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*

Problem Definition

Hospital Readmission Risk Prediction

- Background of the problem

Patients, discharged from the hospital, often face numerous challenges that can lead to unplanned readmissions, significantly impacting their health, well-being, and overall quality of life. This readmission issue is compounded by the growing volume of patients with chronic conditions, making it essential for healthcare organizations to implement effective strategies for identifying high-risk patients.

- Why is it important?

It is a key challenge in the healthcare space, Addressing this problem ultimately leads to a healthier population and a more sustainable healthcare system.

- Objectives of the project

Objective of the project is finding patient who has high risk for readmission

- How can Data Science solve the problem?

Machine learning models can analyze large datasets, including patient demographics, medical history, and clinical variables, to identify patterns and predict which patients are at high risk for readmission.

Data Collection and Preparation

These processes involve gathering, organizing, and refining of the data to ensure it is suitable for analysis or model training.

- Data source(s) (where it's from, how it was collected)
  - Kaggle – source link
- Description of the data (features, size, format)
  - 100k rows of patient's hospitalization data
  - 50 features
  - CSV format
- Any preprocessing steps required
  - Handling null and missing values
  - Encoding categorical columns
  - Expanding id columns? Diagnosis codes, foreign keys

Data Preprocessing

The process of transforming raw data into a clean, structured, and usable format for analysis or machine learning.

General cleaning

- ? Present in 7 columns replaced with NAN
- Updated type to categorical for relevant columns
- Updated categorical columns with NAN to unknown
- If the patient has multiple encounter, kept the **most recent** and dropped remaining (29% rows deleted)
- Dropped rows which have discharge disposition id mapping to either **hospice or expired**

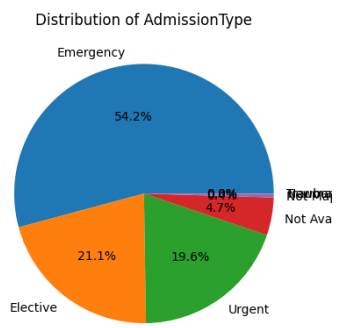
Column level cleaning

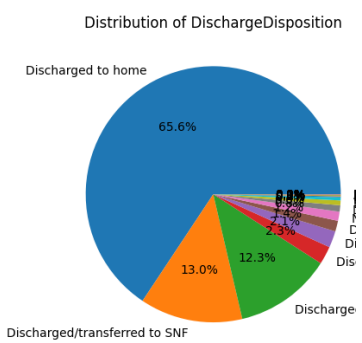
- Diagnosis columns has too many unique values, we categorize the values to specific groups for simpler analysis
- **Readmitted** – Dependent variable, >30 and <30 combied to 'Yes'
- **Weight, payer_code** column is dropped – Height % of values missing
- max_glu_serum, A1Cresult, Medications – ordinal encoding
- 15 medications with >99% not administered dropped

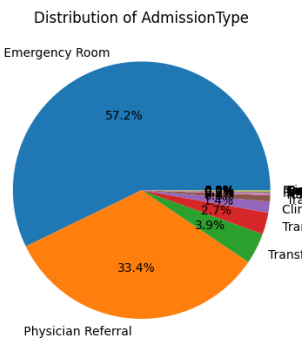
EDA

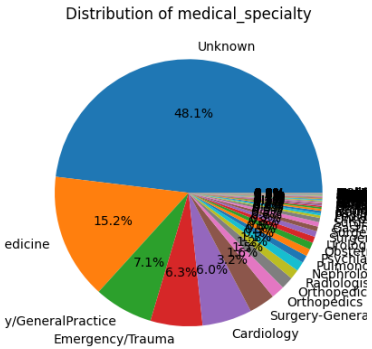
Univariate Analysis

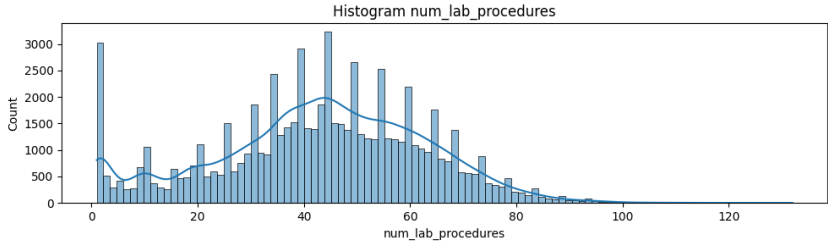
Univariate analysis involves analyzing the distribution and characteristics of individual variables in the dataset to understand their impact and patterns. Below are the detailed actions and observations during the univariate analysis phase.

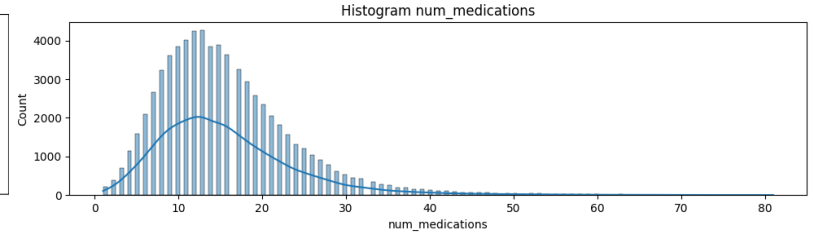


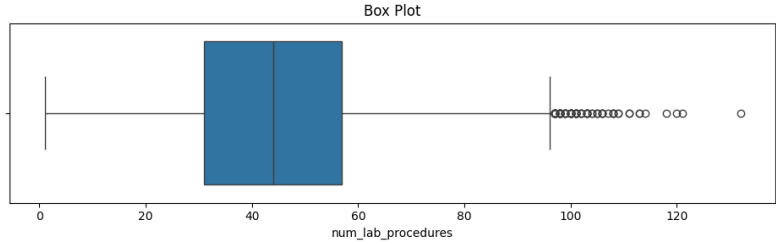


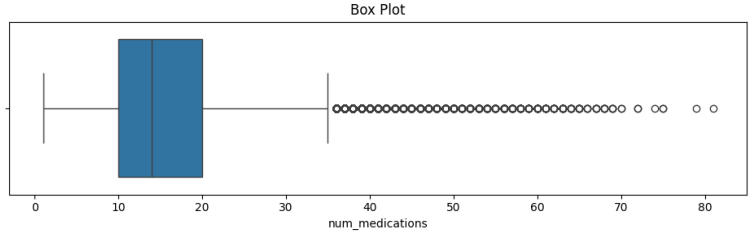












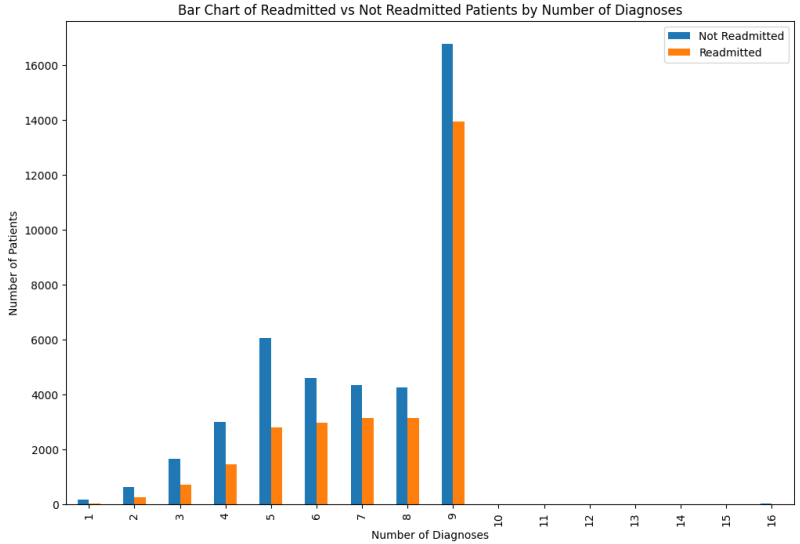
Observations:

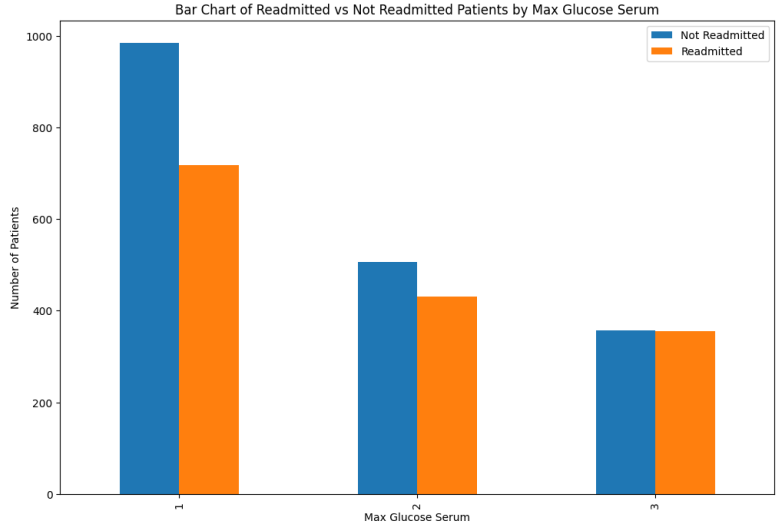
- Most of the categorical columns were highly imbalanced.
- Few of the category columns have very high cardinality - We grouped minority categories in a few cases.
- Outliers in numerical columns were dropped
- Few of the numerical columns were right skewed, we could consider distribution transformation techniques

EDA

Bivariate Analysis

Bivariate analysis examines relationships and patterns between two variables to uncover associations or trends. Below are the key insights derived from the bivariate analysis conducted on the dataset.





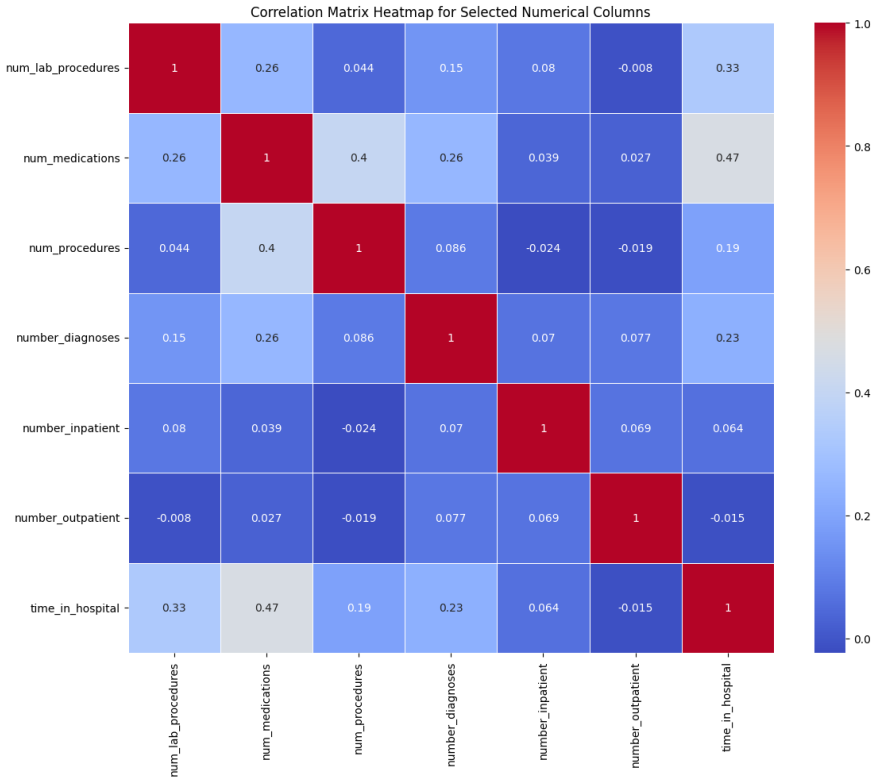
Observations:

- No differentiating features contributing significantly to the outcome
- Many medicine columns did not contribute to the outcome, these features were dropped
- Identified few features that had positive influence on readmission: Specific dispositions, number of procedures, number of medications, number of diagnosis, Age.

EDA

Multivariate Analysis

Multivariate analysis explores the relationships and interactions among multiple variables to uncover patterns and insights that may not be evident in univariate or bivariate analysis.



Observations

- Highly Correlated Features:
  - No strong correlations between features
  - Weak correlation between
    - Number of Medications and Time in Hospital: 0.47
    - Number of Procedures and Number of Medications: 0.4
    - Admission Type and Discharge Disposition: 0.4

Model Development

Decision Tree

The Decision Tree model was developed and evaluated to predict patient readmission, focusing on interpretability and performance.

Model Overview

- Approach:
  - A plain binary decision tree was trained using the Gini impurity criterion to split nodes.
  - Both categorical and numerical features were used, with preprocessing steps ensuring optimal feature scaling and encoding.

Performance Metrics

- Overall Accuracy:
  - 59% of cases were classified correctly, indicating moderate predictive power.
- Class-wise Performance Precision:
  - Class 0 (Not Readmitted): 56% of predicted non-readmissions were correct.
  - Class 1 (Readmitted): 70% of predicted readmissions were correct, showing significant improvement in identifying readmitted cases.
- Recall:
  - Class 0 (Not Readmitted): 86%, demonstrating the model's ability to effectively identify non-readmitted patients.
  - Class 1 (Readmitted): 31%, indicating some readmitted patients were missed.
- F1-Score:
  - Class 0 (Not Readmitted): 0.68, reflecting balanced performance.
  - Class 1 (Readmitted): 0.43, highlighting areas for improvement in recall.

Model Development

Decision Tree

Continue…

Feature Importance

- Top Contributing Features:
  - Number of Medications: Strongly correlated with readmission likelihood.
  - Time in Hospital: Longer stays increased the probability of readmission.
  - Number of Procedures: Higher counts were associated with higher readmission rates.
- These features align with domain knowledge, emphasizing their relevance in readmission prediction.

Insights and Challenges

- Strengths:
  - **High interpretability due to the tree structure**.
  - Significant improvement in precision for readmission cases (Class 1).
- Challenges:
  - Low recall for readmitted patients indicates missed cases.
  - Potential overfitting due to the tree's sensitivity to training data, necessitating hyperparameter tuning.

Key Metrics Summary

- **Accuracy:** 59%
- **Precision:** 56% (Class 0), 70% (Class 1)
- **Recall:** 86% (Class 0), 31% (Class 1)
- **F1-Score:** 0.68 (Class 0), 0.43 (Class 1)

Model Development

Logistics Regression

Logistic Regression is a widely used statistical method for binary classification problems.

Model Overview

- Approach:
  - Features included numerical and categorical predictors, preprocessed for compatibility with the model.
  - Regularization techniques (L1 and L2 penalties) were explored to balance complexity and performance.

Feature Selection

- Key Features:
  - Age (Encoded): Encodes age group patterns.
  - Number of Medications: Reflects treatment complexity.
  - Time in Hospital: A measure of illness severity.
  - Number of Diagnoses: Indicates overall health complexity.
  - Emergency Visits: Captures urgent care needs.
- Feature Impact:
  - Significant correlation observed with readmission likelihood, particularly for ***Time in Hospital*** and ***Number of Medications***.

Hyperparameter Tuning

- Process:
  - Regularization (***C***) values tested logarithmically from 10^{-4} to 10^{4} .
  - Penalties (L1 and L2) and solvers (e.g., ***liblinear*** and ***saga***) were evaluated.
  - Best configuration identified using 5-fold cross-validation.
- Outcome:
  - The optimal configuration yielded a significant improvement in accuracy and generalization.

Model Development

Logistics Regression

continue…

Model Performance

- Overall Accuracy:
  - The model achieved a maximum accuracy of 62%.
- Classification Metrics:
  - Precision: High precision for non-readmitted cases, indicating effective identification with minimal false positives.
  - Recall: Moderate recall for readmitted cases, highlighting areas for improvement in capturing true positives.
  - F1-Score: Demonstrated a balance between precision and recall.

Challenges and Future Improvements

- Challenges:
  - Moderate accuracy indicates room for feature engineering and dataset augmentation.
  - Class imbalance may affect model predictions.
- Future Directions:
  - Employ SMOTE or other techniques to address class imbalance.
  - Incorporate interaction terms and non-linear transformations for enhanced predictive power.

Key Metrics Summary

- **Accuracy:** 62%
- **Top Features:** Age, Number of Medications, Time in Hospital, Number of Diagnoses, Emergency Visits.
- **Regularization Techniques:** Grid search optimization identified the best parameters for generalization.

Model Development

XGBoost

XGBoost (Extreme Gradient Boosting) is a high-performance implementation of gradient boosting, widely used for predictive tasks due to its scalability and accuracy.

Model Overview

- Approach:
  - A tree-based boosting algorithm was used to combine weak learners iteratively.
  - Numerical and categorical predictors were preprocessed for optimal model compatibility.

Feature Selection

- Key Features:
  - Age (Encoded): Encodes age group patterns for patient demographics.
  - Number of Medications: Indicates treatment complexity.
  - Time in Hospital: Reflects the severity of conditions.
  - Number of Diagnoses: Captures overall health complexity.
  - Emergency Visits: Highlights urgent care needs.
- Feature Impact:
  - All selected features showed significant correlation with readmission likelihood, validating their inclusion.

Hyperparameter Tuning

- Process:
  - Hyperparameters like learning rate, tree depth, number of estimators, and subsampling ratio were tuned using grid search with cross-validation.
  - Iterative optimization balanced model complexity and performance.
- Best Parameters Identified:
  - Learning rate (0.1), max depth (6), number of estimators (100), and subsampling ratio (0.8).

Model Development

XGBoost

continue…

Model Performance

- Overall Accuracy:
  - The XGBoost model achieved an accuracy of 64%, outperforming baseline models.
- Classification Metrics:
  - **Precision**: High precision across all classes, minimizing false positives.
  - **Recall**: Balanced recall values, effectively capturing both readmitted and non-readmitted cases.
  - **F1-Score**: Demonstrated robust performance across metrics, indicating a well-balanced model.
- ROC-AUC Score:
  - A value of 0.67 reflects strong discriminatory ability.

Insights and Challenges

- Strengths:
  - Outperformed simpler models due to its ability to capture complex feature interactions.
  - Robust to overfitting with regularization and early stopping.
- Challenges:
  - High computational cost during training.
  - Sensitivity to hyperparameter settings requires careful tuning.
- Future Directions:
  - Explore feature engineering and automated tuning methods for further improvements.
  - Compare performance against neural networks or other advanced algorithms.

Key Metrics Summary

- **Accuracy:** 64%
- **ROC-AUC Score:** 0.67
- **Top Features:** Age, Number of Medications, Time in Hospital, Number of Diagnoses, Emergency Visits.

Summary

The project involved a systematic and technical approach to predicting patient readmissions. Key steps included:

Feature Engineering: *Comprehensive preprocessing, including feature scaling and selection, ensured that input variables were optimized for model training.*

Model Development: *Trained three different models—Logistic Regression, Decision Tree, and XGBoost—to evaluate performance and identify the best predictor.*

Class Imbalance Handling: *Addressed the imbalance in the "readmitted" class using Synthetic Minority Oversampling Technique (SMOTE), ensuring better representation of minority cases in the training process.*

Model Validation: *Utilized fold cross-validation to mitigate overfitting and evaluate model robustness. Performance metrics were assessed using confusion matrices, providing detailed insights into precision, recall, and accuracy.*

Metric	Decision Tree	Logistic Regression	XGBoost
Accuracy (%)	59	62	64
Precision (Macro Avg)	70 (Class 1)	61	High
Recall (Macro Avg)	86 (Class 0)	55	Balanced
F1-Score (Weighted Avg)	43 (Class 1)	56	High
ROC-AUC Score	-	-	0.67

Key Takeaways

- Best Performing Model:
  - XGBoost outperformed Logistic Regression and Decision Tree models with the highest accuracy (64%) and ROC-AUC score (0.67). Demonstrated superior performance by capturing complex feature interactions and achieving robust generalization.
- Balanced Trade-Off:
  - Decision Tree: Provides high interpretability with clear feature importance insights but struggled with low recall for minority class predictions.
  - Logistic Regression: Offers simplicity and interpretability with moderate accuracy (62%). Suitable for real-time predictions or scenarios where model explainability is a priority.
  - XGBoost: Outperforms in accuracy and robustness but requires significant computational resources, making it ideal for high-stakes scenarios.

Conclusion

- Final Result:
  - The overall best accuracy achieved was 64% using the XGBoost model. Despite rigorous efforts in feature engineering and class balancing, the accuracy of the Logistic Regression and Decision Tree models remained moderate (~59-62%), reflecting challenges in the dataset.
- Reasons for Low Accuracy:
  - Class Imbalance: Even after applying SMOTE, the "readmitted" class remained challenging to predict accurately due to inherent dataset biases.
  - Feature Limitations: Some key features may not fully capture the complexity of patient readmissions, limiting the models' predictive power.
  - Non-linear Relationships: Logistic Regression and Decision Tree struggled to capture non-linear patterns that XGBoost handled more effectively.
- Insight:
  - This project highlights the importance of robust preprocessing, feature engineering, and advanced algorithms like XGBoost for handling imbalanced datasets. Future improvements, such as incorporating additional domain-specific features and exploring deep learning methods, could significantly enhance performance.

Data Science Canvas

Project: Hospital Readmission Risk Prediction

Team: Abhitosh, Amit Nitin Joshi, Naik Raghavendra Narottam, Vignesh S

Problem Statement

Business Case & Value Added

Predicting diabetes patient readmission risk to improve care, reduce costs, optimize resources, and enhance hospital performance, supporting value-based healthcare.

Data Landscape

Dataset available from Kaggle.
Contains clinical records from 130 hospitals, detailing factors affecting diabetes patient readmissions from years 1999 – 2008.
No additional data needed.

Model Selection

Predictive Modeling: Logistic regression, decision trees, neural networks for predicting readmission.

Classification Techniques: Multiclass classification, SVM for handling diverse outcomes.

Feature Engineering: PCA, feature importance to identify key predictors.

Exploratory Data Analysis & Visualization: Detect patterns and relationships in the data.

Survival Analysis: Understand time until readmission and guide patient care management.

Model Requirements

Ensure data quality, balanced classes, feature relevance, proper model selection, and thorough validation for reliable predictions.

Software & Libraries

Exploratory data analysis – pandas, numpy

Data Preprocessing – pandas

Data Visualization – matplotlib, plotly

Machine Learning – scikit-learn, Tensorflow

Skills

Data Preprocessing & Cleaning: Handling missing data, outliers, and data transformations.
Feature Engineering: Selecting and creating relevant features for model training.
Machine Learning Development: Building, training, and fine-tuning predictive models.
Statistical Analysis: Understanding data distributions and relationships.
Programming Skills: Proficiency in Python.
Data Visualization: Creating charts and plots to interpret model results.
Healthcare Domain Knowledge: Understanding medical terms and factors influencing patient outcomes.
Model Evaluation & Optimization: Validating models and improving performance.

Execution & Evaluation

Model Evaluation

Indicators Requiring Quality Control:

Accuracy: Measure of correct predictions; interpret as model effectiveness.

Precision & Recall: Assess true positive rates; balance for class imbalances.

F1 Score: Harmonizes precision and recall for overall performance.

In general. real-time monitoring is necessary for model performance adjustments. But this is out of the scope of the project

Data Storytelling

Actionable Insight for clinicians and administrators to determine if patient is at re-admission risk.

Data Collection & Preparation

Data Selection & Cleansing

Data shall require clean-up and pre-processing for effectively training the model.

Data Integration

Federated learning clients located in hospital premises.
This ensures that patient data is not leaked outside the hospital IT infrastructure.

Data Collection

Collect data via surveys, EHRs, and sensors; ensure accuracy, relevance.

Explorative Data Analysis

Yes. Required descriptive statistics for assessment.

Adopted from: https://github.com/tomalytics/datasciencecanvas