Predicting Diabetes Patient Hospital Readmission

Pavel Zimin, PhD
Mentor: Ramkumar Hariharan, PhD

Problem Statement

- Hospital readmission is a highly preventable cause for high healthcare costs
- The ability to predict hospital readmission will help prioritize patients that will benefit from hospital discharge follow up programs

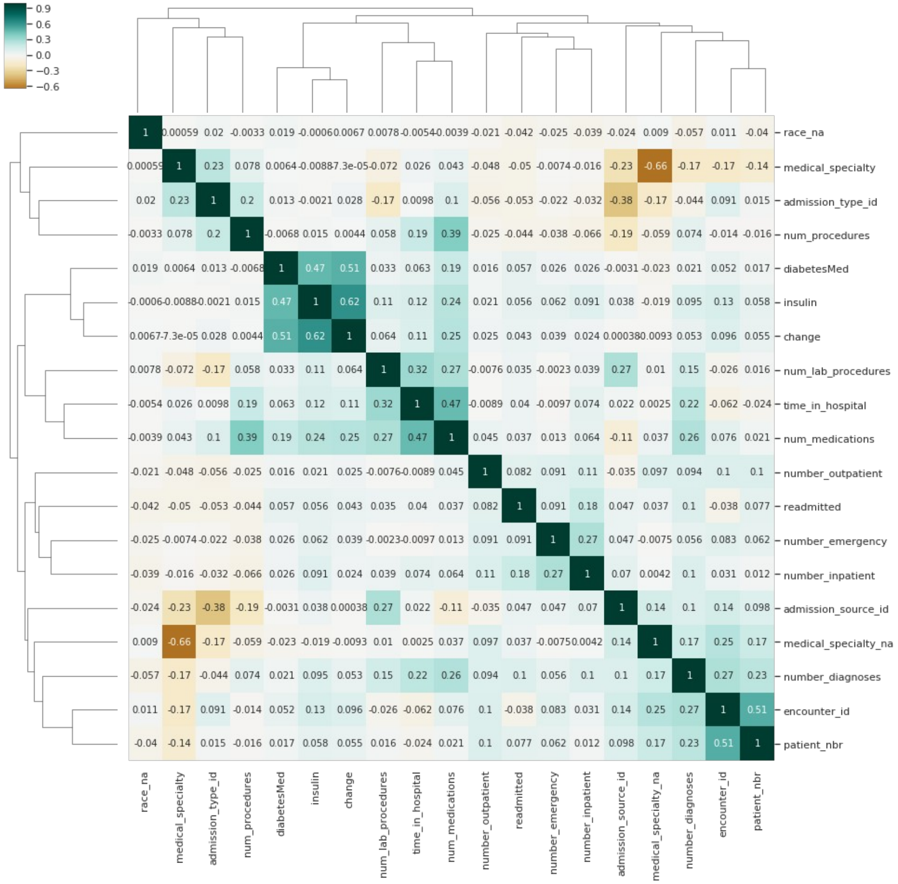
Business use cases:

- The outcome of this analysis will be helpful to the hospital healthcare teams with prioritizing patient support program
- This analysis will benefit patients who will receive improved health care, decreased chances of readmission while incurring smaller cost

Data Wrangling Steps

- Hospital readmission data were downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008)
- The dataset contains 101,766 observations of unique hospital encounters with 50 variables: 13 columns of integer type, 37 columns of object type
- Medical diagnosis codes with their hierarchical groupings were downloaded from a GitHub repository (https://github.com/sirrice/icd9.git) and merged with the readmission data set
- For each variable with missing values a separate column was created with values indicating the missing values
- Each categorical variable was encoded with integer values, the code was saved in a dictionary
- No outliers were removed

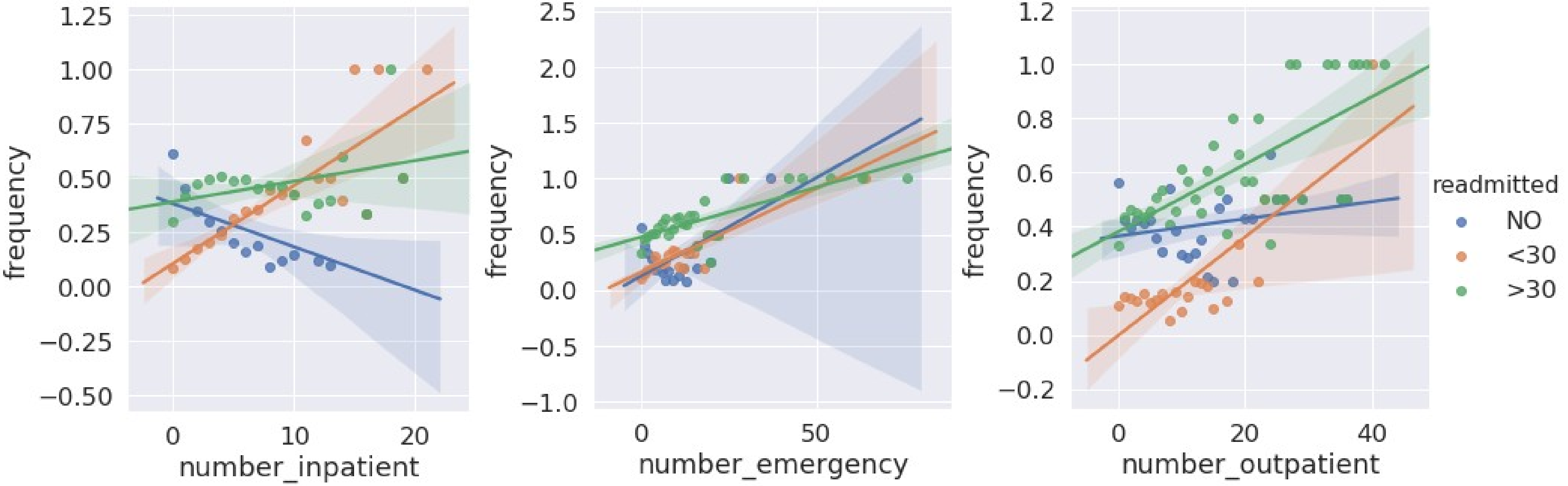
Correlation Clustermap



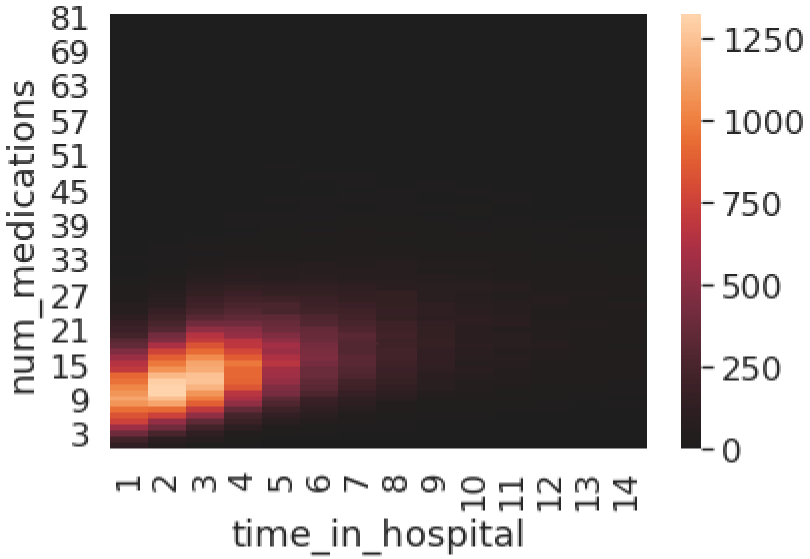
- moderate to low correlation between the readmitted variable and other variables
- variables showing largest correlations are:

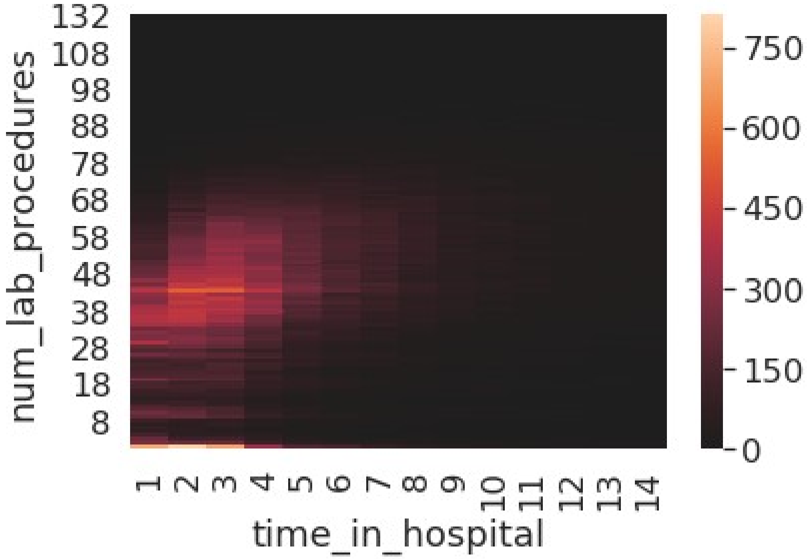
1) number_inpatient (the number of inpatient visits in the year preceding the encounter)
2) number_emergency (the number of emergency visits in the year preceding the encounter)
3) number_outpatient (the number of outpatient visits in the year preceding the encounter)

Frequencies of Selected Variables

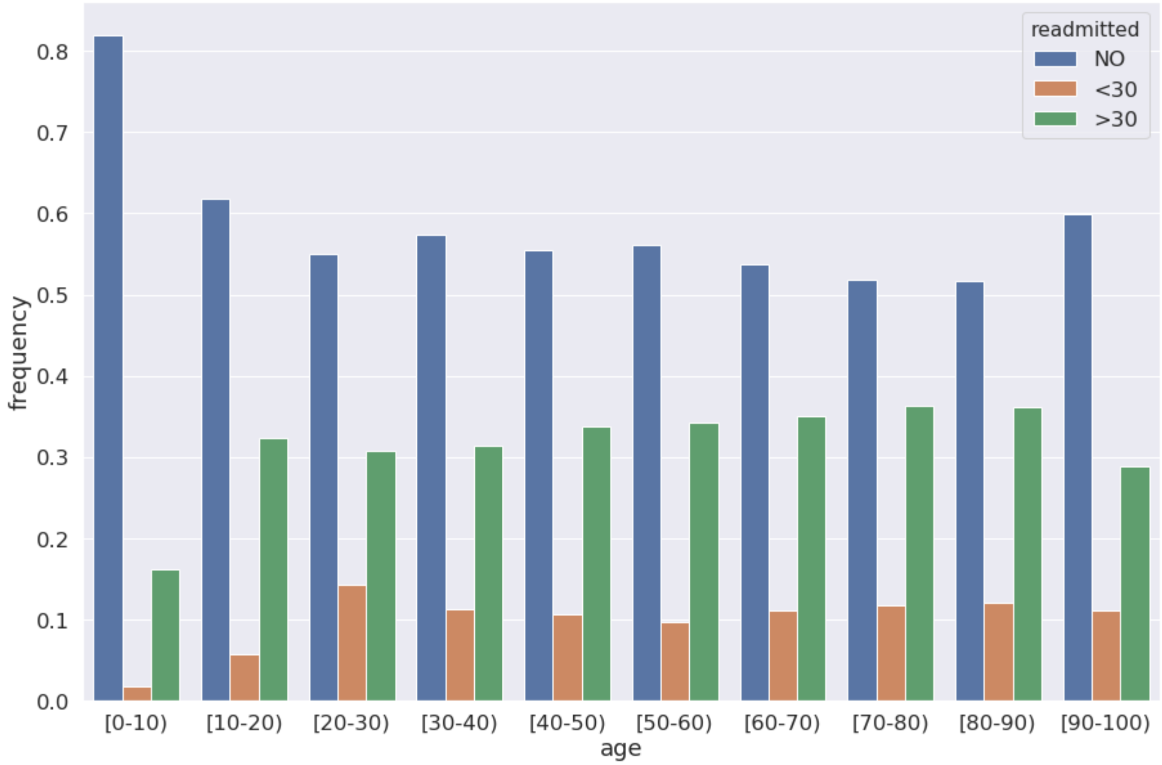


Sanity Check of the Data Set

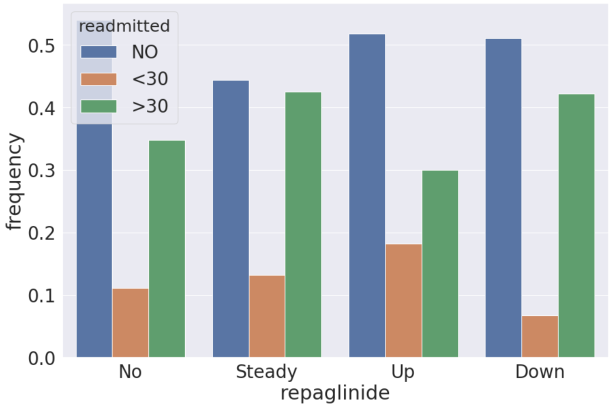


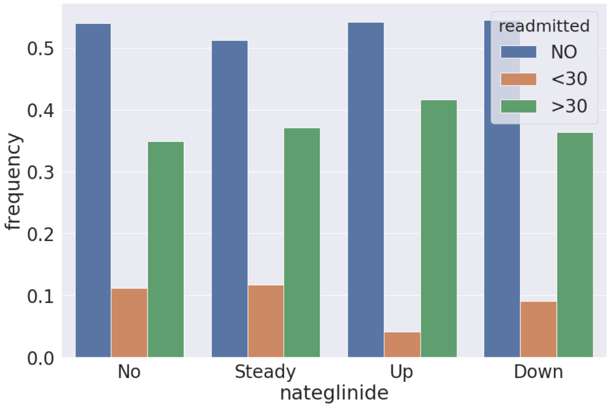


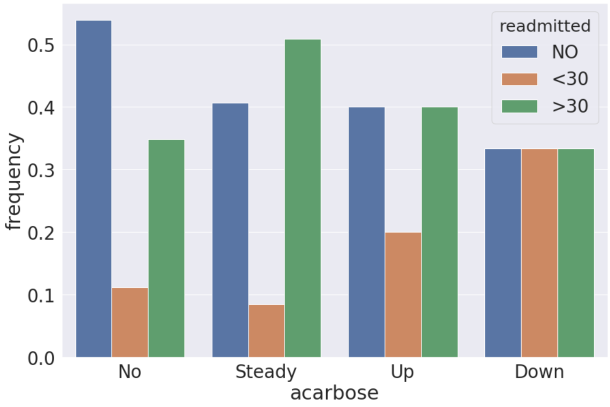
Age Group Frequencies



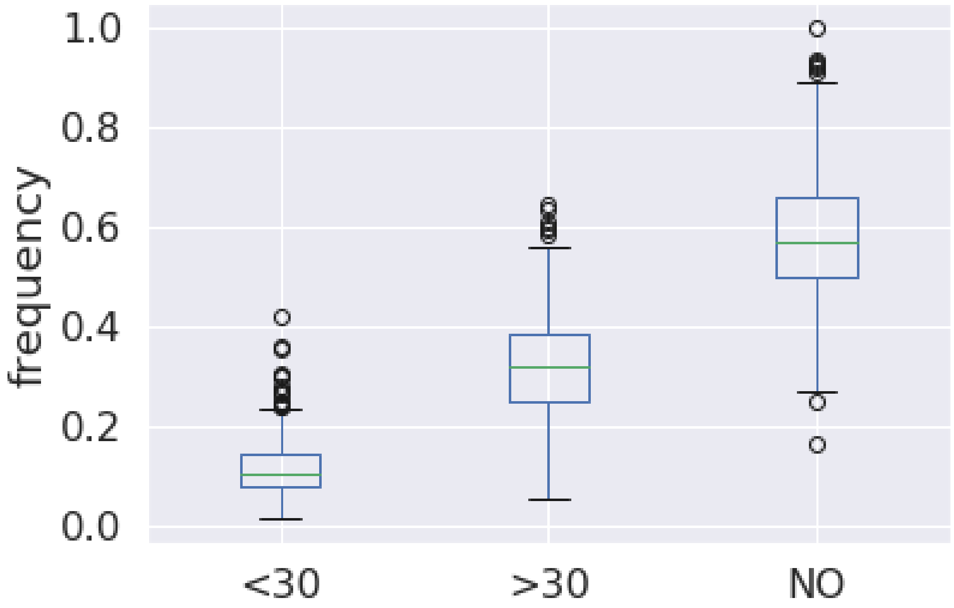
Frequencies of medication dosage changes







Primary Diagnosis Frequency Distributions



The top primary diagnoses in the group with readmission within 30 days:

1) encounter for other and unspecified procedures and aftercare,
2) diabetes with renal manifestations,
3) peritonitis and retroperitoneal infections.

EDA Conclusions

- There is moderate to low correlation between the readmitted variable and other variables.
- The variables showing the largest correlations are: number_inpatient, number_emergency, and number_outpatient.
- The most dramatic changes in the frequencies of medication changes were for the following medications for treating diabetes: repaglinide, nateglinide, and acarbose.
- Distribution of the primary diagnoses shows that for some primary diagnoses the frequency of readmission within 30 days is much higher than the median frequency for that group.
- The top primary diagnoses in the group with readmission within 30 days are

1) encounter for other and unspecified procedures and aftercare
2) diabetes with renal manifestations
3) peritonitis and retroperitoneal infections

Machine Learning

- Data Splitting
  - Data were randomly split into 4 sets: training set (70%), and 3 hold-out sets (10% each). Training set was used for machine learning, hold-out set 1 was used for hyperparameter tuning, hold-out set 2 was used for validating models, hold-out set 3 was used for the final model testing
- Dealing with the Imbalanced Data
  - Positive class represents 11% of the data
  - Random Undersampling, and 2 oversampling methods were tested (SMOTE and ADASYN). Random Unersampling showed the best performance
- F1 score was selected for tuning the model

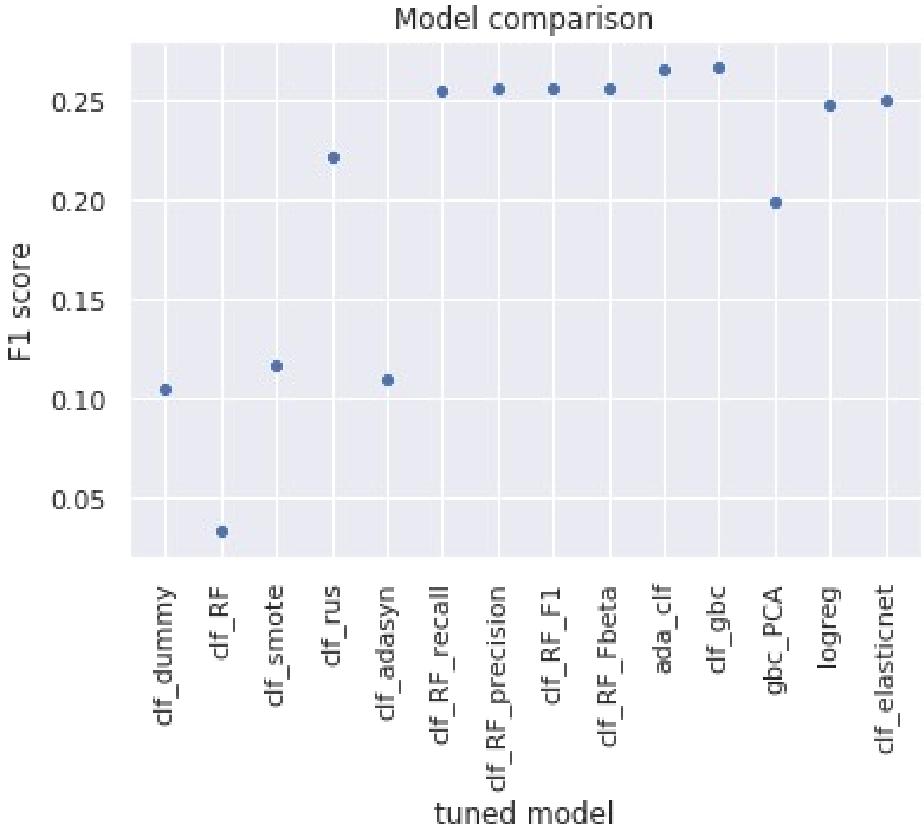
Models fitted

- Dummy classifier
- Random Forest
- Logistic Regression
- Logistic Regression with Stochastic Gradient Descent
- AdaBoost
- Gradient Boosting Classifier
- Gradient Boosting Classifier fitted on principal components

Best Hyperparameters

Model	Hyperparameters
Dummy	None
Random Forest	n_estimators=20, max_depth=5, min_samples_split=5, max_features=25
Logistic Regression	penalty='l2', solver='liblinear', C=1
Logistic Regression, stochastic	alpha=0.001, penalty='elasticnet', l1_ratio=0.3
AdaBoost	max_depth=2, min_samples_split=2, n_estimators=10
Gradient Boosting	n_estimators=100, max_depth=2, min_samples_split=2, max_features=20

Model Comarison

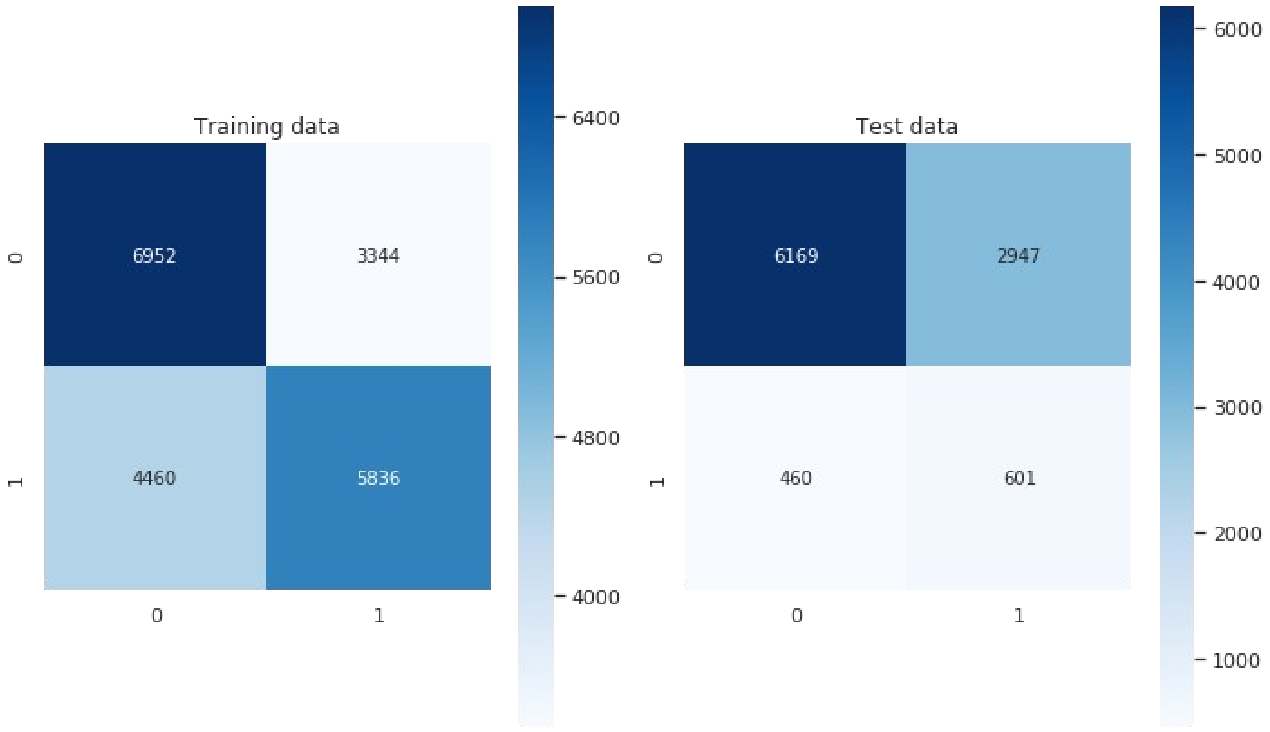


Gradient Boosting Classifier was selected based on the F1 score

Final Model Evaluation

metric	training data	test data
precision score	0.64	0.17
recall score	0.57	0.57
F1 score	0.60	0.26
Fβ score	0.63	0.17
Mattews correlation coefficient	0.24	0.16
accuracy score	0.62	0.67

Final Model Evaluation



Limitations

- Overfitting. Model performes much better on the training set than on the test set.
- Poor precision. Precision of predictions on the test set is ~17%.

Recommendations

- Use the model as is. The model allows to narrow down the number of patients that would benefit from the follow-up program aimed at minimazing the chances of readmission. This might result in significant reduction of the 30-day readmission rates.
- If all patients identified as positive by the model do not return within 30 days of discharge as a result of follow-up program, the 30-day readmission rate would drop by more than half. This is an upper estimate of the benefit of the current model.
- If better precision is required, more data need to be collected. Larger number of features would also be helpful for building the model. This will help with building a model with better precision.