HOTEL BOOKING CANCELLATIONS PREDICTION

by Godfried Junio Matahelemual

CONTENTS

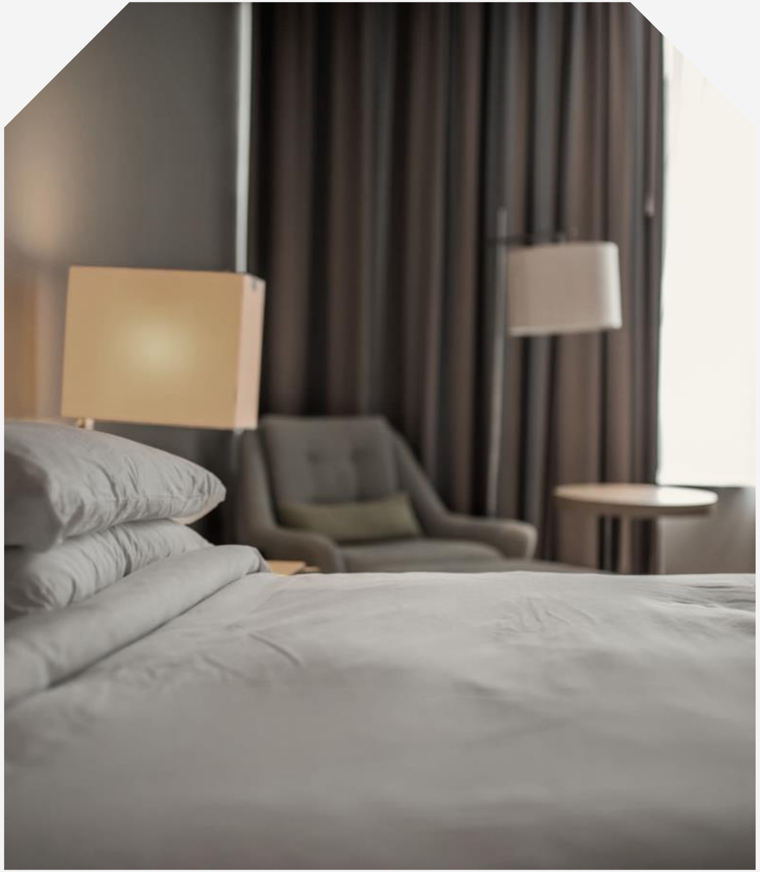
01
INTRODUCTION

ABOUT PROJECT

This project aims to predict hotel booking cancellations. By knowing the guest who is likely to cancel, the hoteliers can do actions to avoid the cancellations, calculate the right demand, or adjust their overbooking tactics and cancellation policies appropriately to increase the revenue.

This project includes the flask dashboard called **Bell-Man** (Booking cancellation prediction machine) which contain some visualizations of the data and the prediction machine itself.

BACKGROUND



Revenue management's objective (increasing revenue) is achieved through demand-management decisions, that is, by estimating demand and its characteristics while implementing price and capacity control to "manage" the demand. Thus, revenue management is concerned with the methodologies and systems required to make decisions regarding demand.

Booking cancellations are one of the topics of hotel revenue management forecasts. They affect the real demand and limit the production of accurate forecasts, a critical tool in terms of revenue management system. To mitigate the effect of cancellations, hotels implement restrictive cancellation policies and overbooking tactics, which in turn can have a negative impact on revenue and on the hotel reputation.

By identifying which bookings are likely to be cancelled, revenue managers and other members of the hotel's staff can take actions to avoid potential cancellations such as offering services, room upgrades, discounts, entrances to shows/amusement parks, or other perks. Hoteliers can reduce the number of people to be contacted and with that, contribute to lower cancellation rates, at controlled costs

02

BUSINESS FORMULATION

PROBLEM

Booking cancellations tend to occur in hospitality industry. **They highly impacted demand forecast accuracy, which later affects revenue management's objective (increasing revenue).**

Hence, it is important to identify which bookings are likely to be cancelled, so hotelier can take actions to avoid the cancellations and develop more effective overbooking and cancellation policies.

ML OBJECTIVE

Predict which bookings are likely to be cancelled to decrease uncertainty and increase revenue

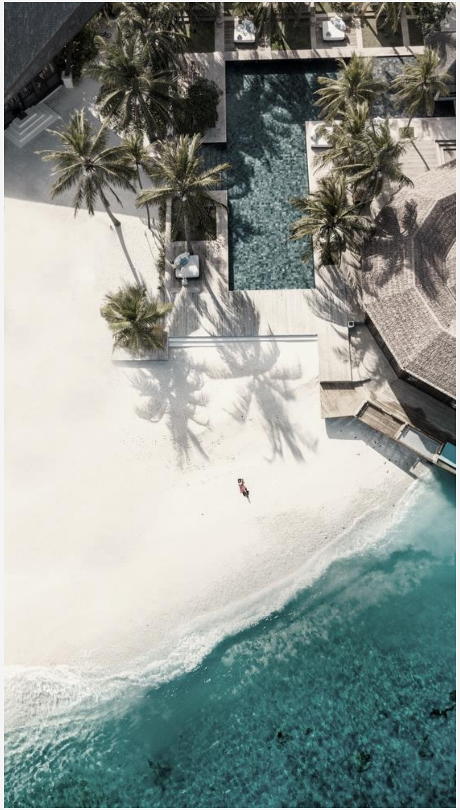
ACTION

Contact potential guests who are likely to cancel to offer services, room upgrades, discounts, entrances to shows/amusement parks, or other perks.

VALUES

Decrease uncertainty, increase revenue, improve hotel reputation

DATA



HOTEL DEMAND DATA

119390 rows x 32 columns

The data is originally from the article Hotel Booking Demand Datasets in Portugal, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020.

DATA

hotel	is_canceled	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
Resort Hotel	0	342	2015	July	27	1	0	0	2	0.0	0	BB	PRT	Direct	Direct
Resort Hotel	0	737	2015	July	27	1	0	0	2	0.0	0	BB	PRT	Direct	Direct
Resort Hotel	0	7	2015	July	27	1	0	1	1	0.0	0	BB	GBR	Direct	Direct
Resort Hotel	0	13	2015	July	27	1	0	1	1	0.0	0	BB	GBR	Corporate	Corporate
Resort Hotel	0	14	2015	July	27	1	0	2	2	0.0	0	BB	GBR	Online TA	TA/TO

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
0	0	0	C	C	3	No Deposit	NaN	NaN	0	Transient	0.00	0	0	Check-Out	2015-07-01
0	0	0	C	C	4	No Deposit	NaN	NaN	0	Transient	0.00	0	0	Check-Out	2015-07-01
0	0	0	A	C	0	No Deposit	NaN	NaN	0	Transient	75.00	0	0	Check-Out	2015-07-02
0	0	0	A	A	0	No Deposit	304.0	NaN	0	Transient	75.00	0	0	Check-Out	2015-07-02
0	0	0	A	A	0	No Deposit	240.0	NaN	0	Transient	98.00	0	1	Check-Out	2015-07-03

03

PROCESSING

DATA PREPARATION

01 DROP ROWS WITH MISSING VALUES

children, country, market_segment, distribution_channel

02 MODIFYING COLUMN

total_guests = adults + children + babies

03 REMOVE WRONG VALUES

- adr feature has negative value. It does not make sense if the price is negative.
- total_guests has the value of 0. It does not make sense if no one booked the room.

04 DROP FEATURES

Considering:
percentage of null values, information contained, high correlation with other features

MODELING

01

PREPROCESSING

Pipeline:
constant imputer and binary encoder for categorical data, robust scaler for numerical data, smote

02

MODELING WITH DEFAULT PARAMETERS

Six models are used: Logistic Regression, DTC, RFC, Gradient Boosting Classifier, XGB Classifier, KNN

Test Score Results:

The two best models are RFC and XGB

	trainScore	testScore	
LogisticRegression	0.681345	0.681037	
DecisionTreeClassifier	0.995342	0.776590	
RandomForestClassifier	0.994080	0.864348	✓
GradientBoostingClassifier	0.753687	0.748542	
XGBClassifier	0.846756	0.823024	✓
KNeighborsClassifier	0.796263	0.705754	

MODELING

METRICS USED
PRECISION

1. FN is a condition when the MODEL predict THE GUESTS ARE COMING (0) but ACTUALLY THE BOOKINGS ARE cancelled (1)

    - High numbers of FN will cause unoptimal revenue due to many rooms are reserved (and can not be booked by other potential guests) but the booking is cancelled

2. FP is a condition when the MODEL predict THE BOOKINGS ARE cancelled (1) but ACTUALLY THE GUESTS ARE COMING (0). High numbers of FP will cause overbooking.

    - overbooking can force the hotel to deny service to a customer; this can be a very bad experience for the guests and may result in online complaints and generation of a negative impact in terms of social reputation.

    - the hotel also lose future revenue from this overbooking, since the guests who experienced bad service have strong tendency to avoid that hotel.

    - overbooking also cause the loss that occurs as a result of the hotel's obligation to compensate the customer, including reallocation costs.

I will focus more to **precision** score because I want the FP as less as possible

MODELING

MODELING WITH DEFAULT PARAMETERS + RFE IN PIPELINE

03

The test score using RFE is slightly drop from the non-RFE model and RFC model still has the best score

	Without RFE	WITH RFE
RandomForestClassifier	0.864348	0.851294
XGBClassifier	0.823024	0.816120

HYPERPARAMETER TUNING ON RFC

04

The tuning is failed to make the score better. So, the default parameter RFC model pipeline will be used

```
print(f'Score Before Tuning: {testScore[2]}')
print(f'Score After Tuning: {precision_score(y_test, y_pred_rfc_tuned)}')
```

```
Score Before Tuning: 0.8643481168795232
Score After Tuning: 0.8519078641228478
```

MODELING

THRESHOLD ADJUSTMENT

05

The threshold of 0.590833 increase precision score as much as almost 4% while losing recall score 5.7%

```
# Comparison between best threshold with default threshold (0.5)
pr_df.iloc[[828, 916]]
```

	precision	recall	threshold
828	0.860297	0.834277	0.500000
916	0.900092	0.776620	0.590833

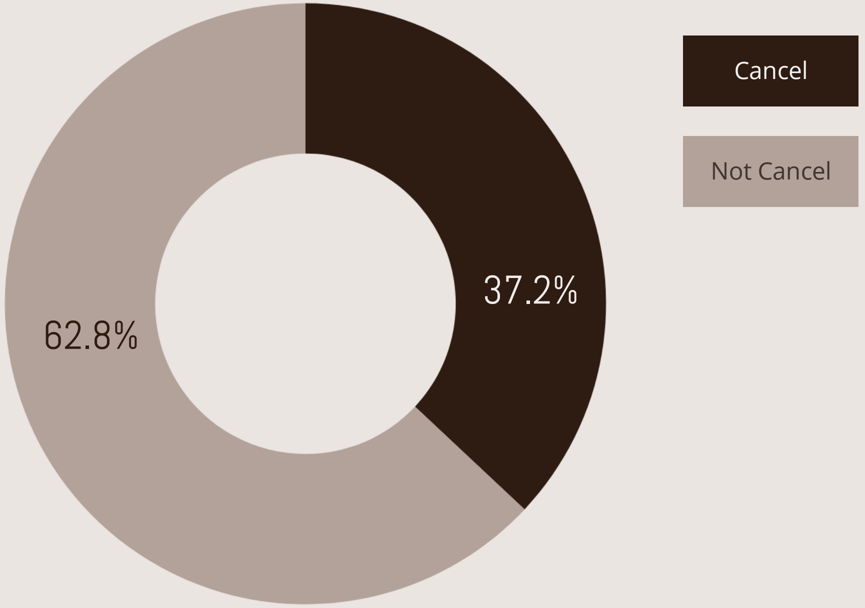
```
print(f'Precision score improvement: {(0.900092-0.860297)*100}')
print(f'Recall score reduction: {(0.834277-0.776620)*100}')
```

```
Precision score improvement: 3.9795000000000025
Recall score reduction: 5.765700000000007
```

04

BUSINESS INSIGHTS

Cancellation Rate

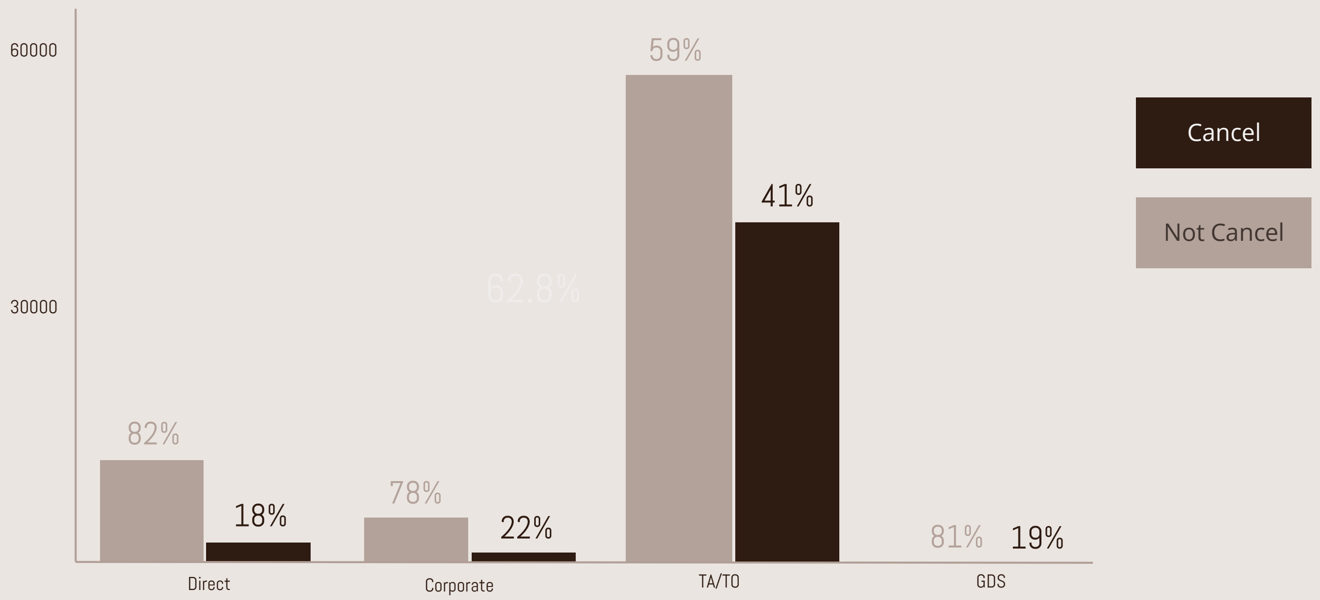


Cancellation Rate:

37.2%

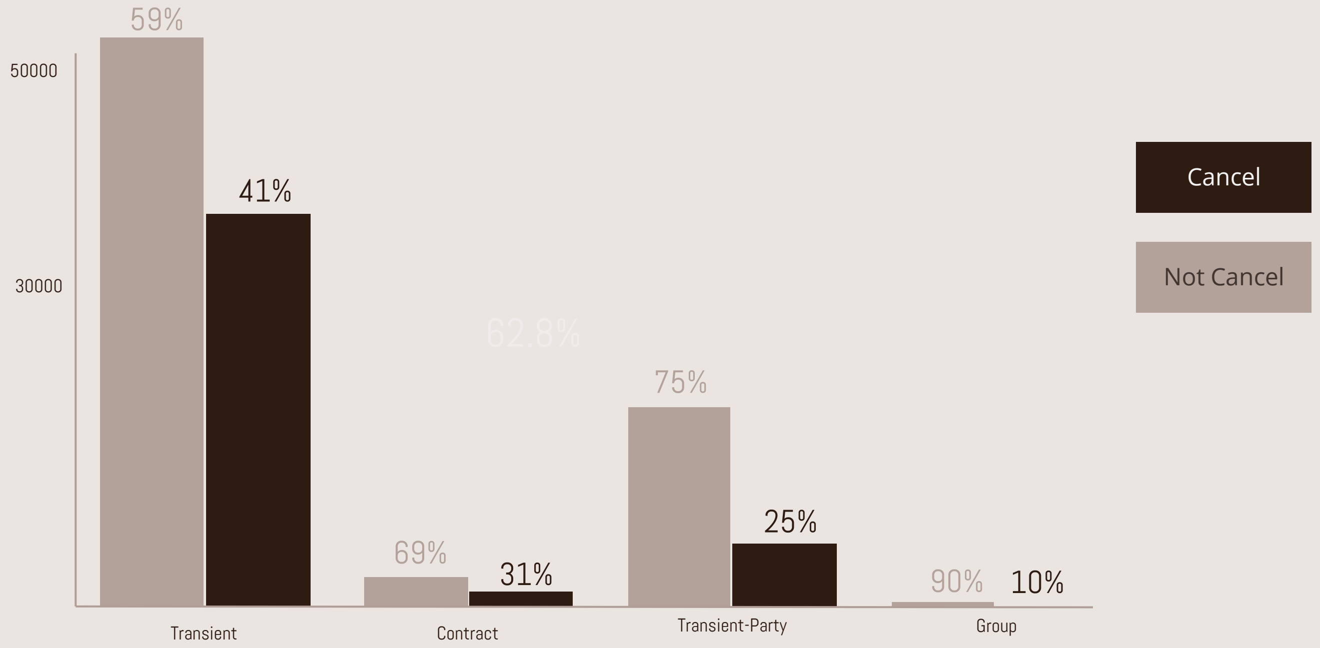
cancelled class shares around 37% proportion of the data. In general, it can be said that there is almost 2 cancellation in 5 bookings.

Cancellation by Booking Channel



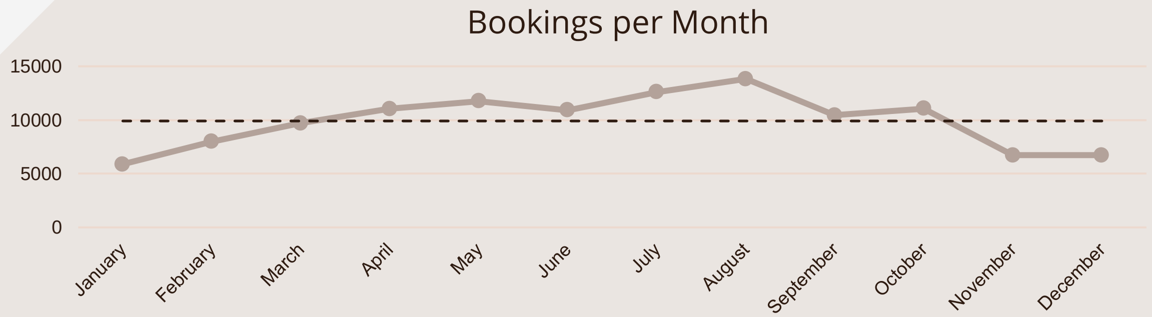
The cancelled booking mostly occurred in Travel Agent/Tour Operator booking channel. This booking channel also has the highest cancellation rate (41% of the bookings are cancelled) compared with the other booking channels. Hotel can formulate a new agreement with TA/TO to minimize cancellation

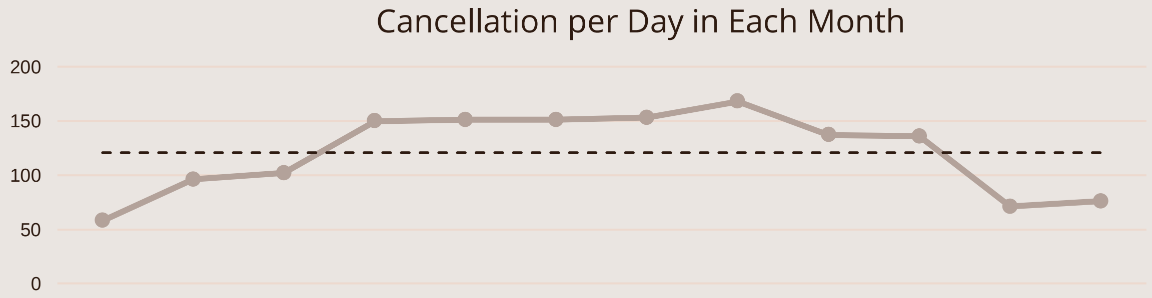
Cancellation by Customer Type



Transient is the customer type with the most booking and the most cancelation rate (41%). Hotel must concern more with this type of customer.

Monthly Analysis

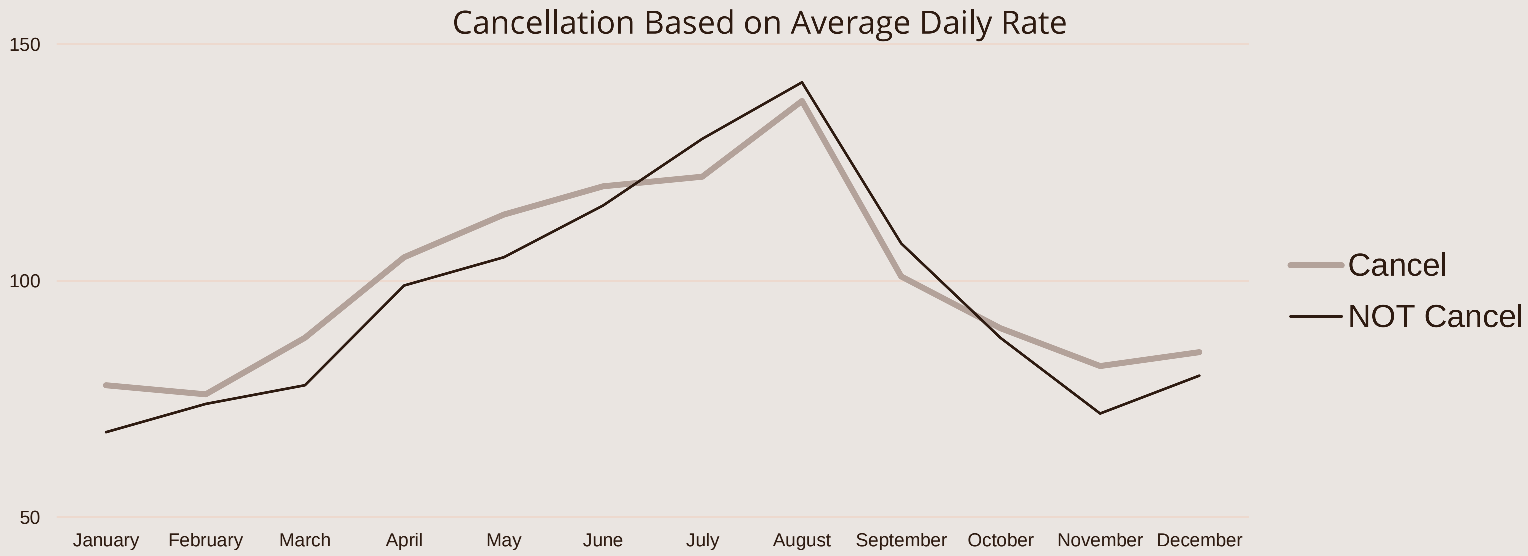




month	cancel	not_cancel	total_booking	total_day	avg_cancel
January	1805	4061	5866	31	58.225806
February	2693	5304	7997	28	96.178571
March	3147	6565	9712	31	101.516129
April	4510	6528	11038	30	150.333333
May	4677	7091	11768	31	150.870968
June	4533	6384	10917	30	151.100000
July	4731	7879	12610	31	152.612903
August	5232	8604	13836	31	168.774194
September	4099	6360	10459	30	136.633333
October	4228	6854	11082	31	136.387097
November	2120	4611	6731	30	70.666667
December	2363	4347	6710	31	76.225806

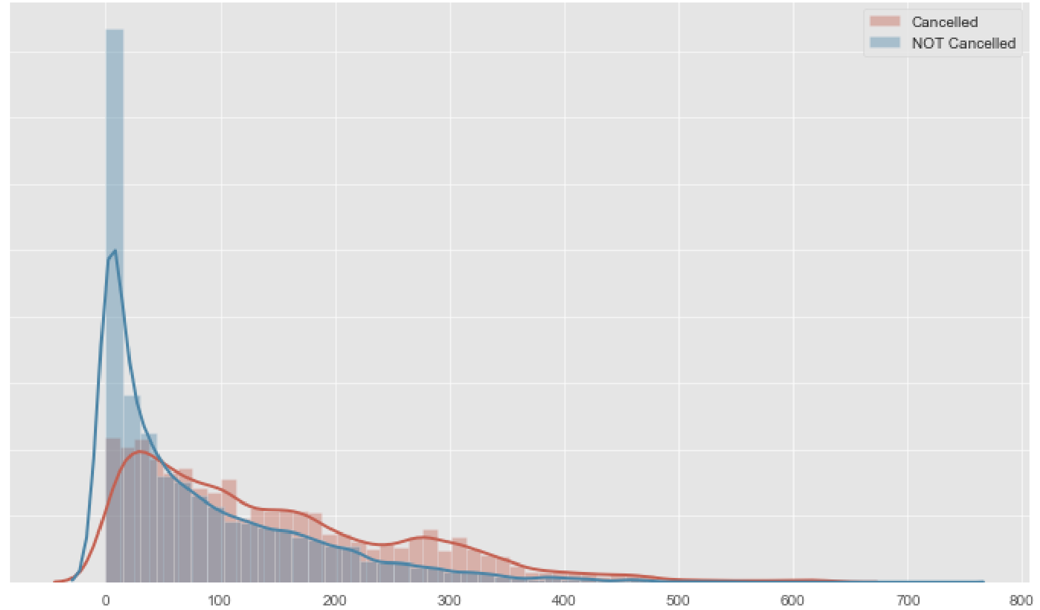
Months of April to October have cancellation per day above monthly average. August is the peak. Since August is the month with the most bookings/demands, it is worth to try setting restrictive booking cancellation policies in August or even April to October.

Monthly Analysis



Average price per night of the reservation (ADR) tends to peak during the summer, which is possibly due to high demand during the summer holiday. In 9 of 12 months (apart of July-September), the cancellation occurs in the higher average price. It can be said that price probably has impact of the cancellation decision.

Booking Lead Time

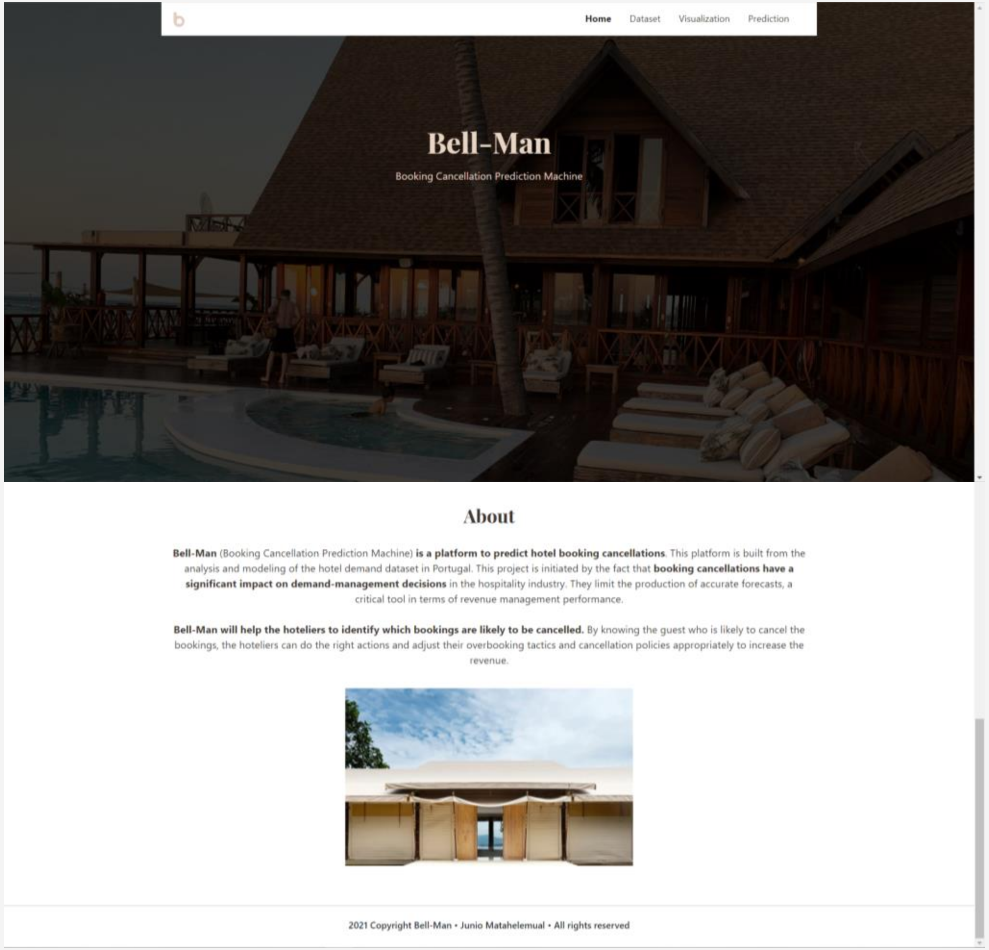


The longer the lead time, the higher the chance the booking is cancelled. It is probably because there is plan change or other uncertainties since there are many things happen during that long lead time.

05

FLASK DASHBOARD

HOMEPAGE


Home
Dataset
Visualization
Prediction
Bell-Man
Booking Cancellation Prediction Machine
About
Bell-Man (Booking Cancellation Prediction Machine) is a platform to predict hotel booking cancellations. This platform is built from the analysis and modeling of the hotel demand dataset in Portugal. This project is initiated by the fact that booking cancellations have a significant impact on demand-management decisions in the hospitality industry. They limit the production of accurate forecasts, a critical tool in terms of revenue management performance.
Bell-Man will help the hoteliers to identify which bookings are likely to be cancelled. By knowing the guest who is likely to cancel the bookings, the hoteliers can do the right actions and adjust their overbooking tactics and cancellation policies appropriately to increase the revenue.
2021 Copyright Bell-Man • Junio Matahelemual • All rights reserved

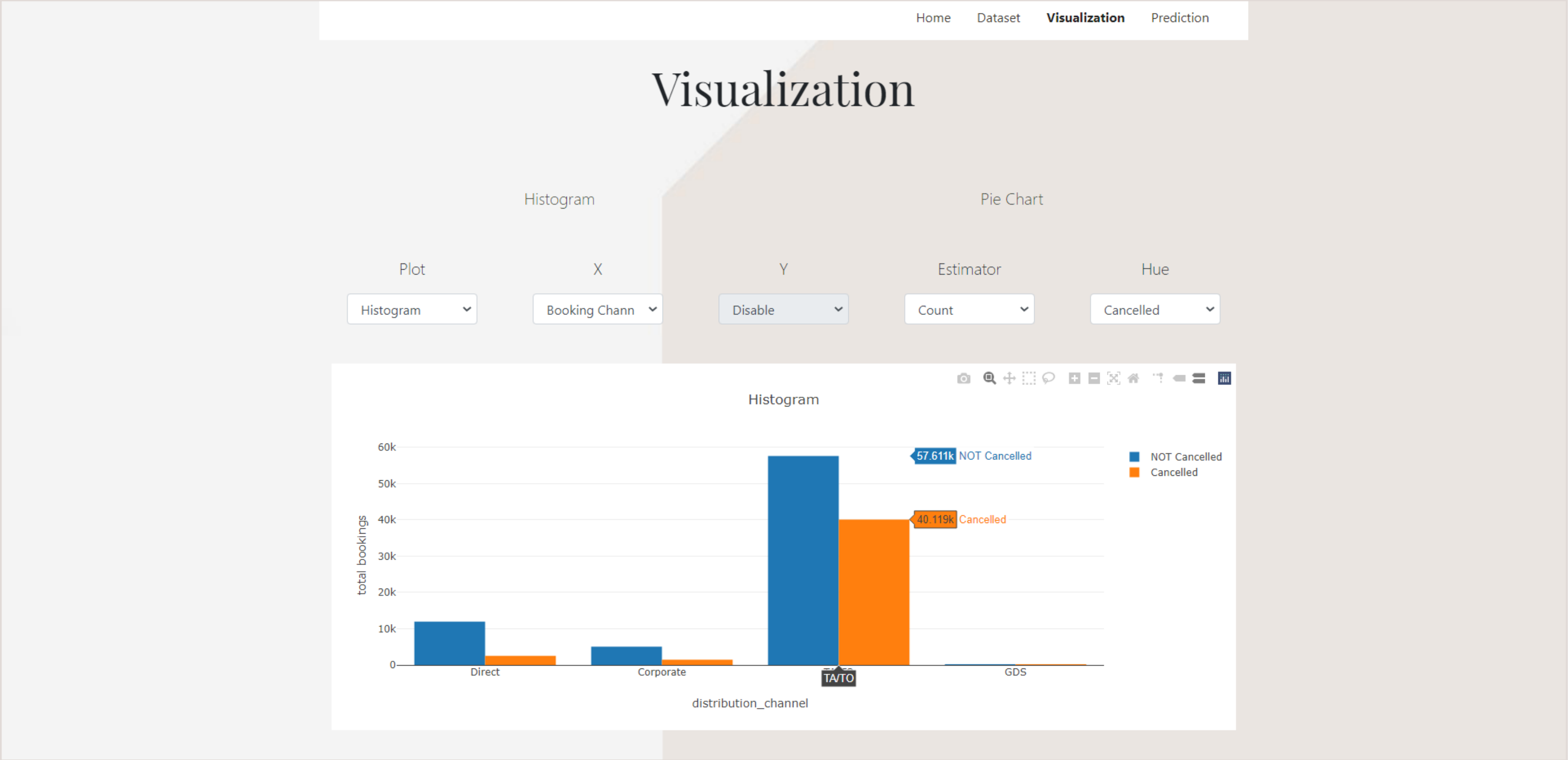
DATASET PAGE

Home **Dataset** Visualization Prediction

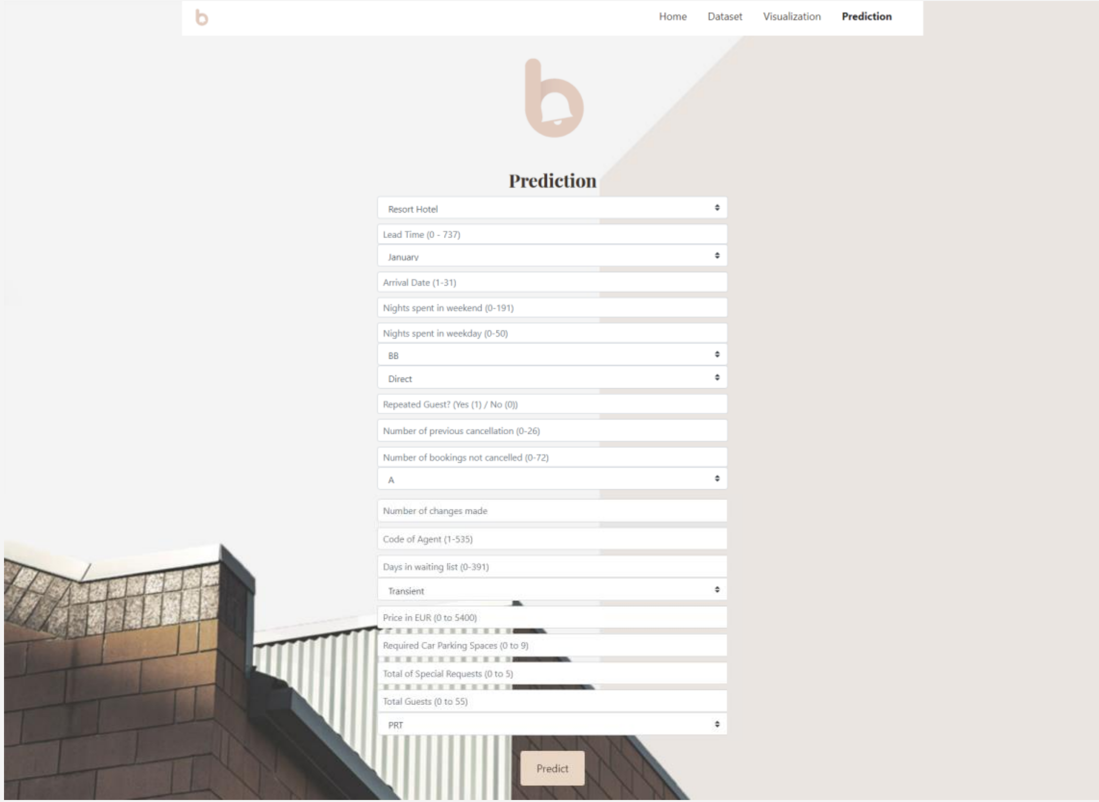
Hotel Demand Dataset

Unnamed: 0	hotel	is_canceled	lead_time	arrival_date_month	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	meal	distribution_channel	is_repeated_guest	previous_ca
0	Resort Hotel	0	342	July	1	0	0	BB	Direct	0	
1	Resort Hotel	0	737	July	1	0	0	BB	Direct	0	
2	Resort Hotel	0	7	July	1	0	1	BB	Direct	0	
3	Resort Hotel	0	13	July	1	0	1	BB	Corporate	0	
4	Resort Hotel	0	14	July	1	0	2	BB	TA/TO	0	
5	Resort Hotel	0	14	July	1	0	2	BB	TA/TO	0	
6	Resort Hotel	0	0	July	1	0	2	BB	Direct	0	
7	Resort Hotel	0	9	July	1	0	2	FB	Direct	0	
8	Resort Hotel	1	85	July	1	0	3	BB	TA/TO	0	

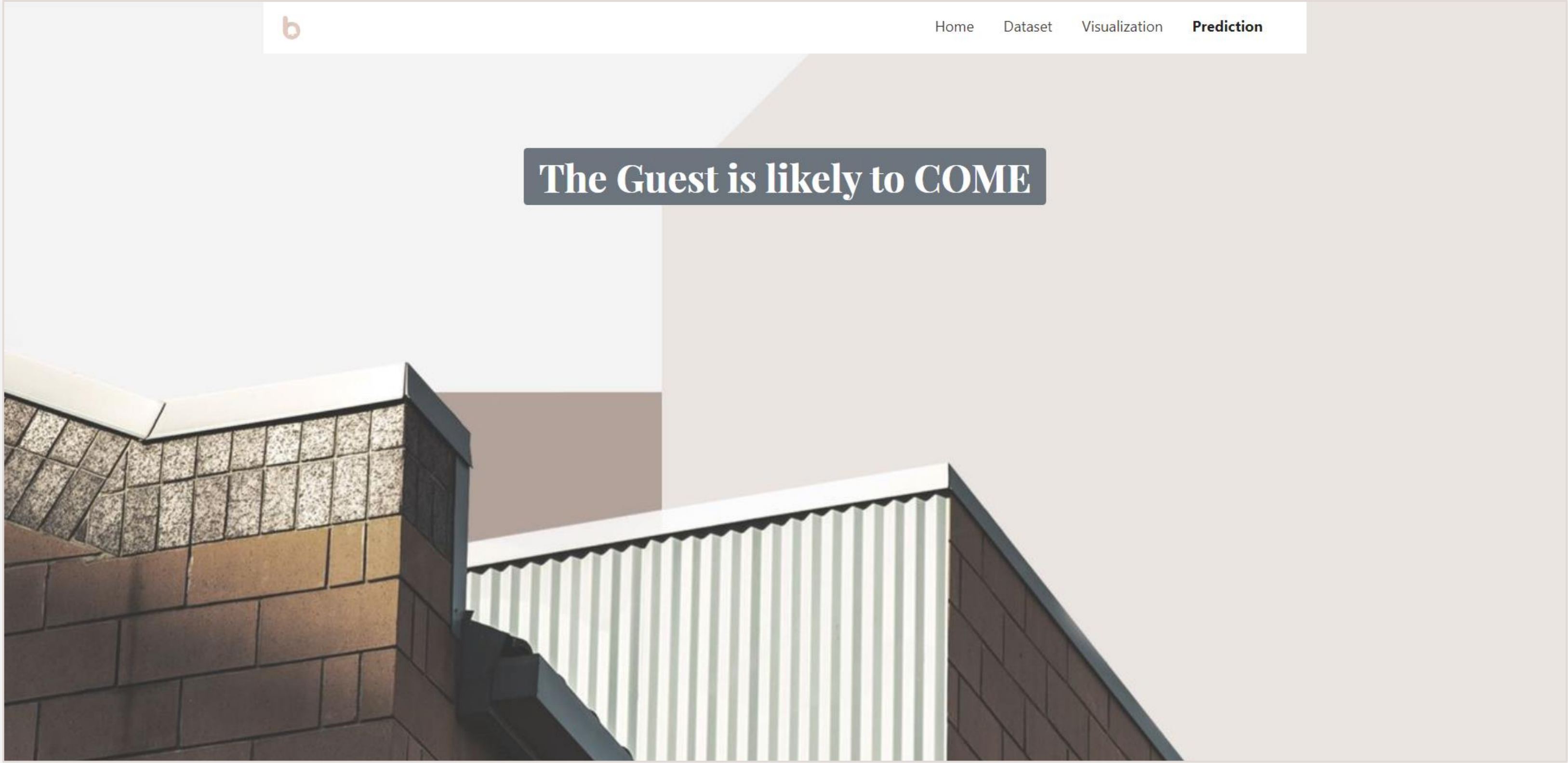
VISUALIZATION PAGE



PREDICTION PAGE



RESULT PAGE


Home
Dataset
Visualization
Prediction
The Guest is likely to COME

05 BUSINESS IMPLICATIONS

NUMBER OF PEOPLE CONTACTED

According to Antonio, de Almeida, and Nunes (2017), guests contacted by hotels, even without being offered nothing substantial, cancel much less than guests not contacted. So, the knowledge of cancellation can help hoteliers **reduce the number of people to be contacted** and with that, contribute to lower cancellation rates, at controlled costs.

Supposed there are 100 resort hotel bookings for certain days with no deposit for the same type of room (to make it easier to count the price). To increase the certainty of the bookings, hoteliers do not have to contact all the 100 booking makers, but only 37 of them (assuming the cancellation rate is still the same for the future: 37%).

BOOKINGS	CONTACTED
100	37

WHO ARE 37 OF THEM?

THE MODEL WILL HELP

DECREASING NUMBER OF PEOPLE TO CONTACT UP TO

+63%

INCREASING REVENUE

From the 37 booking makers contacted, assumed 1/3 of them decide not to cancel. From the 0 revenues expected (because all bookings are no deposit), hoteliers managed to increase revenue to 33%.

INCREASE REVENUE

1/3 of total bookings potentially cancel x Average price for resort hotel
= 12 x 76 EUR

912 €

INCREASING REVENUE UP TO +33%

CALCULATE ACCURATE DEMAND

By running the models daily against all reservations on-the-books, we can get **the number of room cancelled for each of the following days**. With this amount, hotels can deduce its value from their demand by calculating their net demand. **Equipped with an accurate demand value, hotel managers can develop more effective overbooking and cancellation policies**.

06

IMPROVEMENT RECOMMENDATIONS

IMPROVEMENT RECOMMENDATIONS

OTHER FEATURE ENGINEERING TECHNIQUES

Other techniques of feature engineering are worth to try to improve the model accuracy such as one hot encoding, binning, other techniques of feature selection, or other techniques that not included in this project

ADD OTHER INFLUENTIAL DATA

Other data such as events, weather, social reputations will potentially make it better for the hoteliers to understand the cancellation pattern.
(Anderson, 2012; Chan & Wong, 2006; McGuire, 2016; C.-M. Chen & Lin, 2014; Day, Chin, Sydnor, & Cherkauer, 2013)

HELP HOTELS TO FORMULATE THE EXACT STRATEGY

This project not really recommend the price to certain market segment for certain days, cancellation policy, overbooking strategy, etc. So, it will be nice if future work could cover them. Remember the core of revenue management is to sell the right product to Right customer for the right place at the right time through the right channels

THANK YOU