Business Analytics Report

Analysis of the Hotel Industry

Eero Pietikäinen

2002939

Åbo Akademi University

29.10.2025

Abstract

Data analytics have become an increasingly sought-after field in the world of business during the last few years, as the amount of digital data is larger than ever before. It is a crucial part of several industries like the finance, retail and e-commerce sectors. Another industry that benefits from analytics is the hotel industry. Traveling has become more accessible to people from all over the world due to cheaper flights and easy online booking services. Furthermore, with the emergence of social media, people can share content from all over the world, which makes people curious to explore. Due to high demand, hotels face difficulties like last-minute cancellations and double bookings which cause loss of revenue and difficulties in capacity handling. Moreover, hotels must keep up with competition and stick out from the crowd. This analysis will make a deep dive into the hotel industry by analyzing hotel booking and customer review data. This paper follows the CRISP-DM framework which is a structural data mining process. It is composed of six steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

Table of Contents

Business Understanding

Customer booking and cancellation decisions in the hotel industry are spread on a large spectrum of different kinds of factors. These include booking lead time, price sensitivity, deposit policies and market segments. According to Liu et al. (2025) longer lead times and online bookings increase the risk of cancellations, as customers face less barriers to cancel their trip. Seasonal trends and global happening are again external factors that might influence customers decision-making, which furthermore adds to list of demand uncertainty. Understanding these factors is crucial for hotel businesses as they can lead to loss of revenue and distorted forecasting.

Accurate predictions of cancellations reinforce revenue management, pricing and capacity handling (Liu et al. 2025). Furhermore, predictive models such as XGBoost and hybrid approaches have shown positive impact on predicting cancellations accurately. These predictions have helped hotels to allocate their resources more efficiently and by cutting last-minute losses (Ampountolas, 2025).

This analysis' main business objective is to reduce last minute cancellations, enhance occupancy and pricing, and improve customer retention through proactive engagement with the customers. The insights gathered from the prediction models and sentiment analysis will guide marketing, pricing and service improvement strategies by identifying customers segments with high risks. Moreover, the analysis will recognize the dissatisfactory elements that customers experience during their stays. As these strategies help hotels to tailor their incentives, communication and cancellation policies, which increase customer satisfaction and revenue optimization (Liu et al., 2024).

Data Understanding

Available data

The dataset includes information that describes customer, booking and stay characteristics. It includes numerical variables such as the number of days before arrival, whether the customer has children or not and what the date of arrival was. The categorical variables, on the other hand, describe factors such as arrival month, the type of meal booked and what kind of a room the customer booked. The target variable describes whether the customer cancelled the booking (0 = not cancelled and 1 = cancelled).

The full data description is presented below.

Outcome:

Target: Outcome variable that identifies cancelled bookings as 1 and non-cancelled bookings as 0.

Numeric

Days_before_arrival: Numerical variable that represents the number of days between the booking date and the arrival date.

Arrival_day_of_month: Numerical variable representing the day of the month of the arrival date.

Weekend_nights: Numeric variable with the number of weekend nights the guest has stayed or booked.

Week_nights: Numerical variable representing the number of nights between Monday to Friday, the guest stayed or booked.

Adults: Numeric variable of the number of adults in the booking.

Children: Numeric variable of the number of children in the booking.

Cancellations: Numerical variable representing the number of previous bookings cancelled by the client before the current booking.

Bookings_not_cancelled: Numerical variable representing the number of previous bookings not cancelled by the client before thecurrent booking.

Booking_changes: Numerical variable representing the number changes made to the booking.

Average_daily_rate: Numerical variable, calculated by dividing the sum of all lodging transactions by the total number of staying nights.

Parking_spaces: Numerical variable related to the number of car parking spaces required by the customer.

Special_requests: Numerical variable that corresponds to a number of special requests made by the customer.

Categorical

Arrival_month: Numerical variable representing the month of arrival.

Meal: Categorical variable representing the type of food booked. 0: BB – Bed & Breakfast; 1: FB – Full Board; 2: HB – Half Board; 3: SC –Self-Catering; 4: Undefined.

Market_segment: Categorical variable representing market segment designation: 0: Direct; 1: Corporate; 2: Online TA; 3: Offline TravelAgents/Tour Operators; 4: Complementary; 5: Groups; 6: Undefined; 7: Aviation.

Repeated_guest: Categorical variable with 1 if the booking name is of a repeated guest and 0 if it is not repeated.

Reserved_room_type: Categorical variable of room type code booked.

Assigned_room_type: Categorical variable of room type code assigned.

Deposit_type: Categorical variable indicating whether the client deposited to guarantee the booking.0: No Deposit – no deposit was made;1: Refundable – a deposit was made with a value under the total cost of the stay;2: Non Refund – a deposit was made in the value of the total stay cost.

Information about the dataset

The dataset consists of 15 000 rows and 20 columns. The datatypes include floats, integers and categorical objects.

```
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Days_before_arrival     15000 non-null  int64
 1   Arrival_day_of_month    15000 non-null  int64
 2   Weekend_nights          15000 non-null  int64
 3   Week_nights             15000 non-null  int64
 4   Adults                  15000 non-null  int64
 5   Children                14250 non-null  float64
 6   Cancellations           15000 non-null  int64
 7   Bookings_not_cancelled  15000 non-null  int64
 8   Average_daily_rate      15000 non-null  float64
 9   Booking_changes         15000 non-null  int64
 10  Parking_spaces          14250 non-null  float64
 11  Special_requests        15000 non-null  int64
 12  Arrival_month           15000 non-null  object
 13  Meal                    15000 non-null  object
 14  Market_segment          15000 non-null  object
 15  Repeated_guest          14250 non-null  float64
 16  Reserved_room_type      15000 non-null  object
 17  Assigned_room_type      15000 non-null  object
 18  Deposit_type            15000 non-null  object
 19  Target                  15000 non-null  int64
dtypes: float64(4), int64(10), object(6)
```

Summary statistics

Summary statistics:

	Days_before_arrival	Arrival_day_of_month	Weekend_nights	Week_nights	Adults	Children	Cancellations
count	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	14250.000000	15000.000000
mean	103.893067	15.652000	0.918267	2.488533	1.851333	0.106456	0.088333
std	106.828004	8.812805	0.994411	1.910945	0.600824	0.404334	0.819375
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	18.000000	8.000000	0.000000	1.000000	2.000000	0.000000	0.000000
50%	69.000000	16.000000	1.000000	2.000000	2.000000	0.000000	0.000000
75%	159.000000	23.000000	2.000000	3.000000	2.000000	0.000000	0.000000
max	629.000000	31.000000	19.000000	50.000000	27.000000	3.000000	26.000000

Bookings_not_cancelled	Booking_changes	Average_daily_rate	Parking_spaces	Special_requests
15000.000000	15000.00000	15000.000000	14250.000000	15000.000000
0.135733	0.22140	102.309155	0.062316	0.562133
1.421447	0.67321	64.726648	0.242895	0.789629
0.000000	0.00000	0.000000	0.000000	0.000000
0.000000	0.00000	69.000000	0.000000	0.000000
0.000000	0.00000	94.500000	0.000000	0.000000
0.000000	0.00000	126.000000	0.000000	1.000000
66.000000	17.00000	5400.000000	2.000000	5.000000

The descriptive statistics of the dataset show huge variations. *Days_before_arrival* and *Average_daily_rate* show a wide range, which indicates that customers have dissimilar planning horizons and budgets. *Special_requests and Booking_changes* have low averages, but they might be meaningful factors, because customers who change their bookings show higher engagement.

Missing values per column

Only 3 of the 20 columns include missing values. These will be handled later on in the data preparation stage.

```
Missing values per column:
Days_before_arrival         0
Arrival_day_of_month        0
Weekend_nights              0
Week_nights                 0
Adults                      0
Children                  750
Cancellations               0
Bookings_not_cancelled      0
Average_daily_rate          0
Booking_changes             0
Parking_spaces            750
Special_requests            0
Arrival_month               0
Meal                        0
Market_segment              0
Repeated_guest            750
Reserved_room_type          0
Assigned_room_type          0
Deposit_type                0
Target                      0
dtype: int64
```

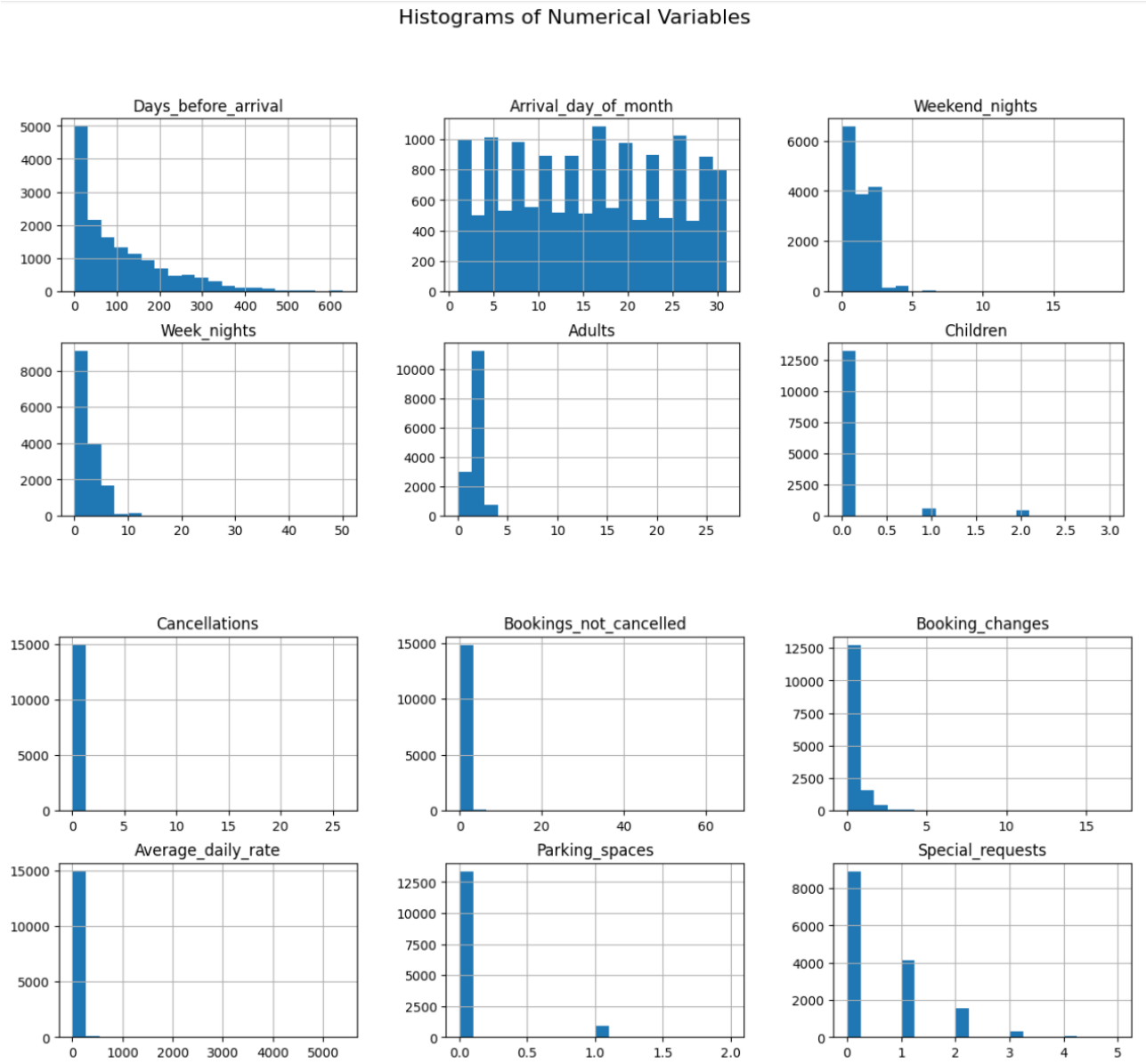
Unique values per categorical column

```
Arrival_month: 12 unique values

Meal: 5 unique values

Market_segment: 7 unique values

Repeated_guest: 2 unique values

Reserved_room_type: 10 unique values

Assigned_room_type: 11 unique values

Deposit_type: 3 unique values

Target: 2 unique values
```

Arrival_month includes the most values as it includes the 12 months in a calendar year. *Target* includes the least values at it is a binary variable. The categorical variables number of unique values are relatively small.

Histograms of numeric variables

Histograms were used to examine the numerical variables shapes and spread.



Days_before_arrival shows a right-skewed distribution, which indicates that most customers book their hotel some days before their arrival date. The assumption is that bookings made far in advance are more vulnerable to cancellations.

Booking_changes and *Special_requests* show peaks in relatively low numbers, which means that customers make a few to none changes or special requests at their stay.

Weekend_nights and *Week_nights* illustrate moderate variations and show that most customers keep their trips short, possibly reflecting casual leisure stays and business trips.

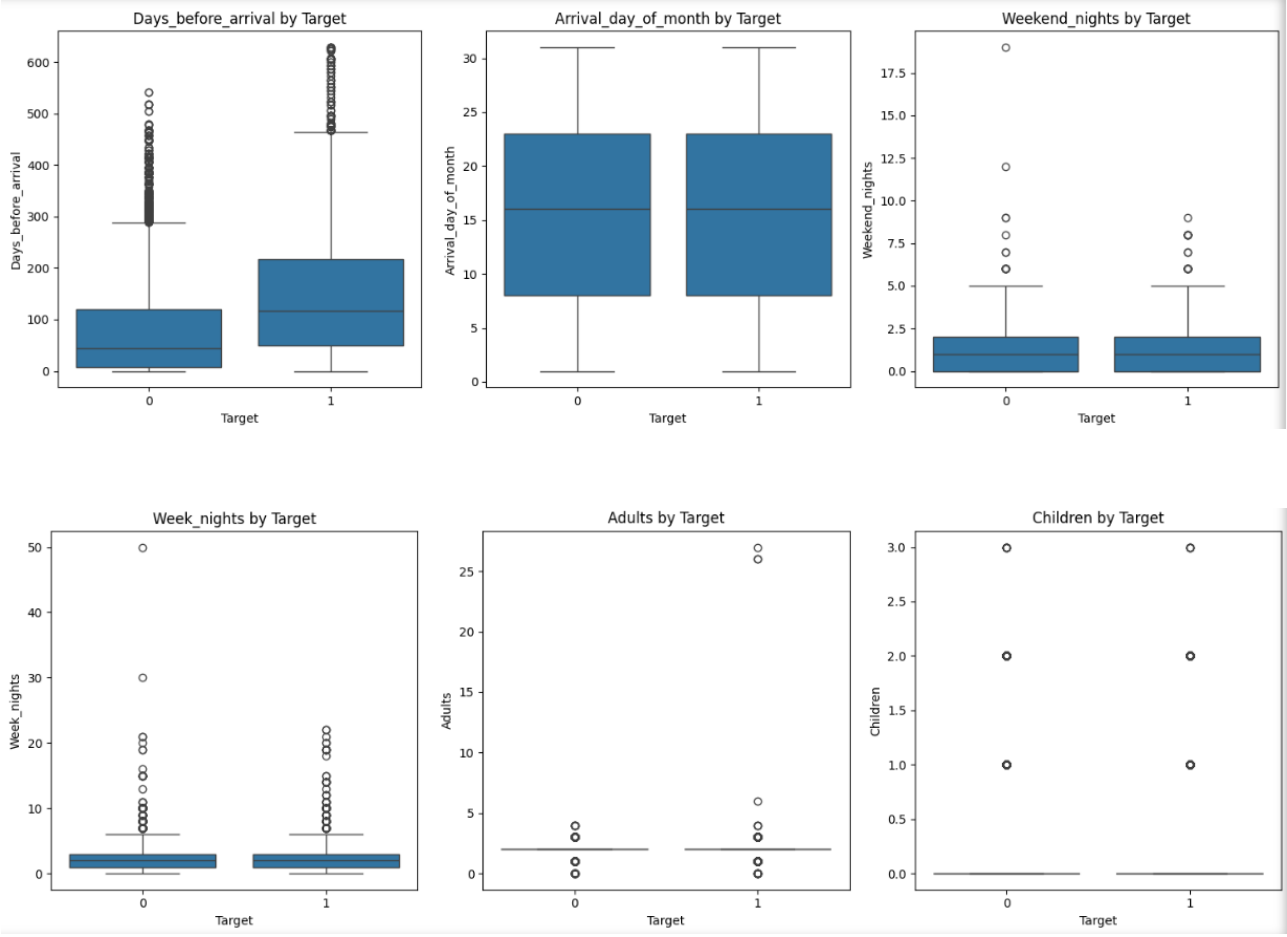
The illusrations of *Adults* and *Children* show that most of the customers are adults and a very small percentage of bookings include any children. All the bookings include under five people, with a large peak at approximately two people, which can indicate that couples are a big part of the customer base.

Parking_spaces show that very few customers require parking spaces, which indicates that people do not arrive by car or that they park elsewhere, possibly because of higher prices at the hotel.

Overall the histograms show that numerical variables are dominated by smaller and more frequent values with only a few outliers.

Boxplots of numerical variables versus Target

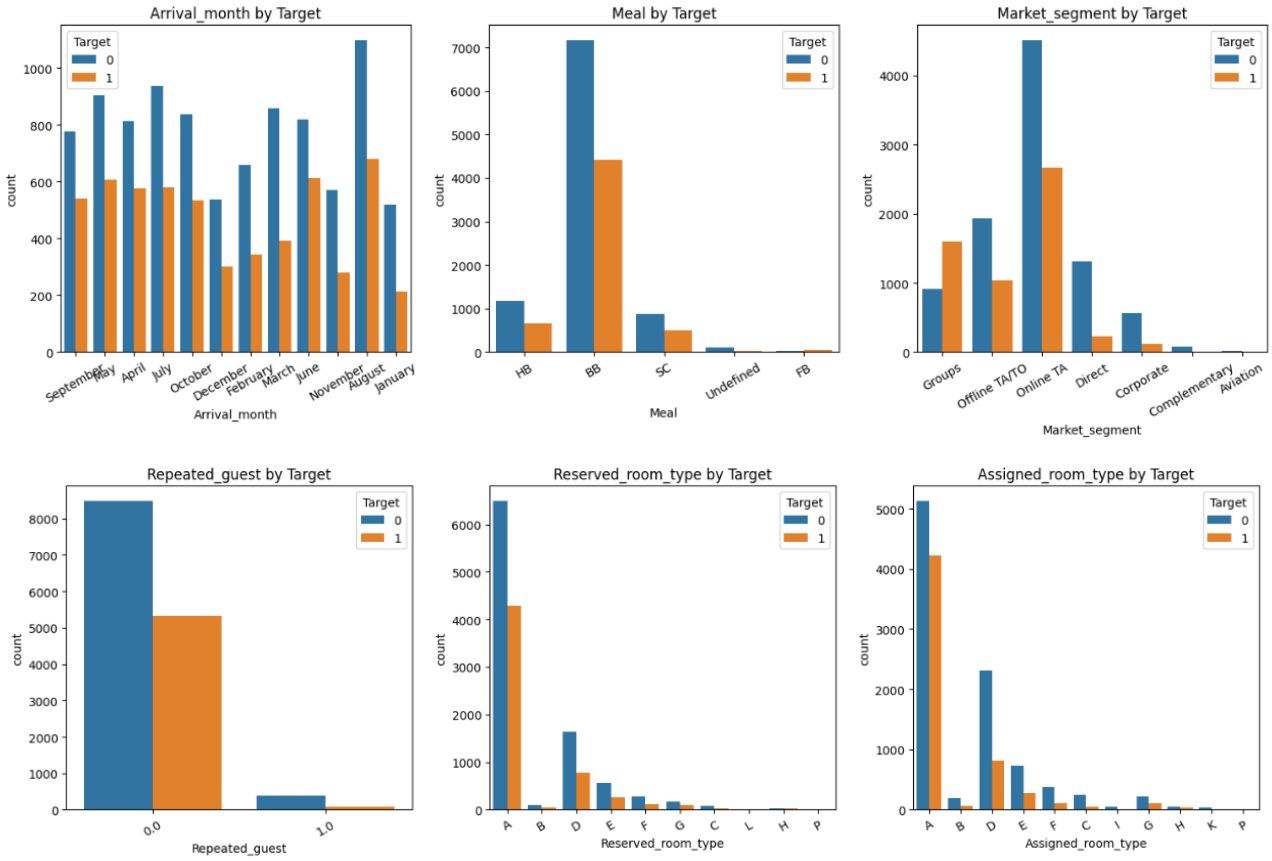
Boxplots were also used to visualize data spread and outlier detection, this time comparing the results with the target variable.



Most boxplots do not give out useful information because the values of the variables are spread at a very small range. However, the boxplot of *Days_before_arrival* shows that people that book far in advance tend to cancel their bookings more often than people who book closer to their arrival date. The figure also illustrates clusters of outliers, which means that many people still book their stays extremely early on.

Countplots for categorical variables

Countplots were used to visualize the frequency of the categorical variables and to compare them with the target value.



Market_segment shows that most of the bookings are made through online travel agencies. Online travel agencies also show the highest proportion of cancellations, but this happens naturally as they are the most popular market segment by a landslide. However, groups as a market segment are the only ones that have more cancellations than non-cancellations. This is also natural, as group reservations are usually made for events, conferences etc. and have a lot of logistical uncertainties.

In the *Meal_plan* countplot bed & breakfast dominates the data. Full board, half board and self service are not that frequent but tend to be more prone to cancellations.

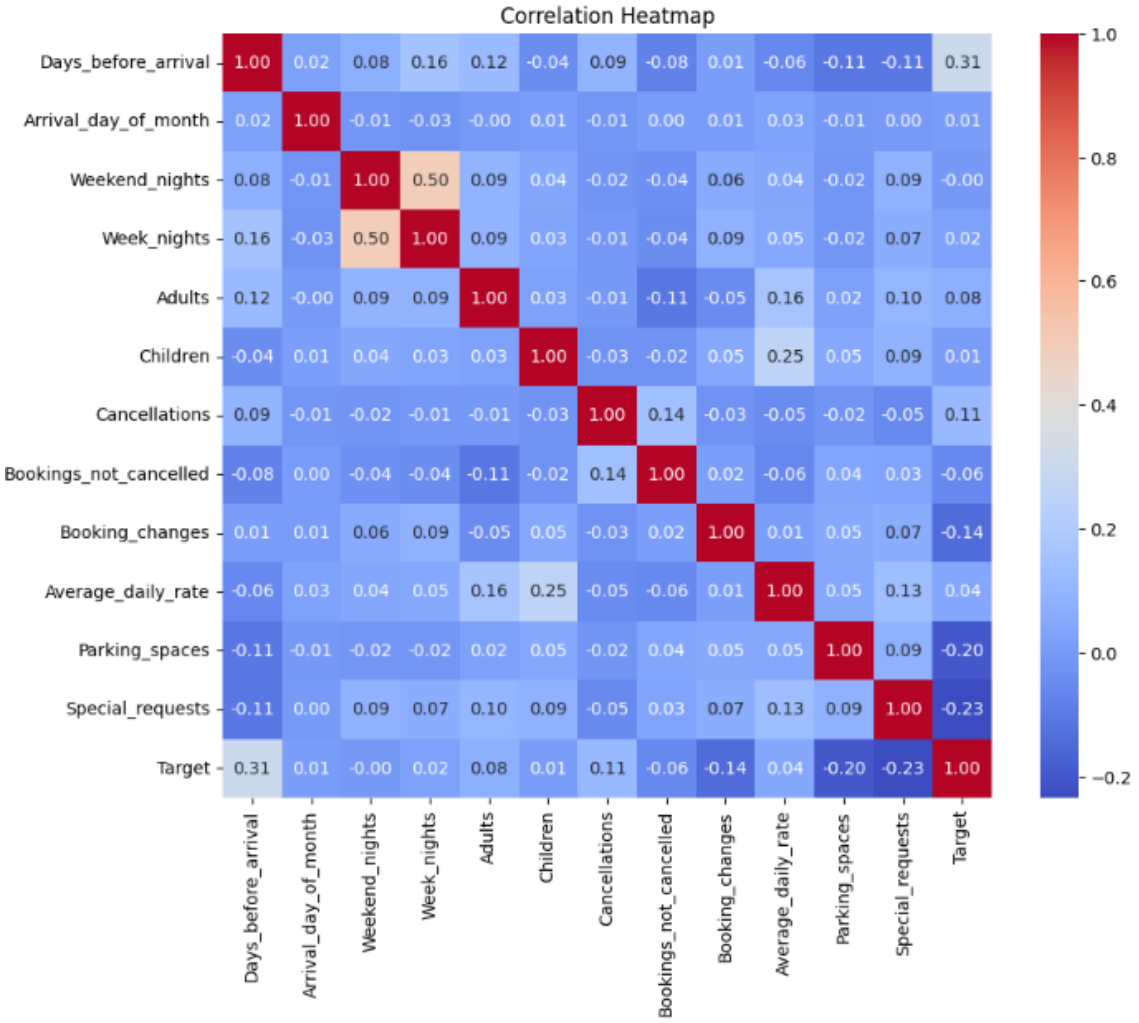
Reserved_room_type and *Assigned_room_type* show that room type A is the most popular type by far, but when assigned it gets tons of cancellations. This can mean that customers who booked better rooms want to cancel their booking if they get assigned to the type A rooom. The opposite happens with the other rooms. For example, room type D gets a lot less cancellations when assigned. A probable cause can be that they got assigned a better room than they originally booked.

The *Repeated_guest* countplot visualizes the fact that regular customer makes cancellations rarely, which is expected.

Arrival_month shows that the most popular month is August, which is a usual month for people to be on summer holidays.

Correlation heatmap for numerical variables

A correlation heatmap illustrates the relationships between variables.



Days_before_arrival show a positive correlation with *Target*, which goes hand in hand with the earlier decriptions.

Special_requests and *booking_*changes are negatively correlated with *Target* which means that customers that make requests make less cancellations. A possible explanation for this could be that they are more invested in their stay.

Average_daily_rate has a mild positive correlation with *Target*, possibly due to price sensitivity.

Average_daily_rate and *Children* have a positive correlation. Families tend to use more money as they need bigger rooms and have more mouths to feed.

In conclusion, the main early insights from the descriptive analytics are that most people book their stay close to the arrival day, but people who book their stay early are more likely to make cancellations. In addition, the mean for average daily rate is moderate, but a few outliers increase the average significantly. The maximum average daily rate of 5400 is nearly 43 times higher than

the 75th percentile with an average daily rate of 126. This indicates that the customer behavior is price sensitive with most people booking the affordable room type.

The variable *room_type_assigned* shows that a lot of cancellations happen when assigned to a worse room, which means that double-bookings and reassigning rooms need to be kept at a minimum. Additionally, people who modify their booking or make requests are less likely to cancel, which can be a sign for the hotel to somehow keep the customers more engaged. The illustration also shows that the most typical bookings consist of two adults, no children and parking spaces, which indicates that the typical guests are small leisure or business parties. Finally, the numerical variables fall within logical ranges and have a relatively small number of missing values, which is promising for the modelling part of the analysis.

Data Preparation

The data preparation part ensures that the data is complete, clean and ready for the machine learning models.

Missing values per column (%)

```
Missing values (%):
                        Count  Percent
Parking_spaces            750      5.0
Repeated_guest            750      5.0
Children                  750      5.0
Days_before_arrival         0      0.0
Week_nights                 0      0.0
Weekend_nights              0      0.0
Arrival_day_of_month        0      0.0
Cancellations               0      0.0
Bookings_not_cancelled      0      0.0
Average_daily_rate          0      0.0
Booking_changes             0      0.0
Adults                      0      0.0
Special_requests            0      0.0
Arrival_month               0      0.0
Meal                        0      0.0
Market_segment              0      0.0
Reserved_room_type          0      0.0
Assigned_room_type          0      0.0
Deposit_type                0      0.0
Target                      0      0.0
```

Only three variables: *Parking_spaces, Repeated_guests* and *Children* include missing values. In all the three columns, only 5% are missing values.

Handling missing values

Because the number of missing values was minimal, an imputation strategy was applied instead of removing the values. The minimal amount of the missing values ensures that the imputation won't distort the results of the machine learning models.

For the numerical variables (*Parking_spaces* and *Children*), the missing values were replaced by each column's median. This maintains the distribution while not being affected by outliers. For the categorical variable (*Repeated_guest*) missing values were replaced by the column's mode. This imputation strategy ensures that no data is lost while keeping realistic values in each column.

After performing the imputation, we must make sure it worked. There indeed are no missing values left:

```
Missing values after imputation:
 0
```

Outliers

Outliers were detected with the help of the Interquartile Range (IQR) method.

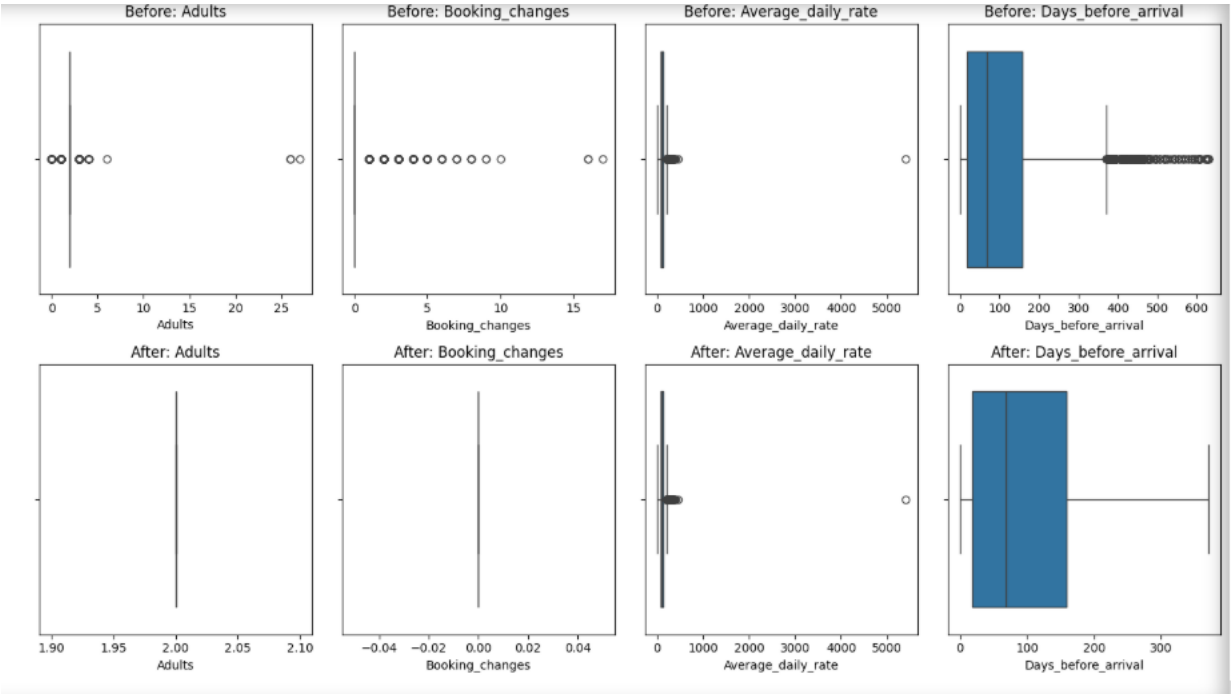
	Outlier_count	Outlier_%
Adults	3768	25.120000
Booking_changes	2246	14.973333
Cancellations	848	5.653333
Bookings_not_cancelled	460	3.066667
Week_nights	412	2.746667
Days_before_arrival	371	2.473333
Special_requests	353	2.353333
Weekend_nights	33	0.220000
Arrival_day_of_month	0	0.000000
Target	0	0.000000

Several variables showed extreme values. Instead of removing them, capping was used. Here we set the values upper bound or below the lower bound in each of their thresholds. This strategy keeps all observations but reduces the influence of the extreme values on the upcoming models.

Again, we need to make sure that the strategy worked. The list of outliers is now blank:

Remaining_outliers	Remaining_%

It can also be visualized through boxplots:



Infrequent categories

Categorical variables were analyzed to find categories that appear in less than 1% of the observations. These rare categories may produce noise and instability in the learning models as they receive too little data to learn reliable patterns in those cases.

The following categories appeared in less than 1% of the observations:

```
Meal - Rare categories:
Meal
Undefined    0.009400
FB           0.005333
Name: proportion, dtype: float64

Market_segment - Rare categories:
Market_segment
Complementary    0.005800
Aviation         0.001667
Name: proportion, dtype: float64

Reserved_room_type - Rare categories:
Reserved_room_type
B    0.009133
C    0.006933
H    0.004200
P    0.000133
L    0.000067
Name: proportion, dtype: float64

Assigned_room_type - Rare categories:
Assigned_room_type
H    0.005200
I    0.003400
K    0.002733
P    0.000133
Name: proportion, dtype: float64
```

Handling Infrequent Categories

To address the problem, all rare categories were combined in to a “Other” category. This strategy keeps information about the rare case while preventing distortion in the models estimates. It also makes the structure of the variables simpler by making it more evenly distributed, which is better for generalization and interpretability as we can see:

```
Meal value counts after replacement:
Meal
BB       0.771
HB       0.122
SC       0.092
Other    0.015
Name: proportion, dtype: float64

Market_segment value counts after replacement:
Market_segment
Online TA        0.478
Offline TA/TO    0.199
Groups           0.167
Direct           0.103
Corporate        0.045
Other            0.007
Name: proportion, dtype: float64

Reserved_room_type value counts after replacement:
Reserved_room_type
A        0.719
D        0.162
E        0.055
F        0.025
Other    0.020
G        0.018
Name: proportion, dtype: float64

Assigned_room_type value counts after replacement:
Assigned_room_type
A        0.623
D        0.208
E        0.067
F        0.033
G        0.022
C        0.019
B        0.017
Other    0.011
Name: proportion, dtype: float64
```

Feature Selection

A variance threshold analysis was conducted to detect redundant columns that do not give any useful information. Low variance columns show little to no variation in their observations, which means that all observations practically include the same values or very similar values. They provide no ground for the learning models to detect patterns, and they add unnecessary complexity.

Highly correlated variables, on the other hand, show a strong linear relationship, which means that one variable can be largely predicted from the other. It is known as multicollinearity that occurs when variables include overlapping information, which may distort results of the learning models.

After conducting the variance threshold and correlation analysis, the results look the following:

```
Low variance columns: ['Adults', 'Cancellations', 'Bookings_not_cancelled', 'Booking_changes']
Highly correlated columns: []
```

As we can see, four columns came back as having zero-variance. No highly correlated variables were found.

It is also good to check if there are any columns that do not have zero-variance but still have a low variance. The results came back the following:

```
Cancellations                0.000000
Bookings_not_cancelled       0.000000
Adults                       0.000000
Booking_changes              0.000000
Repeated_guest               0.029914
Parking_spaces               0.056229
Children                     0.155839
Target                       0.235051
Special_requests             0.546385
Weekend_nights               0.919831
Week_nights                  2.294926
Arrival_day_of_month        77.660363
Average_daily_rate        4189.259604
Days_before_arrival      10208.824666
dtype: float64
```

As we can see, *Repeated_guests* have a low variance, but we will keep it in our dataset for interpretability as it describes core booking attributes.

Handling redundant columns

The zero-variance columns (*Cancellations, Bookings_not_cancelled, Adults* and *Booking_changes)* will be dropped as they are not useful to us.

After dropping, we have 16 columns instead of 20 as we can see:

	Days_before_arrival	Arrival_day_of_month	Weekend_nights	Week_nights	Children	Average_daily_rate	Parking_spaces
0	339.0	26.0	3.0	6.0	0.0	80.00	0.0
1	323.0	17.0	0.0	4.0	0.0	86.00	0.0
2	26.0	5.0	2.0	5.0	0.0	64.00	0.0
3	193.0	28.0	0.0	3.0	1.0	98.50	0.0
4	15.0	11.0	2.0	4.0	0.0	92.17	0.0

```
(15000, 16)
```

Encode categorical variables

Machine learning models usually require numerical inputs. That is why categorical variables need to be converted to numerical formats in the form of encoding.

Two of the categorical variables are binary-encoded (*Repeated_guest* and *Target*):

```
Checking binary-encoded columns:

Repeated_guest: unique values = [0 1]
Target: unique values = [1 0]
```

Variables with more than two categories are encoded by transforming them into multiple binary columns. Examples of the new columns can be seen below:

```
 One-hot encoded columns (sample):
['Arrival_month_August', 'Arrival_month_December', 'Arrival_month_February', 'Arrival_month_January', 'Arrival_mont
h_July', 'Arrival_month_June', 'Arrival_month_March', 'Arrival_month_May', 'Arrival_month_November', 'Arrival_month
_October', 'Arrival_month_September', 'Meal_HB', 'Meal_Other', 'Meal_SC', 'Market_segment_Direct', 'Market_segment_
Groups', 'Market_segment_Offline TA/TO', 'Market_segment_Online TA', 'Market_segment_Other', 'Reserved_room_type_
D']
```

Scaling Numeric Features

All numerical variables were standardized using z-score normalization. Variables are now on a comparable scale, which improves the performance of models such as logistic regression and neural networks. All numerical variables now have a mean of 0 and a standard deviation of 1:

```
Feature means after scaling:
 Days_before_arrival     0.0
Arrival_day_of_month    0.0
Weekend_nights         -0.0
Week_nights            -0.0
Children                0.0
Average_daily_rate      0.0
Parking_spaces          0.0
Special_requests       -0.0
Repeated_guest         -0.0
Target                  0.0
dtype: float64

Feature standard deviations after scaling:
 Days_before_arrival     1.0
Arrival_day_of_month    1.0
Weekend_nights          1.0
Week_nights             1.0
Children                1.0
Average_daily_rate      1.0
Parking_spaces          1.0
Special_requests        1.0
Repeated_guest          1.0
Target                  1.0
dtype: float64
```

After the data preparation, we now have different datasets that we can use depending on the model: A clean dataset and a clean, encoded and scaled dataset.

Splitting data into training and testing sets

To start off model training and unbiased evaluation, the dataset was split into training (80%) and testing (20%) sets. The target variable was stratified during splitting with the intention of maintaining the original balance between cancelled and non-cancelled bookings. The training data is used for model fitting and parameter tuning, while the testing data is used for performance validation. Below you can see the sizes of the sets:

```
Training set size: (12000, 42)
Test set size: (3000, 42)
```

Modeling

In the modeling phase we aim to predict the likelihood of cancellations using three machine learning models: A logistic regression model, tree-based models and neural networks. Each model was evaluated using a set of measures (Shmueli et al., 2023):

Accuracy: Accuracy is the overall classifier of effectiveness and is found by (TP+TN)/sample size.

Precision: Precision value shows all instances predicted as positive and it is crucial when false positives are costly.

Recall: This value is calculated by TP/(TP+FN).

F1 Score: F1 Score gives the harmonic mean of precision and recall. We can calculate it by 2 x [(Precision x Recall)/(Precision + Recall).

AUC-ROC: ROC (Receiver Operator Characteristic) shows the trade-off between correctly identifying positives and incorrectly classifying negatives as positives. AUC (Area Under the Curve) is a single number summarizing the ROC-curve.

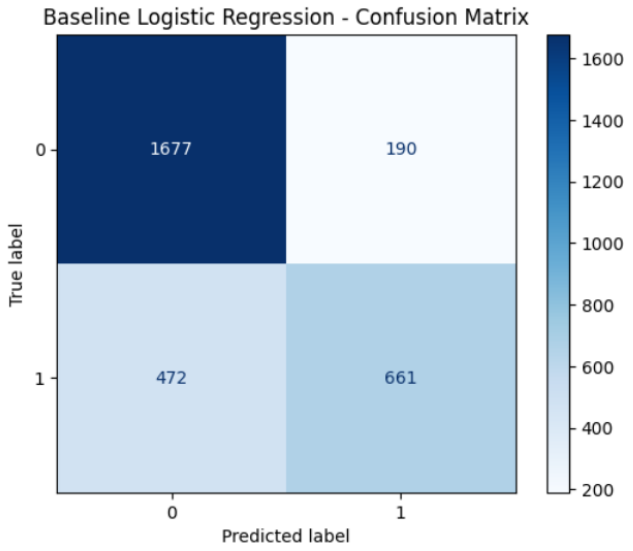
MSE: Calculates the average of the squared differences between observed and predicted values in a model.

Model Building Part 1

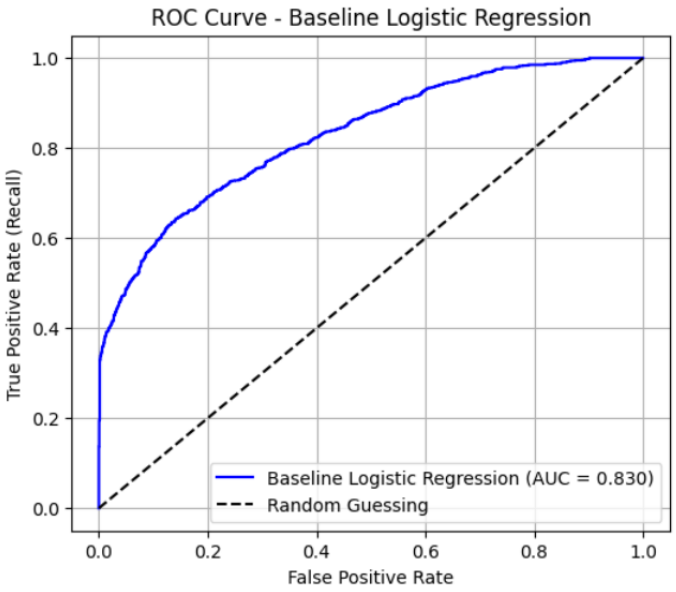
Baseline Logistic Regression Model

A simple regression model was done as the baseline because of its simplicity. It estimates the probability of a booking being cancelled based on weighted contributions from the input variables.

```
Baseline Logistic Regression Performance:
Accuracy:  0.7793
Precision: 0.7767
Recall:    0.5834
F1 Score:  0.6663
MSE:       0.2207
```

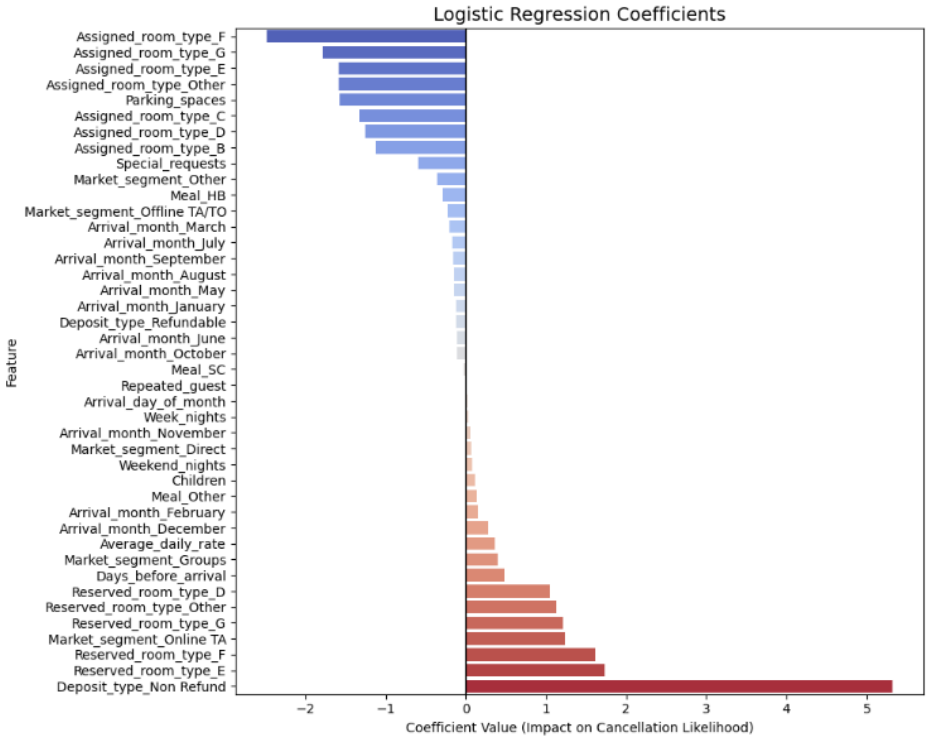


The model achieved an accuracy of 77.9%, which is good for a baseline model. The confusion matrix above is a visualization on the distribution of (going in the order from top left to bottom right) true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). The regression model has 1677 true positives, 190 false positives, 472 false negatives and 661 true negatives.



ROC AUC Score (Baseline Logistic Regression): 0.8301

The ROC-curve rises steeply at low false-positive rates, which shows that the model catches a large number of true cancellations without getting false alarms quite effectively. The model's relatively high AUC proves that the model's performance is quite reliable. The model is not perfect, but it forms a solid baseline for comparing the more complex models, where we will use parameter tuning as well.



The coefficient plot reveals the influence that each feature has on the possibility of a booking being cancelled. Positive coefficients increase the probability of cancellations, while negative coefficients

reduce the probability. The strongest positive predictor is the non-refund deposit type, which is followed by certain room types and online travel agencies. These can possibly reflect on higher risk online bookings. The strongest negative predictors are certain room types, parking spaces, special requests among others. These may be loyal customers that plan their stays more attentively. Features that have a moderate influence include seasonal and such as arrival month and days before arrival. It shows that the number of cancellations may fluctuate depending on travel seasons and the timing of the bookings.

Tree-Base Models

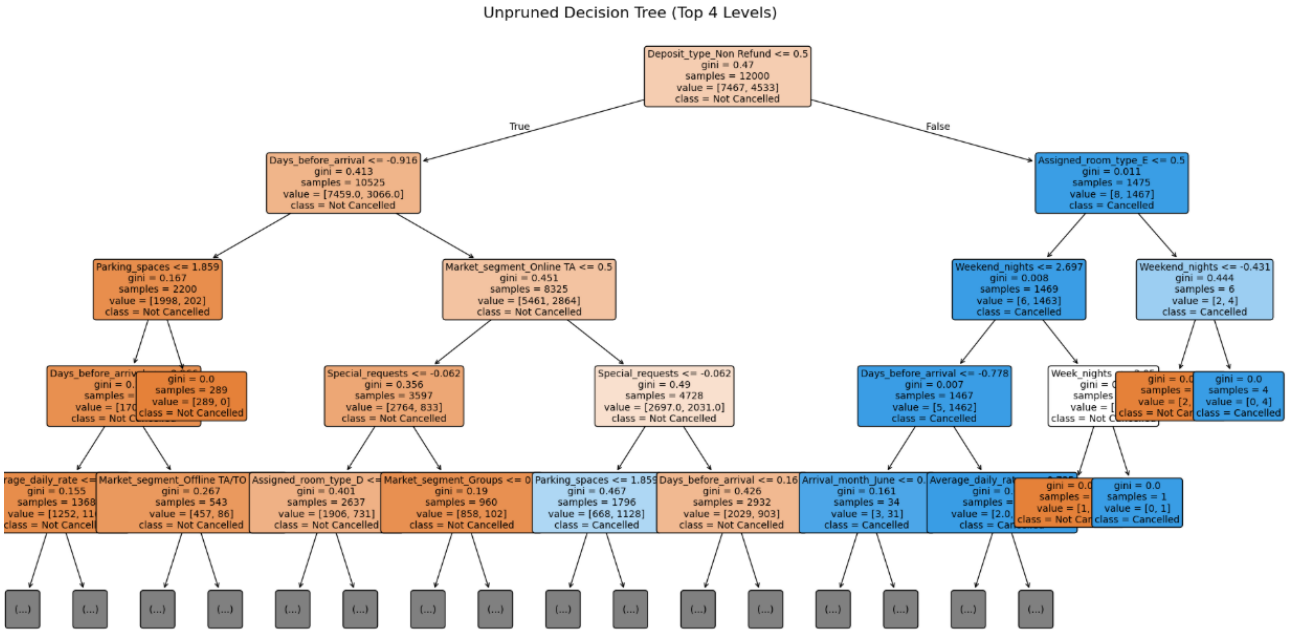
Decision tree and random forest models were developed with the intention of catching possible non-linear relationships and interactions between booking behavior and cancellations.

A new measure is introduced as well for the tree-based models, which is the Gini-coefficient. Gini coefficient gives information about how well a model distinguishes between the two classes. In our case it is cancellations and non-cancellations. It's closely related to the AUC and calculated by: Gini = 2 × AUC – 1 (Shmueli et al. 2023).

Unpruned decision tree

We will start off with an unpruned decision tree as it provides contrast to the upcoming optimized decision trees. An unpruned decision tree grows until there are no further splits possible. It minimizes bias but maximizes variance.

```
Unpruned Decision Tree Performance:
Accuracy:  0.7393
Precision: 0.6456
Recall:    0.6867
F1 Score:  0.6655
ROC AUC:   0.7288
MSE:       0.2607
Avg Gini:  0.1823
```

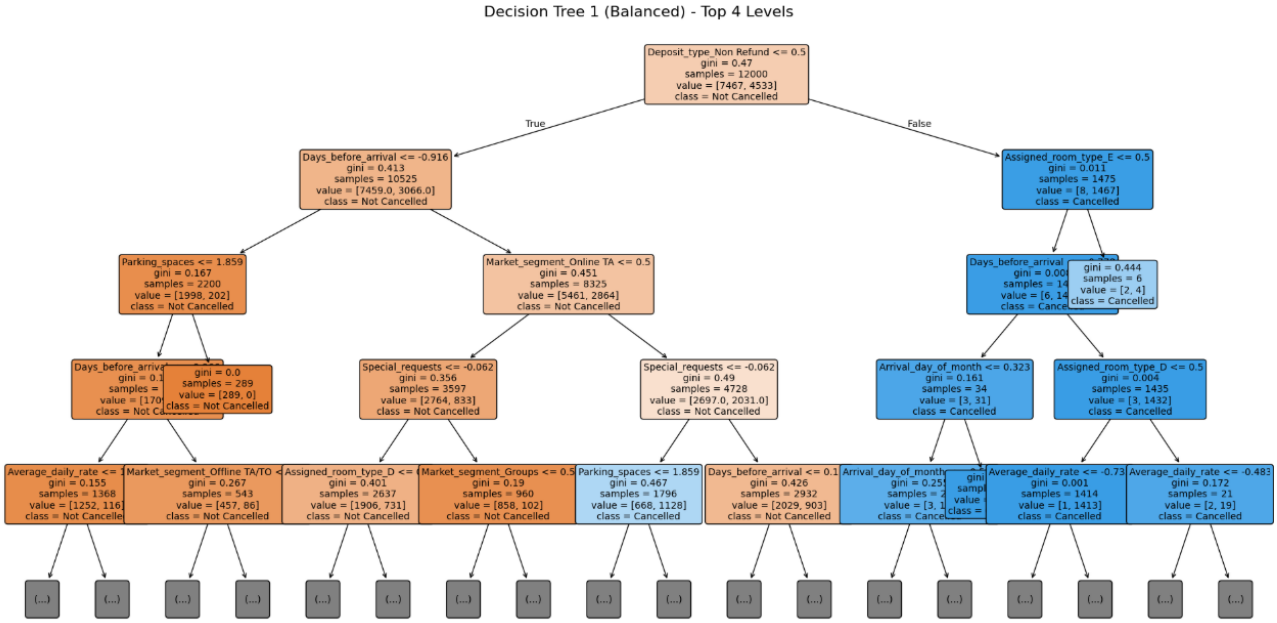


As we can see, the performance measure of 73.9% accuracy is average at best, as are the other measures. Especially the gini-coefficient is quite bad, as the value ranges from 0 to 1and a higher value indicates better separation.

Balanced Decision Tree

Next up we will try out a “balanced” decision tree with a maximum depth of 5 and the minimum number of samples being 5 as well.

```
Decision Tree 1 (Balanced) Performance:
Accuracy:  0.7757
Precision: 0.7861
Recall:    0.5578
F1 Score:  0.6526
ROC AUC:   0.8169
MSE:       0.2243
Avg Gini:  0.2319
```

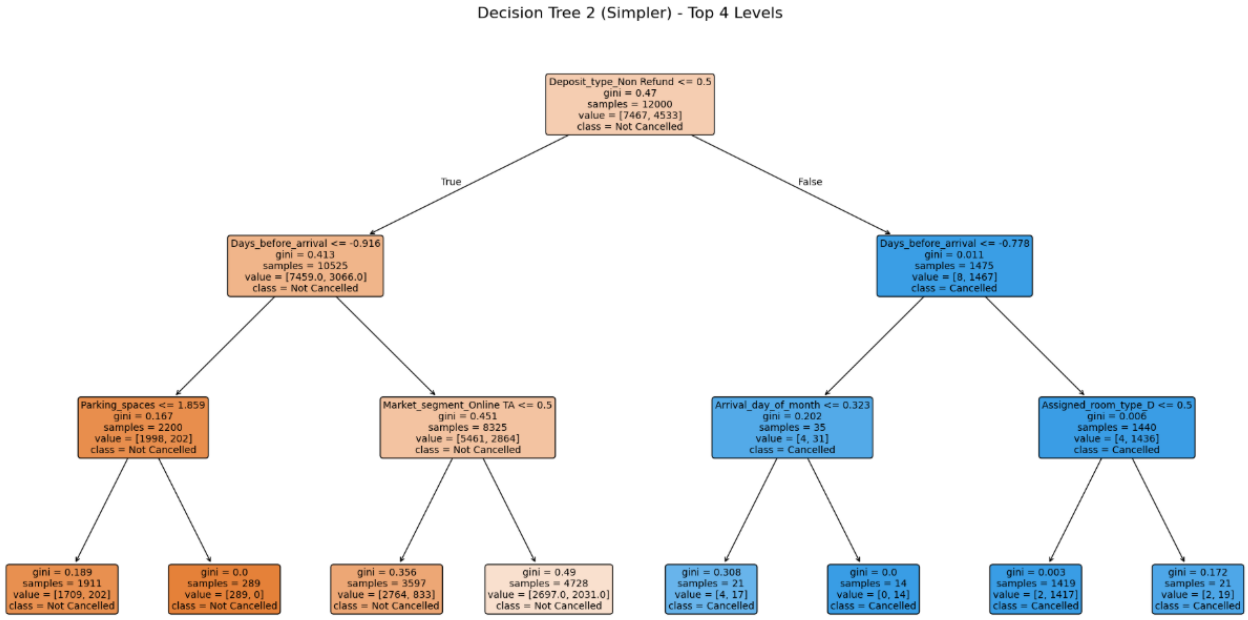


The balanced decision tree achieved an accuracy of 77.6%, which is a slight improvement. Many other measures improved as well. Precision and ROC-AUC increased significantly, while MSE and Gini got moderate improvement. However, recall dropped quite a lot and the F1-score dropped minimally.

Simpler Decision Tree

Now we will try out a simple decision tree with the maximum depth being 3 and minimum samples being 10.

```
Decision Tree 2 (Simpler) Performance:
Accuracy:  0.7443
Precision: 0.9867
Recall:    0.3274
F1 Score:  0.4917
ROC AUC:   0.7785
MSE:       0.2557
Avg Gini:  0.2159
```

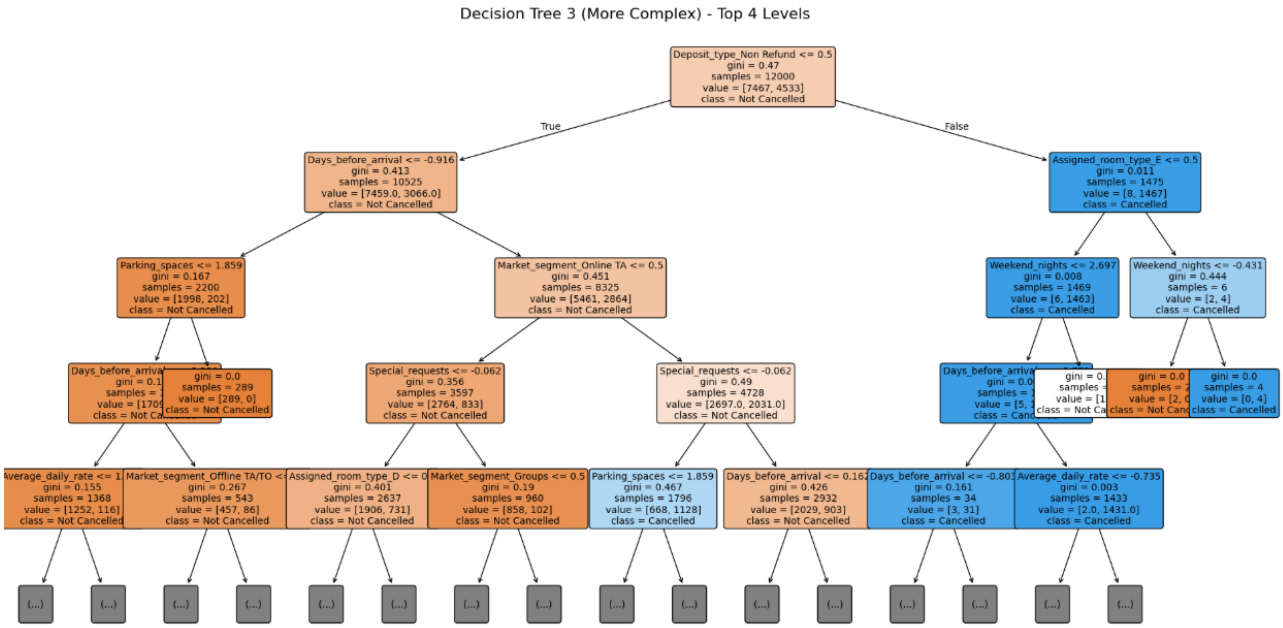


The precision on this tree is excellent with a score of 98.7%. However, the accuracy dropped to 74.4%. All of the other measures decreased as well.

Complex Decision Tree

Finally, we will try out a more complex decision tree with a maximum depth of 8 and the minimum samples being 2.

```
Decision Tree 3 (More Complex) Performance:
Accuracy:  0.7817
Precision: 0.7859
Recall:    0.5799
F1 Score:  0.6673
ROC AUC:   0.8364
MSE:       0.2183
Avg Gini:  0.2661
```



This decision tree got the best accuracy out of the three, with an accuracy of 78.2%. It also got the best recall, MSE and ROC-AUC score. The simple decision tree can be dropped out of the competition as only its precision is better than its counterparts. Let's compare the best balanced and complex trees to figure out the best model. The better score is marked with a *.

Balanced Decision Tree	**Complex Decision Tree**
Accuracy: 0.7757	**Accuracy**: 0.7817*
Precision: 0.7861*	**Precision**: 0.7859
Recall: 0.5578	**Recall**: 0.5799*
F1 Score: 0.6526	**F1 Score**: 0.6673*
ROC–AUC: 0.8169	**ROC–AUC**: 0.8364*
MSE: 0.2243*	**MSE**: 0.2183
Avg Gini: 0.2319	**Avg Gini**: 0.2661*

As we can see, the more complex model is the better performing of the two, however there is a risk for overfitting, Overfitting happens when the model fits the training data too well and fails to generalize unseen data. As the deeper tree contains more terminals and nodes, it makes it challenging to extract insights and increases the risk of overfitting some datapoints or rare customer groups. That is why we will use the balanced model in our feature analysis.

Baseline Random Forest

Next, we will utilize the random forest models, which combines the results of multiple decision trees. The baseline model will not use any optimizers.

```
Baseline Random Forest Performance:
Accuracy:  0.8077
Precision: 0.7970
Recall:    0.6584
F1 Score:  0.7211
ROC AUC:   0.8733
MSE:       0.1923
Avg Gini:  0.1830
```

All of the measures already are better than any of the single decision trees.

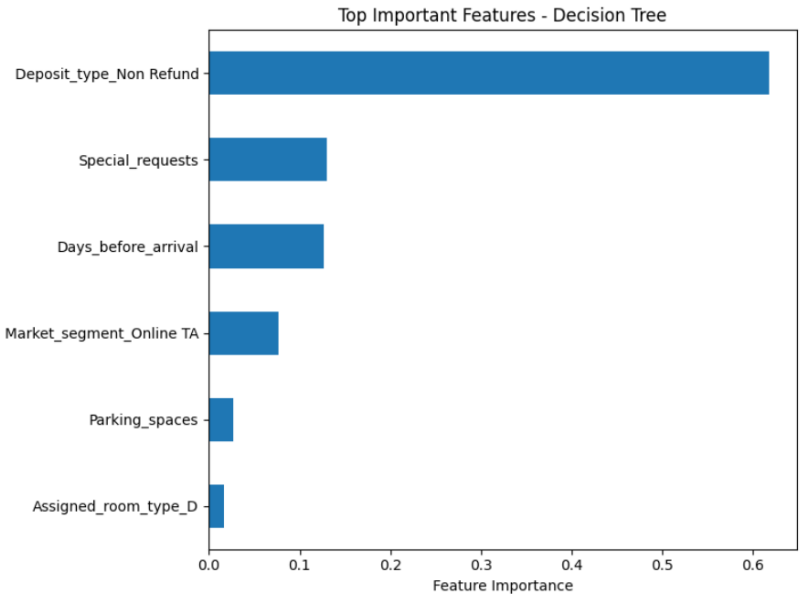
Optimized Random Forest

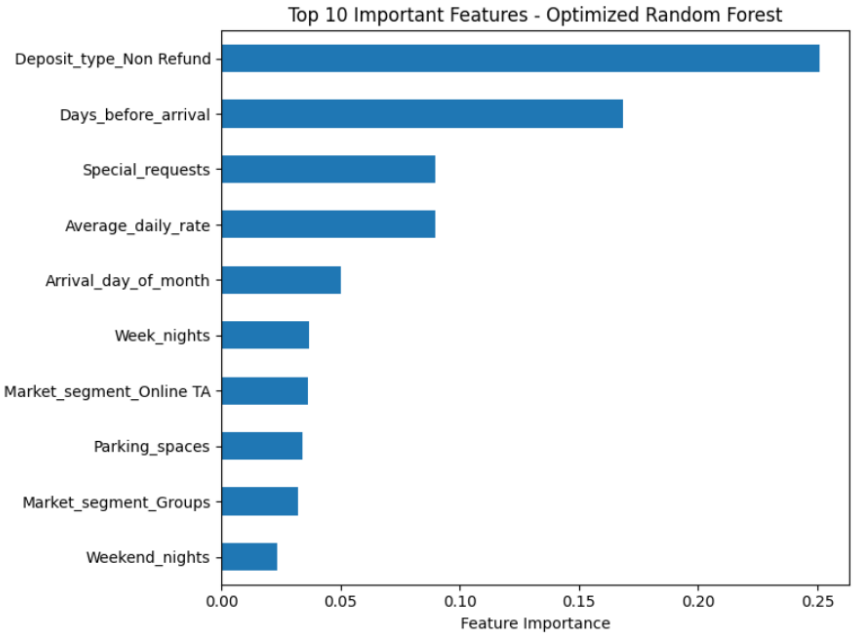
Now we will optimize the random forest with estimators, maximum depth and minimum samples.

```
 Optimized Random Forest Performance:
Accuracy:  0.8057
Precision: 0.8198
Recall:    0.6222
F1 Score:  0.7075
ROC AUC:   0.8715
MSE:       0.1943
Avg Gini:  0.2059
```

The results are very similar. The baseline model has a notable increase in Recall and F1-score, but the optimized version has a notable increase in Gini, and precision. It is a tie between the two, but we will use the optimized model in our feature analysis.

Feature Importance for the Best Decision Tree and Random Forest Model





The insights we have gotten from the tree-based models show that deposit types and booking behavior variables like market segment, days before arrival and special requests are the biggest predictors of cancellations. It goes hand in hand with the regression model and the descriptive analysis.

Neural Networks

To further explore non-linear relationships and interactions. We will use neural networks that can capture highly complex and dimensional patterns that algorithms like logistic regression and decision trees cannot.

We will start by finding the learning rate that gives the best accuracy:

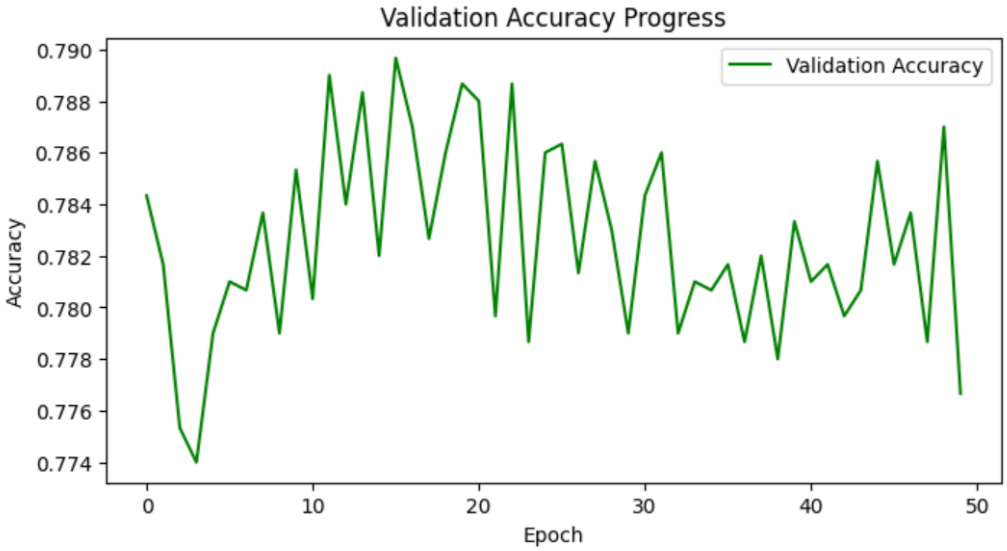
```
Learning Rate: 0.1 → Accuracy: 0.7780
Learning Rate: 0.01 → Accuracy: 0.7887
Learning Rate: 0.001 → Accuracy: 0.7780
Learning Rate: 0.0001 → Accuracy: 0.7813
```

As we can see the learning rate of 0.01 has a slightly higher accuracy than the rest and will therefore use it in our neural networks.

Baseline neural network

The baseline model will only have 1 layer, which is a basic structure.

The training loss decreases, while validation loss increases at around 10-15 epochs, which is classic overfitting.



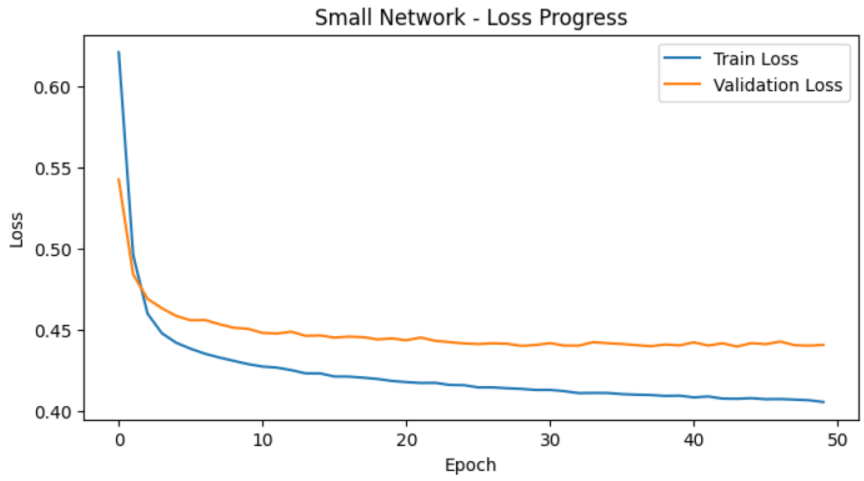
The validation accuracy fluctuates around 0.788 with no upward trend. Learning has stabilized and no more epochs won't help.

```
Neural Network Performance
Accuracy:  0.7767
Precision: 0.7327
Recall:    0.6434
F1 Score:  0.6852
ROC AUC:   0.8450
MSE:       0.2233
```

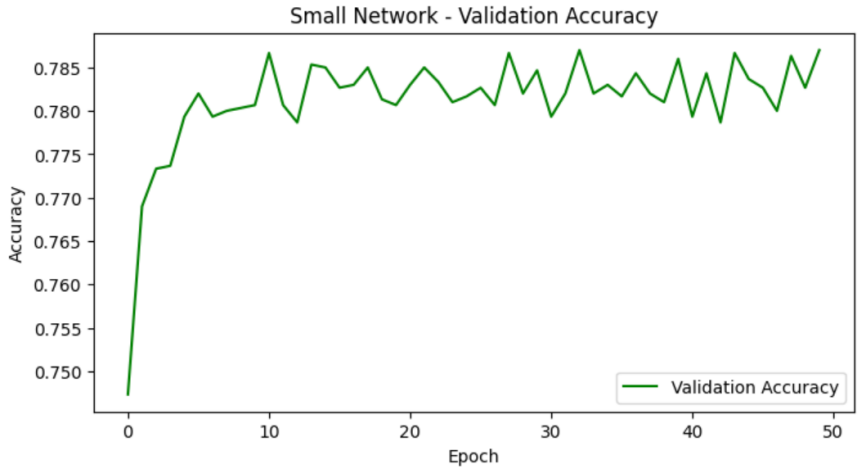
The performance is solid for a basic neural network.

Small neural network

The smaller neural network has 2 hidden layers with 16 neurons.



The training and validation loss both decreased fast in the first 10 epochs. It shows that the model learned efficiently without a lot of overfitting.



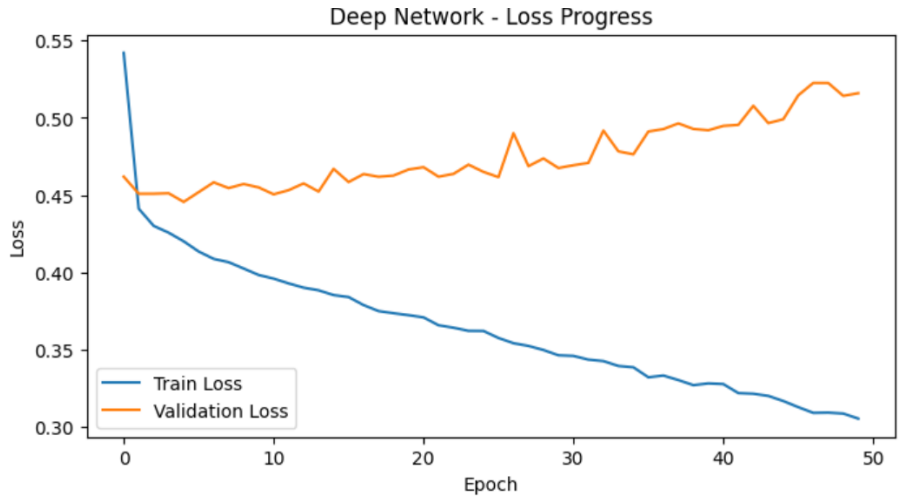
Validation accuracy increases quite sharply, which shows the models consistent predictive capabilities.

```
 Small Neural Network Performance:
Accuracy:  0.7870
Precision: 0.7813
Recall:    0.6055
F1 Score:  0.6822
ROC AUC:   0.8485
MSE:       0.2130
```

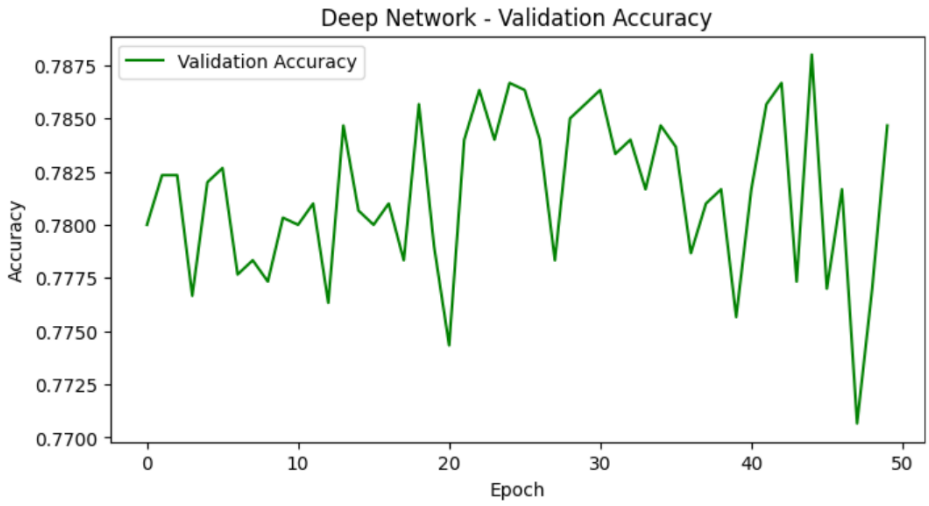
The small neural network performs slightly better than the baseline model in many of the measures.

Deep neural network

The deep neural network has 2 hidden layers with 32 and 16 neurons respectively.



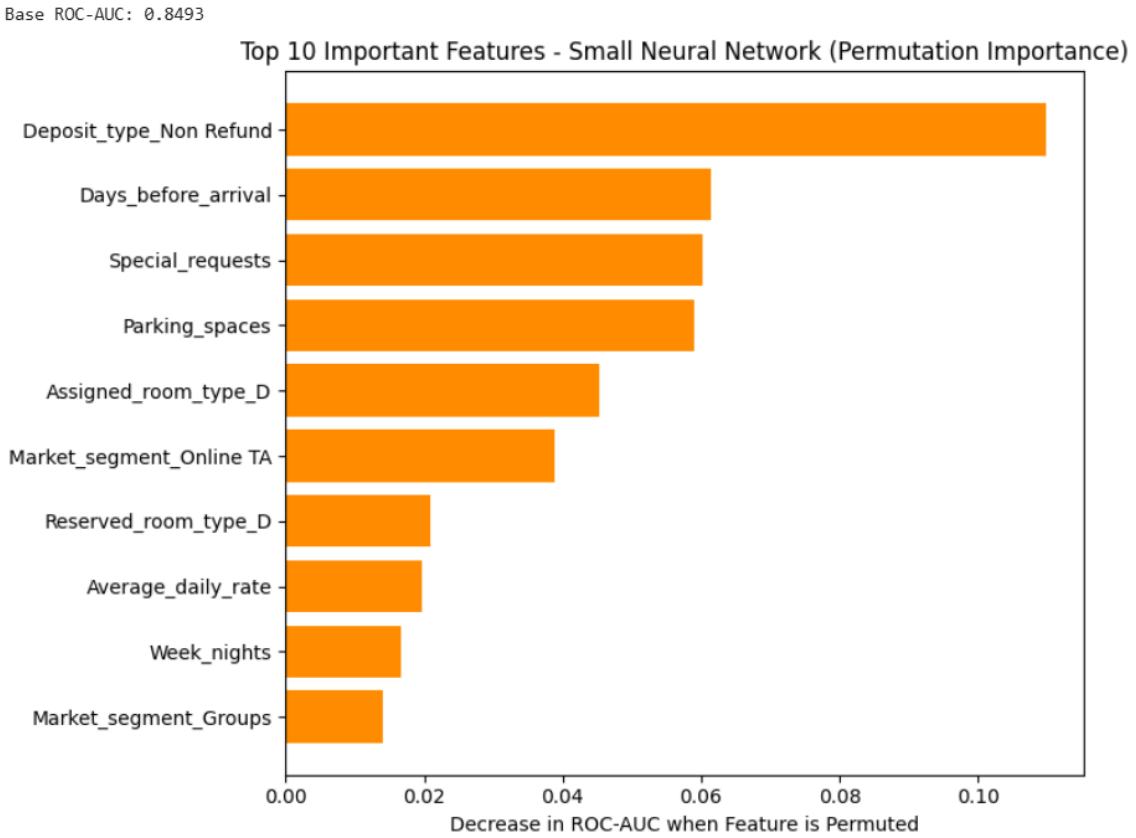
The training loss shows a stable and continuous decrease which indicates the model keeps while going through epochs. However, the validation loss starts to diverge after about 10 epochs, which can be a sign of overfitting.



The validation accuracy fluctuates at around 0.777 – 0.787. It remains stable but shows no improvement.

```
 Deep Neural Network Performance:
Accuracy:  0.7847
Precision: 0.7423
Recall:    0.6584
F1 Score:  0.6978
ROC AUC:   0.8493
MSE:       0.2153
```

The deep version gets the best ROC-AUC and F1-score of the three models, which means it is the most balanced and effective at finding cancellations vs non-cancellations. Its higher recall means it captures more true cancellations, which is positive for the business case. However, the smaller network generalizes a bit better as it has a lower validation loss and MSE, which makes the model more interpretable and stable. The deep network also showed small signs of overfitting. Therefore, the small neural network is the best of the bunch. The baseline model was solid but does not idenitfy deep relationships effectively.



As we can see, the most important features include all the same features as the previous models. Non-refundable deposit types, days before arrival, online travel agencies and special requests dominate.

Evaluation

In the evaluation stage, the aim is to interpret the results of the predictive models and explain the factors that influence booking cancellations the most.

The models confirmed that booking lead time, average daily rate, deposit type, market segment and customer engagement are the most important predictors when it comes to cancellations. These findings confirm the authors' Liu et al. (2025) claims that lead time and deposit policies are big indicators of possible cancellations, because early bookings and refundable deposits add to the uncertainty. Similarly, Ampountolas (2024) wrote that booking channels affect likelihoods of cancellations, with among the top doers being online platforms.

Explainable AI in the form of permutation importance proved that people who book far in advance and customers who have refundable or non-deposit are indeed more likely to cancel. On the contrary people who are old customers or people who make special requests are less likely to cancel. These are findings that align with Liu et al. (2025) claim that customer engagement and loyalty are important factors. Furthermore, a moderate positive correlation was found between the average daily rate and cancellation, which implies that higher prices might discourage customer who are more price-sensitive, also going in line with Liu et al. (2025) research.

The insights from the predictive models can be applied to improve revenue management and customer retention strategies. For example, hotels could start applying stricter rules in their refund policies. Long lead-time or high-risk bookings should have strict or non-refundable policies to minimize the risk of cancellation. Regarding customer engagement, hotels must encourage customers to make special requests and join loyalty programs that can increase commitment. Most importantly, interactions with the customers are encouraged. Furthermore, price-sensitive customers could be campaigned with offers, while loyal customers can receive personalized incentives.

Overall, the results of the predictive models go well in line with the literature and give valuable insights. The results were consistent in all the models, and they add to the evidence that data-driven predictions of cancellations aid hotels in customer retention, pricing etc.

Model Building Part 2

Market Basket Analysis

With the market basket analysis, the goal is to identify frequent combinations of transaction patterns that occur when customers make their cancellations. This model does not predict cancellations, but rather reveals associative patterns.

The measures used include confidence, lift and support. Support count is the number of transactions in which the item is present. Confidence is the support of the combined itemset divided by the support of the individual items. Lift measures the strength of associations between to items.

Top Itemsets

The Apriori algorithm was used to identify frequent itemsets among cancellations.

```
Transactional shape: (5666, 53)
Frequent itemsets @2%: 2579  | @5%: 1148
```

	support	itemsets	length
317	0.745852	(Assigned_room_type=A, Reserved_room_type=A)	2
503	0.578009	(Meal=BB, Reserved_room_type=A)	2
1838	0.567949	(Meal=BB, Assigned_room_type=A, Reserved_room_...	3
314	0.567949	(Meal=BB, Assigned_room_type=A)	2
523	0.557889	(Weekend_nights=Low, Reserved_room_type=A)	2
509	0.556830	(Meal=BB, Weekend_nights=Low)	2
1855	0.549594	(Weekend_nights=Low, Assigned_room_type=A, Res...	3
321	0.549594	(Assigned_room_type=A, Weekend_nights=Low)	2
450	0.502471	(Meal=BB, Deposit_type=No Deposit)	2
521	0.499824	(Week_nights=Low, Reserved_room_type=A)	2
1853	0.490646	(Week_nights=Low, Assigned_room_type=A, Reserv...	3
319	0.490646	(Week_nights=Low, Assigned_room_type=A)	2
507	0.472467	(Meal=BB, Week_nights=Low)	2
449	0.469291	(Market_segment=Online TA, Deposit_type=No Dep...	2
534	0.468408	(Week_nights=Low, Weekend_nights=Low)	2

With 2% support 2579 combinations of attributes were detected and with a 5% support 1148. Many of the cancellations include assigned room type A and Meal type bed & breakfast, which shows that typically lower priced combinations dominate cancelations because of their popularity. To find distinctive cancellation patterns, we are going to look at associations with high lift values.

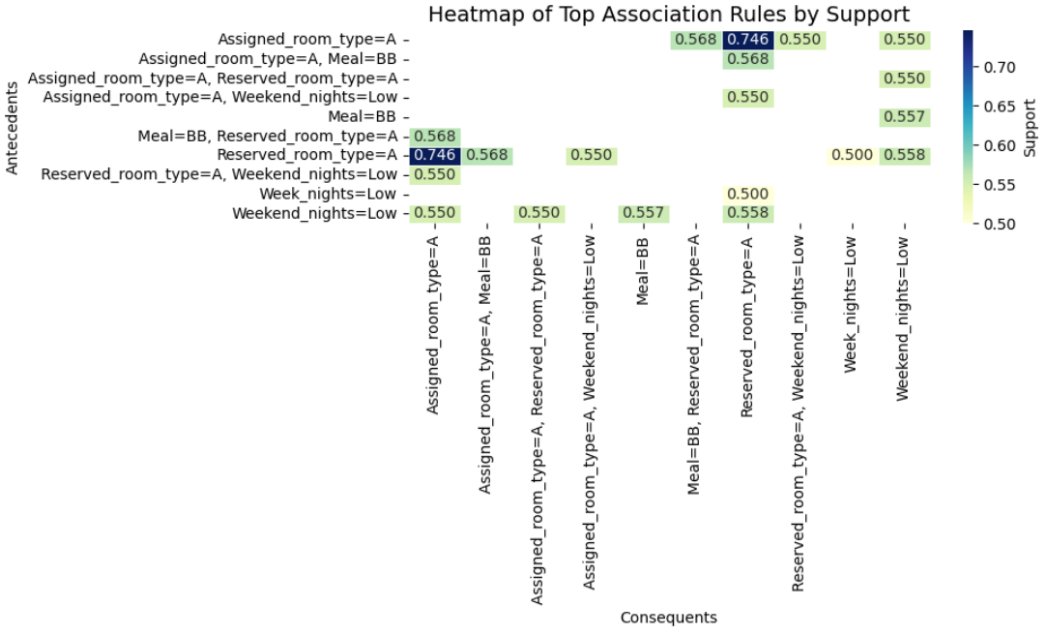
Rules Based On 2% Support Itemsets

	antecedents_str	consequents_str	support	confidence	lift
5963	Assigned_room_type=E, Weekend_nights=Med	Reserved_room_type=E	0.021708	0.976190	20.715713
5907	Assigned_room_type=E, Average_daily_rate=High	Reserved_room_type=E	0.027886	0.963415	20.444597
5947	Assigned_room_type=E, Market_segment=Online TA	Reserved_room_type=E	0.033180	0.959184	20.354812
5950	Reserved_room_type=E	Assigned_room_type=E, Market_segment=Online TA	0.033180	0.704120	20.354812
5964	Reserved_room_type=E, Weekend_nights=Med	Assigned_room_type=E	0.021708	1.000000	20.308244
5908	Average_daily_rate=High, Reserved_room_type=E	Assigned_room_type=E	0.027886	0.993711	20.180519
5948	Market_segment=Online TA, Reserved_room_type=E	Assigned_room_type=E	0.033180	0.989474	20.094473
5949	Assigned_room_type=E	Market_segment=Online TA, Reserved_room_type=E	0.033180	0.673835	20.094473
347	Reserved_room_type=E	Assigned_room_type=E	0.046594	0.988764	20.080061
346	Assigned_room_type=E	Reserved_room_type=E	0.046594	0.946237	20.080061
5924	Deposit_type=No Deposit, Reserved_room_type=E	Assigned_room_type=E	0.045358	0.988462	20.073918
5925	Assigned_room_type=E	Deposit_type=No Deposit, Reserved_room_type=E	0.045358	0.921147	20.073918
5926	Reserved_room_type=E	Assigned_room_type=E, Deposit_type=No Deposit	0.045358	0.962547	20.050699
5923	Assigned_room_type=E, Deposit_type=No Deposit	Reserved_room_type=E	0.045358	0.944853	20.050699
5953	Meal=BB, Reserved_room_type=E	Assigned_room_type=E	0.035475	0.985294	20.009593
5954	Assigned_room_type=E	Meal=BB, Reserved_room_type=E	0.035475	0.720430	20.009593
5959	Reserved_room_type=E, Weekend_nights=Low	Assigned_room_type=E	0.024885	0.979167	19.885155
5952	Assigned_room_type=E, Meal=BB	Reserved_room_type=E	0.035475	0.934884	19.839143
5955	Reserved_room_type=E	Assigned_room_type=E, Meal=BB	0.035475	0.752809	19.839143
5958	Assigned_room_type=E, Weekend_nights=Low	Reserved_room_type=E	0.024885	0.921569	19.556584
5659	Assigned_room_type=D, Days_before_arrival=High	Reserved_room_type=D	0.023297	1.000000	7.217834
5606	Assigned_room_type=D, Average_daily_rate=High	Reserved_room_type=D	0.087893	0.996000	7.188963
5607	Reserved_room_type=D	Assigned_room_type=D, Average_daily_rate=High	0.087893	0.634395	7.188963
5720	Assigned_room_type=D, Days_before_arrival=Med	Reserved_room_type=D	0.058419	0.991018	7.153004
5794	Assigned_room_type=D, Market_segment=Online TA	Reserved_room_type=D	0.115602	0.989426	7.141513

Room type E showed strong self-association, meaning customers who book the room also always get assigned the room (lift 20%). This highlights predictability and consistency. For example, if a customer books the room type E through an online platform and has high rates, it is a strong predictor on whether the booking will get cancelled or not. High rates and online platforms also co-occur a lot in cancellations, which can possibly be explained by price sensitivity and customers who are dependent on third parties.

Heatmap of Top Association Rules by Support

This heatmap was created to visualize which combinations were the most frequent overall



From the heatmap we can tell that bed and breakfast, room type A and dominate cancellations.

Key insights and managerial implications

Online travel agencies and no deposits are frequently associated with a higher cancellation frequency. Standard room types dominate cancellations, but this is because of their popularity. In addition, higher rate bookings made through online platforms may be flexible reservations that are often cancelled later. Finally, no-refund deposits were rare among cancellations, which confirms that stricter refund policies discourage cancellations.

As a hotel owner, encouraging non-refundable or partially refundable options for online travel agencies to reduce the risk of cancellations. Monitoring price-sensitive segments like high-rate bookings through online travel agencies might reduce overbooking. Finally, analyzing cancellation timing with room types A/D/E to optimize reallocations and offering discounts/incentives for direct bookings through own websites might help lower cancellation risks.

Model Building Part 3

Text Analytics

The purpose of text analytics is to identify what aspects customers value the most at their stay at the hotel by analyzing textual reviews left by them. This is done by natural language processing techniques.

Overview

	Review	Rating	cleaned_review
0	great location wife recently spent second week...	4	great location wife recently spent second week...
1	loved breakers stayed breakers recommendation ...	5	loved breakers stayed breakers recommendation ...
2	nice hotel girlfriend booked hotel students, d...	4	nice hotel girlfriend booked hotel students, n...
3	terrible customer service second stay resort, ...	1	terrible customer service second stay resort, ...
4	old style waikiki, reading reviews tripadvisor...	5	old style waikiki, reading reviews tripadvisor...

First, we cleaned up the textual data by lowercasing everything, removing punctuation, numbers and special characters. Next, tokenization was performed by splitting text into words. Then stopwords like “the” and “is” were removed. Finally, we reduced the words to their baseform. For example, “stayed” got reduced to “stay”.

Wordcloud of Most Popular Words

The analysis of the most popular words revealed that the customers mostly mentioned:

Positive terms: “clean”, “friendly”, “comfortable”, “excellent”, “location”.

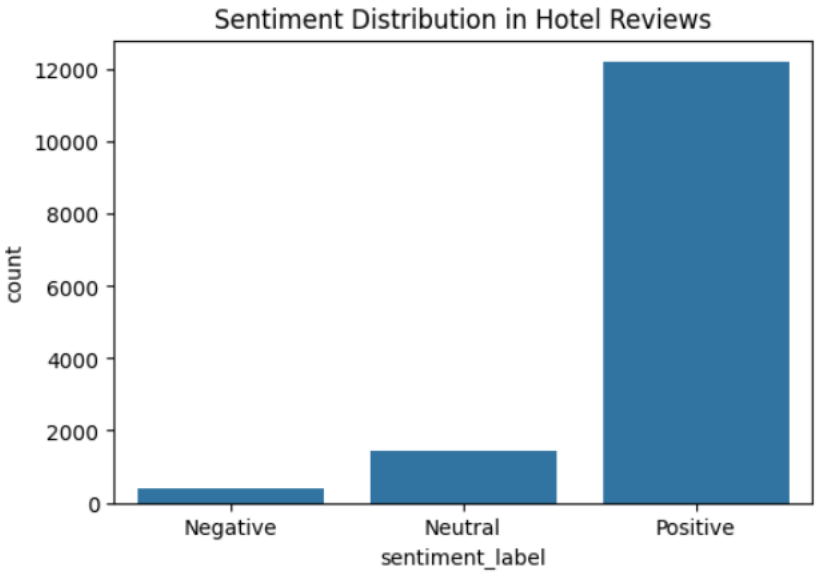
Service-related words: “staff”, “breakfast”, “room”

Negative terms: Not frequent but “noisy”, “small”, “expensive”

Customers were most drawn by cleanliness, staff’s behavior and comfort in their reviews. The negative reviews emphasized more on noise, room sizes and price related factors. The wordcloud proves that positive reviews dominate, which indicates an overall satisfactory impression from the customers.

Sentiment Distribution in Hotel Reviews

Each review was assigned a polarity sentiment with a lexicon-based kind of approach. Positive sentiments polarity > 0, negative sentiments polarity > 0 and a neutral sentiments polarity = 0.




```
Correlation between Rating and Sentiment: 0.63
```

As we can see, the majority of the sentiments were positive, which aligns well with the high ratings. A smaller portion were negative sentiment reviews, which corresponded well with the lower ratings. The sentiment model makes sense and proves that it effectively captured the consistency between words and ratings. This also proves that customer's mostly have had favorable experiences with the hotel.

An example of a review with a positive label is shown here:

```
Review: great budget hotel stayed days hotel second weekend jazzfest pleasantly surprised, room spacious exceptior
lly clean, desk cleaning people nice willing help, roaches room reported previous post, ihop convenient, budget hc
el place exceeded expections,
Sentiment score: 0.2785714285714286
Label: Positive
```

Topic Modeling

With the use of LDA (Latent Dirilecht Allocation), words could be divided into different topics. Each topic represents a cluster of words that occur together frequently, which highlights different factors about customer experience.

```
Topic 1:
['people', 'day', 'room', 'time', 'pool', 'good', 'great', 'food', 'resort', 'beach']

Topic 2:
['like', 'bed', 'desk', 'rooms', 'service', 'staff', 'night', 'stay', 'hotel', 'room']

Topic 3:
['stayed', 'clean', 'breakfast', 'stay', 'good', 'staff', 'location', 'room', 'great', 'hotel']
```

Topic 1 includes leisure and amenities experiences. Customers emphasize facilities like the pool, beach and resort. Words like "good", "great" and "food" indicate positive experiences. This topic quite likely relates to holiday travelers that appreciate comfort and activities, rather than people on business trips.

Topic 2 includes service and room quality experiences. Things like staff interactions, comfort and overall room experience are a large part of this topic. Neutral words like "night" and "stay" mean general descriptions of their stay.

Topic 3 includes tidiness, breakfast and location experiences. "good", "great" and "staff" implies positive experiences with the staff, breakfast and location.

All in all, the topics did not introduce negative themes, suggesting that most of the feedback was positive.

In conclusion, a lot of managerial implications can be derived from the sentiment analysis and topic modeling, although they might seem obvious. High standards of cleanliness and customer service is an important factor for customers. Investing in breakfast and leisure facilities is smart, as they get a lot of positive attention. Marketing campaigns could emphasize happy and friendly staff, tidy rooms and location convenience. Monitoring feedback on these factors is also important.

Deployment

Deploying the results of the predictive and sentiment analysis into a hotel environment requires combining technical elements with business objectives. The enabled models should be integrated into the hotels existing booking management and CRM systems, with the objective of improving pricing, operations and communication.

The predictive models could flag incoming high-risk bookings based on the market segment, lead time and deposit type, which enables the hotel to send out reminder notifications of upcoming trips or offer personal incentives for committed customers. In similar fashion, sentiment analysis models could nonstop process customer reviews from sites like Tripadvisor or Booking.com, as Ameur et al. (2023) describe booking sites like Tripadvisor as continuously providing data analysts with consumer opinions and behavioral data.

The main business value of the deployment would be reduced cancellation rates, more accurate demand forecasting and improved customer retention. Liu et al. (2025) research demonstrated that unoccupied inventory and overbooking precision improved, and therefore raised profitability, when predictive cancellation forecasting was applied into revenue optimization frameworks. Furthermore, Ameur et al (2025) show that using sentiment analysis on customer reviews improves e-reputation and monitoring of customer satisfaction, which supports marketing and reputation strategies.

The models should be connected to the hotels management systems, which enables that predictions generate automatically when a new booking arrives. It would also be important to have automated data updates and regular retraining of the models so that the predictions are accurate even when customer behavior or market circumstances change. Moreover, a simple dashboard interface makes the insights interpretable for the hotel management and non-technical stakeholders. The dashboard would visualize risk segments, sentiment trends and recommended steps with simplified metrics. For example “Cancellation risk = high/medium/low”. Consequently, managers and non-technical stakeholders can apply data driven insights in their daily decision making.

Finally, limitations and risks for deployment include model bias and data drifting. Market trends and customer behavior change over time, which mean the models would need continuous retraining and validation. Moreover, the models must comply with the EU’s GDPR laws. Personal booking and review data must be processed securely and need to be anonymized when necessary. The customers also need to be made explicitly aware that their data is processed securely to keep their trust. In addition, highly complex models may not generalize well and might end up overfitting

References

Ameur, A., Hamdi, S., & Ben Yahia, S. (2023). Sentiment analysis for hotel reviews: A systematic literature review. *ACM Computing Surveys*, *56*(2), 1-38. doi: https://doi.org/10.1145/3605152

Ampountolas, A. (2025). Predicting hotel booking cancellations: a comprehensive machine learning approach. *Journal of Revenue and Pricing Management*, 1-12. doi: https://doi.org/10.1057/s41272-025-00532-x

Liu, Z., De Bock, K. W., & Zhang, L. (2025). Explainable profit-driven hotel booking cancellation prediction based on heterogeneous stacking-based ensemble classification. *European Journal of Operational Research*, *321*(1), 284-301. doi: https://doi.org/10.1016/j.ejor.2024.08.026

Shmueli, G., Bruce, P. C., Gedeck, P., Yahav, I., & Patel, N. R. (2023). *Machine Learning for Business Analytics. Concepts, Techniques and Applications in R*. John Wiley & Sons, Inc.