Rishabh Shrestha

Analysis of Medical Appointment Show-Ups

Introduction and Description of the features of Dataset:

People register for medical appointments, receive all the instructions but no show up. So who to blame and what factors might play role in it? What if we can priorly predict if the patient shows up or not. Here is a dataset of 110,528 medical appointments with 14 characteristics(variables) of each. The variables of the data set are PatientId, AppointmentID, Gender, ScheduledDay, AppointmentDay, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMS_recieved, and No-show.

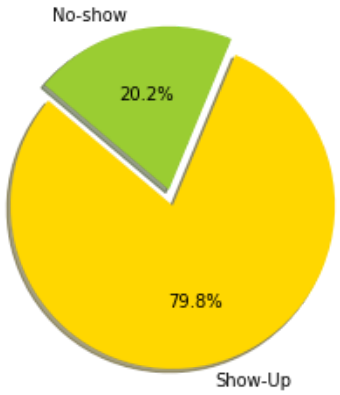
Patient ID contains really high numerical values which is same for the same patient when he/she shows up multiple times. The 'AppointmentID' also contains numerical values which varies by the appointment. We have Gender of the patient as 'F' or 'M' for female and male respectively. 'Scheduled Day', 'Appointment Day' , 'Age' , 'Neighbourhood' are self-descriptive. We have data of medical history like 'Hypertension', 'Diabetes', 'Alcoholism' in form of '0' if false and '1' if true. The variable 'Handicap' has values from 0 to 4 representing the handicap levels of patients. Similarly, we have data if the patient has scholarship for the medical appointment or not. This data column contains 0s and 1s as like of medical history data. Similarly, this dataset also contains data of SMS reminders in the 0s and 1s. Moreover, we have the most important data which is if the patient shows up or doesn't for the appointment(if a patient does not show up then No-show= 'Yes') which is taken as the label for the machine learning algorithm I have used.

The data is pre-processed for the statistical analysis and for creating machine learning model. At first, I have corrected the spelling errors in the field name like 'Hipertension' to 'Hypertension' and 'Handcap' to 'Handicap' to prevent future confusions. The 'No-show' field's values were changed to 0 and 1 for 'Yes' and 'No' respectively. Similarly, the data type of 'AppointmentDay' and 'ScheduledDay' were changed into date and time format. The Gender was also converted into 0 and 1 for female and male respectively. Similarly, the values of the field 'Neighbourhood' was also converted into numerical values. The

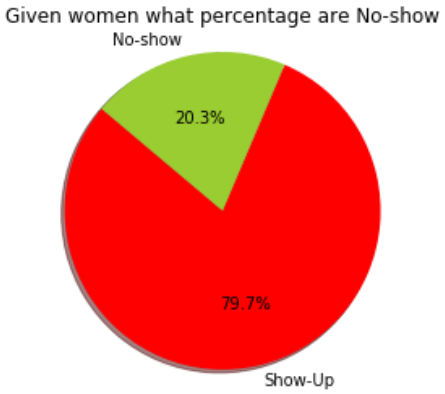
data of negative ages were dropped. Some of the fields were dropped whereas some were added for the further analysis of the data

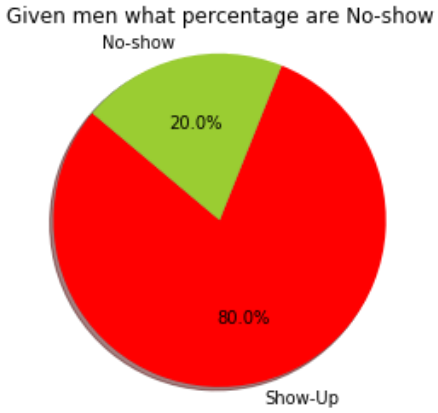
Simple/Descriptive Statistics

In this analytical study of the medical appointment data set, I have used several data analytical methods and machine learning tools and library on the raw data. From the total data set, we can see from the pie-chart below that 79.8% people showed up for their appointment and 20.2% did not.

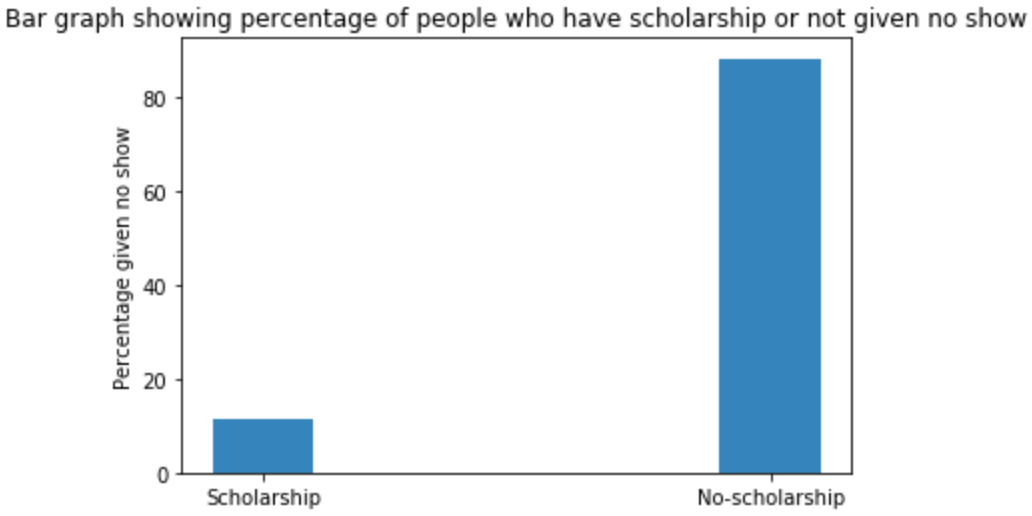


On the basis of gender, we can see that among all females 20.3 % did not show up for the appointment and among males 20 % of them did not show up. So, we can see pretty even distribution among males and females that did not show up.

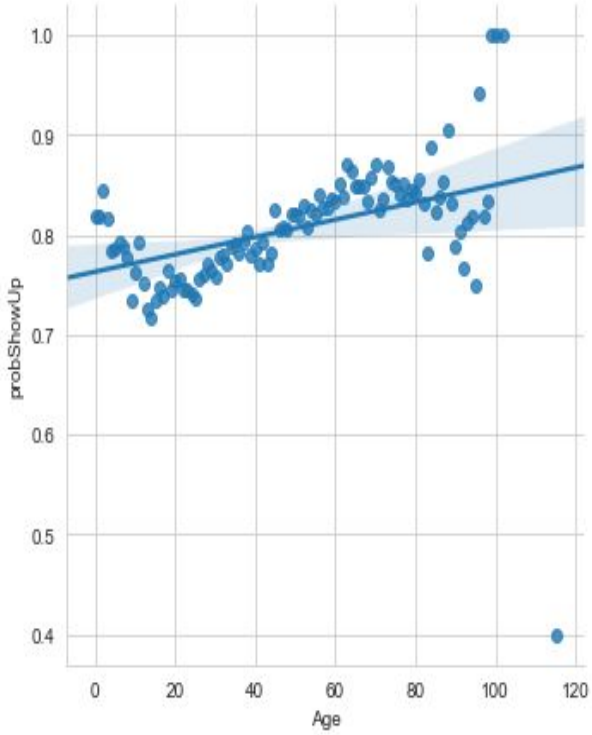




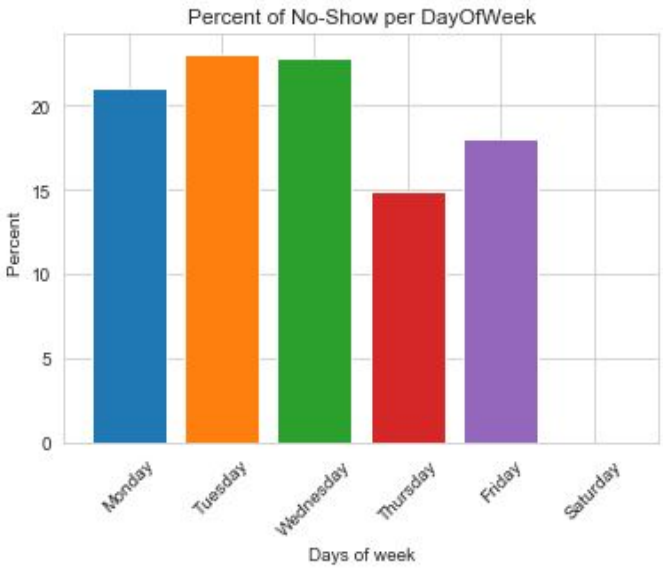
I wanted to analyze and see how does the scholarship affect the likelihood of showing up. From the below bar graph we can see that given a no-show, around 90% people does not have a scholarship whereas only around 10 % people have scholarship. We can see that not having the scholarship increases the likelihood of not showing up.



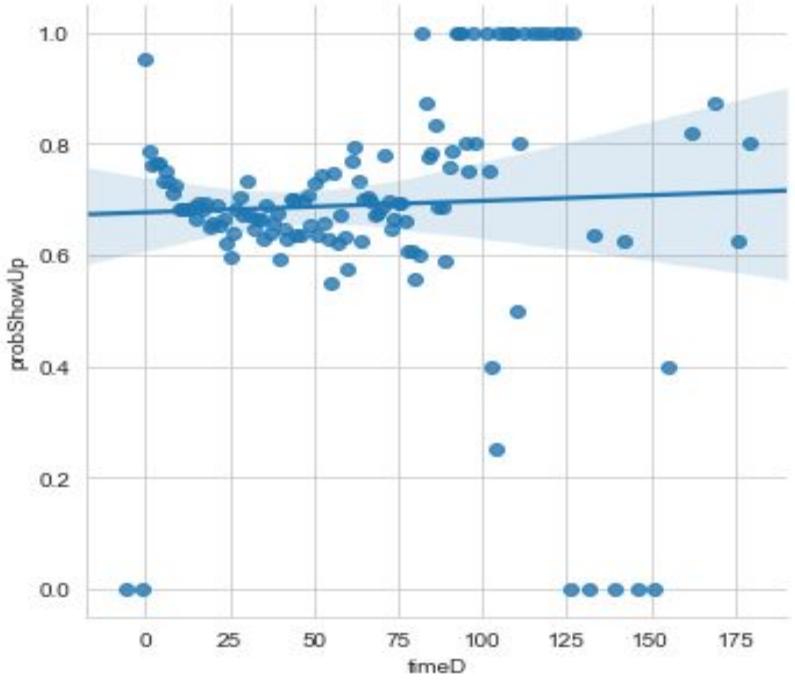
While looking through the age of the patient that signed up for the medical appointment, there were ages in negative which is not possible so I have dropped the data whose age is in negative. For further analysis, I checked how does the age affects the probability of showing up. And, from the graph below, we can see with the help of best fit line that higher the age, higher the probability of showing up.



I wanted to check how the days of the week might affect the probability of no show. For this, I had to do slight pre-processing of the data. I used the appointment date from the data set and converted it into date and time format and then used datetime library to find the weekday of the given date. After analyzing the data, it can be seen from the below bar graph that the percentage of no shows are higher during the start of the week in compare to the later days of the week like Thursday, Friday, Saturdays.



I also analyzed if the time difference between appointment day and scheduled day affects the probability of showing up but it seems like it does not play a big role in it. From the graph below, we can see there is no such relationship between the time difference and the probability of show up.



Prediction using Machine learning:

To predict if a patient shows up for the medical appointment, I have used machine learning algorithms. First, I filtered out the variables like Patient ID, Appointment ID, intuitively that does not affect my prediction model. Then, I analyzed what variables play and don't play role in affecting the likelihood of patient not showing up for the appointment like 'AppointmentDay', 'ScheduledDay', 'timeD',weekday variable and came up with few features to build my model.

The variables I used to build my prediction models are Age, Gender, Neighbourhood, Scholarship, Hypertension, Diabetes, Handicap, SMS_received and No-show. I tried different machine learning algorithms like Naive Bayes, Decision Tree, Random forest classifier, Multiple Perceptron Neural Network and came up with Neural Network to build my model with accuracy of around 76.5 %.

After transforming the dataset into a format suitable for modeling, I built a neural network multiple layer perceptron classifier by importing it from sci-kit learn library. I tried to build my model using Naive Bayes and Decision tree but could not get accuracy as good as using this neural network. The target and labels used were same for the train data and test data.

The pre processed data was then trained using Neural network Multiple Layer Perceptron Classifier. The parameters of the classifier like 'activation' was set to 'relu' which boosted my accuracy than using other activation attributes like 'logistic'. The 'solver' was set to 'adam' which is also the default solver and it works relatively good in pretty large datasets like used in this program. The 'hidden_layer_size' was set to (100,) and 'max-iter' was set to 200 which both are default values. I first started with low values for both considering it might take long time to run but it did not so I ended up with max values. Thus I ended up with the attributes that provide me higher accuracy percentage than any combination I tried. Similarly, the 'random_state' was set to 1. The training data was fit into this model and the accuracy of the test data was calculated. The final accuracy percentage was determined to be around 76.5 %

Future analysis:

For future potential analysis, that could be done using this dataset, is to conduct a causal analysis about the effect of sms messaging on appointment attendance. It might be done by running a simple logistic regression to test the correlation of sms messaging with appointment attendance, controlling for the covariate factors provided (gender, age, scholarship, hypertension, diabetes, alcoholism, handicap, and neighborhood) and a few of my own that I constructed (days between appointment day and scheduled day). Also, I wanted to do an analysis of how SMS messaging system affect the one with low time difference between appointment day and scheduled day in compare to the high time difference.

I would also dig on how the medical history affect the probability of no-show ups. I could check the medical history relation with respect to age and then check the probability of no show ups. The gender and age disparity for medical history like alcoholism, diabetes can be checked in future analysis

Moreover, I could try running my model in another similar type of data set with some more features and check the accuracy of my model.

Note: I have commented in some of the section of the code to help understand better the purpose of the code
I have included two jupyter notebook peview files- one for just prediction model and another for statistical analysis

References:

https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

https://www.kaggle.com/rjonesca/appointment-no-show-notebook

https://www.kaggle.com/joniarroba/noshowappointments/downloads/No-show-Issue-Comma-300k.csv/notebook