**Machine Learning Group Project**

**GROUP 3**

**201621456 Sim YoungHoon**

**201721926 Kim HyunJin**

**201821441 Baek SiYeon**

**201823869 Cho SeongWoo**

**202021540 Park YouSun**

**202132632 Jan-Paul Davidsen**

**Table of Contents**

1. **Introduction**

* Reason for data selection(Related to our situation)
* Business questions(About ML)

1. **Dataset**

* Data Description
* Data Utilization Plan

1. **Analyses & Findings**

* Basic visual data (EDA)
* ML Analysis for Business Questions

1. **Implications**

* Insights(About our problems, The situation we need to develop)
* Solution(How to solve our business problems, How to develop our business-attract more customers, prevent cancellation)

1. **Conclusion**

* Limitations(Parts that need development)
* Conclusion(The conclusion of whether to proceed with the business)

1. **References**
2. **Introduction**

We are the management of a hotel chain. As the number of travelers is expected to increase due to the easing of COVID-19 restrictions, we are trying to expand our business globally. Among the many continents, we especially want to expand our business into southern European city hotels. In this region climate and culture are Mediterranean, which could attract a lot of customers. Moreover, as one of the continents where the COVID-19 blockade has been lifted and with an abundance of sights market entry should be more easily. We decided to start our expansion in Portugal.

Before entering this country, we would like to address business issues that are expected. There are many cancellations at Portuguese hotels right now, the reason being that it is made very easy by Websites and mobile reservations. As a result, people often cancel their reservations and use other hotels even if they pay a small fee when they find a cheaper or nicer room. Some hotels don´t even charge for cancellation. This situation costs these companies a lot of money. Therefore, we have decided to make a prediction by analyzing our dataset to estimate the cancellation of our hotel. By making this analysis, we seek to increase the efficiency of operations. We plan to ensure that room cancellations do not occur as much as possible through strategic operations such as intentionally overbooking or providing additional benefits to prevent cancellations.

Moreover, we need to think of additional ideas to increase profitability. We are trying to increase sales through strategies such as having a high number of booked rooms and decreasing average accommodation costs. We hope that data analytics will help us come up with solutions to accomplish this.

1. **Dataset**

We looked for the dataset with our previously mentioned goals in mind, especially considering the local situation in Portugal. The data we found displays data on reservation and cancellation of city and resort hotels from July 2015 to August 2017. Data for customer identification has been deleted, because of privacy reasons, but still, the set contains a total of 119,390 observations with 31 variables. As a case study, we will analyze Portugal's hotel industry data to find solutions to our business questions.

The full dataset contains booking information for City and Resort hotels, including booking time, length of stay, number of adults, children and/or babies, and available parking spaces. In total there are 31 variables, but we reduced the data to the following:

- hotel: kind of hotel (City Hotel or Resort Hotel)

- is\_canceled: Boolean value stored as int64, indicates the hotel is canceled or not

- lead\_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

- arrival\_date: Information of the arrival date

- adults/ children/ babies: Number of adults/ children/ babies

- stays\_in\_week/weekend\_nights: Number of week/weekend nights the guest stayed or booked to stay at the hotel

- meal: Type of meal booked. Categories: Undefined/SC – no meal package / BB – Bed & Breakfast / HB – Half board (breakfast and one other meal – usually dinner) / FB – Full board (breakfast, lunch and dinner)

- country: Country of origin

- deposit\_type: Indication on if the customer made a deposit to guarantee the booking

- market\_segment: Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"

- distribution\_channel: Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"

- is\_repeated\_guest: Value indicating if the booking name was from a repeated guest (1) or not (0)

- preivious\_cancellations/not\_canceled: Number of previous bookings that were canceled/not

- room\_type: Code of room type reserved

- customer\_type: Type of booking, assuming one of four categories (Transient, Transient - Party, Contract, Group)

- adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

- required\_car\_parking\_spaces: Number of car parking spaces required by the customer

- booking\_changes: Number of booking changes

- total\_of \_speical\_requests: Number of special requests made by the customer

1. **Analyses & Findings**

***Finding 1: is\_canceled***

## **1.1 Cancellation by Special Requests**

Chart, bar chart

Description automatically generated

If there are more than one special request, you can see that the cancellation rate drops sharply.

## **1.2 Cancellation rate by time of booking**

## **Chart, scatter chart Description automatically generated**

<European countries>

There is a big difference in the correlation between European countries and non-European countries. For customers in European countries, the cancellation rate is higher as they made reservations long ago. This can be attributed to the greater the difference between the date of stay and the time of reservation, the greater the possibility that the plan will change in the middle. Traveling in Europe is not a burden, so you don't have to book a hotel too long in advance.

Chart, scatter chart

Description automatically generated

<European countries>

## In the case of non-Europeans, there was little correlation between the cancellation rate according to the reservation period. Here the burden and time of travel to Europe are considerable compared to Europeans, and overseas travel is planned for a relatively long time.

## **1.3 Cancellation rate based on average price**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

<European countries> <non-European countries>

Contrary to lead time, differences between non-European and European countries in the correlation between the average price and the cancellation ratio are much bigger. The reason is that the overall cost of going to Europe is much higher, so the sensitivity of the price is inevitably high, and therefore, the rate of cancellation is higher when adjusting the budget after the initial plan or when finding a hotel for the same price that is better.

***Finding 2 ADR(Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights)***

## **2.1. Guest Per Price by Room Type**

Chart

Description automatically generated

As a result of examining whether there is a difference in price by room type, it was confirmed that the ADR of rooms A to F was high, and outliers exist.

## **2.2 ADR by Arrival Month and Hotel Type**

Chart, line chart

Description automatically generated

## In the case of resort hotels, ADR was the highest in July and August, the peak travel season, and ADR in city hotels was similarly high in May and August. Customers are likely willing to use the hotel even if they must pay more in these time periods.

## **2.3 ADR by market segment and distribution\_channel**

Chart, bar chart

Description automatically generated

The ADR of customers booking through GDS (Travel agencies and services that offer everything from rental cars to plane tickets to hotel rooms) in Online TA and Offline TA/TO was significantly higher. Contrary to that, the price was relatively low when making a reservation through Direct and TA/TO, indicating that the price was more likely to be low when choosing a hotel directly or through an agency.

***Finding 3 Length of stay***

**3.1 Length of Stay by hotel type**

Chart, histogram

Description automatically generated

In the plot, many observations are distributed between the 1st and 4th days of accommodation in City Hotels, and resort hotels are like City Hotels overall, but the tail distribution is higher than that of City Hotels. Resort hotels differ significantly from city hotels in the sense that the number of observations that stayed for seven days is significantly higher, which seems to be visited by many customers who have secured a long vacation period because resort hotels are relatively less accessible than city hotels.

### **3.2 Distribution of the number of domestic/foreigners according to the number of days of stay (0-10)**

Chart, bar chart

Description automatically generated

It examines the distribution of observations according to how many people stayed at the resort hotel and the number of days they stayed, respectively. This graph shows the percentage change of Portuguese and foreigners depending on the number of days the guest stayed from 0 to 10. For the 0th, 1st and 2nd day, the number of Portuguese (domestic) who stayed was high, but after the 3rd day the number of foreigners staying showed a significant percentage change.

***Finding 4 ML Analysis for Business Questions***

### **4.1 Cancellation(Classification)**

***- Preprocessing***

1. Delete rows with all 0 ‘adults’ & ‘children’ & ‘babies’ (There are 180 rows of all zero)
2. Only use rows with 3 or less ‘adults’
3. 'Convert ‘Booking\_changes’, ‘special\_requests’, ‘required\_car\_parking\_spaces’ into binary variables because the number of reservations and cancellation rates are lower than 1
4. 0 <= ADR <= 200
5. Convert Reservation date to datetime variable
6. Convert ‘country’ to PRT/Foreign
7. Normalize numerical variables

***- Result***

Diagram

Description automatically generated with medium confidenceTable

Description automatically generated

Since the data is imbalanced data, performance evaluation was conducted using AUC and f1-score, and as a result, Cat Boost Classifier model with a value close to 1 had the best performance.

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

It can be seen that the month of travel affects cancellation the most. In particular, it can be seen that there are more cancellations during the off-peak season than in July, August, and December, which are the peak seasons.

### **4.2 Revisit(Classification)**

***- Preprocessing is identical to 4.1..***

***- Result***

Diagram

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

Since the data is imbalanced data, performance evaluation was conducted using AUC and f1-score, and as a result, Random Forest Classifier model with a value close to 1 had the best performance.

Chart

Description automatically generated

Shape, square

Description automatically generated

It can be seen that the lead time affects revisit the most. It was found that the lead time was shorter for customers who visited again. If the room with a short reservation period is promoted mainly for repeat customers, it will be able to be sold at a higher probability.

### **4.3 ADR\_1(Multiple Linear Regression)**

***- Preprocessing***

1. Is identical to 4.1.
2. Categorize number of guests by 1/2 = (couple), 3,4,5
3. Create ‘adr\_1’ by dividing adr by total number of guests
4. In the case of the cancelled reservation team, it is considered to be meaningless information that does not help increase sales, so excludes it

***- What is ADR\_1***

### ADR is Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights. **ADR\_1** is a derivative variable calculating the average daily usage amount per person by dividing it by the total number of people staying. A variable that has a significant positive correlation with the **ADR\_1** will be identified and a marketing target of high sales per day will be selected.

***- Performance & Verifications and Remodeling methodology***

1) Adjusted R-squared : 0.548 / MAE : 12.8 / MSE: 317.4 / RMSE: 17.8

2) Figured out Multicollinearity

2.1) Repeat variable selection

-> Filtered and independent variables (based on T-stat and VIF)

2.2) Applied standard Scaler for scaling independent variables

-> solved (VIFs of all of independent variables are less than 5)

***- Result***

Graphical user interface, text

Description automatically generatedChart, scatter chart

Description automatically generatedChart, histogram

Description automatically generated

### The coefficients of ADR\_1 and major independent variables were checked in ascending and descending order, respectively. When booking with GDS channels, there is a significant positive correlation with ADR\_1, and the number of guests & adults in a team has a negative correlation with ADR\_1 values. Considering that certain room\_types are located at the top of the major variables that show a positive correlation with ADR\_1, it is thought that the proportion of rate changes due to the choice of room type among the amounts mainly included in ADR\_1.

### **4.4 Additional Association Rules (Apriori)**

***- Preprocessing***

1. Is identical to 4.3.
2. 0 <= ADR\_1 <= 400
3. Replace Missing values with appropriate values, such as 0
4. The Undefined variable in 'meal' is the MEAL of SC type and is replaced
5. In the case of a team with one 'is\_canceled' (canceled team), it is excluded because it is deemed meaningless to interpret the relevant rules related to actual accommodation
6. Categorize 'addr\_1' as low, mid, high

***- Result***

It is intended to additionally conduct related analysis using Apriori to check case-by-case cases and indicators between variables and use them as additional promotion marketing for each case.

### **- Couple (2-person guest team)**

# Data Listing criteria - conditional, dependent clause, support, confidence, lift

Graphical user interface, text

Description automatically generated

Couples often did not eat hotel meals. It is necessary to figure out why couples do not include meals and think of strategies to increase sales through promotions that include meals.

### **- Transient (Individual customers who are not connected to a group or company)**

Text

Description automatically generated

Transient customers show high ADR. Therefore, transient customers not only increase the room rotation rate because they stay short, but also have a high ADR, so it can be seen as a target with a very high potential value. It will be possible to increase sales by focusing on that type of customer.

### **- Corporate (Customers staying at hotels at reduced corporate rates)**

Graphical user interface, text

Description automatically generated

Customers who stayed at the hotel through corporate reservations are single-person customers with high reliability.

Graphical user interface, text

Description automatically generated

In the case of corporate reservations, there is a high probability of choosing the BB type meal type.

Graphical user interface, text

Description automatically generated

Single-person customers are more likely to have low ADR.

Graphical user interface, text

Description automatically generated

For single-person customers, it belongs to the section with high ADR\_1.

Business customers through the corporate channel are often single-person customers. Since single-person customers naturally have lower accommodation costs than multi-person customers, ADR (total daily accommodation expenditure) is small, but ADR\_1 (average daily accommodation expenditure per a person) is high, indicating that they are high value-added customers.

### **- Customers with a child**

Text

Description automatically generated

Customers with children are more likely to have high ADR.

Text

Description automatically generated

Customers with children are more likely to be Transient type customers.

In the case of guests with children, it was found that ADR was high, and transient-type customers were also high. It was found that the customer with children was a high value-added customer who paid a high ADR with a fast rotation rate.

### **- Customers by country**

Text

Description automatically generated

Customers from the UK are more likely to choose resort hotels.

Text

Description automatically generated

Couple (2-person) customers from the UK are more likely to choose resort hotels.

Text

Description automatically generated

Customers from France are more likely to choose a city hotel.

Text

Description automatically generated

For customers from Germany, there is a high probability that a special request existed.

Text

Description automatically generated

If the ADR belonged to the high section, there is a high probability that the City Hotel was selected.

Text

Description automatically generated

If customers choose a resort hotel, customers are more likely to belong to a low ADR section.

When looking at customers by country, the UK tends to prefer resort hotels, and major countries from France and Germany tend to prefer city hotels. Considering that if the ADR belongs to the high section, the city hotel is selected, and if the resort hotel is selected, the ADR belongs to the low section, if the promotion is conducted mainly on the city hotel for France and Germany, it will attract many customers with high added value. In addition, more effective marketing results will be achieved by more active promotion of resort hotels for UK customers. In addition, it is possible to plan a strategy to provide delicate services to French customers with many special requests.

1. **Implications**

Through the above data analysis, we were able to obtain meaningful insights and solutions for business problems and success.

***Insight 1 Prevent cancellation***

For customers with special requests, the cancellation rate is low. Therefore, if sophisticated customized services are provided for each customer, the cancellation rate will be lowered. It provides more sophisticated services for each customer, such as providing special room services, providing customized breakfast menus for children, and extending checkout time.

In addition, the longer the lead time, the higher the cancellation rate for customers in Europe, and the higher the accommodation cost, the higher the cancellation rate for customers in non-European countries. Accordingly, different strategies for each country will be implemented to lower the cancellation rate. It offers to shorten the period of time that reservations can be made to customers in Europe and discount accommodation costs to customers in non-European countries.

***Insight 2 Selling Remaining Rooms***

Through the Random Forest Classifier model, it was found that the lead time was shorter for revisiting customers. Revisiting customers have high satisfaction with our hotel and have enough information, so if we sell the hotel with a short lead time to the revisiting customers, we will be able to reduce the number of vacancies. Revisiting customers have less resistance to our hotel and make reservations even if the lead time is short, so they will be able to sell more rooms if they sell short lead time rooms at a discounted price to revisiting customers rather than new customers.

***Insight 3 Increase sales***

Through EDA and multi-linear regression analysis, customers who book through the GDS channel and customers who book hotels during the peak season of July and August have high ADRs. In collaboration with the GDS channel, it provides package rooms that include travel services such as airline tickets and rental cars and sells rooms with premium services during the peak season to achieve higher sales.

In addition, we were able to select the main target for marketing through the association rule. Couples (2 people) often did not eat at hotels. Sales can be increased by analyzing the reasons why couples do not eat and planning strategies to induce them to eat at hotels. Transient customers and customers with children have high ADR and Corporate customers are high value-added customers with high ADR\_1. If you provide promotions for the customer, you will get a higher sales growth rate than other targets. For example, selling short-term luxury package room services to Transient customers and partnering with multiple companies to attract corporate customers and increase child comfort and amusement facilities to increase child-accompanied customer satisfaction.

Finally, we can get more effective marketing results by offering a variety of promotions by country. Reflecting the different types of hotels preferred by each country, France and Germany mainly promote city hotels and provide discounts, while the UK mainly promotes resort hotels and offers discounts.

1. **Conclusion**

* ***Limitation***

Since this model is based on data from a particular hotel, there is no guarantee that it will be applied equally to other hotels. In addition, predicting the future with previous data may not be accurate as consumer needs are becoming more diverse and trends are changing rapidly. Therefore, efforts are needed to continuously develop services and promotions by collecting customer data in real time.

* ***Conclusion***

So far, models have been created and interpreted based on Portuguese hotel data for hotel business expansion and success.

Based on the EDA, cancellation, and revisit model, it was possible to see how behavior appeared differently for each customer characteristic. Through this, each customer will be able to operate the room efficiently by creating positive results such as reducing reservation cancellation and reducing vacancies by providing different operating methods and promotions.

Plus, based on the ADR\_1 prediction model, sales can be predicted in advance by grasping the potential value per person according to the customer's pre-booking information. Based on the prediction results of this model, it will be able to help objectify expected sales and sales goals in the mid- to long-term and adjust the prices of rooms and meals in the current period to pave the way for additional analysis strategies such as analyzing changes in ADR\_1.

In addition, various marketing strategies and significant goals have been identified through a number of significant rules, and relatively detailed insights into the various feature values favored by the target have been obtained. Identifying customers with high potential according to related rules and establishing efficient marketing strategies will greatly help achieve significant performance improvement compared to other hotels that do not use AI.

1. **References**
2. <https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy>
3. <https://hongl.tistory.com/136>
4. <https://romg2.github.io/dss/01_%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%82%AC%EC%9D%B4%EC%96%B8%EC%8A%A4-%EC%8A%A4%EC%BF%A8-4.9-%EB%AA%A8%ED%98%95%EC%9D%98-%EC%A7%84%EB%8B%A8%EA%B3%BC-%EC%88%98%EC%A0%95/>+
5. <https://datascienceschool.net/03%20machine%20learning/04.03%20%EC%8A%A4%EC%BC%80%EC%9D%BC%EB%A7%81.html>
6. <https://mingtory.tistory.com/m/140>
7. <https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?sort=votes>
8. lecture note 10 ~13, 16~17