Week 5\_2 Predictive Case Study: Hospital Readmissions

Brian Reppeto

Bellevue University

DSC 630

Frank Neugebauer

September 29, 2024

Week 5\_2 Predictive Case Study: Hospital Readmissions

Introduction

The rising cost of healthcare in the United States continues to be a growing concern, with unplanned patient readmissions being a significant contributor. Unplanned readmissions, particularly those within 30 to 90 days post-discharge, are seen as indicators of the quality of care a hospital provides and represent a major challenge for healthcare system. With the implementation of the Hospital Readmission Reduction Program (HRRP) by the Center for Medicare and Medicaid Services (CMS) in October 2012, reducing readmission rates has become a priority for healthcare providers. The program imposes financial penalties on hospitals with higher-than-expected readmission rates, further emphasizing the need to address this issue.

Studies have shown that about 15 to 25 percent of discharged patients are readmitted within 30 days, and in 2011, unplanned readmissions cost hospitals approximately $41.3 billion (Shinkman, 2014). This staggering cost highlights the importance of accurately predicting and preventing unplanned readmissions to improve healthcare quality and reduce financial burdens on the system. Predictive analytics offers a promising solution by leveraging historical data to identify patients at high risk of readmission, enabling healthcare providers to intervene proactively.

In this case study, Avishek Choudhury and Dr. Christopher M. Greene developed a predictive model to forecast unplanned readmission risks among patients, using various machine learning techniques. Their goal was to build a model accurate enough to be implemented in real-world clinical settings, allowing hospitals to identify high-risk patients and take preventive measures.

The data used in this study was acquired from a large dataset utilizing a sample of 100,000 instances from 130 U.S. hospitals over ten years. The data contains 55 attributes related to patient demographics, clinical factors, and hospital stays. The inclusion of such a diverse and extensive data set ensures that the model captures a wide range of factors that could influence readmission risk.

**Methods and Results**

**Data Preparation**

Data preparation is a critical step in any predictive analytics project, as the quality of the input data significantly impacts model performance. In this study, the dataset contained numerous challenges typical of real-world healthcare data, such as missing values, noisy information, and class imbalance. To address these issues, the researchers employed several data preprocessing techniques:

1. **Handling Missing Values**: The dataset had some attributes with a high percentage of missing values, such as "weight" (97%), "payer code" (40%), and "medical specialty" (47%) (Choudhury & Greene, 2018). The "weight" attribute was excluded from further analysis due to its high level of missing data. The "payer code" was also ignored as it was deemed irrelevant to the outcome. For the "medical specialty" attribute, missing values were replaced with the label "missing" to ensure the completeness of the data.
2. **Feature Selection**: Large datasets can hinder algorithm speed and accuracy, so selecting relevant features is essential. The study used the Boruta algorithm and stepwise regression to identify the most important attributes. The Boruta algorithm is a wrapper built around the random forest classification algorithm that identifies the relevance of each attribute by measuring the impact on classification accuracy. Stepwise regression, an automatic computational procedure, was also implemented to refine the selection further.
3. **Data Balancing**: One of the major challenges with the dataset was the imbalance between readmitted and non-readmitted cases. The researchers used several techniques, such as oversampling, under sampling, and rose sampling, to create a more balanced dataset. This ensured that the model would not be biased toward predicting the majority class.
4. **Data Partitioning**: After preprocessing, the dataset was split into training (70%) and testing (30%) sets (Choudhury & Greene, 2018). This division allowed the researchers to train the model on one subset of data and then evaluate its performance on a separate, unseen dataset.

**Modeling Techniques Used**

The study employed several machine learning algorithms to build predictive models, including:

1. **Random Forest (RF)**: A type of decision tree algorithm that selects a random subset of features at each split during training. This process, known as "feature bagging," helps identify the most important predictors and reduces the risk of overfitting.
2. **Support Vector Machine (SVM)**: A discriminative classifier that creates a hyperplane to separate data points into different classes. The SVM was chosen for its ability to handle high-dimensional data and create complex decision boundaries.
3. **Recursive Partitioning and Regression Tree (Rpart)**: This algorithm recursively splits the dataset based on feature values, making it suitable for capturing non-linear relationships.
4. **Gradient Boosting Method (GBM)**: An ensemble learning technique that combines multiple weak predictive models to produce a more robust and accurate model. GBM iteratively builds decision trees and adjusts them based on the residual errors of the previous trees.
5. **General Linear Model (GLM)**: A flexible generalization of ordinary linear regression that allows for different distributions of the response variable.

**Why These Methods Were Chosen**

The selection of these algorithms was based on their ability to handle the complexities and nuances of the dataset. For instance, Random Forest is well-suited for large datasets with many features, while SVM is effective for high-dimensional data. GBM was chosen for its ability to boost the performance of weak learners, and the GLM offered a more interpretable model for understanding relationships between features and the target variable.

**Evaluation Metrics Used**

The models were evaluated based on accuracy, sensitivity, and specificity:

* **Accuracy** measured the overall correctness of the model's predictions.
* **Sensitivity** (true positive rate) assessed the model's ability to correctly identify patients at high risk of readmission.
* **Specificity** (true negative rate) measured the model's ability to correctly identify patients who were not at risk of readmission.

These metrics were chosen to ensure balanced performance, as focusing solely on accuracy could lead to overlooking important false-negative cases, which could have severe consequences in a healthcare setting.

**Results**

The Gradient Boosting Method (GBM) emerged as the most accurate model, achieving a 98.5% prediction accuracy (Choudhury & Greene, 2018). Genetic Algorithm and Greedy Ensemble techniques were used to optimize the models, further enhancing their performance. The study identified "number of inpatients," "number of emergencies," "number of diagnoses," "diabetes medication," "number of outpatients," and "number of procedures" as the most influential factors affecting readmission risk.

**Conclusion**

The predictive model developed in this case study demonstrated exceptional accuracy and potential for real-world application. The implementation of this model in clinical settings could allow healthcare providers to identify patients at high risk of readmission and take preventive measures, such as offering intensive post-discharge care or scheduling follow-up appointments.

By integrating this predictive analytics model into the hospital's decision support system, hospitals can proactively manage patient care, reduce readmission rates, improve patient outcomes, and minimize costs. The study also highlighted the importance of data preprocessing, as techniques such as feature selection, data balancing, and handling missing values played a critical role in improving model accuracy.

The case study provided valuable insights into the factors contributing to patient readmission and demonstrated how predictive analytics could be used to enhance healthcare quality. The researchers recommended exploring alternative variable selection techniques like LASSO (Least Absolute Shrinkage and Selection Operator) and incorporating additional healthcare cost measures to refine the model further.

In conclusion, this case study emphasizes the potential of predictive analytics in healthcare, offering a viable solution for addressing unplanned readmissions and improving the quality of care. As the healthcare industry continues to collect more data, such models can become even more accurate and offer actionable insights, ultimately contributing to better patient outcomes and reduced healthcare costs.

References

Choudhury, A., & Greene, D. M. (2018). Evaluating patient readmission risk: A predictive analytics approach. American Journal of Engineering and Applied Sciences, 11(4), 1320–1331. Retrieved September 23, 2024, from <https://doi.org/10.3844/ajeassp.2018.1320.1331>

Shinkman, R. (2014, April 20). Readmissions lead to $41.3B in additional hospital costs. Fierce Healthcare. Retrieved September 23, 2024, from <https://www.fiercehealthcare.com/finance/readmissions-lead-to-41-3b-additional-hospital-costs>