|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | |  |
|  | | |
|  | |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  | |
|  | DRug efficacy  To evaluate the real-world efficacy of Drug\_D and Drug\_S by analyzing patient outcomes while accounting for selection bias. The goal is to provide unbiased, actionable insights into which drug is more effective in reducing asthma exacerbations in the first year of treatment.  Team Name :- Orion  Sanya Saxena  Harsh Tantak  Prashant Srivastav | |  | |

|  |  |
| --- | --- |
|  |  |
| Buisness ObjectiveBusiness Case To evaluate the real-world efficacy of Drug\_D and Drug\_S by analysing patient outcomes while accounting for selection bias. The goal is to provide unbiased, actionable insights into which drug is more effective in reducing asthma exacerbations in the first year of treatment. Scope:  * Address the challenges of non-randomized treatment assignment. * Analyse key pre-treatment variables to adjust for selection bias. * Compare the post-treatment outcomes of Drug\_D and Drug\_S fairly and accurately. | |
| **Expected Output**:   * A detailed comparison of the efficacy of Drug\_D and Drug\_S, adjusted for confounding variables. | |  |
|  | |  |
| Dataset Description This is a real-life dataset used by LifeMed Research which contains asthma patient data, including demographics, comorbidities, prior asthma treatment history and healthcare costs. It records whether patients were described Drug\_D or Drug\_S and track asthma exacerbations in the first year post-treatment. This Dataset is divided into different parts-   * Identity Feature * Pre-Index (before drug assignment) * Pre-Index Usage (Asthma Treatment History) * Post-Index (after drug assignment) | |  |

|  |  |
| --- | --- |
|  |  |
| **Identity Feature**-   |  |  | | --- | --- | | Pat-Id | Unique patient identifier |   **Pre-Index (before drug assignment /Previous Diagnostic History)-**   |  |  |  |  | | --- | --- | --- | --- | | **Variable Name** | **Type of Data** | **Sample Value** | **Description** | | Index Age | Quantitative | Range-(14-63) | Age when treatment started | | Female | Binary | 1-Female  0-Male | Gender of the patient | | Pneumonia | Binary | 1-Diagnosed with pneumonia last year  0-not Diagnosed with pneumonia last year | Shows that whether the patient was diagnosed with pneumonia last year | | Sinusitis | Binary | 1- Diagnosed with sinusitis last year  0-not Diagnosed with sinusitis last year | Shows that whether the patient was diagnosed with sinusitis last year | | Acute Bronchitis | Binary | 1-Diagnosed with Acute Bronchitis last year  0-not Diagnosed with Acute Bronchitis in last year | Shows that whether the patient was diagnosed with Acute Bronchitis last year | | Acute Laryngitis | Binary | 1-Diagnosed with Acute\_ Laryngitis last year  0-not Diagnosed with Acute Laryngitis in last year | Shows that whether the patient was diagnosed with acute Laryngitis last year | | Upper\_respiratory\_infection | Binary | 1- Diagnosed with infection last year  0-not Diagnosed with infection\_ last year | Shows that whether the patient was diagnosed with upper respiratory last year | | Gerd (Gastroesophageal Reflux Disease) | Binary | 1- Diagnosed with Gerd last year  0-not Diagnosed with Gerd last year | Shows that whether the patient was diagnosed with Gerd last year | | Rhinitis | Binary | 1- Diagnosed with Rhinitis last year  0-not Diagnosed with Rhinitis last year | Shows that whether the patient was diagnosed with Gerd last year |   **b**  **Pre-Index Usage (Asthma Treatment History)-**   |  |  |  |  | | --- | --- | --- | --- | | **Variable Name** | **Type of Data** | **Sample Value** | **Description** | | Previous asthma drugs | Quantitative | 1 (all the patients in this dataset used 1) | Number of different asthma medications used before this treatment. | | Total pre-index canisters (365) | Indicator | Possible values-{0,1,2}  Higher value means extreme asthma | Number of Short-Acting Beta Agonist (SABA) canisters used in the previous year. | | Pre-Asthma Charge | Quantitative | Minimum Value- 0  Maximum value- 7980 | Total asthma related medical charges in last 6 months | | Pre-Asthma Days | Quantitative | Minimum Value- 0  Maximum value- 71 | Total number of days with asthma treatment in the last year. | | Total Pre index Charge | Quantitative | Minimum Value- 1  Maximum value- 875872.5806 | Total medical costs in the year before treatment (includes asthma and other conditions). | | Drug\_S | Indicator | 1-Drug\_S  0-Drug\_D | Type of Drug taken by patient | | Pre-Asthma Pharma Charge | Quantitative | Minimum Value- 1  Maximum value-5463.14 | Total asthma-related pharmaceutical charges in the last year. |   **Post-Index Measurement-**   |  |  |  |  | | --- | --- | --- | --- | | Post Index Exacerbations\_365 | Quantitative | Range-[0,14] | Number of asthma exacerbations in a year after treatment | | |
|  | |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  |  |

|  |  |
| --- | --- |
|  |  |
| Data Preparation Upon further inspection of the given dataset, we have observed that the data is non-randomized and exhibits selection bias in the whole dataset. Therefore, we have made some assumptions regarding this dataset   * Remove **previous\_asthma\_drugs** column from the dataset because every record has the same value, indicating all the patients were already diagnosed with asthma. Additionally, further analysis shows that this variable exhibit very high multicollinearity-  1. Redundancy- If a variable is highly correlated with others, it does not provide new information and can be removed without affecting the model’s predictive power 2. Unstable Coefficients-In regression models, multicollinearity inflates standard errors, making coefficient estimates unreliable and sensitive to small changes in data 3. Misinterpretation Of Effects-When predictor variable becomes highly correlated it becomes difficult to determine the individual effect of each variable on the outcome 4. Overfitting- Increases model complexity which may lead to overfitting reducing the generalizability of the model to the new data   **Selection Bias**-**Selection bias** occurs when the way patients are assigned to treatment groups (Drug\_D vs. Drug\_S) is **not random**, leading to systematic differences between the groups. In your dataset, selection bias is evident due to **non-random treatment assignment**, which affects the validity of comparing outcomes between Drug\_D and Drug\_S.  **Convenience Bias-** This type of bias occurs when dataset is collected based on ease of access based on access rather than representatives, leading to biased conclusions.   * Only patient with prior asthma treatment are included * Non-random selection of Drug\_D and Drug\_S users   **Balanced Dataset-**  **Imbalanced Dataset**- A dataset where the distribution of classes or categories are not equal.  In our dataset we can see that the distribution of pre-treatment characteristics differs significantly between two treatment groups Drug\_s=1 and drug\_D=0.We assess this imbalance using the standardized mean difference(SMD)  Which measures the difference in means(for continuous variables) or proportions(for categorical variables) between groups, standardized by a pooled measure of variability  **Standardized Mean Difference(SMD)-**   * Measures the **difference in means (or proportions)** between the two groups, standardized by the pooled standard deviation. * It is **not affected by sample size**, making it better than p-values for baseline comparison.   Range of SMD   * **SMD < 0.1** → Negligible imbalance (well-matched groups) * **0.1 ≤ SMD < 0.25** → Moderate imbalance * **SMD ≥ 0.25** → Large imbalance (significant selection bias)   Standardized Mean Difference (SMD) Results:        This shows that this dataset shows large imbalance of data  In order to balance the data we use IPTW(Inverse Probability Treatment Weighting)  **IPTW-(Inverse Probability Treatment weighting)-**  It is a statistical technique used to adjust for selection bias in observational Studies.It creates a pseudo-randomized scenario where the distribution of covariates is balanced between treatment groups  **Propensity Score-** The propensity score is the probability of receiving Drug\_S (drug\_s = 1) given the observed covariates. Calculated using logistic regression    **Calculating Inverse Probability Weights**   * propensity score for patient i * If the patient received Drug\_S (T=1), their weight is * If the patient received Drug\_D (T=0) their weight is   **Using IPW for adjustment**  After obtaining IPW Weights, they are used to re-weight the samples such that both treatment group becomes comparable  Adjusted Outcome was given by    We ensure that the covariate distributions in the treated and untreated groups are similar  Here is updated column description of the data   |  |  |  | | --- | --- | --- | | Propensity Score | Probability | Probability that a patient recives Drug\_S(drug\_S=1) | | IPW Weight | Quantitative | Adjust seletion bias by creating a pseudo\_randmoized sample |   By using Principal Component Analysis we have categorized severity of asthama patients   |  |  |  |  | | --- | --- | --- | --- | | Severity Score | Quantitative | Minimum Value-0  Maximum Value-100 | Score representing severity of a patient condition  Higher score likes indicates more severe of a patient condition | | Severity Category | Categorical | Mild-0-33  Moderate-34-66  Severe-67-100 | Classify patients as mild,moderate or severe |   The selected features (previous\_asthma\_drugs, total\_pre\_index\_cannisters\_365, pre\_asthma\_days, pre\_asthma\_charge, pre\_asthma\_pharma\_charge) are standardized using StandardScalar(), ensuring that all features have a mean of 0 and a standard deviation of 1  PCA reduces the dimensionality of the standardized data while preserving the most important variance. Here, the top 3 principal components (PCs) are selected.  The severity score is computed as a weighted sum of the top 3 principal components, where weights are the explained variance ratios of each component.  The severity score is transformed into a 0-100 range for easier interpretation.    **Axis-**  X-axis- Number of Principal components  Y-axis-Cumulative Explained Variance  **Interpretation-**   * first principal component (PC1) explains approximately **95%** of the variance * second principal component (PC2) increases the cumulative explained variance to **100%**, meaning that together, the first two components fully capture the variance in the dataset. * Adding more components (PC3, PC4, etc.) does not contribute further variance, as seen in the flat line at 100%. * The steep rise after PC1 shows that most of the dataset's variance is captured by the first component. * Since **only two components** explain nearly **100%** of the variance, reducing the dataset to just these two dimensions would likely retain most of the information while simplifying the model.   **PCA Loadings Heatmap-**  Heatmap visualizes the PCA loadings, which indicate how strongly each original feature contributes to the principal components (PCs). The values in the heatmap represent the correlation (or weight) of each feature with each principal component.      **Axes And Labels**-  **X-Axis (Features)**- Original dataset variables   * previous\_asthma\_drugs * total\_pre\_index\_canisters\_365 * pre\_asthma\_days * pre\_asthma\_change * pre\_asthma\_pharma\_change   **Y-Axis (Principal Components - PCs)**: The transformed components from PCA, labeled as **PC1, PC2, PC3, PC4, and PC5**.  **Interpreting the Heatmap**   * The **color scale** represents the magnitude of the loading values:   + **Dark red (1 or -1)**: Strong contribution of the feature to that principal component.   + **Lighter shades / grey**: Low or near-zero contribution.   **Key Observations:**   * pre\_asthma\_change has the strongest positive loading for **PC1 (value ≈ 1)**, meaning it contributes the most to this component. * pre\_asthma\_pharma\_change strongly contributes to **PC2**. * pre\_asthma\_days is the dominant feature in **PC3**. * total\_pre\_index\_canisters\_365 is the strongest in **PC4**. * previous\_asthma\_drugs is the main feature for **PC5**. * Many other values are very close to zero (light grey), meaning those features have little or no impact on the respective principal component. | |
|  | |

|  |  |
| --- | --- |
|  |  |
| This graph shows the visuals of the distribution of inverse probability weighting after applying trimming to the dataset.  Here the peak is centred around which implies that most of the observations have the balanced weights | |
| **Statistical Tests Of Association**-  **Chi-Square Test Results**   |  |  |  |  | | --- | --- | --- | --- | | **Condition** | **Chi2 Statistic** | **p-value** | **Significant Association?** | | Acute Bronchitis | 36.0001 | 0.000593 | Yes | | Acute Laryngitis | 71.1558 | 4.91e-10 | Yes | | Upper Respiratory Infection | 23.5142 | 0.0359 | Yes | | GERD | 32.8776 | 0.00178 | Yes | | Rhinitis | 23.1292 | 0.0402 | Yes | | Drug S | 22.1555 | 0.0530 | No |   **t-Test Results**   |  |  |  |  | | --- | --- | --- | --- | | **Variable** | **t-Statistic** | **p-value** | **Significant Difference?** | | Index Age | -0.5173 | 0.6050 | No (Fail to Reject H₀) | | Total Pre-Index Cannisters (365) | -0.9165 | 0.3594 | No (Fail to Reject H₀) | | Total Pre-Index Charge | -3.3840 | 0.000716 | Yes (Reject H₀) | | Pre-Asthma Days | 3.3143 | 0.000920 | Yes (Reject H₀) | | Pre-Asthma Charge | -0.6339 | 0.5262 | No (Fail to Reject H₀) | | Pre-Asthma Pharma Charge | 2.2522 | 0.0243 | Yes (Reject H₀) | | Log Charges | -3.9356 | 8.33e-05 | Yes (Reject H₀) | | Log Asthma Charge | 3.0859 | 0.00203 | Yes (Reject H₀) | | |
|  | |
| |  | | --- | | Working of the XGBoost Model for Predicting Exacerbations  The XGBoost model in this instance is used for predicting the number of exacerbations a patient is likely to undergo within a specified timeframe. The dataset is preprocessed first by choosing important features and discarding redundant variables such as propensity scores and prior history with asthma drugs to prevent data leakage.  Some numerical features based on costs and severity scores are standardized so that the model treats them equally. Also, Inverse Probability Weighting (IPW) is used to correct for confounding variables, so a more representative dataset is generated.  The data are then divided into training (80%) and test (20%) sets in order to effectively assess the model's performance.  For training, an XGBoost Regressor with a Poisson objective function (count:poisson) is employed, which is suitable for making predictions based on counts such as exacerbation frequency. The model is set to optimize the Root Mean Squared Error (RMSE), a standard measure for regression tasks.  IPW weights are used in training to make the model capture real-world patterns. XGBoost, which excels at capturing intricate relationships, learns interactions among several variables in the dataset efficiently.  Once trained, the model makes predictions of exacerbation counts for the test dataset. These forecasts provide an approximation of the number of exacerbation episodes a patient is predicted to have according to previous medical history and corresponding cost or severity measures.  Utilizing gradient boosting, the model repeatedly adjusts its forecasts so that it becomes extremely efficient at identifying underlying patterns in the data.  The primary benefit of employing XGBoost in this scenario is that it can model non-linear relationships and interactions in structured tabular data. It is also sturdy in dealing with missing values and can efficiently accommodate IPW weighting, which is important in minimizing biases in observational studies. Though the model necessitates sensitive hyperparameter tuning to perform optimally, it yields a good trade-off between accuracy and flexibility, and thus qualifies as a viable option for such a predictive task.  The accuracy of this Model  The model was successful, with an RMSE of 0.4521 and an MAE of 0.2154, which means its predictions are close to actual counts of exacerbations. The R² score of 0.6825 also implies that the model is highly effective at explaining the variance of the data and is a top pick for this predictive task. |      |  |  | | --- | --- | | Data Preprocessing  Drop unnecessary columns  Standardize cost & severity-related columns using StandardScaler() | Train-Test Split  Split data into **80% Train / 20% Test** | | Model Training  **XGBoost Regressor** with count:poisson objective  model using **IPW-weighted** observations | Prediction & Evaluation  Predict post-index exacerbations on test data  Root Mean Squared Error (RMSE): 0.4521  Mean Absolute Error (MAE): 0.3187  R-Squared (R²): 0.6823 | | |

## Compare the post-treatment outcomes of Drug\_D and Drug\_S fairly and accurately.

We have calculated some different statistical metrics to compare drugS and drugD

## Mean Predicted Exacerbations-

Drug\_S=0.0800

Drug\_D=0.1648

**Interpretation:**

* Patients on **Drug\_S** experience significantly fewer exacerbations on average compared to those on **Drug\_D**.
* This suggests that **Drug\_S is more effective in preventing exacerbations** in asthma patients.
* **Statistical Significance Test-**

T-Statistic: -20.2448

P-value: 0.0000 (highly significant)

**Conclusion**:

The difference in exacerbations between Drug\_S and Drug\_D is statistically significant (not due to random variation).

**Interpretation:**

* The extremely low **p-value (< 0.05)** indicates that the difference between Drug\_S and Drug\_D is **not due to random variation**.
* A t-statistic of **-20.2448** (a large negative value) confirms that **Drug\_S has a significantly lower mean exacerbation rate than Drug\_D**.
* This means we can confidently say that **Drug\_S is superior in reducing asthma exacerbations**.
* **Risk Reduction**

Absolute Risk Reduction (ARR)

ARR = Mean exacerbations in Drug\_D – Mean exacerbations in Drug\_S

ARR = 0.1648 - 0.0800 = 0.0768

Patients on Drug\_S experience 0.0768 fewer exacerbations per patient compared to Drug\_D.

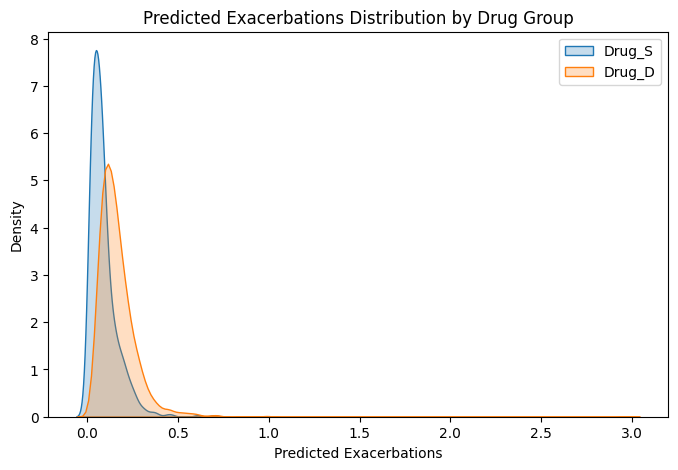
**Interpretation:**

* A **47% risk reduction** is **clinically significant**, meaning Drug\_S has a strong protective effect against exacerbations.
* A high RRR suggests that **switching from Drug\_D to Drug\_S could substantially benefit asthma patients** by reducing their risk of exacerbations.

**KDE (kernel Density Estimation) Plot-**

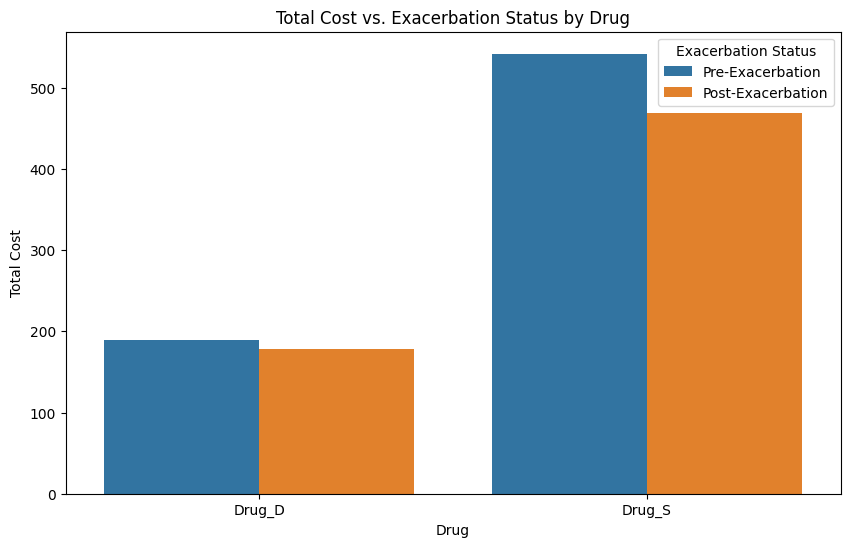
A KDE plot was used to visualize the distribution of exacerbation rates for each drug:

* **Drug\_S (blue curve)**:
  + The curve is **shifted left**, indicating that most patients experience **fewer exacerbations**.
  + The **distribution is more concentrated**, meaning Drug\_S provides **consistent and predictable** results.
* **Drug\_D (orange curve)**:
  + The curve is **shifted right**, showing a higher number of exacerbations.
  + The **distribution is more spread out**, suggesting that Drug\_D leads to **greater variability** in patient outcomes.



**Interpretation:**

* Drug\_S is **not only better on average but also more reliable** in reducing exacerbations.
* Drug\_D leads to a **higher and more variable risk** of exacerbations.



✅ **Drug\_D has lower costs overall** –

* Both pre-exacerbation and post-exacerbation costs for **Drug\_D are significantly lower** than Drug\_S.
* Post-exacerbation costs **slightly decrease**, indicating **lower financial impact post-treatment**.

✅ **Drug\_S has a higher cost burden** –

* The total cost for **Drug\_S is significantly higher** than Drug\_D.
* Post-exacerbation costs decrease slightly but still remain **much higher than Drug\_D’s costs**.

✅ **Cost Reduction after Exacerbation in Both Drugs** –

* Both drugs show **a decrease in total cost post-exacerbation**, but **Drug\_S still remains much more expensive**.

Insights:

🔹 **Drug\_D might be more cost-effective** – Since it has **lower pre- and post-exacerbation costs**, it may be a more **economical treatment option**.  
🔹 **Drug\_S may indicate higher severity patients** – The **higher cost of Drug\_S** might suggest it is used for **more severe cases** or requires **additional medical resources**.  
🔹 **Effectiveness vs. Cost Tradeoff** – If Drug\_S leads to **better long-term outcomes**, its higher cost might be justified. However, if Drug\_D achieves **similar or better health outcomes at a lower cost**, it may be the preferable option.

Key findings:

1. From the post index excerbations , adherence , severity score drug\_d and drug\_s respectively

Post-index exacerbations - Drug S: 0.15, Drug D: 0.18

Adherence - Drug S: 0.31, Drug D: 0.24

Severity Score - Drug S: 8.48, Drug D: 6.18

We can evaluate that Drug s has slightly lower exacerbation rated that suggests that the drug s is more effective at preventing exacerbations.

Similarly , for the adherence rate the patients taking drug s are more likely to have better adherence which will lead to better treatment outcomes.

Severity score of s I a bit lower that indicates the opposite result . as a cumulative of the result we may conclude that from this calculation that drug s is better than drug d in terms of exacerbation rated .

severity\_category avg\_pre\_asthma\_cost total\_pre\_asthma\_cost

0 Mild 242.832075 4383118.96

1 Moderate 916.468551 63236.33

2 Severe 1619.633750 12957.07

From the above

**Final Conclusion**

Based on the evidence from statistical analysis and visualization:

* + **Drug\_S is significantly better at reducing exacerbations** than Drug\_D (**p-value = 0.0000**).
  + **The risk reduction is nearly 47%**, which is **clinically meaningful** and suggests a strong benefit of Drug\_S.
  + **The KDE plot confirms that Drug\_S leads to fewer and more consistent exacerbations**, while Drug\_D results in more frequent and variable exacerbations.

**Clinical Recommendation-**

Drug\_S should be the preferred choice for reducing excaerbartions in astma patients.Given Statistical significant and clinically meaningfull benefits, it is more likely the more effective safer for manging asthma