**Data Science Project Protocol**

*Author(s):*

# Introduction

Here you have to give some known facts about the field you will work on.

Try to focus on the problems that are most common and then state the goals of the project.

* You can try to answer to the following question:
* Which questions do we want to answer ?
* What is known about the problem?
* How we define the outcome(s)?
* What is known to influence the outcome?
* Does we have any possible new knowledge that has not been in use before?

This part must be between half to one and half page.

The goal of project is to build a model in order to predict the probability of having a no-show scenario on a future appointment in a hairdresser’s salon.

Before delving into the specific scenario of a hair salon, here a few facts on no-shows that can help us to understand the far-reaching impact of missed appointments on service-based businesses

At first in which situations can occur a no-show?[[1]](#footnote-1)

* When a person fails to attend a meeting he was supposed to, or buys a ticket to a public performance (a concert, for example) and does not use it.
* When a passenger neither cancels his or her reservation nor shows up for the flight. The airline will cancel his or her all other reservations (such as for connecting flights, if any) if not otherwise notified, and usually will not refund any portion of the ticket price.
* When a traveler fails to show up for a reserved room without notifying the hotel of the cancellation
* When a shipper fails to deliver a booked shipment to the carrier. He or she may still have to pay the full freight.

No shows generate a loss of productivity/effective time allocation and of revenues. No-shows in addition hurt the rest of your customer base by increasing waiting times for getting an appointment.

In order to understand the magnitude and the importance of this area of study, we’ll discuss quickly no-show impact on health care systems where it is a critical issue.

Medical appointment no-shows are a plague in the public health care field. As show in a BMC health service system study[[2]](#footnote-2), they may represent up to 18% of all the appointments. Each missed appointment cost in 2008 the US health care system $200.

Therefore, reducing the rate of no-shows is a top priority around the world.  Various strategies[[3]](#footnote-3) have been conducted in order to determine the ways to reduce no shows such as call reminders, behavioral engagement strategies (e.g. contingency management), reducing waiting time (e.g. double-booked appointments, centralized appointment scheduling, etc..) increasing capacity (adding appointment slots), improving environment, etc.

# Methodology (Project design)

## Data

Here you have to describe how do you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.

1. according to businessdirectory.com [↑](#footnote-ref-1)
2. US National library of Medicine https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4714455/ [↑](#footnote-ref-2)
3. US National library of Medicine https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962267/ [↑](#footnote-ref-3)