**Hair Salon NoShow**

**Data Science Project Protocol**

*Author(s):*

*Eti Mayer*

# Introduction

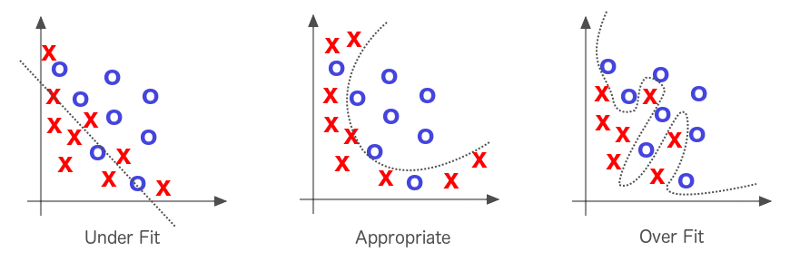
T

he phenomenon of a client missing an appointment without a notification or cancelled the same day as the appointment, is called a “No-show”. No-shows are missed revenue. It’s a waste of time and resources and could have a big impact on business moreover on a small business relying on each customer. Most articles are dealing with the medical no-show phenomenon. There are dozens of studies analyzing it with various machine-learning techniques. Going to a hair salon is kind of a luxury, it’s not a lifesaving treatment therefore not showing up to an appointment or cancelling in the last minute became the reality for hair stylists. A reality I want to explore and see if any outcome can help a hair salon reduce the damages, by predicting which client’s appoint have a high probability to be a no-show.

By showing a probability for a binary results, where NoShow = 1, Show = 0, the outcomes can help the salon understand its no-show and take active steps to diminish it.

There are three main issues in my dataset which can influence the outcomes:

1. Small Dataset – Dealing with small dataset can lead to an overfitting. Generalizing patterns from a small train set can produce amazing outcomes which become soon to a poor preforming in predicting new data:



Source: https://medium.com/@shubhapatnim86/generalisation-training-validation-test-data-machine-learning-part-6-1de9dbb7d3d5

1. Imbalanced dataset - I found that 170 out of 2184 observations are no-shows, only 7.7% of the data. This can leads to profoundly good outcomes becoming useless in production
2. Multicollinearity – The target variable can be extracted from some variables in the data, which can cause a bias and overfitting and require treating the data carefully avoiding over/under fitting.

# Methodology (Project design)

## Data

The Hair Salon No-Show Dataset was taken from Kaggle and came from a small hair salon in Toronto, Canada. Since most of the hair salons are small businesses, gathering data from several salons can enrich our data. Conducting questionnaires among clients can enrich our data with socioeconomic data which is missed in the dataset and can show a different aspect to the no-show phenomenon.

The dataset includes 6 csv file with total of 68 columns, from which I extracted flat-file of 2,184 observations and 60 columns, with a time-frame of 11 months, including the final appointment status indicating if a client was seen, cancelled the appointment, or no-showed which became the binary target variable. The data consists primarily of attributes of the appointment booking itself and excludes client details or details on the staff member providing the service. As mentioned above, it’s a small, imbalanced dataset therefore partitioning the data was not based on time-frame.

The data includes bookings and cancellation information to determine whether a given booking resulted in a no-show wherein the client either didn't show up at all or canceled the appointment within 48 hours of the planned booking (i.e., an out-of-policy cancellation). This ‘cancel\_days’ feature affect directly on the target variable ‘NoShow’, which can cause a bias or overfitting issues. I will examine this during the project and will get rid of the ‘cancel\_days’ feature in case of an overiftting, since it’s a great predictor (maybe too good).

For exploratory data analysis (EDA) I used the ‘Mechkar’ library in R and pandas ProfileReport to create a full report of distributions, outliers and correlation heatmap. I’ve analyzed the outliers and missing values to clean the data and prepare it for feature engineering and modeling. At the beginning of ome of the outliers can be extract

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 the limitations of the model, when it can be used and when not.