Abstract

This project focuses on the comprehensive data validation and analysis of the MOVER dataset, a repository capturing hospital visits for surgery patients at the University of California, Irvine Medical Center. The initial focus is on conducting data validation procedures to ensure the dataset's reliability. Followed by conducting in-depth exploratory data analysis to extract meaningful insights and trends, constructing machine learning models to guide hospitals throughout the patient's treatment journey, from pre-surgery to post-surgery procedures. Additionally, the project aims to develop models facilitating efficient hospital resource management, encompassing beds, utilities, and other critical elements. Addressing these objectives, the project seeks to make a significant contribution to healthcare optimization by providing insights derived from data analysis and ML models.

Introduction

In the contemporary world of healthcare, the efficient management of medical resources has become increasingly important. With challenges like the COVID-19 pandemic, the significance of operational efficiency has become even more pronounced. It's essential to understand how healthcare resources are used, especially during situations like pandemics, to be well-prepared and respond effectively. Active management of patient length of stay contributes to the overall efficiency of healthcare systems. Reduced patient wait times, optimized bed utilization, and discharge processes can collectively reduce strain on hospital resources, leading to improved healthcare delivery. This holds the potential to significantly benefit both healthcare providers and the individuals they serve.

The primary motivation is driven by the idea of leveraging Machine Learning, Statistics and Data Science to assist hospital administrators in making smarter decisions. This initiative is particularly relevant given the wealth of data encapsulated in the MOVER repository. The MOVER repository contains detailed information about patients who had surgery at the University of California, Irvine (UCI) Medical Center in California over a four-year period. It includes records from electronic medical files and waveforms for each patient visit. The important details from these records are organized into different tables for easy access and analysis. This repository gives a comprehensive look at the experiences of patients undergoing surgery at UCI Medical Center. The analysis focuses on key tables, such as Patient History, Patient Labs, Patient Postoperative Complications, Patient Procedure Events, and Patient Information, to derive comprehensive and impactful conclusions.

Method

In our detailed exploration of the MOVER dataset, our focus was on data related to surgical patients at the University of California, Irvine Medical Center. The dataset, spanning a period of four years, provided a rich tapestry of electronic medical records (EMRs) and patient waveforms, which were integral to our exploratory data analysis (EDA) and the development of machine learning models. We embarked on a rigorous selection process of the data tables, ultimately choosing those that offered the most comprehensive insight into the patients' medical journeys. These included tables detailing Patient History, which encompassed all available patient diagnoses from the electronic health records; Patient Labs, which provided a detailed account of laboratory tests and results; Patient Postoperative Complications, offering insights into complications following surgery; Patient Information, encompassing demographic data and specifics of the surgical procedures; and

Patient Procedure Events, which chronicled a timeline of events before, during, and after surgery.

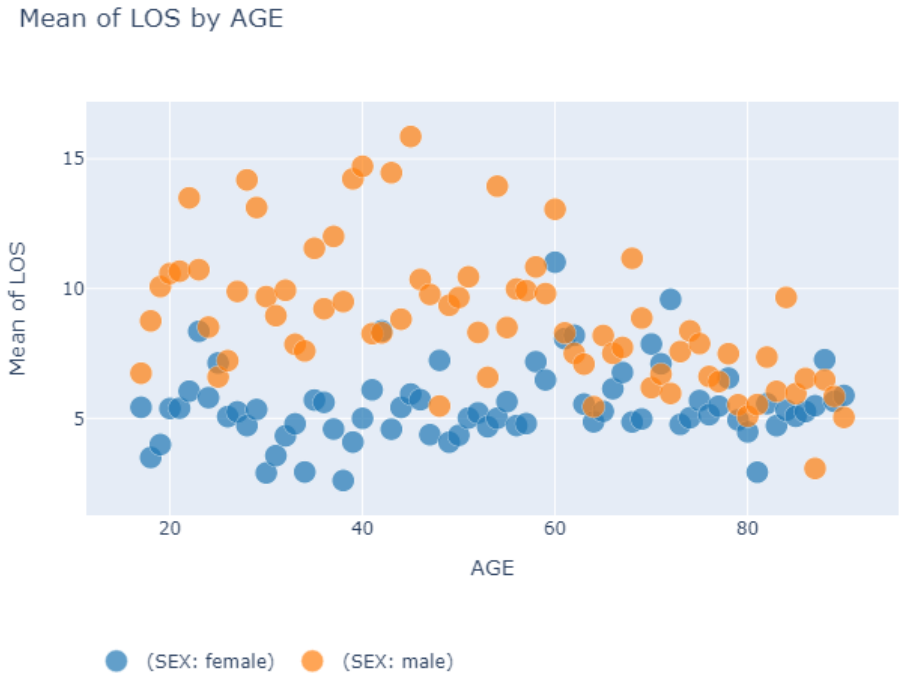
Our processing of this extensive data began with a methodical reduction of its original 20 GB size. This was achieved through the strategic label encoding of categorical features, a move that significantly reduced the data volume without losing critical information. In parallel, we engaged in a thorough cleaning of the data. This included the conversion of physical measurements to a standardized unit (for instance, patient height was converted from feet and inches to centimeters), and the exclusion of less quantifiable, text-based columns like Doctor's Notes and Postoperative Context. This not only streamlined the data but also made it more conducive for analytical purposes.

To facilitate an effective EDA, we amalgamated various data tables at the LOG ID level, creating a 'master table' that offered a panoramic view of the data. For the machine learning aspect, our approach was more nuanced. We aggregated and transformed the data at the LOG ID level, but with a keen eye on creating new, insightful features. These features included quantifiable metrics such as the number of abnormal lab reports, alongside demographic and clinical data like patient age, weight, and ASA(American Society of Anaesthesiologists) rating.

Confronting challenges was a significant aspect of our methodology. The dataset was replete with intricate medical terminology, necessitating collaboration with medical professionals to ensure accurate interpretation and usage. Another major challenge was the potential for data leakage in our machine learning model, primarily due to the inclusion of post-surgery information in tables such as Patient Procedure, Patient Labs, and Patient Postoperative Complications. To mitigate this, we established a cutoff at the Surgery Start Time, ensuring that only pre-surgery data was utilized for feature creation. This was critical in maintaining the integrity of the predictive modeling. Lastly, our strategic approach to data size management, specifically the label encoding of categorical features, was pivotal in making the dataset manageable and conducive for our analytical objectives. This multi-faceted approach ensured that we not only efficiently handled a large dataset but also extracted maximum value from it for insightful analysis and robust model development.

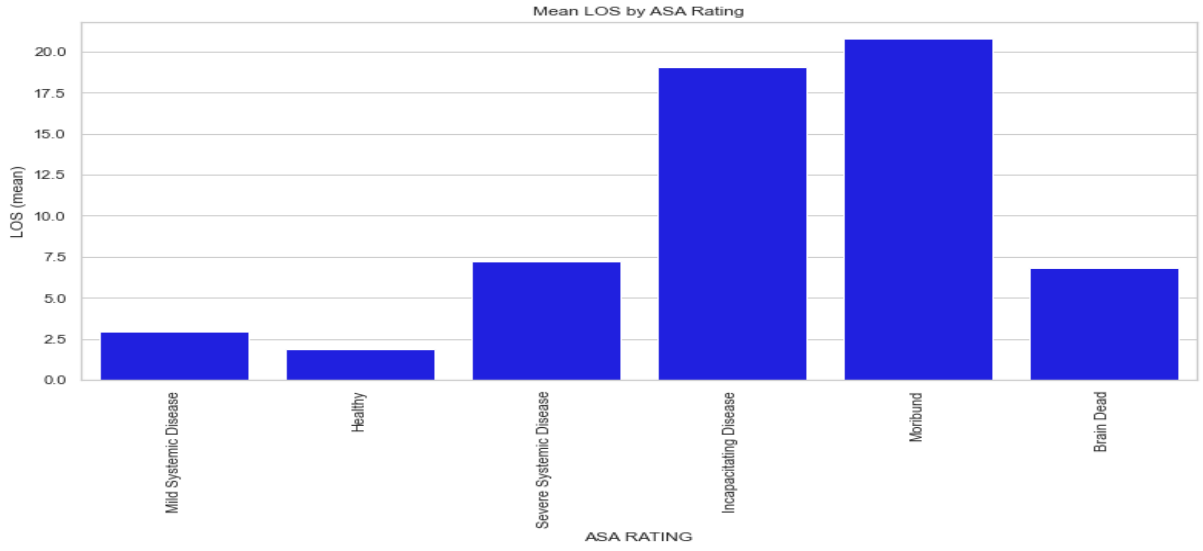
Results

After running the exploratory data analysis on the dataset, following insights were found:
Length of Stay trend with males and females across their age:



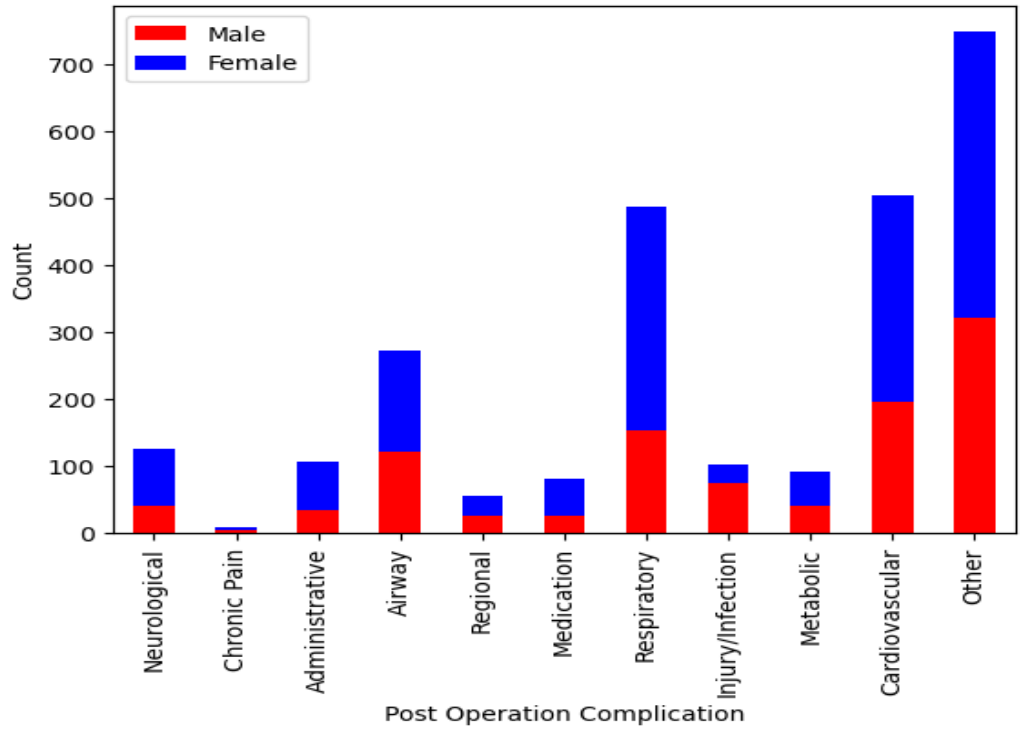
After conducting exploratory data analysis on the dataset, distinct trends in Length of Stay (LOS) between males and females across different age categories were found. The scatter plot clearly suggests that, on average, males exhibit a longer LOS compared to females, indicating potential differences in health conditions or recovery periods. This gender-based distinction implies that hospitals may need to allocate resources for male patients for a longer period.

Impact of ASA Classification on Patient Length of Stay:



The ASA rating guides anesthesiologists in surgery preparation, assessing patient health. The bar plot confirms a clear correlation: 'Healthy' ASA aligns with lower LOS, as expected. Conversely, 'Moribund' patients have notably higher LOS, emphasizing meticulous care importance. Additional LOS insights for other ASA categories can be gleaned from the plot, offering a comprehensive overview of ASA's role in patient recovery times.

Postoperative Complications shown among Males and Females:



The plot illustrates postoperative complications in both genders, with a predominant occurrence among females. Cardiovascular and respiratory issues are primary concerns. This visualization offers a concise overview of gender-specific complication patterns, emphasizing the need to address cardiovascular and respiratory health in postoperative care.

The performance results of the model performing prediction for length of stay:
The length of stay is categorized as given below:

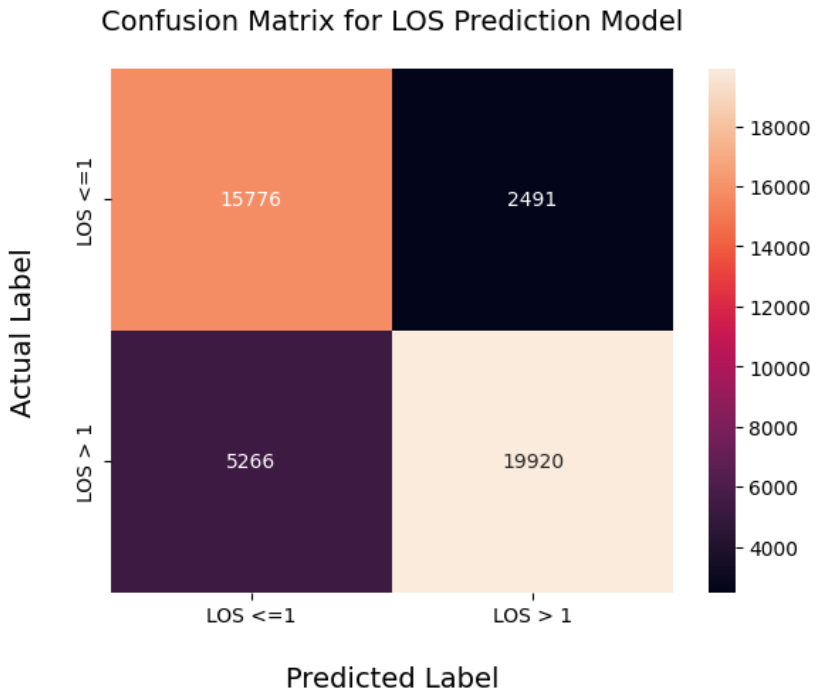
1.	Length of stay < 1 day (42%)
2.	Length of stay > 1 day (58%)

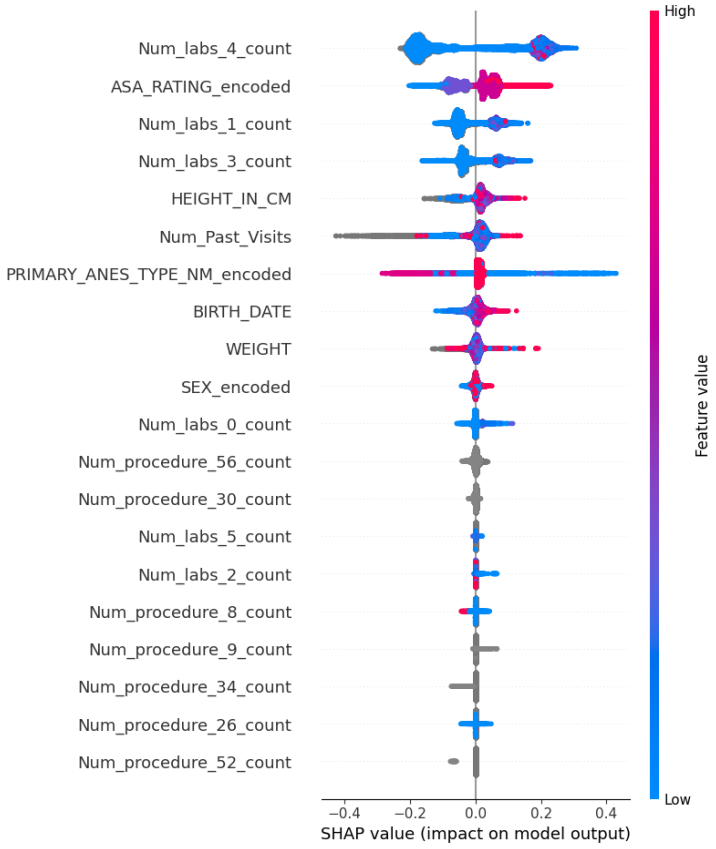
Models trained: Logistic Regression, Support Vector Classifier, Random Forest and Gradient Boosted Trees
Best Model: Bayesian hyperparameter optimization was used to select best Gradient Boosted Model

The accuracy metrics with the best performing model(Gradient Boosted Tree) is given below:

Dataset Type	Accuracy	Sensitivity	Specificity	AUC
Train (43453)	82.40%	79.09%	86.36%	0.90
Validation (5796)	79.57%	77.15%	82.92%	0.88
Test (8692)	80.24%	76.87%	84.93%	0.88

The confusion matrix, ROC curve and SHAP Values for categorical prediction for LOS in gradient boosted tree:





Conclusions

In our project focused on leveraging patient measurement data for real-time diagnostic assistance, we successfully accessed and utilized a diverse set of patient metrics to enhance diagnostic capabilities. By employing advanced machine learning models, we were

able to extract meaningful insights from real-time patient data, aiding healthcare professionals in making more informed decisions.

Moreover, the collaborative engagement with medical students and peers has proven to be invaluable in gaining a better understanding of the data. Their insights have contributed to the creation of more intuitive features, developing a user-centric approach to our MOVER system. This collaboration has not only enhanced the technical aspects of the project but has also bridged the gap between technological innovation and medical expertise.

Moving forward, one critical aspect to address is the validation of model bias and the implementation of robust testing in real-world scenarios. It is imperative to assess the model's performance across diverse patient demographics to ensure fairness and accuracy. This step is essential for gaining the trust of healthcare professionals and ensuring that the MOVER system is reliable in various clinical settings.

Looking ahead, the integration of Natural Language Processing (NLP) to further process free-text data, such as procedure descriptions, within the dataset is a promising avenue for future development. By incorporating NLP, we aim to enhance the system's ability to extract nuanced information from textual records, providing a more comprehensive understanding of patient histories and treatment plans.

Additionally, expanding the scope of our analysis by including alternative Electronic Medical Record (EMR) systems and health data sources is crucial for a more holistic approach. By incorporating a wider range of data inputs, we can improve the depth and accuracy of our MOVER system, ensuring it remains relevant and adaptable to the evolving landscape of healthcare technology. This step aligns with our commitment to achieving a comprehensive and robust analytical framework that can contribute meaningfully to efficient decision-making processes.

In the future, with additional time and resources, we envision further refinement and expansion of our MOVER system. This includes ongoing collaboration with medical professionals to incorporate their evolving insights, the continual validation of the model to adapt to emerging healthcare challenges, and exploring innovative technologies to enhance the system's capabilities. Overall, this project sets the foundation for a dynamic and adaptive healthcare technology that prioritizes precision, collaboration, and ethical considerations, ultimately contributing to the advancement of diagnostic assistance in the medical field.

References

1.Samad M, Angel M, Rinehart J, Kanomata Y, Baldi P, Cannesson M. Medical Informatics Operating Room Vitals and Events Repository (MOVER): a public-access operating room database. JAMIA Open. 2023 Oct 17;6(4):ooad084. doi: 10.1093/jamiaopen/ooad084. PMID: 37860605; PMCID: PMC10582520.

2.Markelle Kelly, Rachel Longjohn, Kolby Nottingham,The UCI Machine Learning Repository, https://archive.ics.uci.edu

3.Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

4.Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.