**IE 359 Machine Learning Using Statistical Analysis System**

**Spring 2021 Final Project**

**Predicting Hotel-Booking Cancellation**

**Arda Can Baysar**

**S006320**

Contents

[**1 Introduction** 2](#_Toc74005535)

[1.1 Project Description 2](#_Toc74005536)

[1.2 Data Set 2](#_Toc74005537)

[**2 Data Exploration and Visualization** 5](#_Toc74005538)

[2.1 Overview 5](#_Toc74005539)

[2.2 Top Customer Nationalities 6](#_Toc74005540)

[2.3 Hotel traffic on Monthly basis 7](#_Toc74005541)

[2.4 Distribution of Cancelled Reservation 7](#_Toc74005542)

[2.5 Change in Average Daily Spent Based on Months 7](#_Toc74005543)

[2.6 Relationship Between Cancelling and Special Requests 8](#_Toc74005544)

[2.7 Relationship Between Lead Time and Duration of Stay 8](#_Toc74005545)

[**3 Data Preprocessing** 9](#_Toc74005546)

[3.1 Deleting of missing (Null) Data 9](#_Toc74005547)

[3.2 Deleting of Variables 9](#_Toc74005548)

[**4 Exploring Important Variables** 10](#_Toc74005549)

[4.1 Target Variable 10](#_Toc74005550)

[4.2 The Most Important Numeric Predictive Variables 10](#_Toc74005551)

[4.3 Variable Importance 10](#_Toc74005552)

[**5 Machine Learning Algorithms** 10](#_Toc74005553)

[5.1 Random Forest Baseline Model 11](#_Toc74005554)

[5.2 Random Forest Tuned Model 12](#_Toc74005555)

[5.3 Gradient Boosting Machine Baseline Model 12](#_Toc74005556)

[5.4 Gradient Boosting Machine Tuned Model 14](#_Toc74005557)

[**6 Conclusion** 15](#_Toc74005558)

[6.1 Conclusion about variables 15](#_Toc74005559)

[**7 Case Study** 15](#_Toc74005560)

[**8 References** 16](#_Toc74005561)

# **1 Introduction**

## 1.1 Project Description

The high cancellation rate for online hotel bookings causes pain for many businesses and a need to take precaution. As a result, forecasting cancellations provides a surplus value for hotels, and hotels may take steps to prevent these cancellations.

In this project, I will try to investigate the data set and show how machine learning algorithms may be used to predict future cancelled bookings. In other words, I will try to predict is\_canceled variable (whether a booking is canceled or not), which is target variable of the dataset.

Project is being described as follows:

Booking management is one of the critical components of revenue management of the airline and hospitality industries. Depending on the booked capacity of a business, one can overbook the inventory to maximize the profit. Therefore, predicting the outcome of current bookings (whether a booking is canceled or not) plays a significant role in a business's profitability. For example, an airline company can book more tickets than the capacity of the flight. However, when the booking office acts too greedily, the airline may have to compensate bookers that cannot take their seats due to excessive overbooking. Such a booking strategy causes loss of profit and loss of customer satisfaction. Therefore, predicting the customers' behaviors is quite significant in booking-based industries. The hospitality industry is no different in terms of the importance of customer behavior prediction.

## 1.2 Data Set

This dataset comprises booking information for a city hotel and a resort hotel, including date the booking was made, duration of stay, number of people, children, and/or infants, and number of available parking spots, among other things.

According to Kaggle, the data is originally from the article [Hotel Booking Demand Datasets](https://www.sciencedirect.com/science/article/pii/S2352340918315191), written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.[2]

In total, there are 32 columns and around 119.4K rows in the data.

Description of the Dataset:

**hotel**: Hotel (H1 = Resort Hotel or H2 = City Hotel).

**is\_canceled**: Value indicating if the booking was canceled (1) or not (0).

**lead\_time**: Number of days that elapsed between the entering date of the booking into the PMS (Property Management System) and the arrival date.

**arrival\_date\_year**: Year of arrival date.

**arrival\_date\_month**: Month of arrival date.

**arrival\_date\_week\_number**: Week number of year for arrival date.

**arrival\_date\_day\_of\_month**: Day of arrival date.

**stays\_in\_weekend\_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.

**stays\_in\_week\_nights**: Number of weeknights (Monday to Friday) the guest stayed or

booked to stay at the hotel.

**adults**: Number of adults.

**children**: Number of children.

**babies**: Number of babies.

**meal**: Type of meal booked. Categories are presented in standard hospitality meal packages: 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. Categories are represented in the ISO 3155–3:2013 format.

**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". The term "GDS" means "Global Distribution System".

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

**previous\_cancellations**: Number of previous bookings that were canceled by the customer prior to the current booking.

**previous\_bookings\_not\_canceled**: Number of previous bookings not canceled by the customer prior to the current booking.

**reserved\_room\_type**: Code of room type reserved. Code is presented instead of designation for anonymity reasons.

**assigned\_room\_type**: 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.

**booking\_changes**: 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.

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

This variable can assume three categories:

No Deposit – no deposit was made

NonRefund – a deposit was made in the value of the total stay cost

Refundable – a deposit was made with a value under the total cost of the stay.

**agent**: ID of the travel agency that made the booking.

**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.

**days\_in\_waiting\_list**: Number of days the booking was on the waiting list before it was confirmed to the customer.

**customer\_type**: Type of booking, assuming one of four categories:

Contract - when the booking has an allotment or other type of contract associated with it

Group – when the booking is associated to a group

Transient – when the booking is not part of a group or contract and is not associated to another transient booking

Transient-party – when the booking is transient but is associated to at least another transient booking.

**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.

**total\_of\_special\_requests**: Number of special requests made by the customer (e.g., twin bed or high floor).

**reservation\_status**: Reservation's last status, assuming one of three categories:

Canceled– booking was canceled by the customer

Check-Out – customer has checked in but already departed

No-Show – the customer did not check-in and did inform the hotel of the reason why

# **2 Data Exploration and Visualization**

## 2.1 Overview

In total, there are 119 thousand observations and 32 variables and one of them is the target variable(is\_cancelled). Let's take a quick look at the data.

Descriptive Statistics of the Dataset Table

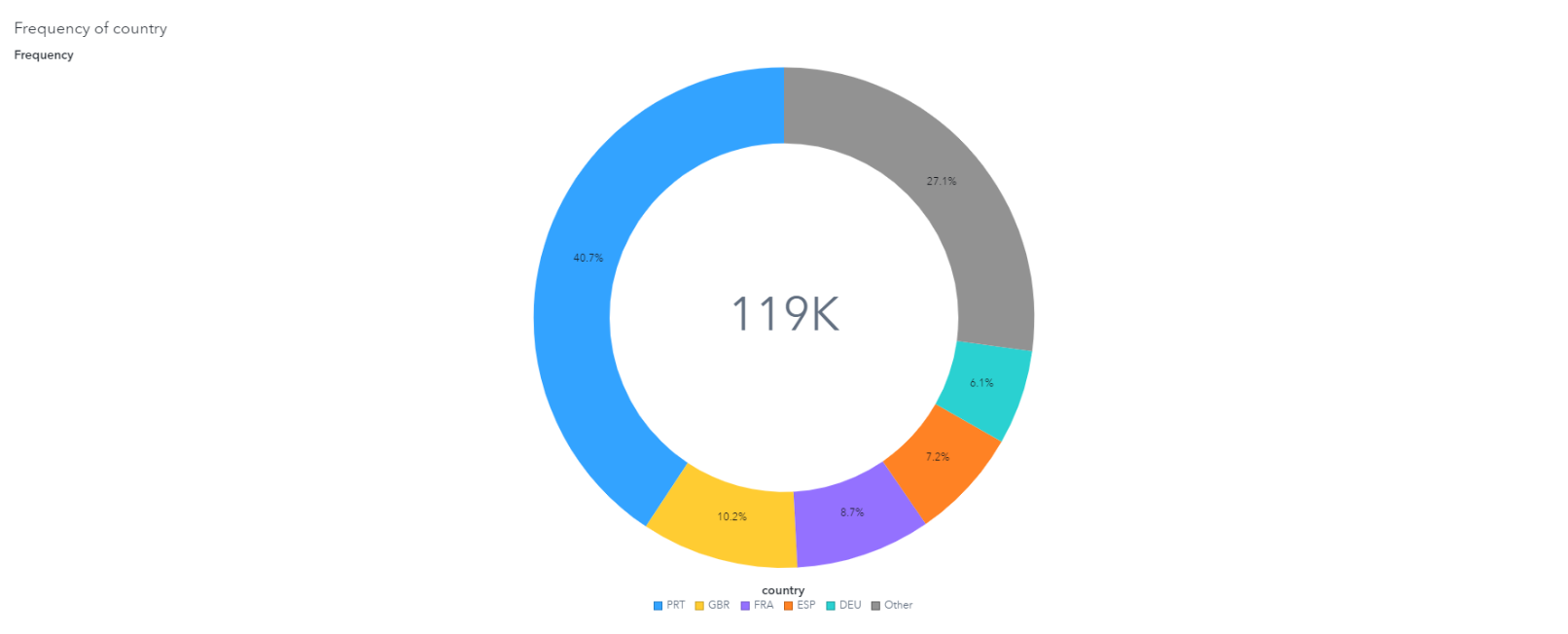
Description automatically generated

Table

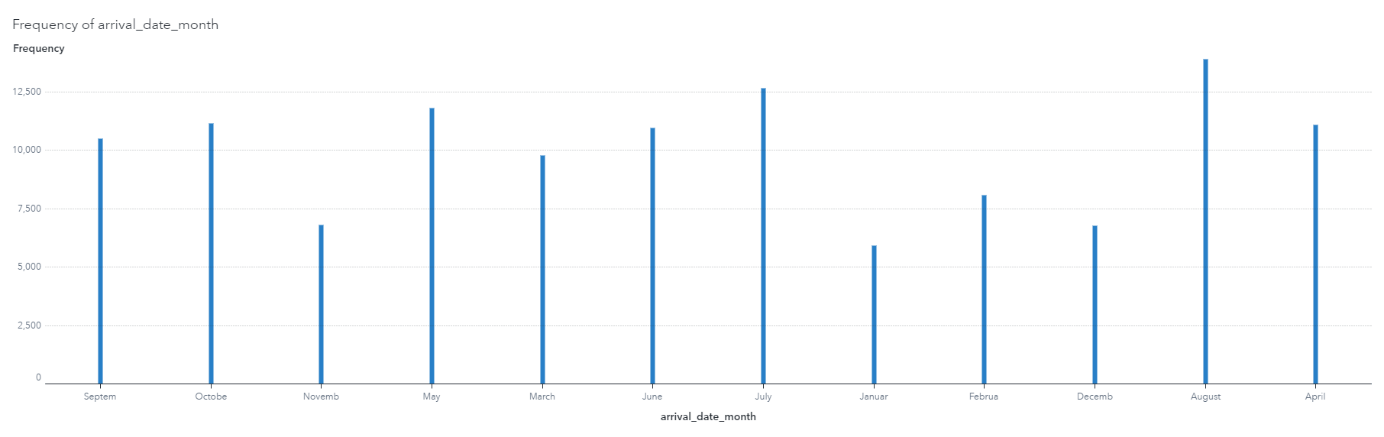
Description automatically generated

## 2.2 Top Customer Nationalities

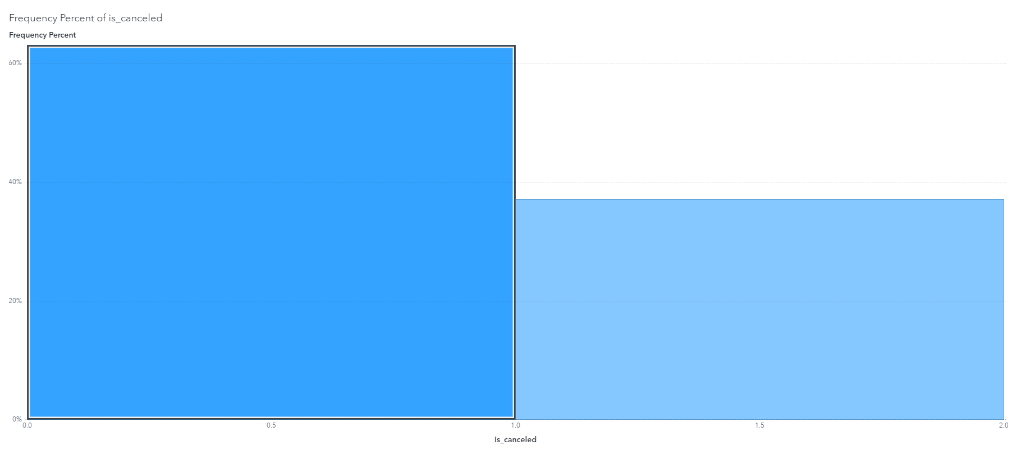
Top nationalities in the dataset are Portugal (40.7%), Great Britain (10.2%), France (8.7%), Spain (7.2%), Germany (6.1%) and other countries (27.1%). One explanation for this, I think, the data probably was collected in some hotels which are in Portugal. That is why most of the customers are Portuguese in the dataset.



## 2.3 Hotel traffic on Monthly basis

We can determine which month the customer plans to check into the hotel by using the variable arrival date month. According to the monthly basis data, the majority of hotel booking requests occurred in the months of July and August, followed by May and June. One explanation for this, I think, is the influence of the weather, as these are the months of nice weather.

## 2.4 Distribution of Cancelled Reservation

Not cancelled= %62.96, cancelled = %37.04

## 2.5 Change in Average Daily Spent Based on Months

Average of Adr which represents the amount paid by the visitor on average for each night over the duration of their stay is relatively high in the summer season. It is probably because of, the hotels that the data was taken from, are in a summer seasonal area. That is why they might charge more in the summer months.

Chart

Description automatically generated

## 2.6 Relationship Between Cancelling and Special Requests

Chart

Description automatically generatedWhen Special request of customer increases, they are less likely to cancel the reservation.

## 2.7 Relationship Between Lead Time and Duration of Stay

Chart, histogram

Description automatically generatedAccording to the heat map, the longer the lead time, the more likely the duration of stay will be shorter.

# **3 Data Preprocessing**

This dataset consists of numeric, categorical variables. In total, there are 32 variables (19 numeric variables and 13 categorical variables) and one of them is the target variable (is\_canceled).

## 3.1 Deleting of missing (Null) Data

Graphical user interface, application

Description automatically generatedFirst, I simply removed missing observations from columns containing a tiny number of NULL values. I discovered just 4 missing values in the children column and removed these four observations.

## 3.2 Deleting of Variables

If two variables are highly correlated, I need to drop one those variables. Because I want the dataset to be independent. For Example, “reservation\_status” is highly correlated to “is\_canceled”. When reservation\_status= Cancelled or No Show, it means that is\_cancelled=1. If reservation\_status= Checked out, it means that is\_cancelled=0 automatically. They are 100% dependent each other. This variable may leak some information about the target variable. That is why I am dropping the “reservation\_status” variable. In Sas Viya it does not show the correlation matrix between categorical variables. That is why I checked the correlation of these variables by using another software.

# **4 Exploring Important Variables**

## 4.1 Target Variable

“Is\_canceled” is a target variable whose value indicates if the booking was canceled (1) or not (0).

## 4.2 The Most Important Numeric Predictive Variables

First, I analyzed numerical variables. I wanted to see which numerical features are most significant to the target variable. To do that I used correlation matrix between numerical variables. From the matrix, we can see that “lead\_time”, “total\_of\_special\_requests”, “required\_car\_parking\_spaces”, previous\_cancellations and is\_repeated\_guest are the 5 most significant numerical features.

## 4.3 Variable Importance

Even though the correlations provide a great perspective of the most significant numeric variables, we still have categorical data; thus, I wanted to take an overview of the most important variables, including all the categorical variables. According to Güneş et al. (2020) “Variable importance tables indicate the statistical contribution of each feature to the underlying model. There are various ways of calculating model-agnostic feature importance. One method includes fitting a global surrogate decision tree model to the black-box model predictions and using the variable importance table that is produced by this simple decision tree model” [2]. In order to do that, I used random forest classification. According to random forest classification the six most important variables, including categorical, are “deposit\_type”, “lead\_time”, “total\_of\_special\_request”, “previous\_cancellations”, “market\_segment” and “required\_car\_parking\_spaces”. Table

Description automatically generated

# **5 Machine Learning Algorithms**

In this project, to predict the target variable Random Forest and Gradient Boosting models were used. As metrics of interest, ROC and F1 score were considered for comparing the model’s success. 70% of the data were used as training split, 30% of the date were used test split. In addition, binary classification cutoff was considered as 0.5.

## 5.1 Random Forest Baseline Model

The Forest node creates a model that contains a number of decision trees. For each tree in the forest, the training data is sampled without replacement from all available observations. In addition, the input variables that are considered for splitting a node are randomly selected from all available inputs, and the best available split is chosen.

All analyzes in this article were made in the SAS Viya software.

I selected all variables in the data set (without “reservation\_status”) to train random forest model.

Hyperparameters of Random Forest Baseline Model in SAS Viya software as below:

**Class target voting method**: Probability

**Class target criterion**: Information gain ratio

**Interval target criterion**: Variance

**Number of trees**:100

**Maximum number of branches**: 2

**Maximum depth**: 20

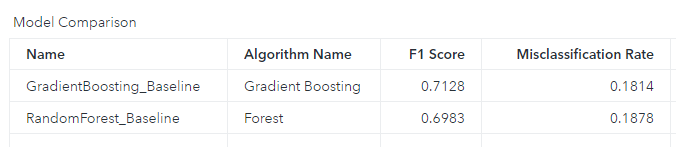
**Minimum leaf size**: 5

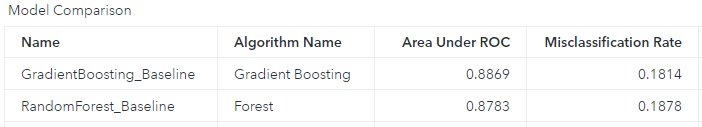
**Number of Bins**: 50

**Missing values**: Use in search

Results for Random Forest Baseline:

**F1**: 0.6983

**ROC**: 0.8783



## 5.2 Random Forest Tuned Model

After Hyperparameter tuning and many [sequential experimentation](https://tureng.com/en/turkish-english/sequential%20experimentation), I significantly improved Random Forest Baseline Model by tuning.

Hyperparameters of Random Forest Tuned Model in SAS Viya software as below:

**Class target voting method**: Probability

**Class target criterion**: Information gain ratio

**Interval target criterion**: Variance

**Number of trees**: 78

**Maximum number of branches**: 2

**Maximum depth**: 50

**Minimum leaf size**: 5

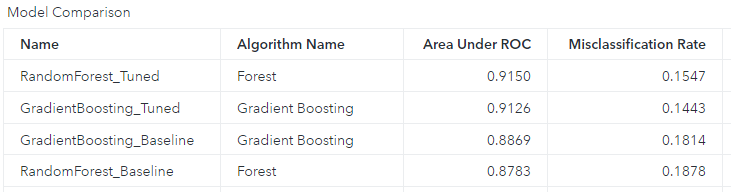
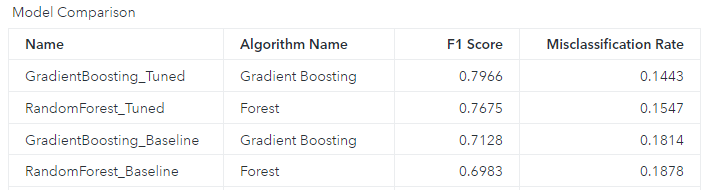
**Number of bins**: 73

**Missing values**: Use in search

Results for Random Forest Tuned Model:

**F1**: 0.7675

**ROC**: 0.9150



## 5.3 Gradient Boosting Machine Baseline Model

The Gradient Boosting creates a gradient boosting model. Starting with a baseline tree, the model hones its predictions by minimizing a specified loss function. The model then iterates, building subsequent trees via sampling without replacement, and thus minimizing the loss function of the previous tree.

I selected all variables in the date set (without “reservation\_status”) to train Gradient Boosting Machine model.

Hyperparameters of Gradient Boosting Baseline Model in SAS Viya software:

**Number of trees**:100

**Learning rate**: 0.1

**Subsample rate**: 0.5

**L1 regularization**: 0

**L2 regularization**: 1

**Interval target distribution**: Normal

**Maximum number of branches**: 2

**Maximum depth**: 4

**Number of bins**: 50

**Minimum leaf size**: 5

**Number of inputs per split**:

**Missing values**: Use in search

Results for Gradient Boosting Baseline Model:

**F1**:0.7128

**ROC**:0.8869

A picture containing graphical user interface

Description automatically generated

A picture containing application

Description automatically generated

## 5.4 Gradient Boosting Machine Tuned Model

After Hyperparameter tuning and many [sequential experimentation](https://tureng.com/en/turkish-english/sequential%20experimentation), I significantly improved Gradient Boosting Machine Baseline Model by tuning.

Hyperparameters of Gradient Boosting Tuned Model in SAS Viya software:

**Number of trees**:150

**Learning rate**: 0.5050

**Subsample rate**: 1

**L1 regularization**: 0

**L2 regularization**: 5

**Interval target distribution**: Normal

**Maximum number of branches**: 2

**Maximum depth**: 11

**Number of bins:** 60

**Minimum leaf size**: 11

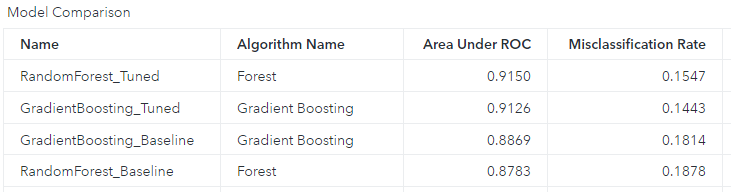
**Missing values**: Use in search

Results for Gradient Boosting Tuned Model:

**F1**: 0.7966

**ROC**: 0.9126

Graphical user interface

Description automatically generated with medium confidence

# **6 Conclusion**

All the tuned models had AUC score of over 90%, which can be considered excellent. This illustrates that combining datasets and machine learning techniques, such as the decision forest and GBM method, to develop booking cancellations prediction models is an effective strategy.

## 6.1 Conclusion about variables

A longer lead time means a greater chance of cancellation. This should make obvious sense, both because it gives the buyer more chances to cancel. When reserving two years in advance, rather than two days, a guest's plans are more likely to alter.

A greater Average Daily Rate, which represents the amount paid by the visitor on average for each night over the duration of their stay, indicates a higher chance of cancellation. This means that the more money a consumer spends on a hotel, the more likely they are to cancel.

The more specific demands a booking has, the less likely it is to be canceled. Customers are less likely to cancel if they have many specific demands. It is important to try to get as many requests as possible by salesperson of hotels.

# **7 Case Study**

According to this project, these findings show that it is feasible to detect bookings that are likely to be canceled. This allows hotel administrators to take preventative actions to avoid future cancellations. These machine learning algorithms can help hoteliers avoid revenue loss due to cancellations and avoid the hazards of overbooking. Hospitality industry may also use booking cancellation models to adopt less strict cancellation policies without creating uncertainty. Less strict cancellation rules result in more bookings, which might lead to increased revenues. Furthermore, these models enable hotel management to act on bookings that have been flagged as "possibly canceled," as well as create more precise demand estimates. On the other hand, there are several points that should not be accepted as foolproof. For example, it would not make logical sense to double-book any room that our algorithm predicts would be canceled.

In addition to these, Hotels can use this information to approach clients that the model predicts would cancel, giving them more time to resell the room. Alternatively, hotels may approach the client in such a manner that they feel unique, allowing you to preserve their reservation. Because in the project it was founded that the more special requests the less likely cancellations.

My suggestion to minimize room cancellations is 'decreasing room prices for the customers who was predicted as cancelled by the algorithm.' This indicates that if savings from fewer cancellations are passed on to customers, the hotel might see a good effect, resulting in fewer cancellations. Because in the project we proved that greater average amount paid by visitors for one night indicates a higher chance of cancellation.

We should be aware of that according to each hotel, different features may have varying weights and relevance. Each hotel is different and the circumstances that lead to cancellations at each hotel may differ from one another. As a result, one model will not fit all hotels, and each hotel will need its own model.

I think this model can be extremely valuable to hotel industry due to the reliable AUC and F1 score. The proper use of this approach, which includes utilizing it to forecast probable cancellations and take precautionary actions, will increase income while lowering costs.

# **8 References**

[1] https://www.kaggle.com/jessemostipak/hotel-booking-demand/metadata

[2]Güneş, F., Tharrington, R., Abbey, R., & Hunt, X. (2020). Paper SAS4502-2020. *How to Explain Your Black-Box Models in SAS® Viya®*.