BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Predicting Hotel booking Cancellations

Ana Marta Silva: M20200971
Natalia Cristina Castañeda: M20200575
María Luisa Noguera: M20201005
Gustavo Tourinho: M20180846

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

1. BUSINESS UNDERSTANDING

1.1 Introduction and Background

One of the main concerns in the hotel industry, regarding how bookings are made as of now, is the number of cancellations (especially if they are last minute) since hotels profitability increases with occupation. Balancing overbooking to keep the highest occupation possible has become an expensive challenge for H2 Hotel, since none of the two main strategies they have applied have reduced cancellations nor kept good levels of occupation.
Therefore, H2 has requested us to construct a model for being able to predict cancellations and better manage overbooking.

One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2.

1.2 Business Objectives

The business objectives were defined by the Revenue Manager Director as follows:

a. Accurately calculate the Net Demand based on current bookings.
b. Understand the main characteristics of the bookings that result in cancellations.
c. Implement contingency strategies to reduce cancellation risks over bookings considered as potential cancellations.

1.3 Business Success criteria

Reduce cancellations from a 41.7% to a 20%.

1.4 Situation Assessment

a. General comments and Resources

The dataset provided contains information for H2 which had the highest cancellation rate (41,7%). This set contains 79,330 records and 31 variables, out of which 16 are numerical (15 Integers, 1 numeric), 14 are categorical and one in date format.
Because there is no key in the dataset, 15.874 records seem like duplicates, however they are bookings that happen to have the same characteristics.

b. Terminology

VARIABLE	MEANING
ADR	Average Daily Rate
Adults	Number of adults
Agent	ID of the travel agency that made the booking
ArrivalDateDayOfMonth	Day of the month of the arrival date
ArrivalDateMonth	Month of arrival date with 12 categories: "January" to "December"
ArrivalDateWeekNumber	Week number of the arrival date
ArrivalDateYear	Year of the arrival date
AssignedRoomType	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
Babies	Number of babies
BookingChanges	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
Children	Number of children
Company	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons

Country	Country of origin. Categories are represented in the ISO 3155-3:2013 format
CustomerType	Type of booking, assuming one of four possible categories (presented below)
DaysInWaitingList	Number of days the booking was in the waiting list before it was confirmed to the customer
DepositType	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories (presented)
DistributionChannel	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
IsCanceled	Value indicating if the booking was canceled (1) or not (0)
IsRepeatedGuest	Value indicating if the booking name was from a repeated guest (1) or not (0)
LeadTime	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
MarketSegment	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
Meal	Type of meal booked. Categories are presented in standard hospitality meal packages (presented below)
PreviousBookingsNotCanceled	Number of previous bookings not cancelled by the customer prior to the current booking
PreviousCancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
RequiredCarParkingSpaces	Number of car parking spaces required by the customer
ReservationStatus	Reservation last status, assuming one of three categories (presented below)
ReservationStatusDate	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel
ReservedRoomType	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Number of special requests made by the customer (e.g. twin bed or high floor)

c. Costs and Benefit

Once the predicted potential cancellations are identified, different offers will be made to this segment to reduce the risk of cancellation. Moreover, the model will also allow to calculate a more precise Net Demand, that will allow at the same time the adjustment (potential reduction) of the overbookings. Consequently, the relocation costs and loss of future revenue will be reduced. The improved booking management will also help controlling the amount of rooms offered under "non-refundable" (or restrictive policies).

d. Risks and Contingencies

Risk	**Contingency**
Cancellation policy details are not clear	Assume all cancellations are made within the allowed cancellation period
15.874 records look like duplicates	Include all records (including those that seem duplicates) in the training dataset.

1.5 Machine Learning Goals

a. Build a predictive model to identify potential cancellations

1.6 Machine Learning Success Criteria

1.6.1 Accuracy of the model which measures the capacity of the model to predict correctly the cancellations and no cancelations considering the total of bookings.

2. DATA UNDERSTANDING

When taking a closer look at each variable, we found the following:

IsCanceled: 58% of the reservations do not get cancelled.
LeadTime: Mean is 109.35, Median 74, Max is 629 – So there is indication of possible outliers. Min is 0 (Consider walk-ins and same day bookings)
ArrivalDateYear: All three main central tendency measures result in 2016. (Consider concatenating with month and day to analyze seasonality- or use just the month).
Most people arrive on the second week of June and on the 15th (mean)
Staysinweekendnights: Median is 1 night, at least 50% of the bookings include 1 weekend night.

Staysinweeknights: average stay of 2 weekday nights.
Adults: Median is 2. However, there are some cases that show no adults and only children in the reservation. We also found reservations with no adults nor children (possible outliers).
Children: At least 75% of the customers don't come in with children, and the max is 3.
DepositType: Most of the customers **do not make** deposits before check-in date. Over 66,000 have this deposit type or 83.7%
Agent: 40.2% of bookings area made through travel agency **ID9**. It seems a very powerful/convenient partnership.
Company: Only a 4,6% of the bookings are made by companies.
Daysinwaitinglist: Number of days in waitlist. Max 391, min 0, Mean is 3. There is clear outliers. 3443 of the bookings are waitlisted only 342 were cancelled.
ADR: average daily rate is 105Eur/night. Min: 79, median 99, Max 5400 (consider a presidential suite or a special occasion).
Requirecarparkingpsaces: average is 0, max is 3.
TotalofSpecialRequests: 50% make no requests. Max is 5 requests. Median also 0.
ReservationStatus: Cancelled, Checkout, no show. Checkout is the mean. All No shows have no payment included.
ReservatioStatusDate: Date when the last status was set (Cancelled, checkout, no show).
Meal: The majority only reserve Breakfast.
Country: Not using for the analysis as it is an "unreliable" variable.
DistChannel: TA/TO is the majority.
RepeatedGuest: Majority of our guests are new.

When checking for missing values/Nan, the variables Country and Children showed to have them. Likewise, we evaluated the correlation between all the variables using the Phik method. The conclusions were the following:

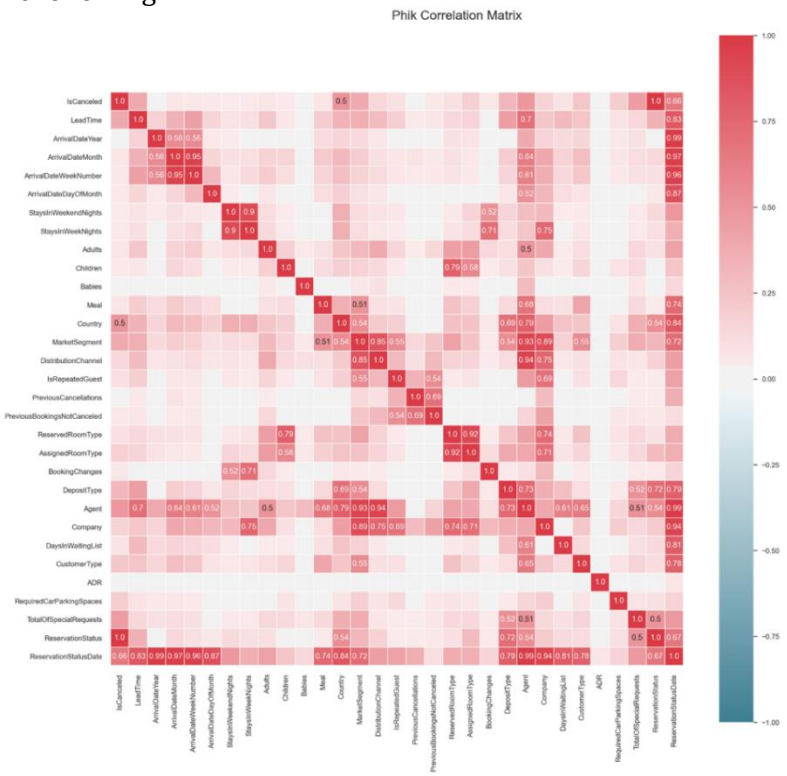

Figure 1: Phik correlation matrix

- ReservationStatus is highly correlated with the target variable isCanceled
- Arrival dates are highly correlated with ReservationStatusDate
- Weeknights and Weekendnight highly correlated which is why we decided that we would create a new variable from this two that would inform on the duration of each staying and another binary variable that would indicate if the booking includes weekend nights or not. Consequently, Weeknights and Weekendnights would be eliminated.
- Reseveroomtype and assignedroomtype higly correlated. We removed the latter since we want to show the room type interest of the client at the time of the booking.
- We should remove Daysinwaitinglist because based on the data it is unlikely that this happens. Only 4.3% of all bookings experienced having to be in a waiting list.
- Eliminate DistributionChannel and MarketSegment because they are highly correlated with the variable Agent. (This needs a number to support the argument)

Moreover, boxplots were built to take a closer look at the outliers of each numerical variables.

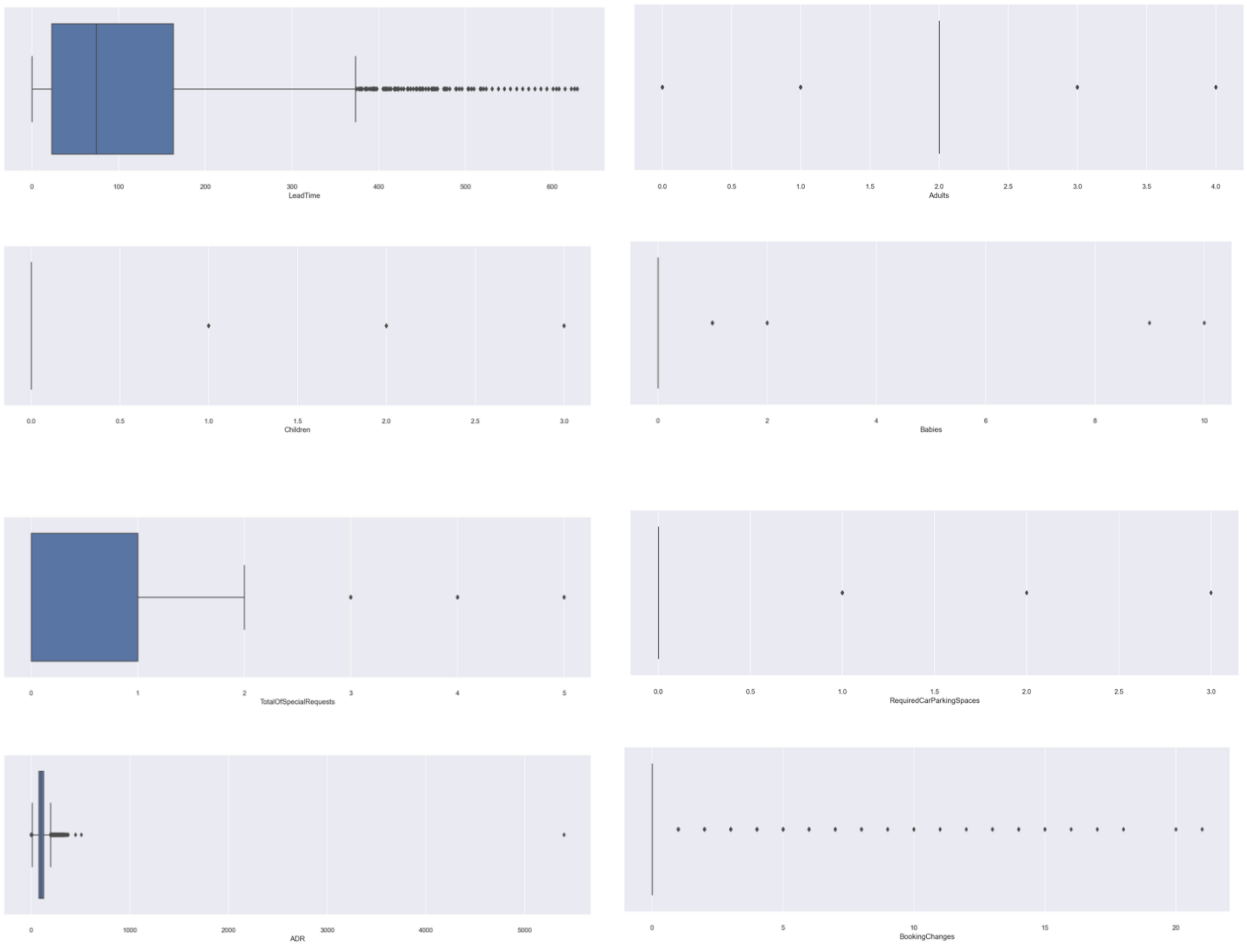

Figure 2: Boxplot for each metric feature

Outliers for the following variables were identified:
LeadTime, Adults, Children, Babies, TotalSpecialRequests, Requiredcarparkingspaces, ADR and BookingChanges, to which we will apply the adequate techniques to treat them.

3. DATA PREPARATION

3.1 Data Quality

Initially we applied the interquartile method but it did not manage to remove any data. Thus, we decided to apply the manual removal to the aforementioned variables.

3.2 Feature Engineering

a. Applied one-hot encoding to the non-metric variables: Meal, Agent, Company, ArrivalDateMonth, ReservedRoomType, DepositType, CustomerType, IsRepeatedGuest, Weekendnight, NoShow.
b. Created the variable StayingDuration which corresponds to the sum of stays in weekend nights and stays in weeknights.
c. Converted weekendnight to binary variable (anything over 0 was converted to 1, others remained as 0).
d. Agent was converted to a binary variable that defines if the booking was made directly with the guest or was made through and agency (all nulls became 0 and others were converted to 1).

3.3 Feature Exclusion

a. Country was excluded since it is an unreliable variable.
b. ArrivaldateYear, ArrivalDateWeekNumber, DayofMonth were dropped.
c. DistributionChannel, MarketSegment, AssignedRoomType, DaysinWaitingList, ReservationStatusDate, Staysinweeknight, Staysinweekendnigt, ReservationStatus were dropped.

3.4 Feature Correlation (Phik)

After feature engineering, we evaluated the correlation between the variables, but none showed to be highly correlated.

4. MODELING

4.1 Modeling Techniques, Assumptions, and Modeling.

Following these steps, we divided our dataset in train (60%), validation(20%) and test(20%) and applied the robust scaler normalization technique to each part of the dataset to avoid data leakage. Then we tried different predictive models in order to be able to obtain the best one. The different models applied are described below.

a. Neural Network

Also known as the Artificial Neural Nnetwork (ANN) algorithm, has a structure built like the human brain. ANN is composed of input, hidden, and output layers of neuron-like structures. Each node or neuron connects to another and associates a weight and threshold, that allows it to learn by itself and improve the predictions.

In this model, we start with the default parameters, we used the MLPClassifier from the sklearn library to perform that task, and we got the following results for the test set:

```
Recall: 0.7466
Precision: 0.8279
Accuracy: 0.8298
ROC_AUC: 0.8179
F1 score: 0.7852
```

Figure 3: Performance statistics of the Neural Network model.

Afterwards, tuned the model using three hidden layers with 100 nodes, with an adaptive learning rate and increased the number of interactions to 500. However, the tunning of the model showed no improvement.

```
Recall: 0.7651
Precision: 0.8126
Accuracy: 0.8287
ROC_AUC: 0.8196
F1 score: 0.7882
```

Figure 4: Performance statistics of the Neural Network model tuned.

With the ANN model, we got a good result when predicting the observations with and without cancelation.

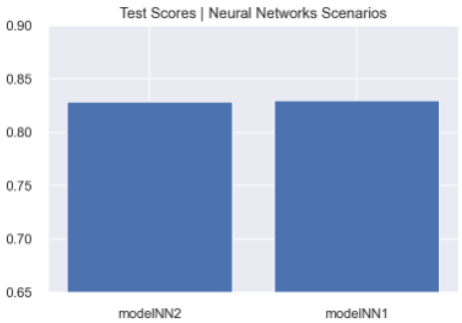

Figure 5: Neural Network Scenarios, tuned (modelNN2) and default (modelNN1)

b. Random Forest

Random Forest is a supervised learning algorithm that builds multiple decision trees and merges them to get a more accurate and stable prediction.

For this model, we start applying the default parameters, we used the RF classifier from the sklearn library to perform that task, and we got the following results for the test set:

```
Recall: 0.7832
Precision: 0.8629
Accuracy: 0.8578
ROC_AUC: 0.8472
F1 score: 0.8211
```

Figure 6: Performance statistics of the Random Forest model.

Afterward, use the GridSearchCV function for tunning the model. This function executes an exhaustive search over specified parameter values for an estimator to improve the model's performance through cross-validation techniques. The selected parameters chosen by the GridSearchCV process were:

```
{'max_depth': 50, 'max_features': 20, 'n_estimators': 200}
```

After applying those parameters, the results of the test set were:

```
Recall: 0.7888
Precision: 0.8579
Accuracy: 0.8576
ROC_AUC: 0.8477
F1 score: 0.8219
```

Figure 7: Performance statistics of the Random Forest model with GridSearchCV

The GridSearchCV tuning parameters have not significantly improved the evaluation metrics, (as observed below). With the RF model, we got a great result identifying and predicting the observations with and without cancelation.

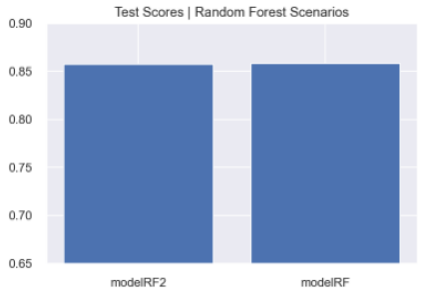

Figure 8: Random Forest Scenarios with (modelRF2) GridSearchCV and without (modelRF)

c. Gradient Boosting

Gradient Boosting is an excellent algorithm for dealing with bias-variance trade-off. GB classifier is an ensemble of regression or classification trees, unlike Random Forest models, in which all trees are built independently from one another. In GBC the setting of a sequential learning procedure to improve accuracy is employed, in which every new tree tries to correct the errors of previously built trees.

The GBC used for this task was pulled from the sklearn library. We started by fitting a GradientBoostingClassifier with the default parameters to get a baseline for the same test set.

```
Recall: 0.6508
Precision: 0.8836
Accuracy: 0.8188
ROC_AUC: 0.7948
F1 score: 0.7495
```

Figure 9: Performance statistics of the Gradient Boosting model.

After applying the GridSearchCV function for tunning the model, the selected parameters chosen by the GridSearchCV process were:

```
{'max_depth': 19, 'min_samples_split': 50}
```

The results of the test with the tuned model set were:

```
Recall: 0.7852
Precision: 0.8589
Accuracy: 0.8568
ROC_AUC: 0.8465
F1 score: 0.8204
```

Figure 10: Performance statistics of the Gradient Boosting Decision Trees

The GridSearchCV tuning parameters have increased by 4 points our evaluation metrics, as we can observe below. In the Gradient Boosting Decision Tree model, we got excellent results identifying and predicting the observations with and without cancelation.

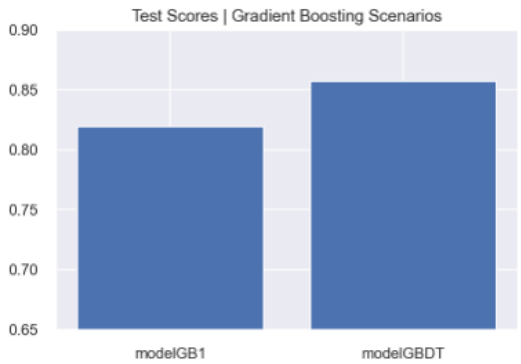

Figure 11: Gradient Boosting Scenarios, GB x GBDT

d. Stacking

Stacking is an ensemble machine learning algorithm. For this task, we used the sklearn library.
We started by using as estimators for the stacking model our best models for reducing bias (modelGBDT) and variance (modelRF2). For the final estimator, we used the neural network model tuned (modelNN2) with the most complex algorithm and the capability to learn, improve, and make the prediction more accurate.
We can observe below the results from the combination (stacking) of our all best models on the test set which allowed us to obtain our best results:

```
Recall: 0.7617
Precision: 0.8859
Accuracy: 0.8599
ROC_AUC: 0.8458
F1 score: 0.8191
```

Figure 12: Performance statistics of the Stacking

To better understand the customers/bookings in the data set, we applied K-means method to the data and obtained four different clusters of customers.
These clusters are very similar regarding number of adults considered in the booking, normally is 2 and in terms of children the majority of bookings include none. None of the clusters contemplates bookings with babies. Moreover, bookings are mainly from new customers, not company sponsored, they rarely make changes in their bookings and they don't require a car parking space at the hotel.
Besides, all clusters indicate that the customers stay for an average of 3 nights. The most opted room is the room A. Most costumers are transient.
Regarding the variables with higher discriminatory power, the clusters had the following characteristics:

Cluster 0	35587 bookings in these cluster
	43% of probability of cancelling
	Leadtime of 59 days
	1 adults & no children
	76.5% of bookings through agent
	ADR is 99 euros
	0 special requests
	58.6% stay during the weekend
	Book more in March, April and May
	Higher incidence of non refund deposit type

Cluster 1	5035 bookings in these cluster
	36% of probability of cancelling
	Leadtime of 86 days
	2 adults & 1 child
	93% of bookings through agent
	ADR is 151 euros
	1 special request
	49.3% stay during the weekend
	Book more in August and July
	Higher incidence of room type F

Cluster 2	23014 bookings in these cluster
	21% of probability of cancelling
	Leadtime of 74 days
	2 adults & no child
	93% of bookings through agent
	ADR is 114 euros
	1.4 special requests
	59% stay during the weekend
	Book more in April and May
	Higher incidence of rooms type D

Cluster 3	15420 bookings in these cluster
	70% of probability of cancelling
	Leadtime of 283 days
	2 adults & no child
	67% of bookings through agent
	ADR is 93 euros
	0 special requests
	48% stay during the weekend
	Book more from July to October
	Higher incidence of non-refund deposit type

Figure 13: Clusters

From this analysis we can conclude that the cluster 3 is the one with higher probability of cancelling followed by the cluster 0.

4.2 Assess the model

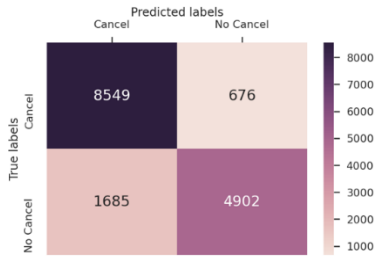

Figure 14: Confusion matrix

Since our success criteria is Accuracy, the most accurate model was stacking with a score of 86% it means that the model predicts correctly cancelations and no cancelation 86% of the times.
From the confusion matrix above we only overlooked 686 cancellations (false negatives).
For the segmentation we obtained an R2 score of 0.516, a Davies_Bouldin score of 1.134, and a Calinski-Harabasz index of 28113.3. Below can be seen the profiling of these clusters.

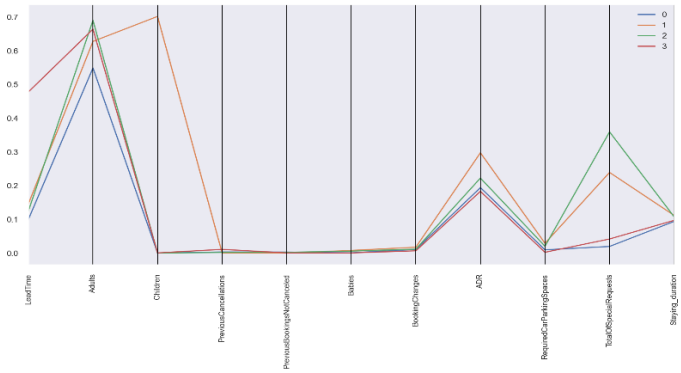

Figure 14: Cluster profiling

5. EVALUATION

5.1 Evaluation results

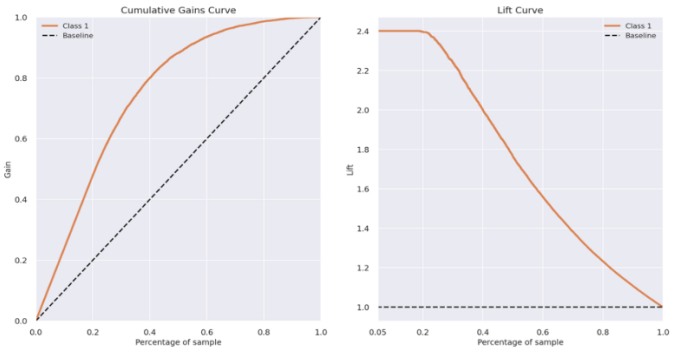

Figure 15: Cumulative gains curve and Lift Curve

When applying our Stacking model to the data set, we can predict 86% of the cancellations, which means that we will be able to decrease the rate by 35.8%. The new cancellation (after implementation) rate should be of 5.8%.
The Lift Curve shows that if we select the 20% of bookings with the highest probability of being cancelled, we will be able to predict cancellations 2.4 times better than a random guest. Within the

same 20% (of those with the highest probability of cancellation) we can find at least 50% of the actual cancellations.

Moreover, we concluded that the variable LeadTime is relevant in determining cancellation since the longer the period of booking time prior to the arrival is, the higher the probability of cancelation. There is also higher probability of cancelling when the booking is done directly. It is less probable when the average daily rate is much higher than 100 euros. This probability decreases when take place demanded special requests.
A non-refund deposit type doesn't seem to prevent cancellations, which indicates that many of them might not be planned and as it is known that the deal-seekers avoid this type of bookings because it troubles their strategy of managing simultaneous bookings, the hotel should avoid this type of policy.

5.2 Review process

Although the results of the model were satisfactory, there is a possibility that using different variable sets could have improved the performance. However, we did not explore this option. We believe that the existence of a booking ID variable would be valuable for proceeding with the deployment steps.

5.3 Next steps

Once the model is delivered to the Manager, the analysis should be performed every week (assuming that cancellations are allowed up until 72 hours prior to arrival).

6. DEPLOYMENT

With the insights given by applying the predictive model to the pool of bookings, the manager of the hotel obtains a more accurate indication of his hotel net demand. Using the knowledge of the bookings that might be cancelled the manager can act on them in order to prevent them.
We suggest two ways of approaching this based on the variable of LeadTime:
If the LeadTime is between 80 days and 150 days (the corresponding average LeadTime for no cancelation and cancelation) meaning the booking is made between 5 to 2 and half months prior to arrival time, the action to take is to remind to the costumers about the special services included in their booking and the extra ones that the hotel an provide. In low season periods, small treats (with low cost impact) can be offered to retain the bookings.
If the LeadTime is higher than 150 days (5 months) then we propose that these costumers should be contacted to confirm if they are still considering staying in your hotel or if something have changed since they first made their booking.
With these actions the hotel will be able to keep at least a week before a certain arrival date to implement overbooking policies with more certainty that they won´t fall in relocation or reputation costs since they have a clearer picture by using the model of the bookings that won´t translate in an actual arrival of a costumer to the hotel.
Regarding a more restrictive refunding policy we concluded that it doesn't prevent cancellation so the hotel should avoid implementing a non-refund policy as it decreases revenue and number of bookings and instead should focus on the results provided by the model to constraint the number of cancelations.