

Machine Learning & Data Mining

Final Project

**Hotel Booking Cancellations**

Professor: Boaz Lerner

Students: Oriel Perets & Dafna Meron

Date:

**Abstract**

unexpected cancellation of hotel bookingsis a disturbing issue for online reservation websites such as Booking.com, Trivago, and Hotels.com, since it results in unwanted vacancy and lost profits. Anticipating whether an order will be canceled, even hours in advance, can help these companies act, for example: double-book certain rooms, create incentives for arriving customers with a high cancellation chance, send reminders and confirmation emails, alter cancelation terms, fees, and deposit types.   
In this project, we will explore these problems and solutions with classification models.

Contents

[**Business Understanding** 4](#_Toc91702028)

[**Data Understanding** 4](#_Toc91702029)

[Collect data 4](#_Toc91702030)

[Describe data 4](#_Toc91702031)

[Explore data 4](#_Toc91702032)

[Verify data quality 4](#_Toc91702033)

[**Data Preparation** 5](#_Toc91702034)

[Select data 5](#_Toc91702035)

[Clean data 5](#_Toc91702036)

[Construct data 5](#_Toc91702037)

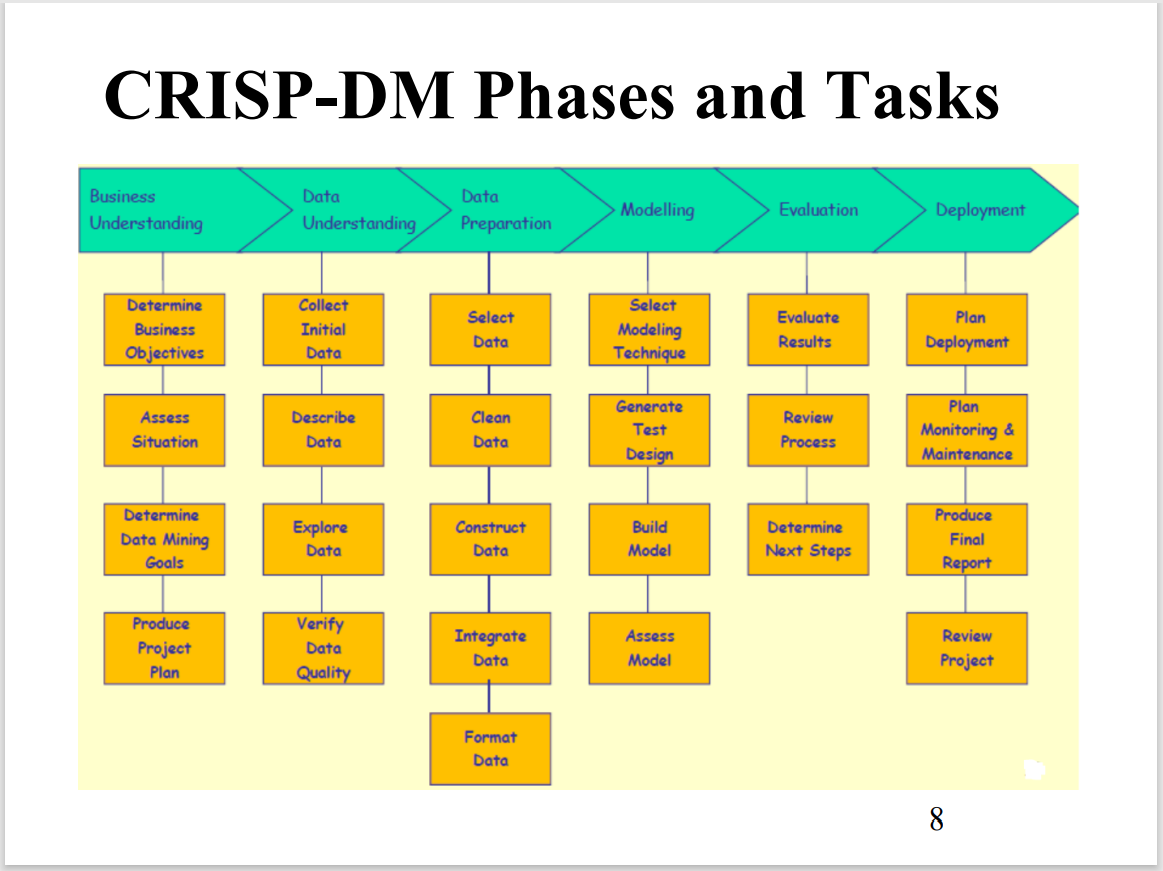
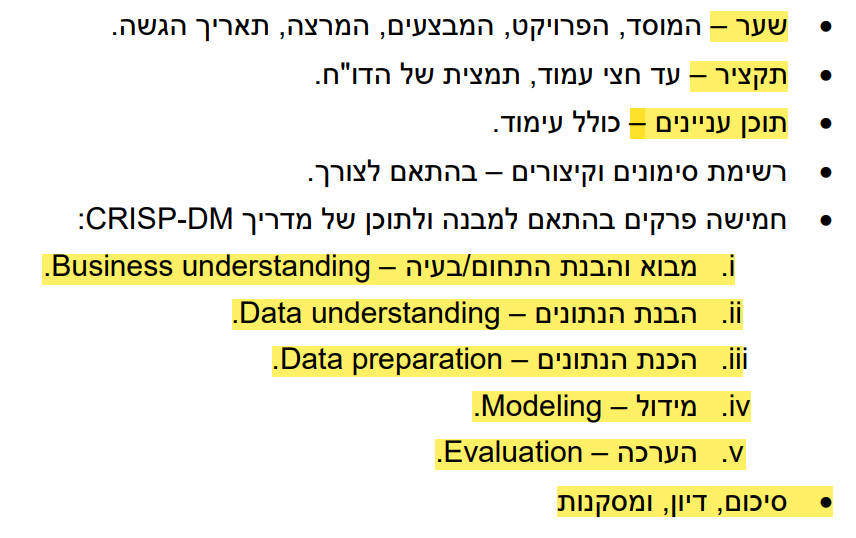
[Integrate data 5](#_Toc91702038)

[Format data 5](#_Toc91702039)

[**Modeling** 5](#_Toc91702040)

[**Evaluation** 5](#_Toc91702041)

[**Discussion and Conclusions** 5](#_Toc91702042)



# **Business Understanding**

**Industry Overview**

The hospitality industry is a 4.1 trillion dollar a year industry as of 2021(Statista.com) – Comprising of Hotels, Amusement parks, lodging, food and drinks services, travel, tourism and more, in said industry, the hotel sector alone comprises of about 35% of the global industry, averaging 1.5 trillion dollars a year.

**Business Objective**

In the hotel industry, demand forecasting and modeling is a crucial factor to revenue, hoteliers’ and accommodation websites’ objective is primarily maximizing revenue, said objective can be achieved using accurate demand forecasting and room pricing techniques accordingly, by doing that, hoteliers aim to maximize occupancy and minimize vacancy of their hotels, while accommodation websites aim to maximize fulfilled bookings (total bookings excluding booking cancellations).

**Situation Assessment**

Accurate demand forecasting is highly impacted by booking cancellations, causing demand management decisions to become difficult and risky, to mitigate booking cancellations, hoteliers and accommodation websites may employ strict cancellation policies or overbooking tactics which in turn reduces the number of bookings, and reduce revenue (N. Antonio, A. Almeida, L. Nunes, 2019), making it an ineffective solution.

Based on the data collected in this work, over 35% of bookings end up being cancelled, 50% of which (or 17.5% of all bookings) are cancelled within the 70 days (t-70) prior to the arrival date, 25% of cancelled bookings (or 8.75% of all bookings) are cancelled within the 18 days (t-18) prior to the arrival date, making it increasingly difficult to accurately predict demand in advance, and showcasing the increasing need for a dynamic, accurate cancellation forecasting model.

In recent years, as uncertainty rises due to the COVID-19 pandemic and rapid changing restrictions, booking cancellations are increasingly affecting hotels’ and accommodation websites’ demand forecasting techniques by introducing new, often very hard to predict factors to the customer’s booking behavior, large accommodation website “Trivago” recently reported booking cancellations reaching 35%.

Booking cancellation policies and penalties are important factors in customer booking decisions (C. Chen, Z. Schwartz, P. Vargas), accurate booking cancellation forecasting can assist hoteliers and accommodation websites accurately price and allocate booking cancellation policies, per customer, based on historic actions and current booking properties.

**Data Mining Goals**

We intend to collect data from Kaggle.com, a highly popular data science website, the dataset contains data of over 119,000 bookings, collected between the years 2015-2017, including 30 features, such as lead time, date of the booking, the hotel type, number of people staying, and more.

**Our Solution**

We aim to use the data collected, to train a machine learning model, based on a Random Forest Classifier which is able to predict a booking cancellation with decent accuracy.

**Sources**

Cite them

<https://journals.sagepub.com/doi/pdf/10.1177/1938965519851466>

<https://www.sciencedirect.com/science/article/pii/S0278431910000320?casa_token=mfr9Ne_Wnp8AAAAA:jqZUmDd-CcqsbrwuI8bcTaYS4jA6pzqq_RSyyP8d4ybnoAUW13uYmdFZYLBXMTQCVd68S9Ymzg>

<https://www.tmstudies.net/index.php/ectms/article/viewFile/1000/pdf_51>

<https://www.cabdirect.org/cabdirect/abstract/19871847153>

# **Data Understanding**

## Collect data

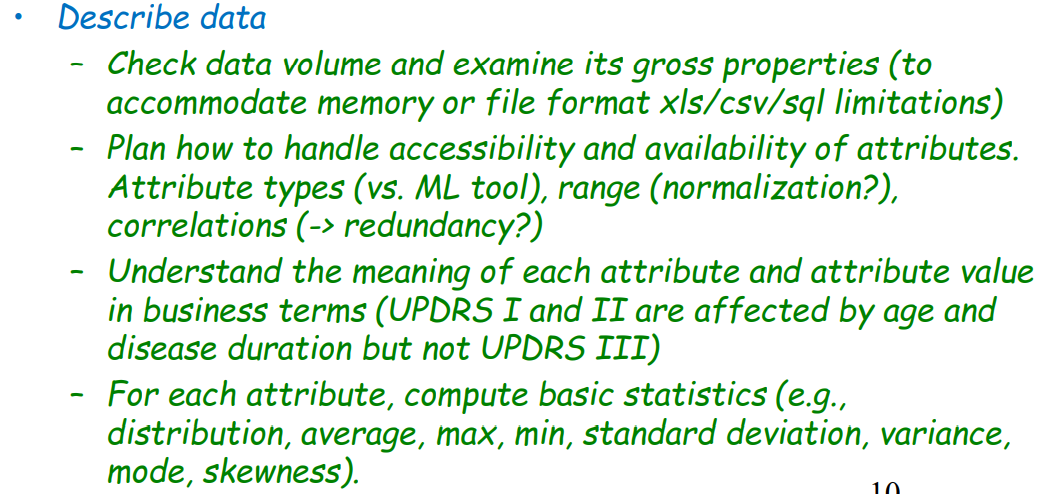
The data was collected from Kaggle.com, “Hotel Booking Demand”.  
link: <https://www.kaggle.com/jessemostipak/hotel-booking-demand>

## Describe data

The dataset contains 119,390 rows, and 33 columns, and is inside a csv file – each row represents a single booking instance, described by 30 distinct features.

**Features**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Name** | **Type** | **Cont/Cat/Unique** | **Values** | **Explain** |
| **ID** | **Int** | **Unique** | **(0,119390)** | **Key** |
| **Hotel** | **String** | **Cat** | **[Resort, City]** | **Represents the type of the hotel, Resort or City** |
| **Is\_canceled** | **Int** | **Cat** | **[0, 1]** | **Was the booking canceled** |
| **Lead\_time** | **Int** | **Cont** | **(0, 463)** | **Days before hotel arrival date** |
| **Arrival\_date\_year** | **Int** | **Cat** | **[2015,2016,2017]** | **The year of the hotel arrival date** |
| **Arrival\_date\_month** | **String** | **Cat** | **[January, …, December]** | **The month of the hotel arrival date** |
| **Arrival\_date\_week\_number** | **Int** | **Cat** | **(1,52)** | **The week of the year of the hotel arrival date** |
| **Arrival\_date\_day\_of\_month** | **Int** | **Cat** | **(0,31)** | **The day of the month of the hotel arrival date** |
| **Stays\_in\_weekend\_nights** | **Int** | **Cont** | **(0,19)** | **# of weekend nights in the booking** |
| **Stays\_in\_week\_nights** | **Int** | **Cont** | **(0,50)** | **# of weekday nights in the booking** |
| **Adults** | **Int** | **Cont** | **(0,55)** | **# of adults in the booking** |
| **Children** | **Int** | **Cont** | **(0,10)** | **# of children in the booking** |
| **Babies** | **Int** | **Cont** | **(0,10)** | **# of babies in the booking** |
| **Meal** | **String** | **Cat** | **[BB, HB, FB, SC, Undefined]** | **The accommodation meal plan** |
| **Country** | **String** | **Cat** | **177 Countries** | **The Country of the hotel** |
| **Market\_segment** | **String** | **Cat** | **[Online TA, Offline TA/TO, Groups, Direct, Corporate, Complementary, Aviation, Undefined]** | **-** |
| **Distribution\_channel** | **String** | **Cat** | **[]** | **-** |
| **Is\_repeated\_guest** | **Int** | **Cat** | **[0,1]** | **Is the guest a repeat customer** |
| **Previous\_cancellation** | **Int** | **Cont** | **()** | **# of previous cancellations for the guest** |
| **Previous\_not\_cancaled** | **Int** | **Cont** | **()** | **# of previous bookings not canceled** |
| **Reserved\_room\_type** | **String** | **Cat** | **[]** | **The type of room reserved by the customer** |
| **Assignmed\_room\_type** | **String** | **Cat** | **[]** | **The type of room actually assigned to the customer** |
| **Booking\_changes** | **Int** | **Cont** | **()** | **# of booking changes done by the customer** |
| **Deposit\_type** | **String** | **Cat** | **[]** | **The type of deposit made by the customer** |
| **Agent** | **Float** | **Cont** | **(1,535)** | **-** |
| **Company** | **String** | **Cat** | **(6,543)** | **-** |
| **Days\_in\_waiting\_list** | **Int** | **Cont** | **()** | **Days the customer has waiting in the waiting list for a room** |
| **Customer\_type** | **String** | **Cat** | **[Transient, Transient-party, Contract, Group]** | **-** |
| **Adr** | **Float** | **Cont** | **(62, 157)** | **The average daily rate** |
| **Required\_car\_parking\_spacs** | **Int** | **Cont** | **(0,8)** | **# of parking spaces the customer asked for** |
| **Total\_s\_requests** | **Int** | **Cont** | **(0,5)** | **# of special requests asked** |

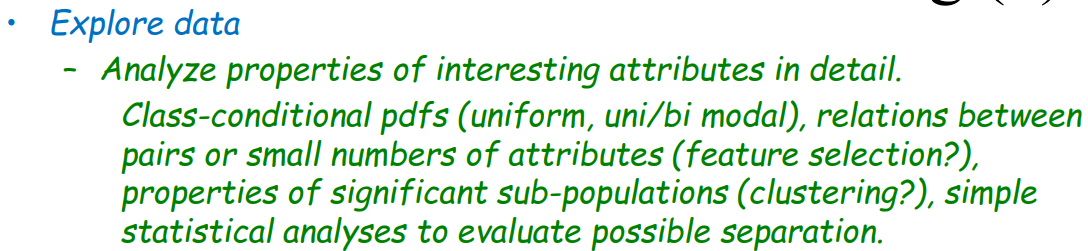


* The dataset contains 119390 samples and 32 attributes that describe them. Add screen shot.
* Atrributes: lead\_time, arrival\_date\_year, arrival\_date\_month, arrival\_date\_week\_number, arrival\_date\_day\_of\_month, stays\_in\_weekend\_nights, stays\_in\_week\_nights, adults, children, babies, meal, country, market\_segment, distribution\_channel, is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled, reserved\_room\_type, assigned\_room\_type, booking\_changes, deposit\_type, agent, company, days\_in\_waiting\_list, customer\_type, adr, required\_car\_parking\_spaces, total\_of\_special\_requests, reservation\_status, reservation\_status\_date.

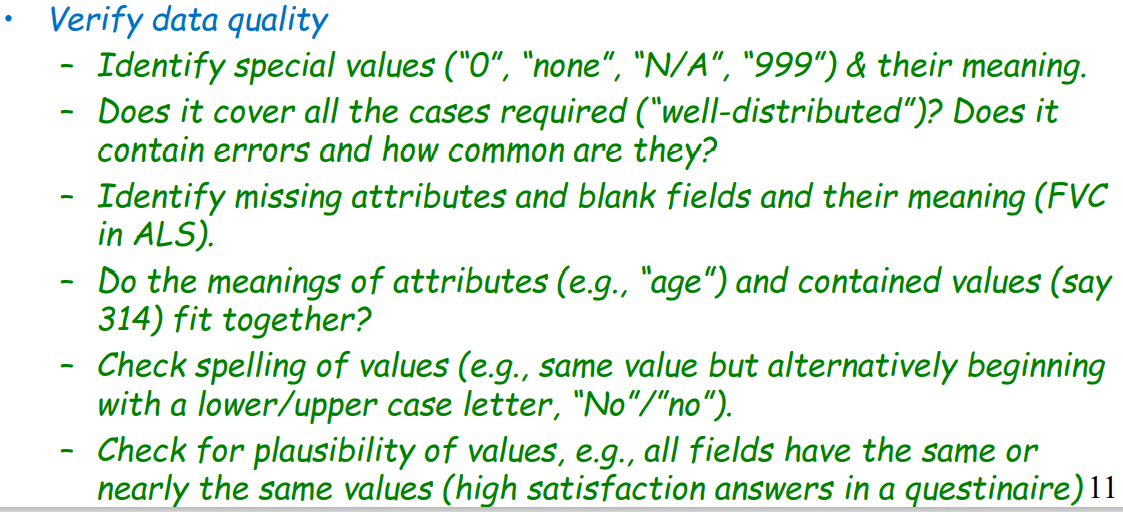
We plan to…, attributes types: Numeric and categorical

* Redundancy: Our learning task is classification of cancelled and not cancelled reservations. There are 3 attributes that identify the result of the reservation: is canceled, reservation\_status, reservation\_status\_date. We chose to train and evaluate on **is\_canceled** as a target variable and discard the other two.
* Understanding business terms: adr?
* Basic statistics of each attribute: Max, min, mean, median, std, distribution. add screen shot.

## Explore data



## Verify data quality

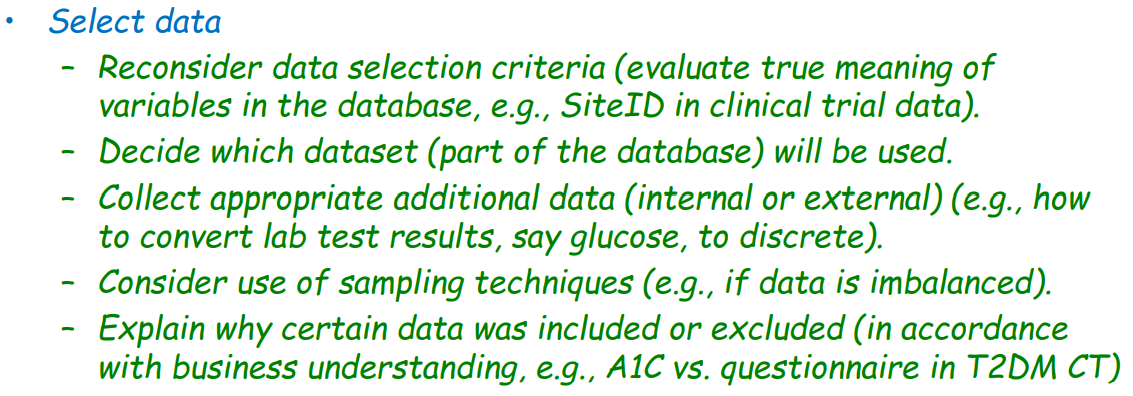


Missing values

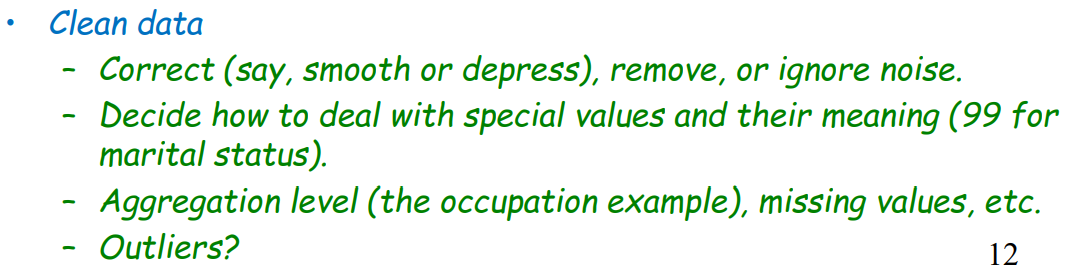
Undefined category

# **Data Preparation**

## Select data



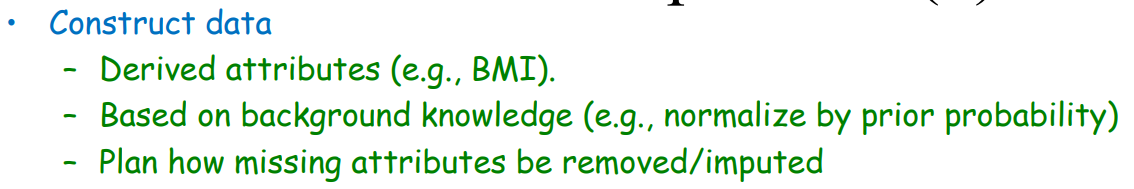
## Clean data



How did we handel missing data?

How did we handle ‘undefined’ categories?

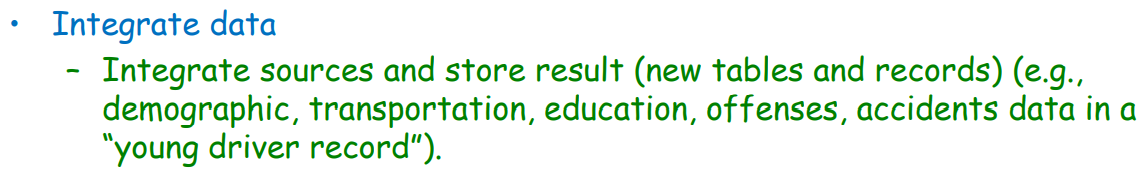
## Construct data



All added attributes

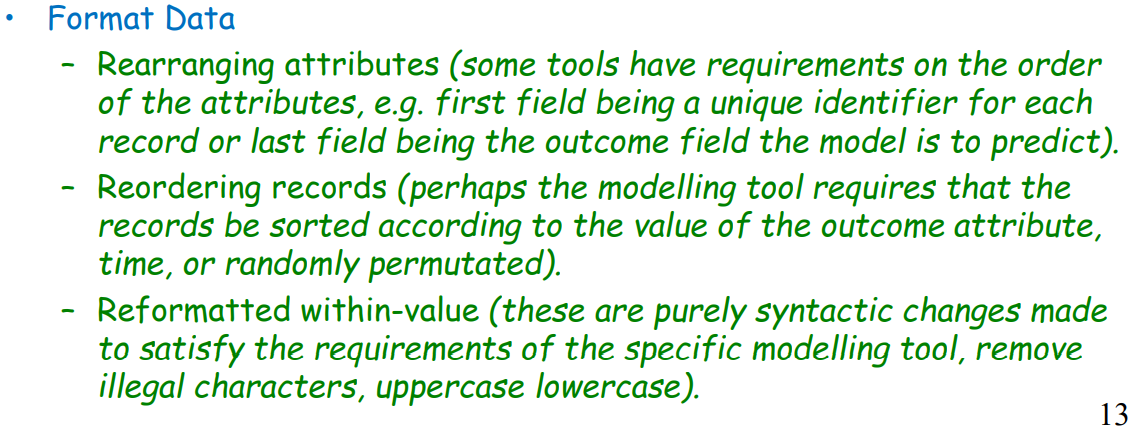
No missing attributes

## Integrate data



Not relevant

## Format data



?

Binnig?

# **Modeling**

RF

Bayes

Logistic regression

More?

# **Evaluation**

Import tree

# **Discussion and Conclusions**

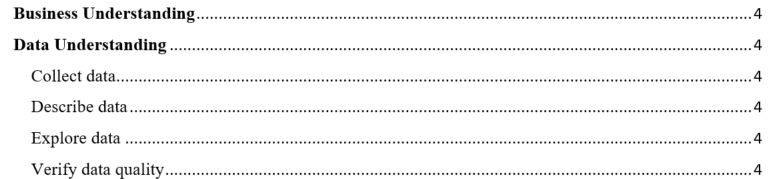
**Code:**

* Scatter plot, don’t use percentage cancellation but cancellation count – Dafna – low priority
* Add confusion matrix to the best kfold split - Dafna
* Import tree – Dafna
* Other methods: check if the data preparation is similar to logistic regression - Oriel
* Other methods: XGboost? DT? – Oriel
* Evaluation – search specific on RF – Dafna

**Document**

**Oriel**

* Presentation template



**Dafna**



**Together – discussion and conclusions**