FINAL PROJECT:
SUPERVISED MACHINE LEARNING: CLASSIFICATION

Antonio Jimenez

IBM ON COURSERA March, 2023

Table of Content

Introduction

Online reservations of hotel rooms have change customer's behavior. The Main objective of this project is to determine whether or not a customer is going to cancel his hotel reservation. The focus is on prediction capabilities – the model is going to be used by management to adjust the capacity of the hotel based on demand and adjusted for cancelations. This is going to help to allocate resources more efficiently and help to determine the number of employees to be hired.

Description of the Data

The data chosen was downloaded from Kaggle, it is a hotel reservation Dataset, uploaded by Ahsan Raza, with 36275 observations. The attributes of customer's reservations are the following (Ahsan):

Booking_ID: unique identifier of each booking

no_of_adults: Number of adults

no_of_children: Number of Children

no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

type_of_meal_plan: Type of meal plan booked by the customer:

required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.

lead_time: Number of days between the date of booking and the arrival date

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking_status: Flag indicating if the booking was canceled or not.

The Goal of the analysis is to get a model that can predict with accuracy whether or not a customer is going to cancel their reservation.

Exploratory Data Analysis

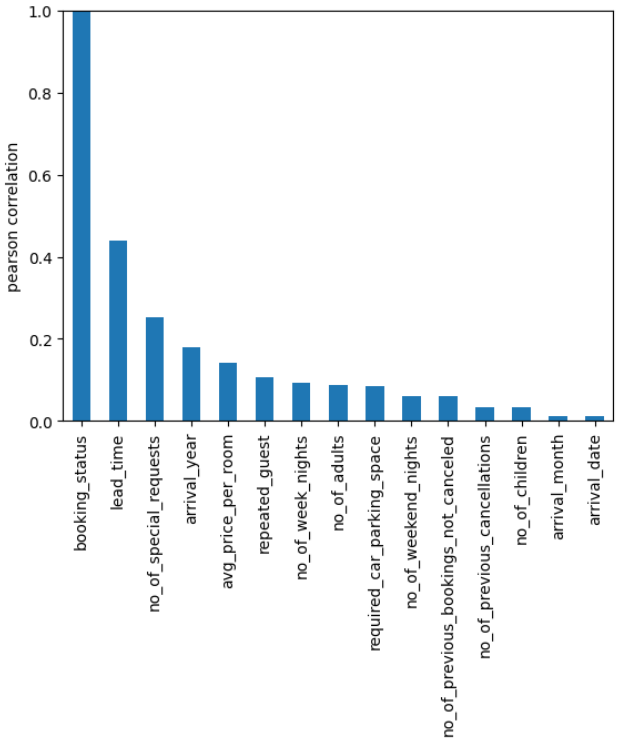
The first step was to load the csv file into a Pandas Dataframe. The next step was to understand the data types of each features and deal with missing values. Luckily, the there wasn't any missing value. Then, the unique values of categorical data were analyced. In the case of actually using them for the model, they must be one-hot encoded.

Also, the booking status, the variable we want to predict, has been encoded:

0: No cancelation

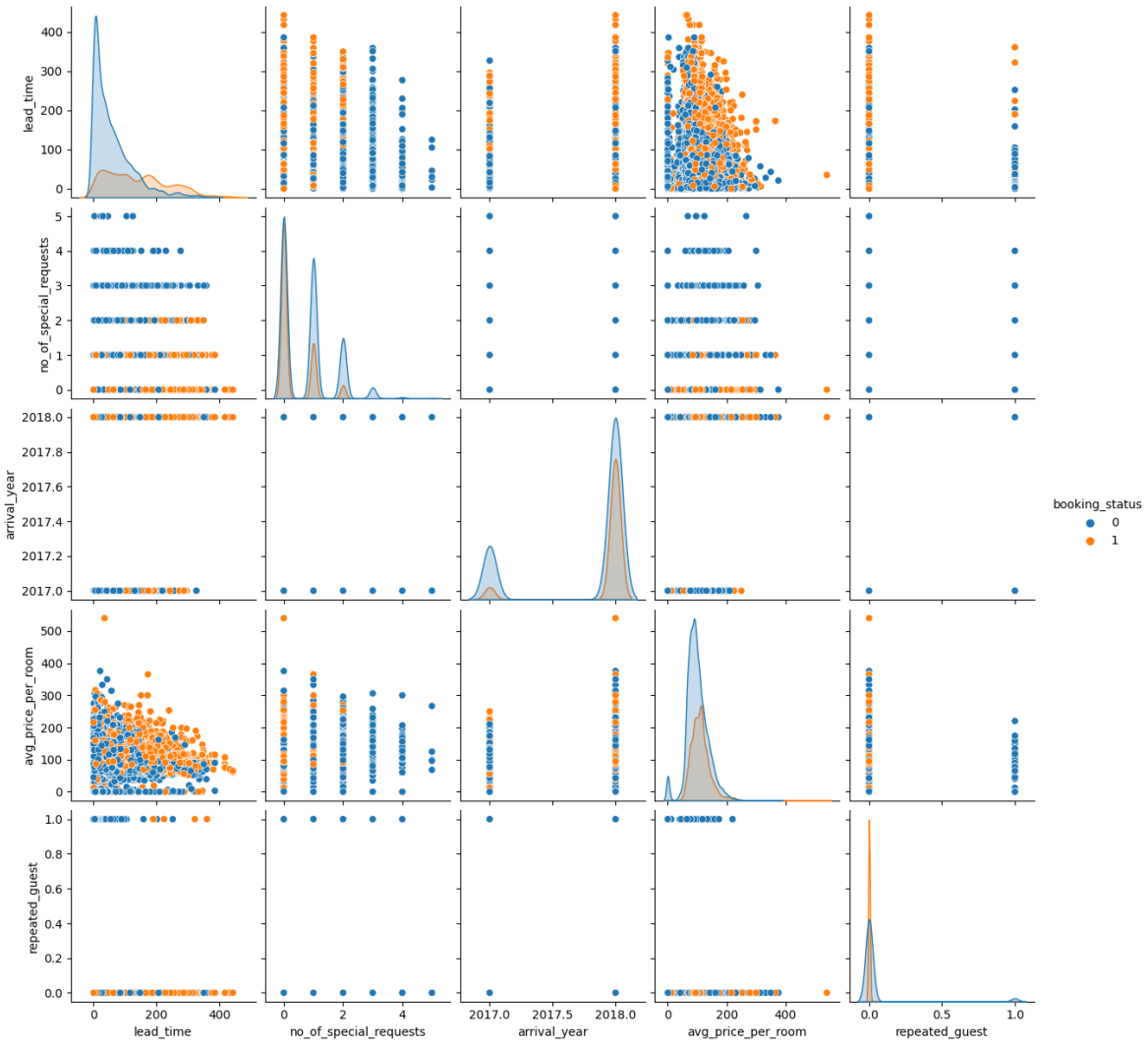
1: Cancelation

The next step was to understand the correlations between the features and the target variable. A bar chart was plotted for this purpose:

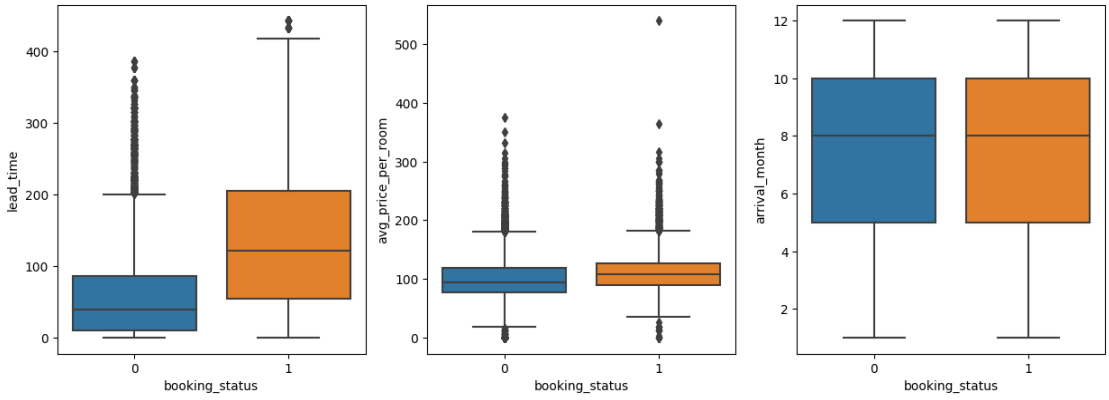


From this chart, the most important variables can be identified in order to use them in the model. The selected features were: lead_time, no_of_special_requests, avg_price_per_nights, repeated_guest, and arrival_month. It was decided to use this last variable, arrival_mont, instead of arrival_year because it was considered that the former is an indicator of the seasons. Also, the data was sampled in 2017 and 2018, so there weren't enough data for yearly projection.

Below is a visualization of the correlations between the chosen features:



Also, boxplot for the most correlated variables were plotted:

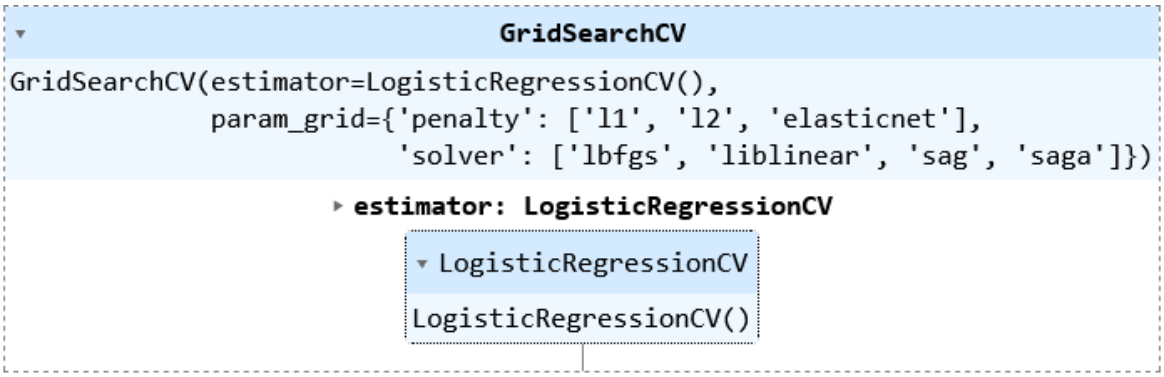


It can be seen that most of the reservations are scheduled between the months of May and October. In the northern hemisphere, this is the summer season.

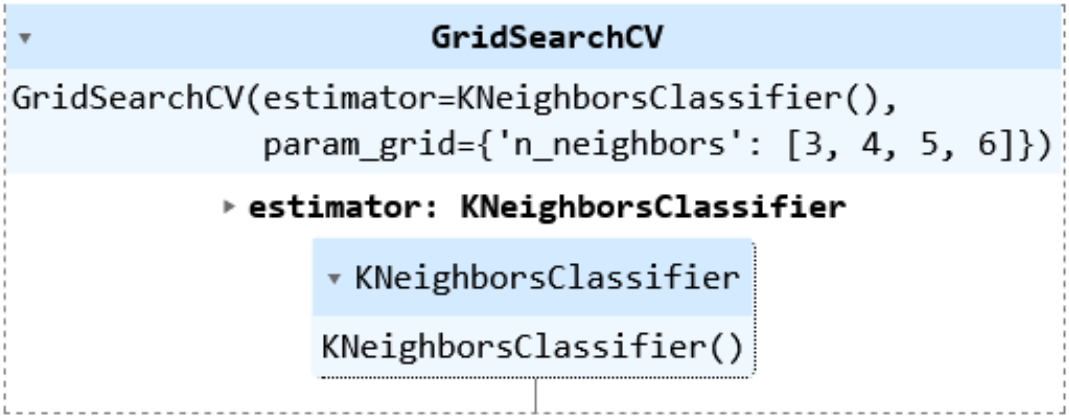
Modeling

Four different models were trained, using GridSearchCV to get the best parameters. These models were:

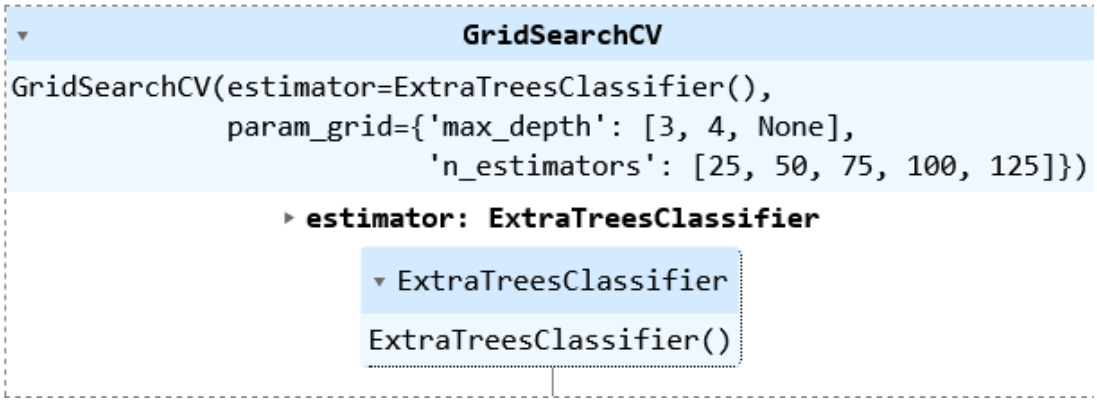
Logistic Regression,



K-Nearest Neighbor,



Extra Trees Classifier,



and Gradient Boosting,

```
GridSearchCV(estimator=GradientBoostingClassifier(),
             param_grid={'learning_rate': [0.1, 1],
                         'n_estimators': [500, 600, 700]})
```

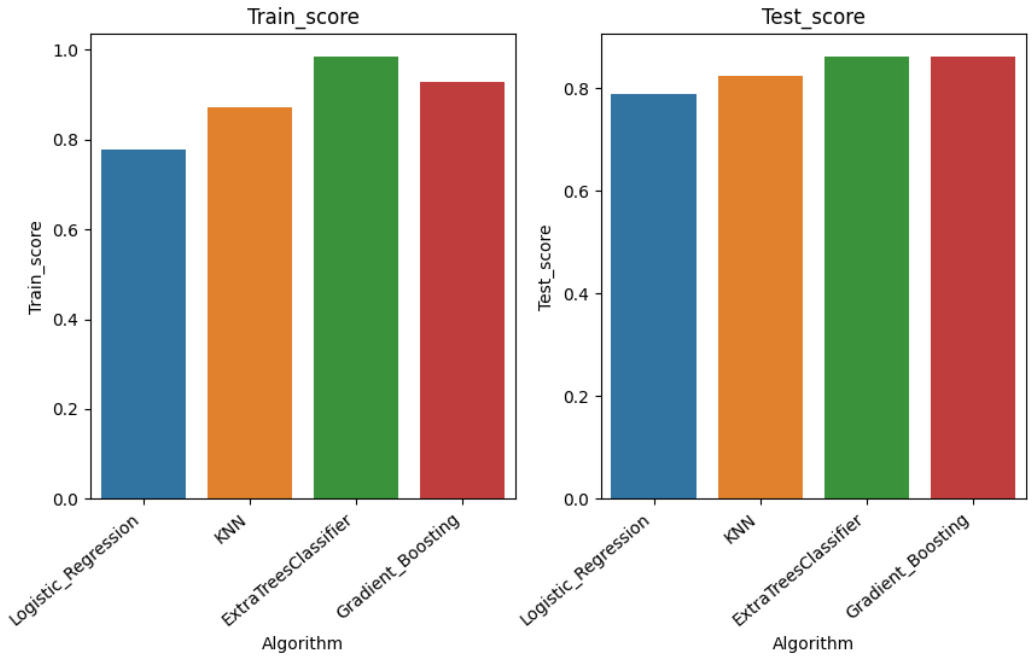
▸ **estimator: GradientBoostingClassifier**

▸ GradientBoostingClassifier

Final model

The focus is on accuracy, so the model to be selected is going to be the one with the highest prediction score. Below, a table and barplots comparing the scores are depicted:

	Algorithm	Train_score	Test_score
0	Logistic_Regression	0.778513	0.787926
1	KNN	0.873385	0.823670
2	ExtraTreesClassifier	0.986216	0.862446
3	Gradient_Boosting	0.929624	0.862262



It can be seen that the Extra Tree Classifier is the one that performs the best, though Gradient Boosting is pretty close. So, The Extra Tree Classifier is choses as the model to be used.

Key findings and insights

From the exploratory data analysis, it can be seen that the hotel reservations are concentrated in the summer season, where it receives the most of its customers.

It also could be seen than lead time, i.e. the time between reservation and arrival, and cancelation are correlated.

The number of special request is negatively correlated with the booking status. This can be observed in the coefficients of the logistic regression model. The more special requests the customer make, the less likely is the cancelation. This negative correlation can also be seen with repeating guesses.

In the correlations chart, it can be seen that the correlation indexes are relatively low, indicating a poor relationship between individual variables and the target variable. However, the model performs good when including more variables, indicating that no individual features, but rather the combination of them is what makes prediction possible.

The chosen model, Extra Trees Classifier, is the one that performs the best, but gradient boosting is pretty close. To get the hyper parameters of the model, Grid Search was used.

Suggestions and next steps

As stated before, this data is from only two years, 2017 and 2018. It would enhance the prediction capabilities to have data from more years. Also, a yearly projection could be made with this additional data.

Finally, appending a question asking for the actual reason for cancelation in the data collection procedure could provide more light into the true causes for reservation cancelation, increasing prediction power of a future model.

References

Ahsan Raza. 2022. Hotel Reservations Dataset. Kaggle. www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset