Gaining Insight on Hospital Readmissions

I) Problem Identification

Using patient data to make personalized healthcare decisions has become a standard of care in medicine over recent years. This is because healthcare providers capable of delivering personalized solutions stand to benefit both the patient and the hospital. Hospitals are charged a fee imposed by the federal government if a patient they release from their care is readmitted to the hospital within 30 days of discharge. Knowing what patients are likely to be readmitted offers clinical insight to allow healthcare providers make more informed decisions. This can offer financial advantages to the hospital as well as the patient since both parties avoid excessive costs.

Machine learning models can be used as a tool to aid physicians and nurses in developing treatment plans. In this project, I explore hospital readmissions as it pertains to a healthcare providers decision to discharge a patient from their care. The machine learning models I build are optimized and scored on how well they are able to predict whether a patient is readmitted or not. The goal is to identify those patients with as little error as possible in order to offer insights to medical practices. The best performing model is able to make predictions with 77% accuracy. The models capabilities, drawbacks, and potential are explored here.

II) Data Wrangling

The raw dataset is made available through Kaggle and includes data from 25,000 patients. The origin of the data is not indicated by the user who provided it, however, it is said to include real patient data from actual hospital encounters.

The original dataset contained 65 column features which I was able to reduce to 44. The data contained some missing values that I labeled as ‘unknown’ where no inference could be made. The categorical features were represented as boolean values which I converted into categorical values in order to conduct EDA. The diagnoses were written in ICD-9 format which I converted into the actual diagnoses names to facilitate interpretability. Here I have broken the data down into categories.

- **Patient demographics:** age, race, and gender
- **Hopital stay information:** days spent in the hospital, number of lab procedures performed, and number of procedures performed
- **Hospital history information:** number of inpatient, outpatient, and emergency room visits the patient had over the past year

- **Medications in use:** number of medications in use and whether they were taking metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone at the time

- **Diagnoses:** number of diagnoses and whether they were diagnosed with heart failure, other heart disease, respiratory symptoms, pH/fluid imbalance, diabetes mellitus, cardiac dysrhythmias, and essential hypertension

- **Other miscellaneous information:** billing paycode, lab results, and medical specialty seeing the patient

- **Target variable:** whether the patient was readmitted to the hospital within 30 days after initial discharged

III) EDA

About 46% of the patients in this dataset were readmitted within 30 days of discharge. The data is unlikely to be randomly sampled from the population since government sources conclude this percentage to be anywhere from 10-20%. This discrepancy, however, should not impact my analysis since I am exploring what features result in readmission.

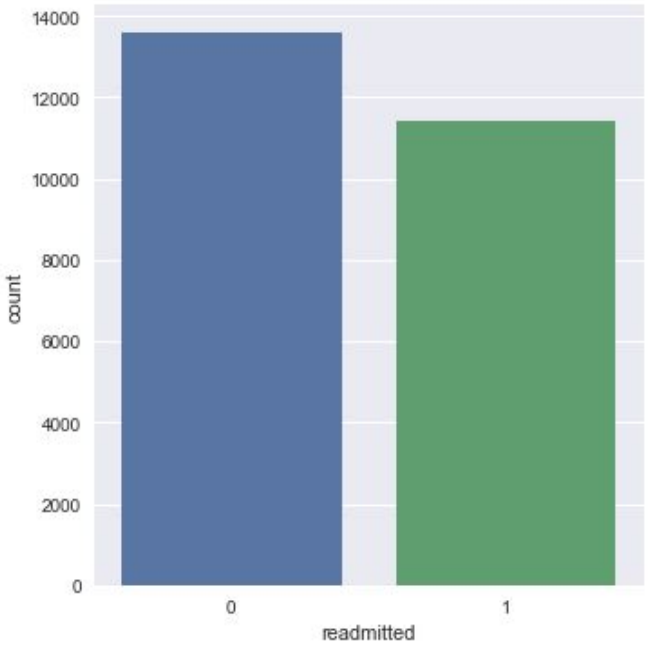

Figure1: count of how many patients were readmitted

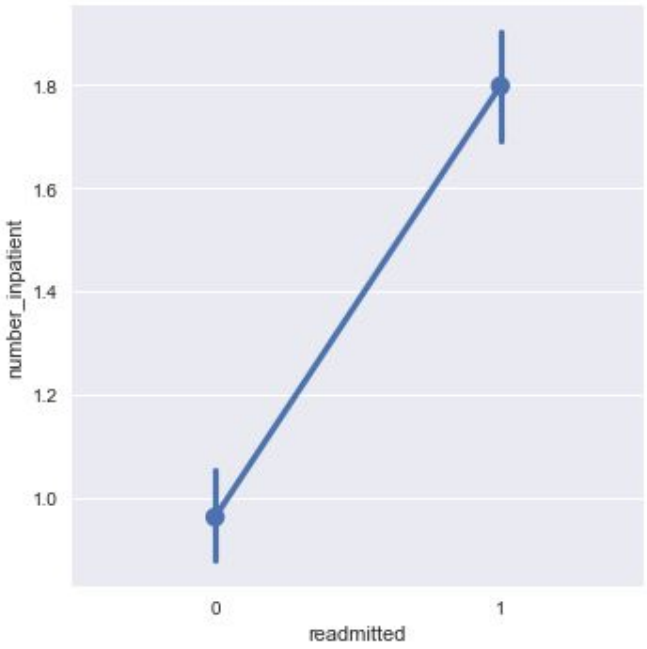

Figure 2: how number of inpatient visits over the past year relates to whether the patient was readmitted or not

IV) Pre-Processing

Categorical data was converted into binary values for machine interpretability using OneHotEncoder. Continuous variables were scaled to values between 0 and 1 using StandardScaler.

	race_African_American	race_Caucasian	race_Other	gender_Female	gender_Male
0	0.0	1.0	0.0	0.0	1.0
1	0.0	1.0	0.0	1.0	0.0
2	0.0	1.0	0.0	1.0	0.0
3	0.0	1.0	0.0	1.0	0.0
4	0.0	1.0	0.0	1.0	0.0

Figure 3: categorical variables converted into binary values

days_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient
1.000000	0.320	0.0	0.1250	0.000000
0.076923	0.232	0.0	0.1375	0.000000
0.307692	0.520	0.0	0.2625	0.027778
0.153846	0.496	0.0	0.0875	0.000000
0.307692	0.312	0.0	0.0625	0.000000

Figure 4: continuous variables converted to scaled data

V) Modeling

I explored four different binary classification machine learning model algorithms to determine the best predictor. The algorithms used in this study are logistic regression (LR), k nearest neighbors (KNN), random forest classification (RFC), and gradient boosting (XGB). A dummy model was also created for comparison.

VI) Model Selection

The model algorithm that performed the best was RFC. It had the highest accuracy of 77% (23% higher than the dummy model) and the lowest amount of type II errors. In other words, this model reduces the probability of falsely releasing a patient too soon who is later readmitted.

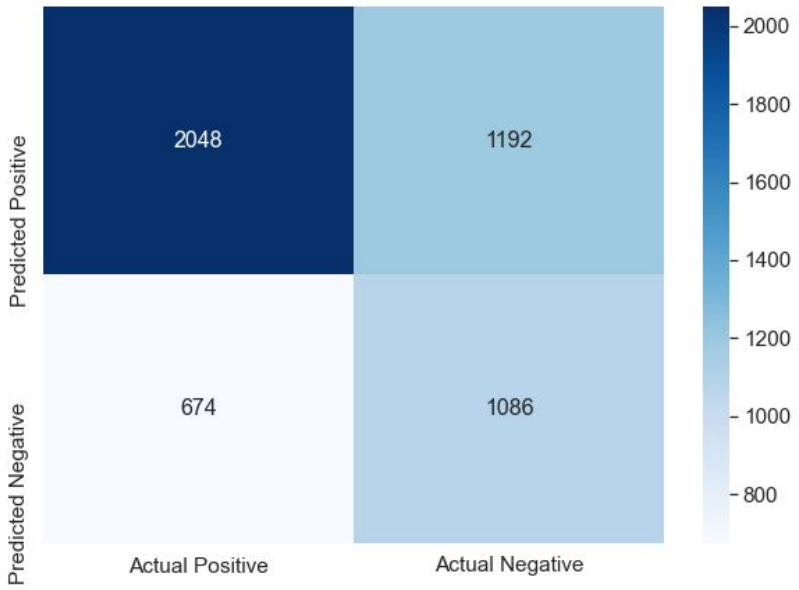

Figure 5: confusion matrix displaying the RFC Gridsearch CV model results

Model performance for all models can be viewed using an ROC curve and computing the AUC score.

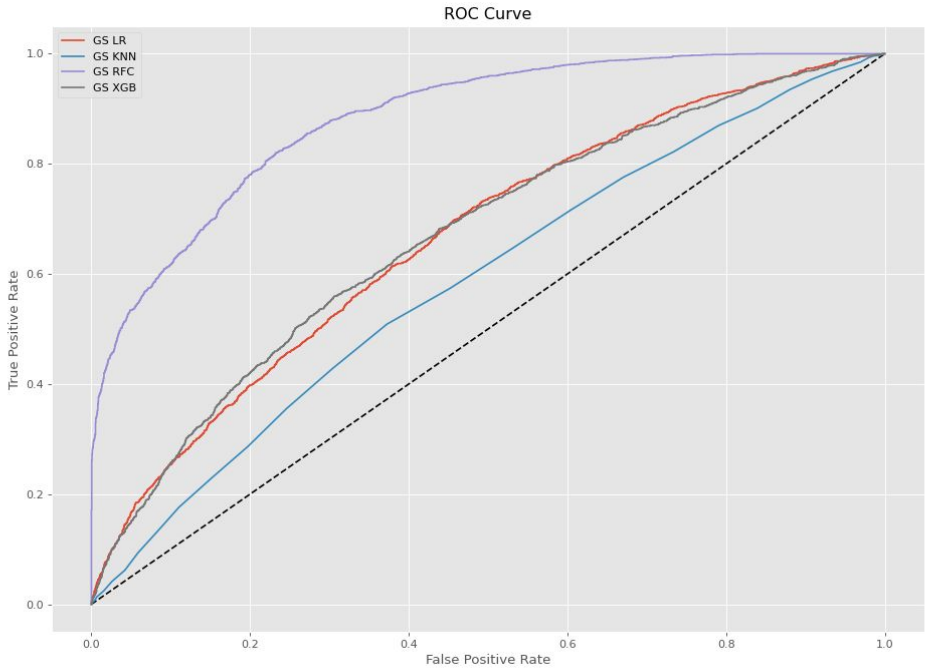

Figure 6: ROC curves for all models after Gridsearch CV for comparison

Labels	Default_accuracy_scores	Fit_times	GS_accuracy_scores	AUC_scores
RFC	0.6118	4.106606	0.7734	0.883675
XGB	0.6202	0.454879	0.6268	0.668189
LR	0.6152	0.558409	0.6146	0.665155
KNN	0.5400	0.557119	0.5740	0.586294

Figure 7: model specs after Gridsearch CV for comparison

	Importance	Feature
0	0.149487	number_inpatient
1	0.090562	num_medications
2	0.083157	num_lab_procedures
3	0.057175	number_diagnoses
4	0.054365	days_in_hospital
5	0.045655	number_emergency
6	0.042379	number_outpatient
7	0.036866	num_procedures
8	0.015015	Dx1_Heart_Failure
9	0.013838	paycode_payer_code_MC
10	0.012966	diabetesMed_Yes

Figure 8: most important features for RFC Gridsearch CV model decision

The number of inpatient visits over the past year was the best predictor to whether a patient would be readmitted to the hospital within 30 days of discharge. The first eight predictors are all integer features with higher numbers likely resulting in greater probability of readmission.

VII) Limitations

While the RFC model offers the greatest predictive capabilities, it's time to fit the model is significantly greater than competing models. With larger datasets, longer fit time introduces a computing challenge given RAM limitations on most machines. This may require the larger datasets to be broken down into small subsets in order to allow for computability.

The RFC model has a relatively high Type I error rate which may result in falsely releasing a patient from hospital care who is later readmitted.

VIII) Areas for Further Exploration

The model may be improved even further by using Bayesian hyperparameter optimization to hypertune some of the hyperparameter values which were not investigated.

It would be worth exploring features not necessarily made available to hospitals such as annual income and living situation, as well as other health features made available to hospitals such as BMI, A1c, blood pressure, etc.