Supervised Learning Capstone

BY: FRANCES CUE

Background

I’ve acquired a Diabetic Readmission Data Set from kaggle. The dataset is composed of integrated data from 130 U.S. hospitals over the course of 10 years. It includes over 50 features such as length of stay, medications, lab results and primary diagnosis.

In the healthcare field, determining the chances of a patient readmission is vital. If risks for readmission is known, better treatment plan can be created for patients.

Objective

To create various models to determine diabetic patient readmission prediction less than 30 days of previous admission.

Exploratory Data Analysis

- ❖Categorical values includes gender, race, age (which were group into counts of 10), medications, diagnosis via ICD9 codes, medical specialty of admitting doctor, medication changes and more.
- ❖Numerical columns included patient identifiers such as encounter id and patient number, admission type, number of times a patient was inpatient, outpatient, in emergency, procedures, time in hospital and number of medications.

Object Types

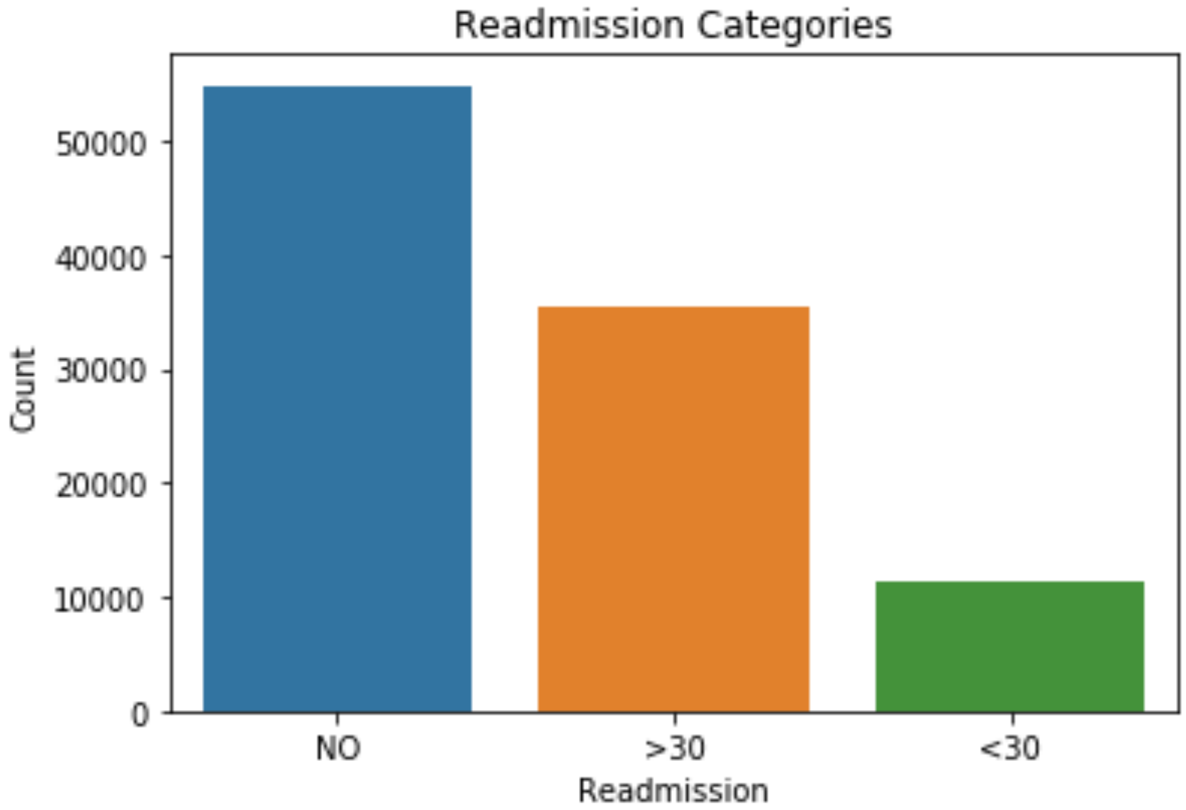
We have a mixture of 37 categorical values and 13 numerical/continuous values:

```
1  num_cols.dtypes

encounter_id                int64
patient_nbr                 int64
admission_type_id           int64
discharge_disposition_id    int64
admission_source_id         int64
time_in_hospital            int64
num_lab_procedures          int64
num_procedures              int64
num_medications             int64
number_outpatient           int64
number_emergency            int64
number_inpatient            int64
number_diagnoses            int64
target                      int64
age_group                   int64
dtype: object
```

```
3  cat_cols.dtypes

race                        object
gender                      object
age                         object
diag_1                      object
diag_2                      object
diag_3                      object
max_glu_serum               object
A1Cresult                   object
metformin                   object
repaglinide                 object
nateglinide                 object
chlorpropamide              object
glimepiride                 object
acetohexamide               object
glipizide                   object
glyburide                   object
tolbutamide                 object
pioglitazone                object
rosiglitazone               object
acarbose                    object
miglitol                    object
troglitazone                object
tolazamide                  object
examide                     object
citoglipton                 object
insulin                     object
glyburide-metformin         object
glipizide-metformin         object
glimepiride-pioglitazone    object
metformin-rosiglitazone     object
metformin-pioglitazone      object
change                      object
diabetesMed                 object
readmitted                  object
dtype: object
```

Readmission Categories

- The target chosen for this project is to predict readmission in less than 30 days after previous admission
- It is worth noting that the readmission > 30 days does not specify exactly when patient was readmitted.



NO	52527
>30	35502
<30	11314

Looking for missing values

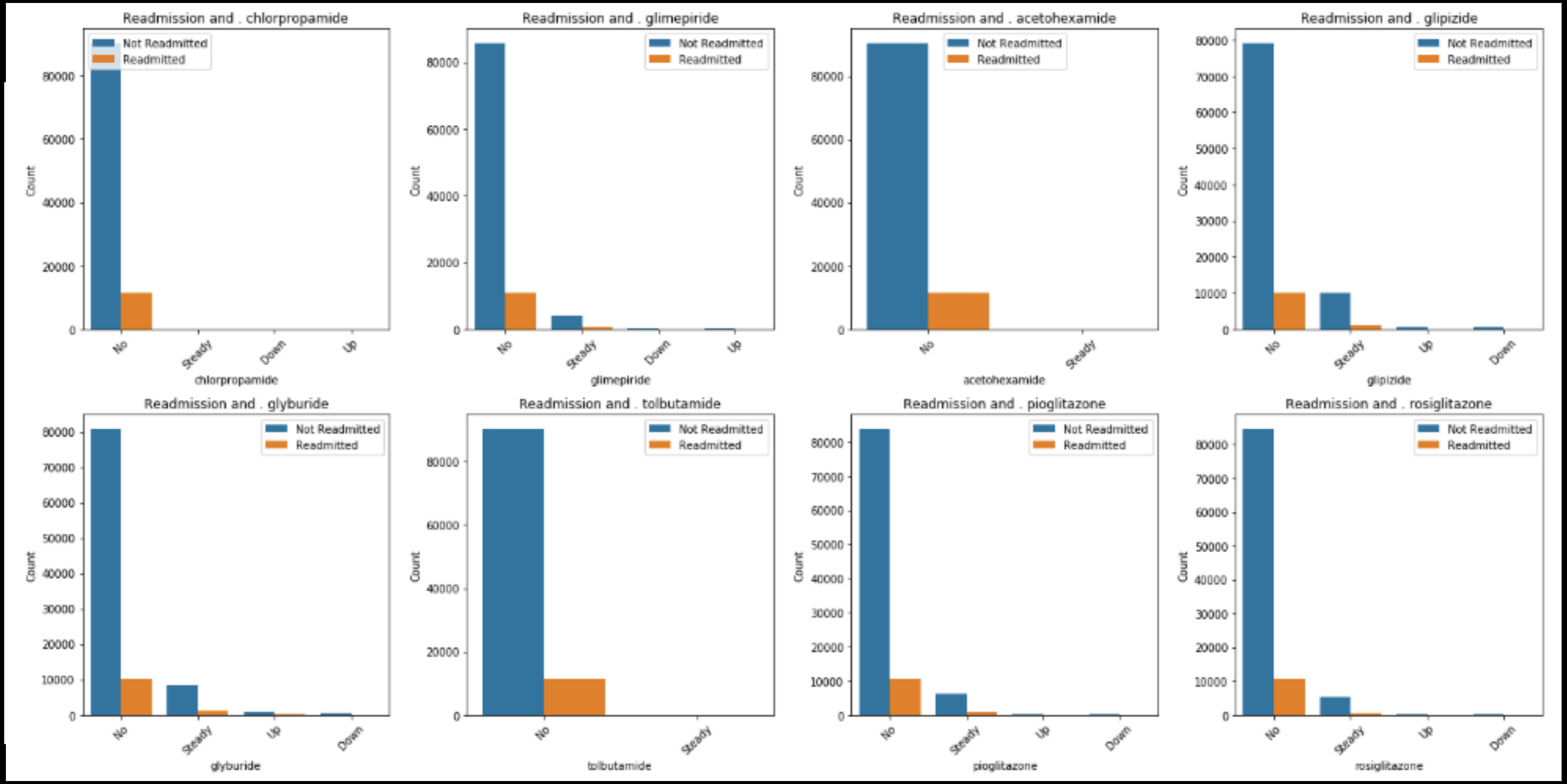
The dataset contained no null values upon first examination. However, it had many '?' values that represent missing values.

Columns with over 90% missing were dropped. Other columns that provide no insight such as insurance payer codes were dropped as well.

```python
#looking for ? values.
for col in diabetes.columns:
    if diabetes[col].dtype == object:
        print(col,diabetes[col][diabetes[col] == '?'].count())
```

```
race 2273
gender 0
age 0
weight 98569
payer_code 40256
medical_specialty 49949
diag_1 21
diag_2 358
diag_3 1423
max_glu_serum 0
A1Cresult 0
metformin 0
repaglinide 0
nateglinide 0
chlorpropamide 0
```

There were many medications (27) included in the dataset. Visualization helped determine relationship between target.



Feature Engineering

```
{'[0-10)':0,
 '[10-20)':10,
 '[20-30)':20,
 '[30-40)':30,
 '[40-50)':40,
 '[50-60)':50,
 '[60-70)':60,
 '[70-80)':70,
 '[80-90)':80,
 '[90-100)':90}
'age_group'] = diabete
```

- New feature was created for age since it was grouped by 10s. This in turn created a new numerical value for age.
- With discharge dispositions, patients who have expired or was sent to hospice were dropped because there will be no chance for readmission.

Grouped diagnosis by counts/occurrence

	Diagnosis	ICD_code	count
0	Heart Failure	428	6862
1	Other forms of Chronic Heart Disease	414	6581
2	Symptoms involving respiratory system and othe...	786	4016
3	Myocardial infarcation	410	3614
4	Pneumonia	486	3508
5	Cardiac Dysrythmias	427	2766
6	Emphysema	491	2275
7	Osteoarthritis	715	2151
8	Cellulitis	682	2042
9	General Symptoms	434	2028

	2nd_diagnosis	ICD_code	count
0	Disorders of fluid electrolyte and acid-base b...	276	6752
1	Heart Failure	428	6662
2	Diabetes mellitus without mention of complicat...	250	6071
3	Cardiac Dysrythmias	427	5036
4	Essential Hypertension	401	3736
5	Chronic airway obstruction, not elsewhere clas...	496	3305
6	Other disorders of urethra and urinary tract	599	3288
7	Hypertensive chronic kidney disease	403	2823
8	Other forms of chronic ischemic heart disease	414	2650
9	Other acute and subacute forms of ischemic hea...	411	2566

	Other_diagnosis	ICD_code	count
0	Diabetes mellitus without mention of complicat...	250	11555
1	Essential Hypertension\t	401	8289
2	Disorders of fluid electrolyte and acid-base ...	276	5175
3	Heart Failure	428	4577
4	Cardiac Dysrythmias	427	3955
5	Other forms of chronic ischemic heart disease	414	3664
6	Chronic airway obstruction, not elsewhere clas...	496	2605
7	Hypertensive chronic kidney disease	403	2357
8	Chronic Kidney Disease	585	1992
9	Disorders of lipoid metabolism	272	1969

Chi-squared Feature Importance

Was able to reduce features that were not

Important using Chi-squared.

```
['discharge_disposition_id_22',
 'discharge_disposition_id_3',
 'number_diagnoses',
 'number_inpatient',
 'number_emergency',
```

Chi-Squared Top 5 features

Have tried the following to address class imbalance:

Oversampling using SMOTE Technique

Undersampling

Class Imbalance

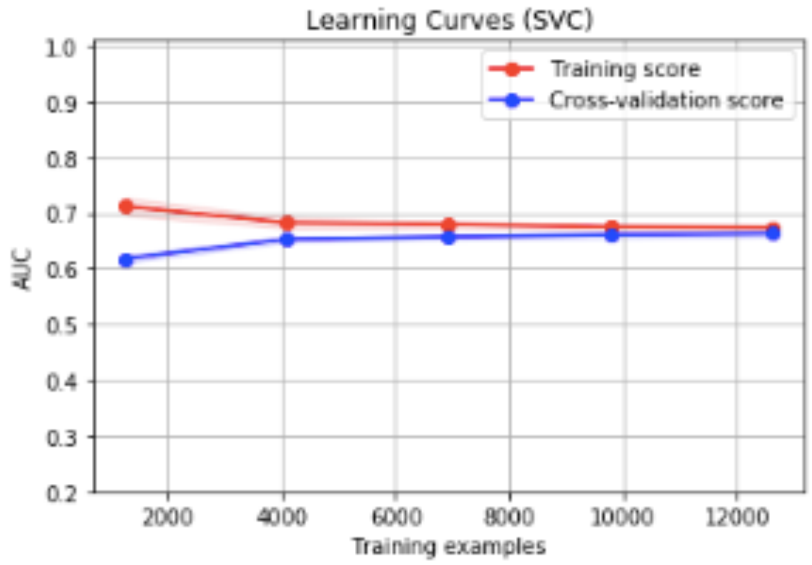
SVC

TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.672	AUC:0.668	AUC:0.94	AUC:0.56
accuracy:0.622	accuracy:0.685	accuracy:0.88	accuracy:0.689
recall:0.527	recall:0.542	recall:0.882	recall:0.345
precision:0.650	precision:0.189	precision:0.91	precision:0.112
specificity:0.693	specificity:0.682	specificity:0.77	specificity:0.712
F1:0.582	F1:0.281	f1:0.89	f1:0.131

UNDERSAMPLING **SMOTE**

Support Vector Classifier

- Longest run times.
- Tuned C and gamma parameter without much improvement.




```
Baseline SVC
Training AUC:0.672
Validation AUC:0.668
Optimized SVC
Training AUC:0.672
Validation AUC:0.668
```

KNN

TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.652	AUC:0.624	AUC:0.923	AUC:0.551
accuracy:0.605	accuracy:0.650	accuracy:0.850	accuracy:0.672
recall:0.518	recall:0.504	recall:0.882	recall:0.333
precision:0.627	precision:0.163 s	precision:0.901	precision:0.131
specificity:0.658	specificity:0.637	specificity:0.768	specificity:0.706
F1:0.567	F1:0.246	f1:0.88	f1:0.08

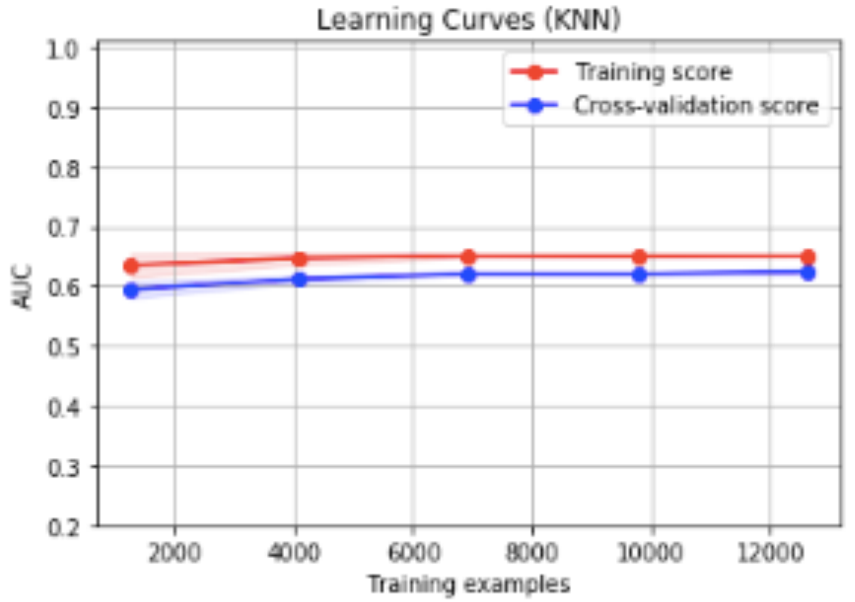
UNDERSAMPLING **SMOTE**

KNN

Moderate running times.

Tuned by increasing n_neighbors, using the minkowski metric, and adding uniform weight.

```
Baseline KNN
Training AUC:0.652
Validation AUC:0.624
Optimized KNN
Training AUC:0.649
Validation AUC:0.629
```



Logistic Regression

TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.675	AUC:0.667	AUC:0.908	AUC:0.543
accuracy:0.623	accuracy:0.668	accuracy:0.844	accuracy:0.773
recall:0.550	recall:0.572	recall:0.833	recall:0.188
precision:0.644	precision:0.186	precision:0.851	precision:0.138
specificity:0.696	specificity:0.681	specificity:0.855	specificity:0.849
F1: 0.59	F1: 0.28	F1:0.891	F1:0.188
UNDERSAMPLING		**SMOTE**	

Logistic Regression

An advantage is model is interpretability. Has feature importance

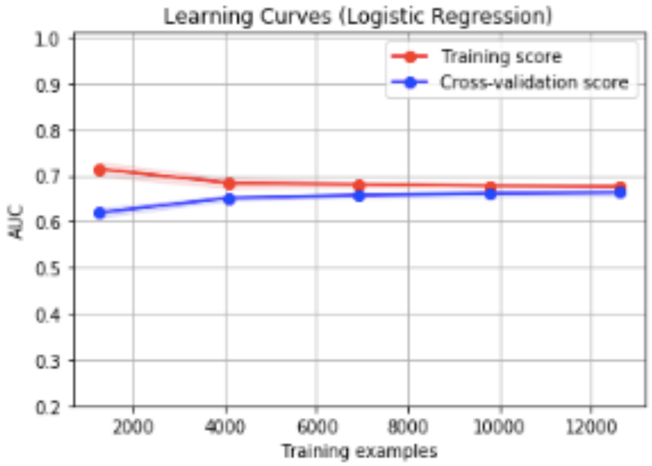
Fast training times.

Tuned by reducing C to 0.1, created l1 penalty, class weight was balanced.

Scores were not much different.

```
feature_importances.head(10)
```

	importance
number_inpatient	0.357715
discharge_disposition_id_22	0.188475
rosiglitazone_No	0.180301
repaglinide_No	0.173352
repaglinide_Steady	0.156140
rosiglitazone_Steady	0.134689
diabetesMed_Yes	0.120780
discharge_disposition_id_3	0.119310
discharge_disposition_id_28	0.110808
discharge_disposition_id_5	0.109521




```
Baseline Logistic Regression
Training AUC:0.675
Validation AUC:0.667
```

```
Optimized Logistic Regression
Training AUC:0.675
Validation AUC:0.668
```

Decision Trees

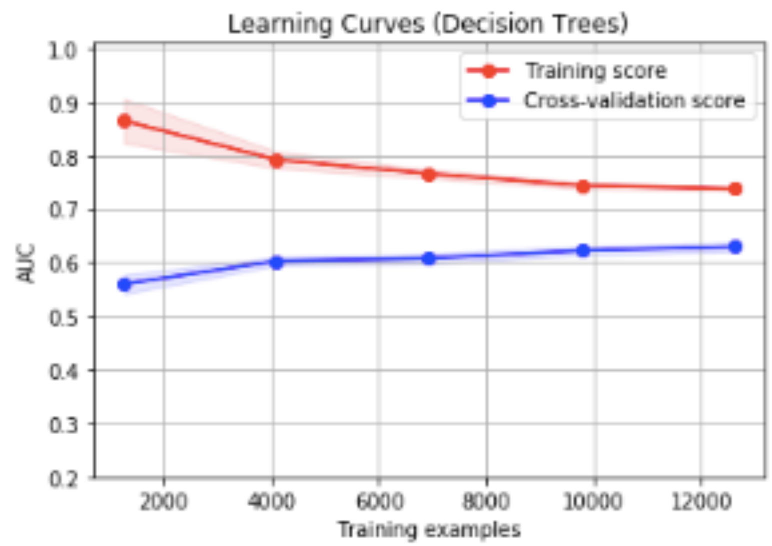
TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.729	AUC:0.637	AUC:0.848	AUC:0.538
accuracy:0.665	accuracy:0.664	accuracy:0.768	accuracy:0.757
recall:0.590	recall:0.539	recall:0.703	recall:0.216
precision:0.693	precision:0.177	precision:0.809	precision:0.138
specificity:0.738	specificity:0.679	specificity:0.833	specificity:0.826
F1: 0.63	F1: 0.26	F1:0.76	F1:0.11
UNDERSAMPLING		**SMOTE**	

Decision Trees

Has tendency to overfit.

Decided not to tune this model.

Moderate run time.



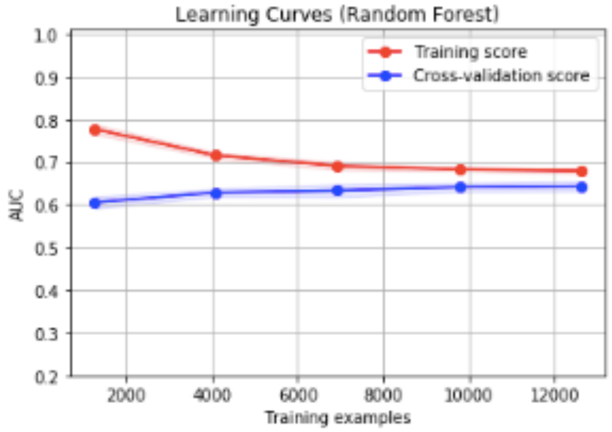
Random Forest

TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.671	AUC:0.635	AUC:0.839	AUC:0.536
accuracy:0.624	accuracy:0.618	accuracy:0.760	accuracy:0.698
recall:0.586	recall:0.587	recall:0.761	recall:0.279
precision:0.634	precision:0.165	precision:0.759	precision:0.126
specificity:0.661	specificity:0.622	specificity:0.759	specificity:0.752
F1: 0.61	F1: 0.26	F1:0.891	F1:0.188
UNDERSAMPLING		**SMOTE**	

Random Forest

- Has feature importance.
- Tuned by increasing number of estimators and depth.
- Moderate time training.
- Optimized score improved in training, but seems overfitted in validation set.

```
Baseline Random Forest
Training AUC:0.671
Validation AUC:0.635
Optimized Random Forest
Training AUC:0.713
Validation AUC:0.662
```



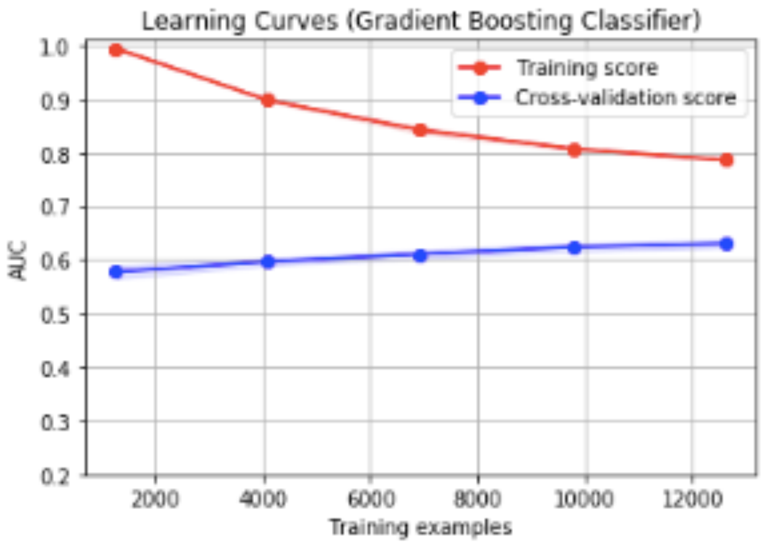
	importance
number_inpatient	0.183701
time_in_hospital	0.098817
number_emergency	0.093810
discharge_disposition_id_22	0.077494
num_medications	0.057466
num_lab_procedures	0.052790
number_diagnoses	0.045715
number_outpatient	0.028754
number_outpatient	0.023830
insulin_No	0.023613

Gradient Boosting Classifier

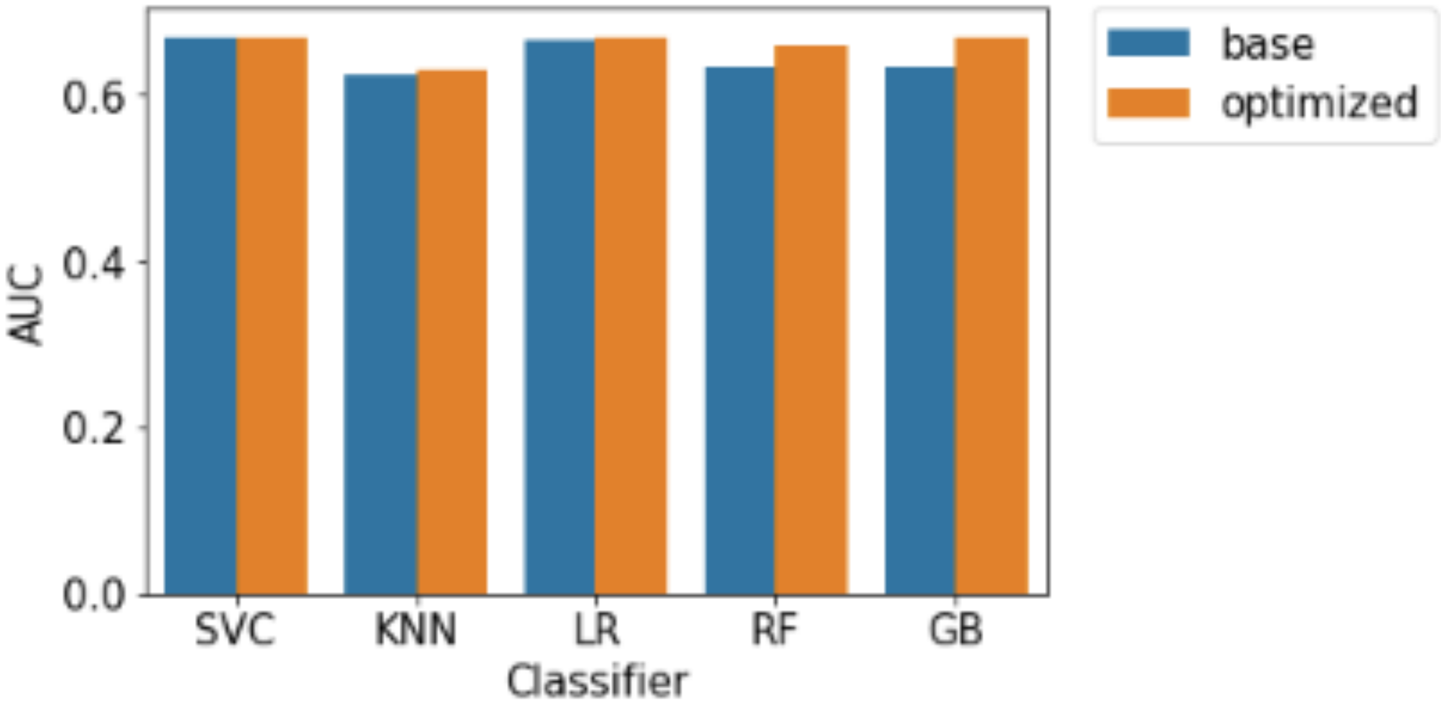
TRAIN	VALIDATION	TRAIN	VALIDATION
AUC:0.770	AUC:0.632	AUC:0.926	AUC:0.553
accuracy:0.694	accuracy:0.614	accuracy:0.861	accuracy:0.793
recall:0.670	recall:0.586	recall:0.839	recall:0.186
precision:0.704	precision:0.163	precision:0.878	precision:0.157
specificity:0.718	specificity:0.618	specificity:0.884	specificity:0.871
F1: 0.68	F1: 0.25	F1: 0.86	F1. 0.116
UNDERSAMPLING		**SMOTE**	

Gradient Boosting Classifier

- Has the best AUC scores for training and validation set.
- Moderate run time.



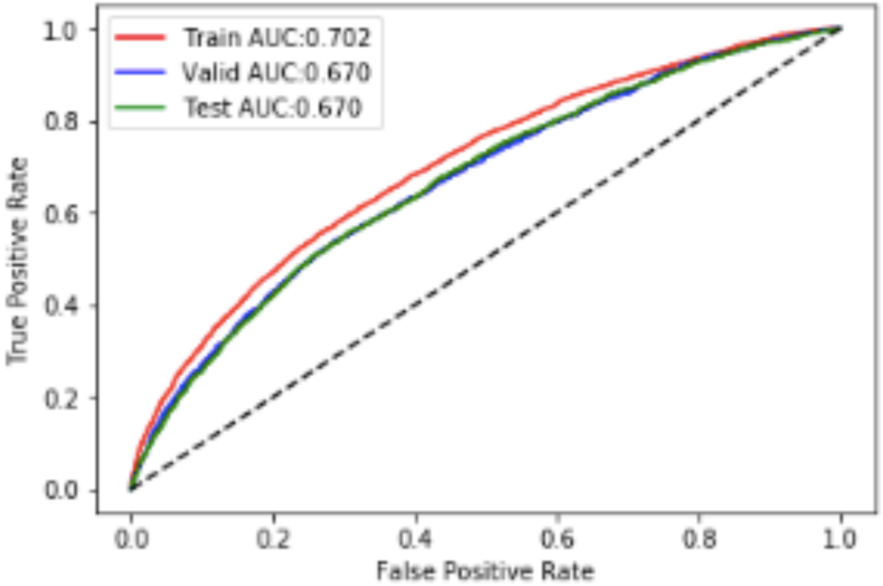

```
Baseline Gradient Boosting Classifier
Training AUC:0.770
Validation AUC:0.632
Optimized GBC
Training AUC:0.702
Validation AUC:0.670
```



Hyperparameter Tuning Results

Model Selection: GBC

```
Training:
AUC:0.702
accuracy:0.644
recall:0.593
precision:0.660
specificity:0.694
prevalence:0.500

Validation:
AUC:0.670
accuracy:0.651
recall:0.584
precision:0.179
specificity:0.660
prevalence:0.113

Test:
AUC:0.670
accuracy:0.644
recall:0.590
precision:0.183
specificity:0.651
prevalence:0.117

gbc f1:0.280
```



Conclusion

What have we learned from exploring this dataset?

1. There is a correlation between number of inpatient visits and being readmitted less than 30 days.

2. A patient being discharged to the rehab or a subacute facility has a higher chance of readmission in less than 30 days.

3. Since many patients have a primary diagnosis of heart related conditions, it is worth looking at studying readmission rates for this population. Diabetes and heart disease have a known correlation, how does this affect readmission rates?

4. If the intention is to truly predict readmission for diabetic patients, it may be helpful to look at diabetes as a primary diagnosis. According to the ICD 10, primary diagnosis requires the most serious attention and is resource intensive while secondary and tertiary diagnosis could be diseases that co-exist during admission or develop thereafter admission.

5. Additional information may be needed for this dataset. Information such as procedures and certain blood work could provide more insight into readmission.

Reference:

Dataset acquired from: https://www.kaggle.com/brandao/diabetes

ICD 10: https://www.icd10watch.com/blog/clearing-confusion-between-principal-and-primary-diagnoses

ICD 9 Codes : http://www.icd9data.com/