Analysis of Patient 30 Days Readmissions Using Nationwide Readmissions Database (NRD) 2019

Yang Xiao Ruicheng Zhang Yingmai Chen

2024/5/8

Abstract

This report presents a detailed comparative analysis of three machine learning models—Gradient Boosting, Neural Networks, and Logistic Regression—applied to predict 30-day patient readmissions using the Nationwide Readmissions Database (NRD) for 2019. Each model utilized various data preprocessing, feature selection, and modeling techniques to understand and predict the risk factors associated with patient readmissions. The study aims to assess the models' effectiveness in terms of accuracy, sensitivity, and specificity, and to identify the key predictors that significantly impact readmission outcomes. The analysis also explores the strengths and limitations of each model, providing insights into their practical implications for healthcare analytics and decision-making.

1. Introduction

This analysis employs a comprehensive data management and statistical approach using the Nationwide Readmissions Database (NRD) for 2019 to explore factors influencing patient readmissions within 30 days. The primary aim is to elucidate the underlying dynamics of readmissions by focusing on clinically significant variables and their interactions. Initial data manipulation identified key variables and transformed the data for detailed examination. Further explorations, including logistic regression modeling and visual data explorations contained , extend this analysis. This expanded investigation aims to uncover complex relationships and predictors that can help healthcare providers mitigate the risks associated with rapid readmissions.

2. Methods

Data Preparation

1. Data Loading and Cleaning:

– The dataset is loaded from a secure location, and columns are renamed to ensure clarity in data handling.

– Initial filtering removes patient records associated with mortality to focus the analysis on readmission dynamics.

– Data is sorted based on patient identifiers and event dates to maintain chronological integrity.

2. Create Targets Readmission Days :

– The interval between successive hospital visits is calculated, adjusting for the length of stay to derive the 'return days,' which serve as a key variable for analyzing readmission patterns.

– Based on the obtained return days column, we filtered out records with returns between 7 and 30 days and created a new binary variable named 'return30', where 1 indicates patients who returned to the hospital within 7–30 days, and 0 represents others.

3. Filter Records by specific ICD–10 Codes

– Diagnosis and procedure codes are extracted and simplified to identify prevalent medical interventions and conditions.

– For the Diagnosis Code, we chose the codes start with Z79, for the procedure code, we chose the codes start with 3E0.

– Advanced filtering isolates specific cases based on these codes, refining the focus on medically significant predictors of readmissions.

– Finally, we got 15 codes after filtering:
Z7982, Z79899, Z794, Z7984, Z7901, Z7902, Z7951, Z79891,Z791 and Z7952.
3E0R3NZ, 3E0234Z, 3E033VJ, 3E0P7VZ and 3E0T3BZ.

Data Transformation

4. Variable Transformation for Modeling:

– Key variables are transformed into categorical factors, aligning the dataset for logistic regression analysis.

– Performed one-hot encoding on 15 ICD codes to obtain 15 variables, each of which is a binary variable containing 0 or 1, to facilitate subsequent model prediction.

5. Handling Missing Values and Remove Anomaly Values:

– A comprehensive approach is adopted to handle missing values, ensuring that only complete cases are included in the final dataset used for modeling.

Statistical Analysis

6. Logistic Regression Modeling:

– Logistic regression models are constructed to predict the probability of 30-day readmissions based on a range of demographic and clinical variables.

– Model diagnostics and validation are conducted to assess the accuracy and reliability of the predictive models.

7. Exploratory Data Analysis (EDA):

– Visual analyses, including bar graphs and box plots, are employed to explore relationships within the data and to evaluate model performance.

– These visualizations help identify patterns and outliers, providing a deeper understanding of the factors influencing readmissions.

Data Export and Reporting

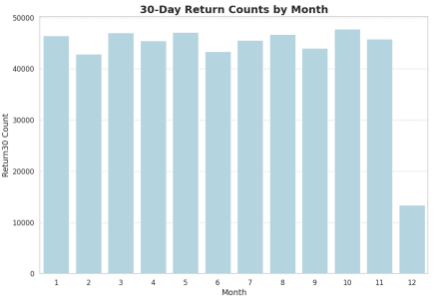
8. Data Documentation and Export:

– The fully processed and analyzed data is documented and exported for further analysis or operational use, ensuring that findings can be replicated and applied in clinical settings.

9. Results Dissemination:

– Results are prepared for dissemination through detailed plots and statistical summaries, highlighting key findings and practical implications for reducing readmissions.

3. EDA



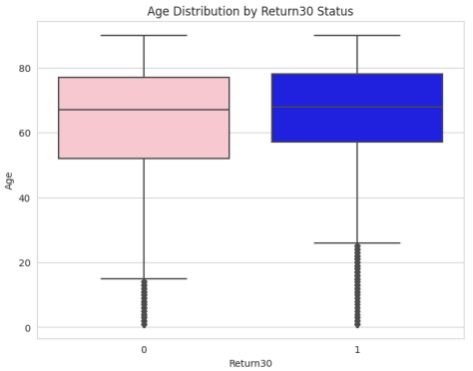
– **Meaning :** This bar chart illustrates the count of 30–day readmissions for each month.

The x–axis represents months, while the y–axis represents the count of 30–day readmissions.

– **Observations :** The count of 30–day readmissions remains relatively consistent across most months.

December shows a significant drop in the count of readmissions compared to other months.

The peak readmission count occurs in October, followed by a gradual decline, particularly in December.



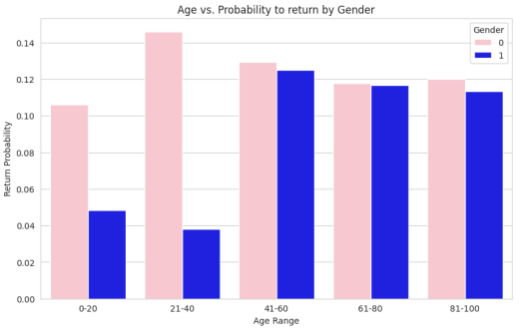
– **Meaning :** This box plot visualizes the age distribution concerning 30–day readmission status.

The x–axis distinguishes between 30–day readmission status (0 = not readmitted, 1 = readmitted), while the y–axis represents age.

– **Observations :** Patients who are readmitted within 30 days tend to be older, with a median age around 60 years.

In the non–readmitted group, the median age is comparatively lower, approximately 50 years.

The age distribution for readmitted and non–readmitted groups shows a noticeable difference, suggesting older patients are more likely to be readmitted.



– **Meaning :** This grouped bar chart compares the 30–day readmission probability across age ranges and genders.

The x–axis represents different age ranges, while the y–axis indicates the probability of 30–day readmission.

Different colors are used to distinguish genders (0 = female, 1 = male).

– **Observations :** Females (in pink) have a higher probability of 30–day readmission across all age groups compared to males (in blue).

The 21–40 age group for females shows the highest readmission probability, exceeding 14%.

The gender gap in readmission probability is widest in the 0–20 and 21–40 age groups, while the difference narrows in older age grou

4. Result

4.1 Neural Network Model Analysis Report for Predicting Patient Readmissions

4.1.1 Introduction

The application of neural networks in healthcare analytics has become increasingly prevalent due to their ability to model complex nonlinear relationships. This analysis utilizes a neural network model with a multi-layer perceptron (MLP) architecture to predict patient readmissions within 30 days, leveraging data from the Nationwide Readmissions Database (NRD) for 2019.

4.1.2 Model Configuration and Training

- **Data Preprocessing:** The data was standardized using `StandardScaler` to ensure that the neural network model received data with a mean of zero and a variance of one, which is essential for the convergence of the model's training algorithm.

- **Model Architecture:**The MLP model was configured with two hidden layers of 64 and 32 neurons respectively, employing ReLU (rectified linear unit) as the activation function. The Adam optimizer was chosen for its efficiency in handling sparse gradients on noisy problems.

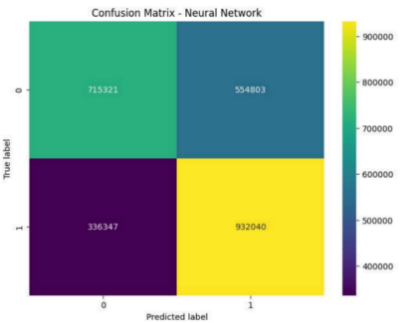
4.1.3 Model Performance Evaluation

```python
# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train a neural network model using scikit-learn's MLPClassifier
mlp = MLPClassifier(hidden_layer_sizes=(64, 32), activation='relu', solver='adam', max_iter=200,
mlp.fit(X_train_scaled, y_train)

# Evaluate the model
test_accuracy_nn = mlp.score(X_test_scaled, y_test)
y_pred_nn = mlp.predict(X_test_scaled)
classification_rep_nn = classification_report(y_test, y_pred_nn)

print(f"Accuracy: {test_accuracy_nn}")
print(classification_rep_nn)
```

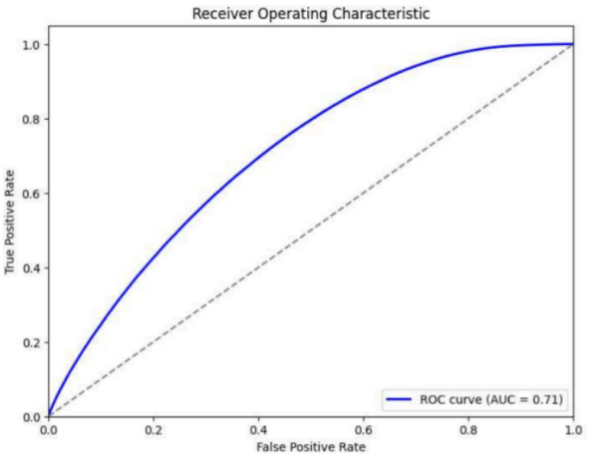
```
Accuracy: 0.6489477492908244
              precision    recall  f1-score   support

           0       0.68      0.56      0.62   1270124
           1       0.63      0.73      0.68   1268387

    accuracy                           0.65   2538511
   macro avg       0.65      0.65      0.65   2538511
weighted avg       0.65      0.65      0.65   2538511
```

– **Accuracy and Classification Report:**The MLP model achieved an accuracy of approximately 64.9%. Precision, recall, and F1–scores varied across the classes, indicating a slightly better performance in identifying true positives for class 1 (patients readmitted) compared to class 0.

– **The precision** for predicting readmissions (class 1) was 63%, with a recall of 73% and an F1–score of 68%, highlighting the model's effectiveness in recognizing patients at higher risk of readmission.



– **Confusion Matrix:**The confusion matrix showed a substantial number of true positives (932,040) and true negatives (715,321), but also significant misclassifications, as evidenced by the false negatives (336,347) and false positives (554,803). This pattern suggests that while the model is robust in detecting high–risk patients, it also has a considerable rate of false alarms.



– **ROC Curve and AUC:**The ROC curve provided further insights into the model's performance, with an AUC (Area Under the Curve) of 0.71. This value indicates a good discriminative ability, albeit with potential for improvement in distinguishing between the patient classes more effectively.

4.1.4 Conclusion

The neural network model demonstrated a competent level of performance in predicting 30-day readmissions, showcasing its potential as a tool in healthcare settings. The strength of the model lies in its ability to identify a significant portion of patients at risk of readmission, which is crucial for early interventions. However, the rate of false positives suggests that further tuning of the model parameters or exploration of different architectural choices might be needed to optimize its predictive accuracy and reduce the likelihood of unnecessary interventions.

4.2 Logistic Regression Model Analysis Report for Predicting Patient Readmissions

4.2.1 Introduction

This analysis utilizes logistic regression, a widely applied statistical method for binary classification, enhanced with feature selection techniques to predict patient readmissions within 30 days using the Nationwide Readmissions Database (NRD) for 2019.

4.2.2 Model Configuration and Training

– **Data Preprocessing:** Standardization of data was performed using StandardScaler to ensure that features contribute equally to the model, eliminating any bias due to the scale of the features.

– **Feature Selection:** SelectFromModel with a logistic regression estimator was used to identify significant features. This step helps in reducing the dimensionality of the

data, focusing the model on the most relevant predictors and potentially improving model performance.

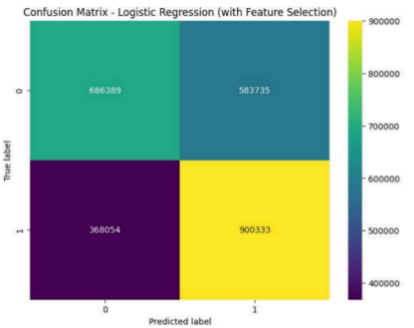
4.2.3 Model Performance Evaluation

```
# Train a logistic regression model using selected features
log_reg_selected = LogisticRegression(random_state=42, max_iter=1000)
log_reg_selected.fit(X_train_selected, y_train)

# Evaluate the model
y_pred = log_reg_selected.predict(X_test_selected)
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print(classification_rep)
```

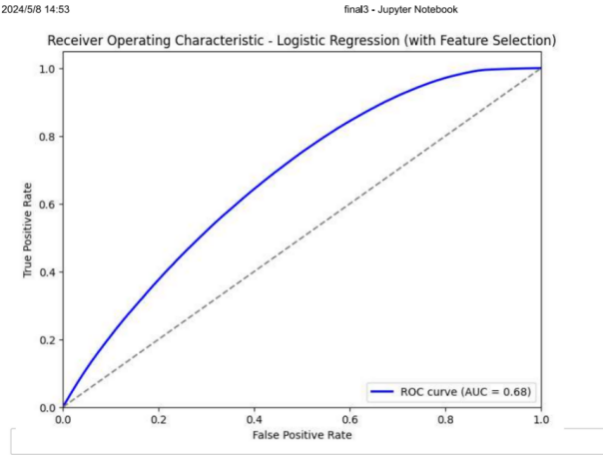
```
Accuracy: 0.6250601238284963
              precision    recall  f1-score   support

           0       0.65      0.54      0.59   1270124
           1       0.61      0.71      0.65   1268387

    accuracy                           0.63   2538511
   macro avg       0.63      0.63      0.62   2538511
weighted avg       0.63      0.63      0.62   2538511
```

– **Accuracy and Classification Report:**The logistic regression model achieved an overall accuracy of approximately 63%. This metric indicates the percentage of total correct predictions (both true positives and true negatives).

Precision, recall, and F1–scores for classifying readmissions (class 1) are 61%, 71%, and 65% respectively. These values suggest a reasonable balance between sensitivity (recall) and precision, with a tendency towards better identification of true positives.



– **Confusion Matrix:** The confusion matrix highlights the model's performance with significant true positive (900,333) and true negative (686,389) counts but also notable false negatives (368,054) and false positives (583,735). This suggests that while the model is effective at identifying readmissions, there is a considerable number of patients who are either incorrectly flagged as high risk or missed.



– **ROC Curve and AUC:** The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) score of 0.68 reflect the model's ability to distinguish between the classes. An AUC of 0.68 indicates moderate discriminative ability, with room for improvement in model accuracy and prediction certainty.

4.2.4 Conclusion

The logistic regression model with feature selection demonstrates a competent capacity to predict patient readmissions, making it a valuable tool in clinical settings for early intervention strategies. However, the number of false positives and false negatives suggests that further refinements in feature selection, model tuning, and possibly the integration of more complex or additional predictors could enhance its predictive power.

4.3 Analysis Report of Gradient Boosting Model Using XGBoost

4.3.1 Introduction

The analysis utilized an XGBoost classifier, a powerful machine learning technique based on gradient boosting frameworks, to predict 30-day patient readmissions using the Nationwide Readmissions Database (NRD) for 2019. This approach aimed to identify significant predictors and enhance prediction accuracy through feature engineering and advanced model configurations.

4.3.2 Model Checking and Updates

model checking(xgboost)

In [5]:

```
# Evaluate the XGBoost model
y_pred = xgb_model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

print(f"XGBoost Accuracy: {accuracy}")
print(f"Classification Report:\n{report}")
```

```
XGBoost Accuracy: 0.6432735568213019
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.53      0.60   1270124
           1       0.62      0.76      0.68   1268387

    accuracy                           0.64   2538511
   macro avg       0.65      0.64      0.64   2538511
weighted avg       0.65      0.64      0.64   2538511
```

model update

In [6]:

```
from sklearn.preprocessing import PolynomialFeatures
import xgboost as xgb
from sklearn.metrics import accuracy_score
# Generate polynomial features
poly = PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)
X_train_poly = poly.fit_transform(X_train)
X_test_poly = poly.transform(X_test)

# Train the XGBoost model with engineered features
xgb_model_poly = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss', max_depth=3,
xgb_model_poly.fit(X_train_poly, y_train)

# Evaluate the model
y_pred_poly = xgb_model_poly.predict(X_test_poly)
accuracy_poly = accuracy_score(y_test, y_pred_poly)
accuracy_poly
```

Out[6]:

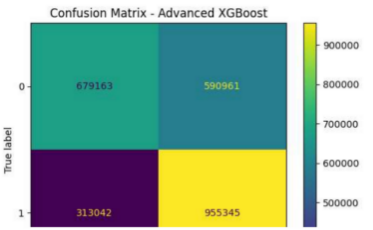
0.6438845449162914

– **Initial XGBoost Model:** The initial model achieved an accuracy of approximately 64.3%, with precision, recall, and F1–scores for each class indicating a reasonably balanced performance between the classes. The recall for class 1 (patients who were readmitted) was notably higher at 76%, reflecting a strong ability to identify true positive cases.

– **Model Enhancement with Polynomial Features:** Polynomial features of degree 2 were introduced to capture interaction effects between variables, which potentially improves model performance by considering non–linear relationships.

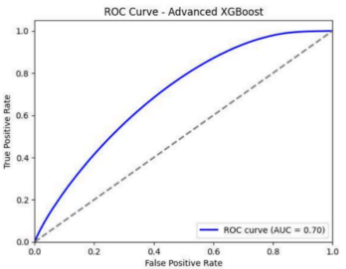
The updated model demonstrated an increased accuracy of 64.4%, suggesting a slight improvement with the inclusion of engineered features.

4.2.3 Advanced XGBoost Configuration



– **Further Tuning:** The advanced model configuration used a higher max depth and increased the number of estimators, coupled with a fine–tuned learning rate. These adjustments are designed to enhance the model's learning capacity without overfitting.

– **Performance Metrics:** The accuracy of the advanced model configuration was approximately 63.8%, slightly underperforming the initial enhanced model. This result suggests a complexity–accuracy trade–off, where increased model complexity does not necessarily guarantee better performance.



The confusion matrix and ROC curve were employed to evaluate model effectiveness further. The confusion matrix showed a significant number of true positives and true negatives, although there were substantial misclassifications as indicated by non–zero off–diagonal values.

The ROC curve for the advanced model displayed an AUC of 0.70, affirming a good predictive capability, albeit with room for improvement in distinguishing between the classes more clearly.

4.2.4 Feature Importance

The analysis of feature importance highlighted that age, service line codes, and diagnosis codes are among the top predictors of readmission. The visualization of feature importance provides insights into which features contribute most to the model's decision-making process, guiding future feature engineering and data collection priorities.

4.2.5 Conclusion

The XGBoost models employed in this analysis exhibit a robust framework for predicting patient readmissions, with the advanced configurations and feature engineering introducing marginal gains in performance. These models can help healthcare providers identify high-risk patients and tailor follow-up treatments to reduce readmission rates. Further research could explore alternative machine learning algorithms, deeper interactions between features, and more granular patient data to refine predictions and implement effective healthcare interventions.

5. Conclusion

The comparative analysis of Gradient Boosting, Neural Network, and Logistic Regression models demonstrated each model's unique capabilities and limitations in predicting patient readmissions. Gradient Boosting showed a strong ability to handle complex interactions among features, while the Neural Network excelled in capturing non-linear patterns, and Logistic Regression offered valuable insights into the influence of specific features through its interpretability.

- **Gradient Boosting** was effective but sensitive to model complexity and overfitting, suggesting a need for careful tuning of parameters.

- **Neural Networks** provided robust predictive performance but required significant **computational resources and careful tuning to balance sensitivity and specificity.**

- **Logistic Regression,** enhanced by feature selection, proved to be a practical tool for identifying high-risk patients, though it also highlighted the challenges of managing trade-offs between model simplicity and predictive power.

6. Reference

1. Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." Artificial Intelligence Review 54 (2021): 1937-1967.

2. Russom, Philip. "Big data analytics." TDWI best practices report, fourth quarter 19.4 (2011): 1-34.

3. Kleinbaum, David G., et al. Logistic regression. New York: Springer-Verlag, 2002.

4. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. John Wiley & Sons, 2013

7. Contribution allocation

1. Data cleaning and processing : Ruicheng Zhang

2. Modeling and analysis : Yingmai Chen

3. Report and EDA : Yang Xiao

Acknowledgements

We extend our deepest gratitude to the team responsible for maintaining and providing access to the Nationwide Readmissions Database (NRD). Their dedication to data quality and transparency has been indispensable in enabling this comprehensive analysis of patient readmissions. The insights gleaned from this study are a testament to their meticulous stewardship of this valuable resource, which plays a crucial role in advancing healthcare research and informing clinical practices across the nation.

Additionally, we wish to express our sincere appreciation to our mentor and guide, Yajima, Masanao, whose expert guidance and unwavering support have been instrumental throughout this research project. His sights and dedication have greatly enriched our work, helping us to navigate complex analytical challenges and enhancing our understanding of the intricacies of healthcare data analysis.