangible rewards on their premium plan. Such a tiered approach provides incremental rewards which will influence sustained engagement.

Strategy Implementation Plan:

1. Launch: The incentivized program will be launched through personalized communication channels which clearly depict the discount on premium tied to therapy duration.
2. Leveraging Communication Channels: The patients will receive notifications on this program throughout the duration through their preferred channel (in app notifications, emails, phone calls) and will be kept informed on their progress and eligibility.
3. Data Driven Targeting: Predictive analytics can be leveraged to perform targeted incentivization and the patients will be lured more during the initial stages of the therapy since that is when most of them opt to discontinue.

4. Seamless Premium Adjustment: The app will be integrated with an automated premium adjustment plan which will work in conjunction with the in-app pill tracking feature. As soon as the patient completed 2 months of therapy their premium for the next 2 months will be discounted by 30%. Similarly if they continue therapy for 4 months, their premium for the next 2 months will be discounted by 40%. Lastly, if they continue therapy for 6 months, their premium for the next 2 months will be discounted by 50%. Hence they will receive a discounted premium for a total of 6 months if they continue the therapy for 6 months.

Cost Analysis:

The costs associated with premium reduction can be calculated for the patients.There will be no substantial cost incurred to add a feature in the app for the incentive based premium reductions. Firstly, we identified that there are 494 patients likely to discontinue the therapy obtained from 2021 Humana's data out of 1765 patients. We are willing to provide incentive based rewards programs to each of these patients. Currently, 76% patients are continuing the therapy for more than 6 months. If we foresee that 85% will continue the therapy for 6 months then they will have the discount applied on their premiums. The projected average premium for a Medicare Advantage plan in 2023 is $18 per month. ("Compare Medicare Advantage Plans Side By Side") Therefore, the costs incurred to Humana will be

$(18 * (0.3) * 2 + 18 * (0.4) * 2 + 18 * (0.5) * 2) * 494 *2 = \$4,268$

7.3.4 Educational Webinars

Strategy Rationale

Educational Webinars are a great method to effectively reach out to a large population guiding members on how to manage side effects while undergoing the therapy. These webinars can have past survivors of Lung Cancer who can share their journeys and how the Osimertinib therapy has helped them which can inspire Humana members to adhere to their medication. For any member to withstand the side effects, they need to know that there is a way to decrease their effect and they also need to know the benefits of completing the therapy for the long term.

Strategy Implementation Plan

Humana can follow this step by step approach to implement the educational webinars solution for its members:

1. Webinar Content Development: Humana can collaborate with Healthcare professionals to make sure that the content of the webinars caters to the needs of the members undergoing the therapy.

2. Identify and Invite Expert Speakers: Survivors of cancer as well as Healthcare Specialists are the perfect people to host these webinars.
3. Conduct the webinars: Host these webinars through the platforms at our disposal.
4. Conduct Survey and Feedback: Receiving feedback on these webinars can contribute to its iterative development.

Cost Analysis

To host a webinar, there is a platform required and there are certain costs associated with it. To develop content , Humana can contact several healthcare organizations and professional doctors who can assist in developing good content for the patients curated as per the requirements facilitated by Humana. The costs incurred to Humana can be calculated by multiplying the number of webinars * the number of occurrences in a year = $500 * 20 = $10000 where we recommend Humana to organize 20 webinars annually , i.e. at least 1 webinar monthly.

7.3.5 ALS Emergency Care

Strategy Rationale

This strategy is grounded in the necessity to provide immediate and specialized assistance to Humana members as they are going through a chronic illness. The key considerations for coming up with this strategy were minimizing the adverse health outcomes that may result from side effect-related treatment interruptions, aligning with Humana's goal of improving health outcomes, ensuring patient safety, promoting adherence, and mitigating risks associated with therapy discontinuation. It underscores Humana's dedication to comprehensive patient care and the importance of optimizing therapeutic outcomes through timely and tailored support.

Strategy Implementation Plan

Humana can follow this step by step approach to implement the emergency care team solution for its members:

1. Care Team Tie-up : Tie-up with emergency care ambulance which has a team of nurses and pharmacists experienced in oncology and managing treatment-related side effects.
2. Emergency Response Protocols and Guidelines: Develop clear and standardized emergency response protocols to guide the care team's actions when responding to patient calls related to side effects. Establish guidelines for determining the severity of side effects and appropriate steps for intervention and escalation.
3. 24/7 Accessibility and Communication: Ensure the care team is accessible 24/7 to respond to patient calls promptly. Set up communication channels (phone lines, online platforms) to facilitate immediate communication between patients and the care team.
4. Integration with the Predictive Model: Integrate the predictive model insights into the emergency response system to prioritize high-risk patients for immediate care team

intervention. Customize response strategies based on the severity of side effects and predicted patient behavior.

5. Patient Outreach and Education: Reach out to eligible patients, informing them of the emergency care team's availability and the importance of reaching out in case of side effects.

Cost Analysis

Humana would need to collaborate with emergency care ambulance providers to provide emergency services to patients if they have ADEs like heart disease or require emergency care. The cost to partner with a healthcare provider is considered where Humana is connecting ambulance services to provide instant support to the patients at their door. Humana can bear the costs for the Advanced Life Support (ALS) emergency ambulance care ride that is $1277 per ride ("How Much Is An Ambulance Ride? Costs and Financing Options").

We obtained that 10% of lung cancer patients might require emergency care in their lives ("Lung cancer in the emergency department - Emergency Cancer Care").

This gives us the calculation as $1277 * 10% * 494(#patients who discontinued in 2021 in Huamana informational document) = $63,083.8

7.3.6 Periodic Follow-Up

Strategy Rationale

Our analysis found that most members drop off the therapy after their first fill. Hence, it is key to communicate with our members during the first few months i,e 30-60 days of the therapy. This strategy is based on the fact that different people respond to different modes of communication. Based on the member's preferred communication channel found through data analytics, we can target those communication channels to reach out to them and increase their response rate. This will in turn increase their medical adherence rate as we are successfully reaching out to them.

Strategy Implementation Plan

1. Data Collection/Analyze and Segmentation: Analyze existing member data or collect it from members, including treatment history, side effects experienced, communication preferences, and other relevant information. Segment members into risk categories based on the predictive model, identifying those at higher risk of therapy discontinuation.
2. Communication Channel Selection: Identify the preferred communication channels for each member segment (e.g., phone, email, text messages, mobile app) based on their communication history and preferences. Ensure the chosen channels align with members' accessibility and convenience.
3. Setup Follow-Ups: Establish a communication schedule based on the treatment timeline and critical points where members may face side effects or make decisions regarding therapy continuation.

4. Integration with Predictive Model: Integrate the engagement strategy with the predictive model to automate and streamline the identification of high-risk members and the delivery of personalized interventions.

Cost Analysis
This approach is an advanced version of expanding communication channels by better understanding the preferences of the patients. As Humana needs to include some changes to the mode of existing communication , there are no additional costs involved by this approach.

7.3.7 In app tracking and alerts

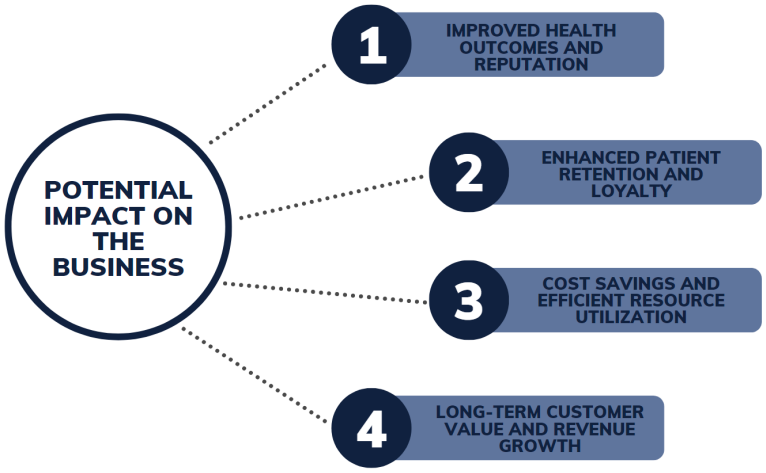
Strategy Rationale
This strategy will help members build accountability to adhering to the Osimertinib drug and allow Humana to send personalized reminders in case a member forgets to take their daily Osimertinib fill. Implementing in-app pill tracking for Osimertinib therapy members predicted to be at high risk of discontinuation due to manageable side effects is a forward-thinking strategy. It combines the power of technology, data analytics, and personalized engagement to improve adherence, empower patients, and ultimately enhance health outcomes by ensuring successful completion of the Osimertinib therapy.

Strategy Implementation Plan
Humana can follow this step by step approach to implement the In-App Pill Tracking and Usage solution for its members:
1. Developing and Integrating the feature: We can leverage the MyHumana application to add this functionality which will be used by not only Osimertinib therapy members but all the Humana members.
2. Education about the feature: Humana can educate its members about this feature allowing them to use it for their tracking purposes.
3. Personalized reminders based on the prediction model: We can harness the insights generated by our prediction model to send customized reminders to members based on their likelihood of dropping out of the therapy score.

Cost Analysis
The Humana app has several features for their members where they can add in-app tracking and usage. Further, analytics and personalized events can be generated based on algorithms to send the patients reminders. To calculate the cost of in-app feature development, we can multiply the development time of the feature * the hourly rate of a developer. We obtained the result by assuming 3 weeks to develop this feature that gives us 120 hours * $100 (Lastovetska) = $1200.

8. Potential Impact On Business

Humana has always stood for value-based care which helps its members manage their health and save money at the same time. The strategies and solutions discussed above will not only help Humana members improve their health and wellbeing but also help Humana as a business.



Improved Health Outcomes and Reputation	Such strategies will position Humana at the forefront since it is living up to its commitment to helping people lead a healthy and happy life. It will not only drive more positive outcomes but also improve patient adherence to the therapy. Such an approach puts the patient first which contributes to enhanced reputation for Humana in the healthcare industry.
Enhanced Patient Retention and Loyalty	With a more customer focused approach and personalized care, Humana can improve the rate of therapy continuity. This, in turn, leads to a higher patient retention rate since the people feel a sense of belonging. Thereby increasing loyalty towards Humana as their primary healthcare provider.
Cost Savings and Efficient Resource Utilization	A reduction in therapy discontinuation leads to a win-win situation for both the patient as well as the healthcare provider. A longer therapy continuation decreases the need for unnecessary treatment interruptions, hence better resource utilization. All of this in turn encourages cost savings for Humana.

Long-Term Customer Value and Revenue Growth	With a curiosity to learn how to better serve the members and patients, Humana fosters better relationships that serve for a long-term purpose - providing a way to lead healthy and meaningful lives. Greater patient adherence and engagement leads to an increase in the long-term customer value, thereby resulting in a sustained revenue growth. Satisfied and loyal customers will contribute to a greater return rate because of their continued association with Humana.

9. Cost Benefit Analysis

The Cost Benefit analysis is done for the Case study where we have obtained the data for the Humana members who might be facing several adverse health conditions and therefore might discontinue the therapy. The table shows the data for patients that might be hospitalized due to not continuing the therapy and might face challenges. They would have to go through several treatments and diagnostic procedures ("Understanding Total Cost of Care in Advanced Non-Small Cell Lung Cancer Pre- and Post Approval of Immuno-Oncology Therapies"). Therefore to analyze the cost outcome if the patients discontinued the therapy is $66,953,819 on the basis of the findings.

Description	Values
Total Humana Members taking Osimertinib therapy	1765
Number of people discontinuing the therapy(24%)	424
Diagnostic Procedures cost	$3,500 [15]
Treatment Procedures cost	$42,000 [16]
Rate of hospitalization	85%
Total hospitalization cost per person	$140,247 [17]
Cost beared by the patient	$8,000
Cost beared by the insurance provider	$132,247
Total Cost to Humana per year	$66,953,819

Further, we came up with several recommendations to prevent the patients from discontinuing the therapy which were explained earlier in the strategies. Each of the strategies can be followed by Humana that would give the patients more motivation and reasons to continue the therapy. We have supported the data with findings from the research and calculated the cost of the strategies.

Cost incurred due to each strategy	
1. 1:1 Mentor	$ 79,200.00
2. Support Groups	$ 6,000.00

Cost incurred due to each strategy	
3. Educational webinars	$ 10,000.00
4. Periodic follow Up	$ -
5. ALS Emergency Care	$63,083.8
6. Incentivized Premium Plans	$ 4,268.00
7. In app tracking and alerts	$ 1,200.00
Total Cost of Recommendations	$ 100,668.00
Net Savings	**$66,853,150.80**

To obtain the net savings, we have subtracted the cost that would incur to Humana by the hospitalization of patients - the cost obtained from the recommended strategies. As per the data, we can observe that the net savings is found to be $66,853,150.8. This compelling data underscores the significant financial benefits awaiting Humana while simultaneously enhancing patient outcomes and survival rates. It is our strong recommendation that Humana embraces these strategies wholeheartedly, aligning both financial prudence and patient welfare in a harmonious synergy.

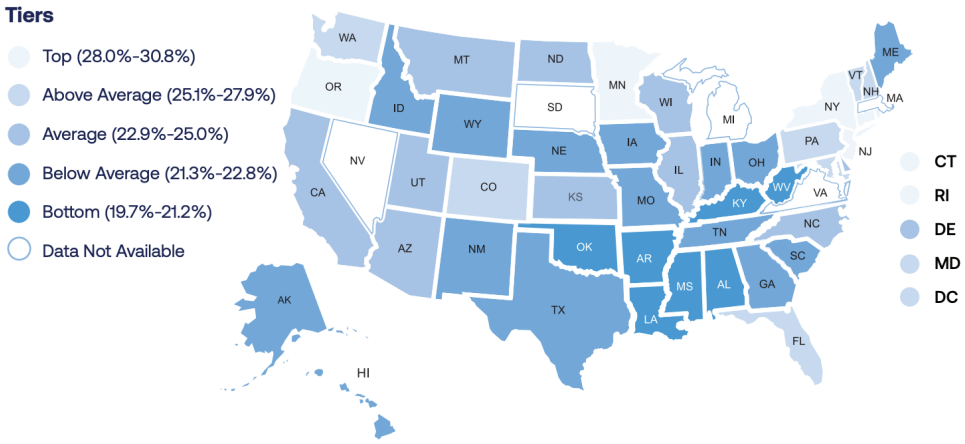
10. Scope for Improvement

In an ideal scenario, we'd possess all of the information we need to enhance our support for patients battling this deadly disease. Yet, reality often presents gaps in knowledge. In this section, we'll explore potential improvements in our strategy given that we are able to collect more information about our patients, and then we will delve into alternative approaches and methods that might have yielded even more effective solutions.

One of the main gaps in knowledge is the absence of information about which stage of lung cancer the patient is undergoing treatment. This information would have helped understand if there is a correlation between the cancer stage and the severity of side effects which eventually affects a patient's decision to discontinue treatment. Sometimes, if the side effects are severe, it is advised by doctors to stop treatment and taking this into consideration within our predictive model could have potentially improved our recommendations.

Secondly, having more information about the habits and pre-existing medical conditions could help improve the model's prediction. For example, smokers are more susceptible to lung cancer and hence, that could have an effect on whether or not a smoker continues treatment. Further, someone with co-morbidities like diabetes or heart conditions could have differences in the severity of their side effects and different tolerance levels when compared to an otherwise healthy human being with just lung cancer. Hence, we believe that such relevant medical information could prove to be helpful in making predictions.

Further, from our research on lung cancer, we understand that the location that a patient receives treatment plays an important role. For example, according to a study by the American Lung Association[18], the survival rate within the south eastern states in the USA is generally lesser than the rest of the country (see image below). Having information regarding the locality and in general the demographic information of each patient can help formulate improved strategies for each patient.



In reflecting upon alternative methodologies, one significant approach comes to mind: incorporating the 'date' field from our tables to transform the data into a time series structure. This would introduce time as a predictive factor, which was overlooked in our initial model. For example, we only considered the total number of times a person visited the pharmacy but not the frequency of visits. While there is no guarantee as to whether this would have improved the

accuracy of the model, it remains an unexplored avenue and represents a potential area for future enhancement.

11. References

1. “About TAGRISSO® (osimertinib) for Early-Stage EGFR+ NSCLC.” *Tagrisso*, https://www.tagrisso.com/early-stage-nsclc/about-tagrisso.html#about-wrapper-6. Accessed 15 October 2023.
2. Brownlee, Jason. “A Gentle Introduction to k-fold Cross-Validation -MachineLearningMastery.com.” *Machine Learning Mastery*, 4 October 2023, https://machinelearningmastery.com/k-fold-cross-validation/. Accessed 15 October 2023.
3. “Deconstructing ADAURA. It is Not Yet Time to Forgo Platinum-based Adjuvant Chemotherapy in Resected Early Stage (IB-IIIA) EGFR-mutant NSCLC.” *NCBI*, 19 May 2022, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9126226/. Accessed 14 October 2023.
4. “Documentation by example for shap.dependence_plot — SHAP latest documentation.” *the SHAP Documentation*, https://shap-lrjball.readthedocs.io/en/latest/example_notebooks/plots/dependence_plot.html. Accessed 15 October 2023.
5. “ICD Code Lists.” *CMS*, 6 September 2023, https://www.cms.gov/medicare/coordination-benefits-recovery/overview/icd-code-lists. Accessed 15 October 2023.
6. Iuga, Aurel O., and Maura J. McGuire. “Adherence and health care costs - PMC.” *NCBI*, 20 February 2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934668/. Accessed 15 October 2023.
7. “Lung Cancer Statistics | How Common is Lung Cancer?” *American Cancer Society*, 12 January 2023, https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html. Accessed 15 October 2023.

8. Nyberg, Kara. “Final ADAURA OS Analysis Reinforces Adjuvant Osimertinib as a Standard of Care for Patients With Stage IB to IIIA EGFR-Mutated Non–Small Cell Lung Cancer.” 4 June 2023, https://dailynews.ascopubs.org/do/final-adaura-os-analysis-reinforces-adjuvant-osimertinib-standard-care-patients-stage. Accessed 14 October 2023.
9. “Osimertinib After Surgery Significantly Improves Survival in Patients With Resected EGFR-Mutated Non-Small Cell Lung Cancer.” *American Society of Clinical Oncology*, 4 June 2023, https://old-prod.asco.org/about-asco/press-center/news-releases/osimertinib-after-surgery-significantly-improves-survival. Accessed 14 October 2023.
10. “Osimertinib in Advanced Lung Cancer with EGFR Mutations.” *National Cancer Institute*, 12 December 2019, https://www.cancer.gov/news-events/cancer-currents-blog/2019/osimertinib-lung-cancer-improves-survival-flaura. Accessed 14 October 2023.
11. O'Sullivan, Conor. “Analysing Interactions with SHAP. Using the SHAP Python package to… | by Conor O'Sullivan.” *Towards Data Science*, 4 December 2021, https://towardsdatascience.com/analysing-interactions-with-shap-8c4a2bc11c2a. Accessed 15 October 2023.
12. Sabate, E. “Adherence to Long-Term Therapies: Evidence for Action.” *World Health Organization*, 2003.
13. “State of Lung Cancer | Key Findings.” *American Lung Association*, https://www.lung.org/research/state-of-lung-cancer/key-findings. Accessed 15 October 2023.
14. “Treatment for Early-Stage EGFR+ NSCLC | TAGRISSO® (osimertinib).” *Tagrisso*, https://www.tagrisso.com/early-stage-nsclc.html. Accessed 15 October 2023.

15. “CT and Other Key Scan Costs.” *Health*, 18 November 2022, https://www.health.com/mind-body/6-key-medical-scans-and-what-they-should-cost. Accessed 15 October 2023.

16. “Lung cancer costs by treatment strategy and phase of care among patients enrolled in Medicare.” *NCBI*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6346221/. Accessed 15 October 2023.

17. American Cancer Society: "Things to Know About the Cost of Your Cancer Treatment," "Key Statistics for Lung Cancer," "The Costs of Cancer in 2015: 8.7 Million Years of Life and $94 Billion in Lost Earnings."

18. State of Lung Cancer, American Lung Association https://www.lung.org/getmedia/647c433b-4cbc-4be6-9312-2fa9a449d489/solc-2022-print-report

19. “Understanding Total Cost of Care in Advanced Non-Small Cell Lung Cancer Pre- and Postapproval of Immuno-Oncology Therapies.” *American Journal of Managed Care*, 19 October 2018, https://www.ajmc.com/view/understanding-total-cost-of-care-in-advanced-nonsmall-cell-lung-cancer-pre-and-postapproval-of-immunooncology-therapies. Accessed 15 October 2023.

20. Lastovetska, Anastasiia. “App Development Cost in 2023 | Types, Examples, Features.” *MLSDev*, 19 April 2023, https://mlsdev.com/blog/app-development-cost. Accessed 15 October 2023.

21. “How Much Is An Ambulance Ride? Costs and Financing Options.” *CareCredit*, https://www.carecredit.com/well-u/health-wellness/ambulance-ride-cost/. Accessed 15 October 2023.

22. “Lung cancer in the emergency department - Emergency Cancer Care.” *Emergency Cancer Care*, 6 March 2023, https://emergcancercare.biomedcentral.com/articles/10.1186/s44201-023-00018-9. Accessed 15 October 2023.

23. “Compare Medicare Advantage Plans Side By Side.” *Humana*, 31 August 2023, https://www.humana.com/medicare/medicare-resources/compare-medicare-advantage-plans. Accessed 15 October 2023.

24. Clark, Dorie. “How Much Should You Charge for a Speech?” *Harvard Business Review*, 3 May 2018, https://hbr.org/2018/05/how-much-should-you-charge-for-a-speech. Accessed 15 October 2023.

25. Gambill, Mac. “How Much Should You Charge As A Health Coach? (Updated for 2022).” *Nudge Coach*, 31 October 2022, https://nudgecoach.com/blog/how-much-to-charge-as-a-health-coach. Accessed 15 October 2023.