greatlearning

Prediction of Early Hospital Readmission of Diabetic patients

Batch : Group 2 Online MARCH - 2022

Team Details

1. H H Arvind
2. Ram Krishna Namdev
3. Yerrabachu Keerthy Rao
4. Prangi Sharma
5. Tarun Tadikonda

Mentor

Mr. Jayveer Nanda

Introduction and Importance of the problem

Hospitals engaging in any model are likely to face penalties if their providers cannot improve hospital readmission rates. In recent years, government agencies and healthcare systems are increasingly focused on 30-day readmission rates as a way to improve quality Health care.

To avoid value-based penalties readmission rates Hospitals should reduce early readmission by identifying the Diabetic patients who are having high probability of compared to other patients who do not have Diabetes.

As cost of inpatient care & readmission rates are higher in patients with Diabetes Mellitus (DM) compared to other diagnosis focusing on reducing early readmissions, cost of readmissions to avoid value-based penalties from the government is primary research goal.

greatlearning

Business objective

The main objective of our work is to come up with the predictive model which helps Hospital Management systems to predict the risk of early readmission of patients who are having Diabetes Mellitus which can further address

- Enhanced patient care, patient Engagement, glucose monitoring, Transitional care & Post discharge follow up

- Reducing cost of early readmission there by reducing penalties to the Hospital which increases its commercial value in terms of reputation in health care & escape value based penalty.

greatlearning

Data set information

'Diabetes 130-US hospitals for years 1999-2008 Data Set'

To achieve our business objective we have considered the Analysis of a large clinical database which was undertaken to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital during the period 10 year period (1999 – 2008) . The dataset has been retrieved from UCI Machine learning Repository which is related to Diabetic inpatient encounters in 130 Hospital across United States.

- ✓ Dataset contains 101766 unique encounters corresponding to more than 70000 patients
- ✓ Dataset has 50 features including target label related to diabetes of which
  - ❑ Numerical features – 8
  - ❑ Categorical features – 41 (23 medicines + 18 patient information)
  - ❑ Target

Analysis of Target Variable

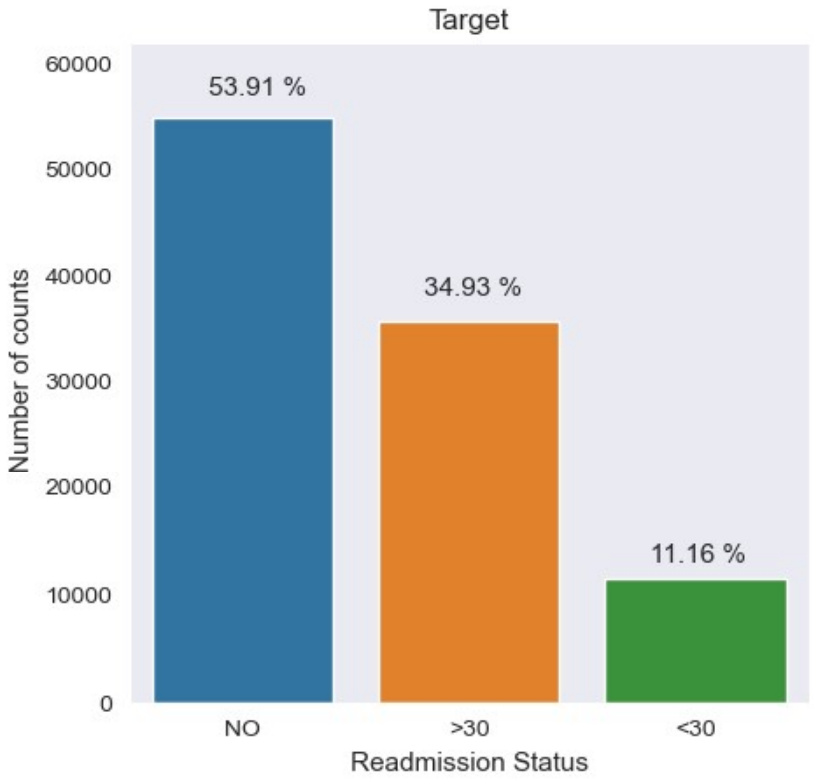
- ✓ 'Readmitted' is the target variable and it is of categorical (Nominal) data type. Hence our problem is a classification problem

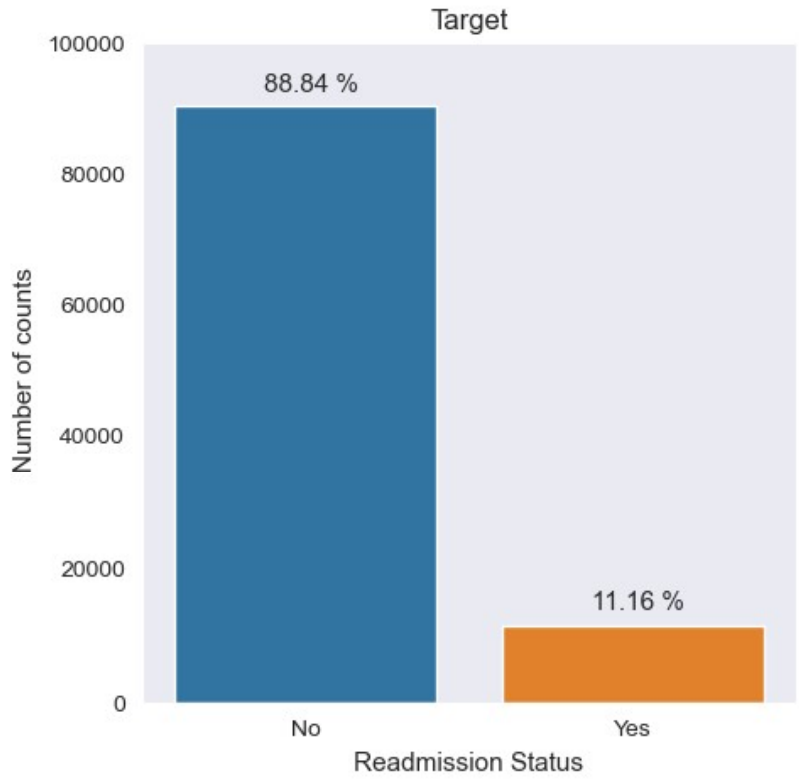
- ✓ There are 3 levels in the Target variable **:**
  - ❑ No : No Readmission
  - ❑ < 30 Days : Readmission within a month
  - ❑ > 30 Days : Readmission after a month

- ✓ A data imbalance is observed for the target variable which might effect the accuracy of the model.
- ✓ Since our business objective is to predict 'Early Hospital Readmissions' preferably (<30 days). we will redefine Multi – class Target into Binary Target.

greatlearning

Redefining Target label (Binary classification)





✓Categorized the target variable into two levels due to class imbalance**:**

- ❑ 0 : Not Readmitted (NO or > 30 days) : 'No'
- ❑ 1 : Readmitted (< 30 days) : 'Yes'

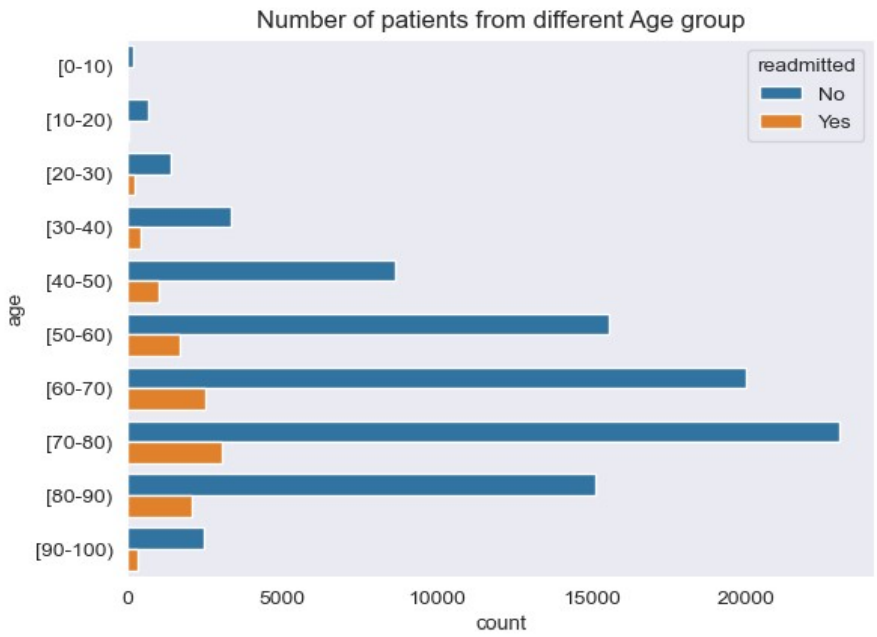
Exploratory Data Analysis

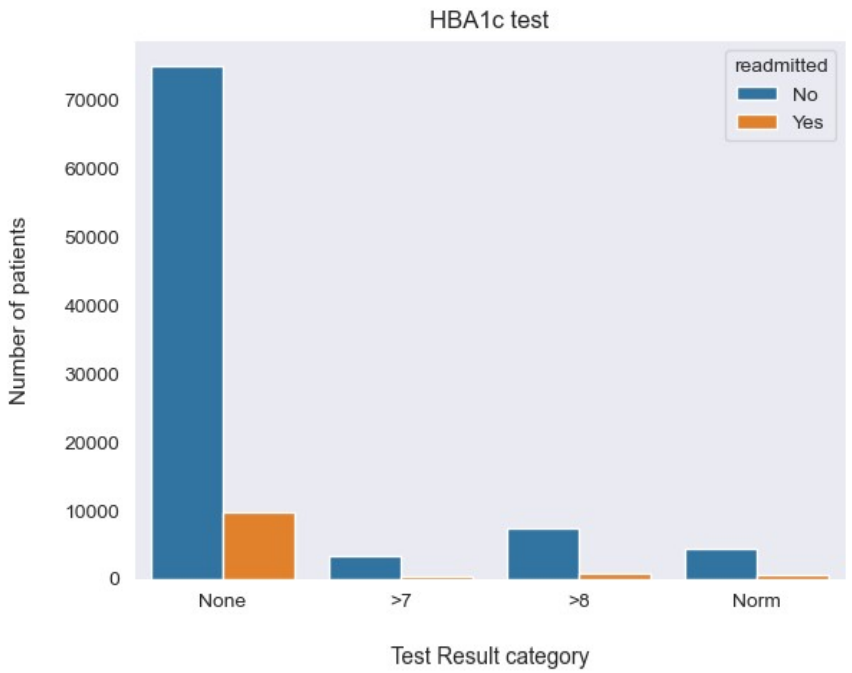
• **Gender :** As 'Unknown/Invalid' are only 3 observations, so we have imputed using mode.

• **Diagnosis :** In diagnosis columns,first three digits are coded from ICD9, so from research study we categorized the data into description using ICD9 codes.

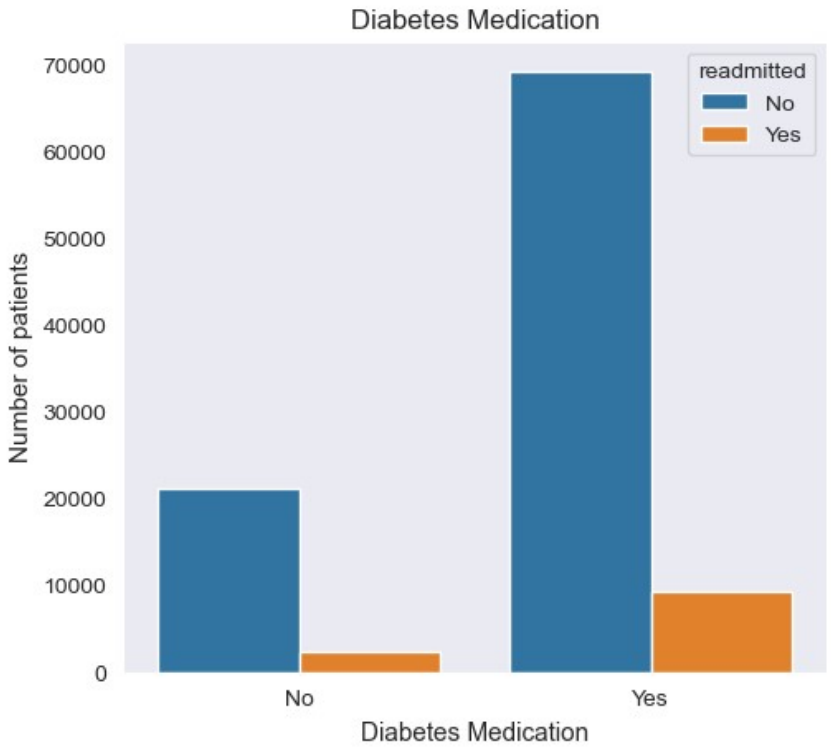
Variable	%
weight	96.85%
Medical specialty	49.08%
Payer code	39.55%
Race	2.23 %
Diagnosis 1	0.02 %
Diagnosis 2	0.35 %
Diagnosis 3	1.40 %

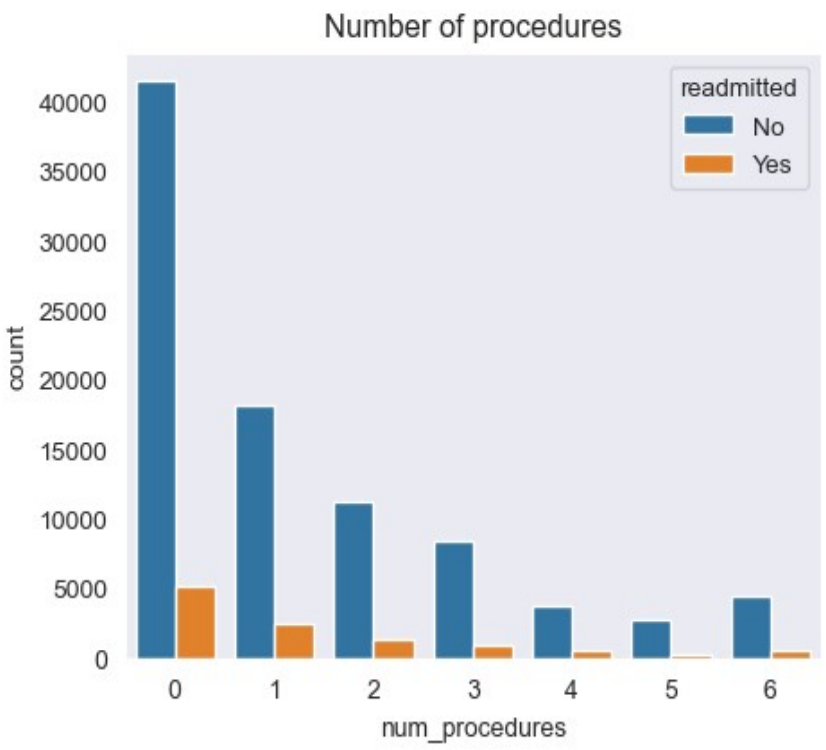
- **Medical specialty** is an important information related to Readmissions, we have considered missing values as it as 'Not mentioned' category
- **Race, Diagnosis1, Diagnosis2 & Diagnosis3** are being imputed using Iterative imputer (KNN Algorithm) using only numeric values as features which found to significant with Target (Features which contain missing values)
- We dropped weight which had more than 90% missing values & Payer code which does not give any information regarding readmissions.





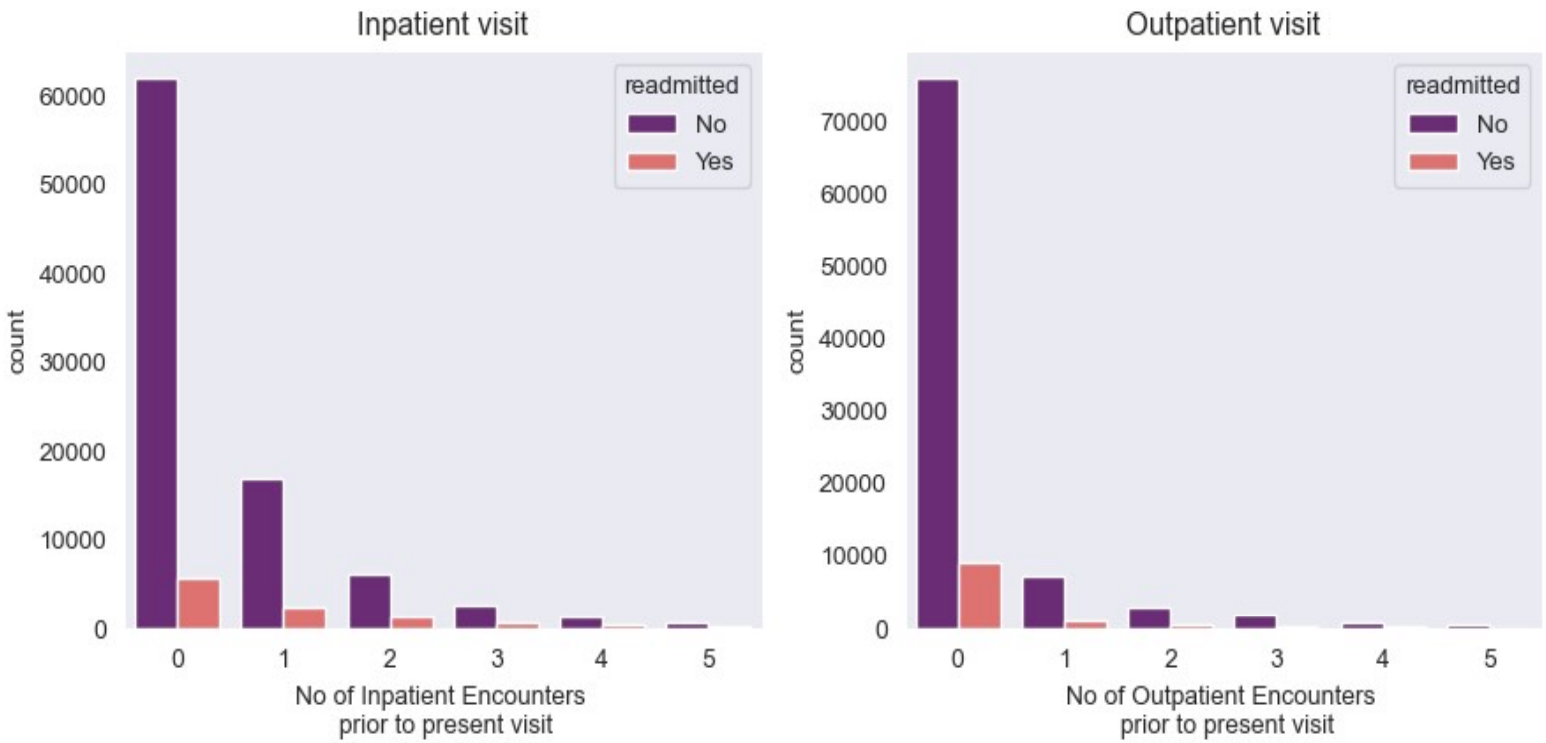
- ✓ Most number of patients early readmissions are in the age range of 40-100
- ✓ People who are between age 70-80 have highest patient encounter as well as re-admission rate. As the age increases the Early readmission cases increases, decreases after 70-80
- ✓ from above observation it shows that people who have not undergone HBA1c test, they have high chance of early readmission
- ✓also compared to '>7','Norm' the people who have '>8' HBA1c result have slightly high early readmissions.





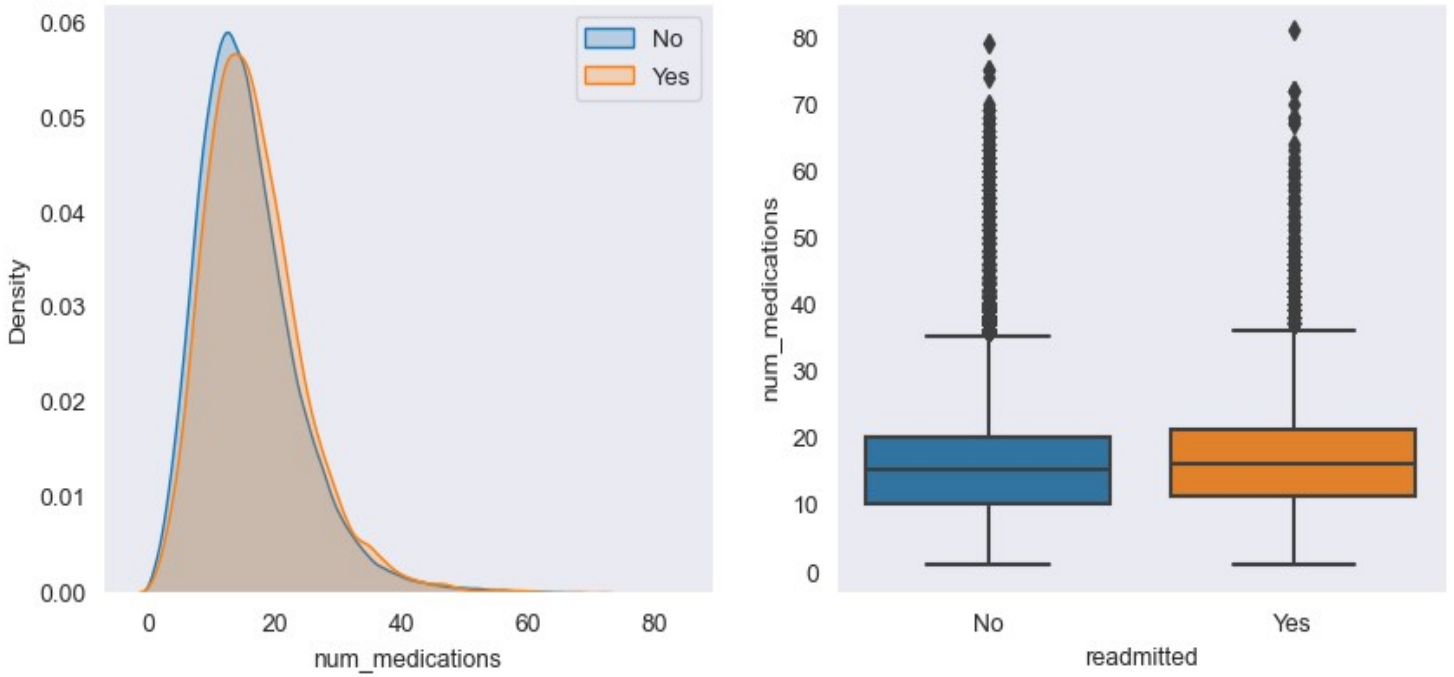
- ✓ From above plot we can say that patients who are taking diabetic medication have higher chance of early readmission which has relevance in our business objective.

- ✓ Early Readmissions in case of Less number of procedures is more compared to more number of procedures



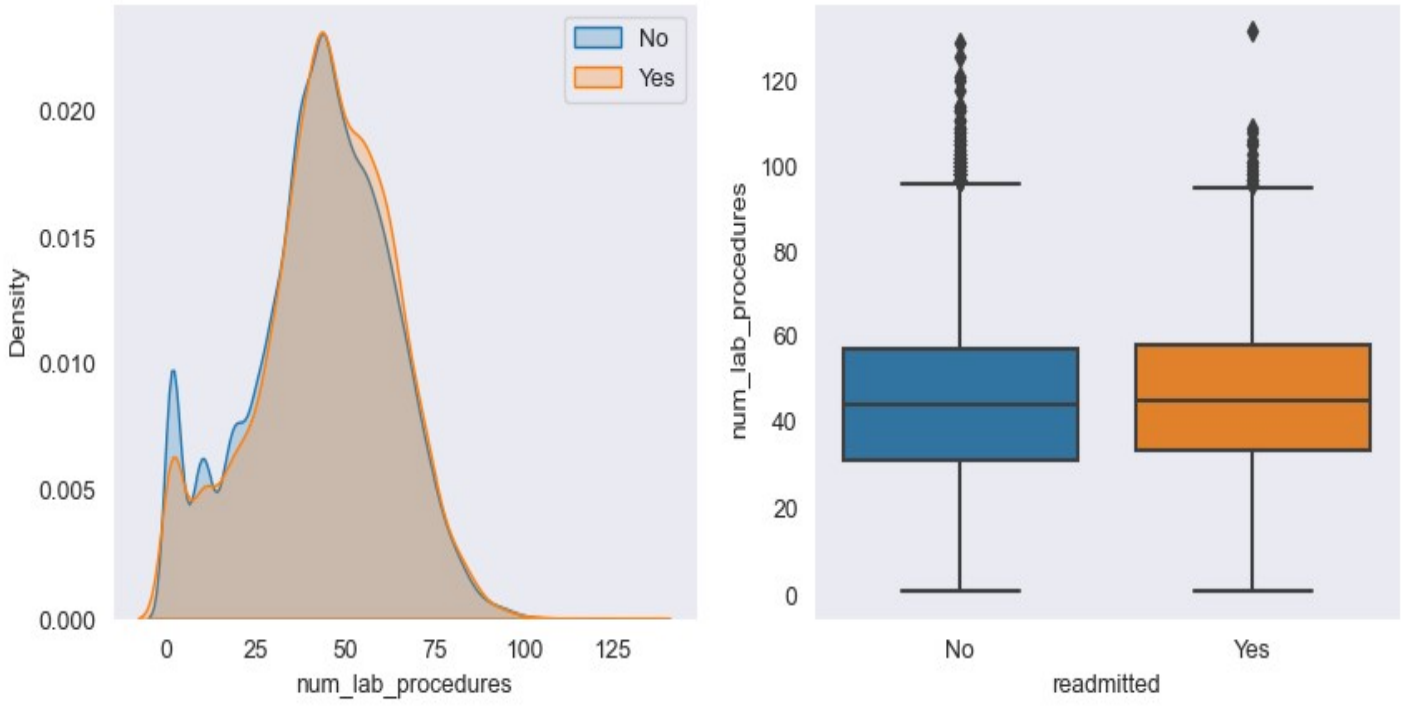
- ✓ From above plot we can say that patients who have no previous visits to Hospitals prior to present encounter they have high chance of readmission
- ✓ Its obvious that as the previous inpatient/outpatient encounters increased, the Early readmissions has been decreased.
- ✓ It means that the risk of readmission decreased over the time with frequent visits.

Number of Medications



- ✓ Both the categories (Readmission & No-readmission) follows the same distribution with extreme values at the right side of distribution.
- ✓ On an average every patient is under 15 medications

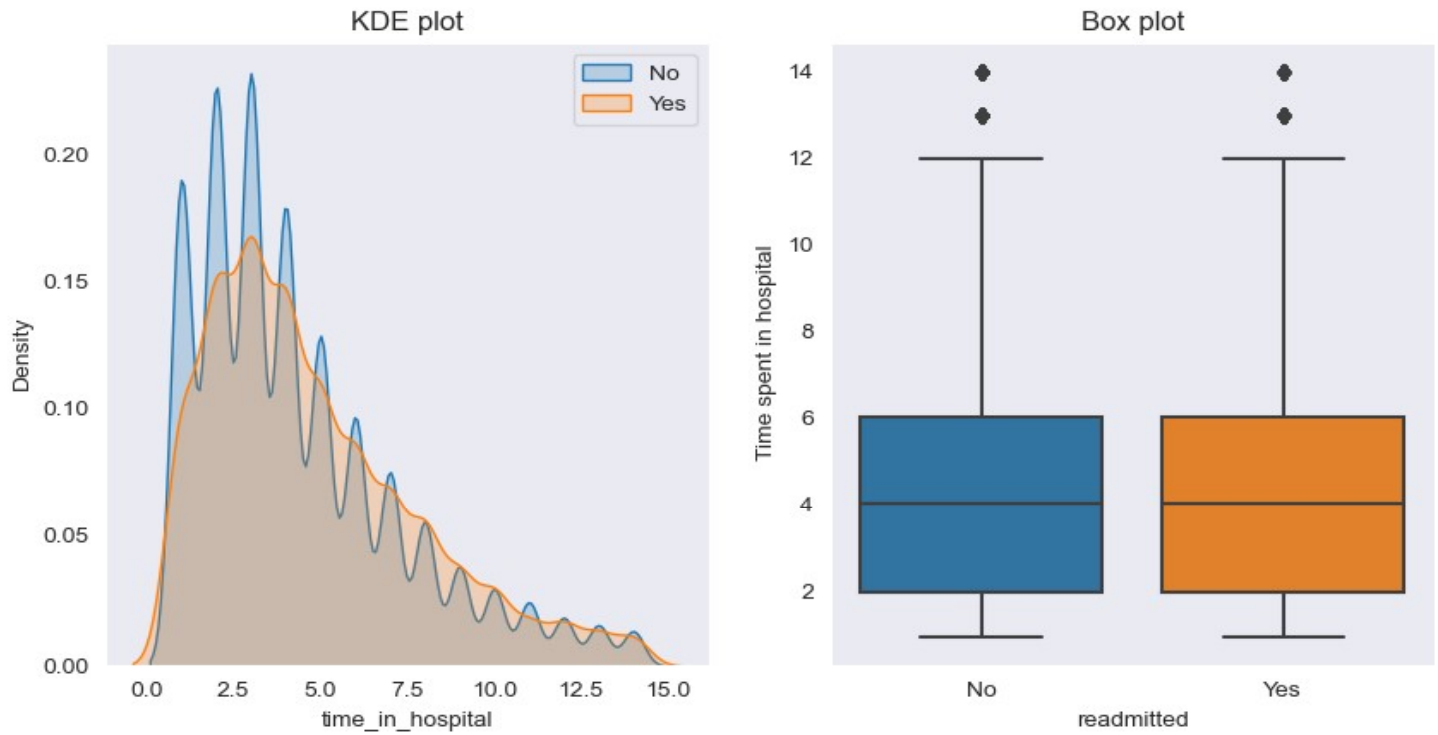
Number of Lab procedures



- ✓ Average lab procedures done for patient in both categories is same i.e. 40 lab procedures

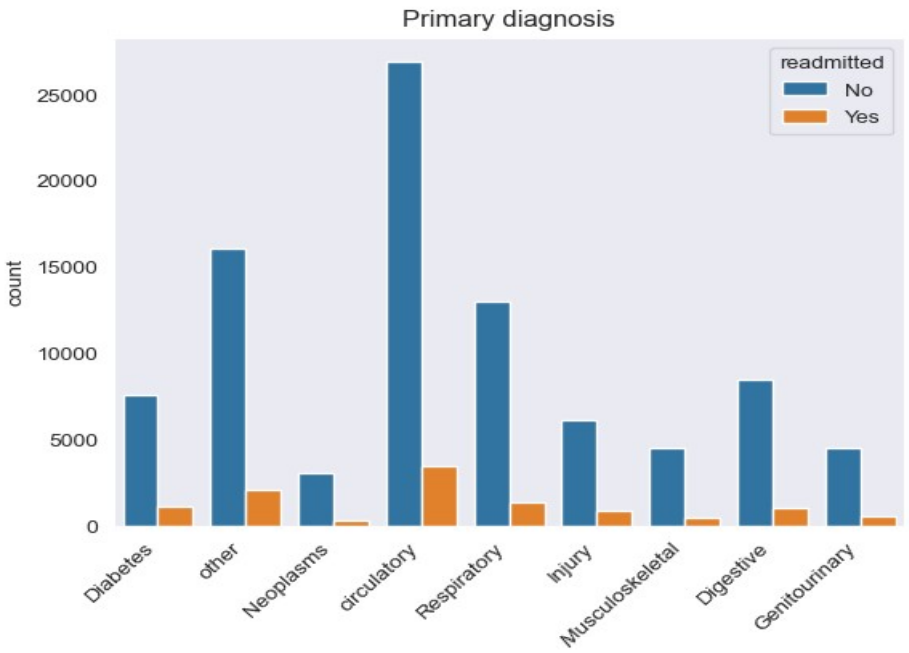
greatlearning

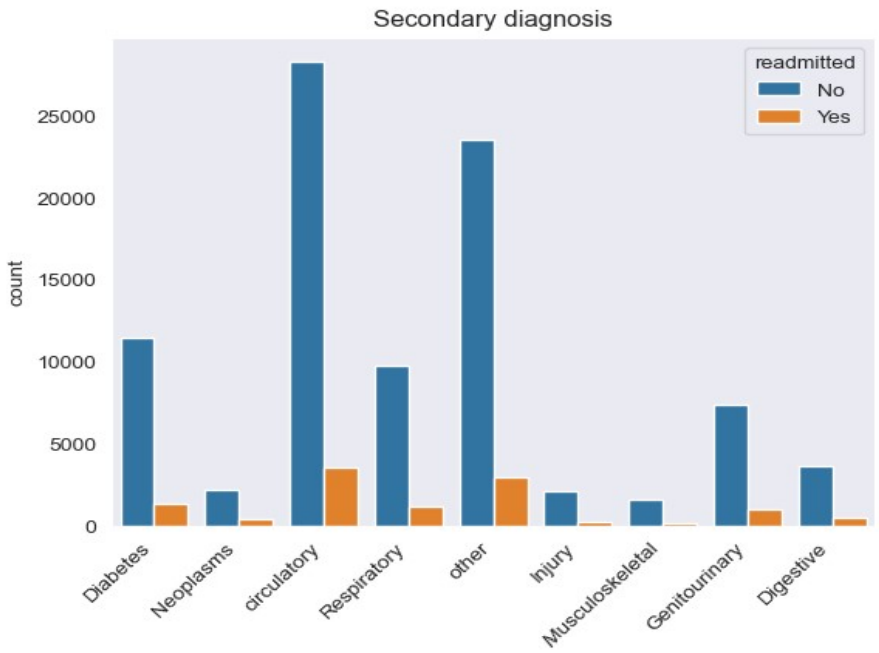
Time spent in hospital

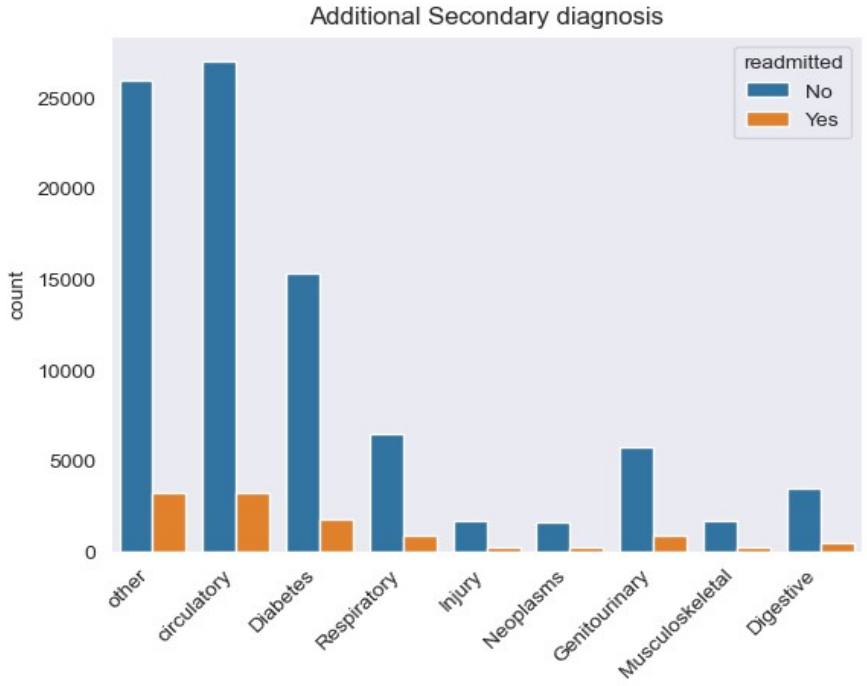


- ✓ Average time spent by patient in both categories is same i.e. 4 days
- ✓ Both the distributions are rightly skewed which shows higher time spent in hospital

greatlearning

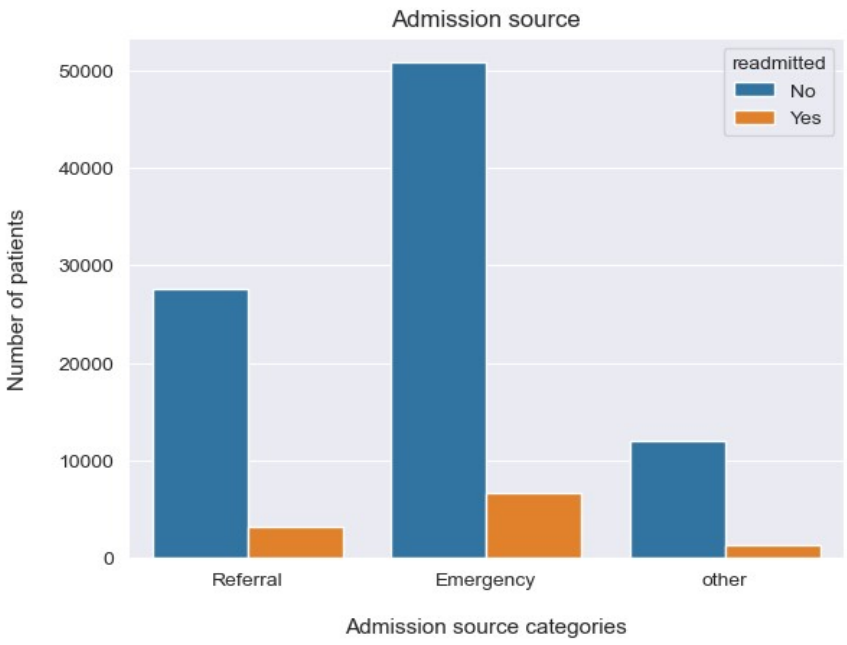


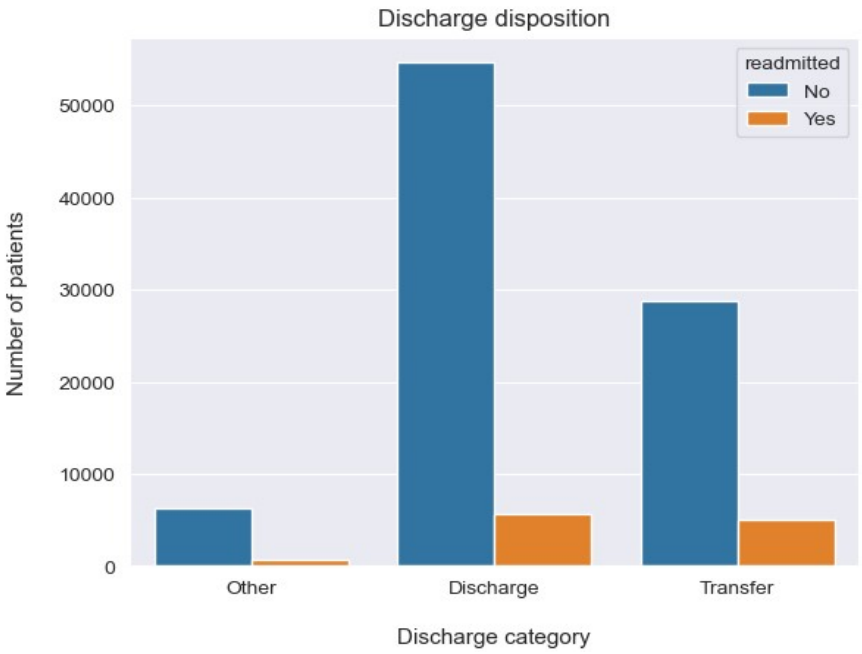


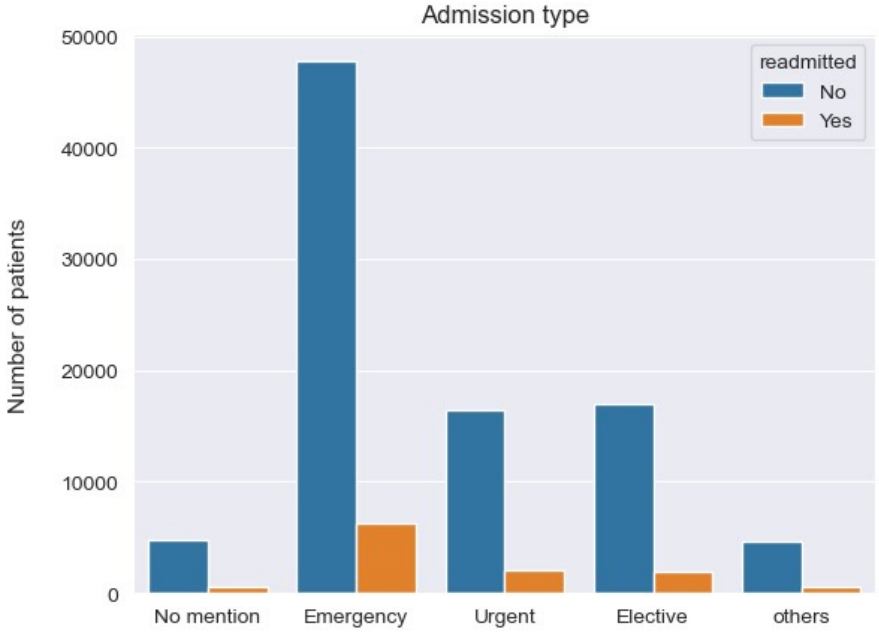


✓ All diagnoses contain more than 800 unique values which are categorized into 9 major categories using ICD-9 codes

✓ Diabetes, Circulatory & Respiratory diagnosis dominate Early readmissions.



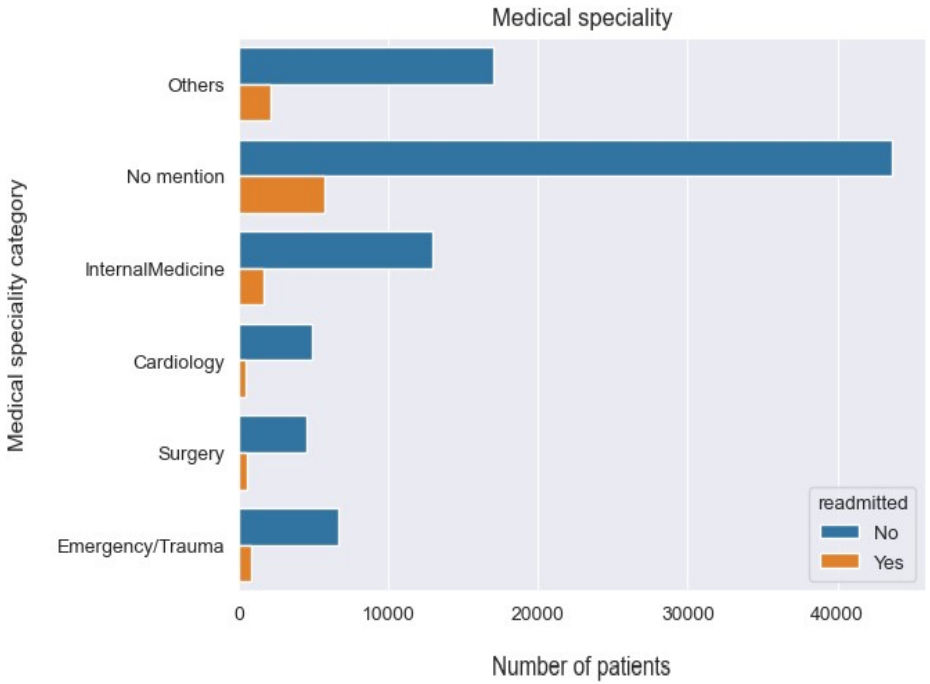




✓ Emergency & Referral categories in Admission sources have high readmissions

✓ Discharge, Transfer have high early readmissions in Discharge categories

✓ Emergency & Urgent Admission types have high early readmissions

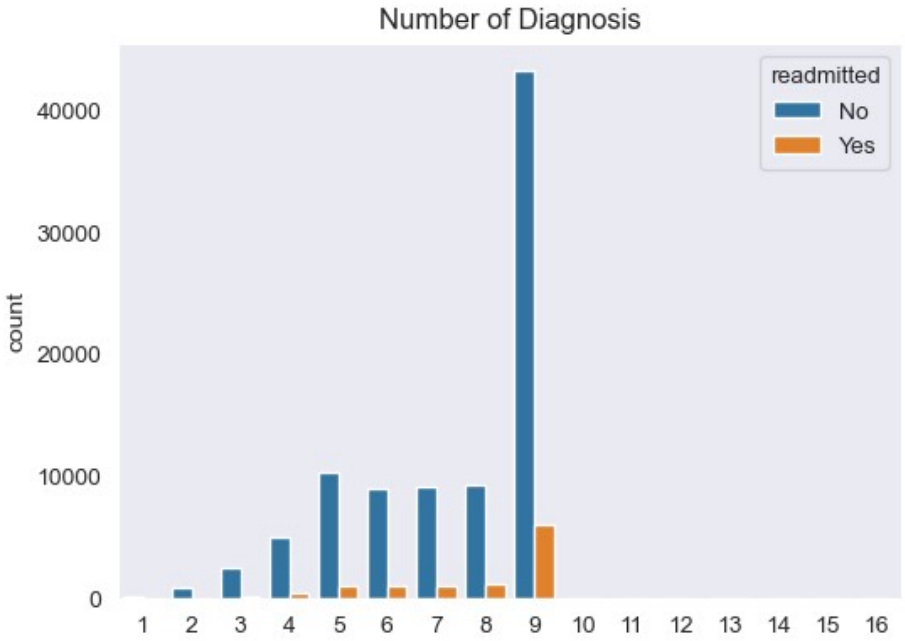


✓ No mention category (Missing values) have high readmissions which tells suspicious entries.

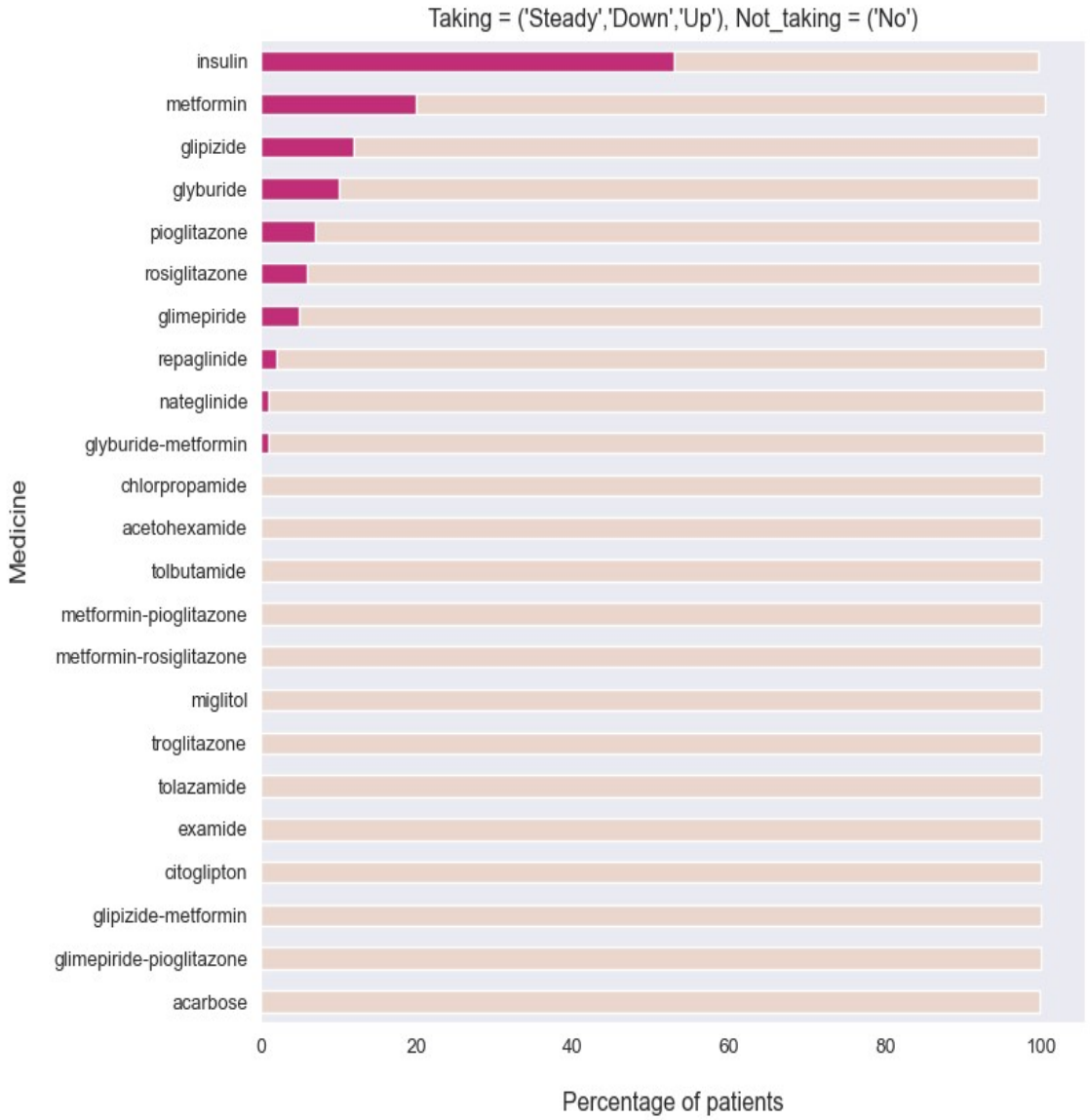
✓ Reasons which are unknown & Internal medicines have considerable amount of readmissions

✓ Patients who have been diagnosed with more illness have high early readmission rates.

✓ Patients with 8 or more diagnosis have high readmissions.



Diabetes Medications



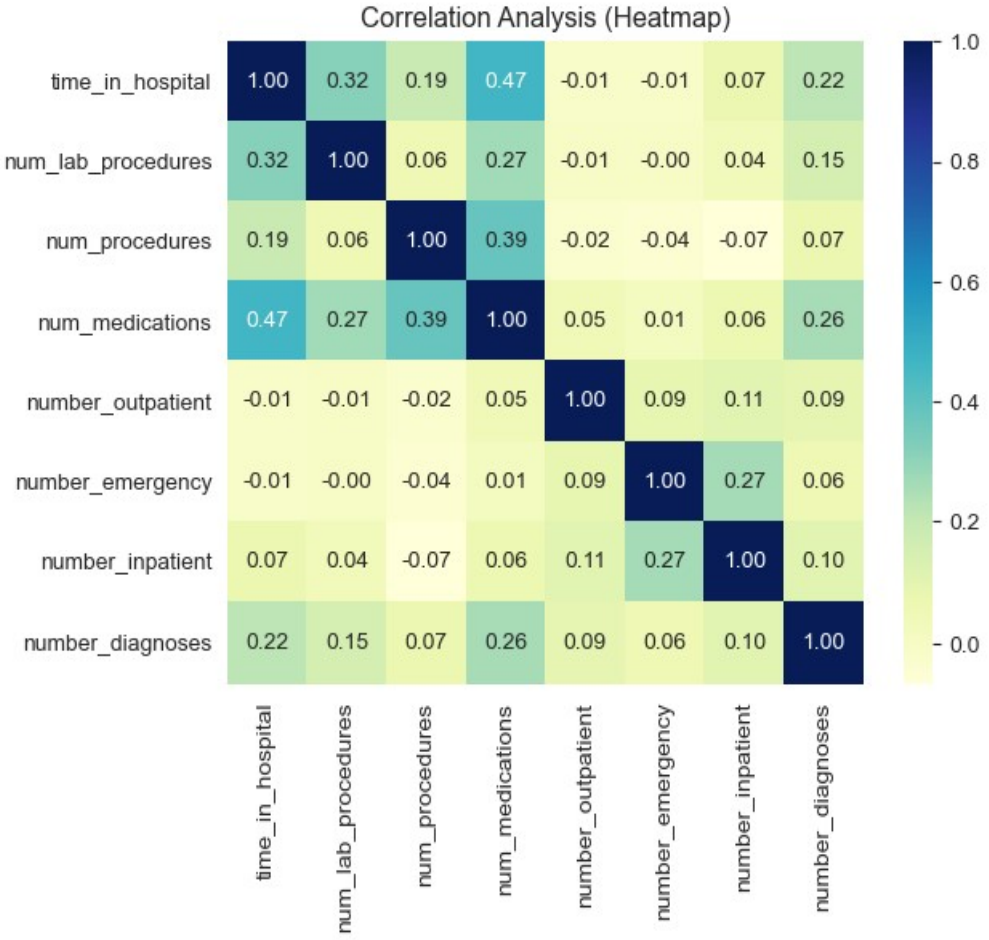
✓ only 10 out of 23 medicines are seen as relevant, other 13 medications have very skewed levels which may or may not provide any information (Rare medicines)

✓ Insulin, Metformin are the major medication prescribed to patients which are taken by more than 70 % of patients

	Not taking	Steady	Up	Down
metformin	81778	18346	1067	575
repaglinide	100227	1384	110	45
nateglinide	101063	668	24	11
chlorpropamide	101680	79	6	1
glimepiride	96575	4670	327	194
acetohexamide	101765	1	0	0
glipizide	89080	11356	770	560
glyburide	91116	9274	812	564
tolbutamide	101743	23	0	0
pioglitazone	94438	6976	234	118
rosiglitazone	95401	6100	178	87
acarbose	101458	295	10	3
miglitol	101728	31	2	5
troglitazone	101763	3	0	0
tolazamide	101727	38	1	0
examide	101766	0	0	0
citoglipton	101766	0	0	0
insulin	47383	30849	11316	12218
glyburide-metformin	101060	692	8	6
glipizide-metformin	101753	13	0	0
glimepiride-pioglitazone	101765	1	0	0
metformin-rosiglitazone	101764	2	0	0
metformin-pioglitazone	101765	1	0	0

- ✓ The medicines which are highlighted green in colour will be dropped in our Analysis due to skewness
- ✓ There are 23 medicines in total out of them we are considering 16 features
- ✓ we have also dropped patients who are expired, patients who are sent to hospice facilities who have no or very less chance of readmissions.

Correlation Analysis



✓ From Correlation Analysis we have no much or severe multi collinearity present in our data

✓ The highest multi- collinearity present in our data is between Number of medications and lab procedures with the correlation of 0.47

✓ From this we can say multi-collinearity will not affect much in linear classification models.

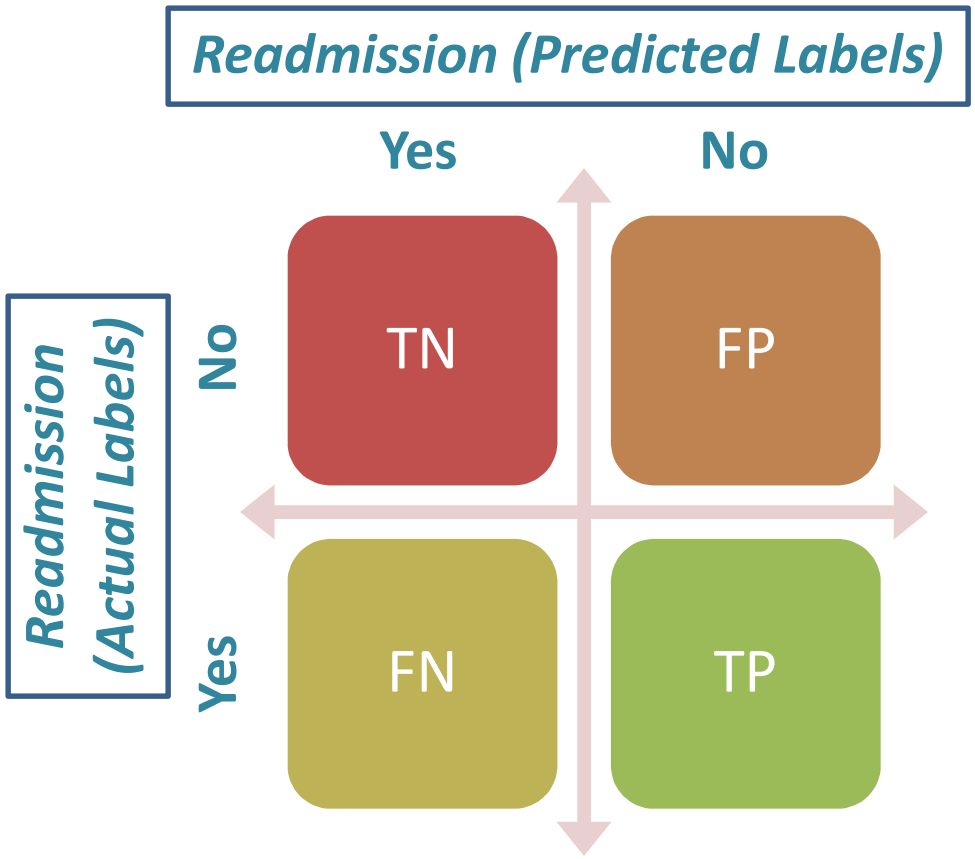
Evaluation Metrics

An evaluation metric which quantifies the performance of a predictive model related to our business objective.

In our case, if readmitted patient is predicted as non-admitted then there is a risk of patient life.

It will leads to hospital law suits, loss of reputation, huge business loss so our main target is to decrease the false negatives which can be done by using

- ✓ **RECALL**
- ✓ **F1 SCORE**



$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

Challenges and solutions

- class imbalance

  Solutions for class imbalance which reduces the bias of model

  - ✓ Synthetic Minority oversampling technique (SMOTENC)
  - ✓ SMOTENC + Tomek link under sampling
  - ✓ Cost-Sensitive Algorithms

- Encoding categorical variables for oversampling & different Algorithms

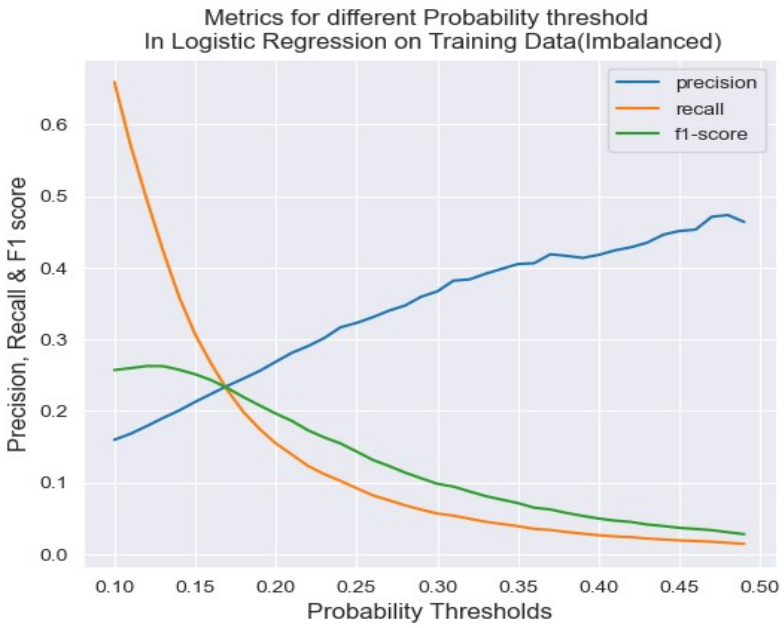
  - ✓ *logistic regression model* –
    Dummy encoding categorical features (dropping first level) for oversampling & scaling numerical features
  - ✓ *Tree based Models, Ensemble techniques & Naïve Bayes algorithm* –
    Dummy encoding categorical features (without dropping first level) for over sampling & applying algorithm

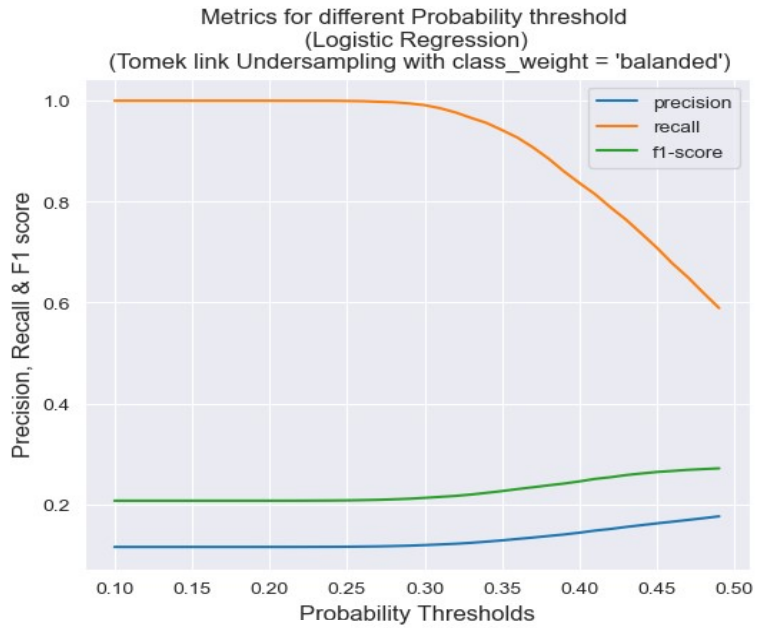
greatlearning

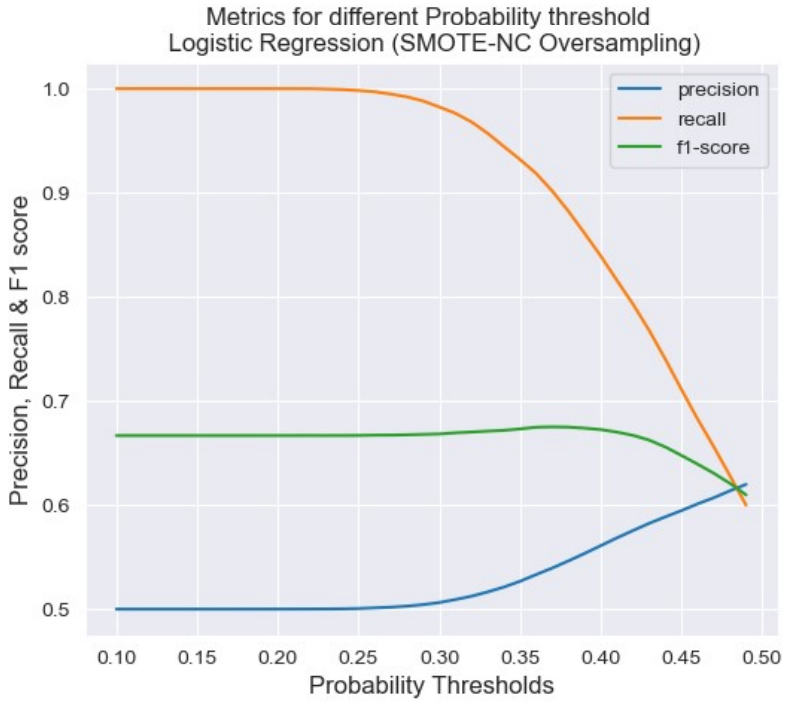
Logistic Regression Algorithm (Full predictors)

- Features dropped from analysis
  - 7 medication features
  - 4 patient information features (Encounter Id, weight, Patient_nbr, payer code)
- Total features considered - 38 out of 49 features & 1 Target
  - 8 numerical (standardized)
  - 30 categorical (dummy encoded)
- Total features after dummy encoding = 110 features

Logistic Model	*Accuracy*	*Precision*	*Recall*	*F-1 score*	*AUC*
Imbalanced data	0.89	0.47	**0.04**	**0.03**	0.505
class_weight = 'balanced'	0.60	0.1687	**0.5441**	**0.2575**	0.60
Tomek-link undersampling	0.6072	0.1763	**0.5504**	**0.2671**	0.615
SMOTENC oversampling	0.8402	0.8666	**0.8039**	**0.8341**	0.90


Metrics for different Probability threshold
In Logistic Regression on Training Data(Imbalanced)
precision
recall
f1-score
Precision, Recall & F1 score
Probability Thresholds


Metrics for different Probability threshold
(Logistic Regression)
(Tomek link Undersampling with class_weight = 'balanded')
precision
recall
f1-score
Precision, Recall & F1 score
Probability Thresholds


Metrics for different Probability threshold
Logistic Regression (SMOTE-NC Oversampling)
precision
recall
f1-score
Precision, Recall & F1 score
Probability Thresholds

greatlearning

Random Forest Algorithm(Full predictors)

- Features dropped from analysis
  - 7 medication features
  - 4 patient information features (Encounter Id, weight, Patient_nbr, payer code)
- Total features considered - 38 out of 49 features & 1 Target
  - 8 numerical (standardized)
  - 30 categorical (dummy encoded)
- Total features after dummy encoding = 140 features

Random Forest Model	*Accuracy*	*Precision*	*Recall*	*F-1 score*
Imbalanced data (class_weight = 'balanced')	0.5022	0.5553	**0.03**	**0.011**
SMOTENC oversampling (Ordinal encoded)	0.8294	0.8555	**0.7911**	**0.8220**

greatlearning