DSA/ISE 5013

Fundamentals of Engineering Statistical Analysis

Fall 2019

Title:

Medical Appointment No Shows

By

Akhila Podupuganti (113461562)

---

1. Abstract:

A no-show occurs when a scheduled patient neither keeps nor cancels the appointment. A cancellation happens when individuals contact the clinic and cancel their scheduled appointments. Such disruptions not only cause inconvenience to hospital management, they also have a significant impact on the revenue, cost and resource utilization for almost all of the healthcare systems [2]. In this project I developed three algorithms namely – ANN, random forests, gradient boosting from which Gradient boosting is giving better results as I concluded with accuracies. And also, ANN and G boosting are in same distribution from T-test on cv results.

2. Introduction:

Patient no-shows permeate medical practices, across specialties, locations, and practice models. While no-shows consistently cause problems for practices, not all practices track their no-show rate or realize the impact that even a couple daily no-shows can have on both their processes and their revenue. There are many reasons that patients fail to make their appointments and there are demographic similarities across patients who no-show more consistently than others. Practice managers work hard to reduce no-shows using a variety of strategies, but often those strategies involve manual processes or difficult-to-enforce policies, resulting in a low impact.

3. Problem Definition and Formulation:

This Project Model helps to predict if patient shows up or not. This would help then to take appropriate actions which can help them from the above-mentioned impacts.

Data Description:

110,527 medical appointments its 14 associated variables (characteristics). The most important one if the patient show-up or no-show to the appointment.

I'm taking data from Kaggle

Data Dictionary

01 – PatientId: Identification of a patient

02 - AppointmentID: Identification of each appointment

03 - Gender: Male or Female. Female is the greater proportion, woman takes way more care of the health in comparison to man.

04 – DataMarcacaoConsulta: The day of the actual appointment, when they have to visit the doctor.

05 - DataAgendamento: The day someone called or registered the appointment, this is before appointment of course.

06 – Age: How old is the patient.

07 – Neighborhood: Where the appointment takes place.

08 – Scholarship: True of False . Observation, this is a broad topic, consider reading this article https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia

09 – Hypertension: True or False

10 – Diabetes: True or False

11 - Alcoholism: True or False

12 - Handicap: True or False

13 - SMS_received: 1 or more messages sent to the patient.

14 - No-show: True or False.

Exploratory data analysis:

	Gender	Age	Scholarship	Hipertension	\
count	110527.000000	110527.000000	110527.000000	110527.000000	
mean	0.649977	37.088874	0.098266	0.197246	
std	0.476979	23.110205	0.297675	0.397921	
min	0.000000	-1.000000	0.000000	0.000000	
25%	0.000000	18.000000	0.000000	0.000000	
50%	1.000000	37.000000	0.000000	0.000000	
75%	1.000000	55.000000	0.000000	0.000000	
max	1.000000	115.000000	1.000000	1.000000	

	Diabetes	Alcoholism	Handcap	SMS_received	\
count	110527.000000	110527.000000	110527.000000	110527.000000	
mean	0.071865	0.030400	0.022248	0.321026	
std	0.258265	0.171686	0.161543	0.466873	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	1.000000	
max	1.000000	1.000000	4.000000	1.000000	

	No-show	missed_appointment_before	...	SANTOS REIS	\
count	110527.000000	110527.000000	...	110527.000000	
mean	0.201933	0.389443	...	0.004949	
std	0.401444	0.487626	...	0.070175	
min	0.000000	0.000000	...	0.000000	
25%	0.000000	0.000000	...	0.000000	
50%	0.000000	0.000000	...	0.000000	
75%	0.000000	1.000000	...	0.000000	
max	1.000000	1.000000	...	1.000000	

	SEGURANÇA DO LAR	SOLON BORGES	SÃO BENEDITO	SÃO CRISTÓVÃO	\
count	110527.000000	110527.000000	110527.000000	110527.000000	
mean	0.001312	0.004243	0.013019	0.016611	
std	0.036197	0.065003	0.113358	0.127811	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	

	SÃO JOSÉ	SÃO PEDRO	TABUAZEIRO	UNIVERSITÁRIO	\
count	110527.000000	110527.000000	110527.000000	110527.000000	
mean	0.017887	0.022148	0.028337	0.001375	
std	0.132541	0.147167	0.165934	0.037059	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	

	VILA RUBIM
count	110527.000000
mean	0.007699
std	0.087409
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

Expected outcomes:

What if that possible to predict someone to no-show an appointment? It is Yes or No.

Statistical Analysis:

Applied few other classification models along with Artificial Neural Network to train on the given data after doing proper data engineering to make the classification.

4. Solutions and Discussion:

Steps Before creating models:

- Reading data from csv file
- Converting No and Yes to 0 and 1 (out target variable *No-Show*)
- Counting the no of time that the Patient didn't show up before and taking as one the feature which is also giving good correlation with target variable.
- Applying one hot encoding for *Neighborhood* variable and Gender to 0 and 1
- Preparing train test data for modeling.

Applied Artificial Neural Network along with random forest and gradient boosting trees.

Artificial Neural Network

```
NN - Accuracy: 83.65%
NN - Precision score: 60.84%
NN - Recall score: 51.35%
NN - F1-score: 55.70%
```

Random Forest

```
RF - Accuracy: 82.73%
RF - Precision score: 57.35%
RF - Recall score: 53.43%
RF - F1-score: 55.32%
```

Top 10 important variables according to Random Forest modeling

	0
missed_appointment_before	0.518212
Age	0.284241
SMS_received	0.023686
Gender	0.021100
Scholarship	0.011664
Hipertension	0.008901
Diabetes	0.007960
Alcoholism	0.005091
Handcap	0.003906
RESISTÊNCIA	0.002887

Gradient Boosting

```
GB - Accuracy: 83.71%
GB - Precision score: 61.41%
GB - Recall score: 50.05%
GB - F1-score: 55.15%
```

Top 10 important variables according to Gradient Boosting modeling

	0
missed_appointment_before	0.941771
SMS_received	0.038758
Age	0.011410
ITARARÉ	0.001534
Scholarship	0.000730
ROMÃO	0.000677
FORTE SÃO JOÃO	0.000640
REPÚBLICA	0.000519
REDENÇÃO	0.000455
Alcoholism	0.000320

To find best algorithm model applying hypothesis T testing on cross validation accuracies

Where we got to know

NN and RF - we accept null hypothesis, are in different distributions

GB and NN - we reject the null hypothesis that assumes that the two samples have the same distribution.

RF and GB - we reject the null hypothesis that assumes that the two samples have the same distribution.

We got better scores in both NN and GB. From hypothesis also, we could say they both are in the same distribution of cv accuracies. And from feature importance *missing appointment before, age, SMS received* are playing more impact on the end result.

5. Conclusions:

Summing up the results from above section we can conclude this data is working great with GB and NN then with Random forests.

6. References:

[1]. Mohammadi,1 Huanmei Wu,1 Ayten Turkcan,2 Tammy Toscos,3 and Bradley N. Doebbeling4 'Data Analytics and Modeling for Appointment No-show in Community Health Centers Iman', J Prim Care Community Health, Nov 17th 2018.

[2]. Adel Alaeddini, Kai Yang, Pamela Reeves, Chandan Reddy, A hybrid prediction model for no-shows and cancellations of outpatient appointments, IIE Transactions on Healthcare Systems Engineering, 5:1, 14-32.