**Implementation Project - Pt 3 of 3**

Eswar Vardhan

Rivier University

COMP-616-AO1.SP25: DATA MINING

Dr. Darlene Barker, Ph.D February 28, 202

**FINAL REPORT**

**1. Executive Summary**

This project implements a machine learning solution to predict hospital readmissions for diabetic patients within 30 days of discharge. Using advanced ensemble methods and addressing class imbalance challenges, we achieved an **AUC of 0.642** and **F1-score of 0.256** on the test dataset. The optimized **XGBoost** model correctly identifies **41.5% of patients** who will be readmitted, with **a precision of 18.6%.** This solution can help healthcare providers proactively identify high-risk patients and implement targeted interventions to reduce readmission rates, ultimately improving patient outcomes and reducing healthcare costs.

**2. Implementation Process**

**2.1 Data Exploration and Preprocessing**

The implementation began with exploring the diabetic\_data.csv dataset containing 101,766 patient records with 50 features. Key preprocessing steps included:

* **Feature Selection**: Selected 17 essential features based on domain knowledge and relevance
* **Missing Value Handling**: Replaced '?' placeholders with NaN and filled missing categorical values with mode
* **Categorical Encoding**: Applied Label Encoding to transform categorical variables
* **Target Definition**: Converted the 'readmitted' column to a binary classification problem (readmission within 30 days = 1, otherwise = 0)

Analysis of class distribution revealed significant imbalance with only 11.16% of cases belonging to the positive class, making this a challenging predictive task.

**2.2 Model Development Strategy**

The model development followed a systematic approach:

1. **Data Splitting**:
   * 70% training data
   * 15% validation data (for model selection and hyperparameter tuning)
   * 15% test data (for final evaluation)
2. **Feature Scaling**:
   * Applied StandardScaler to normalize numerical features
3. **Class Imbalance Handling**:
   * Implemented SMOTE (Synthetic Minority Over-sampling Technique) to create a balanced training dataset
4. **Model Selection and Training**:
   * Phase 1: Implemented Decision Tree, Neural Network, and Logistic Regression
   * Phase 2: Implemented Gradient Boosting and XGBoost
   * Final Phase: Optimized XGBoost, Gradient Boosting, and added Random Forest
5. **Hyperparameter Optimization**:
   * Fine-tuned model hyperparameters to improve performance
   * Implemented classification threshold optimization to balance precision and recall
6. **Evaluation**:
   * Used comprehensive metrics including AUC, precision, recall, F1-score, and confusion matrices
   * Implemented visualization of ROC and Precision-Recall curves
   * Analysed feature importance to understand key predictors

**3. Challenges Faced**

**3.1 Class Imbalance**

The most significant challenge was the highly imbalanced nature of the dataset, with only 11.16% of patients being readmitted within 30 days. This imbalance caused models to tend toward negative predictions, achieving high accuracy but poor recall on the positive class.

**Solution**:

* Implemented SMOTE to create synthetic samples of the minority class
* Adjusted class weights in model configurations
* Optimized classification thresholds to balance precision and recall

**3.2 Model Performance Trade-offs**

Each model presented different trade-offs:

* **Gradient Boosting**: High precision (0.308) but extremely poor recall (0.048)
* **XGBoost**: More balanced performance with better recall (0.451) but lower precision (0.196)

**Solution**:

* Developed an evaluation framework prioritizing F1-score and AUC
* Implemented threshold optimization to find the optimal balance
* Added feature importance analysis to improve model interpretability

**3.3 Computational Efficiency**

The Gradient Boosting model was particularly slow (56.72 seconds for training) compared to XGBoost (1.17 seconds).

**Solution**:

* Optimized hyperparameters for better efficiency
* Implemented execution time tracking to monitor performance
* Identified XGBoost as the best balance of predictive power and efficiency

**4. Results and Analysis**

**4.1 Model Comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | AUC | Precision | Recall | F1-Score | Training Time |
| XGBoost | 0.661 | 0.196 | 0.451 | 0.274 | 1.17s |
| Gradient Boosting | 0.659 | 0.182 | 0.525 | 0.270 | 56.72s |
| Random Forest | 0.655 | 0.167 | 0.618 | 0.263 | 23.71s |

**4.2 Feature Importance Analysis**

The most predictive features for readmission risk were:

1. **Number Inpatient**: Previous Inpatient Visits
2. **Discharge Disposition Id**: Where the Patient Was Discharged To
3. **Number Diagnoses**: Number of Diagnoses Entered
4. **Time in Hospital**: Duration of Hospital Stay
5. **Number Emergency**: Number of Emergency Visits Before the Current Encounter

This analysis provides valuable clinical insights for understanding readmission risk factors.

**4.3 Clinical Significance**

The final model's performance translates to:

* 41.5% of patients who will be readmitted within 30 days are correctly identified
* When the model predicts readmission, it is correct 18.6% of the time
* The model demonstrates moderate discriminative ability (AUC = 0.642)

While these metrics may appear modest, they represent a significant improvement over random chance (which would yield approximately 11% precision) and can provide meaningful clinical decision support.

**5. Future Improvements**

1. **Data Enhancement**
   * Add clinical variables (lab values, medications, comorbidities)
   * Introduce temporal features for patient history
   * Collect social determinants of health (insurance status, living situation)
2. **Methodological Improvements**
   * Use advanced sampling techniques (ADASYN, Borderline-SMOTE)
   * Explore deep learning for automated feature extraction
   * Implement stacked ensemble methods for model combination
3. **Deployment Considerations**
   * Develop an API for real-time hospital predictions
   * Create an interpretability layer for healthcare providers
   * Set up a monitoring system for ongoing model performance tracking

**6. Conclusion**

This project successfully developed a predictive model for hospital readmissions among diabetic patients. Despite the inherent challenges of class imbalance and complex clinical data, our optimized XGBoost model achieved meaningful predictive power that could translate to clinical utility. By identifying patients at high risk for readmission, healthcare providers can implement targeted interventions, potentially reducing readmission rates and associated healthcare costs while improving patient outcomes.

The implementation balances predictive performance with computational efficiency and provides interpretable results that can guide clinical decision-making. Future work will focus on enhancing the model with additional data sources and more sophisticated methodologies to further improve predictive performance.

.