**Predicting Hospital Readmission – Preliminary Analysis**

# Andrea Fox, Jolene Branch, & Sam Sears

# Data Science Program, Bellevue University

# DSC 630: Predictive Analytics

# Dr. Brett Werner

# October 4, 2020

**Abstract**

In healthcare, hospital readmissions can create undue costs and decrease the perceived quality of care at the facility. With a data set obtained through the UCI Machine Learning Repository, hospital readmissions for diabetic patients from over 130 US hospitals was examined. Variables surrounding demographic information, diagnoses, lab results, and more were analyzed for their relationship with the target variable; whether the patient was readmitted. Feature variables were identified and manipulated to be used in predictive modeling. The data was then split into different groups and fit to various machine learning classification models including K-Nearest Neighbors, Logistic Regression, Naïve Bayes, Stochastic Gradient Descent, Decision Tree, Random Forest, and Gradient Boosting Classifier. The models were then tested on both the training data for their fit and the testing data for their generalization. Key metrics including AUC, accuracy, precision, and recall were obtained and a confusion matrix was constructed in order to evaluate the models. Finally, next steps for model improvement and selection were given.

*Keywords:* K-Nearest Neighbors, Logistic Regression, Naïve Bayes, Stochastic Gradient Descent, Decision Tree, Random Forest, Gradient Boosting Classifier, Machine Learning, Classification Model, Predictive Analytics

**Predicting Hospital Readmissions – Preliminary Analysis**

The rising cost of healthcare in the United States has the potential to stall the economy as well as threaten national security (due to need to balance the budget by making cuts elsewhere). “Rising health-care costs stall Americans’ dreams of buying homes, building families and saving for retirement.” (Leonhardt, 2019). This project will look at demographic and clinical variables commonly collected on hospital admissions and determine their effectiveness in predicting hospital readmissions.

**Background of the Problem**

Inadequate social resources and suboptimal care transitions, especially in the form of poor or misunderstood discharge instructions, have been implicated for years as possible causes of readmission to the hospital within 30 days. Readmissions to the hospital also cause capacity challenges that can force medical centers to expand the number of beds (at great cost to the facility) (Hospital, n.d.).

**Problem Statement**

The need for repeated episodes of hospital care disrupts the patient's life, costs the healthcare industry billions of dollars each year, places demand on hospital bed capacity, and threatens the viability of hospitals with high readmission rates in the form of Medicare linking payment to the quality of hospital care.

Through analysis of the features associated with hospital readmissions, the goal was to determine which factors correlate strongly with hospital readmissions and create a predictive model which can accurately predict whether a patient will be readmitted based on those features.

# **Methods**

Since this is classification problem, model selection was based accordingly. Further, data preparation and feature engineering were done with the assumption that a classification algorithm would be used. The data was randomly split into training, testing, and validation datasets. Using scikit-learn, models explored were K-nearest neighbors, Logistic Regression, Stochastic Gradient Descent, Naïve Bayes, Decision Tree Classifier, Random Forest, and Gradient Boosting Classifier.

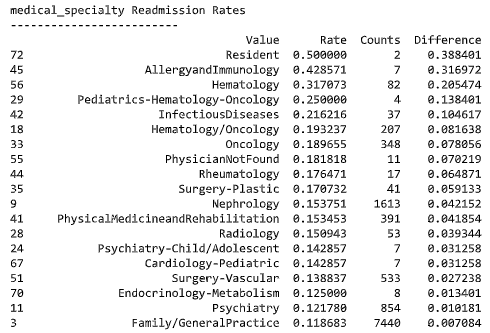
**Data Exploration**

Based on domain knowledge, a derived binary feature was created early in the data exploration process that differentiated between patients with a readmission within 30 days and those without one. This required combining the non-readmissions and the readmissions greater than 30 days, which we treated as a ‘0.’ This binary categorical variable was treated as our target variable for predicting whether a patient was readmitted or not.

With several categorical variables in the dataset, readmission rates for each category in the categorical variables were analyzed. The mean readmission rate for the entire data set was 11.2%. This was then compared to the readmission rates of each category to see if there were any that were significantly above or below average. Our thought here was that a below or above average readmission rate would signify some effect on the target variable. In looking through the data, there were some areas where we saw promise. For example, certain medical specialties for doctors showed higher readmission rates. Hematology had 82 occurrences in the dataset and had a 20% higher readmission rate than the mean. Another interesting bit to note was when looking at value counts of the age group by readmission, the 60-70 category had the highest number of readmissions. However, looking at the rate of readmission, the 20-30 age group had the higher rate of admission.

**Figure 1**

***Readmission Rates by medical\_specialty***

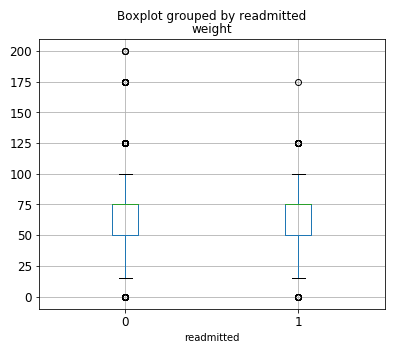


*Note: Difference was calculated by subtracting the mean readmission rate from readmission rate of value*

The next step was exploration of correlations of various features against the target feature, including weight, admission and discharge, length of stay, and diagnosis. Through group discussion based off domain knowledge of two of the group members, a few variables were tested to see if they had an association with readmission. First, we looked to see if weight had a correlation with readmission as it was hypothesized that higher weights would lead to more readmissions. This was not the case as the distributions of weights was almost the same for patients who were readmitted and those who were not as seen in Figure 2. Similar findings came from examining admission id, discharge id, time in hospital, and diagnosis correlations. In all these explorations, there was no clear evidence of a variable or value of a variable having strong association with readmission.

**Figure 2**

***Boxplot of weight grouped by readmission***



*Note: A patient that was readmitted is represented as a 1 and one that was not readmitted a 0.*

**Data Preparation**

Early irregularities in data exploration warranted further examination, resulting in the discovery that hospice patients (or those expected to expire within the next six months) and patients who died were contained within the dataset. Fortunately, this discovery occurred early in the process, and all instances with one of the six hospice/death classifications were removed.

The realization that the diagnoses are not based on the current ICD-10 medical diagnosis categorization system prompted retrieval of the previous (ICD-9) system codes. It was determined that the system of subcategories within categories presented too many diagnoses to use each as a separate variable. Since there were so many diagnosis codes, we plan on grouping them in a later iteration. For the preliminary modeling, they were removed.

Next, the medical specialty variable was examined. Even through certain medical specialties exhibited a higher rate of readmission than the mean, they were unequally represented. For example, when the patient was seen by a Resident, the readmission rate was 50% as seen in Figure 1, well over the mean rate. However, there were only two patients that had that medical specialty. In later iterations, we will continue to discuss how to group these effectively, but for preliminary modeling, we used only the top 10 represented medical specialties and grouped the remaining into another category labeled ‘other’.

Type conversions required included exploding out categorical variables and converting age to numeric bins using the lowest age of the range. Exploding categorical variables created a new variable for each unique value of the categorical variable. Because many of these categorical variables only had four of fewer unique values, it did not add an overwhelming number of variables. Since nearly all weight values were absent, the weight value was converted to a binary response, indicating simply presence of weight in the record or not.

Movement back and forth between the data preparation and data exploration stages ultimately resulted in a dataset with nine numeric predictor variables and 133 (binary) categorical predictors. With the target variable, the dataset used for predictive modeling contained 143 variables.

**Modeling**

To start, the data was randomly split into 15% test, 15% validation, and 70% training. Because our data set had far more patients not readmitted than readmitted, the training data was balanced so it would have equal proportions of the target variable. The training data was then scaled using unit variance to improve the performance of the classification models. Finally, the training data was used to fit 7 different classification models including K-Nearest Neighbors, Logistic Regression, Stochastic Gradient Descent, Naïve Bayes, Decision Tree, Random Forest, and Gradient Boosting. To evaluate the models, a function was created to calculate and display AUC, accuracy, precision, recall, and a confusion matrix. The fit models were used to make predictions on the training dataset to evaluate the fit of the model and then again on the test data set to evaluate the generalization of the model to new data.

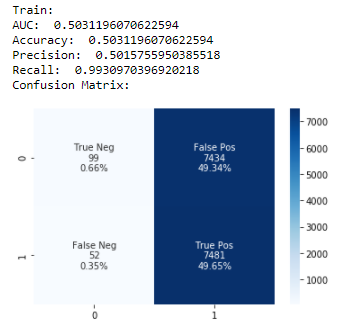
# **Results**

Since failing to predict a hospital readmission, or a type 2 error, is much worse for the hospital, insurance company, and patient than predicting a hospital readmission when there would not have been one, the focus was to reduce type 2 errors. This meant a model with high recall was desired.

The Naïve Bayes Classifier initially looked the best, with the highest recall, but review of the confusion matrix in Figure 3 found this model to have an unacceptable rate of true negatives. It was predicting almost all patients as ones that would need to be readmitted. Therefore, it did not add much value and had poor accuracy. The Gradient Boosting Classifier model performed well on the train dataset, but validation scores dropped a lot, which informed us that the model did not generalize well. The model was likely overfitting. Additional hyperparameter tuning is likely in order, such as adjusting the number of nearest neighbors to use to calculate probability for a given record’s result. In the decision tree model, tuning to specify how deeply to go in the tree is needed.

**Figure 3**

*Train Dataset Prediction Results of Naïve Bayes Classifier*



*Note: Results on the testing dataset were similar.*

# **Discussion/Conclusion**

So far, our evaluation of the data has challenged pre-conceived ideas that some of us had regarding our data and what we felt was going to influence the readmission rate. Many times, socio-economic factors play a role in healthcare data, but we have proved with this dataset that has not played a huge factor. Further, at this stage two models look to be performing slightly better than the rest: Random Forest and Gradient Boosting Classifier. Both have accuracies greater than 60% on both the validation and test sets and have higher recalls than the other models aside from Naïve Bayes. Our next steps will include the following: bin and include the diagnoses as a predictor variable, re-evaluate the binning of the medical specialty variable, analyze feature importance scores to see which variables to take out, research and optimize hyperparameters for each model, test the new models on the third dataset, and evaluate and select the best performing model.

# **Acknowledgements**

We want to acknowledge the University of California Irvine as their Machine Learning Repository is where we found our dataset that was used for this project. We would also like to acknowledge Andrew Long. His code gave us a reference point and guided our preliminary work through data preparation and modeling as well as sparking ideas to test regarding feature engineering.

**References**

Hospital Readmissions Reduction Program (HRRP). (n.d.). Retrieved September 11, 2020, from

<https://www.cms.gov/Medicare/Medicare-Fee-for-Service->

Payment/AcuteInpatientPPS/Readmissions-Reduction-Program

Leonhardt, M. (2019, November 04). Rising health-care costs stall Americans' dreams of buying homes, building families and saving for retirement. Retrieved from <https://www.cnbc.com/2019/11/04/health-care-costs-are-preventing-many-americans-from-hitting-life-milestones.html>

Long, A. (2020, January 30). Using Machine Learning to Predict Hospital Readmission for Patients with Diabetes with Scikit-Learn. Retrieved from <https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0>