# Predicting Hospital Readmissions with Population-Based Data

**By Cara Shrivastava**
**September 2020**

Each year, readmissions within 30 days of hospitalization cost the **average community hospital** in the US approximately **$17 million**.*

Over the course of a year, rehospitalization costs may reach **over $34 million**.*

* Numbers based on data from:
- Fast Facts on US Hospitals
- Characteristics of 30-Day All-Cause Hospital Readmissions, 2010-2016

Research has found that over 25% of readmissions could be prevented[1] through targeted interventions.

[1] https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2498846

# Why this matters

High hospital readmission rates are associated with:

- Poor patient outcomes
- High costs for healthcare systems and private and government-sponsored insurance programs such as Medicare.

# Medicare's HRRP

The Hospital Readmission Reduction Program (HRRP) was designed as a Medicare value-based purchasing program.

It **penalizes hospitals** by **decreasing payments** to those that have disproportionately high readmissions within 30 days of a patient's discharge.

# The Client

Results from this project would be particularly beneficial to a **healthcare organization** looking to reduce its readmission rate.

# The Solution

Reduce readmission risk
as much as possible.

# The Solution

By using data they already have, hospitals can screen patients prior to discharge, selecting those at ‘high risk’ for readmission.

- ★ Patients tagged as ‘high risk’ will receive evidence-based interventions before discharge and in the two weeks following discharge.

# The Process


Data collection
Data Preparation
Exploratory data analysis (EDA) and initial findings
Prediction using machine learning

# Step 1

## Data Collection

# Data Collection

Data was obtained from four datasets from the National Health and Nutrition Examination and Survey (NHANES) 2015-2016:

- Demographics
- Laboratory
- Examination
- Questionnaire

NHANES provides a cross-sectional survey of the health status of the US population each year.

Data can be found on the CDC website.

# The Data

This project was based on readily-available population-based data. However, the analysis and models are highly scalable to electronic medical records data that health systems routinely collect.

# Step 2

## Data Preparation

# Data Preparation

## Reading in the data

Imported the SAS files into pandas dataframes, then merged the dataframes together.

## Initial exploration

Initial dataset included 9971 rows and 479 columns. Columns were limited to those features of interest.

## Renaming columns

Column names in original dataset were coded and hard to interpret. They were renamed.

## Nulls and Outliers

Null values were retained, since they made sense, and also outliers were kept since none were noticeable errors.

## Feature Types

Initially all columns were of the 'float' data type. Categorical variables were coded to 'category' objects.

## Feature Engineering

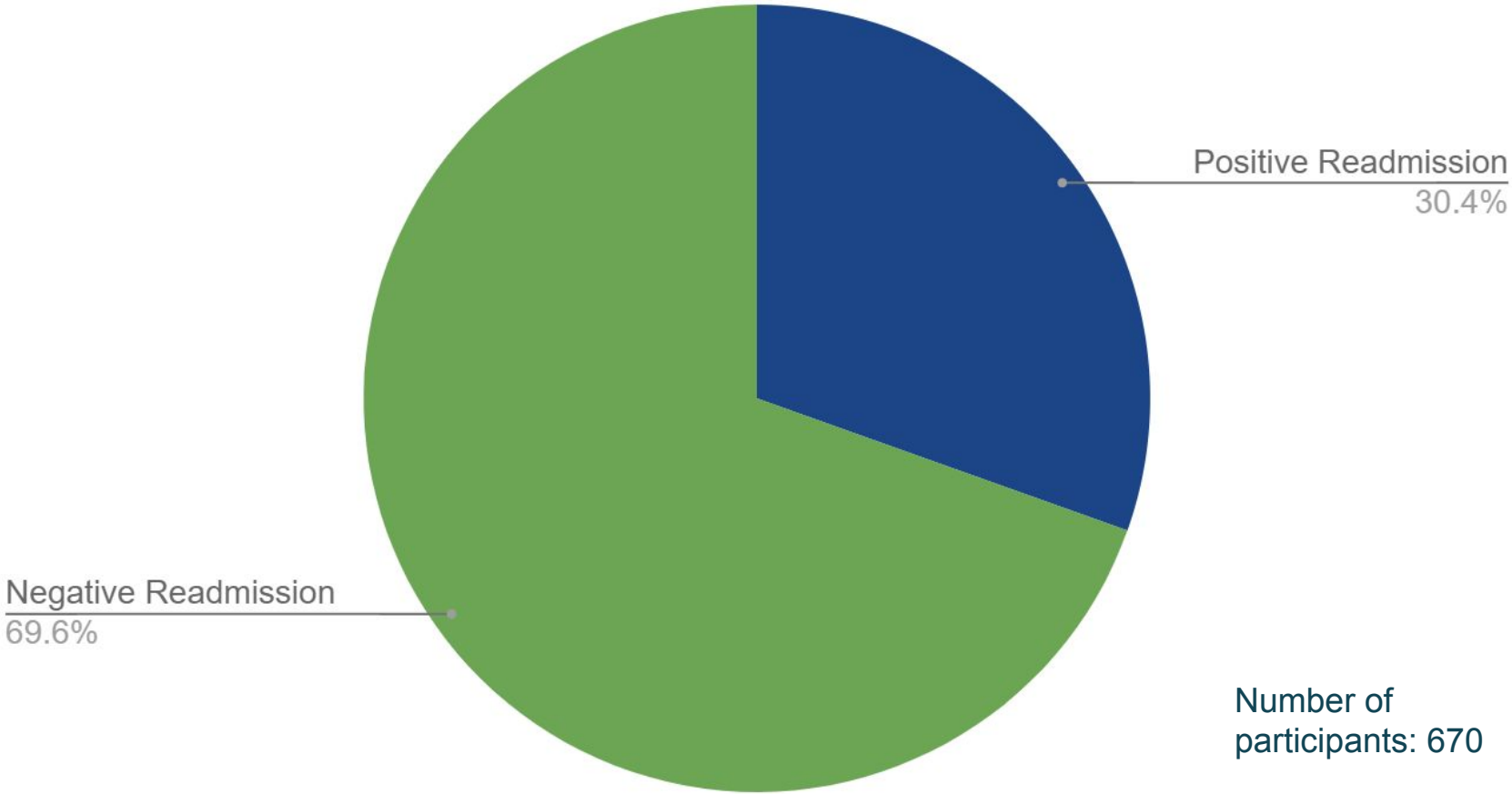
The final target variable, 'readmissions' and five new features were created based on available data.

# The Final Dataset

There were **670 individuals** in the final dataset after removing those under age 18 and participants who had no hospitalizations in the last year.

These participants were classified into two groups:

| Positive readmissions | Negative readmissions |
| --- | --- |
| ≥ 2 hospital stays in the last year | 1 hospital stay in the last year |

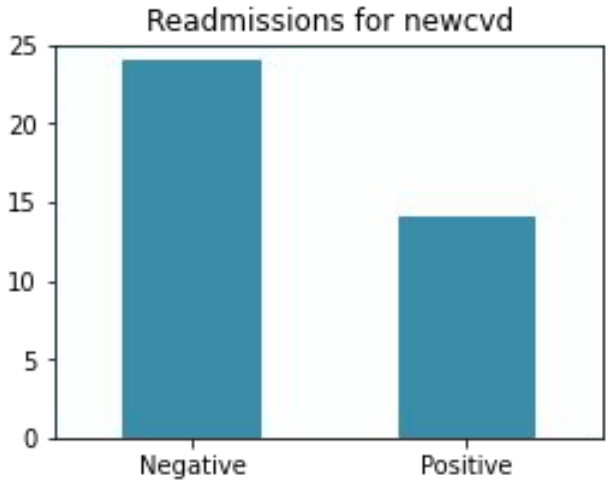



Number of participants: 670

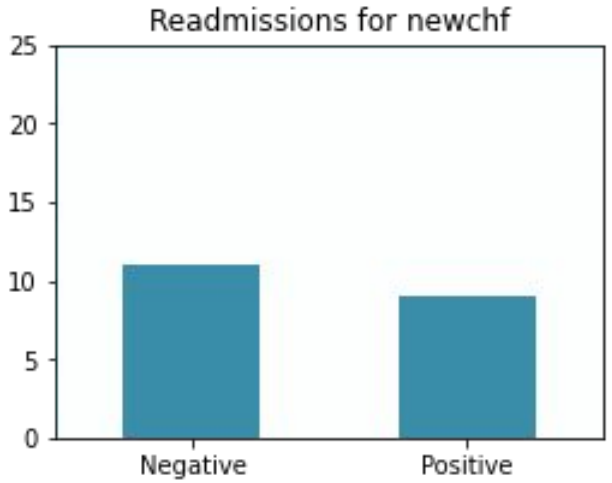
# Step 3

## EDA & Initial Findings

# Readmissions due to new-onset cardiovascular disease (CVD) and congestive heart failure (CHF)



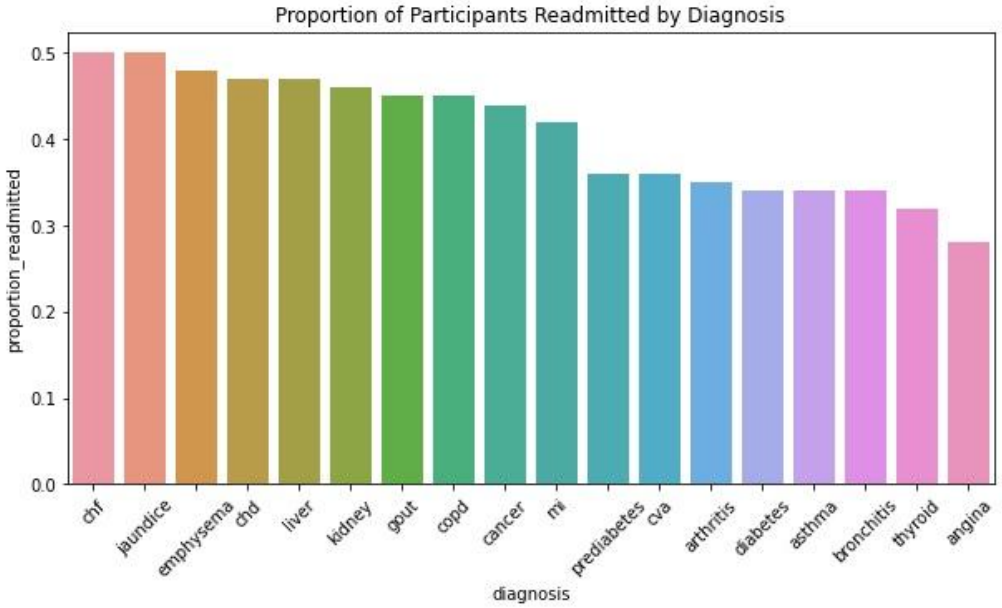

37% of new CVD patients were readmitted.




55% of new CHF patients were readmitted.

Since the number of participants with new onset CHF was small, statistical tests and models were not modified to include only new onset CHF participants.

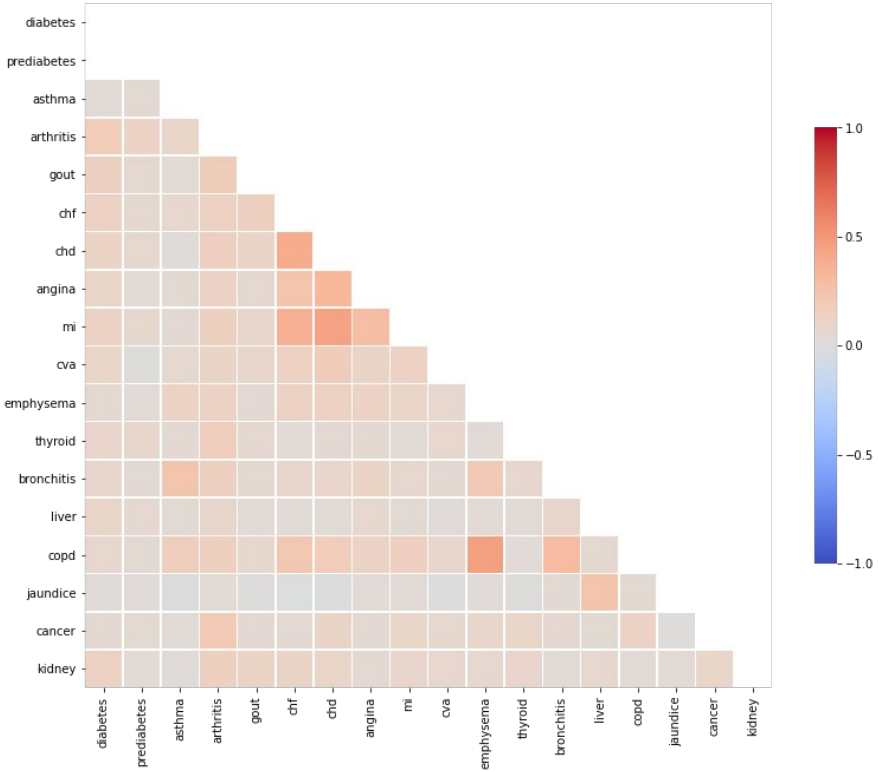
# Proportion of readmissions by diagnosis




Based on results of a Chi-square test, the following diagnoses had statistically significant readmissions:

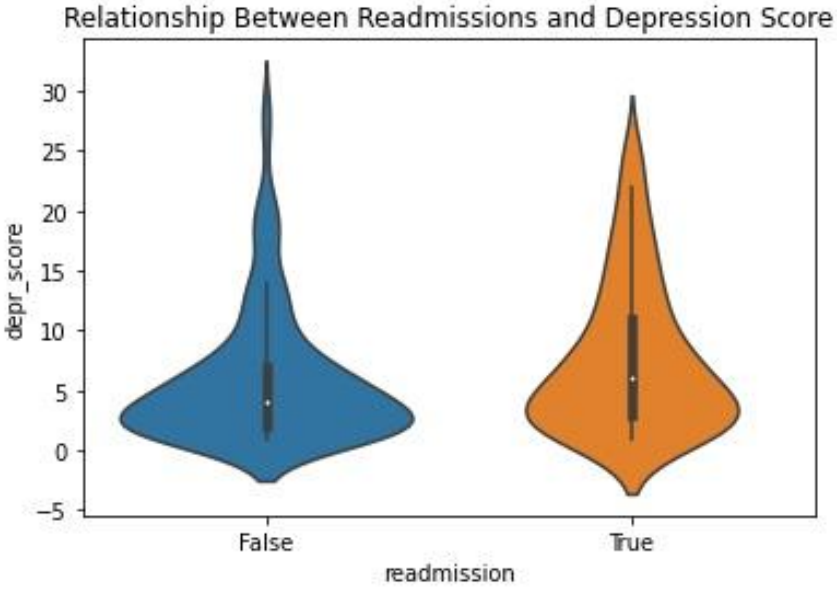
- CHF
- Kidney disease
- COPD
- Stroke
- Emphysema
- Cancer

# Correlation between medical conditions




None of the Pearson correlation coefficients were above 0.9, so all medical condition features were kept in the final dataset.

# Depression score and readmissions




A depr_score >4 indicates mild depression. The higher the score, greater severity of depression.

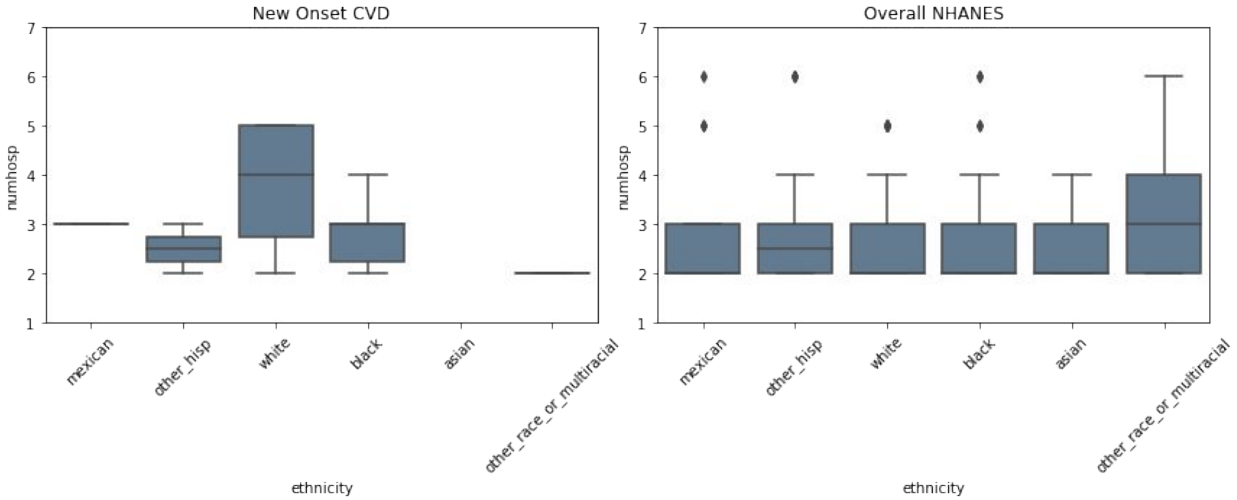
For statistical analysis, participants were divided into two groups:

- Group 1 (little to no depression):
  - depr_score < 5
- Group 2 (mild depression or worse:
  - depr_score ≥ 5

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).

**Therefore, depression appears to be a good indicator for readmission.**

# Ethnicity and readmissions




For new onset CVD, ethnicity appeared to have an effect on proportion of readmissions (p=0.03), with participants who identified as White having a greater incidence.

**For the overall group, a Chi-square test implied that ethnicity did not affect number of readmissions (p=0.65).**

# Food security and readmissions


Mean foodsec for positive readmissions: 2.4

Mean foodsec for negative readmissions: 2.6

Based on the means, there appeared to be little to no relationship between food security and readmissions

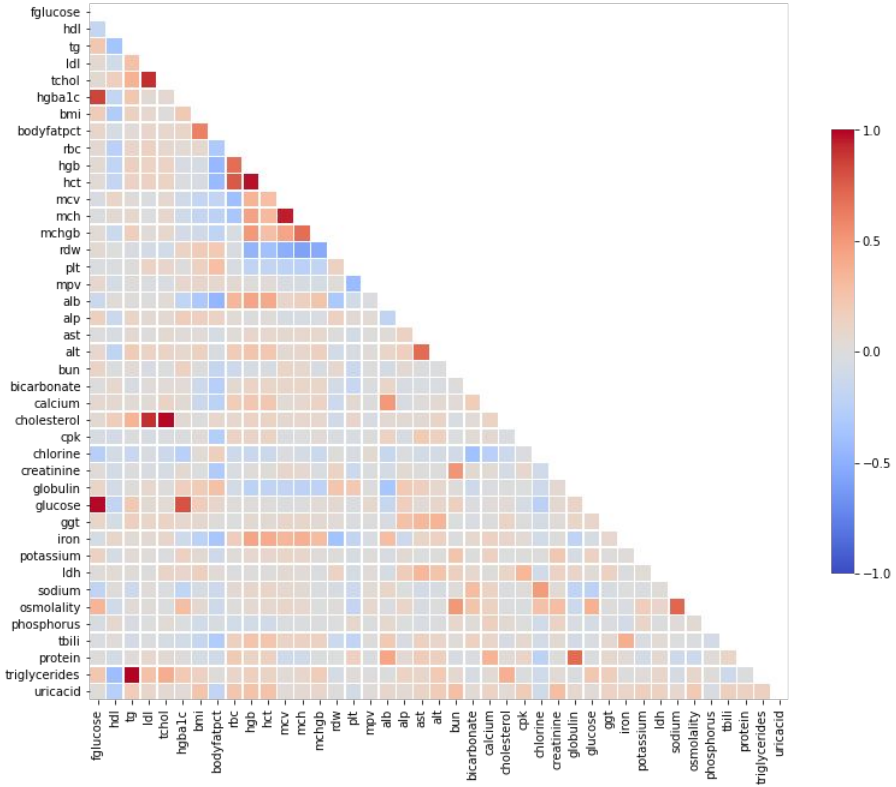
# Laboratory data and readmissions

The relationship between lab values and readmissions was explored using a 2-sided t-test.

**The following were significant or approached significance between + and – readmissions:**

- BUN ($p<0.01$)
- creatinine ($p<0.05$)
- osmolality ($p<0.01$)
- hemoglobin ($p=0.052$)
- Total cholesterol ($p=0.055$)

# Correlation between laboratory values




Laboratory values with an absolute correlation coefficient of 0.9 or greater were dropped to avoid multicollinearity.

# Summary of Findings

- New onset CHF is a strong indicator of readmission, but the sample size of participants is quite small. Predictions will not be able to be made based on a subset of new CHF patients.
- 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.

# Step 4

## Machine Learning

# Data Preparation for Machine Learning

To further prepare the data for predictive modeling, the following were completed:

1. Dummy coding of categorical variables
2. Removal of redundant features and features with insufficient data
3. Replacement of null values with mean value of column
4. Splitting of data into train and test sets
5. Normalization of data (necessary for some models)

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

# Model Selection

These supervised learning models were implemented and compared:

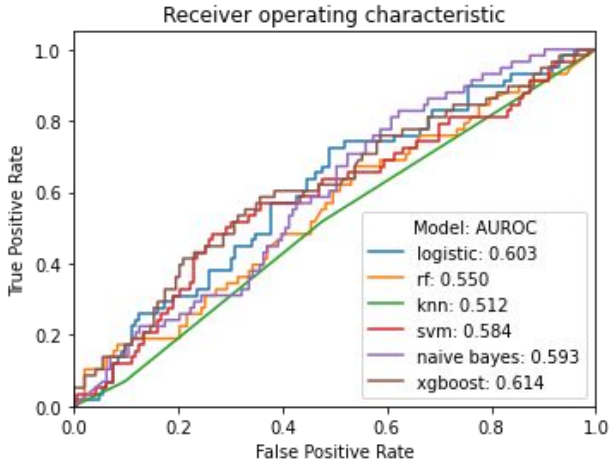
- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Support Vector Machine (SVM)
- Naive Bayes
- XGBoost (Extreme Gradient Boosting)

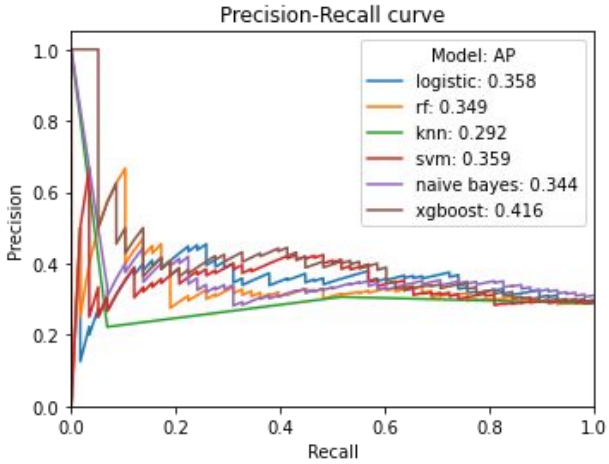
# Model Evaluation

Steps taken to evaluate model performance:

1. Identified range of hyperparameters for each model, and performed a gridsearch to determine the best hyperparameters for each
2. Test data was run through each machine learning model using the best hyperparameters
3. Classification reports and confusion matrices were completed to allow for model comparison.
4. The models were ranked in terms of performance based on their precision and recall rates.
5. Based on the ranking, the model that best fit the training data was chosen.

# Evaluation of the models




To compare the models' performance, an ROC curve was plotted. **Classifiers that give curves closer to the top-left corner indicate a better performance.**

The area under the ROC curve, **AUROC**, is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. **The higher the AUROC, in general, the better the model.**

The precision-recall curve shows the tradeoff between precision and recall for different thresholds. **Average precision (AP) summarizes a precision recall curve. A balanced dataset would have AP rate close to 0.5.**

# Evaluation of the models

| 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 |

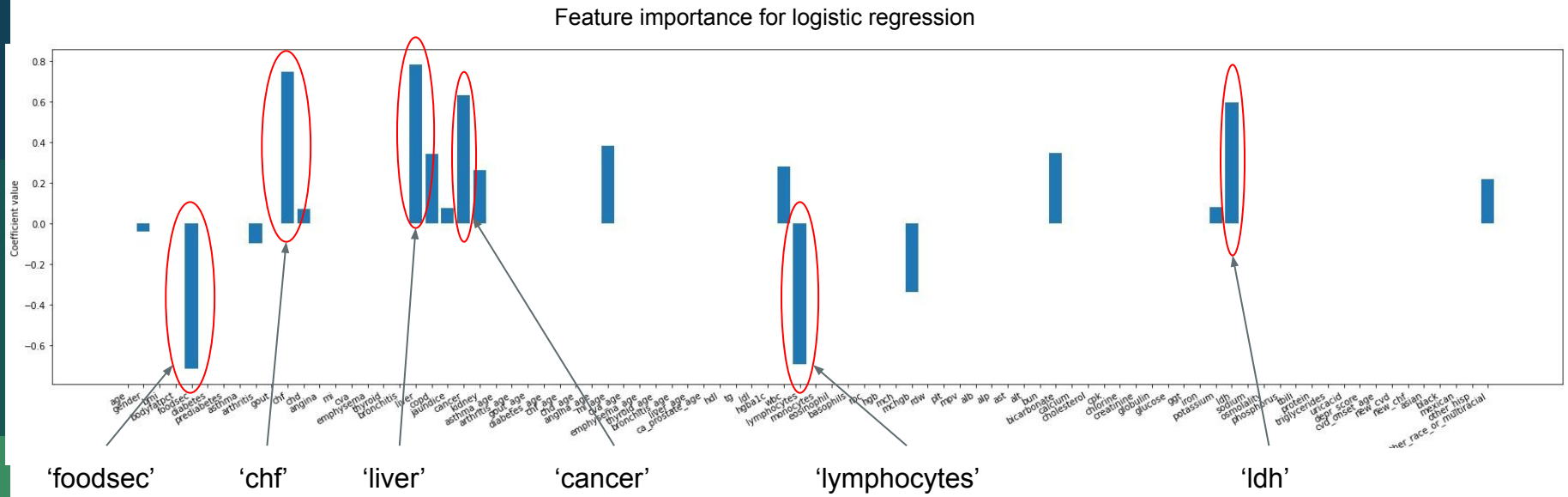
Best performing model is XGBoost, an ensemble model

Second best is the simple logistic regression

# Feature importance

Feature importance can play an important role in predictive modeling, providing insight into the data and the model.

Feature importance was examined in the top performing models.




Top features for logistic regression included 'liver', 'foodsec', 'chf', 'lymphocytes' (a cancer biomarker), 'cancer', and 'ldh' (a liver disease biomarker).

# Feature importance for XGBoost

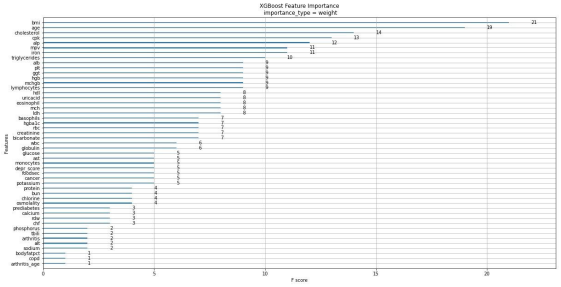
XGBoost is much more complex compared to a linear model like logistic regression.

Therefore, its feature importance is more difficult to interpret.

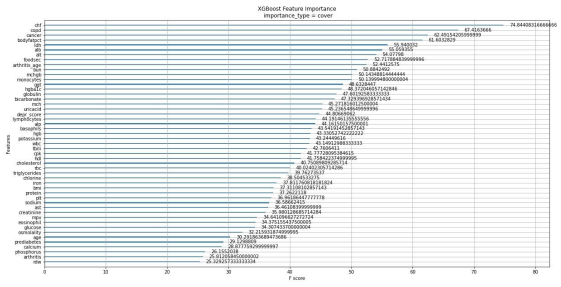
There are three basic options for measuring feature importance in XGBoost:

- Weight
- Cover
- Gain

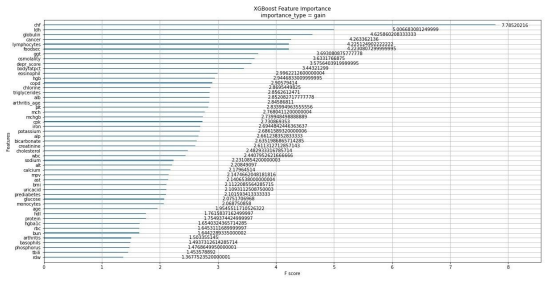
# Feature importance for XGBoost




Weight




Cover




Gain

While there is a little bit of  overlap in the ‘cover’ and ‘gain’ feature importances --both include ‘chf’ and ‘cancer’ in their top four features -- overall, there is a lot of variation between results of the three XGBoost options.

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

# Step 5

## Interpretation and Scalability

# Average US hospital readmissions

Using the results from the XGBoost model's confusion matrix, outcomes of an imaginary, 'average' community hospital in the US were projected.

The average hospital[5] in the US has:

- 6589 admissions/year.
- Proportion of readmissions (based on the data): 0.28.
- 1901 of the admissions are readmissions

| | Positive Predicted | Negative Predicted | Total |
|---|---|---|---|
| Positive readmission | 656 | 1245 | 1901 |
| Negative readmission | 688 | 3999 | 4688 |
| Total | 1344 | 5245 | 6589 |

[5] https://www.aha.org/statistics/fast-facts-us-hospitals

# Assumptions

- Average cost of readmission: $14,400
- 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

# Cost Savings Analysis

The cost savings analysis was 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)
- Option 3: Intervention for only patients who are Predicted Positive by the ML model

# Cost Savings Analysis

| | 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% |

# $15.6 Million in savings

By employing the ML model to patient data, with minimal upfront cost to the hospital system,  the average hospital in the US could save **nearly $15.6 Million (46% savings)**. With further refinement of the model, the savings would be even greater.

# Conclusion

There are many demographic, social-emotional and biological factors that can contribute to hospital readmissions.

Using these factors, a better-than-random classification model can be built to predict hospital readmissions.

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.

# THANKS!

## Any questions?

Contact me at:
carashri@hotmail.com

Python code and savings calculations can be found at:

https://github.com/carashri/Predicting-Hospital-Readmissions/tree/master/Code

# Credits

Special thanks to all those who helped!

- Tommy Blanchard (my Springboard mentor)
- Springboard