Project Phase-2

Akhil Venkata Shiva Sai Gorugantu
Person Number: 50606819

How do the post-operative trends accompany the patients being admitted?

Problem Statement and Dataset Complexity

Question: How do the post-operative trends accompany patients being admitted?

Objective: Understanding post-operative trends is directly linked to patient outcomes. Complications leading to extended stays can affect hospital bed availability, especially in departments like the ICU. Analyzing these trends can assist in predicting patient recovery trajectories and enable hospitals to anticipate patients requiring more attention, thereby helping in resource allocation and patient flow optimization.

Dataset Complexity:

- The dataset contained multiple complex features, including patient-specific identifiers and post-operative condition labels.
- **Data Preparation**:
  1. **Merging Datasets**: Two datasets were merged using a unique key combining `MRN` (patient unique identifier) and `Log_ID`, which required careful handling due to duplicate entries for certain patients.
  2. **Cleaning and De-duplication**: There were duplicate records with identical unique keys. I cleaned the data to ensure only unique records for each patient, which required extensive data manipulation.
  3. **Target Labels**: The target variable (`Post_OP_type_Category`) initially had redundant classes (e.g., no entries for "chronic pain"), necessitating adjustments in target labels to ensure correct multiclass classification was possible because the one-hot encoding for the "chronic pain" was completely 0 so, I had to rearrange the classes present in the dataset.
  4. I also had to resample the data present in the dataset as the data distribution of each class varied a lot. Hence, I had to oversample (classes with more samples) or undersample(classes with fewer samples) so that we had a uniform distribution of all the classes in the dataset.

Model Selection: Why XGBoost?

XGBoost was selected as the primary algorithm due to its:

- **Strength in Tabular Data**: XGBoost is well-known for handling structured data, especially when there are complex feature interactions.
- **Feature Importance**: The ability of XGBoost to compute feature importance directly makes it ideal for healthcare applications, where understanding the impact of each feature on the outcome is crucial.
- **Handling Imbalanced Data**: XGBoost has parameters to handle imbalanced datasets, making it effective for scenarios where certain post-operative conditions were underrepresented.

Additionally, the interpretability of the model through feature importance and other evaluation metrics aligns with the project's goal to understand influential factors affecting post-operative patient trends.

Training and Tuning the Model

- **Model Configuration**:
  - The model was configured with the `multi:softprob` objective to handle multiclass classification, which outputs probability distributions across all classes.
  - **Booster**: `gbtree` was used as the booster, ideal for classification problems on structured data.
- **Hyperparameter Tuning**:
  - Tuning involved adjusting the learning rate, maximum tree depth, and number of boosting rounds to prevent overfitting and achieve optimal results.
  - Through cross-validation, we identified the best parameters to ensure the model generalizes well.
- **Training Loss vs. Test Loss**:
  - The loss curves indicate steady learning with both training and validation losses decreasing over boosting rounds, demonstrating effective training without overfitting.

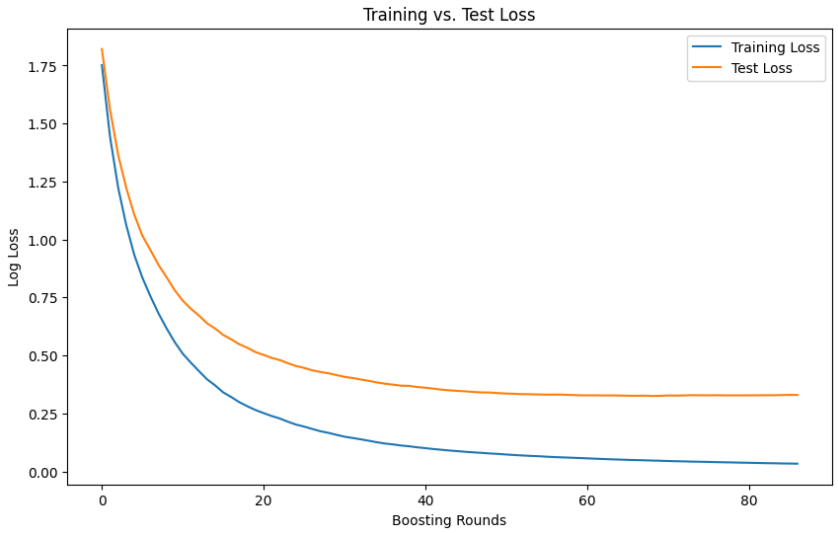

Figure 1: Training vs. Test Loss

Model Evaluation and Insights

To evaluate the XGBoost model's effectiveness, multiple metrics and visualizations were used:

1. **Feature Importance**:
   - The feature importance plot shows that `OR_LOS_HOURS`, `WEIGHT`, `HEIGHT_METRES`, and `LOS` are the most significant features, aligning with medical expectations.
   - From the importance Graph we can see that our hypothesis of the effect of whether a patient is admitted to ICU will affect the Post-operative conditions rather it is dependent on the above conditions.

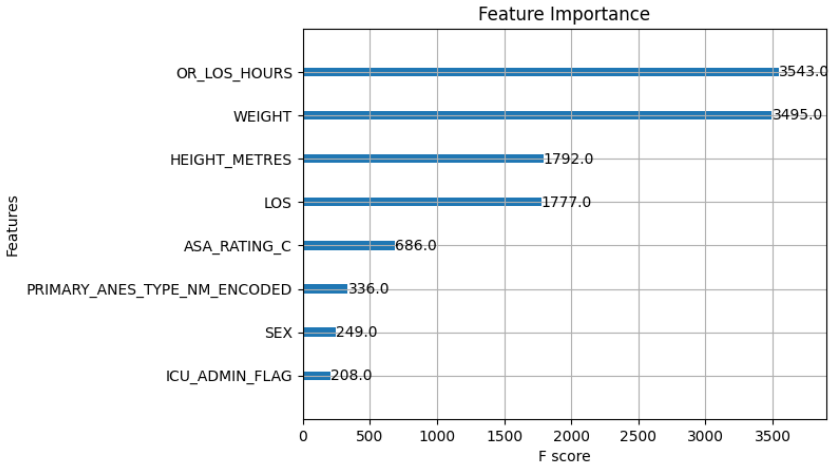

Figure 2: Feature Importance Plot

2. **Correlation Matrix**:
   - The correlation matrix reveals dependencies between features, such as a moderate correlation between `WEIGHT` and `HEIGHT_METRES`.

Figure 3: Correlation Matrix

3. **Precision-Recall Curve**:

- High average precision for most classes indicates effective discrimination, which is crucial for detecting rare post-operative complications.

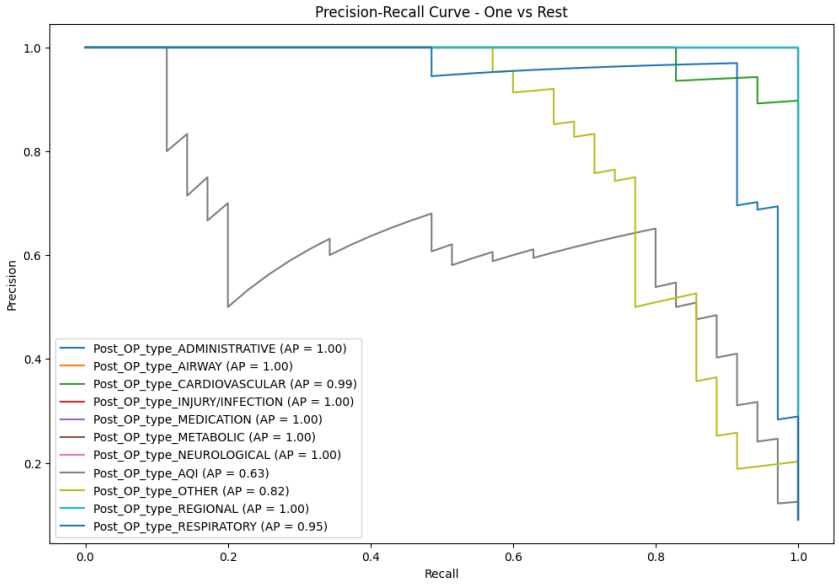

Figure 4: Precision-Recall Curve for Each Class

4. **ROC-AUC Curve**:

- The high AUC values for each class demonstrate strong classification performance across multiple categories.

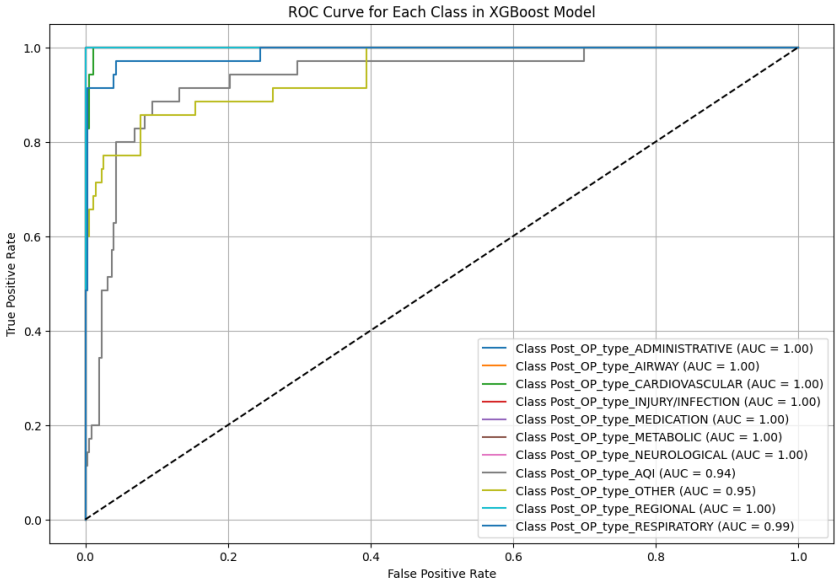

Figure 5: ROC Curve for Each Class in XGBoost Model

5. **Confusion Matrix**:

- The matrix shows high accuracy for dominant classes and highlights areas for potential improvement in underrepresented classes.

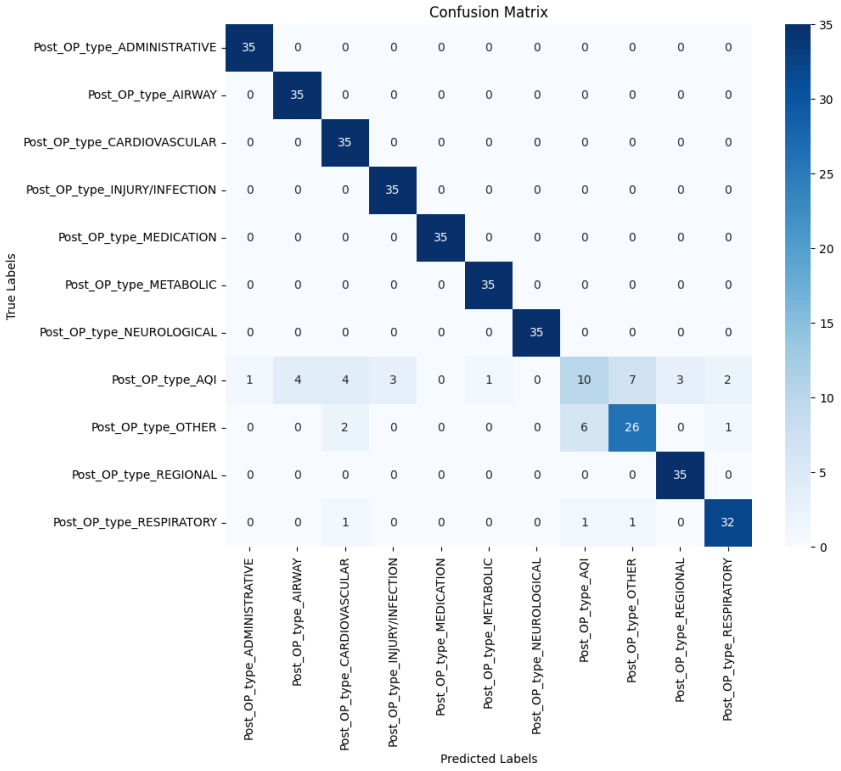

Figure 6: Confusion Matrix for XGBoost Model

Justification of Model Effectiveness

The effectiveness of the XGBoost model was demonstrated through:

- **Robust Multiclass Classification**: The model achieved high precision, recall, and AUC values, proving its suitability for complex multiclass predictions in healthcare.
- **Interpretability**: Feature importance and visualization metrics allowed insights into the most influential factors affecting post-operative outcomes.
- **Adaptability**: XGBoost's ability to handle structured tabular data and imbalanced class distributions made it an ideal choice for this medical dataset.

Conclusion and Insights Gained

The application of XGBoost to this dataset enabled us to derive valuable insights into post-operative trends:

- **Critical Factors**: Factors like `OR_LOS_HOURS` and `WEIGHT` were identified as critical predictors for post-operative outcomes.
- **Resource Optimization**: By predicting likely post-operative complications, hospitals can allocate resources more efficiently.
- **Actionable Predictions**: The model provides a reliable way to flag patients who might need additional post-operative care, aiding in preemptive actions.

The XGBoost model proved to be a powerful tool for analyzing and predicting post-operative trends, providing actionable insights that align well with the goals of enhancing patient outcomes and streamlining hospital resource management.

Analyzing Sex, ICU Admission, and Discharge Disposition

Question 2: How does the Sex of a particular person affect the Discharge disposition (where they go after discharge) and who is admitted to the ICU more times? What is the relation between the Length of Stay vs whether the patient is admitted to the ICU, with the help of the Sex feature?

Model Choice and Justification

For this question, we selected the **CatBoost** model. CatBoost is designed to handle categorical features efficiently without requiring extensive preprocessing like one-hot encoding, which makes it ideal for tabular data containing both categorical and numerical features. Given the categorical nature of variables such as `Sex` and `Discharge Disposition`, and the need to understand complex interactions with ICU admissions and Length of Stay, CatBoost was a suitable choice.

Differences between XGBoost and CatBoost

While both XGBoost and CatBoost are gradient boosting algorithms, they differ in handling categorical data and managing overfitting:

- **Handling Categorical Data**: CatBoost natively supports categorical variables, allowing automatic processing without manual encoding, preserving data structure. XGBoost requires explicit transformations like one-hot encoding.
- **Overfitting Prevention**: CatBoost uses ordered boosting to reduce prediction bias and enhance generalization, while XGBoost primarily relies on regularization techniques like L2 and L1.

CatBoost was therefore chosen for Q2, as it facilitates better interpretation of categorical variables and performs robustly with minimal preprocessing.

Model Training and Tuning

To capture relationships between `Sex`, `ICU_ADMIN_FLAG`, `DISCH_DISP`, and `LOS` (Length of Stay), we incorporated additional features like `WEIGHT`, `HEIGHT_METRES`, `OR_LOS_HOURS`, and `ASA_RATING_C`. This expanded feature set enabled the model to form meaningful connections, improving interpretability and accuracy.
I have truncated the number of classes present in the Discharge Disposition Class as there are many classes which are having very less presence in the dataset (like 2 rows in a dataset of 54k rows) so i truncated the classes to classes which are having rows of above 500 so that if we even sample we will not have samples which are not identical

Key tuning steps included:

- **Iterations**: Setting the number of boosting rounds for optimal convergence.
- **Learning Rate**: Adjusted to balance convergence rate and avoid overfitting.
- **Depth**: Set to capture sufficient complexity while avoiding excessive computation.

Metrics and Effectiveness

The CatBoost model's performance in predicting discharge disposition and understanding ICU admission trends by sex was assessed through various metrics and visualizations. These insights helped in analyzing the effectiveness of the model and deriving actionable intelligence.

- **Confusion Matrix**: The confusion matrix (Figure 7) illustrates the model's predictions against the actual discharge dispositions, offering a clear view of where misclassifications occur. High accuracy for major classes such as routine discharges and specific ICU-related categories highlights the model's ability to distinguish between common discharge outcomes. However, some misclassifications in less frequent categories indicate that further tuning may be needed for rare discharge types, which could be impacted by patient sex and ICU stay patterns.
- **ROC Curve**: The ROC curve (Figure 8) provides a detailed breakdown of the model's ability to distinguish between discharge disposition categories. Each curve represents one discharge type, with high Area Under Curve (AUC) values indicating that the model is effective at correctly predicting each class. The nearly perfect AUC scores for specific classes demonstrate the model's capability to differentiate cases, especially in understanding how patient sex influences outcomes like ICU admission and discharge location. This helps in validating that the model is reliable for decision-making in hospital resource allocation.
- **Training vs. Validation Loss**: The training and validation loss curves (Figure 9) showcase the model's learning progression over iterations. A steady decline in both curves indicates that the model is learning effectively, without significant overfitting. This is crucial for complex medical datasets, as overfitting could lead to inaccurate predictions in real-world applications. The convergence of training and validation losses suggests that the model generalizes well, which is essential for providing reliable predictions in medical settings where patients vary greatly.
- **Feature Importance**: The feature importance plot (Figure 10) identifies key predictors influencing the model's decisions. Notably, `LOS` (Length of Stay) has the highest importance, suggesting it's a strong determinant for discharge disposition and ICU admission patterns. This is followed by `HEIGHT_METRES` and `WEIGHT`, indicating that physical characteristics are also significant in predicting outcomes. The importance of `ICU_ADMIN_FLAG` supports the hypothesis that ICU admission status is directly linked to discharge outcomes, and patient `SEX` further influences these trends, which validates the use of these features in addressing the question.

Insights and Practical Relevance

This analysis provides critical insights for hospital resource planning and personalized patient care:

- **Discharge Disposition Predictions**: Understanding discharge disposition based on sex and ICU admission trends enables hospitals to better anticipate bed availability, particularly for patients likely to require prolonged ICU stays. This helps in managing patient flow and ensuring that resources are allocated effectively.
- **ICU Admission Trends by Sex**: The model's predictions reveal sex-based patterns in ICU admissions and discharge locations, which could be valuable for personalized healthcare strategies. For instance, female and male patients might have differing recovery trajectories and discharge needs post-ICU, necessitating tailored care plans.
- **Resource Allocation and Length of Stay**: Accurately predicting the Length of Stay for ICU patients aids in discharge planning and resource allocation. Given the high importance of `LOS` and `ICU_ADMIN_FLAG` in the feature importance plot, hospitals can make informed decisions about patient management, reducing overcrowding and optimizing turnover.
- **Patient Care Improvements**: By understanding the discharge patterns and ICU admissions, hospitals can enhance patient satisfaction and care quality. For example, high-risk patients identified by the model could receive prioritized care, reducing complications and improving outcomes.

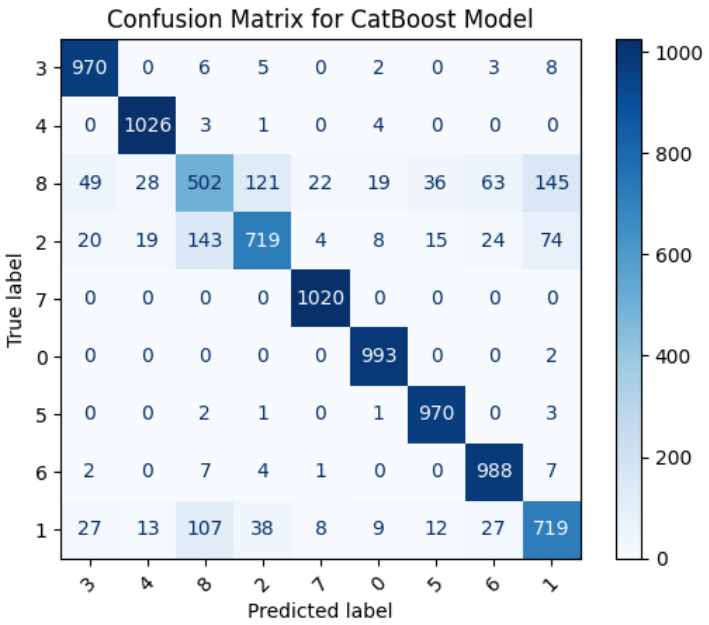

Figure 7: Confusion Matrix for CatBoost Model

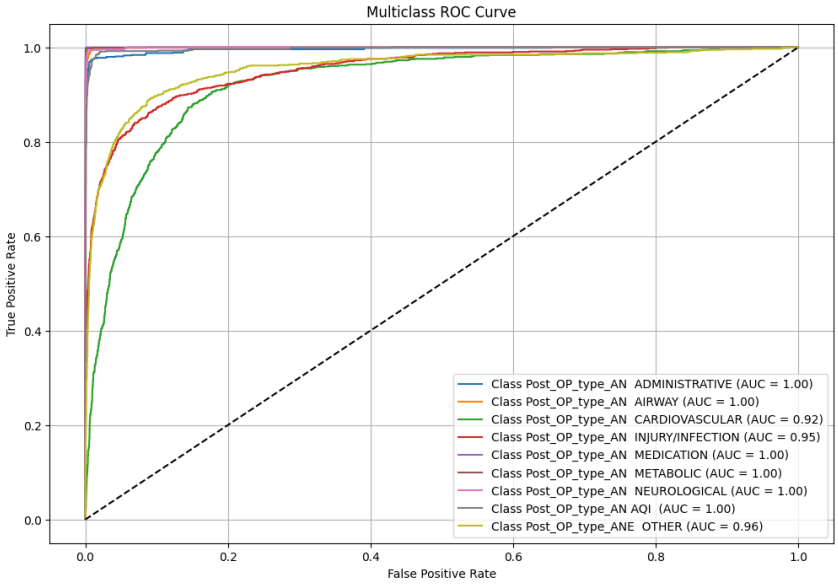

Figure 8: Multiclass ROC Curve for CatBoost Model

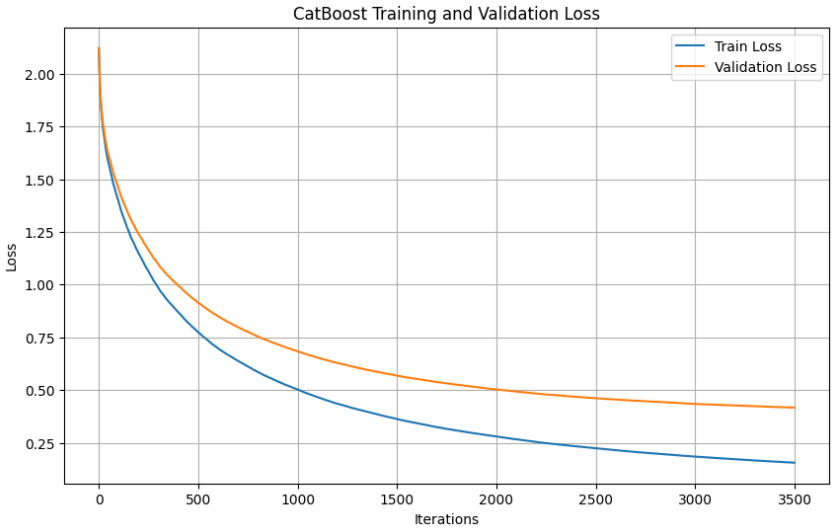

Figure 9: CatBoost Training and Validation Loss

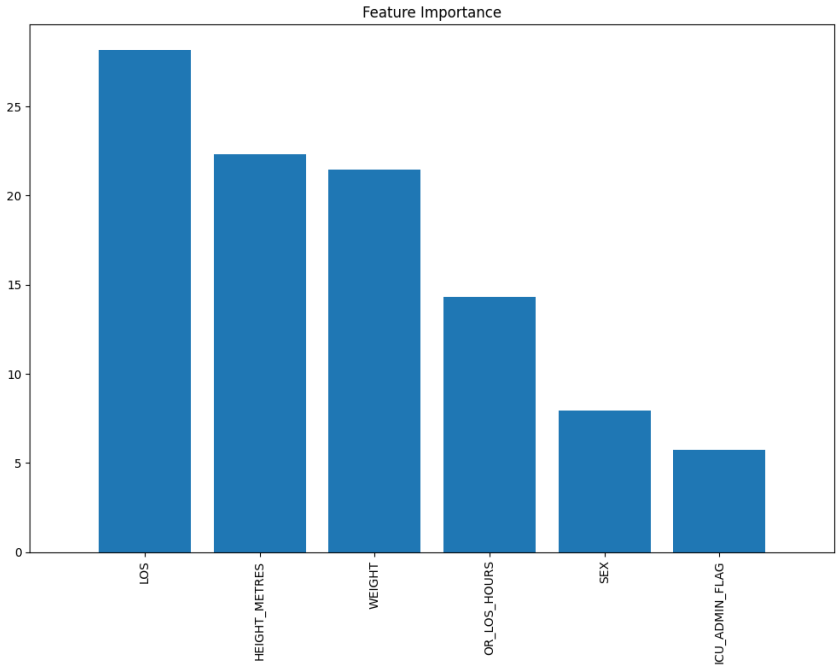

Figure 10: Feature Importance in CatBoost Model

Differences between XGBoost and CatBoost

XGBoost and CatBoost are both powerful algorithms but are fundamentally different in their approach, architecture, and feature handling capabilities. Here's how they stand apart:

- **Handling Categorical Data**: CatBoost has native support for categorical features, allowing it to process them directly without requiring manual transformations like one-hot encoding. This is especially advantageous for datasets with significant categorical data, as it reduces preprocessing complexity and preserves feature relationships. On the other hand, XGBoost lacks native categorical handling and requires explicit transformation, typically through one-hot encoding, which can lead to a larger feature space and potentially slower training times.
- **Algorithmic Approach - Ordered Boosting vs Standard Boosting**: CatBoost employs a unique *ordered boosting* technique that mitigates overfitting by ensuring that each data point's prediction is based on prior observations. This sequential approach helps CatBoost achieve better generalization on unseen data, especially in cases with limited samples. XGBoost, while powerful in standard gradient boosting, lacks this ordered boosting mechanism, making it more susceptible to overfitting in some cases unless additional regularization is applied.
- **Efficiency and Training Speed**: Due to CatBoost's ordered boosting and its efficient handling of categorical variables, it tends to perform faster on datasets with a high proportion of categorical features compared to XGBoost, especially for large datasets with complex categories. XGBoost, while fast, can become computationally expensive with one-hot encoded categorical variables, resulting in longer training times for high-dimensional data.
- **Handling Imbalanced Data**: CatBoost includes mechanisms to handle imbalanced data effectively, such as class weights and ordered boosting, which can help it perform better on minority classes without overfitting. While XGBoost also supports class weighting and other regularization techniques, it may require additional tuning and external techniques like SMOTE for handling significant class imbalances effectively.
- **Interpretability and Feature Importance**: CatBoost provides intuitive feature importance scores, which are easy to interpret due to its handling of categorical data. This can make it easier to gain insights from models trained on mixed data types. XGBoost, while offering feature importance, may present challenges when working with high-dimensional, one-hot encoded categorical features, potentially complicating interpretability.

Viability, Importance, and Profitability of Predictions in Healthcare Settings

Pertaining to the complexity scale defined in Phase 1 of the DIC project, our work addresses several impactful areas, especially in terms of patient outcomes, healthcare economics, and operational efficiency. Below are key aspects demonstrating how these predictions can benefit hospitals, patients, and the healthcare ecosystem as a whole:

1. **Cost-Effective Resource Allocation**:
    - By accurately predicting post-operative trends, including patient length of stay and likely post-operative complications, hospitals can optimize the allocation of critical resources, such as ICU beds, specialized staff, and equipment.
    - This optimized allocation not only minimizes unnecessary resource utilization but also translates to reduced costs for patients. Hospitals can increase profitability by avoiding underutilization or overuse of resources, ultimately cutting operational expenses and increasing patient throughput.
2. **Relevance to Patients, Hospitals, and Insurance Companies**:
    - Anticipating hospitalization duration, medication types, and anesthesia needs is essential for several stakeholders:
        - **Patients** benefit from cost predictability and enhanced care quality.
        - **Hospitals** can better plan for patient flow, ensuring adequate care levels for each patient.
        - **Insurance Companies** can use accurate predictions to design coverage plans and streamline reimbursement processes.
    - These factors make the problem both highly relevant and popular, addressing a real need within the healthcare industry for data-driven patient and resource management.
3. **Rich Background with a Comprehensive Dataset**:
    - This project utilizes the MOVER (Medical Informatics Operating Room Vitals and Events Repository) dataset from UC Irvine, a comprehensive dataset containing hospitalization data for 58,799 patients and 83,468 surgeries. Leveraging such a rich dataset enhances the reliability of our model and supports predictions with a robust foundation of real-world data.
    - Reference to the MOVER Dataset: *[Add citation or link to paper here if possible]*
4. **Advancing AI Integration in Healthcare**:
    - The integration of AI-driven predictions in healthcare settings facilitates data-driven personalized care, improving patient outcomes by anticipating needs and delivering tailored care.
    - Predicting the length of stay and complications allows for better planning, especially in cases requiring specialized post-operative care. This fosters patient-centric healthcare and supports the transition towards evidence-based practices.
    - Optimized resource management, as discussed in Q1, enhances hospital efficiency, reduces patient wait times, and alleviates strain on critical care resources, making AI-driven healthcare systems more effective and sustainable.

Significance of AI-Driven Predictions in Modern Healthcare

In a healthcare landscape increasingly oriented towards efficiency, AI models like ours are instrumental in:

- **Predictive Resource Allocation**: AI allows for the anticipation of resource demands, which is crucial in high-stakes areas like ICU management.
- **Patient-Centric Care**: Predictions on hospitalization duration and post-operative care support personalized healthcare, reducing the chances of complications and improving recovery outcomes.

- **Economic Benefits for Healthcare Providers**: AI optimizes hospital operations, reducing idle resources and allowing hospitals to accommodate more patients without increasing costs, thus enhancing profitability.
- **Scalability and Application**: While our model currently applies to a specific dataset, the approach is scalable across various healthcare settings, with potential applications in other areas such as chronic disease management, emergency care, and hospital admissions forecasting.

In conclusion, the predictive insights generated by our model represent a meaningful advancement in healthcare, improving the alignment of resources to patient needs, enhancing care quality, and providing significant economic benefits to healthcare providers and insurers. This work aligns with current trends in healthcare that prioritize data-driven, patient-centric, and efficient care models.