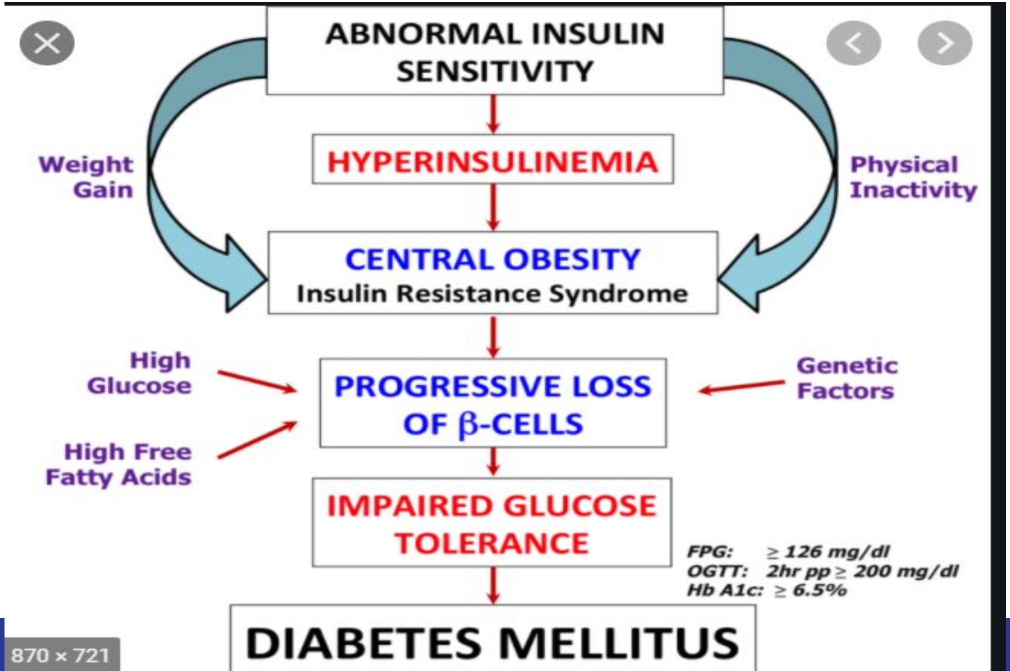
Predicting Hospital Readmission of Diabetes Patients

DSI - Capstone Project
December 2020
Veronica Phillip

What is Diabetes?

Diabetes, is a metabolic disease that causes high blood sugar. The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. With **diabetes**, your body either doesn't make enough insulin or can't effectively use the insulin it does make.



Problem Statement

- Diabetes is a condition that can be effectively treated in primary care facilities.
- High and increasing levels of hospital readmission are cause for concerns.

Impact

- High readmission rates can point to quality of hospital care
- Impact on the cost to the Healthcare providers - Medicare and Medicaid
- Cost to the patient and their families.
- Psychological impact of repeated admissions, especially if it relates to the same illness.
- Impact on staffing levels within the hospital
- Impact on hospital budgets .

Project Goals

- Identify the factors that drive hospital readmission of diabetes patients
- Build a model that can predict the likelihood hospital readmission of diabetes patients occurring.
- Propose a strategy for reducing hospital readmission using the model and other insights from the analysis.
- Metrics - Precision and Recall

Data Set

- UCI Dataset
- 10 years of clinical care at 130 US hospitals and integrated delivery networks.
- Inpatient admission encounters for diabetes patients.
-

```
diabetes.shape
```

```
(101766, 50)
```

Variables in the dataset

- Encounter_id
- Patient_nbr
- Race
- Gender
- Age
- Weight
- Admission_type_id
- Discharge_disposition_id
- Admission_source_id
- Time_in_hospital
- Payer_code
- Medical_specialty
- Num_lab_procedures
- Num_procedures
- Num_medications
- Number_outpatient
- Number_emergency
- Number_inpatient
- Diag_1
- Diag_2
- Diag_3
- Number_diagnoses
- Max_glu_serum
- A1C result
- Metformin
- Repaglinide
- Change
- Diabetes Med
- Readmitted
- + 23 diabetes medication

Data Cleaning & Feature Engineering

Data pre-processing.

- Removal of unwanted variables
- Missing value treatment
- Variable transformation
- Addition of new variables
- Data split into test and train

Transforming Variables

The code indicating the type and priority of an inpatient admission associated with the service on an intermediary submitted claim.

Comments: Source: NCH

Code	Code value
0	Blank
1	Emergency - The patient required immediate medical intervention as a result of severe, life threatening, or potentially disabling conditions. Generally, the patient was admitted through the emergency room.
2	Urgent - The patient required immediate attention for the care and treatment of a physical or mental disorder. Generally, the patient was admitted to the first available and suitable accommodation.
3	Elective - The patient's condition permitted adequate time to schedule the availability of suitable accommodations.
4	Newborn - Necessitates the use of special source of admission codes.
5	Trauma Center - visits to a trauma center/hospital as licensed or designated by the State or local government authority authorized to do so, or as verified by the American College of Surgeons and involving a trauma activation.
6 THRU 8	Reserved
9	Unknown - Information not available.

Inpatient Admission Type Code (FFS)

- Used to convert the numbers for Admission type Id to categorical variables

Transforming Variables

Diag_1, Diag_2 & Diag_3

These are medical diagnosis codes that are used In health care. Diagnosis codes are used as a tool to group and identify diseases, disorders, symptoms, poisonings, adverse effects of drugs and chemicals, injuries and other reasons for patient encounters.

Unique values in each Variable were as follows:

diag_1	101766	716
diag_2	101766	748
diag_3	101766	789

Converted to 9 Categories using the ICD9 Codes from Biomed Research International.

BioMed Research International

Group name	icd9 codes
Circulatory	390–459, 785
Respiratory	460–519, 786
Digestive	520–579, 787
Diabetes	250.xx
Injury	800–999
Musculoskeletal	710–739
Genitourinary	580–629, 788
Neoplasms	140–239
	780, 781, 784, 790–799
	240–279, without 250
	680–709, 782
	001–139
Other (17.3%)	290–319
	E–V
	280–289
	320–359
	630–679
	360–389
	740–759

The Target Variable

Readmitted:

This was initially:

```
diabetes['readmitted'].value_counts(normalize=True)
```

```
NO     0.539106
>30    0.349292
<30    0.111602
Name: readmitted, dtype: float64
```

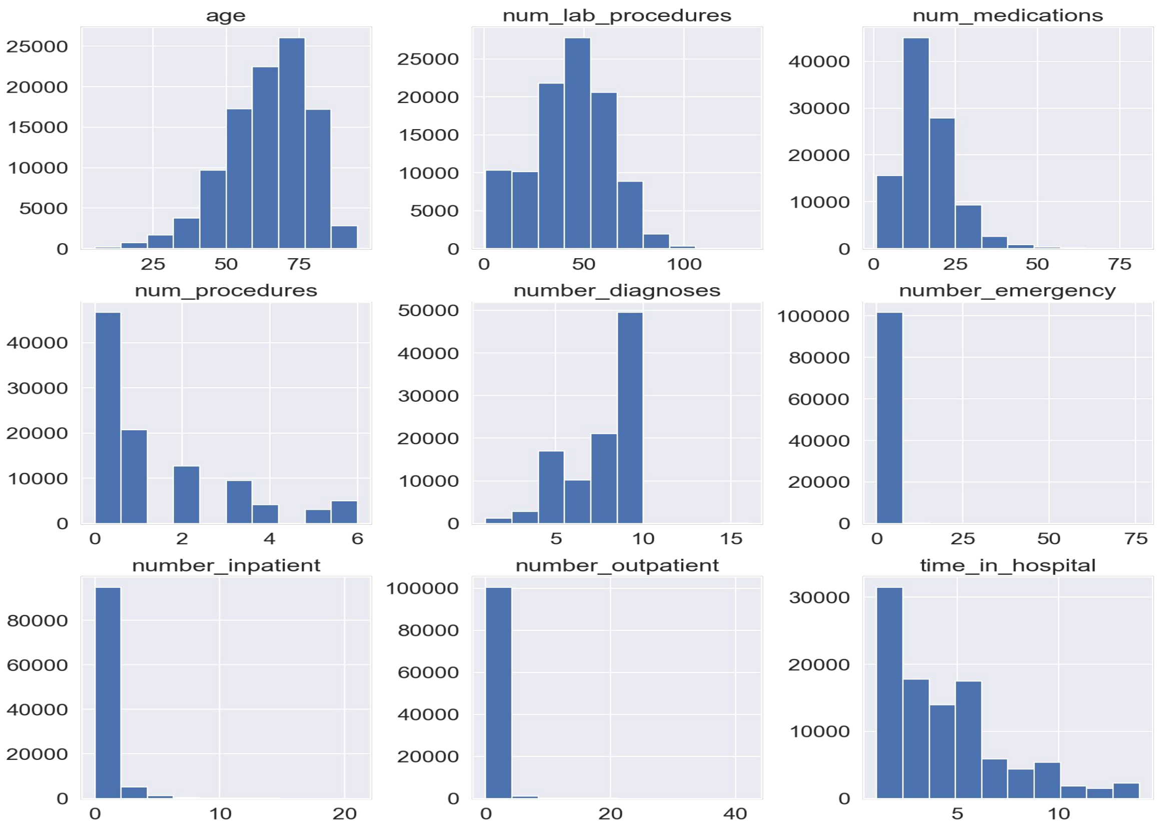
Converted to:

```
diabetes1.readmitted.value_counts(normalize=True)
```

```
0    0.539106
1    0.460894
Name: readmitted, dtype: float64
```

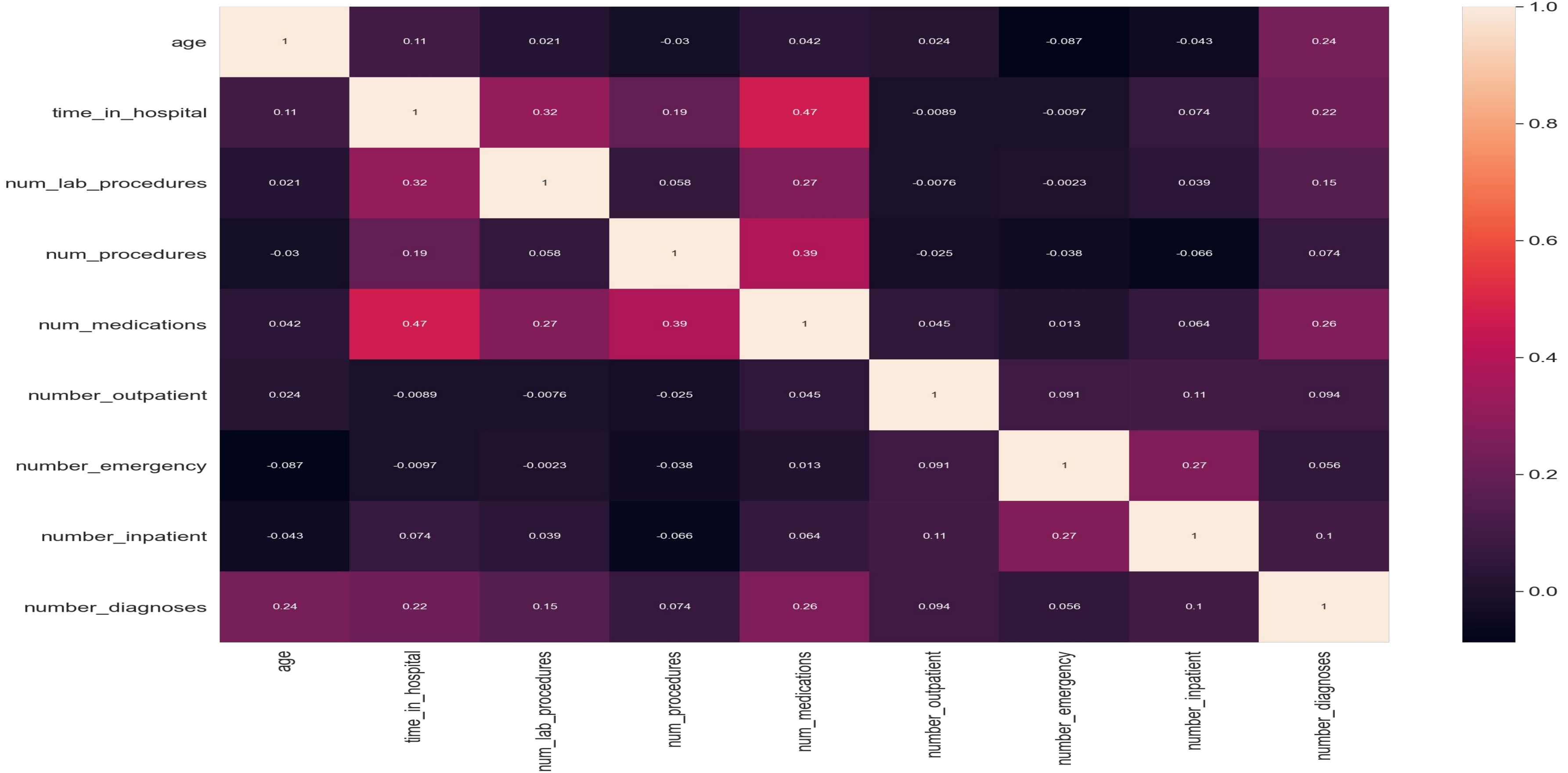
Conversion to binary Categorical variables done to eliminate 2 categories for readmission.

Data Visualizations

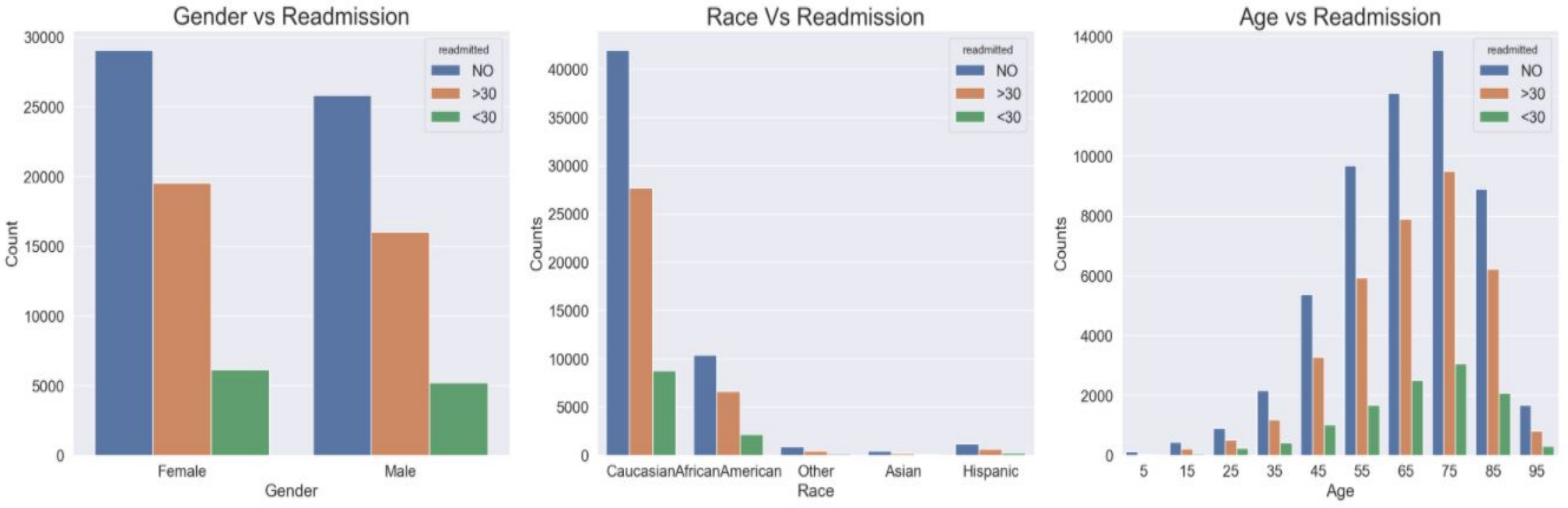


Data log transformed to remove skewness

Heat Map

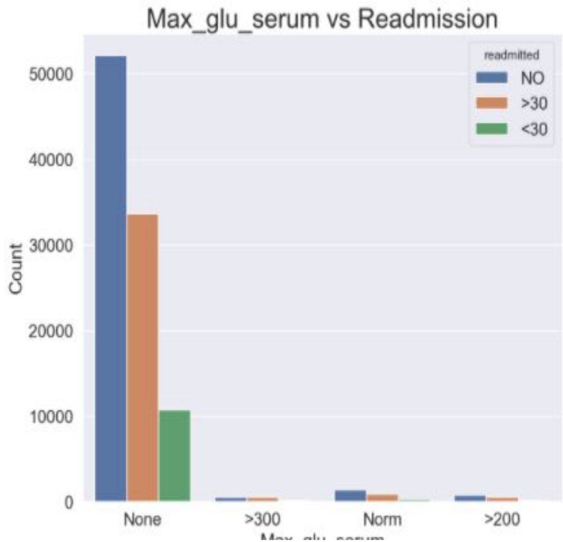

age
time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient
number_diagnoses
1 0.11 0.021 -0.03 0.042 0.024 -0.087 -0.043 0.24
0.11 1 0.32 0.19 0.47 -0.0089 -0.0097 0.074 0.22
0.021 0.32 1 0.058 0.27 -0.0076 -0.0023 0.039 0.15
-0.03 0.19 0.058 1 0.39 -0.025 -0.038 -0.066 0.074
0.042 0.47 0.27 0.39 1 0.045 0.013 0.064 0.26
0.024 -0.0089 -0.0076 -0.025 0.045 1 0.091 0.11 0.094
-0.087 -0.0097 -0.0023 -0.038 0.013 0.091 1 0.27 0.056
-0.043 0.074 0.039 -0.066 0.064 0.11 0.27 1 0.1
0.24 0.22 0.15 0.074 0.26 0.094 0.056 0.1 1
1.0
0.8
0.6
0.4
0.2
0.0

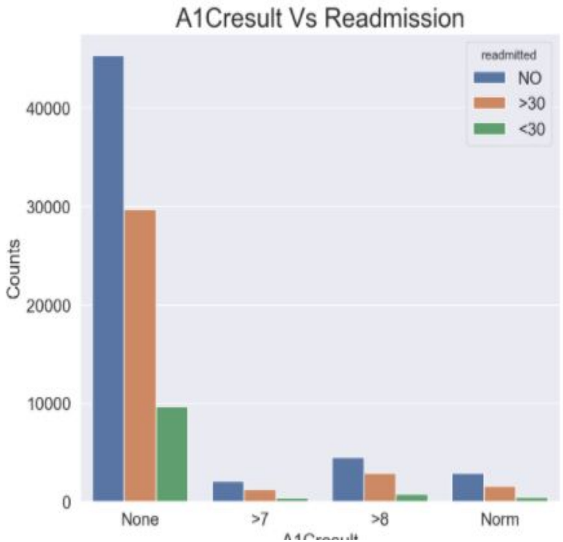
Data Visualizations

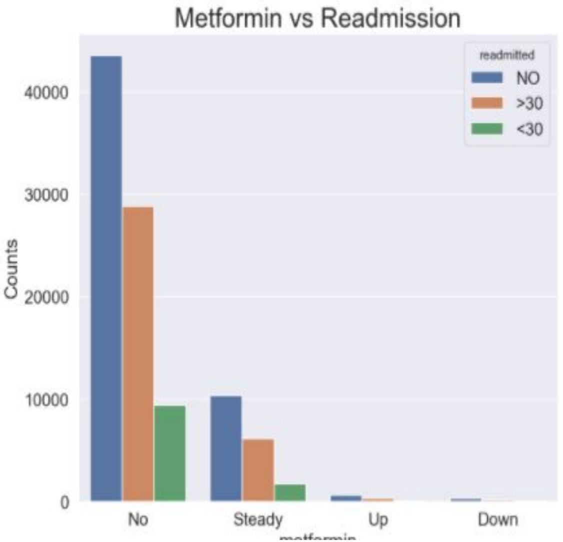


- Readmission among females greater than of male diabetic patients
- Caucasions have the highest level of readmission; but is inline with the number of diabetic patients in that race.
- Readmission is highest among diabetic patients aged 45 - 85 and increases with age.

Data Visualizations (Cont’d)

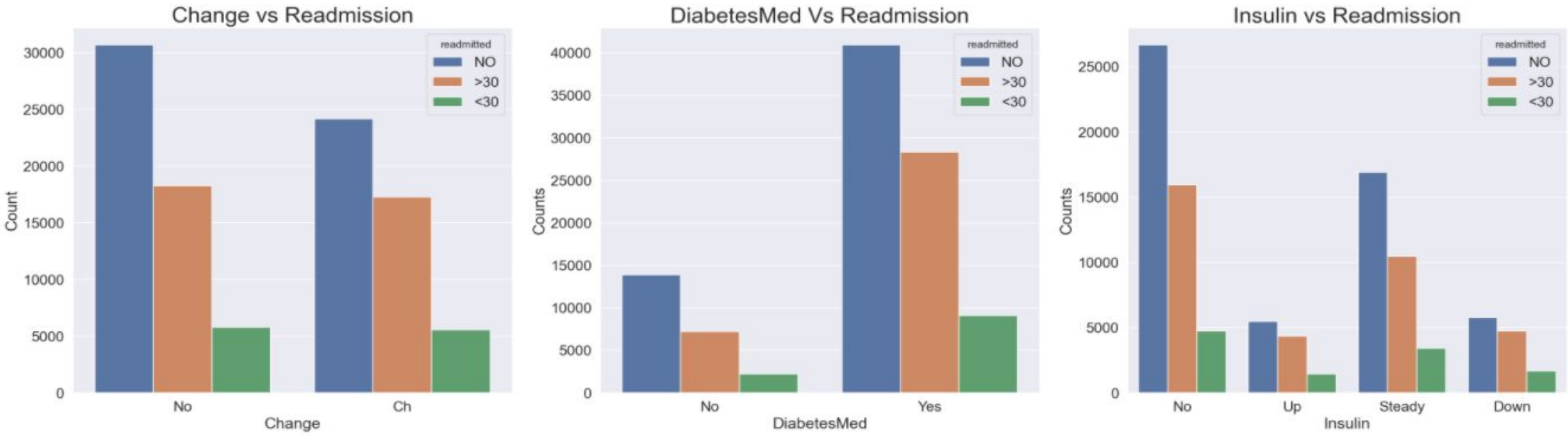






- High readmission for significant number of patients for which the Max_glu_Serum and A1C were not measured.
- 23 Diabetes medication in the dataset - None stood out as being used over another.

Data Visualizations (Cont'd)



- Significant number of patients on diabetes medication than those who are not.
- Hospital readmission higher amongst the patients on diabetes medication than those not on the medication.
- Readmission rate appears to be the same for patients who are changing medication and those who are not.

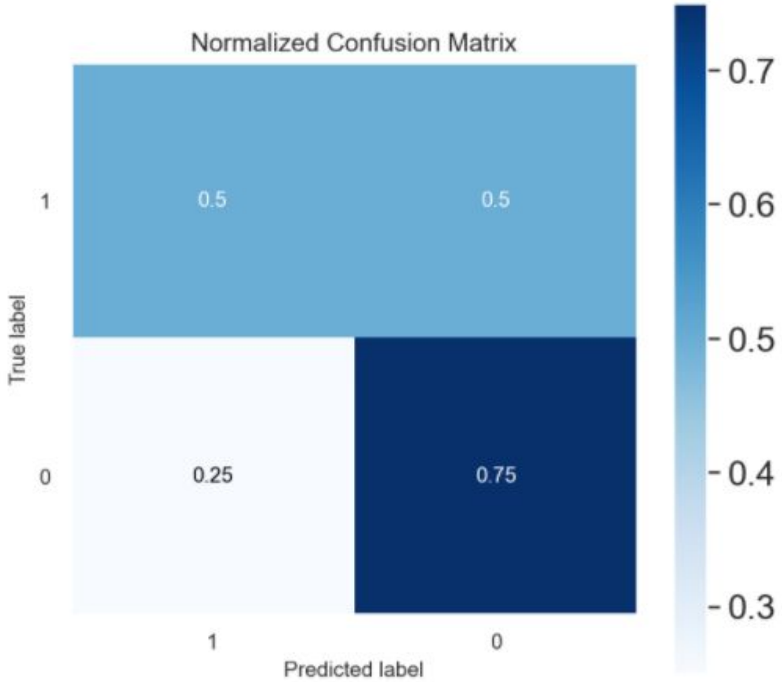
Modelling - Results

Target Variable - Readmitted

Baseline
0 - No readmission - 0.5391
1 - Readmission - 0.4609

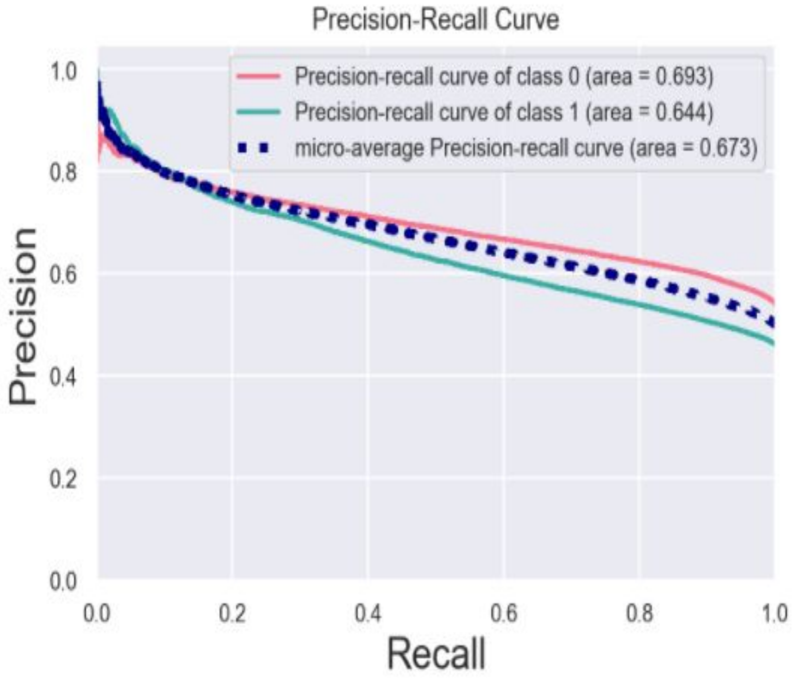
Model	Best Model Params	Mean CV Score	Training Score	Test Score
Logistic Regression	C' : 0.215443, Penalty: l2, Solver: 'liblinear'	0.62272	0.62445	0.62209
Decision Tree Classifier	alpha:0, max_depth: 5, max_features: None, min_samples_split: 25	0.61991	0.62263	0.61971
Adaboost Classifier	max_depth: 5, N_estimators: 10, algorithm: 'SAMME'	0.62198	0.62657	0.62318
Gradientboosting Classifier	max_depth: 5, N_estimators: 10	0.62161	0.62413	0.62052
Random Forest Classifier	max_depth: 10, max_leaf nodes:20,	0.62265	0.62418	0.61817
Linear SVC Classifier	C: 1	0.61934	0.62028	0.61889
MLP Classifier	appha: 0.2154435, hidden_layer_sizes: 42, solver: adam	$\boxed{0.62415}$	0.63238	0.62606

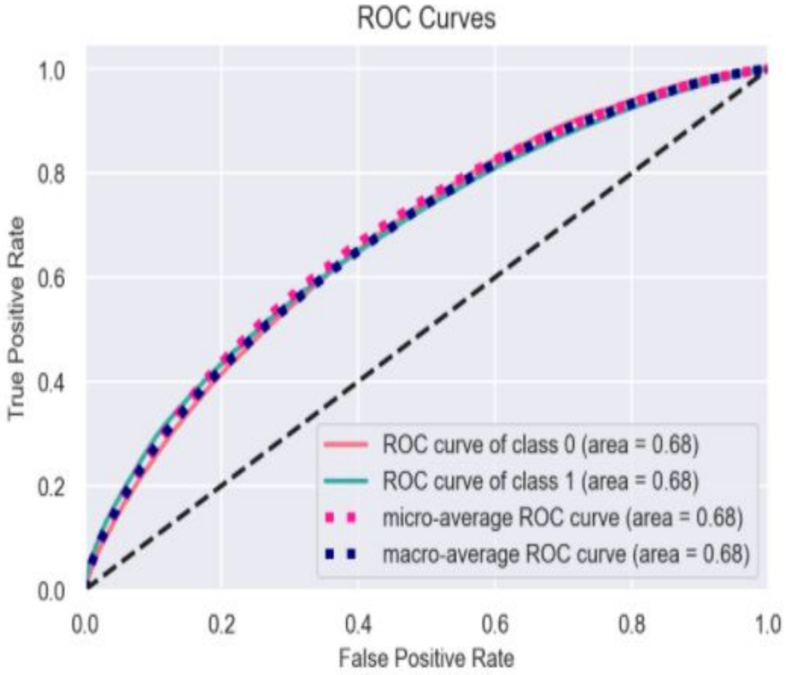
Neural Networks - Confusion Matrix & Classification Report



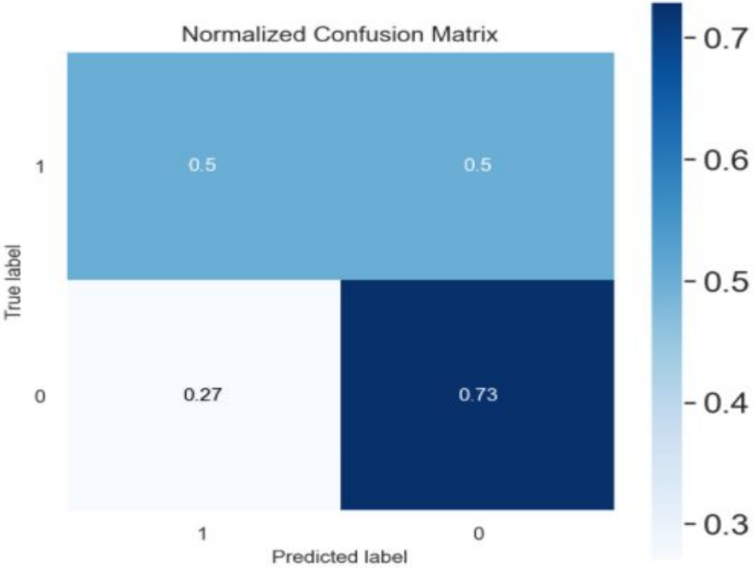
	precision	recall	f1-score	support
0	0.63	0.75	0.69	38403
1	0.63	0.50	0.55	32831
accuracy			0.63	71234
macro avg	0.63	0.62	0.62	71234
weighted avg	0.63	0.63	0.63	71234

Neural Networks - Precision-Recall & ROC Curves



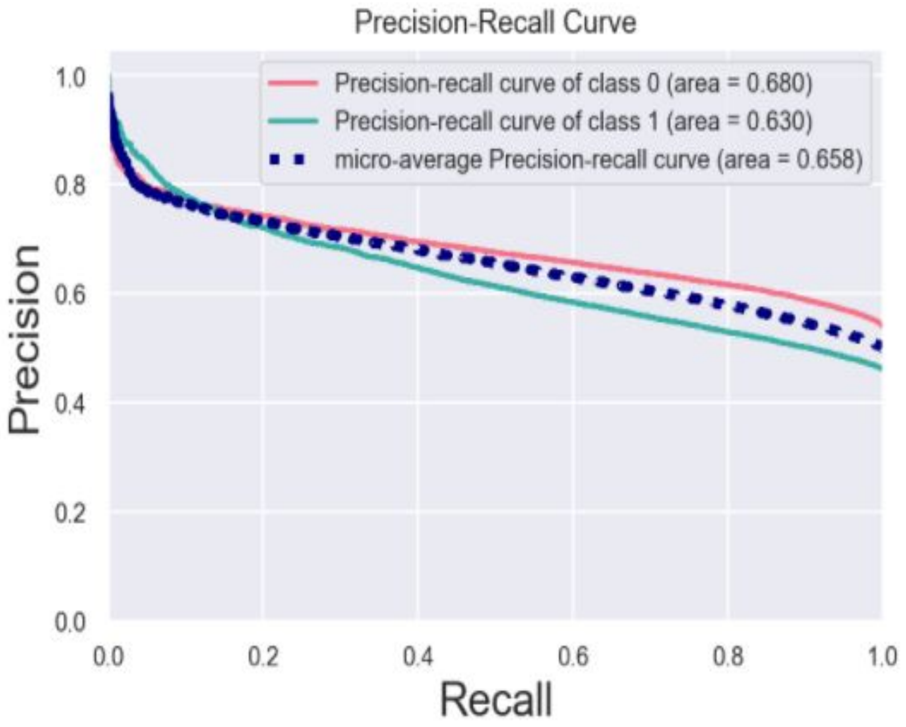


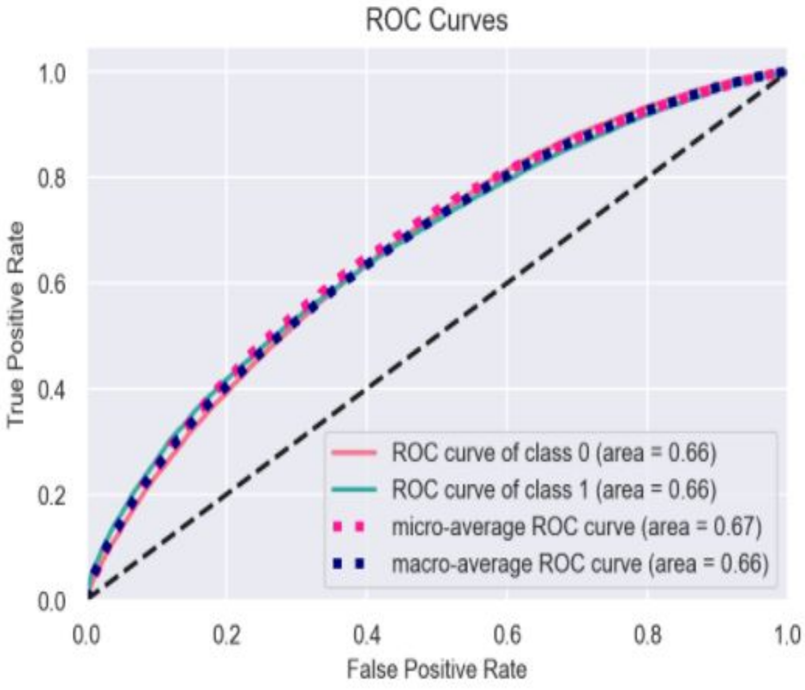
Logistic Regression - Confusion Matrix & Classification Report



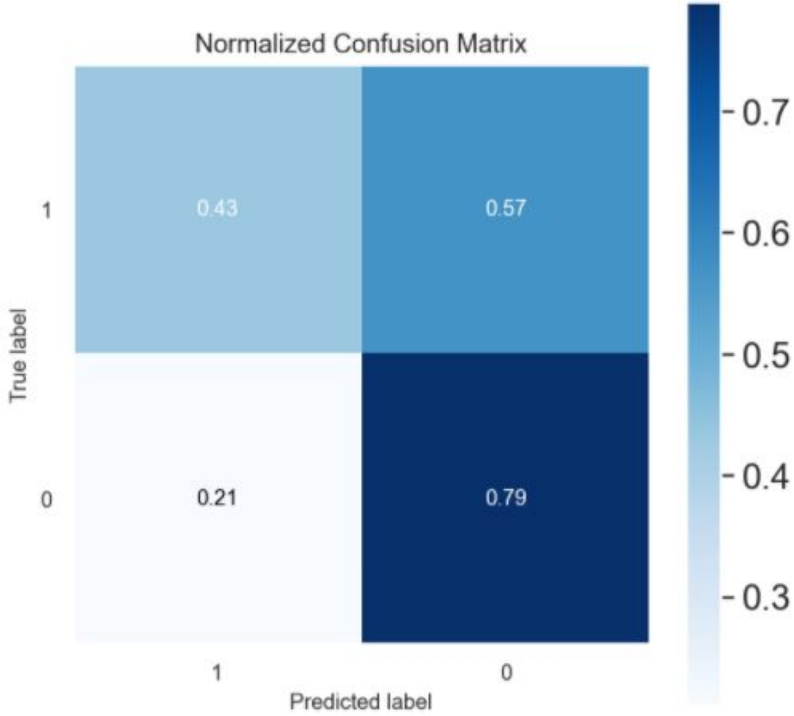
	precision	recall	f1-score	support
0	0.63	0.73	0.68	38403
1	0.61	0.50	0.55	32831
accuracy			0.62	71234
macro avg	0.62	0.62	0.61	71234
weighted avg	0.62	0.62	0.62	71234

Logistic Regression - Precision-Recall & ROC Curves



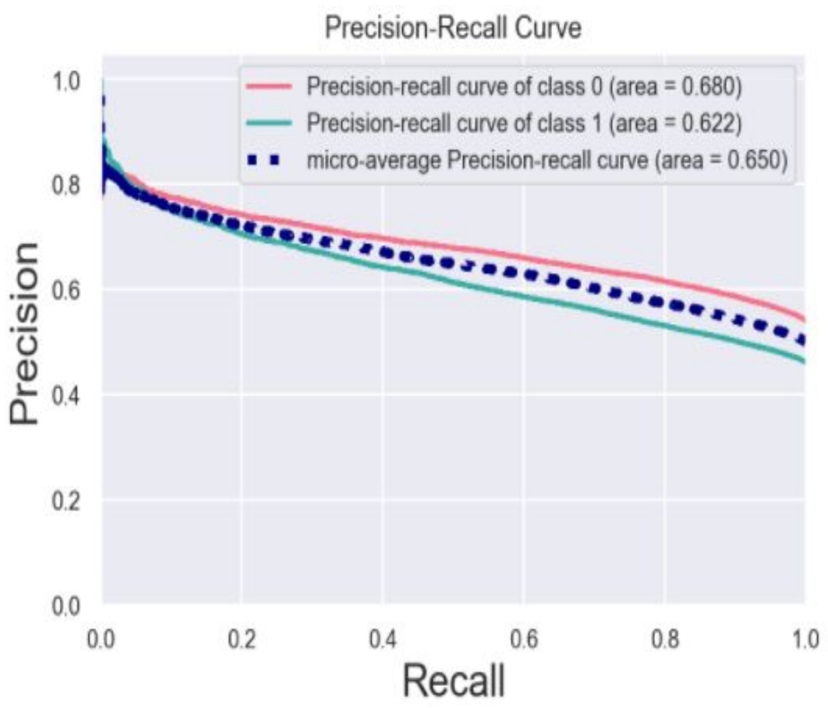


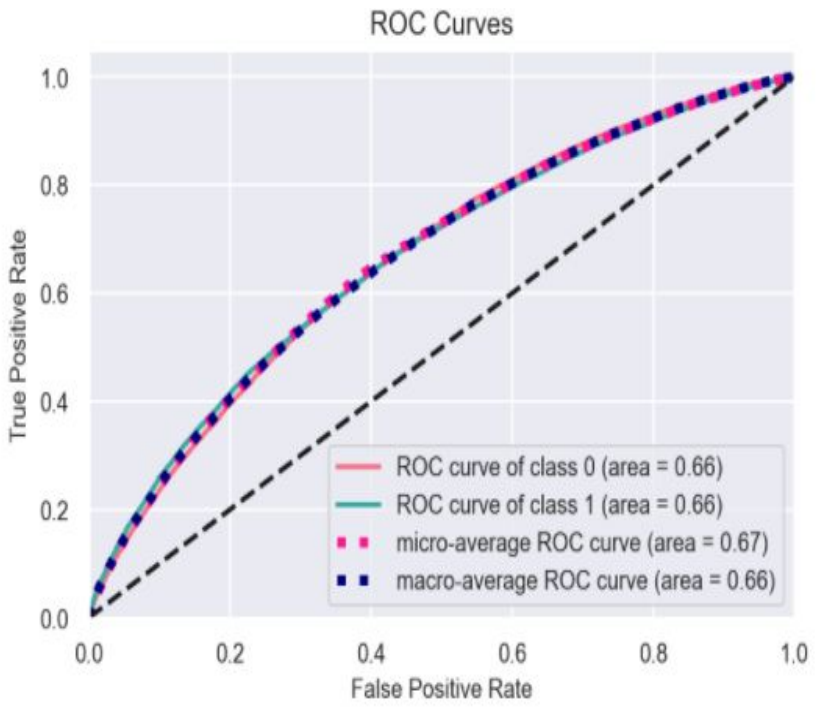
Random Forest - Confusion Matrix & Classification Report



	precision	recall	f1-score	support
0	0.62	0.79	0.69	38403
1	0.64	0.43	0.51	32831
accuracy			0.62	71234
macro avg	0.63	0.61	0.60	71234
weighted avg	0.63	0.62	0.61	71234

Random Forest - Precision-Recall & ROC Curves

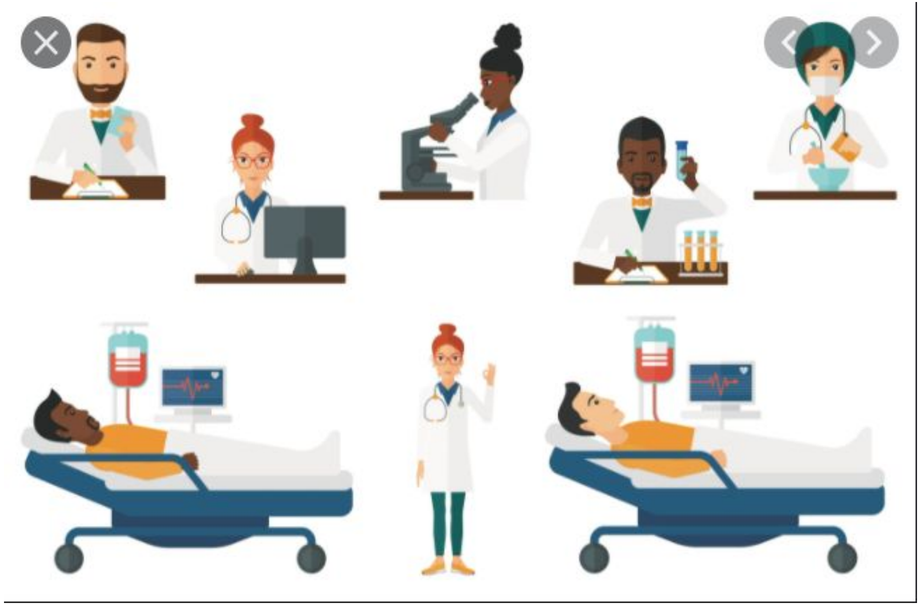




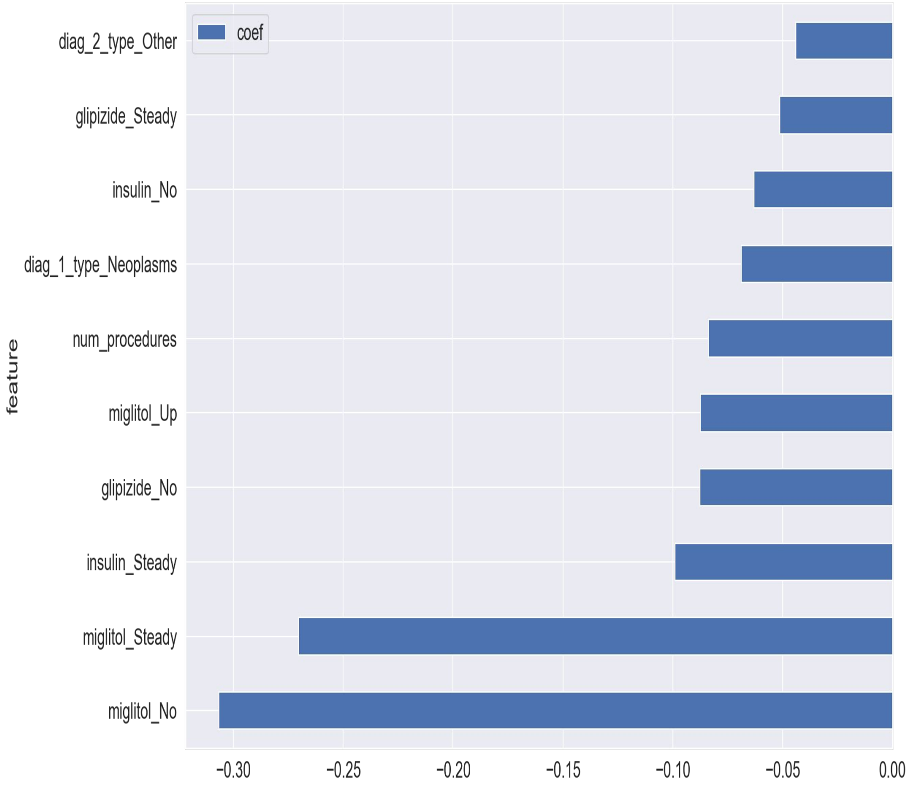
Summary Results

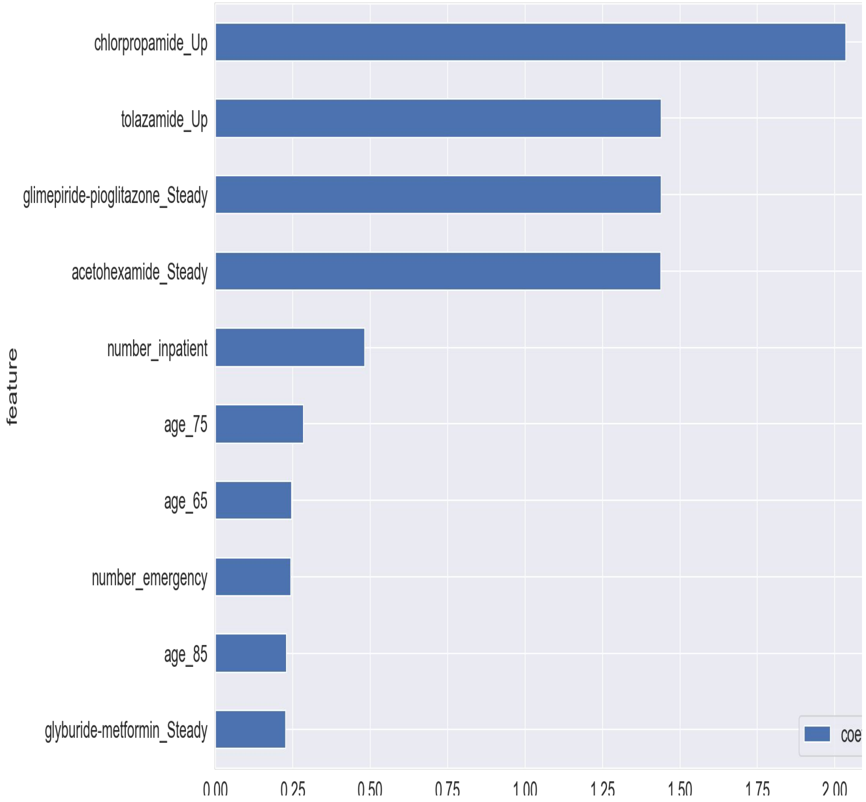
Model	Precision	Recall	Precision-recall curve	ROC Curve
Logistic Regression	0.61	0.5	0.63	0.66
Decision Tree Classifier	0.61	0.5	0.606	0.65
Adaboost Classifier	0.61	0.53	0.633	0.67
Gradientboosting Classifier	0.61	0.5	0.63	0.66
Random Forest Classifier	0.64	0.43	0.625	0.67
Linear SVC Classifier	0.64	0.41		
MLP Classifier	0.63	0.5	0.644	0.68

Using data science to help reduce hospital admissions

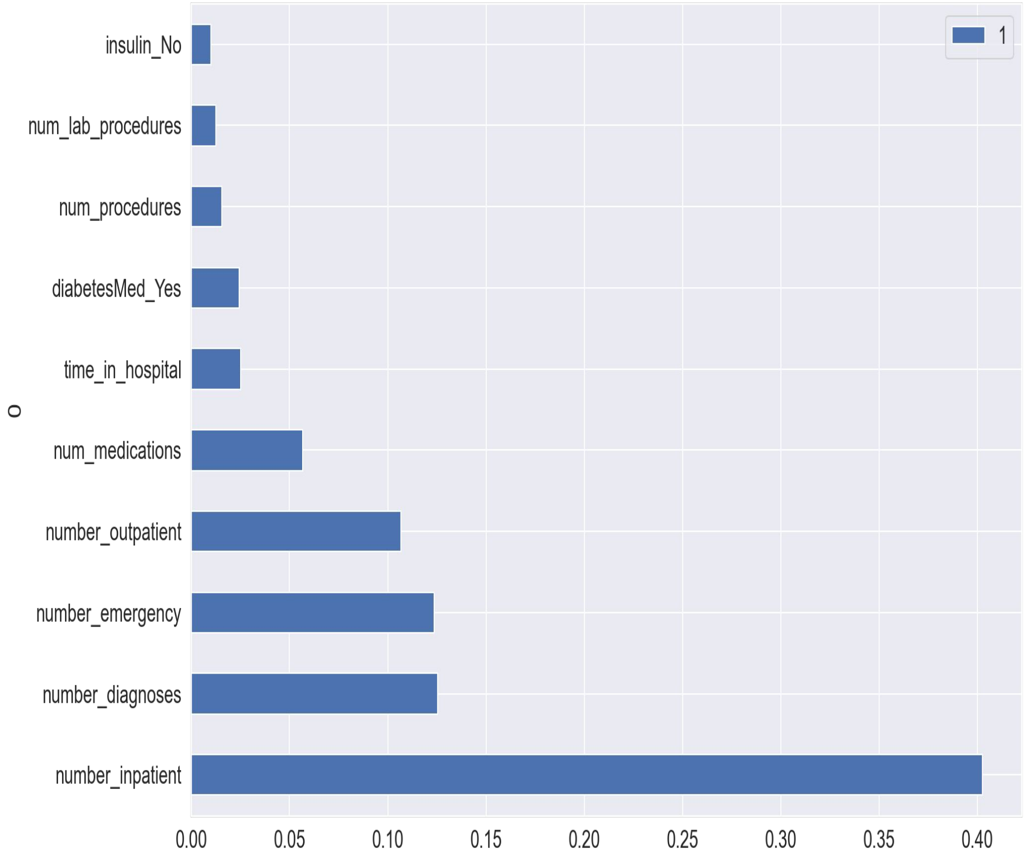


Logistic Regression - Drivers of Readmission





Random Forest Model - Drivers of Readmission



Limitations

- Old Dataset, and as medicine is constantly changing, new developments in patient management has resulted in the management of hospital readmissions.
- Insufficient information on patient medical history - readmission could result from other underlying health issues.
- Patients respond differently to the same medication. No medication history provided.
- This is an insurance claim dataset and is lacking compared to the different from the metrics the hospital would hold.

Recommendations

- Ensure that the Max_Glu_Serum and A1C test are carried out for all patients diagnosed with diabetes as this will identify the extent of the diabetes and the treatment can then be modified to the patients propensity towards the illness.
- Patients aged 55 and over should be managed in the primary care facilities and where necessary prioritised for home care visits.
- Medication management should be carried out to ensure that the medications prescribed are working effectively together and are benefitting the patient.

Takeaways

- Models took long time to fit……..had to drop some of the initial parameters to get results out.
- Need a develop a clear understanding of how to adjust the different parameters to reduce code run time whilst at the same time get good scores out.
- Difficult to tune parameters when you have a time constraint with a model that takes long to fit……..was just happy to be able to get results
- Improving the model needs to be done………………..to be tackled in "What's next"

What’s Next?

- Improve the predictive power of the model by removing variables that have no impact on readmission
- Fitting other models using regression technique
- Use the three classes of readmissions and use clustering to predict readmission
- It would be interesting to get a current UK dataset to carry out a similar prediction and see how it compares.

Thank You.

Questions?