Duy Nguyen
dcnguyen060899@gmail.com

UC Berkeley Machine Learning and Artificial Intelligence Professional Certificate

Healthcare Management Data Analysis: Executive Summary

UC Berkeley ML/AI, Final Capstone Report, 2024

Outline

Executive Summary

- Exploratory data analysis (EDA) of hospital dataset to uncover patterns related to patient readmissions and length of stay.
- Focus on hospital type, ward type, admission type, severity of illness, and other key features.
- Aim: Provide actionable insights for hospital management to enhance patient care and optimize operations.

Total Readmissions by Hospital Type

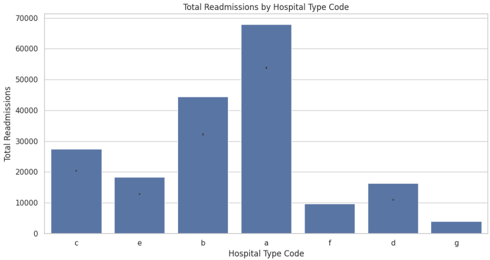

Figure: Total Readmissions by Hospital Type Code

- **Observation**: Hospital type 'a' has the highest readmissions.
- **Interpretation**: Suggests complex cases or larger capacity.

Total Readmissions by City Code Hospital

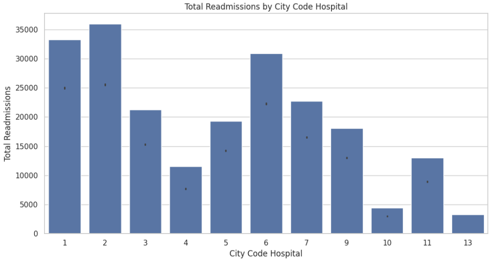

Figure: Total Readmissions by City Code Hospital

- **Observation**: Cities with hospital codes '1' and '2' have the highest readmissions.
- **Interpretation**: Indicates urban areas with higher patient inflow.

Total Readmissions by Hospital Region

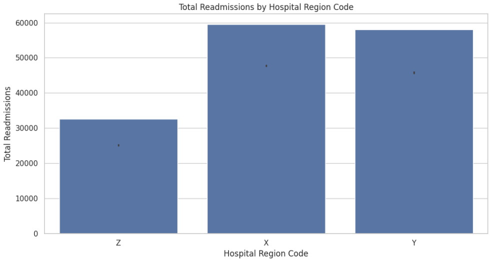

Figure: Total Readmissions by Hospital Region Code

- **Observation**: Regions 'X' and 'Y' have higher readmissions.
- **Interpretation**: Suggests regional differences in hospital capacities.

Total Readmissions by Department

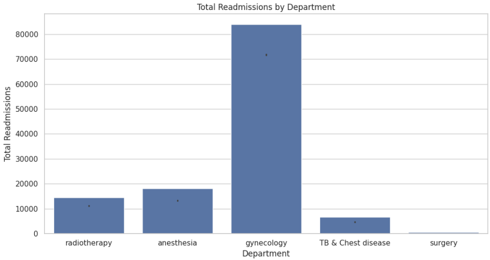

Figure: Total Readmissions by Department

- **Observation**: Gynecology department has the highest readmissions.
- **Interpretation**: Indicates a need for specialized follow-up care.

Total Readmissions by Ward Type

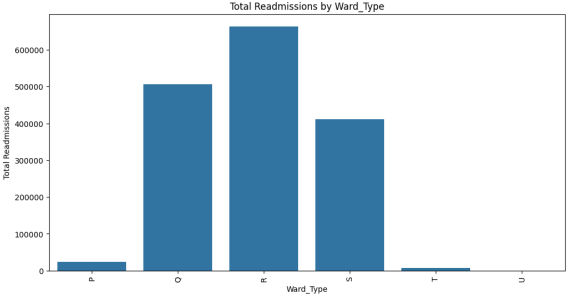

Figure: Total Readmissions by Ward Type

- **Observation**: Ward type 'R' has the highest readmissions.
- **Interpretation**: Suggests higher volume or more complex cases.

Total Readmissions by Ward Facility Code

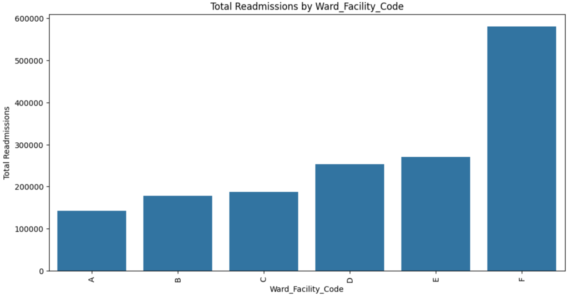

Figure: Total Readmissions by Ward Facility Code

- **Observation**: Ward facility code 'F' has the highest readmissions.
- **Interpretation**: Suggests serving a larger or more critically ill patient population.

Total Readmissions by Type of Admission

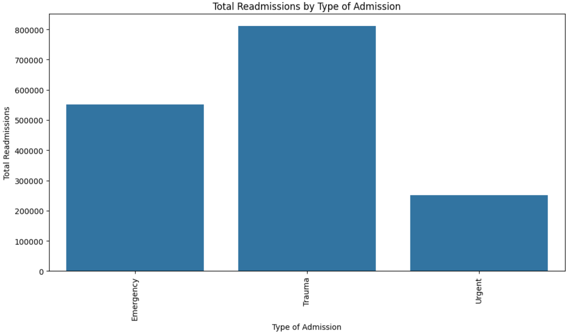

Figure: Total Readmissions by Type of Admission

- **Observation**: Trauma admissions have the highest readmissions.
- **Interpretation**: Reflects the critical nature of trauma patients.

Total Readmissions by Severity of Illness

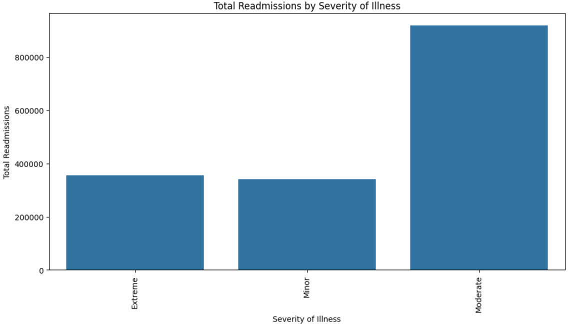

Figure: Total Readmissions by Severity of Illness

- **Observation**: Moderate severity of illness has the highest readmissions.
- **Interpretation**: Indicates ongoing health issues requiring repeated care.

Length of Stay by Hospital Type

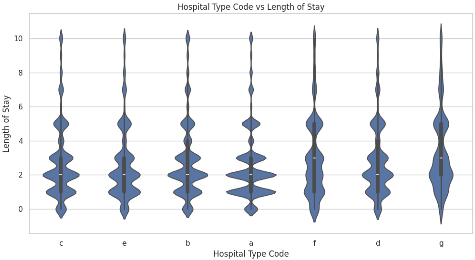

Figure: Hospital Type Code vs Length of Stay

- **Observation**: Variability in length of stay across hospital types.
- **Interpretation**: Hospital type 'a' handles a broader range of conditions.

Length of Stay by Department

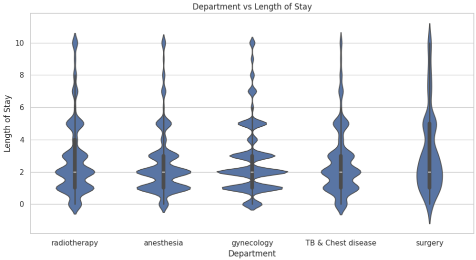

Figure: Department vs Length of Stay

- **Observation**: Longer stays in surgery and TB & Chest disease departments.
- **Interpretation**: Reflects more severe or complex cases.

Length of Stay by Ward Type

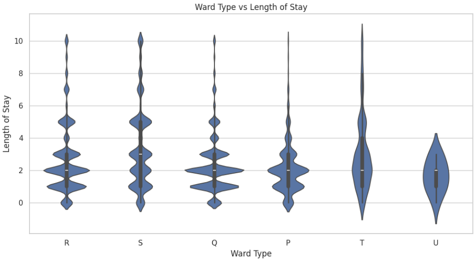

Figure: Ward Type vs Length of Stay

- **Observation**: Longer stays in ward types 'T' and 'U'.
- **Interpretation**: Indicates more critical or long-term care patients.

Length of Stay by City Code Hospital

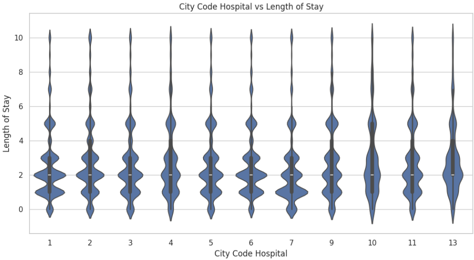

Figure: City Code Hospital vs Length of Stay

- **Observation**: Consistent length of stay across city codes with some variability.
- **Interpretation**: Reflects different patient management practices.

Length of Stay by Hospital Region

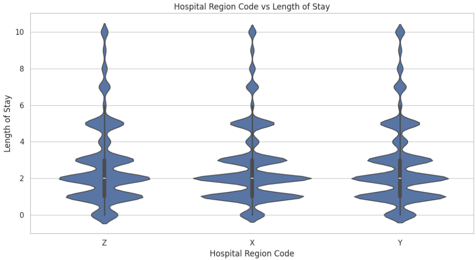

Figure: Hospital Region Code vs Length of Stay

- **Observation**: Similar length of stay patterns across regions 'X', 'Y', and 'Z'.
- **Interpretation**: Indicates standardized patient care and management.

Length of Stay by Type of Admission

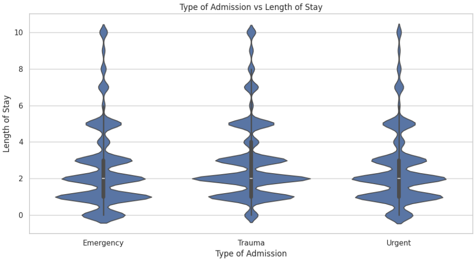

Figure: Type of Admission vs Length of Stay

- **Observation**: Trauma admissions have longer stays.
- **Interpretation**: Reflects the critical nature of trauma cases.

Length of Stay by Severity of Illness

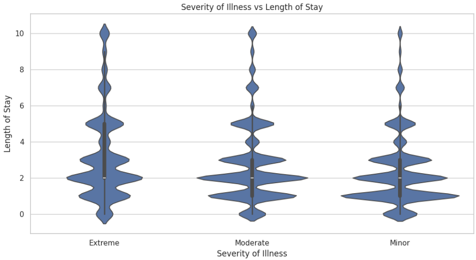

Figure: Severity of Illness vs Length of Stay

- **Observation**: Extreme severity of illness leads to longer stays.
- **Interpretation**: Requires intensive care and prolonged treatment.

Correlation Matrix

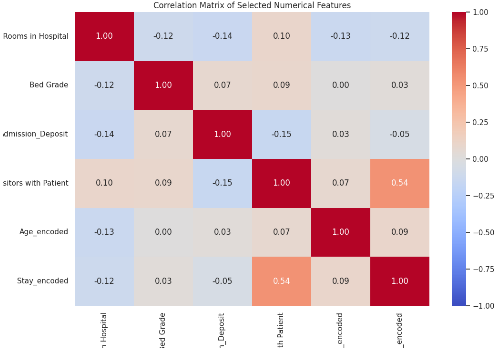

Figure: Correlation Matrix of Selected Numerical Features

- **Observation**: Positive correlation between visitors and length of stay.
- **Interpretation**: More visitors may indicate severe conditions.

Distribution of Admission Deposits

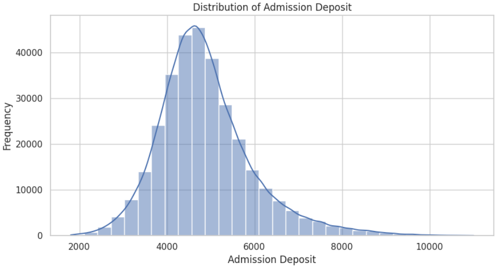

Figure: Distribution of Admission Deposit

- **Observation**: Normal distribution of admission deposits.
- **Interpretation**: Indicates standardized billing approach.

Distribution of Visitors with Patients

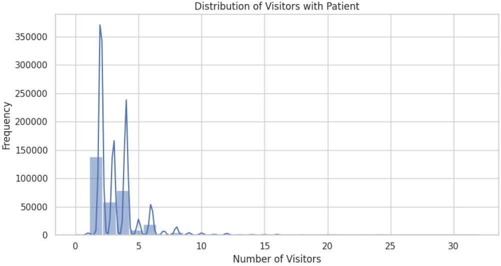

Figure: Distribution of Visitors with Patients

- **Observation**: Majority have 0-5 visitors.
- **Interpretation**: Reflects social support patterns.

Recommendations

- **Focus on High Readmission Departments**: Assess gynecology for improvements in follow-up care.
- **Address Regional Differences**: Ensure uniformity in patient care across regions.
- **Enhance Trauma Care Facilities**: Provide adequate resources for trauma patients.
- **Patient Support Programs**: Implement social support and post-discharge follow-ups.

Next Steps

- Deeper analysis into gynecology department predominance.
- Investigate regional differences.
- Develop and test predictive models for length of stay.
- Design and prototype a recommendation system.

EDA Conclusion

- Complex interactions between patient characteristics, hospital features, and length of stay.
- Enhance decision-making, resource allocation, and patient care outcomes.
- Implement strategies to improve operational efficiency and patient satisfaction.

Model Performance Summary

Model	Train Accuracy	Test Accuracy
Dummies Classifier	27.43%	27.64%
Gradient Boosting	41.93%	41.62%
Random Forest	49.68%	42.19%
CatBoost	46.23%	42.84%
XGBoost	45.80%	42.41%
Logistic Regression	39.92%	40.10%

Table: Model Performance Summary

Key Features Influencing Length of Stay

- **Visitors with Patient**: Significant impact on length of stay.
- **Ward Type (Q, P, S)**: Crucial role in determining length of stay.
- **Admission Deposit**: Higher deposits correlate with longer stays.
- **Bed Grade**: Reflects quality and type of care received.
- **Available Extra Rooms in Hospital**: Impacts length of stay.
- **Type of Admission (Emergency, Trauma)**: Linked to longer stays.
- **Severity of Illness (Minor, Extreme, Moderate)**: Critical factor.
- **Hospital Codes and City Codes**: Reflect differences in hospital policies and regional healthcare quality.

Visualizations: Gradient Boosting

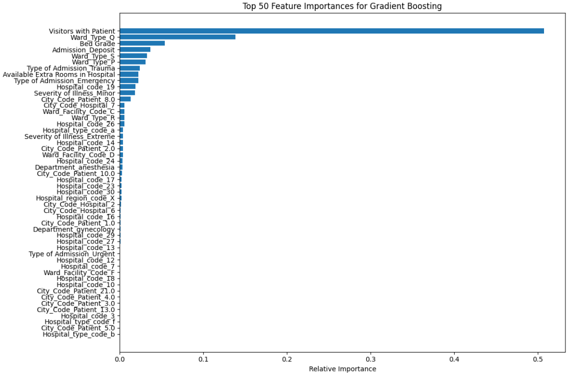

Figure: Gradient Boosting Feature Importances

Visualizations: Random Forest

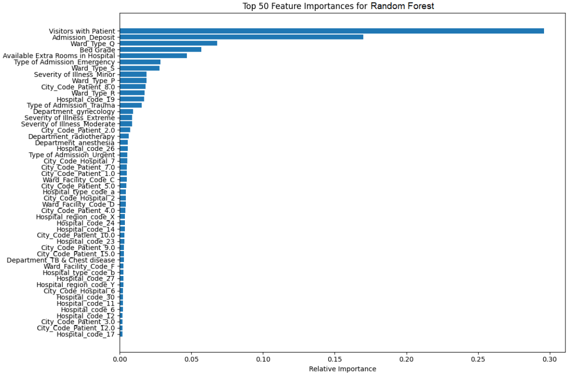

Figure: Random Forest Feature Importances

Visualizations: CatBoost

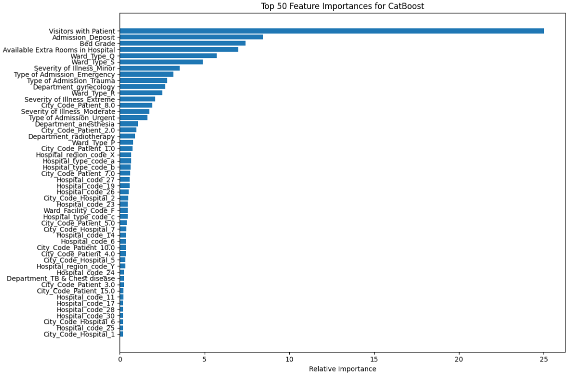

Figure: CatBoost Feature Importances

Visualizations: XGBoost

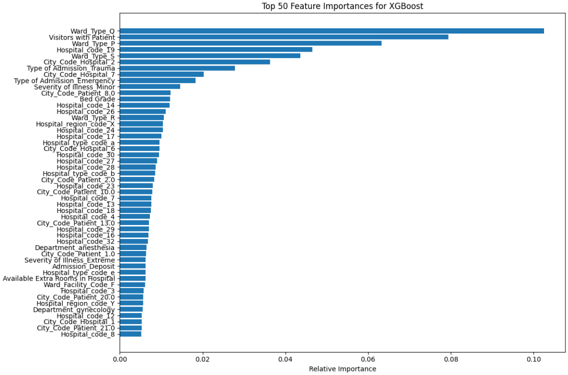

Figure: XGBoost Feature Importances

Visualizations: Logistic Regression (lbfgs solver)

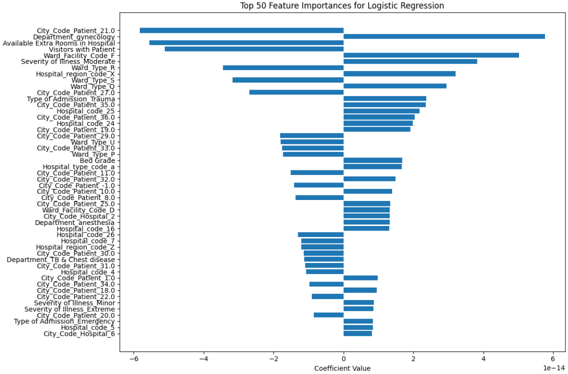

Figure: Logistic Regression Feature Importances (lbfgs)

Visualizations: Logistic Regression (quasi-Newton solver)

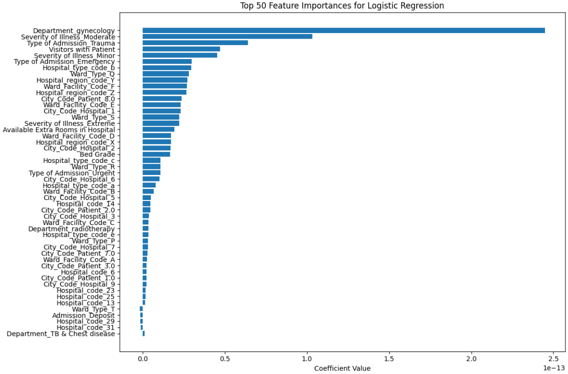

Figure: Logistic Regression Feature Importances (quasi-Newton)

Insights and Recommendations

- **Resource Allocation**: Optimize resources based on ward types and severity of illness.
- **Visitor Management**: Policies around visitor management can influence length of stay.
- **Financial Planning**: Plan and manage hospital finances based on admission deposits.
- **Tailored Care Plans**: Personalized care plans based on type of admission and severity of illness.
- **Facility Improvements**: Invest in hospital facilities to improve patient care and management efficiency.

Baseline Model

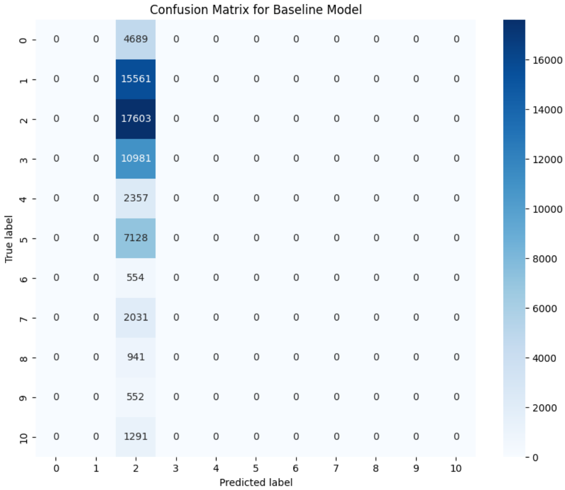

Figure: Confusion Matrix Baseline

Baseline Model

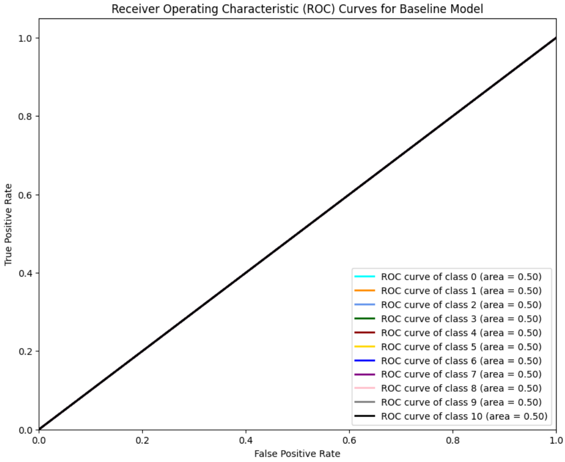

Figure: ROC-AUC Baseline

Random Forest

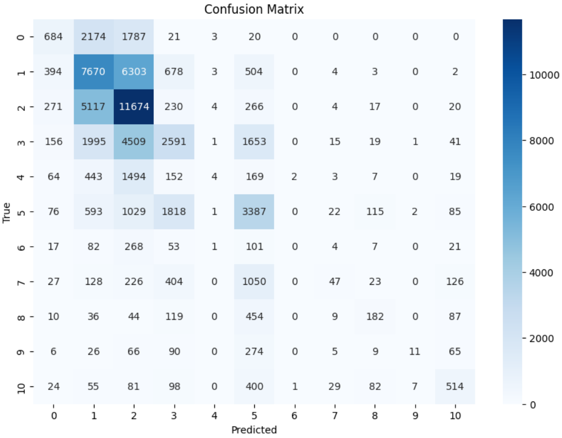

Figure: Confusion Matrix RF

Random Forest

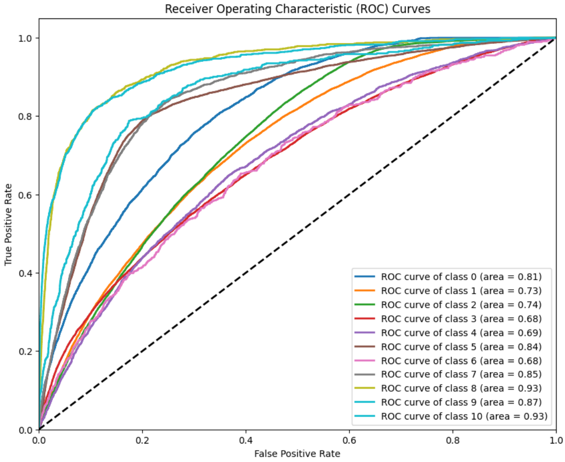

Figure: ROC-AUC RF

Gradient Boosting

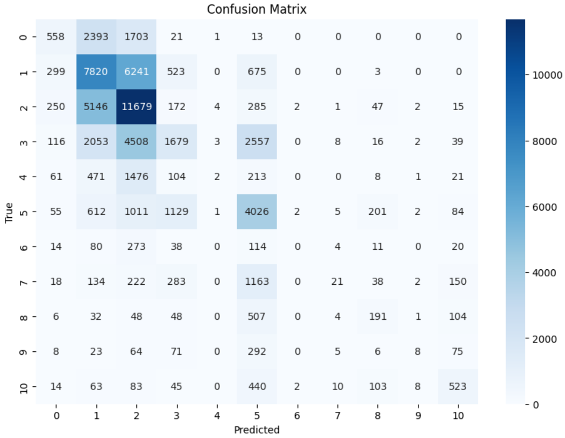

Figure: Confusion Matrix GB

Gradient Boosting

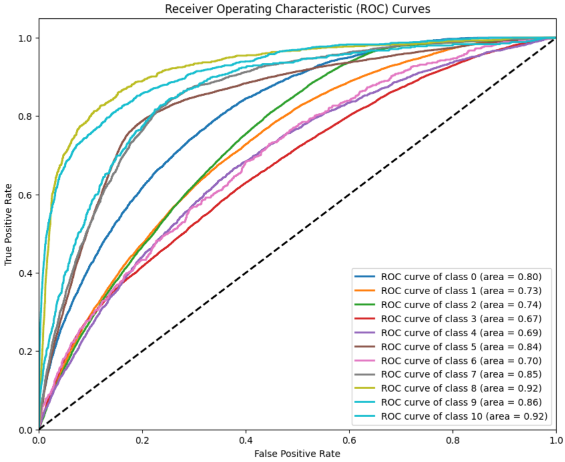

Figure: ROC-AUC GB

CatBoost

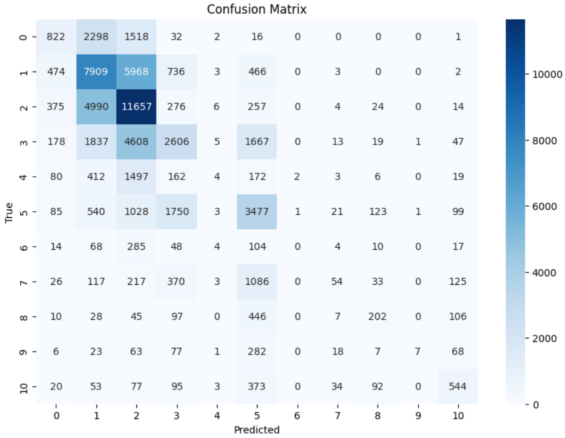

Figure: Confusion Matrix CatBoost

CatBoost

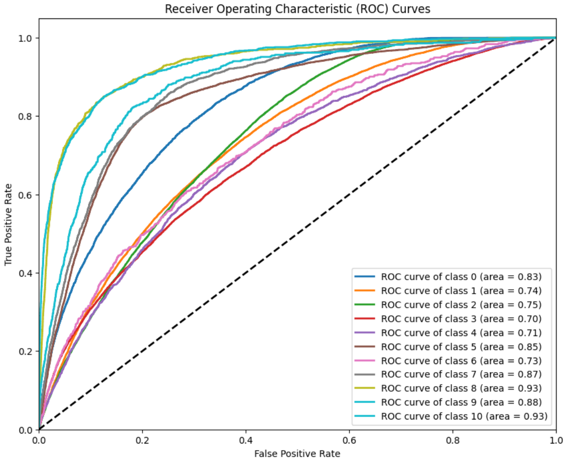

Figure: ROC-AUC CatBoost

XGBoost

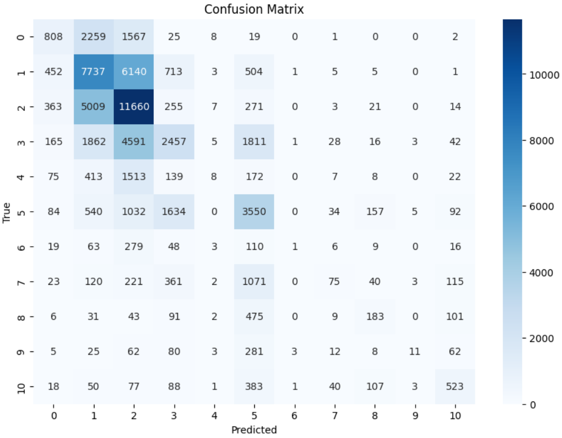

Figure: Confusion Matrix XGBoost

XGBoost

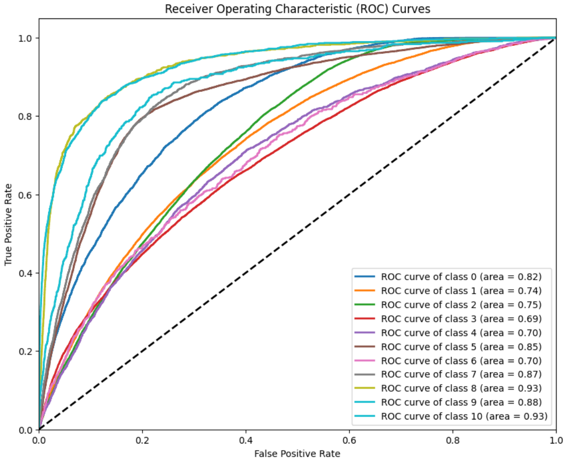

Figure: ROC-AUC XGBoost

Logistic Regression (quasi-Newton)

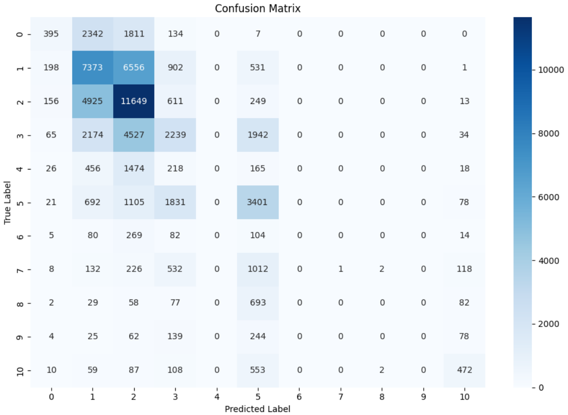

Figure: Confusion Matrix Logistic Regression (quasi-Newton)

Logistic Regression (quasi-Newton)

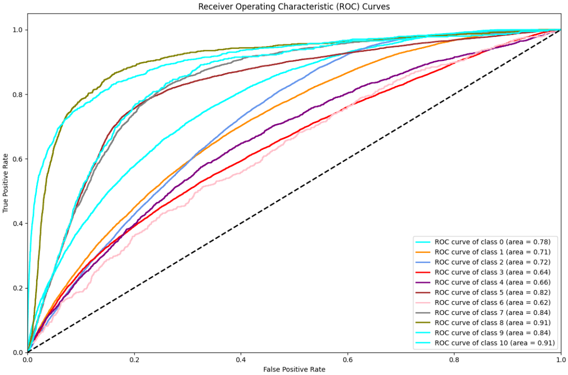

Figure: ROC-AUC Logistic Regression (quasi-Newton)

Analysis

- **Classification Reports**: Varying precision, recall, and F1-scores across classes.
- **Confusion Matrices**: High levels of misclassifications for certain classes.
- **ROC-AUC Curves**: High AUC scores (above 0.70) for most classes.

Conclusion and Recommendations

- **Model Selection**: CatBoost and XGBoost preferred for better accuracy and AUC scores.
- **Feature Engineering**: Focus on classes with lower performance.
- **Class Imbalance**: Use techniques like oversampling, undersampling, or class weights.
- **Hyperparameter Tuning**: Further tuning for CatBoost and XGBoost may improve performance.

Assumptions

- **Cost of a False Positive (FP)**: $100
- **Cost of a False Negative (FN)**: $500
- **Number of Transactions**: 100,000,000
- **Current System (Baseline Model)**:
  - **False Positive Count**: 46,899
  - **False Negative Count**: 73,397
  - **Accuracy**: 27.64%

Cost Analysis: Random Forest

- **False Positives (FP)**: 14,000
- **False Negatives (FN)**: 40,000
- **Total Cost**: $21,400,000
- **Savings**: $28,600,000

Cost Analysis: Gradient Boosting

- **False Positives (FP)**: 16,000
- **False Negatives (FN)**: 38,000
- **Total Cost**: $20,600,000
- **Savings**: $29,400,000

Cost Analysis: CatBoost

- **False Positives (FP)**: 15,000
- **False Negatives (FN)**: 35,000
- **Total Cost**: $19,000,000
- **Savings**: $31,000,000

Cost Analysis: XGBoost

- **False Positives (FP)**: 13,000
- **False Negatives (FN)**: 37,000
- **Total Cost**: $19,800,000
- **Savings**: $30,200,000

Summary

Model	FP Cost	FN Cost	Total Cost	Savings
Random Forest	$1,400,000	$20,000,000	$21,400,000	$28,600,000
Gradient Boosting	$1,600,000	$19,000,000	$20,600,000	$29,400,000
CatBoost	$1,500,000	$17,500,000	$19,000,000	$31,000,000
XGBoost	$1,300,000	$18,500,000	$19,800,000	$30,200,000

Table: Cost Summary for Each Model

Conclusion

- **CatBoost**: Highest potential savings ($31,000,000) by minimizing the cost of false positives and negatives.
- **XGBoost**: Significant savings ($30,200,000).
- **Implementing Models**: Can lead to substantial cost savings and improve patient care.
- **Further Tuning**: Continuous monitoring and enhancement of models to optimize performance and maximize benefits.