DA 204o: ***Data Science in Practice***

Course Project Proposal

Hospital Readmission Risk Prediction

Abhitosh, *DSBA, abhitosh2024@iisc.ac.in*
Amit Nitin Joshi, *DSBA, amitj@iisc.ac.in*
Naik Raghavendra Narottam, *DSBA, nraghavendra@iisc.ac.in*
Vignesh S, *DSBA, vigneshs@iisc.ac.in*

Abstract

Hospital readmission is a critical healthcare issue, impacting patient outcomes and increasing healthcare costs. This project utilizes data science and machine learning to predict high-risk patients, aiming to reduce readmissions through targeted interventions.

Problem Definition

Hospital readmissions significantly affect patients' quality of life and healthcare costs. Identifying high-risk patients using machine learning can enable timely interventions. This project focuses on leveraging clinical and demographic data to predict readmissions.

Data Collection and Preparation

The dataset used is sourced from Kaggle and contains 100,000 rows with 50 features in CSV format. Key features include patient demographics, hospital admission details, medications, diagnoses, and discharge dispositions.

Data Preprocessing

- General Cleaning
  - Replaced missing values in critical columns with 'Unknown'.
  - Converted relevant columns to categorical types for efficiency.
- Handling Missing Values
  - Missing values were imputed based on the context of the column.
  - Irrelevant columns with excessive missing data were dropped.
- Standardizing Diagnostic Codes
  - Diagnostic codes were standardized to reduce complexity and improve analysis consistency.
- Ordinal encoding
  - Applied ordinal encoding for some columns like max_glu_serum, A1Cresult, Medications

Exploratory Data Analysis (EDA)

- Univariate Analysis
  - Analyzed distributions of individual features to identify patterns and outliers.
  - Key Observations: Skewness in numerical variables; class imbalance in categorical variables.

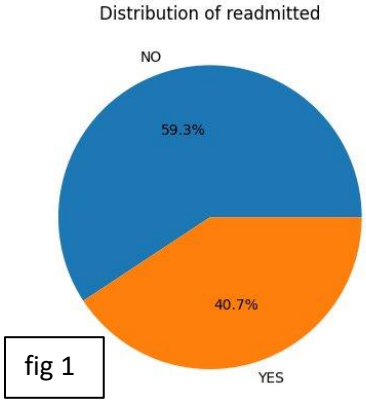


fig 1

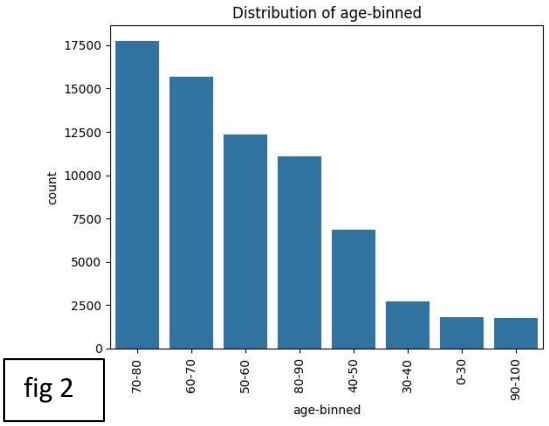


fig 2

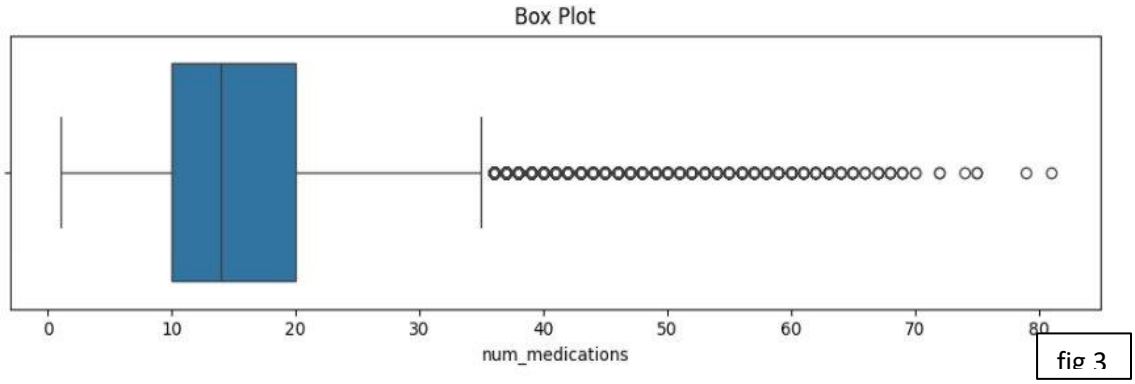


fig 3

Fig 1: Pie chart showing decent class balance in our target feature
Fig 2: Histogram representing more people admitted in higher age categories
Fig 3: Box plot showing data skewed between 10-20, few outliers

- Bivariate Analysis
  - Explored relationships between pairs of features, such as medications vs. readmissions.
  - Key Observations: Higher readmissions in older patients and those prescribed more medications.

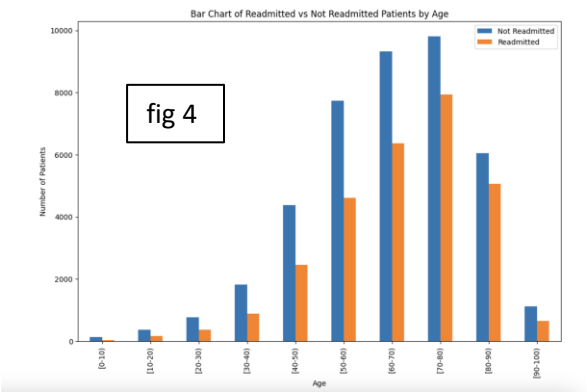


fig 4

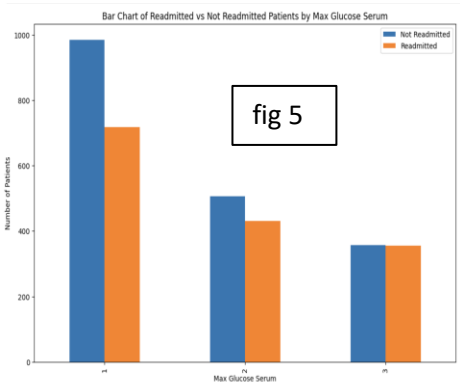


fig 5

Fig 4: With higher age the readmission rate increases gradually
Fig 5: Readmission rate is very high for the feature max_glu_serum >300

- Multivariate Analysis
  - Examined interactions among multiple variables.
  - Key Observations: no strong correlation between the columns,

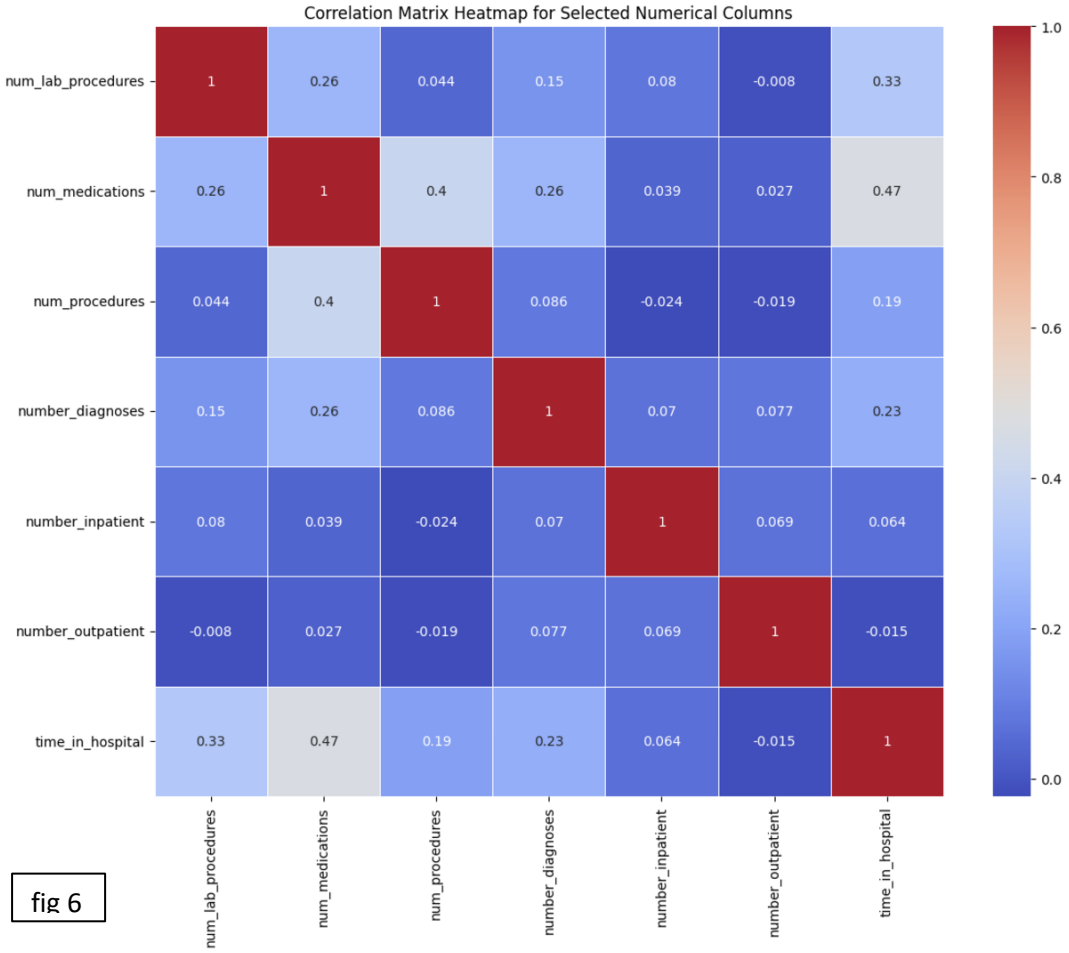


fig 6

Fig 6: weak correlation between medications, diagnoses, and readmissions.

Model Development

- Logistic Regression
  - Simple and interpretable model providing a baseline for comparison.
  - Best hyperparameters with grid search
    - C = 0.004832930238571752,
    - max_iter = 100,
    - Penalty = l2,
    - Solver = liblinear,
    - Tol = 0.001
  - Performance: Moderate accuracy (62%), limited by non-linear feature relationships.

- Decision Tree
  - Provides interpretability but prone to overfitting.
  - Performance: Accuracy of 59%, with improved precision for predicting readmissions.

- XGBoost
  - Robust ensemble method capturing complex feature interactions.
  - Performance: Best accuracy (64%) and ROC-AUC score (0.82).

Model Evaluation

Model performance was assessed using precision, recall, F1-score, and confusion matrices. XGBoost demonstrated superior accuracy and robustness compared to Logistic Regression and Decision Tree.

- Logistic regression

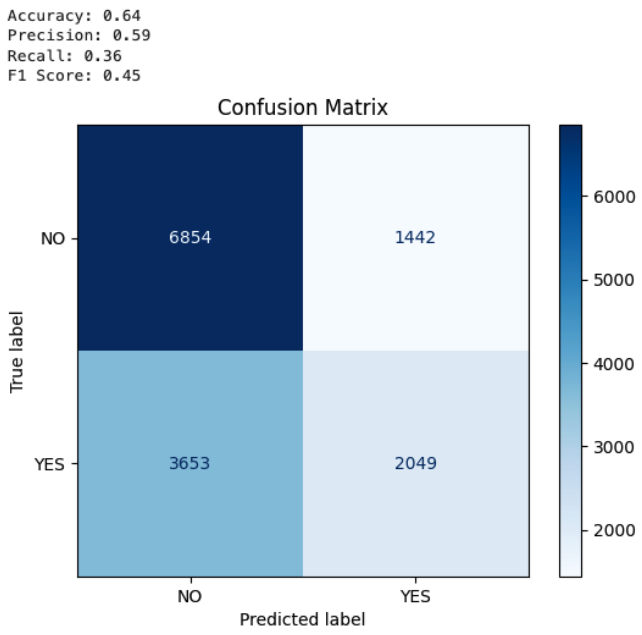
	precision	recall	f1-score	support
NO	0.62	0.91	0.74	8296
YES	0.60	0.19	0.29	5702
accuracy			0.62	13998
macro avg	0.61	0.55	0.51	13998
weighted avg	0.61	0.62	0.56	13998

- Decision Tree

Classification Report:

	precision	recall	f1-score	support
0.0	0.56	0.86	0.68	12468
1.0	0.70	0.31	0.43	12418
accuracy			0.59	24886
macro avg	0.63	0.59	0.55	24886
weighted avg	0.63	0.59	0.55	24886

- XG Boost



Conclusion and Future Directions

The project underscores the importance of preprocessing, feature engineering, and ensemble models in tackling imbalanced datasets. Future work could incorporate additional features and explore deep learning techniques to enhance model performance.

References

Dataset Citation:

- Kaggle. (2023). Hospital Readmission Risk Dataset.
  Retrieved from URL: https://www.kaggle.com/datasets/brandao/diabetes

Tools and Technologies Used:

- **Programming Language:** Python 3
- **Development Environments:** MS Visual Studio, Jupyter Notebook
- **Data Handling and Preprocessing:**
  - *Libraries*: pandas, numpy
- **Data Visualization:**
  - *Tools*: matplotlib, plotly
- **Machine Learning Libraries:**
  - SciKit-Learn, TensorFlow

Team Contributions:

- **Abhitosh**: Data collection, cleaning, and preprocessing.
- **Amit Nitin Joshi**: Model training and evaluation.
- **Naik Raghavendra Narottam**: Exploratory data analysis and feature engineering.
- **Vignesh S**: Exploratory data analysis and feature engineering.

Acknowledgments:

- The dataset, tools, and libraries used in this project significantly contributed to its success. The team acknowledges the creators of these resources for their invaluable work, which forms the foundation for data science and machine learning projects.