**Hair Salon NoShow**

**Data Science Project Protocol**

*Author:*

*Eti Mayer*

# Introduction

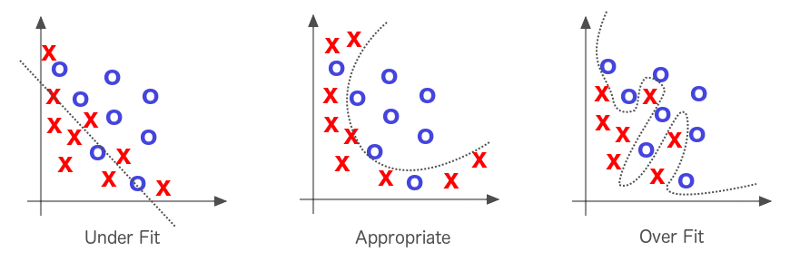
T

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. Preparing the data included categorizing and transforming the outliers and missing values.

## Models

NoShow data contain 2,184 observations – 2,014 are show = 0, 170 are NoShow = 1. A small, imbalance dataset. I will divide the dataset to train and test with proportions of 0.4 to test set and 0.6 to train. Before training the model will use different sampling techniques to balance the data, such as under-sampling, over-sampling, and synthetic sampling (ROSE, SMOTE). Then use the chosen technique for training the train set. By partitioning the available data into three sets, we drastically reduce the number of samples which can be used for learning the model, and the results can depend on a particular random choice for the pair of (train, validation) sets. Therefore using cross-validation techniques, a computationally expensive approach, it does not waste too much data, which is a major advantage in problems such as inverse inference where the number of samples is very small. As mentioned above, I have classification problem with large imbalance in the distribution of the target class. Evaluating different cross-validation techniques shows that stratified sampling implemented in StratifiedKFold (skf) and StratifiedShuffleSplit (sss) can ensure that relative class frequencies is approximately preserved in each train and validation fold.

The scoring parameter for classification problem is depended on data balance. In my case the data is Imbalance and I want to balance between positive and negative predictions, therefor I will use AUC as the metric for choosing the best model and fine tuning.

Model choose - ADAboost

## Deployment of your model

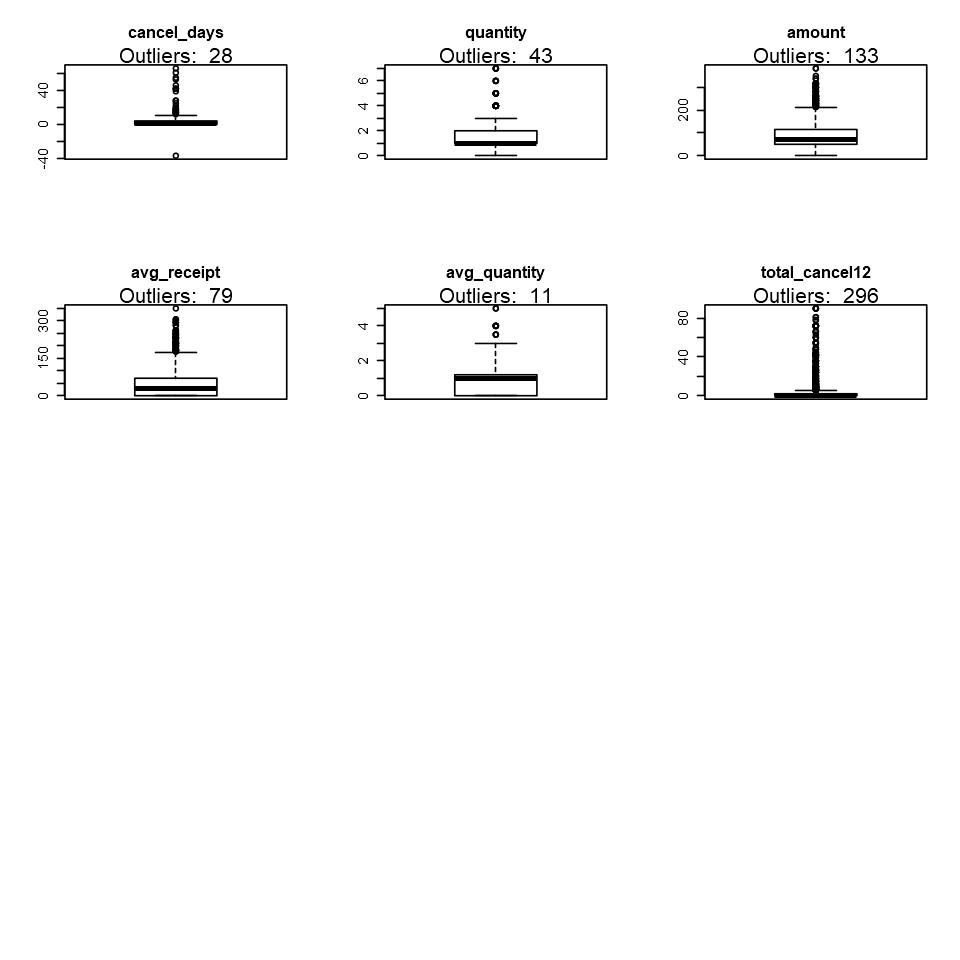
* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* 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

Total observations = 2,184 ; Train set = 1,310 ; Test set = 874

Timeframe = 11 months during 2018

**Outliers**

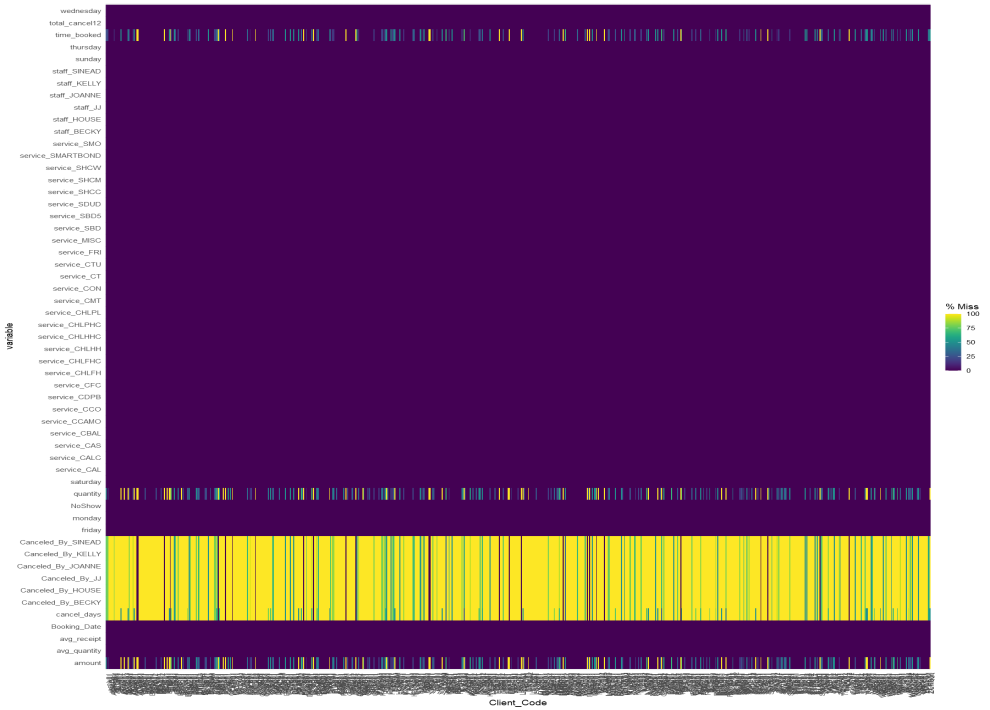


Comparing the distributions with/without outliers showed me that some of the distributions were about to change, transformation was done to avoid such case. In cases where the distribution and the result of the model didn’t change, I used categorization method.

Here are the conclusions for outliers’ treatment:

|  |  |  |
| --- | --- | --- |
| Variable | Treatment | Why |
| cancel\_days | Leave | It changes both the assumption the results |
| Cancel\_days=-37 | Drop | A definite mistake. canceling after booking day passed |
| Quantity | Leave | It changes both the assumption the results |
| Amount | Leave | It changes both the assumption the results |
| Avg\_recipt | Drop | It changes the assumption but dosen't change the results |
| Avg\_quantity | Drop | It changes the assumption but dosen't change the results |
| Total\_cancel12 | Leave | It changes both the assumption the results |

**Missing values**



All missing in 'cancel\_By\_(staff)' should be 0. If the data is missing it means that it wasn't canceled by this staff member. As for the rest of the missing values I determined the missingness generation mechanism

There are three mechanisms that generate Missing values:

1. Missing Completely at Random (MCAR)

2. Missing at Random (MAR)

3. Missing not at Random (MNAR)

By the distribution and the t-test I assume that the missing mechanism is MNAR. Here are the conclusions:

|  |  |  |
| --- | --- | --- |
| Variable | Mechanism | Treatment |
| Amount | MAR | Imputation Using Mice |
| Time\_booked | MAR | Imputation Using Mice |
| Avg\_receipt | MNAR | Categorize |
| Avg\_quantity | MNAR | Categorize |

# Conclusion

As a final project to Data Science course we had to choose data to work with form raw tables until final prediction. At first I thought that only a data with great meaning will produce a high quality project. So I took historical data from the World Bank trying to find out if a renewable energy use will decrease CO2 emissions, sounds great and trendy. Unfortunately I had 99% missing values and no time to concentrate the exact data for my project. I decided to work with hair salon data, although it was far from what I wanted to achieve.

Through exploratory data analysis, feature selection, data preprocessing, modeling and fine tuning the best model I found out I was wrong! It was one big challenge the put me through intensive couple of months of deep learning, by deep learning I mean that I took it as a great opportunity to go deeper in every step necessary for the project. I’d wanted to understand what I was doing. Thou it took me more time, I feel I can submit my outcome knowing I did my best to get the best results I can. Facing collinearity, overfitting, imbalance and me more challenges kept me alert to many kinds of issues that come with a data.

It is now clear to me that facing data, even just a hair salon data, can be fascinating.