Medical Appointment No Shows

Why do 30% of patients miss their scheduled appointments?

Dataset Description

The dataset collects information from more than 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. The Project Medical Appointments No Shows is investigating Medical Appointment No Shows dataset which contains historical data for more than 110K appointments made across different medical facilities in Brazil for more than 60k patients, for each record there are 14 Variables, metadata related to appointments date, patients gender, age, medical condition, social support coverage and facilities and 1 TARGET variable "Wither the patient attended the appointment or not.

Project Target

What is the explanation for a person making a doctor appointment, receives all the instructions and no-show. Who to blame?

Data Definition

PatientId: Identification of a patient that is unique for each person.	Age: How old is the patient.	Alcoholism: T or F.
AppointmentID: Identification of each appointment.	Neighborhood: Where the appointment takes place.	Handcap: T or F.
Gender: Male or Female.	Scholarship: 1 or 0 (this is a program in Brazil to support poor people with their cost of living),	SMS_received: 1 or more messages sent to the patient.
AppointmentDay: The day of the actual appointment, when they have to visit the doctor.	Hipertension: T or F.	No-show: "Yes" or "No" ("No" means they showed up on their appointments while "Yes" means they didn't!).
ScheduledDay: The day someone called or registered the appointment, this is before appointment of course.	Diabetes: T or F.	

Data Definition

```
In [4]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64  
 2   Gender          110527 non-null  object 
 3   ScheduledDay    110527 non-null  object 
 4   AppointmentDay  110527 non-null  object 
 5   Age             110527 non-null  int64  
 6   Neighbourhood   110527 non-null  object 
 7   Scholarship     110527 non-null  int64  
 8   Hipertension    110527 non-null  int64  
 9   Diabetes        110527 non-null  int64  
 10  Alcoholism      110527 non-null  int64  
 11  Handcap         110527 non-null  int64  
 12  SMS_received    110527 non-null  int64  
 13  No-show         110527 non-null  object 
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

```
# Outcomes:
- This shows that we have 14 columns in total besides index, and we have 110527 record.
- None of the data fields have NULL values.
- Datatype should be changed for both "ScheduledDay" and "AppointmentDay".
- PatientId is float while it supposed to be an integer.
```

Data Cleaning

- Edit the "No-show" column to be in positive form instead of negativity.
- Edit the "ScheduledDay", and "AppointmentDay" columns' datatype to be Datetime.
- Edit the "PatientId", "AppointmentID" columns' datatype to be String.
- Remove Row with negative age Value.

Algorithms

- Classification> Logistic Regression Model.
- Regression > Linear Regression Model.

Questions for Analysis

- What is the percentage of patients who show up on their appointments vs. who not?
- Is one gender more committed to medical schedules than another?
- Where do most appointments take place?
- Are patients who received SMS messages reminding them of the appointment likely to attend?
- What is the percentage of patients diagnosed with diabetes, hypertension, alcoholism and disability?
- Is drinking alcohol a cause of missing appointments?
- Is the duration between registration and appointment affect the ability to show up?
- Do older patients more committed to medical schedules than others?

Results

- Percentage of patients who show up on their appointments represents 79.8%
- Percentage of patients who Don't show up on their appointments represents 20.2%-
- Both genders have same commitment to medical schedules.
- "JARDIM CAMBURI" is the most frequent place.
- Patients Who didn't show up have an average of 15 days between registration day and their appointments.
- Patients Who show up have an average of 8 days between registration day and their appointments.
- As Duration increases, the ability of patients to show up on their appointments decreases.
- Older patients are more committed to their appointments' schedules than younger ones.

Tools

- **Pandas**: a library offers data structures and operations for manipulating numerical tables and time series.
- **Numpy**: a library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- **Matplotlib**: a plotting library for the Python programming language and its numerical mathematics extension NumPy.
- **Seaborn**: a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python

```
In [1]: # Import all libraries needed in analysis
        import pandas as pd # data processing, CSV file - Dataframe
        import numpy as np # linear algebra - Arraies
        import matplotlib.pyplot as plt # plotting - Visualization
        import seaborn as sns # Visualization
        from datetime import datetime
        %matplotlib inline
```

Limitations

- Source of data should record timing of sending SMS to Patients to better investigate effect of this on Show ups and why it is not effective as much as expected.
- Most of the data are categorical which made most of plots are meaningless.
- Some patients who marked as no show up, in real they may show up but on another day, this is shown on a different record but this will be better noticed and enhance result if recorded that patient has made a reschedule instead of being recorded as No show.

The End

Thank you