MEDICAL APPOINTMENTS -NO SHOW PREDICTION

SUBHASHINI NATARAJAN

SEPTEMBER 24, 2022

TABLE OF CONTENTS

BUSINESS PROBLEM

The business problem to be solved by the project is to predict the possibility of no shows for the medical appointment by patients. When patients do not show up for medical appointments, it means the loss of time for medical professionals. It also impacts the hospital management's ability to predict supply requirements and revenue. Hence, it helps in predicting possible no-shows for the scheduled appointments, while addressing the root cause.

BACKGROUND/HISTORY

The stated problem appears to have impacted several not for profit medical centers across the country. The clinics in the absence of donors, solely depend on the patients taken care by them. Many clinics in the absence of any financial mistakes or over hiring have faced losses. Such clinics do have adequate number of appointments taken, however the patients don't show up for the appointment resulting in revenue losses. For such clinics the root cause of the no-show will have to be analyzed and addressed and also predict if the no-shows will happen for the appointments taken as well.

DATA EXPLANATION

The project data is sourced from kaggle. The link to the data set is given below –

https://www.kaggle.com/datasets/joniarroba/noshowappointments

The dataset contains 14 variables with the dictionary defined as below –

- PatientId - Identification of a patient
- AppointmentID - Identification of each appointment
- Gender - Male or Female
- DataMarcacaoConsulta - The day of the actual appointment, when they have to visit the doctor
- DataAgendamento - The day someone called or registered the appointment, this is before appointment of course
- Age - How old is the patient
- Neighbourhood - Where the appointment takes place
- Scholarship - True of False
- Hipertension - True or False
- Diabetes - True or False
- Alcoholism - True or False
- Handcap - True or False

- SMS_received- 1 or more messages sent to the patient.
- No-show - True or False.

METHODS & ANALYSIS

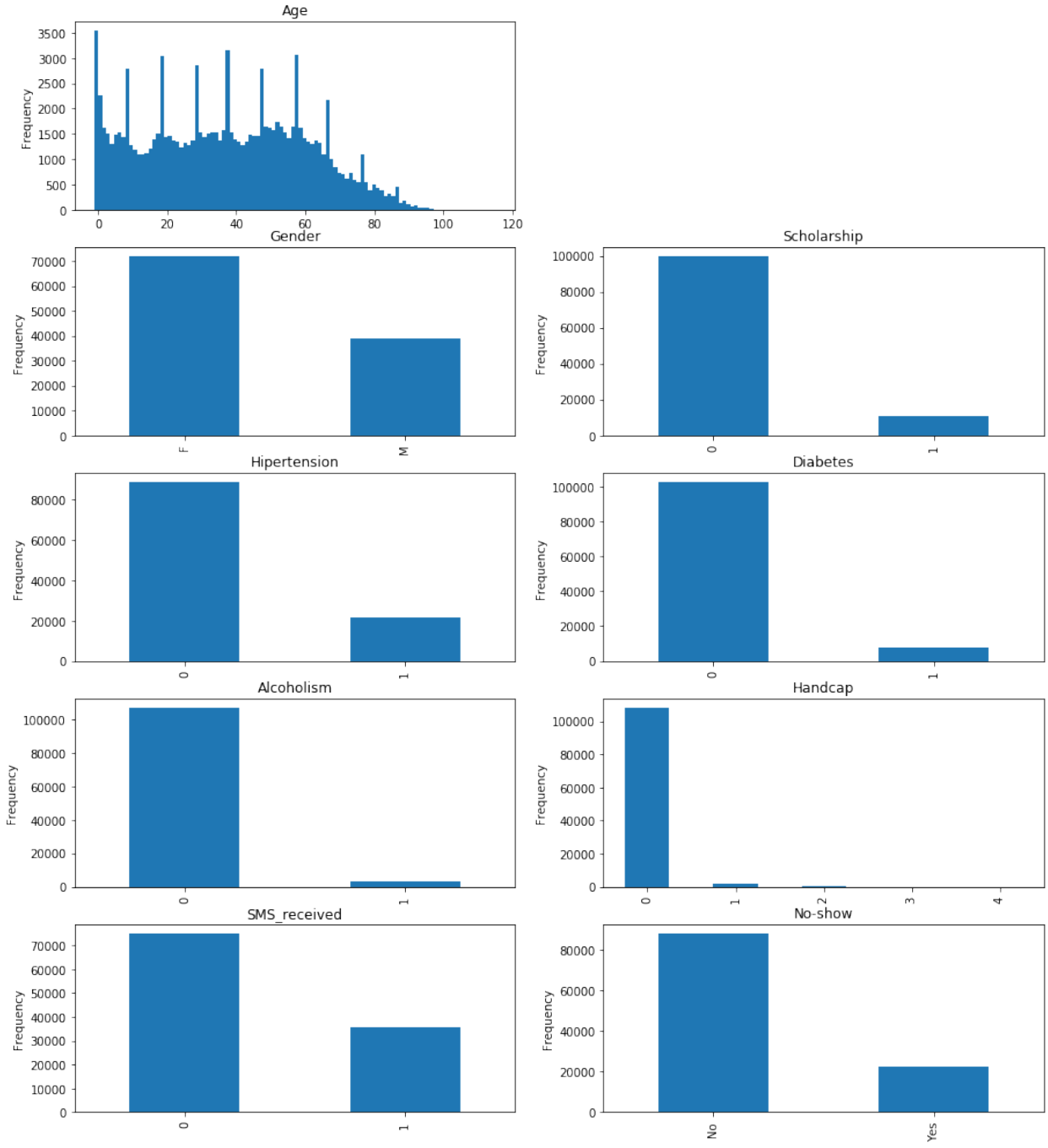
Feature exploration is the first step in solving the problem. The data is loaded into panda's data frame.

The number of unique values in the variables in the dataset is observed to be –

PatientId	62299
AppointmentID	110527
Gender	2
ScheduledDay	103549
AppointmentDay	27
Age	104
Neighbourhood	81
Scholarship	2
Hipertension	2
Diabetes	2
Alcoholism	2
Handcap	5
SMS_received	2
No-show	2

Data Wrangling is then performed necessary data transformations in the dataset. The field, "Age" has a few records with negative values. These records are removed from the dataset. The field, "Gender" with unique values "M", "F" is transformed as 1, 0. The categorical variables in the feature list is converted into 1 or 0 using get_dummies method.

As part of Exploratory Data Analysis – The distribution of dataset to understand the features is obtained and the data looks like the following.



Correlation between the features and No-show is determined as follows –

```
Gender          0.004122
Age            -0.060327
Scholarship     0.029134
Hipertension   -0.035704
Diabetes       -0.015181
Alcoholism     -0.000197
Handcap        -0.006077
```

```
SMS_received      0.126428
No-show_Yes       1.000000
```

MODEL BUILDING & EVALUATION

The variable No-show is regarded the target variable, and the features, Gender, Age, Scholarship, Hipertension, Diabetes, Alcoholism, Handicap, SMS_received are stored as predictor variables. The dataset is then split into training dataset and test dataset using train_test_split() method.

Classification models for the given dataset was built using the following methods- Decision tree classification, Kernel approximation, Random forest classification, Gradient boosting

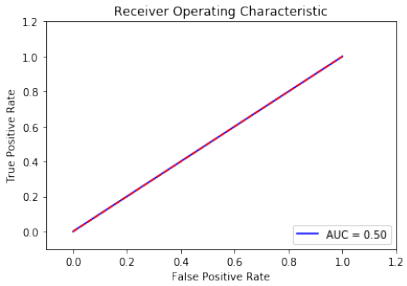
Evaluation for the models is performed by determining the following metrics –

- Accuracy score for test data set
- ROC AUC score for test data set
- Accuracy score on training data set
- Area Under the Precision-Recall Curve

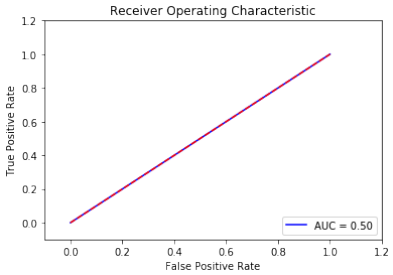
Model	**Accuracy score on test data set**	**ROC AUC score for test dataset**	**Accuracy score on training data set**	**Area Under the Precision-Recall Curve**
Decision tree classifier	0.798601	0.500578	0.798082	0.106990
Kernel approximation	0.798751	0.500000	0.797772	0.399376
Random forest classifier	0.798601	0.500578	0.798082	0.106990
Gradient boosting classifier	0.798480	0.500503	0.798082	0.081912

ROC Curve

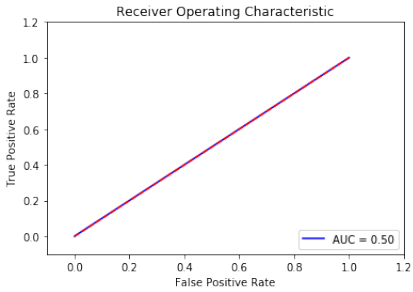
Decision tree classifier



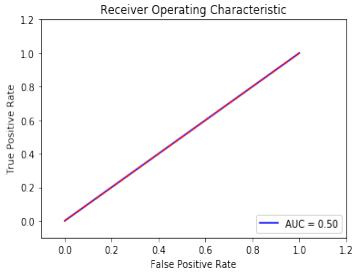
Kernel approximation



Random forest classifier



Gradient Boosting



Feature importance

Among the features used for modeling, the importance of features is determined and is observed to be as follows –

Feature	Score
Gender	0.016332
Scholarship	0.066204
Hipertension	0.070914
Diabetes	0.018665
Alcoholism	0.028442
Handcap	0.033037
SMS_received	0.766407

CONCLUSION

The comparison of the model scores indicate no greater difference among them. Random forest and decision tree can be used for the data set for the prediction. Also, feature importance score indicates that sms_received as the most important feature in determining the no shows. It implies that the no-shows can be significantly improved by making sure that the patients receive sms reminder for their appointments, followed by aiding handicapped patients.

Assumptions

The assumption for the model is that only the stated set of attributes are captured for the patients.

Limitations

The limitation for the model lies in the dataset. The dataset has only limited number of features captured and has only one month's worth of data.

Challenges

The challenge with the models is that the AUC score is just 0.5 even if the model accuracies are close to 80%. The anticipated challenge of class imbalancing has not occurred with the dataset and hence the need to balance the classes is eliminated.

Future uses/Additional applications

Classification models can be applied for use cases like image classification, music classification, email spam filtering, identifying fraud insurance claims etc.,

Recommendations

It can be further assessed to see the possibility of adding more features to improve the model accuracy and AUC score.

Implementation plan

The machine learning models are deployed and scheduled to production in batch mode or online mode. Batch mode is a process scheduled to run daily, weekly or monthly basis. Online mode can be made available to run on demand basis by an online application. This model can be run on a daily basis to predict for the appointment no-show understanding.

Ethical Assessment

Decision making using machine learning models are prone to bias and discrimination. The way to build the model avoiding such bias is by starting with quality/credible dataset. The next step is to build a model that is more interpretable rather than a black box. Conscious and strict evaluation of the model especially when the result interpretation includes age, gender or race.

References

https://machinelearningmastery.com/types-of-classification-in-machine-learning/
https://www.edureka.co/blog/classification-in-machine-learning/
https://www.blackbelt.digital/choosing-the-best-classification-model-for-machine-learning/
https://www.nature.com/articles/s41599-020-0501-9