Medical Appointments No-Show Predictive Analysis

Shaikha Bin Ateeq ,Alanoud Alosaimi

6/1/2022

I. Definition

Project Overview :

The current COVID-19 outbreak has highlighted the scarcity of our healthcare resources, forcing us to shift our mentality and actions to be much more aware of how our personal use of those resources may impact others in need. While the pandemic has brought attention to the problem, it is not new to medical specialists

One aspect of this problem is patients who fail to show up for medical visits, so wasting slots that could have been used by others in need. Only in the United States, medical appointment no shows are estimated to account for 20% to 30% of total visits, with a research estimating their annual cost at $150 billion USD in 2006. (Sviokla, John, Bret Schroeder, and Tom Weakland. 2014).According to scientific studies, when machine learning and data mining approaches were applied therefore these technologies outperformed the traditional management of no-shows (e.g. Srinivas, Sharan, and A. Ravi Ravindran. 2018).

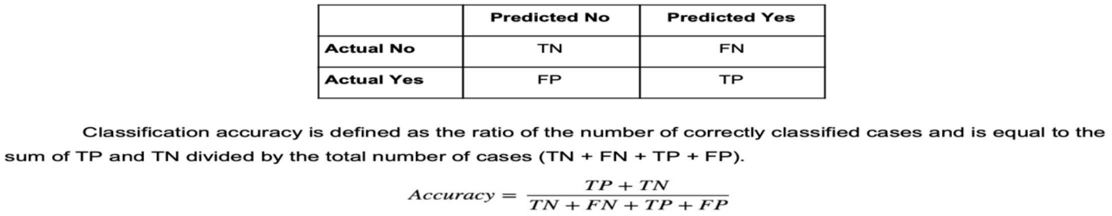
A binary classification task can be used to predict whether or not a patient will show up for a medical appointment. In this project, i will predict whether the patient will show or not, the aim is to help clinics and understand the causes that led to it. The attributes of this model will be addressed in further detail in the following chapter

Problem Statement :

Patients and clinicians both lose time and money when appointments are missed. Patients are frequently required to pay a charge, clinics waste time and effort altering their schedules, and other patients are denied access to care. Clinic staff must spend time contacting all patients with reminders because there is no way to determine which patients are likely to miss their appointments. Clinics can adopt a proactive approach by identifying those patients who are more likely to miss an appointment and focusing their efforts on notifying and/or rescheduling these patients.

Metrics :

The evaluation metrics proposed are appropriate given the context of the data, the problem statement, and the intended solution. The performance of each classification model is evaluated using accuracy



	Predicted No	Predicted Yes
Actual No	TN	FN
Actual Yes	FP	TP

Classification accuracy is defined as the ratio of the number of correctly classified cases and is equal to the sum of TP and TN divided by the total number of cases (TN + FN + TP + FP).

$$Accuracy = \frac{TP + TN}{TN + FN + TP + FP}$$

The evaluation metric for this problem is simply the Accuracy Score , we know the distribution of our target variable is unbalanced, so i will take into account that it must be re-sampled to be balanced to use accuracy , another metric will be F1-Score

II. Analysis

Data Exploration :

The dataset is collects from Kaggle that contain information on more than 110k rows (with no missing values) and 14 variables, and it collects anonymized data and a binary result (show/no-show, our goal variable) of medical appointments in public hospitals in Vitoria, Brazil. The 14 features which are:

Input variables:

- PatientId: Identification of a patient.
- AppointmentID: Identification of each appointment.
- Gender: Male or Female.
- ScheduledDay: The day someone called or registered the appointment.
- AppointmentDay: The day of the actual appointment, when they have to visit the doctor.
- Age: Indicates the age of the patient.
- Neighborhood: Indicates the place of the appointment.
- Scholarship: Indicates whether the patient is enrolled in the Brasilian welfare program.
- Hipertension: True or False
- Diabetes: True or False
- Alcoholism: True or False

- Handcap: True or False
- SMS_received: Indicate if the patients have received SMS massage before their appointment.

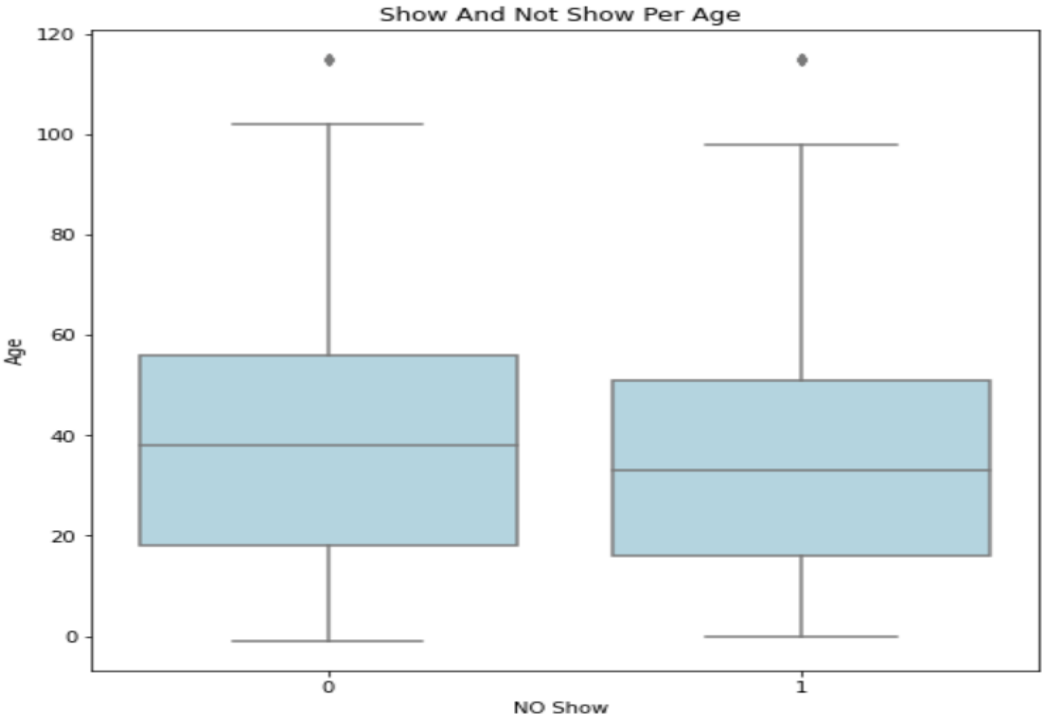
Output variable (desired target):

- No-show: Indicates ‘No’ if the patient showed up to their appointment, and ‘Yes’ if they did not show up.

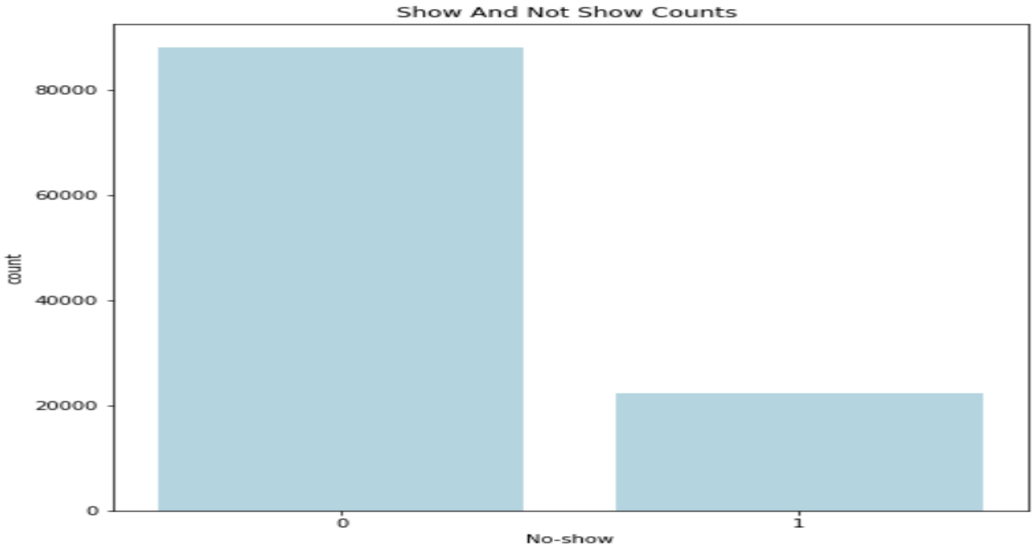
This project focused on investigating the effectiveness of Deep and Wide neural network architectures in predicting medical appointment no-shows

Exploratory Visualization :

The figure below shows us that the age is centered around 20 and 50. Also, any age greater than 100 is considered an extreme value for us



From the figure below, it is clear to us that the distribution of the data is unbalanced, so we will consider that it must be re-sampled to be balanced.



Algorithms and Techniques :

We will prepare the data by splitting feature and target/label columns and also check for quality of given data and perform data cleaning. To check if the model I created is any good, I will split the data into training and validation sets to check the accuracy of the best model As described in above section, there are several non-numeric columns that need to be converted. one of them are simply yes/no or M/F so I will handle it by dummy variable

We can also make subsets of original data using feature scaling techniques to normalize and scale data to try various iterations on the chosen models just to see if we see any differences in performance also I need to do resampling since the data embalmed , So let's pick a few algorithms to evaluate.

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Extra Trees
- Random Forest
- Support Vector Machine

- Naive Bayes (Gaussian)
- Naive Bayes (MultinomialNB)
- XGBoost
- XGBoost (Hyper Parameter Tuning)

Benchmark Model :

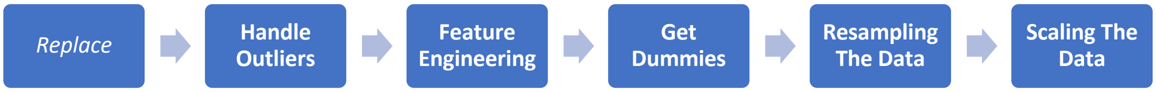
For the benchmark model, we will use the algorithms outlined in the paper "Predicting no-shows in Brazilian primary care " Thepaperdescribesfivedifferent algorithms with the following accuracie

ALGORITHM	ACCURAY
LOGISTIC REGRESSION	0.79
RANDOM FORESTS	0.60
DECISION TREE	0.61
GAUSSIAN NAÏVE BAYES	0.76

III. Methodology

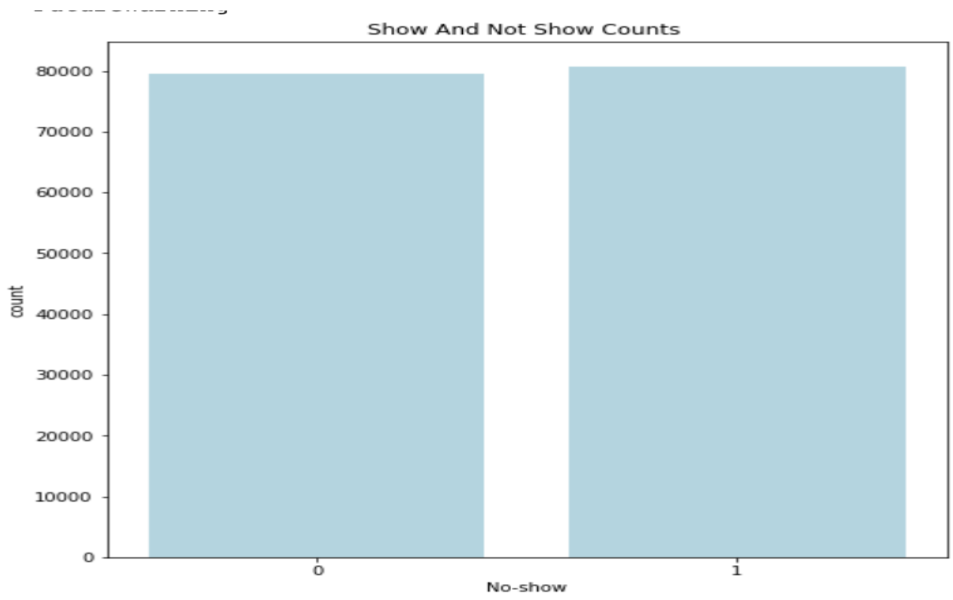
Data Preprocessing

The data cleaning and Preprocessing go through the under the fallowing fig , however Several Data preprocessing steps like preprocessing feature columns, identifying feature and target columns, data cleaning and creating training and validation data splits were followed and can be referenced for details in attached jupyter notebook.



Implementation

The data were split into a training and test, for the training was 0.95 and test 0.05 after that we split again the training into training and validation with validation size 0.05,however since the data imbalanced I start with resampling the data with ratio 4 we can see from the figure below the data is balanced and we can use the accuracy as our matric



After that, i scale the data by using standerscaler() I start with baseline using logistic regression following with rest of algorithms these techniques help to investigate in detail about how is the good classifierRefinement

We used tree-based algorithms to verify accuracy and standard error metrics before selecting the best model. Typically, Random Forest and XGBoost are known to perform well in supervised machine learning applications, and we found that XGBoost has the highest accuracy. One noteworthy finding we

After that i used GridSearchCV to tune the hyperparameters of the XGBoost model in batches. We could keep track of the best results for each parameter and apply them to the next batch in this fashion.

Refinement

i performed hyper tuning of parameters of XGBoost wrapper from scikit-learn library and the parameters tuned were shown in below table:

Parameter	Description	Values Tested	Best Value
max_depth	maximum depth of a tree to control overfitting	[5, 15, 25]	25
n_estimators	The number of trees in the forest.	[100, 300, 500]	500
learning_rate	Learning rate shrinks the contribution of each tree by learning_rate	[0.1, 0.01, 0.05]	0.1

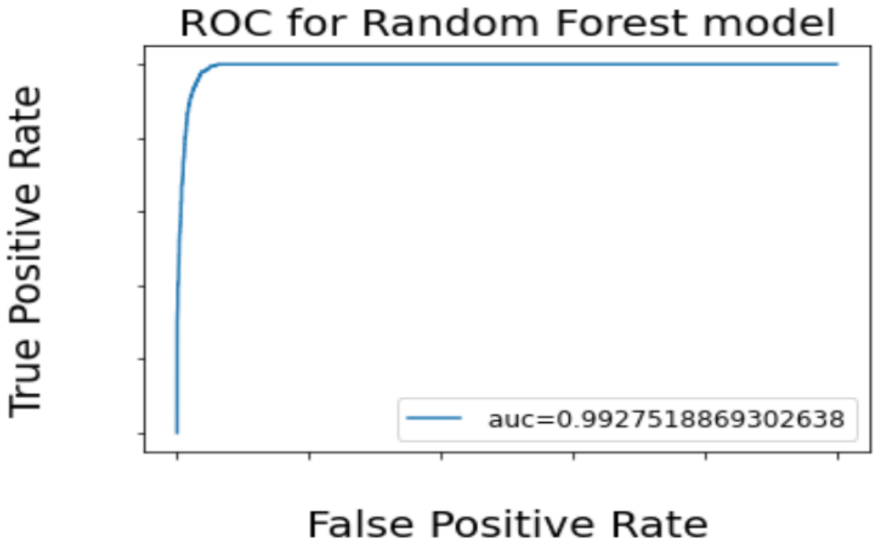
IV. Results

Model Evaluation and Validation

Our end goal was to have a tuned model that have a good accuracy more than the benchmark , however the benchmark try few algorithm also they did not try xgbosst , but we can compare the result with algorithm that we do both of us such as : Logistic Regression , Random Forests, Decision Tree and Gaussian Naïve Bayes

ALGORITHM	ACCURAY BENCHMARK	OUR ACCURAY
LOGISTIC REGRESSION	0.79	0.92
RANDOM FORESTS	0.60	0.96
DECISION TREE	0.61	0.95
GAUSSIAN NAÏVE BAYES	0.76	0.86

So the solution described below is very satisfactory to your initial expectations. we can say that our feature engineering and resampled also scaling play critical rule in our algorithms , However we know that " xgbosst " was the best model for it because it reduces the over-fit in the data , as we see in picture below the ROC show how the area under the curve is high



ALGORITHM	ACCURAY TRAIN	ACCURAY TEST
XGBOSST	0.99	0.96

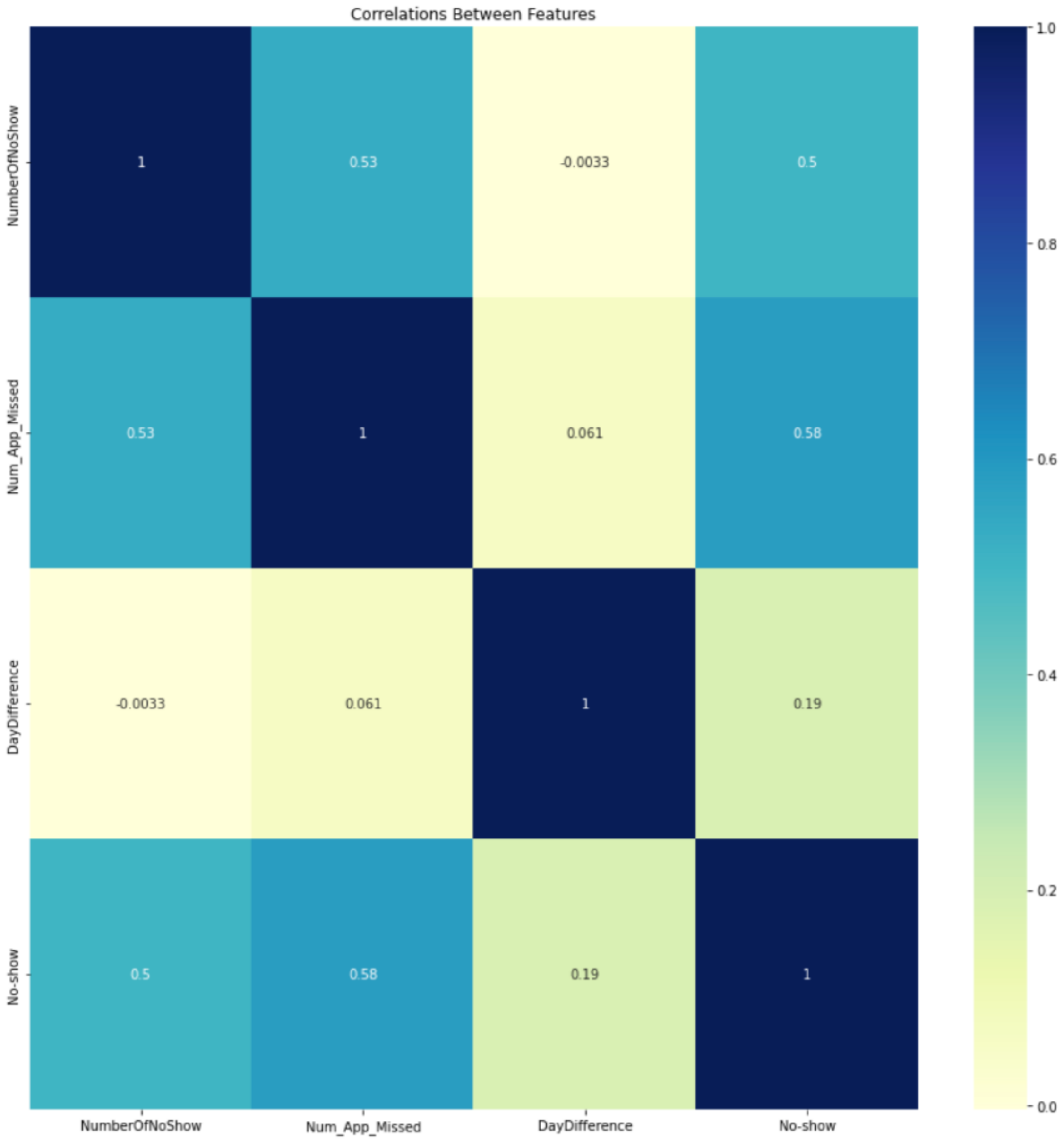
Justification

There is no room for improvement on the final results, the tuned final model made has significant improvement over the untuned model.

V. Conclusion

Free-Form Visualization

The importance of each feature engineering was visualized with the following plot:



Reflection

The most important and time consuming part of the problem was the unbalanced data and its processing. Once the data was prepared and balanced, the accuracy went up. The next challenge was to choose an algorithm that could be best suited to the problem we were choosing to solve. prior knowledge and also the outcome of accuracy on training data, we observed that XGBoost performed the best out of others.

References:

- Sviokla, John, Bret Schroeder, and Tom Weakland. 2014. "How Behavioral Economics Can Help Cure the Health Care Crisis." Harvard Business Review. August 20. https://hbr.org/2010/03/how-behavioral-economics-can-h.
- (2022). from https://theses.liacs.nl/pdf/2017-2018-HelmondsJoep.pdf
- Medical Appointment No Shows. (2022), from https://www.kaggle.com/joniarroba/noshowappointments