**Data Science Project Protocol**

*Author(s):*

# 

# Introduction

The no show customers problem is common to many areas such as Airline, Hotels, Hospitals, Clinics, Medical Facilities, Restaurants, Hair Salons, etc.

No-show appointments, defined as an appointment in which the customer/patient did not present for treatment or cancelled the same day as the appointment.

**The no show problem causes to:**

* Long waiting times
* a lower quality of care.
* Dissatisfied staff

Because you want to satisfy your customers all the time and you don’t want to lose a deal, you could offer your customer the means to rebook the appointment the same day or squeeze some extra time between two other appointments. To do so, you’ll have to ask your team members to work overtime or to change their schedules to cater to another customer. You’ll end up upsetting your staff to keep your customers happy.

* Force the provider to face constant uncertainty.
* Unsuccessful revenue management Specifically, cancellations made close to the time of service are the most damaging because they leave management with no time to react.
* reduce revenue / lost revenue opportunity
* Raises financial loss
* Affects the decision-making process
* Increases the frustration of customers in receiving the service(In cases where a provider make over-Booking)
* Causes customer abandonment

**What can cause no-shows?**

* Forgetfulness
* Wrong slot time

Customers didn't forget they have an appointment to visit one of your condos or that their dog has an appointment for a vaccine. However, they thought it was a day later or at another time on the same day. Chances are, these “new” time slots are already booked. Inevitably, this results in no-shows for you.

* Last-minute emergency

Customers can run into all sorts of last-minute emergencies: they have to stay home to take care of a sick kid, overtime was required at work, water damage occurred at home, etc. There is a lot of things that could happen at the last minute to prevent your customers from showing up to their appointment.

* Service not required anymore

They don’t hear the noise in their car anymore or their cat is miraculously feeling better. By the time the appointment is supposed to take place, they don’t seem to have an issue anymore; their appointments are not needed so they decide to not show up and forget to let you know.

* Financial trouble

They know they won’t be able to handle the bill at the end of the appointment so they simply don’t go to their appointments and don’t tell you.

* Bad weather
* Unexpected traffic domestication
* Customer in a bad mood

**How to prevent no-shows**

A simple way to prevent no-shows by Communication,

Communication is the key to reducing no-shows.

One effective way is to use friendly text reminders. Send a message to customers to confirm that they are coming to the appointment two days before it occurs is an example.

If the appointment is not needed anymore, a customer can easily answer to the text and you will know right away.

The customer could ask for another time and you will be able to find the most convenient spot at a later date.

Another way is to confirm by text the date and time of each appointment. That way, customers know that they always have the info in their cell if they are not sure about the date and time. They can also create an appointment into their phone calendar directly from the message.

Finally, by providing a simple communication channel to your customers, you’ll increase the chances that they will let you know that they can’t make it to their appointments.

**Definition the outcome**

Which client bookings are most likely to either

* no-show
* cancel within 48 hours of the scheduled appointment time.

False positives are okay as long as we can optimize the number of actual no shows predicted in test.

# Methodology (Project design)

## **Data**

**Data Sources Description**

1. Client Cancellations0.Csv
2. Future Bookings (All Clients)0.Csv
3. No-Show Report0.Csv
4. Product Listing (Retail)0.Csv
5. Receipt Transactions0.Csv
6. Service Listing0.Csv
7. Hair\_Salon\_No\_Show\_Wrangled\_Df.Csv
8. Client Cancellations0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| Cancel Date | The date the customer canceled the appointment. |
| Code | The unique client code. |
| Service | The service code for the booking. (e.g., SHCW - women's hair cut) |
| Staff | The staff member to provide the service. |
| Booking Date | The date the service is scheduled to be provided. |
| Canceled By | The staff member who canceled the service. |
| Days | The number of days between the Booking Date and the Cancel Date. |

1. Future Bookings (All Clients)0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| Code | The unique client code. |
| Staff | The staff member to provide the service. |
| Service | The service code for the booking. (e.g., SHCW - women's hair cut) |
| Date | The date the service is scheduled to be provided. |
| Time | The time the service is scheduled to be provided. |
| TimeInt | UNUSED |

1. No-Show Report0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| Date | The date the service was scheduled to be provided. |
| Code | The unique client code. |
| Service | The service code for the booking. (e.g., SHCW - women's hair cut) |
| Staff | The staff member who was to provide the service. |

1. Product Listing (Retail)0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| IsActive | Is this an active product? |
| Code | The unique product code. |
| Description | The product description. |
| Supplier | The product supplier. |
| Brand | The product brand. |
| Category | The product category. |
| Price | The regular price. |
| On Hand | The number of units on hand. |
| Minimum | UNUSED |
| Maximum | UNUSED |
| Cost | The unit cost of the product. |
| COG | The total cost of all units. |
| YTD | YTD sales. |
| Package | UNUSED |

1. Receipt Transactions0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| Receipt | The receipt number. |
| Date | The date of the transaction. |
| Description | The service or product name. |
| Client | The unique client code. |
| Staff | The staff member who provided the service or sold the product. |
| Quantity | The number of services or product sold. |
| Amount | The total dollar amount. |
| GST | Federal tax amount. |
| PST | Provincial tax amount. |

1. Service Listing0.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| IsActive | Is this an active service? |
| Code | The service code. (e.g., SHCW - women's hair cut) |
| Desc | The service name. |
| Cate | The service category (e.g., Balayage is in the COLOR category.) |
| Price | The regular price of the service. (Note: price varies across staff so this is not precise.) |
| Cost | This is the amount the staff pays to the salon for professional product costs. |

1. Hair\_Salon\_No\_Show\_Wrangled\_Df.Csv

|  |  |
| --- | --- |
| Column Name | Column Description |
| # | Booking index. |
| book\_tod | The booking time of day. |
| book\_dow | The booking day-of-week. |
| book\_category | The booked service category (COLOR or STYLE) |
| book\_staff | The staff member to provide the service. |
| last\_category | The client's last booked service category before the current booking or today whichever is greater. |
| last\_staff | The staff member who provided the client's last service before the current booking or today whichever is greater. |
| last\_day\_services | The number of services provided to the client on their last visit before the current booking or today whichever is greater. |
| last\_receipt\_tot | The amount paid by the client on their last visit before the current booking or today whichever is greater. |
| last\_dow | The day-of-week of the client's last booking before before the current booking or today whichever is greater. |
| last\_tod | The time-of-day of the client's last booking before the current booking or today whichever is greater. |
| last\_noshow | Did the client no-show on their last booking before the current booking or today whichever is greater? (0 - no, 1 - yes) |
| last\_prod\_flag | Did the client buy a retail product on their last booking before the current booking or today whichever is greater? (0 - no, 1 - yes) |
| last\_cumrev | The client's cumulative service revenue as of their last booking before the current booking or today whichever is greater. |
| last\_cumbook | The client's cumulative number of bookings as of their last booking before the current booking or today whichever is greater. |
| last\_cumstyle | The client's cumulative number of STYLE bookings as of their last booking before the current booking or today whichever is greater. |
| last\_cumcolor | The client's cumulative number of COLOR bookings as of their last booking before the current booking or today whichever is greater. |
| last\_cumprod | The client's cumulative number of bookings with retail product purchases as of their last booking before the current booking or today whichever is greater |
| last\_cumcancel | The client's cumulative number of appointment cancellations as of their last booking before the current booking or today whichever is greater. |
| last\_cumnoshow | The client's cumulative number of no-shows as of their last booking before the current booking or today whichever is greater. |
| noshow | Did the client no-show or execute an out-of-policy cancellation for this booking? (0 - no, 1 - yes) |
| recency | The number of days since the client's last booking before the current booking or today whichever is greater. |

**DataSet Problem:**

|  |  |  |
| --- | --- | --- |
| Data Set | MinimumThe date the service was scheduled to be provided. | Maximum The date the service was scheduled to be provided. |
| Client Cancellations0 | *29/03/2018* | *21/09/2018* |
| No-Show Report0 | *14/03/2018* | *29/07/2018* |
| Future Bookings | *14/03/2018* | *15/02/2019* |

* The number of records does not match between Hair\_Salon\_No\_Show\_Wrangled\_Df.Csv when noshow variable = 1 and No-Show Report0.Csv and Client Cancellations0.Csv when Days variable <= 2
* Its Unable to know if the records from Hair\_Salon\_No\_Show\_Wrangled\_Df base on No-Show Report0 and Client Cancellations0 and Future Bookings.
* The number of records does not match between Future Bookings dataset and No-Show Report0 dataset and Client Cancellations0 dataset when Days variable <= 2 and Hair\_Salon\_No\_Show\_Wrangled\_Df when noshow variable = 1
* The number of records does not match between Future Bookings dataset and Hair\_Salon\_No\_Show\_Wrangled\_Df

In my opinion If the variables were included in the data set

* **Gender**
* **Age**
* **Neighbourhood**(Gives an indication of socio-economic status)
* **SMS\_received**

**External Data Sources**

* Toronto Weather,Canada DataSet (<https://climatedata.ca/download/>)

I was thinking of adding a feature that would describe the weather on the day of the service.

The goal is to divide the weather into categories and perhaps find a connection between cancellation and weather

**The Outcome Variable**

* noshow

Did the client no-show or execute an out-of-policy cancellation for current booking?

(0 - no, 1 - yes)

The data for model can be use only from Hair\_Salon\_No\_Show\_Wrangled\_Df.Csv file.

When I wrote this paragraph, still do not know how to relate to outliers so I will follow that rules.

rules for droping outliers

* If the outliers is obviously due to incorrectly entered or measured data
* If the outliers not change the results but does affect assumptions
* If the outliers creates a significant association

If not droping outliers I will try to using non-parametric models decision trees to give an answer for outliers problem .

**The Missing Values**

I will try to Imputation missing values in 2 ways:

* Do Nothing , I will let the algorithm handle the missing data.
* Do imputation using deep learning (Datawig <https://github.com/awslabs/datawig>)

After running the model in both modes (Do Nothing and Do Datawig) , I will compare the results.

**data exploration strategy**

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

## Models

## **divide No-Show data**

NoShow dataset containing 1952 record

* Option 1

|  |  |  |
| --- | --- | --- |
| **subsets** | **Data Percent** | **Record Count** |
| train | 65% | 1203 |
| dev | 15% | 277 |
| test | 20% | 390 |
| for late evaluation |  | 82 |

* Option 2

|  |  |  |
| --- | --- | --- |
| **subsets** | **Data Percent** | **Record Count** |
| train | 70% | 1366 |
| test | 30% | 586 |

I will check the differences between 2 options.

I will use cross-validation technique to apply the model.

The techniques I will use to apply the model outcome is Regression

**Data balance**

* Resampling (Oversampling and Undersampling)
* Ensembling Methods (Ensemble of Sampler)

## **I Will try :**

## **Hypothesis 1**

The noshows are spread uniformly accros the weekdays and daytime.

## **Hypothesis 2**

The noshows are spread uniformly across those clients who sets apponintments with the master they’d been working before and those who chooses a new one.

## 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 will the prediction 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 will the model be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which measurements will be used to evaluate if the prediction is decaying?.
* How will the model’s prediction be adapted for the needs of the production system?.

# Results

Here you will present the main results of all the processes. 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:** the best linear model is LASSO

# 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.

.