Machine Learning Engineer Nanodegree

Report Project

Fareeda Alharbi

Overview:

Patient missing or did not come to their appointment is common issue in many country, but this issue could cause a lot of problems. One of them is economic issue in England there was" £1bn is being wasted annually by patients missing appointments"1 and the cost of each missed appointment in 1997 was estimated at £65. The second issue is manpower. Patients' failure to attend increases the time others must wait to see a hospital specialist. Non-attendance means under-utilization of equipment and manpower. The third issue is patient health. A delay in presentation and therefore diagnosis, or haphazard monitoring of chronic conditions, will predispose to avoidable ill health[2].

We will apply the Machine learning algorithms to predict the patients who have no show up to their appointments' so, we can focuses on their situation to find a suitable solution for them.

Problem Statement

In Vitoria the capital city of Espirito Santo State-Brazil the patient set his appointment then he/she don't show up, So I will tray to Know why there are 30% of the patient did

[1] https://www.theguardian.com/society/2018/jan/02/patients-missing-their-appointments-cost-the-nhs-1bn-last-year
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1279909/

not show up, by applied the Machin learning algorithms to predict the patient that he is missing his appointment.

This is common issue in many countries so if I successfully design the predictive model then this will be very helpfully to other people when they deal with the same problem in different dataset. And because this is a classification problem I use Machin learning Supervised algorithms (Gaussian Naive Bayes GaussianNB,AdaBoostClassifier,RandomForestClassifieer and GradientBoostingClassifier) because the dataset has labels and also I want to classify the patent to show or no show up patient, these algorithms have different matric to evaluate I choose the accuracy and F1 to evaluate the performance of the model, the data will feed to the model and the out put will be Show or no Show up.

Analysis:

Datasets and Inputs

Variable	Decision
PatientId	Identification of a patient
AppointmentID	Identification of each appointment
Gender	Male or Female
ScheduledDay	The schedule the appointment
AppointmentDay	The actual day of appointment
Age	How old is the patient
Neighbourhood	Where the appointment takes place
Scholarship	If he tacks scholarship the value will be True and if not it will be false (https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia)

Hipertension	Hypertension chronic disease
Diabetes	Diabetes chronic disease
Alcoholism	If the patient drink Alcohol 1 and 0 for not drinking
Handcap	How many patients has disability
SMS_received	SMS reminder the patient about his appointment
No-show	Value 1 to who missing the appointment 0 for who attend
AgeClass	Child, adult, ,senior
Scheduledyear	The year which set the appointment
Scheduledmonth	The month which set the appointment
Scheduledweek	The week which set the appointment
Scheduledhour	The hour which set the appointment
Appointmentyear	The actual appointment year
Appointmentmonth	The actual appointment month
Appointmentweek	The actual appointment week
WaitingDay	The period between the schedule date and the appointment date
Weekname	The name of the week days Sunday, Monday…etc.

The data set is hosting by Kaggle.com it is comma separated values (CSV), consist of 14 variables and 110527 records before cleaning and data engineering. The final input data is 110527 records and 24 columns,

there are no missing values!! now I will classify the variables into four groups patient information, appointment information, health situation and the general information, I think this will help me in data preparation process.

1- Patient Information

PatientId

Gender

Age

I will ignore Patient Id as I think it has no importance in our analysis. Now let check the values of Gender and Age:

Gender:['F', 'M']

Age : [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 102, 115, -1]

There are tow values for gender M = male and F= Female so no need to perform any cleaning or auditing process, but in the Age variable We can see strang values for some patients like 100 and above and negative values, I wonder if the negative values mean baby before born or it may come from a typo error however as I'm not sure about this value I will delete it, also I check in
the [geoba.se](http://www.geoba.se/country.php?cc=BR&year=2017) and I found the average of life expectancy in Brazil is about 74.06 and there are just 4,388 people have age 100 and above so I will choose to delete patients who have age in this range.In addition we will classify the patient according to their age to child, adult and senior.

2- Appointment Information

AppointmentID

ScheduledDay

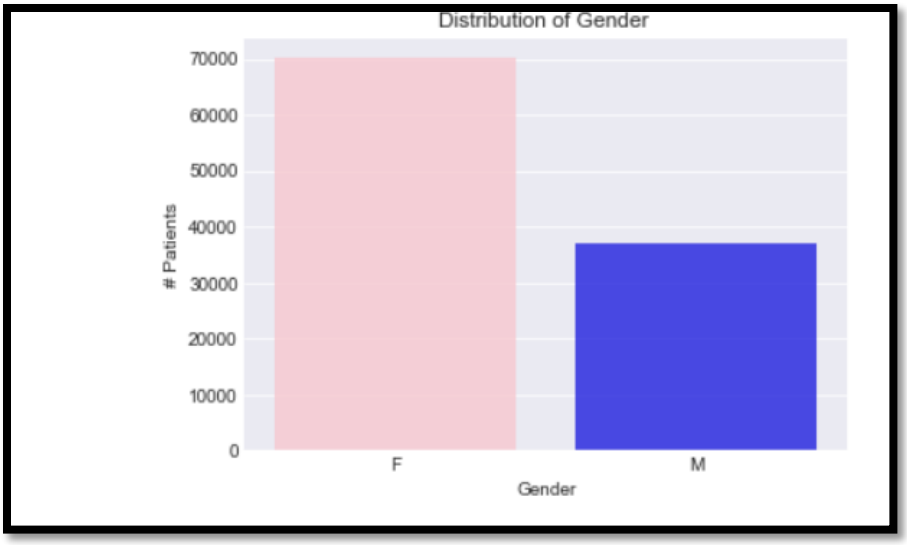
AppointmentDay

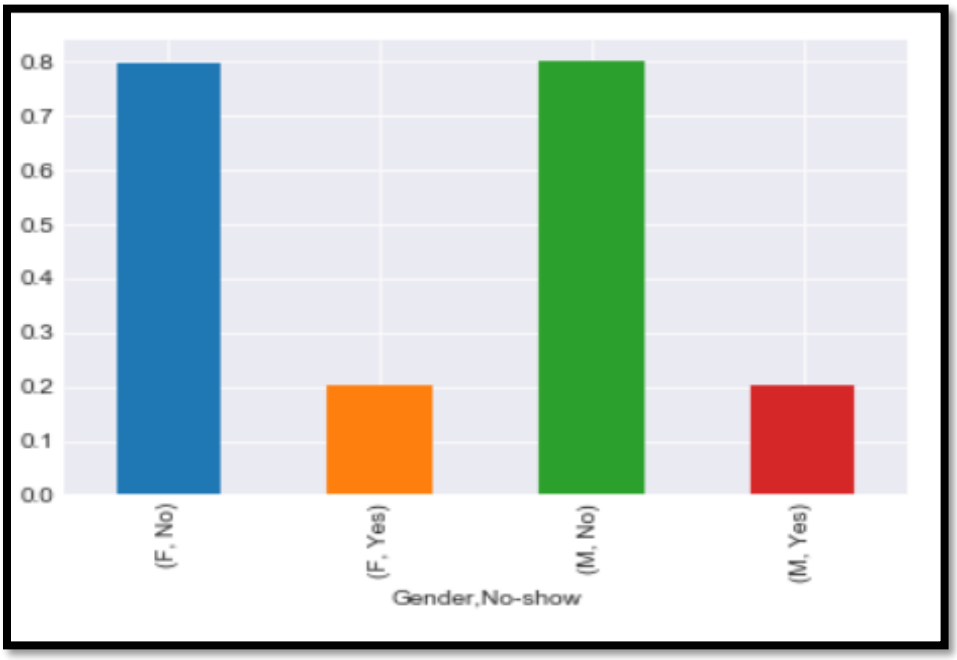
for the group of Appointment Information, I remove the AppointmentID from our dataset and reformate the ScheduledDay, in addition, i create a new variable WaitingDay which mean the duration between the scheduled date and the appointment date. Also I create others columns which is derived from the ScheduledDay and the AppointmentDay .

Exploring Data:

first let see How many patient attende thier appointment :

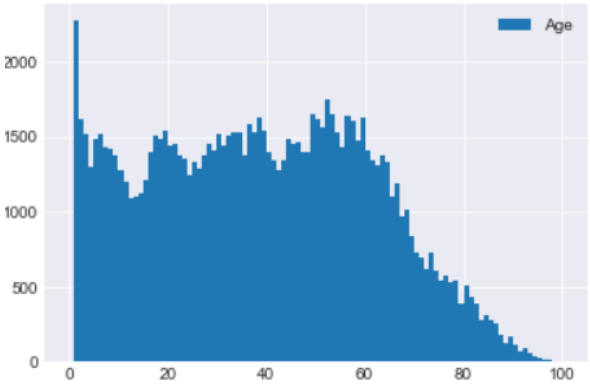
No-show
No 85299
Yes 21677
(20.27%) of the Patients not attend the appointment and approximately (79.73%) of them came to their appointment, so i try to go deeper in the data and identify how each features play roles in this problem.
How the Patient gender play role in this analysis??
Distribution of Gender
Gender
F 70109
M 36867

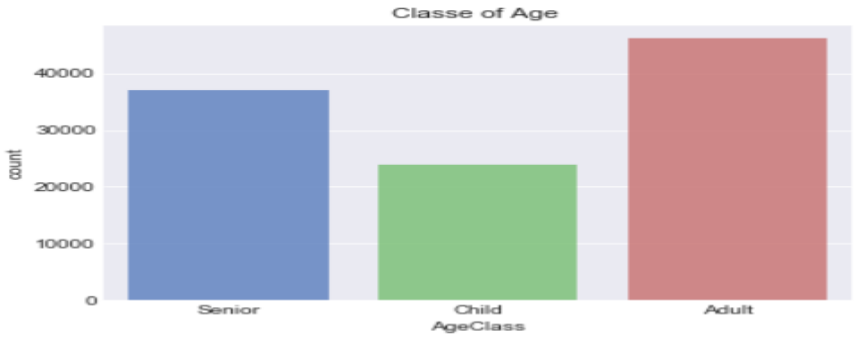




We can see the women are most likely visit the hospitals than men this may due to several reasons : women take care about her health than men and also the pregnant woman usually visit the hospital several time during her pregnancy, and we may consider that the population mean for women is greater than mean in Brazil, but when we foucse on the (Show up) statuse we can see 79.6% of wamen attened to their appointment compared to 79.9% of men, so women and men are most likly to have the same rate of attendance.

Does the Age affected the patient attendance to their appointment? Dose the elderly woman take care about her health more than elderly man?

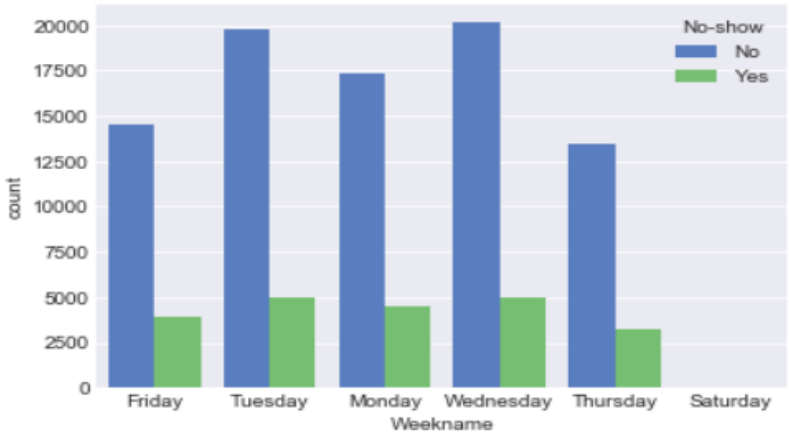





```
AgeClass   No-show
Adult      No          0.776539
           Yes         0.223461
Child      No          0.775252
           Yes         0.224748
Senior     No          0.837542
           Yes         0.162458
Name: No-show, dtype: float64
```

Most patients were between the age of 18 and 49 and the patients below 18 years has the minimum rate of visiting the hospital. However, when we see the show-up status we found patients above 50 years are most likely did not attend their appointment and also the gender did not play any role in this analysis.

Scheduled day , Appointmeint Day and Waiting Day:



in the appointment month it just includes 3 months!! on another hand the Scheduled month missed the months (7 to 10). However, in both diagrams, the most rate appointment was in may which is the end of Autumn in Brazile. The patient tends to attend his appointment in the middle of the week, and no show up at the end of the week, especially in Thursday. If we focus on the waiting day and how the data is distribution we can see there is a drop in the appointment number after the three months of waiting especially for the group who not came to there appointments.

How the Sms reminder and handicap variables affect the patient attendance

The majority of patient did not have handicap so I can't see the affective of handicap level on patient show up. 27% of patient who received sms did not show up and 0.723332% of patient who received the sms came to their appointment.

Algorithms and techniques:

For the Model I use python and scikit learn library which provide a lot of powerful tools and models , In the proposal I mansion that I will used :

- Support Vector Machines (SVM)
- Boosting Algorithm(XGBoost)

  But I have some issue in my machine so I replace it with other supervised algorithms :

Gaussian Naive Bayes (GaussianNB):

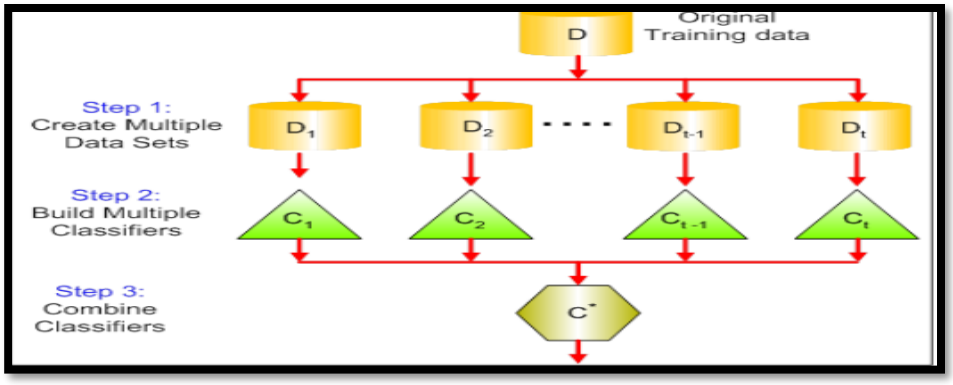
The Naive Bayes Algorithm is a fast, highly scalable algorithm and also It is a simple algorithm that depends on doing a bunch of counts. As it has few parameters it's hard to overfit, In addition As we have to find out model for a non profit hospital a high-end hardware may not be feasible and Naive bayes will be a good choice for lightweight learning.

- LogisticRegression

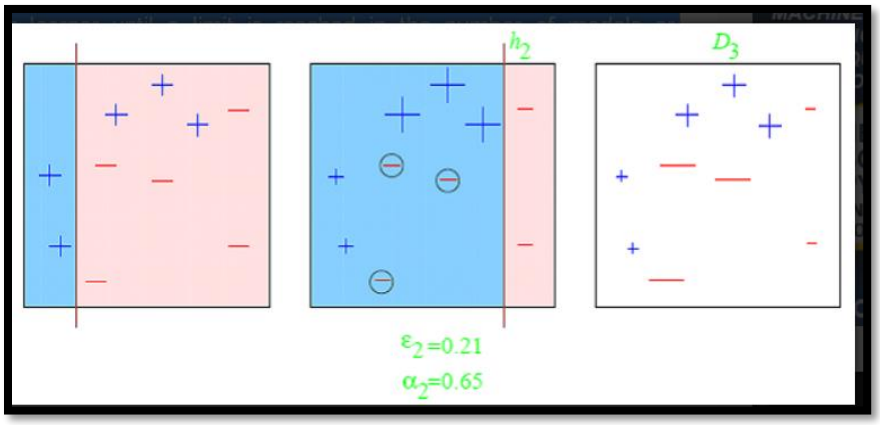
It is simple, fast, efficient for small dataset with limited features and I chose it because yt is the baseline algorithm in most framework and it is widely used in applications, The output of a logistic regression is more informative than other classification algorithms. Like any regression approach, it expresses the relationship between an outcome variable (sow up) and each of its predictors (features), Differently, a random forest (and all other variations of the decision tree method) will only tell you which predictors are more important to build the trees, without any information on the direction of association.

Also I use ensemble algorithms because it is Better prediction and More stable algorithms, the ensemble algorithms can be divide to :

Bagging is an ensemble method. First, we create random samples of the training data set (sub sets of training data set). Then, we build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging helps to reduce the variance error.



Boosting provides sequential learning of the predictors. The first predictor is learned on the whole data set, while the following are learnt on the training set based on the performance of the previous one**.** It starts by classifying original data set and giving equal weights to each observation. If classes are predicted incorrectly using the first learner, then it gives higher weight to the missed classified observation. Being an iterative process, it continues to add classifier learner until a limit is reached in the number of models or accuracy. Boosting has shown better predictive accuracy than bagging, but it also tends to over-fit the training data as well.



I use the ensambling algorithms:

- AdaBoostClassifier:

  Ensemble methods are considered to be high quality classifiers, and adaboost is the one of most popular boosting algorithms. We also have a class imbalance in our dataset, which boosting might be robust to.

GradientBoostingClassifier:

build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive. There are typically three parameters - number of trees, depth of trees and learning rate, and the each tree built is generally shallow.

RandomForestClassifier:

train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data. There are typically two parameters in RF - number of trees and no. of features to be selected at each node.

Benchmark Model

In the data set overview we say there is 30% is missing their appointment so, I will assume the model could predict 70% of theme as baseline for it , also in Kaggle.ocm for the same dataset there are some upvoting model has the accuracy 70% However the benchmark for this model will be 70 for the accuracy and 70% for the F1 score.

Methodology:

I start the data set with 14 variables then I create some variable like the waiting date and the month of appointment.. etc. After I do the data clean and create a new variables , I drop unnecessary features like ('AgeClass','Weekname','ScheduledDay','AppointmentDay'), and because I have some category's variable I use the encodeing technique to transforme the category features to numeric features (Neighbourhood, Gender) , after that all the varibles was numbers but some of them have valuese 0 and 1 and others like age have a big number like 70, 60 so I normalize the data using minmaxscaler() to have all values between the 0 and 1 so, we can deal with features equally ,then I split the data into training and testing set using cross_validation (train_test_split) I choose 20% of the data to be the testing set and the remaining for the training (Training set has 85576 samples. Testing set has 21395 samples).then I use the ShuffleSplit() it randomly sample the entire dataset during each **iteration** to generate a training set and a test set. The `test_size` and `train_size` parameters control how large the test and training test set should be for each iteration. Since we are sampling from the entire dataset during each iteration, values selected during one iteration, could be selected again during another iteration[3]. , a grid search technique was used to optimize the hyperparameters for a learning algorithm. Gaussian Naive Bayes did not have parameters to tuning it

LogisticRegression:

parameters = {'penalty': ['l1', 'l2'], 'C': [0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]}

RandomForestClassifier:

parameters = {'n_estimators':[1,2,3,4,5,6,7,8,9,10], 'criterion': ['gini', 'entropy'], 'min_samples_split': [2,3,4,5,6,7,8,9,10]}

[3] https://stackoverflow.com/questions/34731421/whats-the-difference-between-kfold-and-shufflesplit-cv

daBoostClassifier:

A parameters = {'n_estimators':[75,200,500],'learning_rate':[1.0,1.5,2.0]}.

Then I evaluate the models before the parameter tuning and after tuning the parameter, then I choose – RandomForestClassifier wich is has the highest F1 score then I tray to find the more important features for this data set (I will present it later in this report) and feed the model with the top 5 important features and evaluate it with accuracy and F1 score,

Evaluation Metrics

The Machin learning Supervised algorithms has a lot of Evaluation Metrics one of them is the accuracy which I will use it on evaluate the model performance, the accuracy can be defined as:

(True Negative + True Positive)/Total Papulations

And because in the dataset overview it said that approximately 30% is no show up I assume the dataset is imbalance so I should also use additional matrices so I will use the F1 score which include in his calculation the recall and precision :

*F1 = 2(recall * precision)/ (recall +precision)*

	Benchmark Model	Gaussian Naive	AdaBoostClassifier	RandomForestClassifier	GradientBoostingClassifier
Accurcy	70	0.7664	0.7985	0.7214	0.7934
F1	70	0.19	0.01	0.25	0.01

In general we can say the performance of this models is not good , the accuracy is approximately upove the benchmark model but we focuses on the other measures like f1, recall..

Refinement

In the supervised algorithms I use gride search to find the best hyper parameters and after that I evaluated the different algorithms before and after the tuning parameter using the F1 score and accuracy :

GradientBoostingClassifier

```
Unoptimized model
------
Accuracy score on testing data: 0.7979
F-score on testing data: 0.0087

Optimized Model
------
Final accuracy score on the testing data: 0.7934
Final F-score on the testing data: 0.2093
The best parameters are {'learning_rate': 0.3, 'max_depth': 5, 'n_estimators': 400}
             precision    recall  f1-score   support

          0       0.80      1.00      0.89     17083
          1       0.39      0.00      0.01      4312

avg / total       0.72      0.80      0.71     21395
```

RandomForestClassifier:

```
Unoptimized model
------
Accuracy score on testing data: 0.7883
F-score on testing data: 0.2502

Optimized Model
------
Final accuracy score on the testing data: 0.7214
Final F-score on the testing data: 0.3232
The best parameters are {'criterion': 'entropy', 'min_samples_split': 2, 'n_estimators': 1}
            precision    recall  f1-score   support

         0       0.82      0.94      0.88     17083
         1       0.44      0.18      0.25      4312

avg / total       0.74      0.79      0.75     21395
```

LogisticRegression:

```
Unoptimized model
------
Accuracy score on testing data: 0.7943
F-score on testing data: 0.0401

Optimized Model
------
Final accuracy score on the testing data: 0.7941
Final F-score on the testing data: 0.0409
The best parameters are {'C': 100.0, 'penalty': 'l1'}
             precision    recall  f1-score   support

          0       0.80      0.99      0.88     17083
          1       0.34      0.02      0.04      4312

avg / total       0.71      0.79      0.71     21395
```

AdaBoostClassifier:

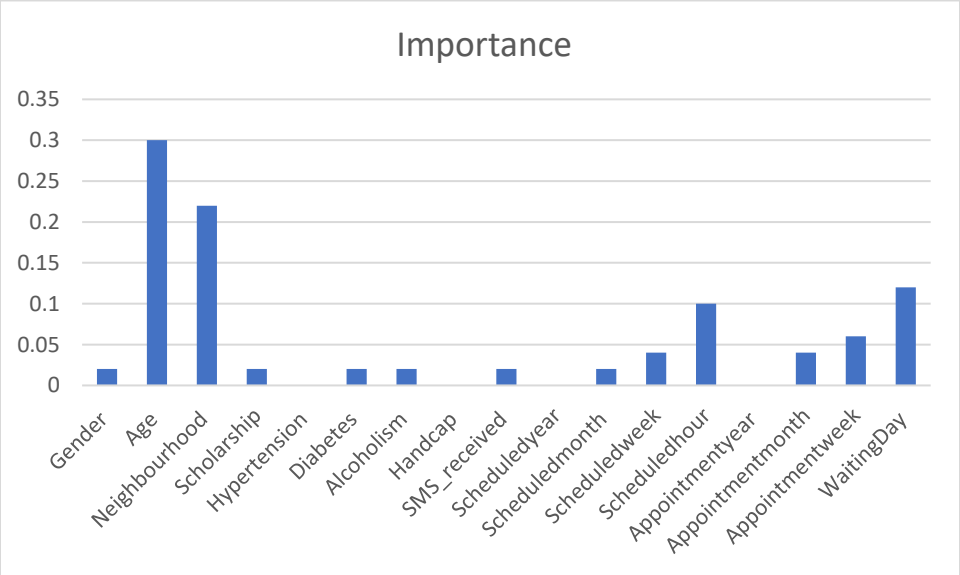
```
Unoptimized model
------
Accuracy score on testing data: 0.7980
F-score on testing data: 0.0096

Optimized Model
------
Final accuracy score on the testing data: 0.7985
Final F-score on the testing data: 0.0000
The best parameters are {'learning_rate': 2.0, 'n_estimators': 200}
The Featuears Importaint:[0.02 0.3  0.22 0.02 0.   0.02 0.02 0.   0.02 0.   0.02 0.04 0.1  0.
 0.04 0.06 0.12]
             precision    recall  f1-score   support

          0       0.80      1.00      0.89     17083
          1       0.41      0.00      0.01      4312

avg / total       0.72      0.80      0.71     21395
```

We can see the algorithms have improvement when we use the best hyperparameter for them.

IV. Results

Model Evaluation and Validation

I choose the RandomForestClassifier to be my final model it this because this regressor generally shows a good balance between results and computational cost, moreover it has important characteristics like variable selection and tunning between variance and bias through parametrization and in general it gives the highest F1 score over the other model, However I try to see how the model work if we reduce the dataset by select just the top 5 features:



And I gives me :

```
Final Model trained on full data
------
Accuracy on testing data: 0.7985
F-score on testing data: 0.0000

Final Model trained on reduced data
------
Accuracy on testing data: 0.7276
F-score on testing data: 0.3251
```

So we can see clearly how the accuracy decreased when we reduce the dataset but the data is not balance so when we focus to the F1 score for the training for the full data was worst but it rise when we reduce the dataset.

Conclusion

Through this project I try to bullied a predictive model to predict the patient who no show up, This model will help to classify the patient to tow groups show and no show up and I evaluate it using confusion matrix :

	precision	recall	f1-score	support
0	0.83	0.82	0.82	17083
1	0.31	0.32	0.32	4312
avg / total	0.72	0.72	0.72	21395

We can see the precision is 31% and this mean there are 31% are no show up patient and the recall 32% which mean this model is capture 32% of the no show up patient , so I need to emphasize my model to raise the recall score to ensure that he will not miss any patient who belong to the no show up class.

Reflection and Improvement

Through this the Capstone Project I I tray to solve the problem of no show up patient for their appointment the dataset was hosted on kaggel and there were 30% of the patient in barizeal has no show up for their appointment so I try to use supervised algorithms to build a predictive model to identify the patient who miss his appointment then this will help use to find the factore that affected this problem so first I clean the data like age I omit the negative values and create a new features like the waiting date which mean the period between the schedule appointment and actual date of appointment , then I did some quick analysis of our data after that I start building the predictive model first I delete some unnecessarily features like the weak name ..etc. then I convert the categories features to number by encoding these features and because now all features are number and these features are start from 0 to 80 so I normalize it by using minmaxscalr() then I splite the data into trining and testing data bu shuffusplit(), after that I choose to deal with : Gaussian Naive Bayes GaussianNB,AdaBoostClassifier,RandomForestClassifieer and GradientBoostingClassifier

First I build these model without tuning their parameters, then by Gridsearch() I find the best hyperparameter for these model, I evaluated by F1 score and the accuracy score however the accuracy score was not useful because the dataset is imbalance so I choose RandomForestClassifier wich has the best F1 score and I choose the top 5 important features by feature_importances_ function and built the model by these features finally I get score low than the benchmark model and really I did not expect this so I think this model need a lot of improvement to be use , in this project I face some Issus one of them was the Neighbourhood name I did not clean this feature because I did not know the real data name and also the dataset it was just for 3 months and this is not a good sample to generalize the solutions, In the future I will

improve theis model by using XGBoost as it is a new and powerful calefaction algorithms and also the hyperparameter need to extend his values to find the good value that will raise the F1 score of the model and also I may split the data using the kfolds as the data is imbalance it will help to divide the dataset , training and teste and validation set.