Diabetes Less than 30 days readmission

Brian Reppeto

Bellevue University

DSC550-T302 Data Mining

Brett Werner

May 30, 2024

Diabetes Less than 30 days readmission

Introducing the Problem

Hospital readmissions pose significant challenges in the management of chronic diseases like diabetes. Readmissions are not only costly for healthcare systems but also indicate poor disease management and suboptimal patient outcomes. Addressing the issue of hospital readmissions can significantly improve patient care and reduce healthcare costs.

Importance of Solving the Problem

Predicting hospital readmissions for diabetic patients is crucial because it allows healthcare providers to identify high-risk patients and intervene before readmission occurs. This proactive approach can improve disease management, reduce hospital readmissions, lower healthcare costs, and enhance the quality of life for diabetic patients.

Pitching the Problem to Stakeholders

To gain stakeholder buy-in, it's essential to highlight both the financial and operational benefits. Predictive modeling can help hospitals reduce penalties associated with high readmission rates and improve patient satisfaction scores. By demonstrating the potential for cost savings, improved patient outcomes, and enhanced hospital performance, stakeholders will see the value in investing in predictive analytics.

Data Source

The dataset used for this project was obtained from the UCI Machine Learning Repository. The dataset, titled "Diabetes 130-US Hospitals for Years 1999-2008," was created by (John Clore, 2014). It contains 101,766 records of diabetic patients, providing a comprehensive listing of demographics, medical history, and readmission status.

Milestones Summary

Milestone 1:

Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset's structure and uncover patterns related to hospital readmissions. Key insights include:

- Age distribution: Most patients were within the age ranges of 60-70 and 70-80.

- Medication usage: Patients were on a wide range of medications, with a median of around 13.

- Readmission rates: Most patients were not readmitted within 30 days.

Visualizations:

**A graph of a number of days in hospital

Description automatically generated**

**A chart of medication versus readmission

Description automatically generated**

**A graph of different colored rectangular bars

Description automatically generated**

A graph of a number of people

Description automatically generated with medium confidence

Milestone 2:

Data Preparation

To ensure the dataset was clean and ready for predictive modeling, several features were dropped.

- Encounter ID and Patient Number: These are unique identifiers for each hospital stay and patient, respectively. They do not provide any predictive value for readmission.

- Weight: The weight column had over 95% missing data, making it unsuitable for analysis.

- Payer Code & Medical Specialty: These columns had a high proportion of missing values. Their relevance to readmission might be limited compared to clinical features.

The data preparation process involved several steps to handle missing values, encode categorical variables, and engineer new features that could enhance the model's predictive capabilities.

Data Cleaning and Missing Values

Missing values were handled appropriately to ensure the integrity of the dataset. Initially, any missing values were identified and replaced with NaN for clarity. Columns with a high percentage of missing data, such as weight, payer\_code, and medical\_specialty, were dropped. Non-relevant identifiers like encounter\_id and patient\_nbr were also removed. Additionally, patients with discharge dispositions indicating they expired or were admitted to hospice care (and thus not readmittable) were filtered out of the dataset. For the remaining missing values, numerical columns were filled with the median, and categorical columns were filled with the most frequent value (mode).

Feature Engineering

Several new features were created to capture interactions between existing variables, which could be significant predictors of readmission risk. The age feature, originally in ranges, was converted to a numerical format representing the midpoint of each range. An interaction feature, meds\_x\_time, was created to capture the interaction between the number of medications a patient is on and their length of stay in the hospital. Additionally, the target variable readmitted was encoded as a binary variable, where 1 indicates readmission within 30 days and 0 otherwise.

Encoding Categorical Variables

Categorical variables were transformed into a format suitable for machine learning models using one-hot encoding. This process involved creating dummy variables for each category, ensuring that the machine learning models could process them correctly.

Summary of Prepared Data

After data preparation, the dataset was structured with numerical and categorical features properly encoded and new interaction features created to enhance the predictive power of the model. The data prep steps ensured that the dataset was clean, relevant, and ready for the next phase of building and evaluating predictive models for hospital readmissions among diabetic patients.

Milestone 3:

Model Building and Evaluation

The goal of Milestone 3 was to build and evaluate a predictive model for hospital readmissions within 30 days using the diabetic dataset. The focus was on selecting the most effective machine learning techniques, tuning model parameters, and evaluating the model's performance on key metrics.

Loading and Cleaning Data

The dataset was loaded, and missing values were handled by replacing them with NaNs and dropping irrelevant columns such as weight, payer\_code, and medical\_specialty. Further, columns with high percentages of missing values were also removed. Patients who could not be readmitted (expired or admitted to hospice) were filtered out.

Feature Engineering

New interaction features were created to capture potentially significant predictors of readmission risk. For example, the num\_medications\_age feature was created to represent the interaction between the number of medications a patient is on and their age. Similarly, the num\_lab\_procedures\_num\_medications feature captured the interaction between the number of lab procedures and medications.

Encoding and Transforming Data

Categorical variables were transformed into a format suitable for machine learning models using one-hot encoding. The target variable `readmitted` was encoded as a binary variable (1 for readmitted within 30 days, 0 otherwise). Remaining non-numeric columns were encoded to numerical codes.

Handling Class Imbalance

The class imbalance issue was addressed using the SMOTE Tomek technique, which combines oversampling of the minority class (SMOTE) with under sampling of the majority class (Tomek links) to balance the training data.

Model Building and Hyperparameter Tuning

The random forest classifier was selected for its robustness and ability to handle many features. Initial models were built to establish a baseline performance.

Feature selection was performed using LassoCV to identify the most important features for the model. This regularization technique helps in reducing overfitting by penalizing less significant features.

Hyperparameters of the random forest model were tuned using RandomizedSearchCV to find the optimal combination of parameters. The parameter grid included variations in the number of estimators, maximum depth, minimum samples split, and minimum samples leaf, among others.

The model's performance was evaluated using multiple metrics including accuracy, ROC-AUC, precision, recall, F1 score, and balanced accuracy. These metrics provided a comprehensive view of the model's predictive capabilities and robustness.

Random Forest with Best Parameters: Accuracy: 0.8863761998090356

ROC-AUC: 0.5113573492136174

Precision: 0.4105960264900662

Recall: 0.02775290957923008

F1: 0.0519916142557652

Balanced Accuracy: 0.5113573492136173

Milestone 4:

The goal of Milestone 4 was to improve on the F1 score from the prior Milestone as the dataset was unbalanced and evaluate additional machine learning models to predict hospital readmissions within 30 days for diabetic patients. The approach involved prepping the data, building and tuning models, and evaluating their performance to identify the best predictive model.

Data Preparation

- The dataset was loaded and cleaned by replacing missing values (?) with NaNs.

- Unnecessary columns such as encounter\_id, patient\_nbr, and payer\_code were dropped.

- Age and weight ranges were converted to average numerical values. Missing values in weight were filled with the mean.

Categorizing Variables

- The medical\_specialty column was categorized into predefined groups such as high\_frequency, low\_frequency, pediatrics, psychic, neurology, surgery, and ungrouped.

Handling Missing Values

- Numerical columns with missing values were filled with the mean.

- Medication columns were filled with 0 if NaN and 1 if not.

- Remaining NaNs in categorical columns were filled with the mode.

Feature Engineering

- New interaction features were created, such as num\_medications\_age, num\_lab\_procedures\_num\_medications, num\_medications\_time\_in\_hospital, and num\_procedures\_time\_in\_hospital.

Encoding Variables

- Categorical variables were encoded using one-hot encoding.

- The target variable readmitted was encoded as a binary variable (1 for readmitted within 30 days, 0 otherwise).

Scaling Features

- Features were scaled using MinMaxScaler to ensure they were on a similar scale.

Model Building and Tuning

- The dataset was split into training and testing sets using a 70-30 split.

- The SMOTE-Tomek technique was applied to the training data to handle class imbalance.

Model Selection and Hyperparameter Tuning

- Several models were considered: XGBoost, Random Forest, Gradient Boosting, and Logistic Regression.

- GridSearchCV was used to tune hyperparameters for each model:

- XGBoost: Parameters such as max\_depth, learning\_rate, n\_estimators, subsample, and colsample\_bytree were tuned.

- Random Forest: Parameters such as n\_estimators, max\_depth, min\_samples\_split, and min\_samples\_leaf were tuned.

- Gradient Boosting: Parameters such as learning\_rate, n\_estimators, subsample, and max\_depth were tuned.

Model Evaluation

- The best models from GridSearchCV were used to create an ensemble model using VotingClassifier.

- The ensemble model combined the strengths of individual models by using soft voting.

Model Evaluation

- The ensemble model was evaluated on the test set using accuracy, precision, recall, F1 score, and a classification report.

- The initial accuracy of the ensemble model was 87.7%.

Ensemble Model Classification Report:

Ensemble Model Accuracy: 0.8769586954333456

Ensemble Model Classification Report:

precision recall f1-score support

0 0.89 0.98 0.93 26362

1 0.37 0.09 0.14 3441

accuracy 0.88 29803

macro avg 0.63 0.53 0.54 29803

weighted avg 0.83 0.88 0.84 29803

Adjusted Threshold Evaluation

The adjusted accuracy was 73.9%, and the adjusted classification report provided insights into the model's precision and recall at the chosen threshold.

Adjusted Accuracy: 0.7386504714290507

Adjusted Classification Report:

precision recall f1-score support

0 0.91 0.78 0.84 26362

1 0.21 0.44 0.28 3441

accuracy 0.74 29803

macro avg 0.56 0.61 0.56 29803

weighted avg 0.83 0.74 0.78 29803

Conclusion

The model implementation and evaluation process demonstrated that an ensemble model combining XGBoost, Random Forest, Gradient Boosting, and Logistic Regression provided robust predictions for hospital readmissions within 30 days for diabetic patients. The use of SMOTE Tomek for handling class imbalance and careful hyperparameter tuning were crucial in achieving good performance.

Recommendations

Model Deployment: Before deployment, further validation on external datasets and real-time testing are necessary. Integration with HER systems will enable seamless use in clinical settings. Additionally, further efforts can be had to help with the class imbalance.

Continuous Monitoring: The model should be continuously monitored and updated to maintain accuracy and relevance.

Further Research: Future work could explore additional features, alternative models, or ensemble techniques to further improve predictive performance. Additionally, a graphical interface could be built to allow for individual patients to be screened real-time.

The insights from this model can aid healthcare providers in identifying high-risk patients and implementing targeted interventions to reduce readmission rates, ultimately improving patient outcomes and reducing healthcare costs.

References

John Clore, K. (2014). Diabetes 130-us hospitals for years 1999-2008 [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/c5230j>