PREDICTING HOSPITAL READMISSION FOR PATIENTS WITH DIABETICS

A Project report submitted to

DEPARTMENT OF

COMPUTER SCIENCE AND ENGINEERING

In partial fulfillment of the requirements for the award of the degree of

Bachelor Of Technology In

Computer Science And Engineering

CHALAPATHI INSTITUTE OF ENGINEERING & TECHNOLOGY

Presented by

CH.RENUKA	Y16CS1219
B.MOUNIKA	Y16CS1212
M.UNNATHA	Y16CS1263
B.GOPI KRISHNA	Y16CS1216

Under the esteemed guidance of

Dr.B.V.V.S.PRASAD,

Associate professor,

Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CHALAPATHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Approved by A.I.C.T.E, Affiliated To Acharya Nagarjuna University)

GUNTUR-522034

2019-2020

Introduction

As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program (HRRP), which reduces reimbursement to hospitals with above average re admissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. According to Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population (Ostling et al 2017). Therefore, in this article, I will focus on predicting hospital readmission for patients with diabetes.

In this project I will demonstrate how to build a model predicting readmission in Python using the following steps

- data exploration
- feature engineering
- building training/validation/test samples
- model selection
- model evaluation

You can follow along with the Jupyter Notebook provided on my github (https://github.com/andrewwlong/diabetes_readmission).

Data Exploration

The data that is used in this project originally comes from the UCI machine learning repository

Feature Engineering

In this section, we will create features for our predictive model. For each section, we will add new variables to the dataframe and then keep track of which columns of the dataframe we want to use as part of the predictive model features. We will break down this section into numerical features, categorical features and extra features.

Numerical Features

The easiest type of features to use is numerical features. These features do not need any modification.

Categorical Features

The next type of features we want to create are categorical variables. Categorical variables are non-numeric data such as race and gender. To turn these non-numerical data into variables, the simplest thing is to use a technique called one-hot encoding.

Extra Features

The last two columns we want to make features are `age` and `weight`. Typically, you would think of these as numerical data.

Model Selection: Baseline models

In this section, we will first compare the performance of the following 7 machine learning models using default hyperparameters:

- K-nearest neighbors
- Logistic regression
- Stochastic gradient descent
- Naive Bayes
- Decision tree
- Random forest
- Gradient boosting classifier

Conclusion

Through this project, 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. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients. Overall, I believe many healthcare data scientists are working on predictive models for hospital readmission.

Through this project, we created a binary classifier to predict the probability that a patient with diabetes would be readmitted to the hospital within 30 days. On held out test data, our best model had an AUC of of 0.67. Using this model, we are able to catch 58% of the readmissions from our model that performs approximately 1.5 times better than randomly selecting patients.

screen shots of out puts:

Home Page - Select or create a n × | diabetes_python_proj - Jupyter N ×

localhost:8888/notebooks/diabetes_python_proj.ipynb

jupyter diabetes_python_proj Last Checkpoint: 21 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

In [20]:
```
df.groupby('med_spec').size()
```

Out[20]:
```
med_spec
Cardiology                        5279
Emergency/Trauma                  7419
Family/GeneralPractice            7252
InternalMedicine                 14237
Nephrology                        1539
Orthopedics                       1392
Orthopedics-Reconstructive        1230
Other                             8199
Radiologist                       1121
Surgery-General                   3059
UNK                              48616
dtype: int64
```

In [21]:
```
cols_cat_num = ['admission_type_id', 'discharge_disposition_id', 'admission_source_id']
df[cols_cat_num] = df[cols_cat_num].astype('str')
```

In [22]:
```
df_cat = pd.get_dummies(df[cols_cat + cols_cat_num + ['med_spec']],drop_first = True)
```

In [23]:
```
df_cat.head()
```

Out[23]:

	race_Asian	race_Caucasian	race_Hispanic	race_Other	race_UNK	gender_Male	gender_Unknown/Invalid	max_glu_serum_>300	max_glu_serum_None	max
0	0	1	0	0	0	0	0	0	1	

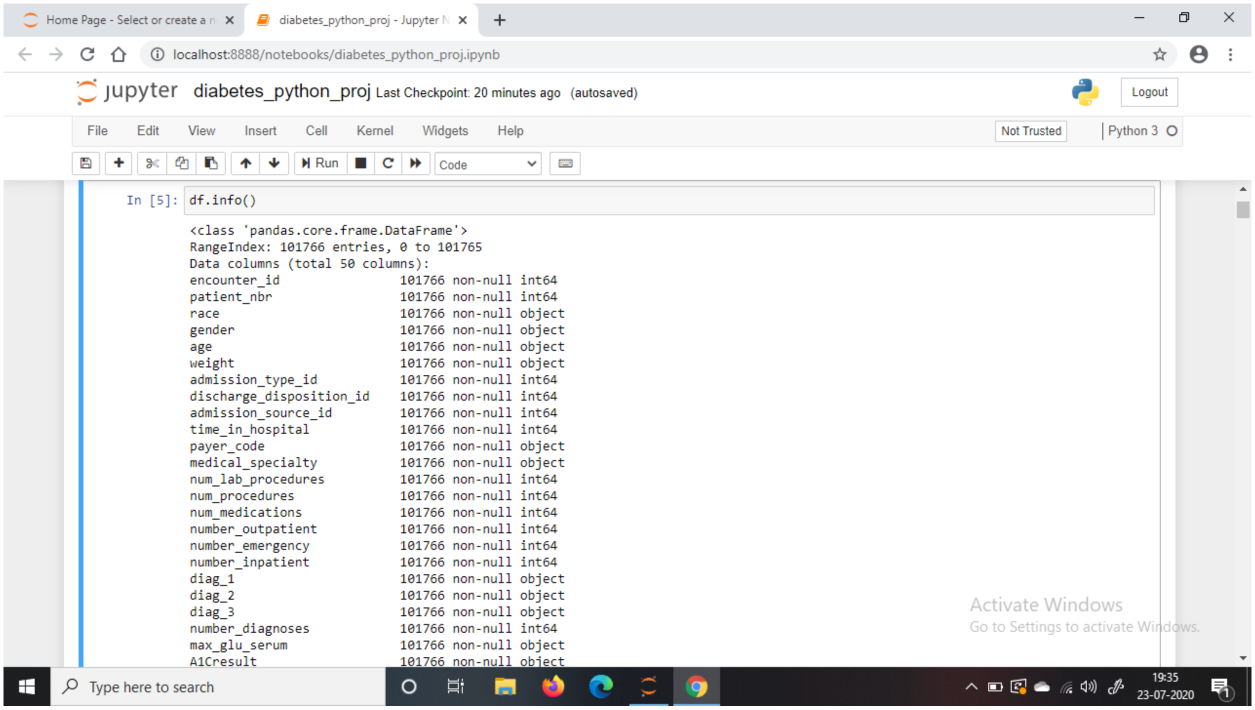
Type here to search

19:35
23-07-2020

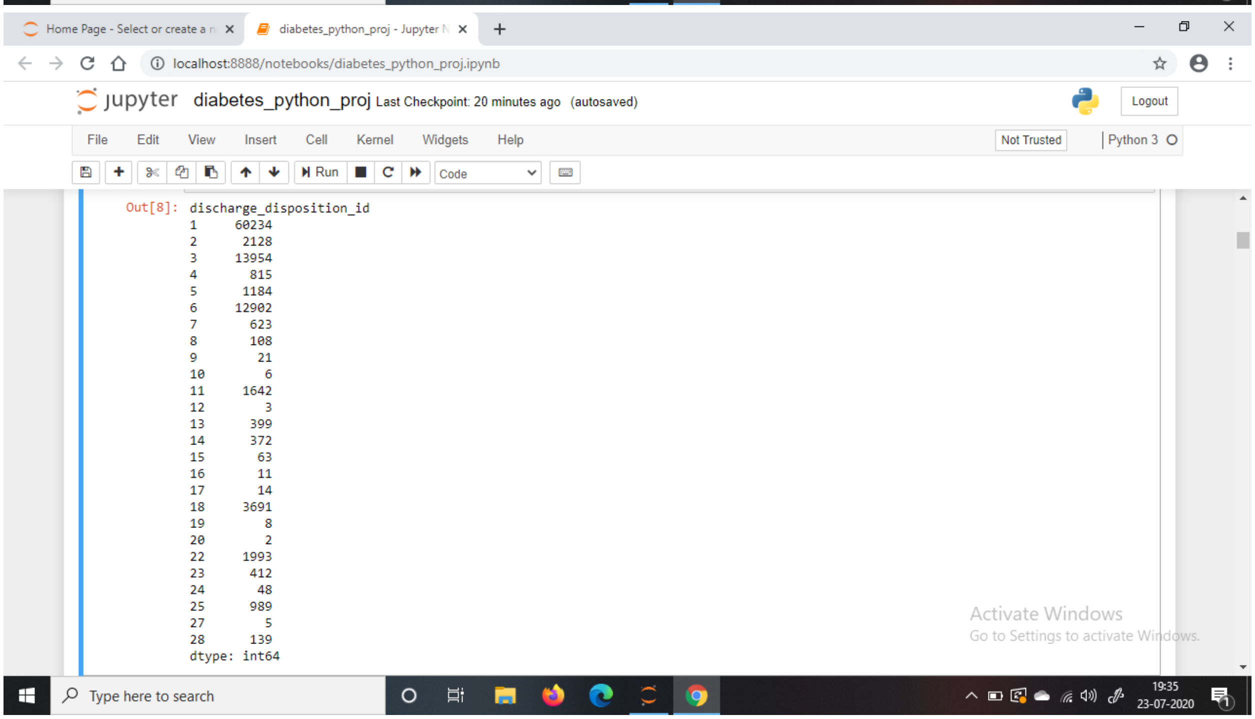
```
In [24]: df[cols_cat].isnull().sum()

Out[24]: race                        2234
         gender                         0
         max_glu_serum                  0
         A1Cresult                      0
         metformin                      0
         repaglinide                    0
         nateglinide                    0
         chlorpropamide                 0
         glimepiride                    0
         acetohexamide                  0
         glipizide                      0
         glyburide                      0
         tolbutamide                    0
         pioglitazone                   0
         rosiglitazone                  0
         acarbose                       0
         miglitol                       0
         troglitazone                   0
         tolazamide                     0
         insulin                        0
         glyburide-metformin            0
         glipizide-metformin            0
         glimepiride-pioglitazone       0
         metformin-rosiglitazone        0
         metformin-pioglitazone         0
         change                         0
         diabetesMed                    0
```

```
df.groupby('medical_specialty').size().sort_values(ascending = False)

Number medical specialty: 73

Out[16]: medical_specialty
         UNK                                     48616
         InternalMedicine                        14237
         Emergency/Trauma                         7419
         Family/GeneralPractice                   7252
         Cardiology                               5279
         Surgery-General                          3059
         Nephrology                               1539
         Orthopedics                              1392
         Orthopedics-Reconstructive               1230
         Radiologist                              1121
         Pulmonology                               854
         Psychiatry                                853
         Urology                                   682
         ObstetricsandGynecology                   669
         Surgery-Cardiovascular/Thoracic           642
         Gastroenterology                          538
         Surgery-Vascular                          525
         Surgery-Neuro                             462
         PhysicalMedicineandRehabilitation         391
         Oncology                                  319
         Pediatrics                                253
         Neurology                                 201
         Hematology/Oncology                       187
         Pediatrics-Endocrinology                  159
         Otolaryngology                            125
```

====




```
In [5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 101766 entries, 0 to 101765
Data columns (total 50 columns):
encounter_id                101766 non-null int64
patient_nbr                 101766 non-null int64
race                        101766 non-null object
gender                      101766 non-null object
age                         101766 non-null object
weight                      101766 non-null object
admission_type_id           101766 non-null int64
discharge_disposition_id    101766 non-null int64
admission_source_id         101766 non-null int64
time_in_hospital            101766 non-null int64
payer_code                  101766 non-null object
medical_specialty           101766 non-null object
num_lab_procedures          101766 non-null int64
num_procedures              101766 non-null int64
num_medications             101766 non-null int64
number_outpatient           101766 non-null int64
number_emergency            101766 non-null int64
number_inpatient            101766 non-null int64
diag_1                      101766 non-null object
diag_2                      101766 non-null object
diag_3                      101766 non-null object
number_diagnoses            101766 non-null int64
max_glu_serum               101766 non-null object
A1Cresult                   101766 non-null object
```




```
Out[8]: discharge_disposition_id
1     60234
2      2128
3     13954
4       815
5      1184
6     12902
7       623
8       108
9        21
10        6
11     1642
12        3
13      399
14      372
15       63
16       11
17       14
18     3691
19        8
20        2
22     1993
23      412
24       48
25      989
27        5
28      139
dtype: int64
```

Out[8]:
```
discharge_disposition_id
1     60234
2      2128
3     13954
4       815
5      1184
6     12902
7       623
8       108
9        21
10        6
11     1642
12        3
13      399
14      372
15       63
16       11
17       14
18     3691
19        8
20        2
22     1993
23      412
24       48
25      989
27        5
28      139
dtype: int64
```

In [15]:
```
df[list(df.columns)[10:20]].head()
```

Out[15]:

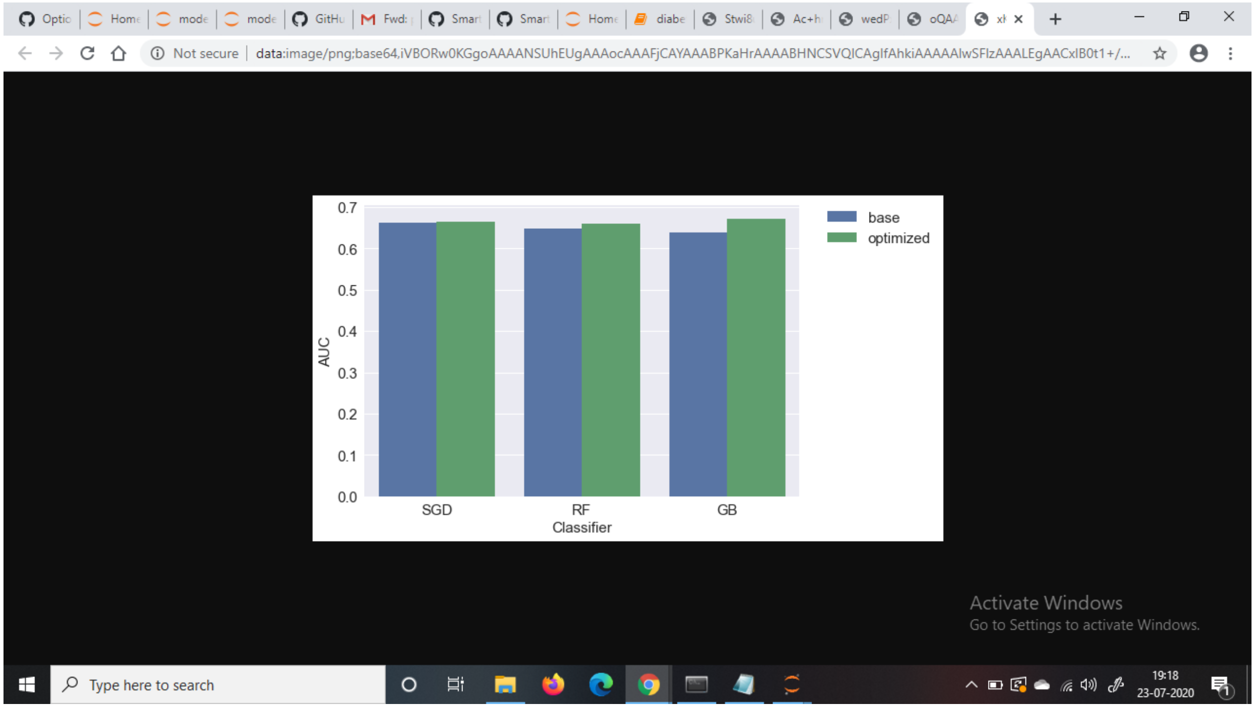
	payer_code	medical_specialty	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	diag_1
0	?	Pediatrics-Endocrinology	41	0	1	0	0	0	250.83
1	?	?	59	0	18	0	0	0	276
2	?	?	11	5	13	2	0	1	648
3	?	?	44	1	16	0	0	0	8
4	?	?	51	0	8	0	0	0	197

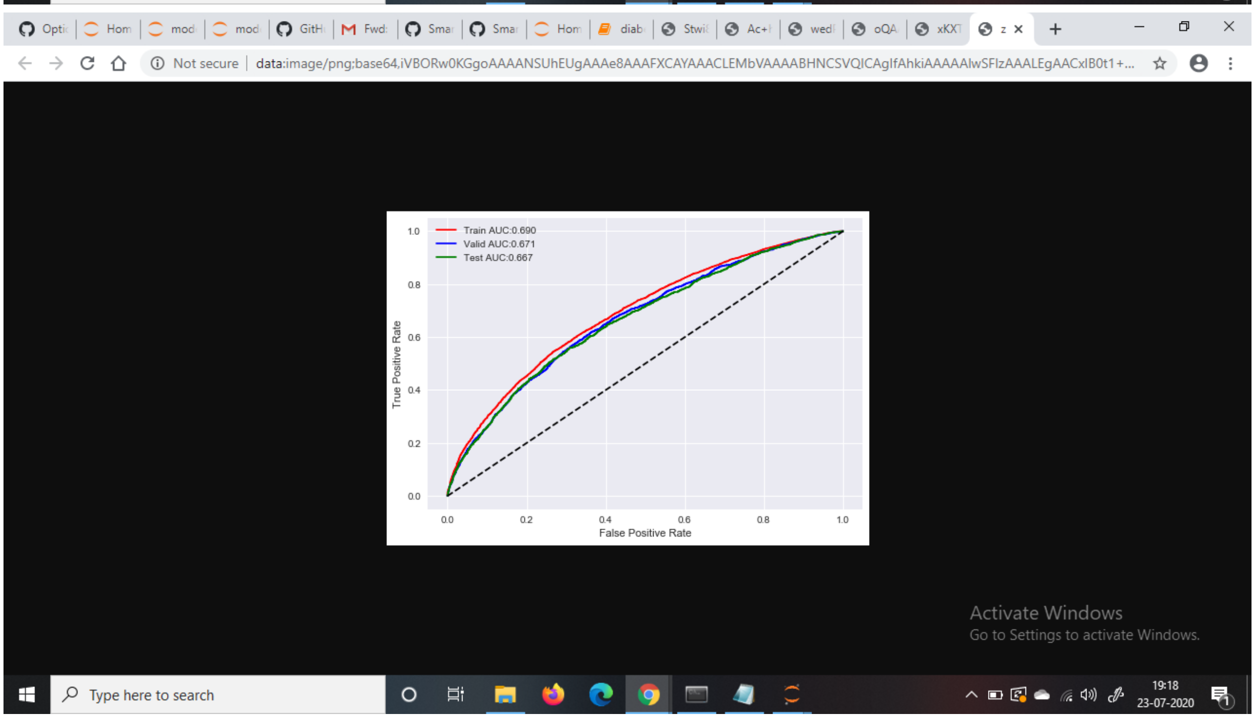
In [16]:
```
df[list(df.columns)[20:30]].head()
```

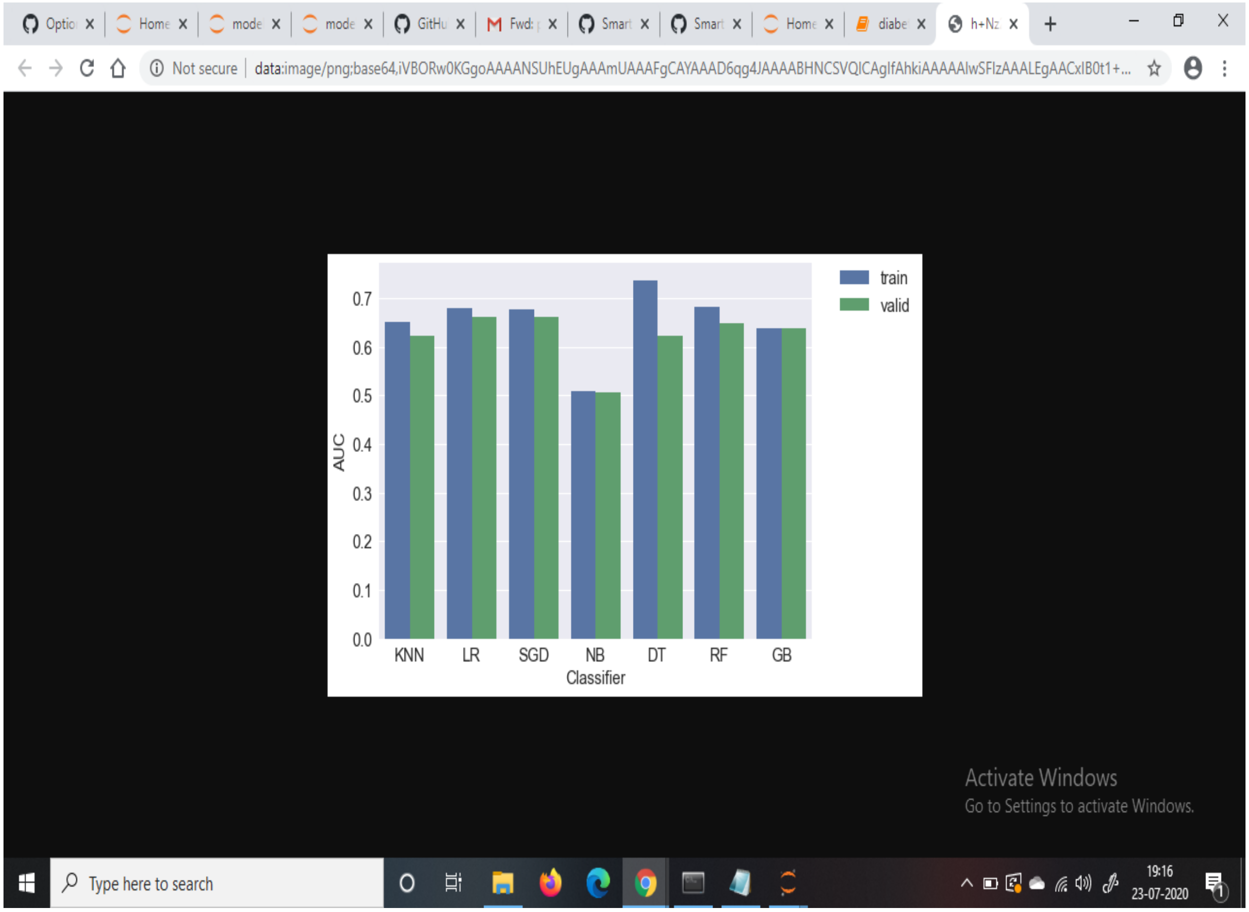
Out[16]:

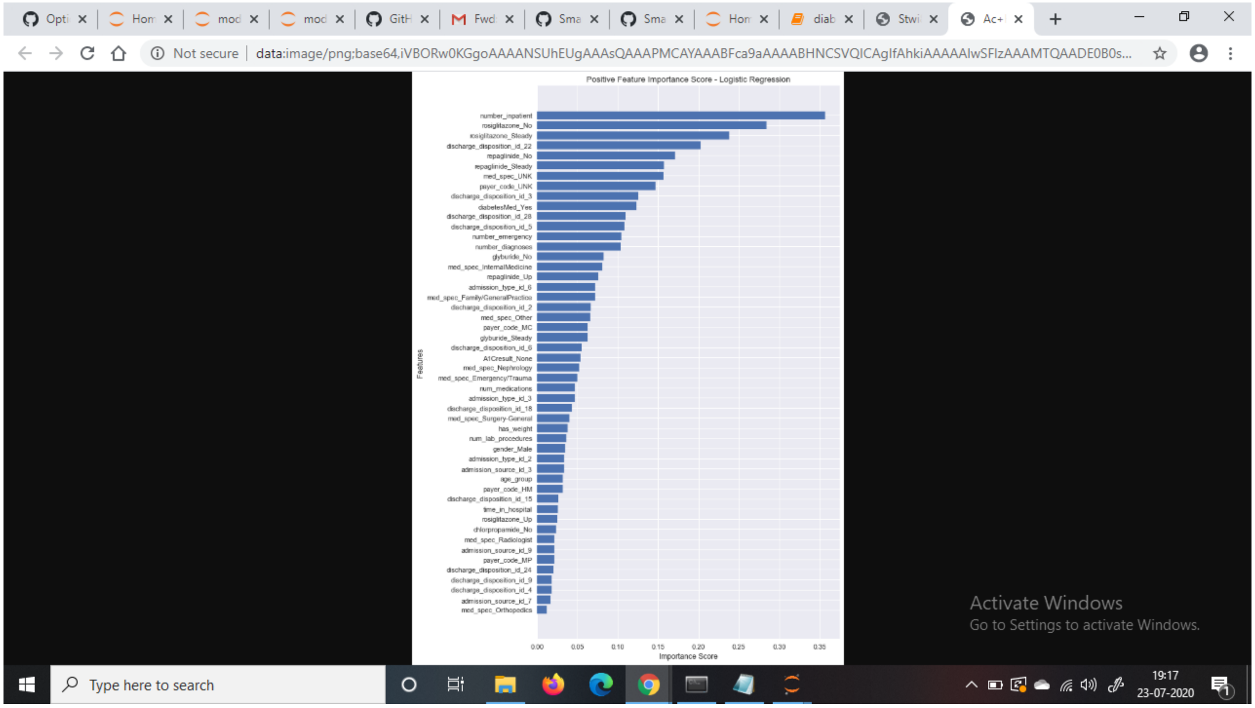
	diag_3	number_diagnoses	max_glu_serum	A1Cresult	metformin	repaglinide	nateglinide	chlorpropamide	glimepiride	acetohexamide
0	?	1	None	None	No	No	No	No	No	No
1	255	9	None	None	No	No	No	No	No	No
2	V27	6	None	None	No	No	No	No	No	No
3	403	7	None	None	No	No	No	No	No	No
4	250	5	None	None	No	No	No	No	No	No

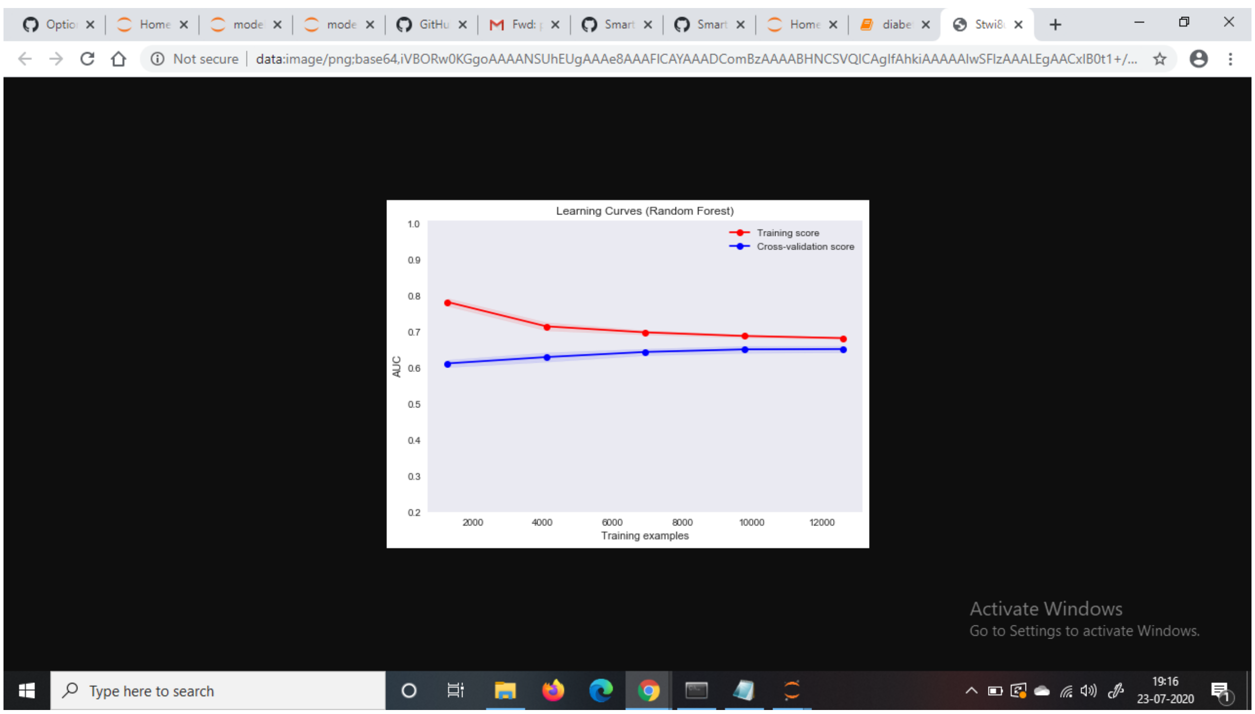
In [17]:
```
df[list(df.columns)[30:40]].head()
```


Not secure
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
AUC
SGD
RF
GB
Classifier
base
optimized
Activate Windows
Go to Settings to activate Windows.
Type here to search
19:18
23-07-2020


Not secure
Train AUC:0.690
Valid AUC:0.671
Test AUC:0.667
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Activate Windows
Go to Settings to activate Windows.
Type here to search
19:18
23-07-2020


Not secure
data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAmUAAAFgCAYAAAD6qg4JAAAABHNCSVQICAgIfAhkiAAAAAlwSFlzAAALEgAACxIB0t1+...
train
valid
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
AUC
KNN
LR
SGD
NB
DT
RF
GB
Classifier
Activate Windows
Go to Settings to activate Windows.
Type here to search
19:16
23-07-2020


Positive Feature Importance Score - Logistic Regression
Features
Importance Score
Activate Windows
Go to Settings to activate Windows.
Type here to search
19:17
23-07-2020


Learning Curves (Random Forest)
Training score
Cross-validation score
AUC
Training examples
Activate Windows
Go to Settings to activate Windows.
Type here to search
19:16
23-07-2020