Predicting Hospital Readmissions Using Population-Based Data

Introduction

High hospital readmission rates are associated with poor patient outcomes and high financial costs for healthcare organizations and private and government-sponsored insurance programs such as Medicare. This project investigates the relationship between a variety of health and demographic variables and hospital readmissions.

I initially began this project with a focus on individuals with recently-diagnosed congestive heart failure (CHF). The goal was to create a machine learning model that predicts the likelihood of a new patient with CHF being readmitted within the next year, by isolating risk factors that are correlated with CHF-related readmissions. However, as you will see, I was not able to fully carry out this goal. I found that the sample size of new onset CHF was inadequate to isolate only CHF-related readmissions. Therefore, I created a more generalized machine learning model to identify risk factors for hospital readmission, applicable to all individuals who are hospitalized at least once in a given year. This model could be used to support decision-making for targeted intervention programs aimed at improving discharged patient outcomes, resulting in the reduction of unnecessary readmissions and substantial cost savings.

Potential Clients

The Hospital Readmission Reduction Program (HRRP) was designed as a Medicare value-based purchasing program that penalizes hospitals by decreasing payments to those that have disproportionately high readmissions. Results from this project would be particularly useful to healthcare organizations looking to reduce readmission rates, thereby improving patient outcomes and reducing associated costs and Medicare penalties.

Data

I used the dataset from National Health and Nutrition Examination and Survey (NHANES) 2015-2016. NHANES provides a cross-sectional survey of the health status of the US population each year. In 2015-2016, 15,327 persons were selected for NHANES from 30 different survey locations. Of those selected, 9,971 completed the interview and 9,544 were examined. The data is publicly available on the CDC's website, at https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015.

There are several datasets available for the 2015-2016 Survey. All datasets are linked by a common identifier variable which is tagged to each individual survey participant.

Data was obtained from four datasets in NHANES 2015-2016: Demographics[1], Laboratory[2], Examination[3], and Questionnaire[4]. Each dataset was further divided into data files. Descriptions of each feature within a Data File can be found in the 'Doc File' that accompanies each Data File.

Data Preparation/Wrangling

Reading in the Data

Each dataset was provided in a SAS transport file format. These files were imported directly into individual pandas dataframes, then the dataframes were merged together.

Data Exploration

Initial data exploration found 9971 rows and 479 columns. Because many of the features did not apply to this project, columns of interest were limited to include: number of hospital stays in the last year (from which the target variable would be based), age, gender, ethnicity, language spoken at home, diagnosis, age of onset of diagnosis, laboratory values, body mass index, body fat percentage, food security, and depression. After removing irrelevant features, 9971 rows and 127 columns were left.

Renaming Columns

Column names in the original dataset were coded and very difficult to interpret. Columns were therefore renamed to make it easier to work with and to understand their interactions.

Addressing Null Values

Null values within the dataset were next addressed. The null values were retained, since they made sense, as all features did not apply to all participants of interest. Data elements that were not coded as 'null' but were, in fact, null (for example, those coded as 9, 7, 999, 777, 99999 and 77777), were replaced with 'NaN'.

Adjusting Feature Types

Initially, all of the variables were of the float data type. Categorical variables were therefore coded as 'category' objects, for ease of use in visualization and later steps.

Target Variable and Feature Engineering

The final target variable and five new features were next created:

- 'readmission': binary target variable indicating positive readmission (2 or more hospital stays) or negative readmission (1 hospital stay)
- 'depr_score': sum of all depression scores
- 'cvd_age': age of onset of a participant's first CVD diagnosis)
- 'chf_new': compilation of all participants who were diagnosed with CHF in the last year
- 'cvd_new': compilation of all participants who were diagnosed with any form of cardiovascular disease in the last year

Removal of participants under age of 18

Since the focus of the study is on adults, all participants under the age of 18 were removed.

Outliers

After visualizing the different features, all outliers were retained, since none of them were noticeable errors.

After removing individuals under age 18, 5992 participants remained, which included participants who had no hospitalizations in the past year. Since the focus of this project was on individuals who were hospitalized at least once in the past year, the dataset was further reduced to include only participants who had at least 1 hospital stay. The final number of participants included in the investigation was 670. Of

these participants, 466 had a single hospital stay during the year of study and 204 had two or more hospital stays.

Exploratory Data Analysis and Initial Findings

CHF, CVD, and Other Medical Conditions

CHF is a common predictor of hospital readmissions, so the initial plan was to examine only readmissions related to new onset CHF. In order to investigate this, all adult NHANES participants with CHF (n=214) were identified (including both those who had and had not been hospitalized). Next, a new feature, new_chf was created, which identified only 30 participants who were diagnosed with CHF in the past year. Of new_chf, 20 had at least one hospital admission, with 55% of these reporting at least one readmission.

Since the number of participants with new_chf was so small, in an attempt to increase the sample size of patients with heart disease, the number of participants with cardiovascular disease (CVD) in general was investigated. There were 635 adult NHANES participants who had at least one type of CVD (including those who had and had not been hospitalized). A new diagnosis feature 'new_cvd' was created. This included new onset 'chf', 'chd', 'mi','angina',and 'cva'. Of the original 635 participants, there were 63 new_cvd. The number of total hospitalizations for new_cvd was 38, with 37% of these as readmissions. Therefore, new_cvd was not as strong an indicator of hospital readmissions when compared with new_chf.

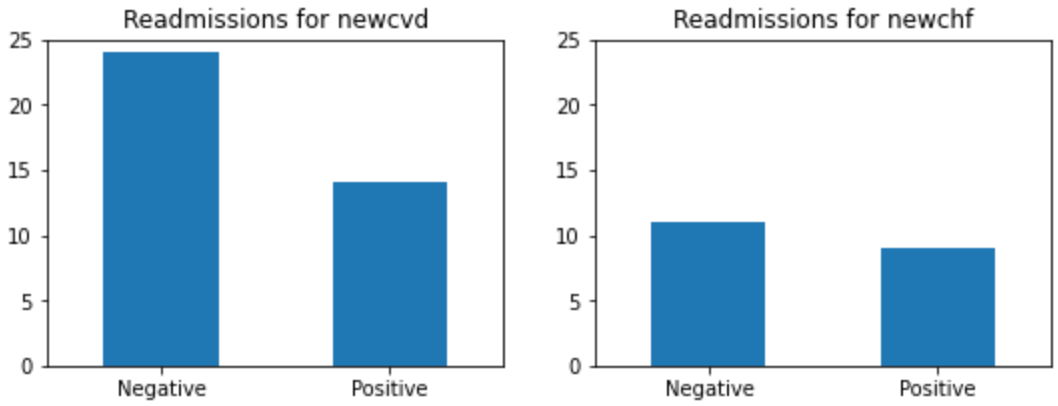

Figure 1. Readmissions for newcvd vs newchf.

Although the 'new_chf' proportion of readmissions was likely statistically significant, the sample size was quite small. Therefore, the investigation was not isolated to only new onset disease.

Since the NHANES dataset contains columns indicating many other chronic disease diagnoses that are related to hospital admissions, those diagnoses were investigated as well. Figure 2 indicates the proportion of hospital-stay participants that were readmitted by diagnosis.

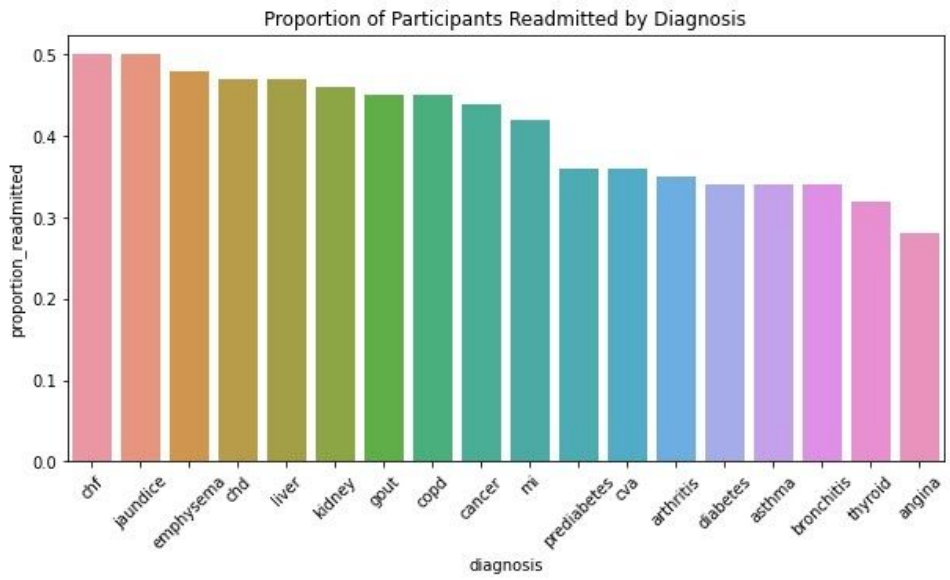

Figure 2. Proportion of readmissions by diagnosis.

To further examine the relationship of these medical conditions with readmission, Chi-square test statistics for individual diagnoses were computed. Diagnoses with significance at the $p<0.01$ level included ‘chf’, ‘kidney’, ‘copd’, ‘stroke’, ‘emphysema’, and ‘cancer’.

A matrix was also created to examine if there was a correlation between medical conditions. None of the absolute Pearson correlation coefficients were above 0.9, so all ‘medical condition’ features were kept in the final dataset.

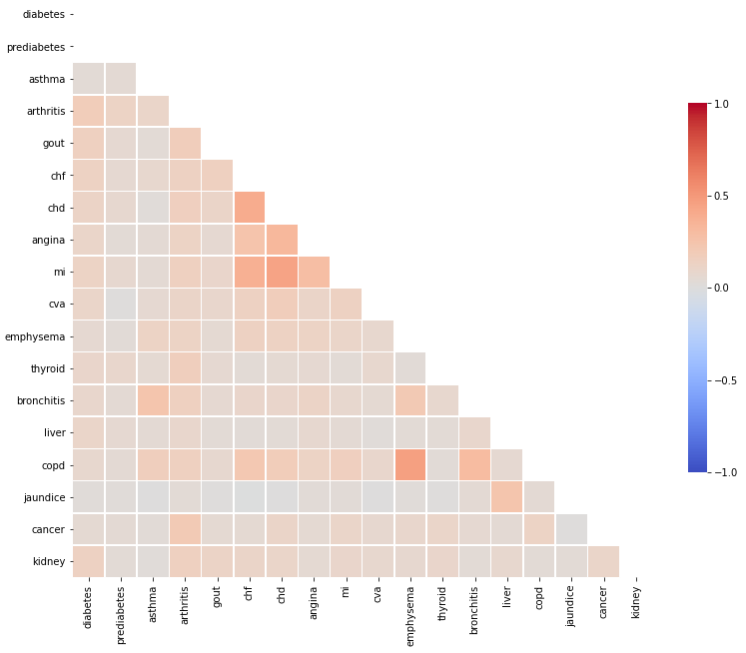

Figure 3. Correlation matrix of medical conditions.

Depression Scores

The relationship between depression score (depr_score) and readmission (Figure 3) was explored next.

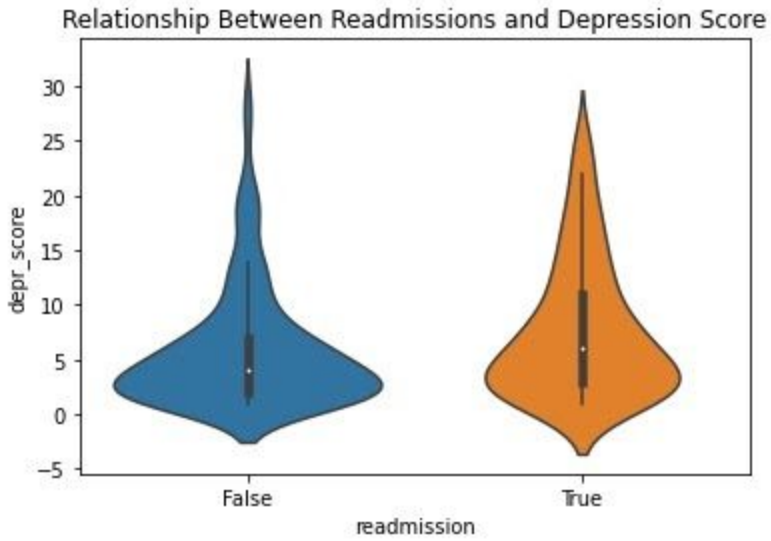

Figure 3. Relationship between depression score and readmission.

A depr_score above 4 indicates mild depression. The higher the score, greater severity of depression. For statistical analysis, participants were divided into two groups based on depression scores:

Group 1: depr_score < 5: little to no depression detected

Group 2: depr_score ≥ 5: mildly depressed or worse

A one-tailed, 2 proportion z-test was performed to detect the potential effect of depression score on readmission. Based on the two-proportion z-test, those with depr_score ≥ 5 had a higher proportion of hospital readmissions when compared with those with depr_score < 5 (p=0.03).

Ethnicity

The potential relationship between ethnicity and hospital readmissions (Figure 4) was also investigated.

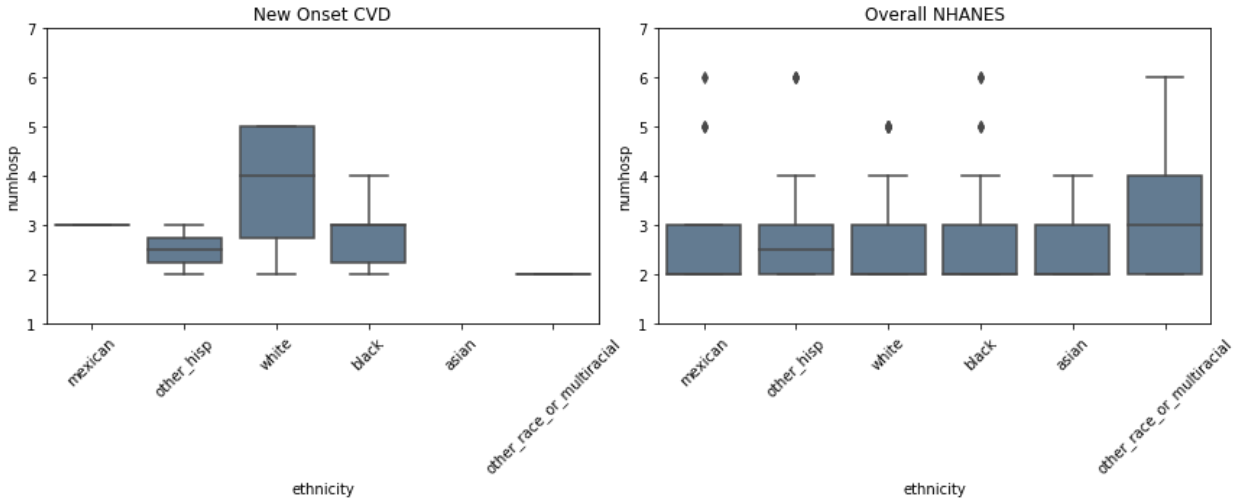

Figure 4. Relationship between ethnicity and number of readmissions.

For the overall nhanes group, a Chi square test implied that that ethnicity did not affect the number of readmissions (p=0.65). However, when new onset CVD was examined, ethnicity did appear to have an

effect on the proportion of readmissions (p=0.03), with participants who identified as White having a greater incidence..

Food Security

Food security refers to the availability of food and an individual's access to it. The relationship between hospital readmissions and food security through the feature 'foodsec' (Figure 5) was examined. Since there seemed to be little difference between positive and negative readmissions, statistical analysis on this feature was not computed.

Mean foodsec for positive readmissions: 2.4
Mean foodsec for negative readmissions: 2.6

Figure 6. Food security ratings for positive vs negative readmissions.

Laboratory Data

The relationship between laboratory values and positive and negative readmissions was explored using a 2-sided t-test. While the majority of laboratory values did not vary significantly between the two groups, mean BUN ($p<0.01$), creatinine ($p<0.05$), and osmolality ($p<0.01$) were affected. These lab values relate to kidney function and fluid balance. Other lab values that approached significance included hgb ($p=0.052$) - which is an indicator of anemia; and tchol($p=0.055$) - an indicator of heart health.

In addition to the t-tests, a correlation matrix and table of all lab values were created to determine which, if any, could be dropped to avoid multicollinearity (Figure 6). Features with an absolute correlation coefficient of 0.9 or greater were dropped.

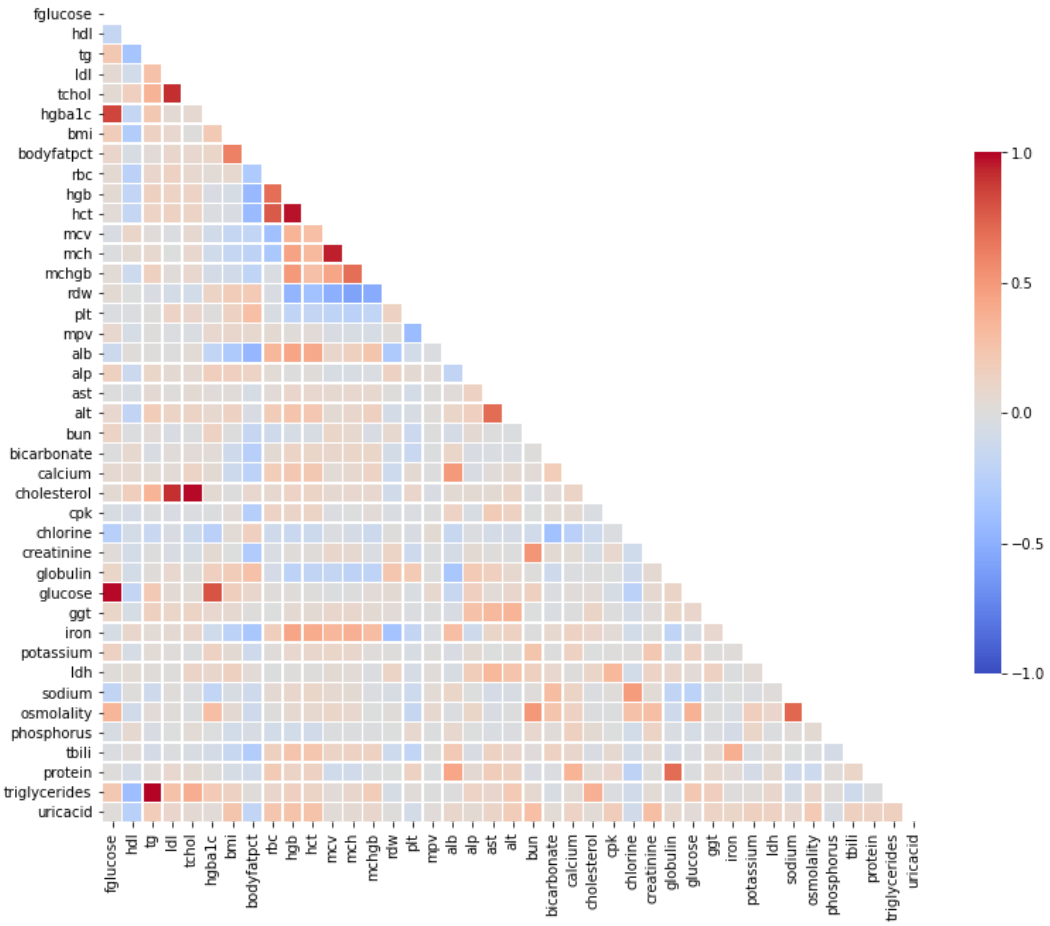

Figure 6. Correlation matrix of laboratory values.

Summary of Findings

- New onset congestive heart failure is a strong indicator of readmission, but the sample size of participants with this is quite small. Therefore, making predictions based on this feature may not be possible.
- Medical conditions, depression score, ethnicity, and laboratory values all appear to have an effect on readmission status. All features (minus laboratory values that are associated with one another) will be included in the machine learning model.

Prediction Using Machine Learning

A variety of supervised machine learning models to train and test the data were examined, since the target variable, hospital readmissions, is categorical and binary.

Preparing the Data

The following steps were taken to prepare the data for processing:

1. Dummy coded categorical variables, so that they could be used in the machine learning model
2. Removed redundant features and features with insufficient data
3. Replaced null values (NaN) with mean value of column
4. Split data into train and test sets
5. Normalized data: Since some algorithms require data to be scaled prior to use, and most of the time normalizing data will not have an effect on the outcome of those that do not require it, the training and test data were normalized.

The final dataset for use in the classification models included 88 attributes, including 1 target variable, 1 identification variable, and 86 explanatory features.

Model Selection

Several supervised machine learning models were implemented and compared, first with a dummy classifier, then with each other. The models were built using scikit-learn or XGBoost.

- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Support Vector Machine (SVM)
- Naive Bayes
- XGBoost (Extreme Gradient Boosting)

Prior to hyperparameter tuning, SVM, Random Forest, XGBoost, Logistic Regression, and KNN appeared to be the most promising models.

Hyperparameter Tuning and Model Evaluation

Next a range of parameters for each of the machine learning models were identified, and a gridsearch was performed to determine the best hyperparameters for each model based on the training data. Once the best parameters were selected, test data was run through each machine learning model. Classification reports and confusion matrices were completed to allow for comparison between the six models. The models were ranked in terms of performance based on the precision and recall rates of the different models. Finally, based on the ranking, the model that best fit the training data was chosen.

To explore the relationship between precision and recall of the different models, curves for ROC-AUC and Precision-Recall were plotted.

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is defined as follows:

Recall = TPR=TP/(TP+FN)

(TP: True Positive, FN: False Negative)

False Positive Rate (FPR) is defined as follows:

FPR=FP/(FP+TN)

(FP: False Positive, TN: True Negative)

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

AUC: Area Under the ROC Curve

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. Figure 6 displays the AUC-ROC Curve for the classifiers.

Precision/Recall Curve

Precision attempts to answer the question "What proportion of predicted positive readmissions was actually correct?" and is defined as TP/(TP+FP). **Recall,** as mentioned above, answers the question "What proportion of actual readmissions was identified correctly?" and is defined as TP/(TP+FN). The precision-recall curve (Figure 7) shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

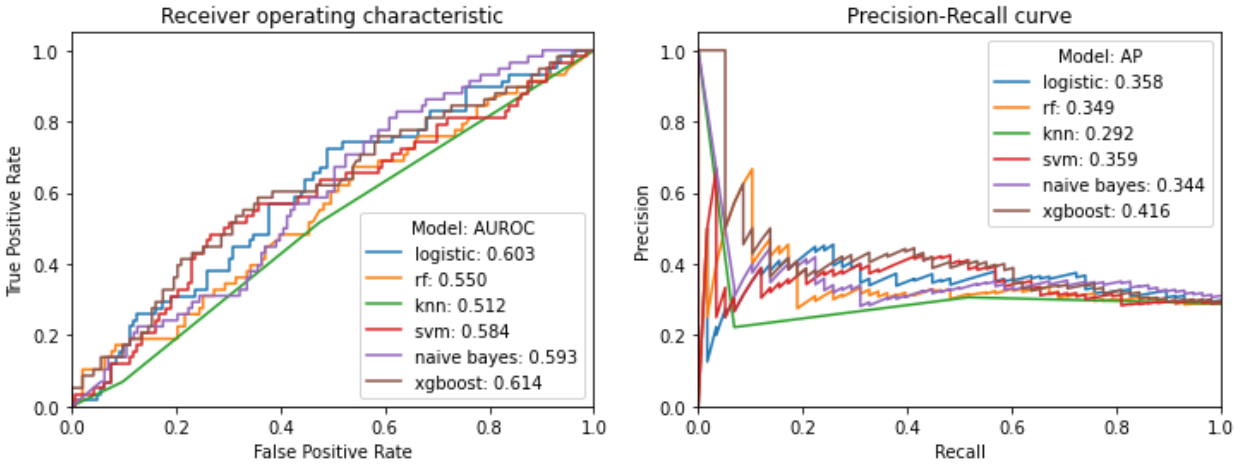

Figure 7. ROC-AUC and Precision Recall for each of the classifiers.

Average Precision (AP) summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. A balanced dataset would have an AP rate close to 0.5.

Figure 8 includes the ROC-AUC scores and AP for each classifier.

Model	ROC AUC Score	AP
Logistic Regression	0.603	0.358
Random Forest	0.550	0.349
K-Nearest Neighbors	0.512	0.292
SVM	0.584	0.359
Naive Bayes	0.593	0.344
XGBoost	0.614	0.416

Figure 8. ROC-AUC Scores and AP for each of the classifiers.

The best performing model is XGBoost, an ensemble model, which has AUROC at 0.614 and Precision-Recall at 0.416. Second best is the simple logistic regression model with AUROC at 0.603 and Precision-Recall of 0.358.

Feature Importance

Feature importance can play an important role in predictive modeling by providing insight into the data and the model, and may also, in this case, provide insight into potential targets for interventions to reduce readmission rates. Feature importance in both of the top performing models was examined.

In logistic regression, coefficients can be used to score feature importance. Since this is a classification problem with classes 0 (negative readmission) and 1(positive readmission), the coefficients are both positive and negative. Positive scores indicate a feature that predicts positive readmission, while negative scores indicate a feature that predicts negative readmission. Figure 9 displays the bar chart of feature importance for the logistic regression model. Top features for the model included 'liver', 'foodsec', 'chf', 'lymphocytes' (a cancer biomarker), 'cancer', and 'ldh' (a liver disease biomarker).

.

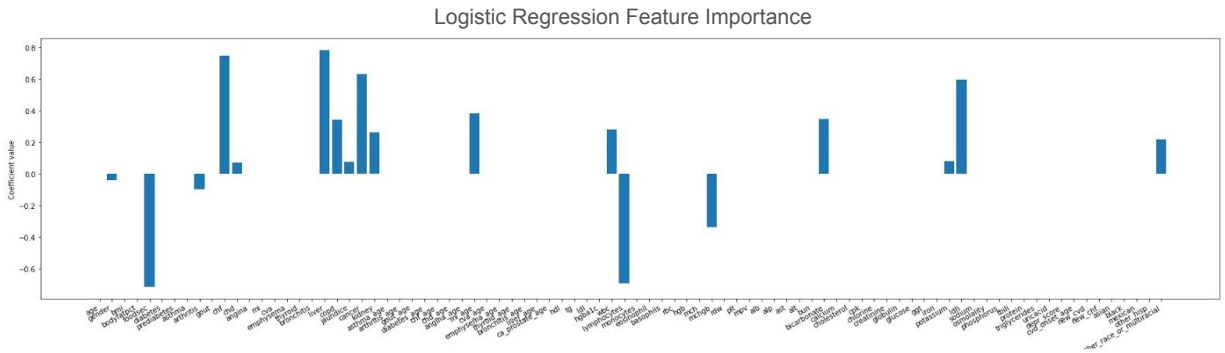

Figure 9. Feature importance for best logistic regression model.

Because XGBoost is much more complex compared to a linear model like logistic regression, its feature importance is more difficult to interpret. There are three options for measuring feature importance in XGBoost: Weight (number of times a feature is used to split the data across all trees), Cover (number of times a feature is used to split the data across all trees weighted by the number of training data points that go through those splits), and Gain (average training loss reduction gained when using a feature for splitting). All three measures were run to examine if there is a trend in feature importance over the three options (Figures 10a-c).

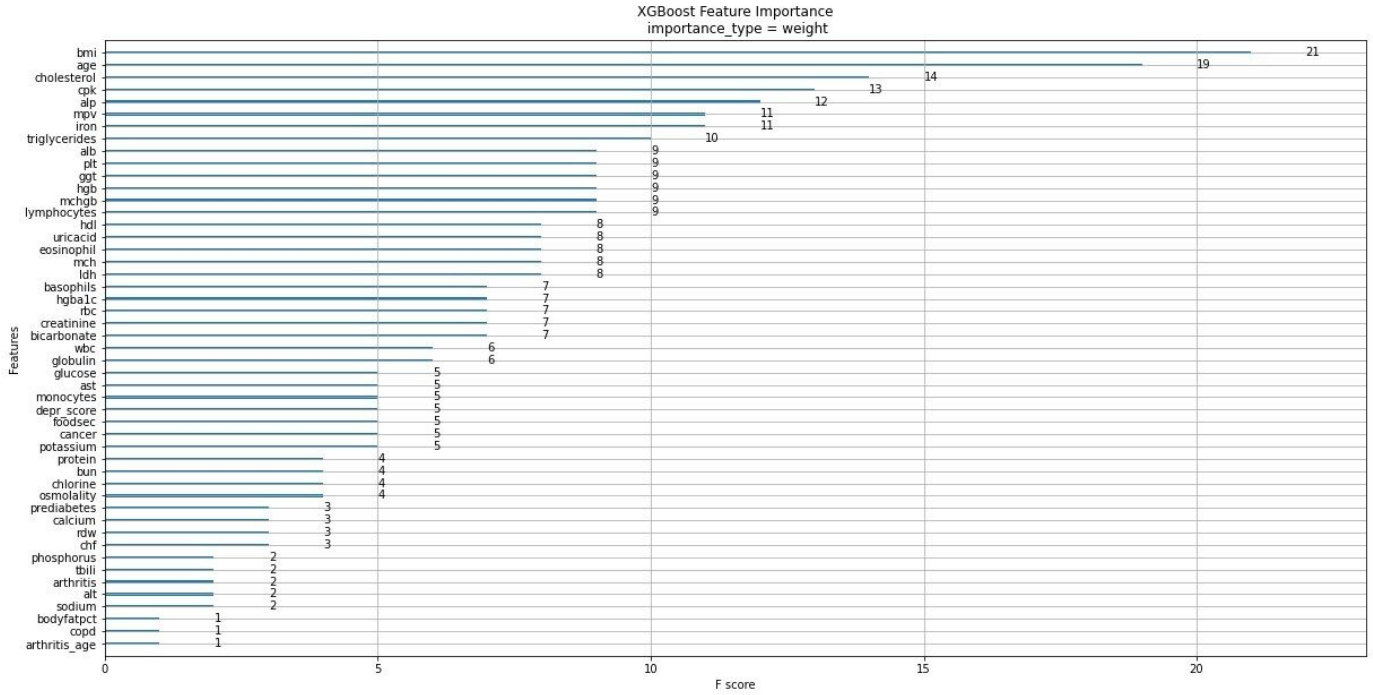

Figure 10a. XGBoost feature importance. Importance type = Weight.

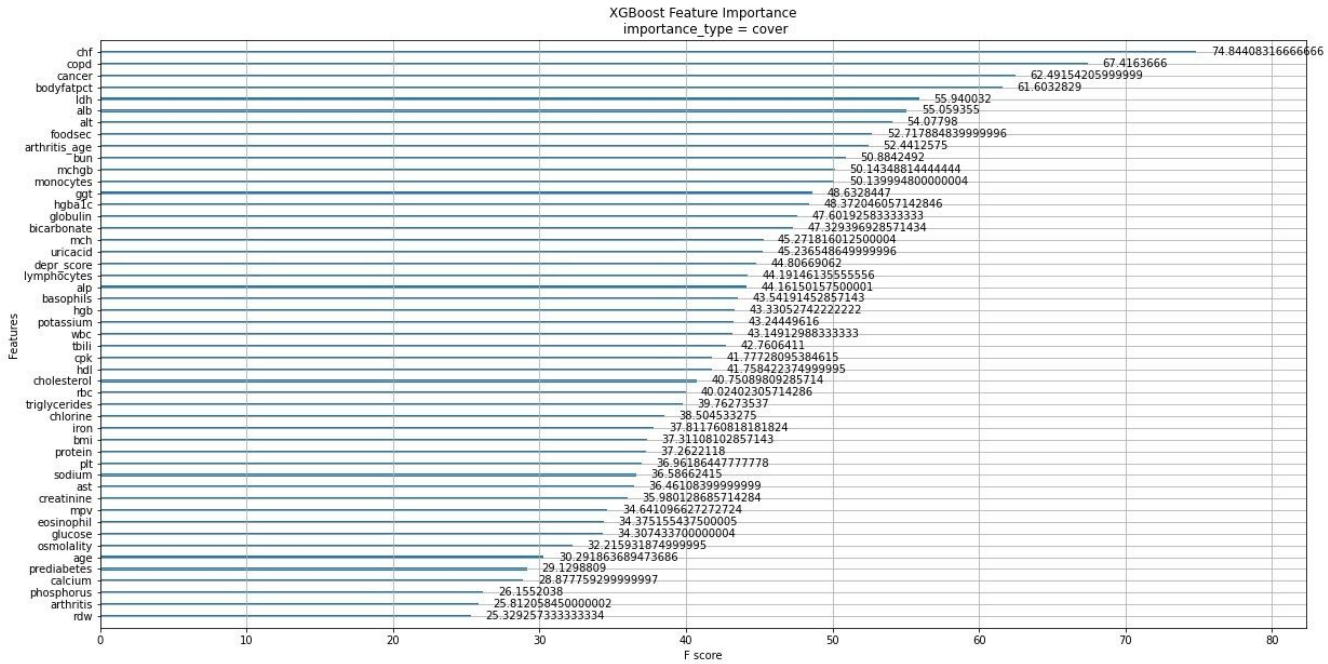

Figure 10b. XGBoost feature importance. Importance type = cover.

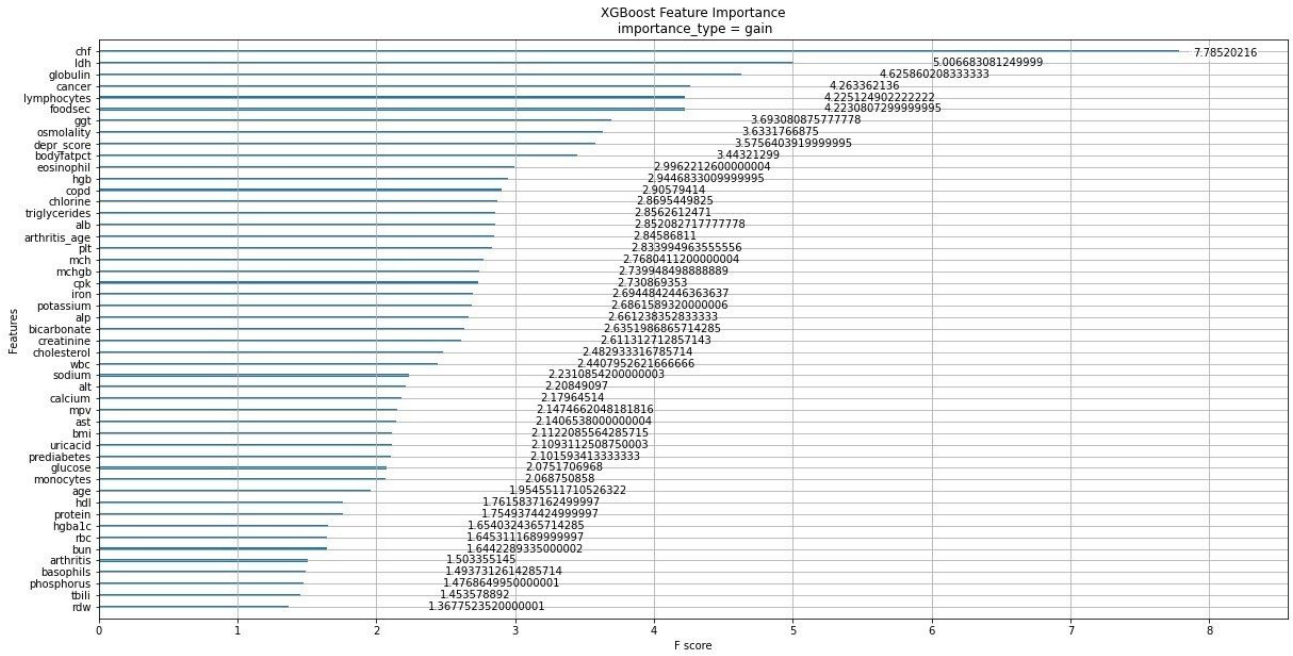

Figure 10c. XGBoost feature importance. Importance type = gain.

The feature importance orderings determined above are very different between the 3 XGBoost options, although there is some overlap in cover' and 'gain', which both include 'chf' and 'cancer' in

their top four features. For a better understanding of the XGBoost model, other feature importance methods would be required which are beyond the scope of the current project.

Confusion Matrix

A confusion matrix is a table layout that allows for visualization of the performance of an algorithm Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). Figure 11 displays the confusion matrix for the two best performing models.

Confusion Matrix of XGBoost (Best Performer)

	Positive predicted	Negative Predicted
Positive readmission	20 (TP)	38 (FN)
Negative readmission	21 (FP)	122 (TN)

Confusion Matrix of Logistic Regression (Second Best Performer):

	Positive predicted	Negative Predicted
Positive readmission	42 (TP)	32 (FN)
Negative readmission	26 (FP)	101 (TN)

Figure 11. Confusion matrix of the best performing models.

Data from the confusion matrix can be extrapolated to a real-world situation to make predictions on potential cost-savings contributed by the classifier, which chooses which patients on whom to focus intervention efforts..

Interpretation and Scalability

The preceding analysis determined that the best performing model for predicting readmissions was XGBoost. Using the results from the model's confusion matrix, the outcomes of an imaginary, 'average' community hospital in the US were projected. Based on data from the American Hospital Association[5], the average hospital in the US has approximately 6589 admissions per year. Based on the NHANES sample (test data), proportion of readmissions : total admissions was 0.28. Therefore, it can be predicted that of the 6589 admissions per year for the average hospital, 1901 are readmissions. Applying the XGBoost model's confusion matrix to this hypothetical example:

	Positive Predicted	Negative Predicted	Total
Positive readmission	656	1245	1901
Negative readmission	688	3999	4688

Total	1344	5245	6589

Assumptions:

- Average cost of readmission: $14,400[6]
- Cost of intervention during first 1-2 weeks post-discharge (home visits from multidisciplinary team for 2 weeks post-discharge): $1000/patient
- If intervention is used, approximately 27% of readmissions will be prevented[7]

The cost savings analysis are based on three possible actions a health system could take:
Option 1: Do nothing - no intervention; costs from readmissions remain as predicted based on status quo.
Option 2: Intervention for all patients (no machine learning model used).
Option 3: Intervention for only patients who are Predicted Positive by the ML model.

Cost Savings from Applying ML Model*

	Option 1	Option 2	Option 3
Cost of Readmissions	$34,061,590	$24,864,961	$17,551,281
Intervention Cost	$0	$4,223,907	$861,593
Total cost	$34,061,590	$29,088,868	$18,412,874
Savings	$0	$4,972,722	$15,648,716
Percent Savings	0%	15%	46%

By employing the ML model to patient data, with minimal upfront cost, the average hospital in the US could save $15.6 Million on average from reduced readmissions, with a 46% savings on readmissions cost. With further refinement of the model, the actual savings would be even greater.

Limitations

The dataset did not include information regarding length of time between initial hospital stays and readmissions. If this information was available, a stronger model could be built to focus on prediction of 30-day hospital readmissions, to more specifically target interventions to these. In addition, the sample size was relatively small, so with a larger sample, a stronger model with more accurate predictions could be made.

*Savings calculations can be found here.

Conclusion

- Data analysis highlights many potential demographic, social-emotional and biological factors that can contribute to hospital readmissions.
- Using these types of factors, a better-than-random classification model can be built to predict hospital readmissions in individuals from the general US population.
- A predictive model such as this could support intervention programs that would result in savings and prevention of Medicare penalties for healthcare organizations.
- More extensive validation and refinement of the model and more specific hospital-derived data would likely result in better targeted interventions and even more substantial savings.

References

1. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2015
2. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2015
3. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&CycleBeginYear=2015
4. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2015
5. https://www.aha.org/statistics/fast-facts-us-hospitals
6. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb248-Hospital-Readmissions-2010-2016.jsp#:~:text=In%202016%2C%20the%20average%20readmission,at%20index%20admission%20was%20%2414%2C400
7. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2498846