PROJECT REPORT ON

Exploration and Modelling of Factors Leading to Readmission of Diabetes Patients

SUBMITTED BY

Group No. 4 [Batch: Feb 2020]

GROUP MEMBERS

1. FaiquaSaman

2.Madhu OG

3.Nikhil Manchanda

4. Pragati Khedkar

5. Sri Hari Reddy Pydala

RESEARCH SUPERVISOR

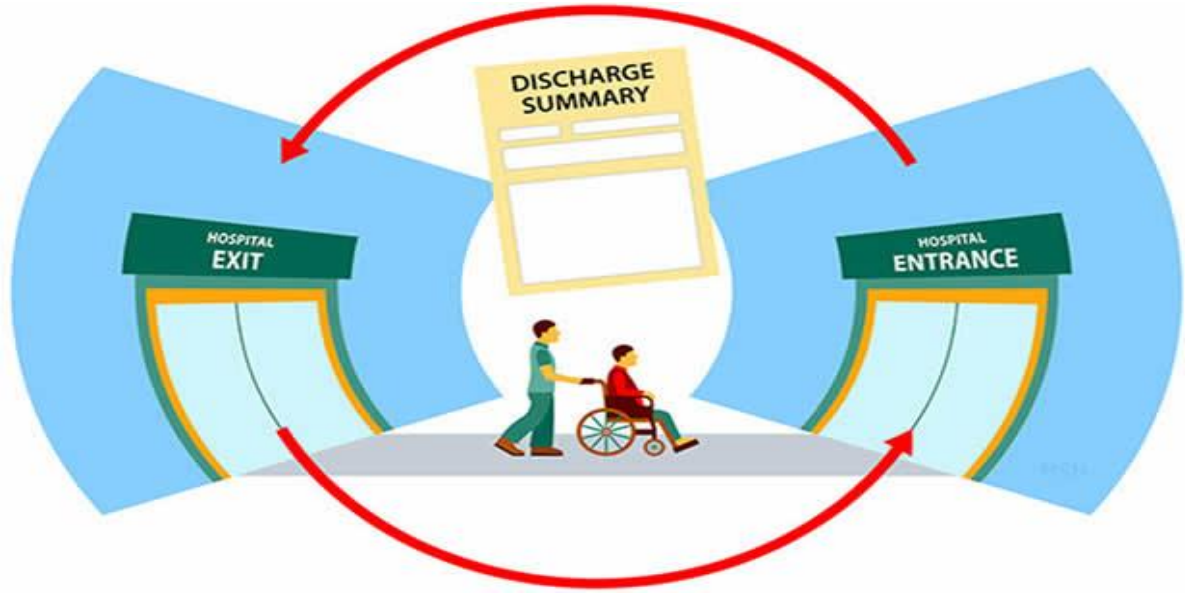
Mr. SrikarMuppidi

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Health care is a particularly important sphere in today's scenario, and a lot of investment and improvements are being done to achieve maximum output. Approximately out of the 100,000 cases, 78,000 are diabetic and over 47% are readmitted. In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge. Healthcare hospitalization (patient re-admission) cost is generally more than the healthcare premium paid by the consumer especially in US healthcare industry. Therefore, it is critical to identify patient readmission rate. Being able to determine factors that lead to higher readmission in such patients, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care.

Most studies employ regression data mining technique and provide a framework for implementing other machine learning techniques in exploring the causative agents of readmission rates among diabetes patients. The primary importance of the algorithm is to help hospitals identify multiple strategies that work effectively for re-admission of a given health condition.



1. Summary of problem statement, data and findings Every good abstract describes succinctly what was intended at the outset, and summarizes findings and implications.

Data Source: https://www.kaggle.com/brandao/diabetes

Problem Statement: Using predictive modelling to help prioritize diabetic patients those who have higher chances of getting readmitted to hospital.

- **Topic:**Exploration and Modelling of factors leading to readmission of diabetes patients
- **Domain:**Healthcare
- **Dataset:** Diabetes 130-US hospitals for years 1999-2008 Data Set

Business Perspective:

One of the problems faced in treatment of diabetes by the healthcare industry in U.S. is readmission of patients, which is often due to improper treatment and diagnosis. We aim at recognizing patterns that explain re-admission so that treatment and diagnosis criteria can be improved, and millions of dollars can be saved while improving quality of healthcare.

Data Findings:

Shape &dtypes:

- The dataset consists of **101766 records and 50 attributes**.

- The dtypes of attributes are a mixture of **int64 and object**.

Feature Categorization (Quantitative and Qualitative)

- **Quantitative: 13 columns**

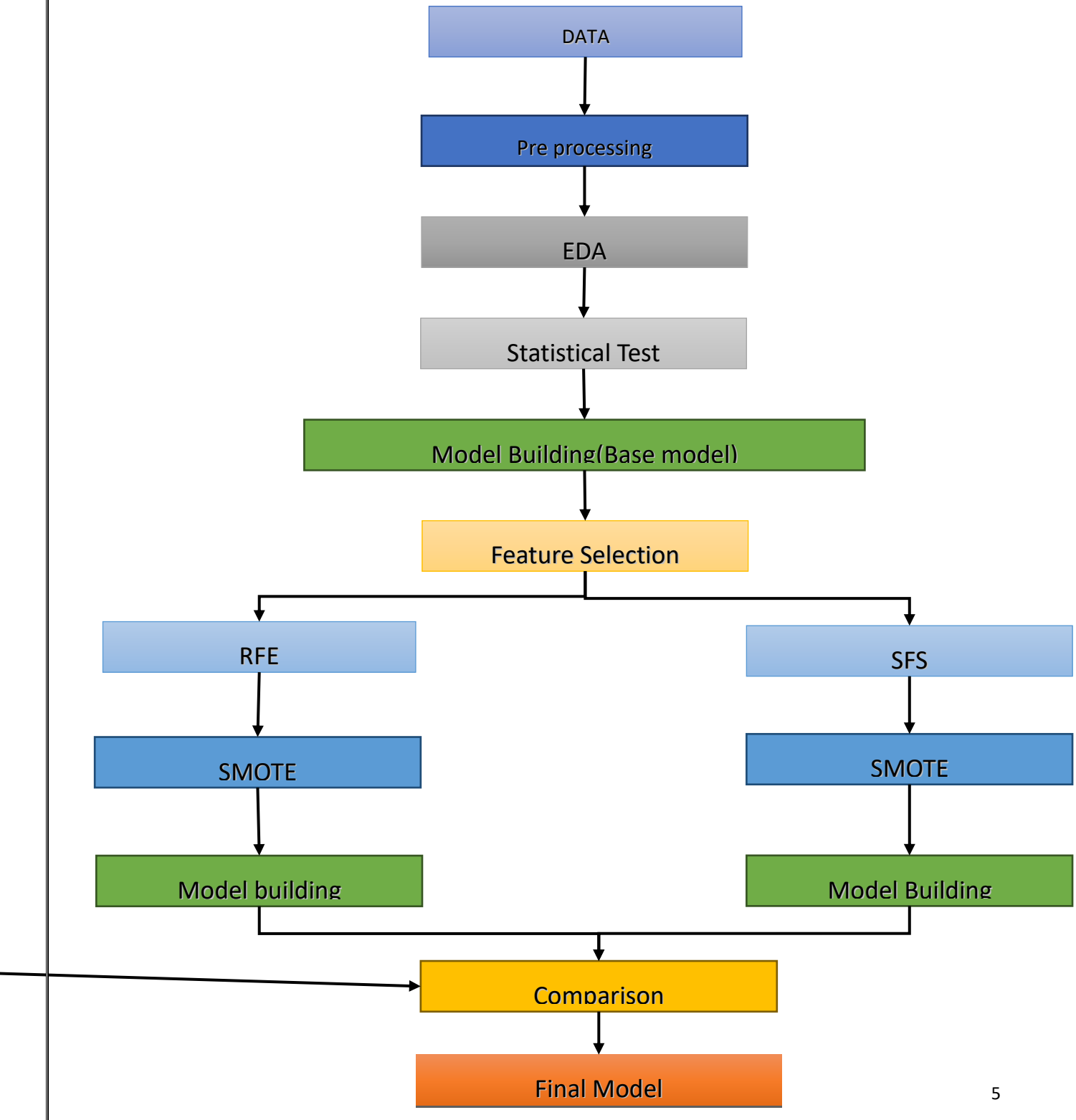
'encounter_id', 'patient_nbr', 'admission_type_id','discharge_disposition_id', 'admission_source_id', 'time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses'

- **Qualitative: 37 columns**

'race', 'gender', 'age', 'weight', 'payer_code', 'medical_specialty', 'diag_1', 'diag_2', 'diag_3', 'max_glu_serum', 'A1Cresult', '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', 'metformin-pioglitazone', 'change', 'diabetesMed', 'readmitted.

2. Overview of the final process Briefly describe your problem-solving methodology. Include information about the salient features of your data, data pre-processing steps, the algorithms you used, and how you combined techniques.

PROCESS OVERVIEW



A stepwise process has been followed in implementation of the supervised learning.

1. The project starts with collation of data from source, with its data description and mapping Ids.
2. It has been imported as a pandas data frame and the structure of the data is noted.
3. A recursive approach has been followed while checking the data for missing values, variable data types and outliers.
4. Exploratory data analysis has been performed including the univariate and bivariate analysis for the continuous variables and for the categorical variables as far as possible.
5. Data has been visualized using different charts including distance plots, histograms, bar charts and box plots. Five-point summary statistics and correlation of data is also noted.
6. Categorical and numerical data separated out and treated differently. Categorical data converted to dummy variables and numerical data is scaled.
7. Statistical test are performed on data to check the relevant features.
8. Three different sets of features are obtained using different selection techniques, namely:
   1)All features
   2)Features using RFE
   3)Features using SFS
9. The target variable has been completely analysed and SMOTE technique has been applied to overcome the class imbalance in the data.
10. Baseline models have been designed using 3 different sets of features obtained in step8.
11. Understanding different metrices based on baseline model(recall and precision score)
12. Based on conclusions on baseline model, different classification algorithms are used to enhance accuracy and recall score.
    Algorithms used are: Random Forest(baseline model),Logistic Regression, Navies bayes,EnsembleTechnique(AdaBoost Classifier), Gboost, Catboost.
    Models implemented recursively in the process to compare the results of each.

3. Step-by-step walk through of the solution Describe the steps you took to solve the problem. What did you find at each stage, and how did it inform the next steps? Build up to the final solution.

Data Pre-Processing:

Step1: Dropping records from 'readmitted'='>30'

This is done keeping in mind that according to business point of view, patients that get readmitted before 30 days is a class that needs the maximum attention. Also '>30' can mean any number of days, months, or years, so this class does not seem relevant enough.

Step2: Finding the percentage of missing values with respect to the actual volume of data

Column Name	Missing Value Percentage
weight	96.85848
medical_specialty	49.08221
payer_code	39.55742
race	2.233555
diag_3	1.398306
diag_2	0.351787
diag_1	0.020636
Gender	0.002948

Step3:Dropping unwanted columns:

1) **Dropping columns having high percentage of missing value:**
   weight
   medical_specialty
   patient_nbr
2) **Dropping ID columns that are not relevant for prediction:**
   encounter_id
   patient_nbr
   admission_type_id
   discharge_disposition_id
   admission_source_id

Step4: Missing Value Imputation:

Column Name	Method of imputation
gender	mode
diag_1	mode
diag_2	mode
diag_3	mode
Medical_condition	Iterative Imputer method
race	Replaced with string 'Unknown'

Step5: Converting data from label to strings:

- Note: This step is done using the information given in data description.
- Converted columns:diag_1, diag_2, diag_3
- Advantage of step 5: Data became more readable and the number of distinct values in diagnosis columns reduced from approx. 700 to5 which is good for EDA purpose.

Step6: Converting columns diag_1, diag_2, diag_3 to respective labels

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, 001–139,290–319,E–V, 280–289,320–359,630–679,360–389,740–759

Step7: Merging all the medication columns into one column.

Medication Columns: 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'glyburide-metformin', 'tolazamide', 'metformin-pioglitazone','metformin-rosiglitazone', 'glimepiride-pioglitazone', 'glipizide-metformin', 'troglitazone', 'tolbutamide', 'acetohexamide'

Merged to form new column: **numchange**

Reason: Medications other than 'insulin' and 'metformin' did not have much variation among class. So, it was better to merge the columns to reduce the number of features there by increasing the comprehensibility.

Step8: Checking for percentage category:

Readmitted	Count of Readmitted	% Readmitted
<30	11,357	17.15%
NO	54,864	82.84%

Here we can see our final Dataset is moderately imbalanced.

Step9: Performing Statistical tests

Shapiro test is performed to check normality in the numerical features.

Column Name	P-value
time_in_hospital	0.000000e+00
num_lab_procedures	0.000000e+00
num_procedures	0.000000e+00
num_medications	1.048171e-42
number_outpatient	0.000000e+00
number_emergency	0.000000e+00
number_inpatient	0.000000e+00
number_diagnoses	0.000000e+00

Inference: Here p-value is <0.01, therefore the data in above columns is not normal.

Since it is not normal, we perform non-parametric test (Mannwhitneyu)

time_in_hospital	7.607614e-84
num_lab_procedures	5.718669e-21
num_procedures	1.962113e-09
num_medications	1.822846e-77
number_outpatient	1.794262e-95
number_emergency	7.779246e-216
number_inpatient	0.000000e+00
number_diagnoses	1.979743e-112

Step10: Applying power transform

Transforming numeric features and comparing the skewness before and after the transform. Huge reduction in skewness is observed after power transform.

Column Name	Skewness before Transform	Skewness after Transform
time_in_hospital	1.133999	0.012906
num_lab_procedures	-0.236544	-0.226734
num_procedures	1.316415	0.207261
num_medications	1.326672	0.018905
number_outpatient	8.832959	1.810634
number_emergency	22.855582	2.462986
number_inpatient	3.614139	0.740986
number_diagnoses	-0.876746	-0.13744

Observation: We can see that skewness reduced to a great extent after applying power transform.

Step11: Baseline model Building

Model: RandomForestClassifier

-Reason to use Random Forest as our base model:
As we have seen our dataset is moderately imbalanced. Decision trees frequently perform well on imbalanced data. In modern machine learning, tree ensembles (Random Forests, Gradient Boosted Trees, etc.) almost always outperform singular decision trees, so directly using Random Forest which is bag of Decision Tree.
Note: Tree base algorithm work by learning a hierarchy of if/else questions. This can force both classes to be addressed.

- **The Problem with Class Imbalance**

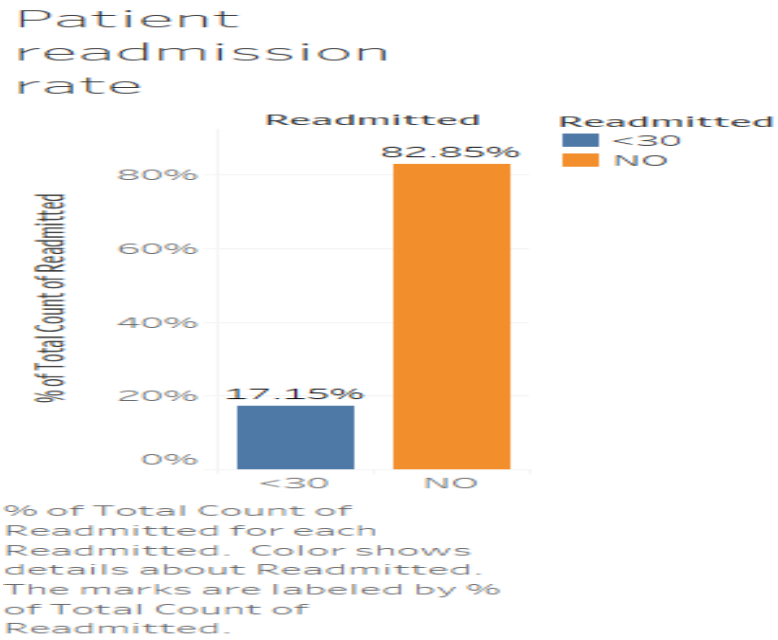
Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce errors.

#Reason to not rely on accuracy score if data has high imbalance.

If the data set in imbalance then In such cases, you get a pretty high accuracy just by predicting the **majority class**, but you fail to capture the **minority class**, which is most often the point of creating the model in the first place.

#Imbalance in our dataset:

Readmitted	Count of Readmitted	% Readmitted
<30	11,357	17.15%
NO	54,864	82.84%



Step12: Feature Selection

1)SFS(Sequential Feature Selector)

After applying SFS we obtained 17 features with high average score.

Features:
'race','gender','age','num_medications','number_outpatient','number_emergency','number_inpatient','diag_1','max_glu_serum','A1Cresult','metformin','examide','citoglipton','insulin','change','diabetesMed','numchange'

2)RFE(Recursive Feature Elimination Technique)

After applying RFE we obtained features with high average score.

Features:

'race','gender','age','num_medications','number_outpatient','number_emergency','number_inpatient','diag_1','max_glu_serum','A1Cresult','metformin','examide','citoglipton','insulin','change','diabetesMed','numchange'

Step13: Implementing technique to eliminate class imbalance.

Technique: **SMOTE method to treat class imbalance**

SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The **synthetic points are added** between the chosen point and its neighbors.

Disadvantage of SMOTE Method: The number of records increased significantly from 66221 records to 108000 records, making computation difficult.

Advantage: We saw little enhancement in the base model metrices after using SMOTE.

Note: Reason of not using Downsampling:

Downsampling reduces the number of majority class records to match it to minority class records. In this process we might lose important information.

4. Model evaluation Describe the final model (or ensemble) in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models(s)? A convincing explanation of the robustness of your solution will go a long way to supporting your solution.

Model Evaluation:Comparing all the models on various metrices based on all three feature selection techniques

1) Considering all features:

Model	Accuracy	Precision	Recall	Bias error	Variance error
RF	0.763276	0.167768	0.097173	0.027683	0.001456
LogReg	0.825137	0.151786	0.005006	0.011802	0.000952
NB	0.829063	0	0	0.018715	0.001374
Boosted RF	0.171339	0.170939	0.999411	0.02679	0.001704
GBoost	0.422761	0.153324	0.525618	0.020799	0.000896
CatBoost	0.574873	0.141641	0.293875	0.013991	0.000415

2) Considering RFE selected features:31 features

- Features: time_in_hospital, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, number_diagnoses, race_Caucasian, race_Other, age_[30-40), age_[40-50), age_[50-60), age_[70-80), age_[80-90), diag_1_Diabetes,diag_1_Genitourinary, diag_1_Injury, diag_1_Neoplasms, diag_1_Respiratory, diag_2_Diabetes, diag_2_Genitourinary,diag_2_Respiratory,diag_3_Diabetes, diag_3_Genitourinary, max_glu_serum_None, A1Cresult_Norm,metformin_No, insulin_Steady, insulin_Up, diabetesMed_Yes

Model	Accuracy	Precision	Recall	Bias error	Variance error
LogReg	0.81497	0.123656	0.013545	0.044273	0.003238
RF	0.591886	0.144324	0.281508	0.062196	0.003361
NB	0.829063	0	0	0.054718	0.005359
Boosted RF	0.560175	0.140318	0.306832	0.058628	0.004581
GBoost	0.727639	0.129732	0.103946	0.054718	0.005359
CatBoost	0.752554	0.127451	0.076561	0.048876	0.004191

3) Considering SFS selected features: 17 features

Features: race, gender, age, num_medications, number_outpatient, numchange number_emergency, number_inpatient, diag_1, max_glu_serum, A1Cresult, metformin, examide, citoglipton, insulin, change, diabetesMed

Model	Accuracy	Precision	Recall	Bias error	Variance error
RF	0.66	0.26	0.5	0.11	0.001
LogReg	0.67	0.22	0.35	0.18	0.001
NB	0.69	0.23	0.34	0.27	0.004
Boosted RF	0.7	0.23	0.31	0.13	0.001
GBoost	0.65	0.25	0.49	0.27	0.004
CatBoost	0.75	0.12	0.07	0.04	0.004

Final Model:

- **Objective: To select a model which gives least number of false negatives and higher recall score:**

Explanation:

1)False negative is more dangerous than false positive because:

A false positive can lead to unnecessary treatment and a false negative can lead to a false diagnostic, which is very serious since a disease has been ignored.

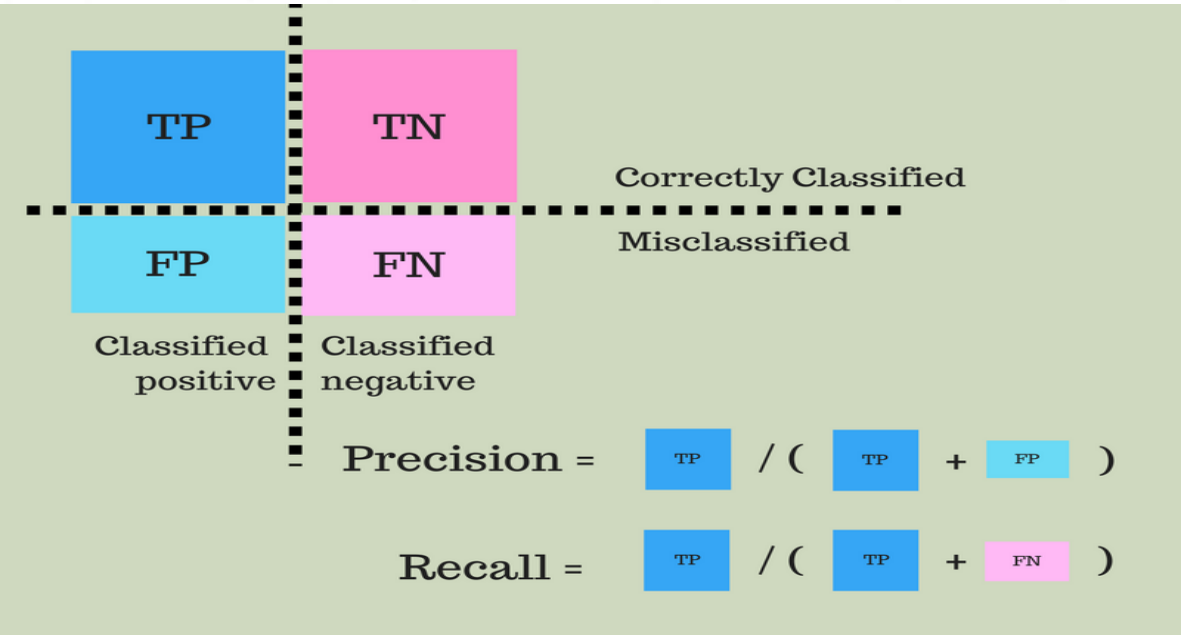
So, model must be such that false negative is minimal.

2)**Reason to use recall and precision score as the matrix for evaluation instead of accuracy score:**

If the number of negative samples is very large (a.k.a imbalance data set) the false positive rate increases more slowly. Because the true negatives (in the fpr denominator — (FP+TN)) would probably be extremely high and make this metric smaller.

Precision:Precision however, is not affected by a large number of negative samples, that's because it measures the number of true positives out of the samples predicted as positives (TP+FP).Precision is more focused in the positive class than in the negative class, it actually measures the probability of correct detection of positive values, while FPR and TPR (ROC metrics) measure the ability to distinguish between the classes.

Recall: Our dataset is based on healthcare data. Here one class(i.e. patient readmitted in less than 30 days) holds much more importance than other class(i.e. patient not getting readmitted). Also, in field of healthcare, high false negatives cant be afforded according to health point of view (false negative in healthcare data means disease not being detected which could me even fatal in some cases). So our model should be such that false negatives are minimal.



As we can see, Low false negative means high recall score. **That's is why we need to give more weightage to recall score than any other matrix.**

Evaluating the success of your models/Final model on different metrices:

Metrices used: Confusion matrix, Precision,recall,accuracy score

- **Confusion Matrix:** a table showing correct predictions and types of incorrect predictions.
- **Precision:** the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
- **Recall:** the number of true positives divided by the number of positive values in the test data. The recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
- **FPR and TPR**

predicted → real ↓	*Class_pos*	*Class_neg*
Class_pos	TP	FN
Class_neg	FP	TN

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

$$\text{FPR (1-specificity)} = \frac{FP}{TN + FP}$$

Final Model Metrices:

Final Model	Accuracy	Precision	Recall	Bias error	Variance error
GBoost	0.422761	0.153324	0.525618	0.020799	0.000896

We created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyperparameters.

5. Comparison to benchmark How does your final solution compare to the benchmark you laid out at the outset? Did you improve on the benchmark? Why or why not?

Benchmark:

The initial expectation from this process is to have a considerable accuracy score, precision and recall, a smaller number of false positives and false negatives and high TPR.

Benchmark Chosen: recall=60%, precision=60%

Note: These are the approx. values that were expected at the outset.

Reasons to choose benchmark:

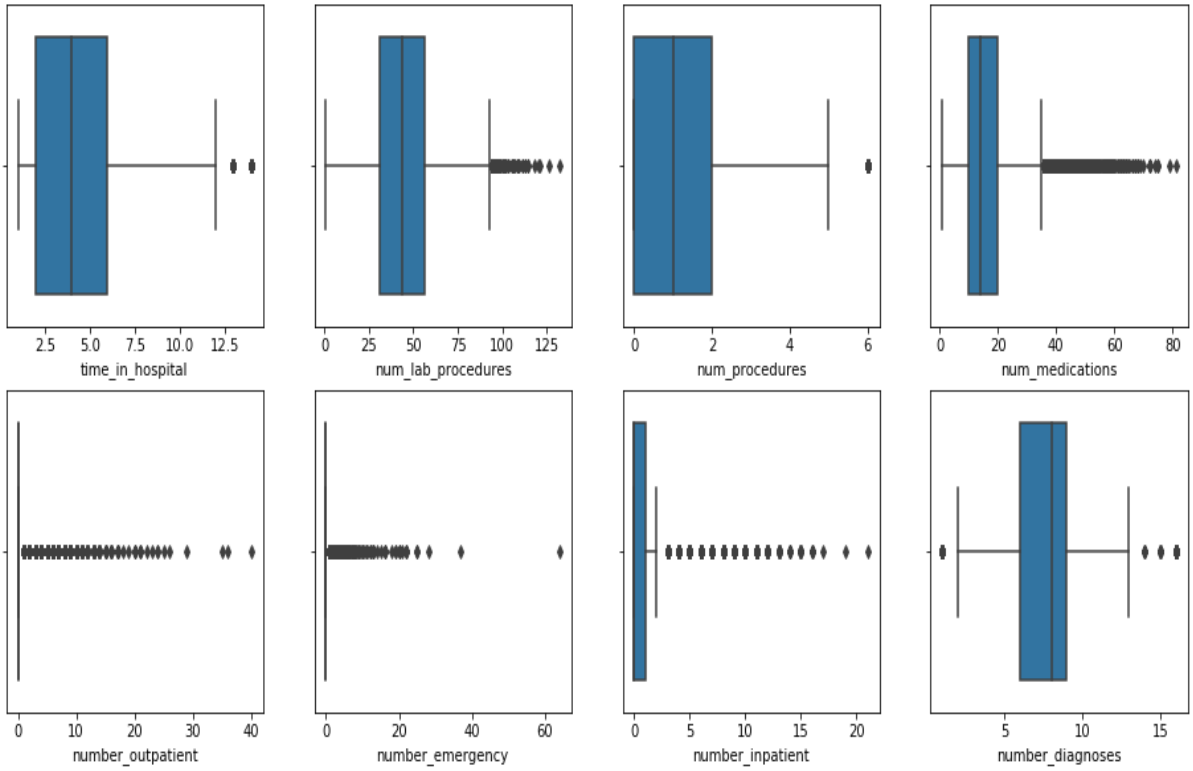
1) Any machine learning model cannot give 100% accuracy. there is always a possibility of misclassification in the output.
2) Number of records are not sufficient to train the model to an extent of getting exceptionally high accuracy.

In the final model :

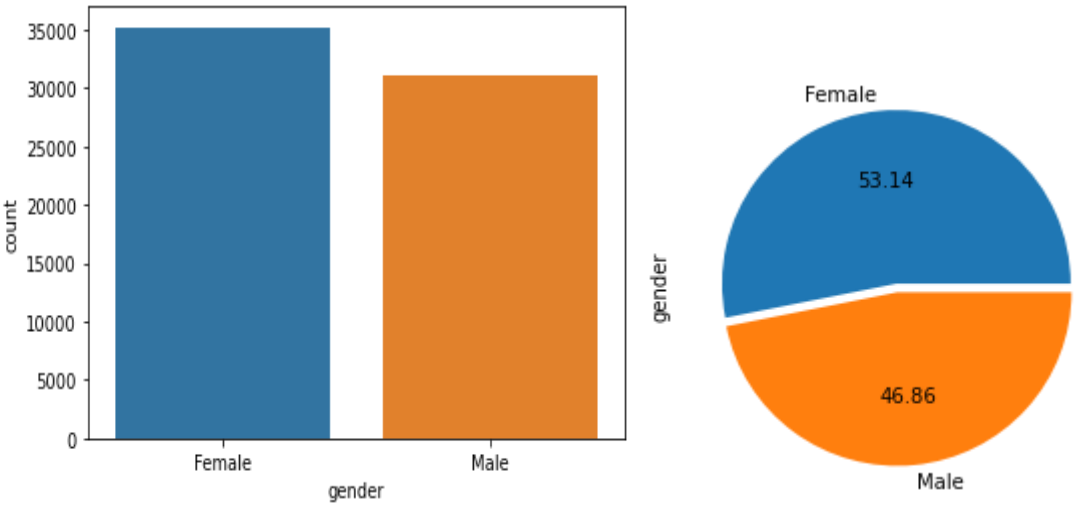
- Accuracy, precision, recall are moderate
- false positives and false negatives reduced
- Bias Error and Variance error also reduced

6. Visualization(s) In addition to quantifying your model and the solution, please include all relevant visualizations that support the ideas/insights that you gleaned from the data.

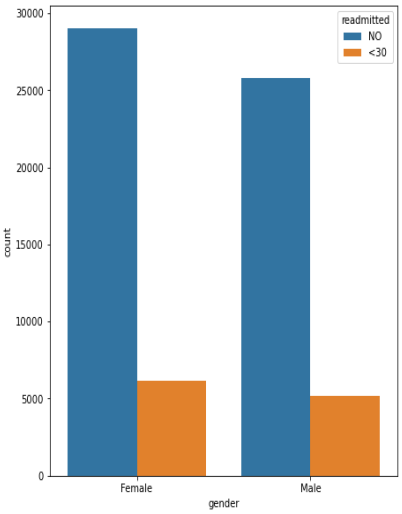
- **Analysing outliers using boxplot:**



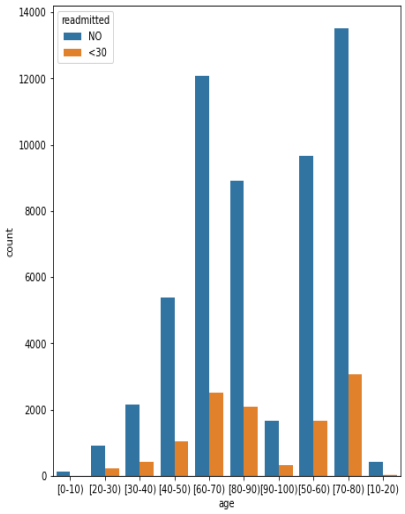
- **Analysing Gender distribution of patients:**



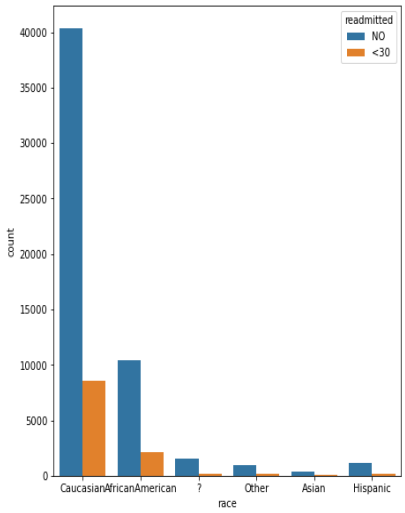
- **Analysing distribution of various features based on target column:**



Gender

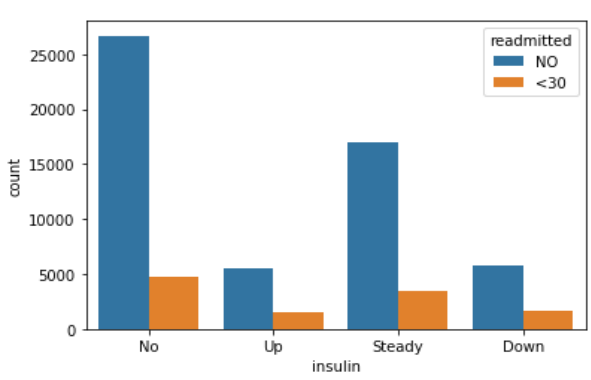


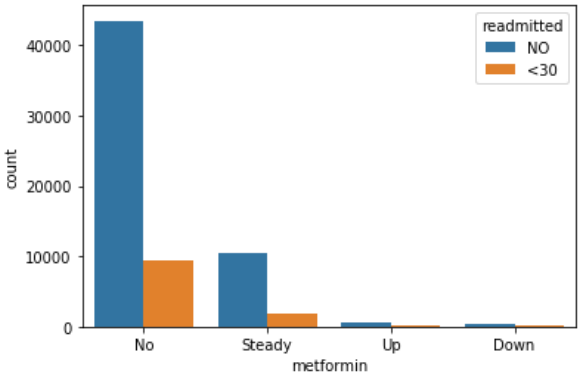
Age



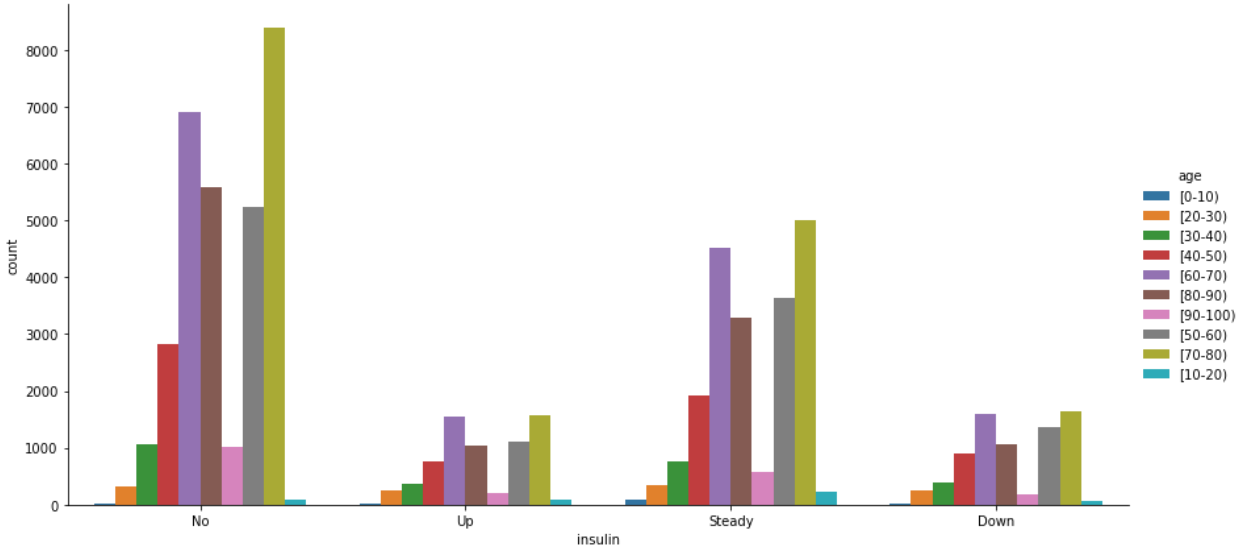
Race

- **Analysing important medications base on target column:**

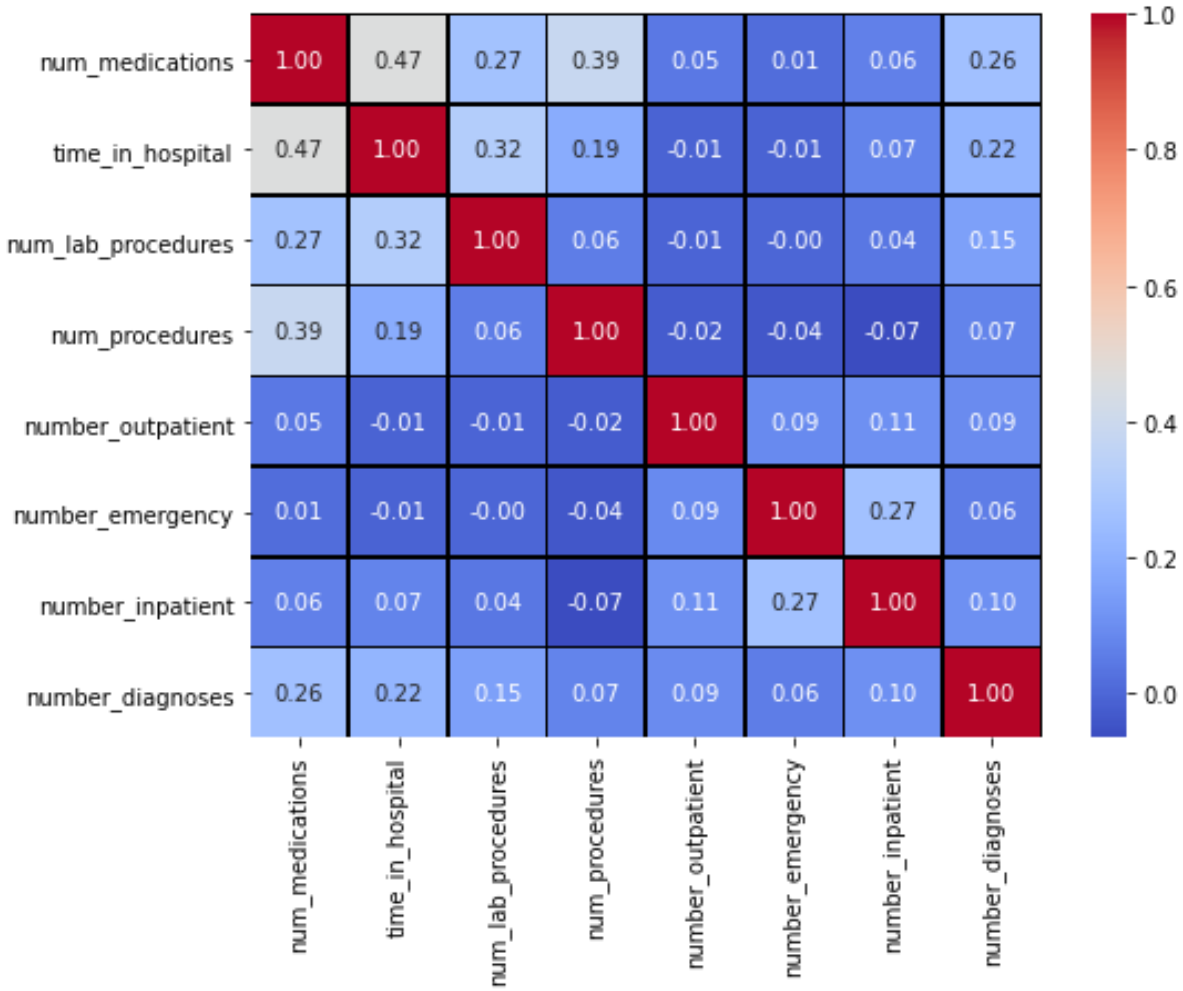




- **Analysing Insulin prescription base on age-group**



#Heatmap to visualise correaltion among various numerical features:



7. Implications How does your solution affect the problem in the domain or business? What recommendations would you make, and with what level of confidence?

Implications(because of dataset):

1) Features in the dataset are not sufficient to have exact predictions. Some more features such as weight,medical background, information about any other major disease,etc. can also be tracked and considered while prediction.
2) Large number of records taking high processing time
3) Missing value: Feature '**weight**' which is one of the most prominent features to keep track of in case of diabetic patients had 97% of missing values.
4) Dealing with class imbalance in data:

Readmitted	Count of Readmitted	% Readmitted
<30	11,357	17.15%
NO	54,864	82.84%

Effect of solution on the problem in the domain or business

Problem in hand:

Health care is a particularly important sphere in today's scenario, and a lot of investment and improvements are being done to achieve maximum output.

Approximately out of the 100,000 cases, 78,000 are diabetic and over 47% are readmitted. In 2011, American hospitals spent over $41 billion on diabetic patients who got readmitted within 30 days of discharge.

Healthcare hospitalization (patient re-admission) cost is generally more than the healthcare premium paid by the consumer especially in US healthcare industry. Therefore, it is critical to identify patient readmission rate.

Project Outcome on business:

This project can help us to be able to determine factors that lead to higher readmission in such patients, and to predict which patient has higher chances of getting readmitted so that they could be taken under supervision for a longer duration of time by the hospital authority. This can help hospitals save millions of dollars while improving quality of healthcare.

Recommendations:

- Healthcare centres (especially the ones where readmission rate is high) can implement this ML predictive modelling technique to check which patient has high chance of getting readmitted thereby prioritising their treatment thus reducing the patient readmission rate and improving the quality of service provided.
- Flaws in the dataset (mentioned earlier like the missing values in 'weight' column) can also be reduced if proper track of patient information is considered by the hospital.

8. Limitations What are the limitations of your solution? Where does your model fall short in the real world? What can you do to enhance the solution?

Limitations of solution in real world:

- Accuracy Score: Accuracy of the final model is 42%, so the predicted output does not have 100% surety to be correct
- Number of False negatives: It is extremely dangerous to have false negatives in the model output (especially in healthcare domain) because this shows that our model may miss out on detecting the disease (here the chance of readmission)
- Number of False positives: This is not as bad as false negatives, but it does have a disadvantage that if a person if falsely detected with chance of getting readmitted, he/she would unnecessarily consume the resources at hospital. Also, the money spent by the patient and hospital in this case will be of no use.
- Despite being highly relevant data, since most of our data is a set of binary class variables, thus we see less scope of visualization and bivariate analysis in terms of categorical variables.

Measures to enhance the solution:

1) Trying to reduce the anomalies in the dataset in the beginning.

Problems such as missing values, outliers,peculiar behaviour of some records is reduced in this project.But, other advanced techniques(such as iterative imputer) can also be used to cleanse data further for better prediction results

2) Introducing new columns in dataset:

Derived features in our dataset (i.e. 'numchange': number of changes made in medication) has proved to be a significant feature. Other derived features can also be introduced to predict the target variable.

3)Use of advanced algorithms:

In this project, there are restrictions on use of fewer models because of high volume of dataset and less processing speed. Butpredictive models (such as KNN model, SVM classifier) can be used to have better prediction with the use of better processors.

9. Closing Reflections What have you learned from the process? What would you do differently next time?

#Learning:

- Learnt to work on different classification algorithms for predictive modelling, different data cleaning processes, feature selection techniques, SMOTE technique to handle class imbalance in the data.
- Learnt the role of model evaluation metric based on business goal(reducing false negative which results in high recall score).
- Based on the dataset, more stringent feature engineering and feature selection process.
- Evaluating each step not only with a mathematical point of view but also from a business perspective.
- The major takeaway is to understand some internal details (such as how a model deals with class imbalance, interpretation of confusion matrix for each model,etc) about the algorithms and how different algorithms work by comparing their performance.
- Saw what kind of problems may arrive in classification modelling and tried to solve them using different approaches. This helped in understanding the use of each approach we tried and pros and cons of all the methods involved.

Things to try:

- Use of SVM classifier.
- Use of different imputation techniques (KNN imputer, Iterative imputer).
- Use of derived features.
- Building models with 90:10 train test split

Conclusion:

The project undertaken involves a real life scenario (i.e. the prediction of diabetic patient readmission) and working on such case, looking at the problems that may arrive and finding different solutionsto it helped in understanding how we can use our knowledge of predictive modelling in real world on a large scale.

References:

1)False Positive v/s False negative:https://blogs.ams.org/mathgradblog/2016/08/06/false-positive-vs-false-negative/

2)SFS:http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

3)Class imbalance treatment:https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

4)Why precision score should be used in case of high data imbalance:
https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

5)Dealing with imbalanced data:https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/