Introduction

This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row.

- 'ScheduledDay' tells us on what day the patient set up their appointment.

- 'Neighborhood' indicates the location of the hospital.

- 'Scholarship' indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família.

- Be careful about the encoding of the last column: it says 'No' if the patient showed up to their appointment, and 'Yes' if the patient did not show up

Describtion of the Dataset

I am going to use a CSV file which contains the data I am going to use.

Questions for Analysis

What are the factors to consider in order to know weather a patient is going to show up for his/her appointment.

Data Wrangling

General Properties

The dataset has 110527 rows and 14 colums

This is the summary statistics of the dataset; it gives us an insight into tha statistics of the dataset. If you notice you'll see -1 as the min of age which is an error; the next cell will query that error.

The .info function shnows a metadata of the dataset showing types, number of rows and colums, colum names etc. it helps us to norrow down the content of the dataset and understand it's structure and shape.

The dtypes function shows us the type of each colum specifically. You may see that there are no 'strings' in the types; that is because python recognize and stores 'strings' as object; so where we see 'object' we'll it's a string.

There are no duplicated values in the dataset

There are 62299 out of 110527 unique patient Id

There are 48228 duplicated patient Id

As you can see the column names for 'Hipertension' and 'No-show' has been corrected and change to 'Hypertention' and 'No_SHow' respectively.

The duplicated values have been droped from 110526 rows to 71816 rows

Data Wrangling Summary

In this stage I was able to gather my Data from a CSV file and explore the first few rows to see the sturcture of my data and it's general properties. I was able to see the dimention of the data, checked the data for duplicated values, missing values and possible errors; I found some duplicated values in the 'PatiendId' column and remove them. checked the dataset type to see if there are any furthr missing data to handle. And lastly I was able to see some statistical information about my data on min, max, and mean. And finally I cleaned my data by removing duplicates, unnecessary data, correcting the columns names and dropping unnecessary columns.

Exploratory Data Analysis

Now that I have trimmed and cleaned my data, it's time to move on to exploration. Computing statistics and creating visualizations with the goal of addressing the research questions that I posed in the Introduction section.

The Number of showed is 54153 which is greater than number of no show 17663

Mean age for show is 37 and the mean age for noshow is 34, but the SMS_received mean of noshow is higher than that of show whnich means that we need to look into our messaging strategies.

Investigation for the Influencing Factors on the Attendance Rate

This histogram shows that patient from 0:10 years of age are the most showing and patient from 45:100 and above are the least showing.

The Above figure show that there is correlation between Age and chronic diseases but there is no correlation between Age and attendance.

You'll notice that there is no change in attendance for both sexes in the pie chart. that show that gender does not affect the attendance.

There is no correlation between Gender and Age affecting the attendance; the mean and the median of both sexes are almost the same.

The number of patient showing without receiving SMS is higher than those wowing after receiving SMS.

this shows that only 5 Neighbourhood response to the SMS with ILHAS OCEANICAS DE TRINDADE having the highest response

Patient attendance differs by neighbourhood; AEROPORTO has the highest attendance rate followed by the rest. Despite having the highest SMS received in ILHAS OCEANICAS DE TRINDADE it has the highest no show rate.

Conclusions

Having finished my analysis, I found that Neighbourhood has the highest effect on attendance having some neighbourhood having the greater number of patients and also having the greater number of showing. And some neighbourhood are clearly affected by SMS received and Age.

Also Age has a greater effect on the showing rate as the age range 0:10 has the highest number of shows and the age range 40:100 and above has lower shows.

Lastly I noticed that the number of showing patient by SMS received is lower than those without receiving the SMS. This shows that we need to revisit our SMS sending strategy.

Limitations

Theirs is no clear correlation between showing and chronic diseases, gender and enrolment in the welfare program.

Thank You