BC02 - Predicting Hotel Booking Cancellations

Group AD
Haitham Abbas
Nicholas Dafnides
Leonardo Figueiredo
Roeland Rensink

1. Table of content

1. Introduction

In the last decades the traditional travel industry has shifted rapidly from physical travel agencies towards large Online Travel Agencies (OTAs) like Booking.com. On these OTAs everything for your travel can be booked through so-called advanced bookings. From your breakfast until a rental car at arrival can be booked through these OTAs. The largest chunk of these bookings consist of hotel bookings, also known as reservations. These reservations are forward contracts between the hotel and the customer at a settled price. Due to the online competition and pressure from OTAs most hotels offer free cancellation policies. Since the number of OTAs has been increasing there was also a change in the online behavior of customers. Due to these free cancellation policies the number of "deal-seeking" customers grew. Deal-seeking customers tend to book multiple reservations at hotels and have a high likelihood to cancel their reservations for numerous reasons (e.g., looking for other hotels with a better price, location, or social reputation). This is also shown in the cancellation rate by reservation value, which rose from 33% in 2014 to 40% in 2018. The hotels have nowadays a problem where the market expects to have free cancellation policies and the hotel needs to work with overbooking their rooms to encounter the high cancellation rates. Overbooking creates problems in terms of reallocations costs, social reputation damage and loss of immediate and future revenue. But restrictive cancellation policies also create problems like a decrease in revenue or a decrease in the number of bookings.

Hotel chain C is a chain with hotels and resorts located in Portugal. As expected they also encounter the difficulties as described earlier. For this reason, the revenue manager director of hotel chain C has limited the amount of rooms sold with restrictive cancellation policies. To encounter the loss of revenue the manager implements a more aggressive overbooking policy, which as expected drove other expenses higher. The manager has provided us with a dataset with the bookings made in the hotel. Our goal is to forecast net demand based on reservations on the books. With the result from the model the manager expects to be able to better price rooms, have a better overbooking system and identify bookings with high likelihood of canceling. The overall goal is to reduce cancellation to a rate of 20%.

In the report we will explain our steps through the project in the following way, 1. explaining the data exploration; 2. How did we handle the findings in the first part and explain other steps taken in the preprocessing of the dataset; 3. Explain which machine learning models we chose and how we evaluate them; 4. How to implement the model created

2. Data Exploration

See Appendix A, B, C

Hotel Chain C has delivered a csv file consisting of 79.330 records. Each record has 31 features and the datatypes are a mixture of integers, floats and objects. For further exploration we splitted our dataframe based on features being numeric or non-numeric. We determine numeric as having a data type integer or float and non-numeric as a feature with a data type object. In appendix A there is an overview of the datatypes per feature and shows

that the distribution of data types is 13 objects, 2 floats and 16 integers. From the descriptive statistics of the numerical features (Appendix B), we found that the standard deviation of the features Leadtime, DaysInWaitingList and ADR are relatively large. Also we encountered 4 missing values in the feature Children. From the descriptive statistics of the non-numerical features (Appendix C), we found that there are 24 missing values for the feature Country. In another deep dive of the data exploration we grouped categorical features by their values to the target feature IsCanceled to check for disparities. Cancellations proves to indeed show disparities however, these differences in cancellation may only be absolute and not significant relative to the number of clients within those categories i.e. The majority of our cancellations are from Transient customers but those also make up the vast majority of our customers. More will be done on this in the feature engineer section of our project.

3. Data Preprocessing & Feature Engineering

After the data exploration we use the insights for the Data Preprocessing. We start with data cleaning and feature engineering and finish the preprocessing by assigning label encoding for categorical features and scaling our data.

3.1 Data Cleaning

See Appendix D, E, F, G, H

In the data cleaning part we started by checking for duplicate values and found that there are 25.902 duplicates. This amount is very large and needs more explanation before deleting these rows. By creating an extra data frame consisting of only the duplicate records we could create a histogram as shown in appendix E. From the graph we see that roughly 20.000 records (80% of all duplicates) comes from the value Groups and Offline OTAs. A possible explanation is that Groups are registered per person in the system or that Offline OTAs register groups as individuals, because of personal financial incentives for creating more sales in terms of count. In the data exploration we found missing values for the feature Children and Country. Since there are only four records related to the Children feature that are missing we will delete these records. In the description of how the feature Country is retrieved we found out that this feature is mostly updated at the moment that the customer is checked in. We cannot use a feature that will have mostly missing values upon checking in when we create a predictive model. Therefore, the feature Country is dropped. In the data exploration we found that the features ADR and LeadTime have very high standard deviations. This might indicate outliers and so we created a boxplot of these two features which are shown in appendix F. We see that there is one very obvious outlier for ADR, which we will delete. On the lower bound we also found that 1174 rows have a value less than 10. These values are very low, but since these values could be promotional offers we will not delete them from the dataframe. But we did make a constraint for ADR at “ADR < 300” to eliminate the largest outliers. This constraint deleted 43 rows. For LeadTime we also put a constraint on the upperbound of the spectrum. “LeadTime < 365” is the constraint and deletes 376 records from the dataframe. In the total data cleaning part we have decided to delete 79.330 - 52.985 = 26345. The largest part of this deleting came from the duplicate values. So taking the duplicate values of the total features gives a better total value to

compare. 79.330 - 25.902 = 53.428. So from the adjusted dataset we deleted (1 - 52.985 / 53.428) * 100% = 0.82% of records during our data cleaning process.

3.2 Feature Engineering

In this part we have created three new features from existing features. The first feature that has been created is ChangeInRoom. The values for its rows are built from two features, AssignedRoomType and ReservedRoomType. It returns a 0 when the AssignedRoomType equals the ReservedRoomType and a 1 if this is not true. The second feature that we created is the feature Dependants. This feature is a combination of the features Babies and Children and returns the combined value of the two columns. The third feature that has been created is TotalBookings. This is a combination of the features PreviousBookingsNotcanceled and PreviousCancellations. It returns a total number of previous bookings, regardless if it is canceled or not. The fourth feature that has been created is StayDuration. It is a combination of the features ReservationStatusDate and ArrivalDate. Lastly, we did for categorical features a label encoder so that these features are able to be used in our machine learning model.

3.3 Feature Selection

Feature selection comes in handy when there are many features introduced to the model. Very few models are robust to high dimensionality, and even if we were to use them it is always better to reduce the amount of features in the model. We started by creating a correlation matrix to check if there is a high correlation between any of the independent variables we have, because a high correlation may indicate multicollinearity and this will reduce the quality of our data.. We didn't find any high correlation between the independent variables apart from the variable StayDuration which we engineered, therefore we decided to drop it, the correlation matrix can be viewed in appendix G. After label encoding our non-numeric features we ended up with more than 600 features, this introduces high dimensionality and further reduction of the quality of our data, for this reason we constructed a feature selection criterion. We used a Decision tree regressor to measure the feature importance, we chose the squared error as criterion over the rest because it is the most accurate and used criterion, after examining our results we chose to set a threshold to reduce the features, our threshold was an MSE value of 0.005 and thus our features dropped to 20 features, the 20 features and their MSE score can be viewed in appendix H.

4. Modeling

At first we attempted a normal train test split, however since we were working with time series data, we wanted to make sure we didn't have data leakage. The other factor was we wanted repeatability in testing our models. This process leads us to use cross validation methods. We used Stratified K Fold at first with the thought process that if we split the dataset into n segments, we would have excellent repeatability in testing our models. However, we later discovered that time series data can be problematic with cross validation. In particular when we observed a diagram for Stratified K Fold, we saw that we were indeed

future looking in our models. So we therefore ended up splitting our dataset with the time series split method. This was a specifically developed model selection for time series data as the training sets are supersets of those that came before them. Indeed, we could avoid the future looking this way.

For evaluation metrics, we knew that false negatives and false positives are not equally costly. We knew that since we are dealing with time series data that we would be using two different evaluation metrics that would be based on the season to judge our models. Therefore, we chose to use an F0.5 score for the high season. This is because we want to reduce false positives, meaning that we want to reduce overbooking. On the contrary, for the low season we used a F1.5 score to evaluate because false negatives are more costly.

Cross validation scores were used to find the ideal number of n splits. We also had to create scoring variables that specifically used the F Score we wanted for the particular season. After we had that value, we went with XGBoost as our model. The reason being, it simply performed much better than other models using the evaluation metrics we decided on. After a lot of parameter fine tuning, we achieved significantly positive results with a F 1.5 score of .84 for Low Season and a .74 F 0.5 Score for High Season.

5. Deployment

To use the model as efficiently as possible we have created a platform for the manager where he can have a technical analysis of the model. Besides that he also has the ability to load new reservations into the model to predict future cancellations. In this part we will mainly focus on the interpretation of the output after putting the new reservations through the model.

After loading new reservations into the platform you will get multiple graphs with output. The first graph is a simple distribution of the amount of expected cancellation. In appendix I you can find an example of this graph. The second graph will show you the net demand for the input given. In appendix J you can see how the module returns the first 14 days with the corresponding expected net-demand. Based on this the manager should be able to decide if he can overbook more or less. Another useful tool for the manager is the output with the certainty of the outcome being cancellations. The model produces the probability for every predicted canceled reservation. This list can be downloaded from the platform and based upon this list the manager is able to determine if he should take action towards people that have a high probability of cancelation.

6. Appendix

A. Info on Name, NoN-Null Count and Dtype of features

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   IsCanceled                      79330 non-null  int64
 1   LeadTime                        79330 non-null  int64
 2   ArrivalDateYear                 79330 non-null  int64
 3   ArrivalDateMonth                79330 non-null  object
 4   ArrivalDateWeekNumber           79330 non-null  int64
 5   ArrivalDateDayOfMonth           79330 non-null  int64
 6   StaysInWeekendNights            79330 non-null  int64
 7   StaysInWeekNights               79330 non-null  int64
 8   Adults                          79330 non-null  int64
 9   Children                        79326 non-null  float64
 10  Babies                          79330 non-null  int64
 11  Meal                            79330 non-null  object
 12  Country                         79306 non-null  object
 13  MarketSegment                   79330 non-null  object
 14  DistributionChannel             79330 non-null  object
 15  IsRepeatedGuest                 79330 non-null  int64
 16  PreviousCancellations           79330 non-null  int64
 17  PreviousBookingsNotCanceled     79330 non-null  int64
 18  ReservedRoomType                79330 non-null  object
 19  AssignedRoomType                79330 non-null  object
 20  BookingChanges                  79330 non-null  int64
 21  DepositType                     79330 non-null  object
 22  Agent                           79330 non-null  object
 23  Company                         79330 non-null  object
 24  DaysInWaitingList               79330 non-null  int64
 25  CustomerType                    79330 non-null  object
 26  ADR                             79330 non-null  float64
 27  RequiredCarParkingSpaces        79330 non-null  int64
 28  TotalOfSpecialRequests          79330 non-null  int64
 29  ReservationStatus               79330 non-null  object
 30  ReservationStatusDate           79330 non-null  object
dtypes: float64(2), int64(16), object(13)
```

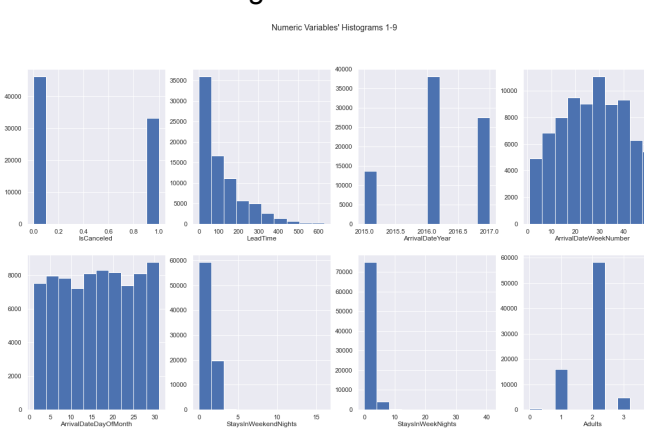
B. Descriptive statistics and missing value count for numeric features

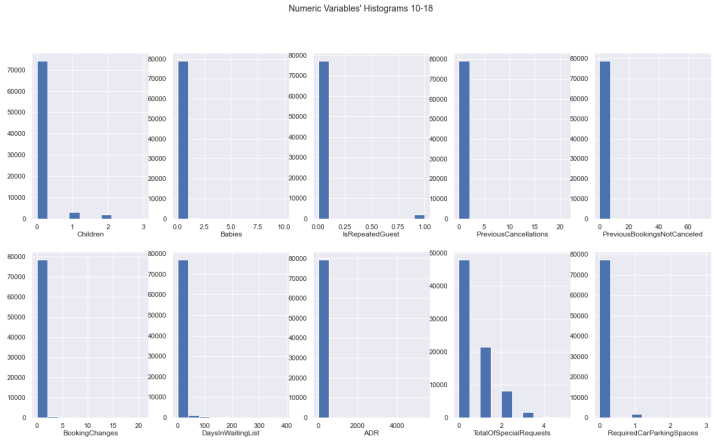
	count	mean	std	min	25%	50%	75%	max	dtype	size	na_count
IsCanceled	79330.000000	0.417270	0.493111	0.000000	0.000000	0.000000	1.000000	1.000000	int64	79330	0
LeadTime	79330.000000	109.735724	110.948526	0.000000	23.000000	74.000000	163.000000	629.000000	int64	79330	0
ArrivalDateYear	79330.000000	2016.174285	0.699181	2015.000000	2016.000000	2016.000000	2017.000000	2017.000000	int64	79330	0
ArrivalDateWeekNumber	79330.000000	27.177449	13.398523	1.000000	17.000000	27.000000	38.000000	53.000000	int64	79330	0
ArrivalDateDayOfMonth	79330.000000	15.786625	8.728451	1.000000	8.000000	16.000000	23.000000	31.000000	int64	79330	0
StaysInWeekendNights	79330.000000	0.795185	0.885026	0.000000	0.000000	1.000000	2.000000	16.000000	int64	79330	0
StaysInWeekNights	79330.000000	2.182957	1.456416	0.000000	1.000000	2.000000	3.000000	41.000000	int64	79330	0
Adults	79330.000000	1.850977	0.509292	0.000000	2.000000	2.000000	2.000000	4.000000	int64	79330	0
Children	79326.000000	0.091370	0.372177	0.000000	0.000000	0.000000	0.000000	3.000000	float64	79330	4
Babies	79330.000000	0.004941	0.084323	0.000000	0.000000	0.000000	0.000000	10.000000	int64	79330	0
IsRepeatedGuest	79330.000000	0.025615	0.157983	0.000000	0.000000	0.000000	0.000000	1.000000	int64	79330	0
PreviousCancellations	79330.000000	0.079743	0.415472	0.000000	0.000000	0.000000	0.000000	21.000000	int64	79330	0
PreviousBookingsNotCanceled	79330.000000	0.132371	1.693411	0.000000	0.000000	0.000000	0.000000	72.000000	int64	79330	0
BookingChanges	79330.000000	0.187369	0.608620	0.000000	0.000000	0.000000	0.000000	21.000000	int64	79330	0
DaysInWaitingList	79330.000000	3.226774	20.870890	0.000000	0.000000	0.000000	0.000000	391.000000	int64	79330	0
ADR	79330.000000	105.304465	43.602954	0.000000	79.200000	99.900000	126.000000	5400.000000	float64	79330	0
TotalOfSpecialRequests	79330.000000	0.546918	0.780776	0.000000	0.000000	0.000000	1.000000	5.000000	int64	79330	0
RequiredCarParkingSpaces	79330.000000	0.024367	0.154919	0.000000	0.000000	0.000000	0.000000	3.000000	int64	79330	0

C. Descriptive statistics and missing value count for non-numeric features

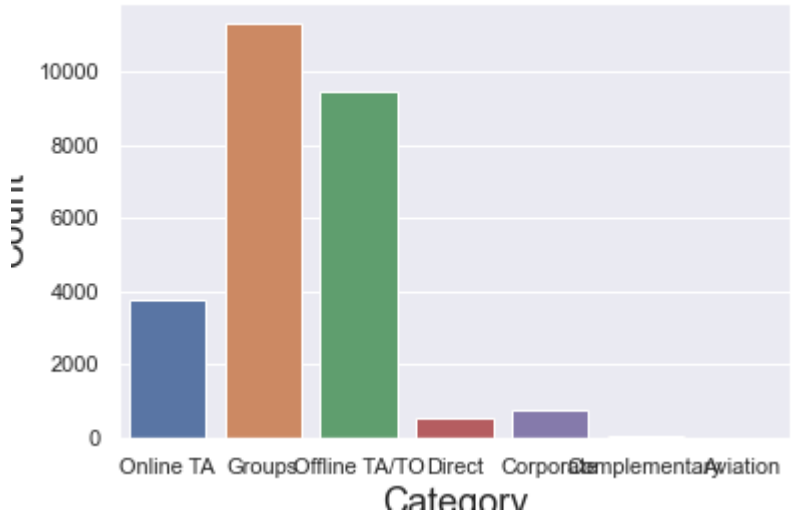
	count	unique	top	freq	dtype	size	na_count
ArrivalDateMonth	79330	12	August	8983	object	79330	0
Meal	79330	4	BB	62305	object	79330	0
Country	79306	166	PRT	30960	object	79330	24
MarketSegment	79330	8	Online TA	38748	object	79330	0
DistributionChannel	79330	5	TA/TO	68945	object	79330	0
ReservedRoomType	79330	8	A	62595	object	79330	0
AssignedRoomType	79330	9	A	57007	object	79330	0
DepositType	79330	3	No Deposit	66442	object	79330	0
Agent	79330	224	9	31955	object	79330	0
Company	79330	208	NULL	75641	object	79330	0
CustomerType	79330	4	Transient	59404	object	79330	0
ReservationStatus	79330	3	Check-Out	46228	object	79330	0
ReservationStatusDate	79330	864	2015-10-21	1416	object	79330	0

D. Histogram of Numeric features

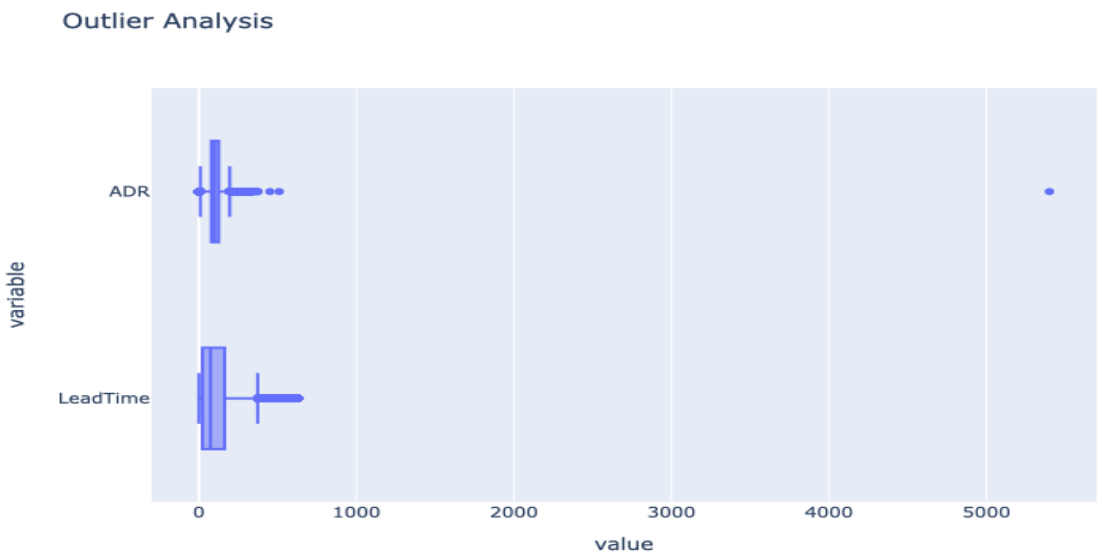




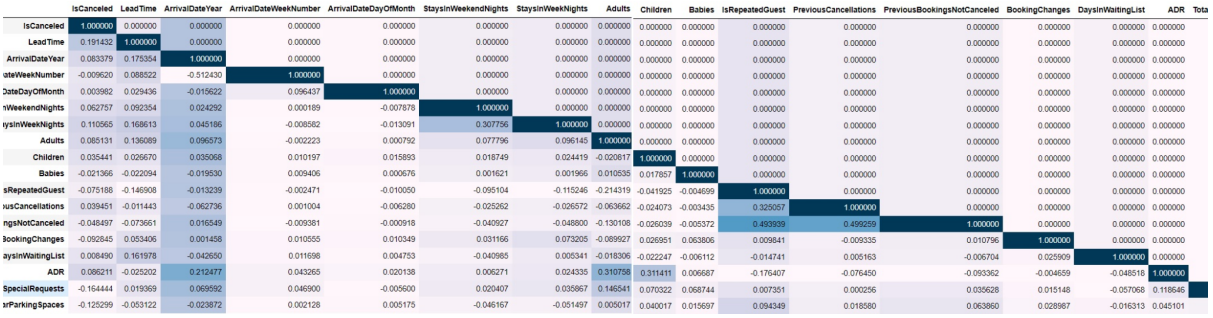
E. Duplicate count per category



F. Boxplot of the feature ADR and LeadTime



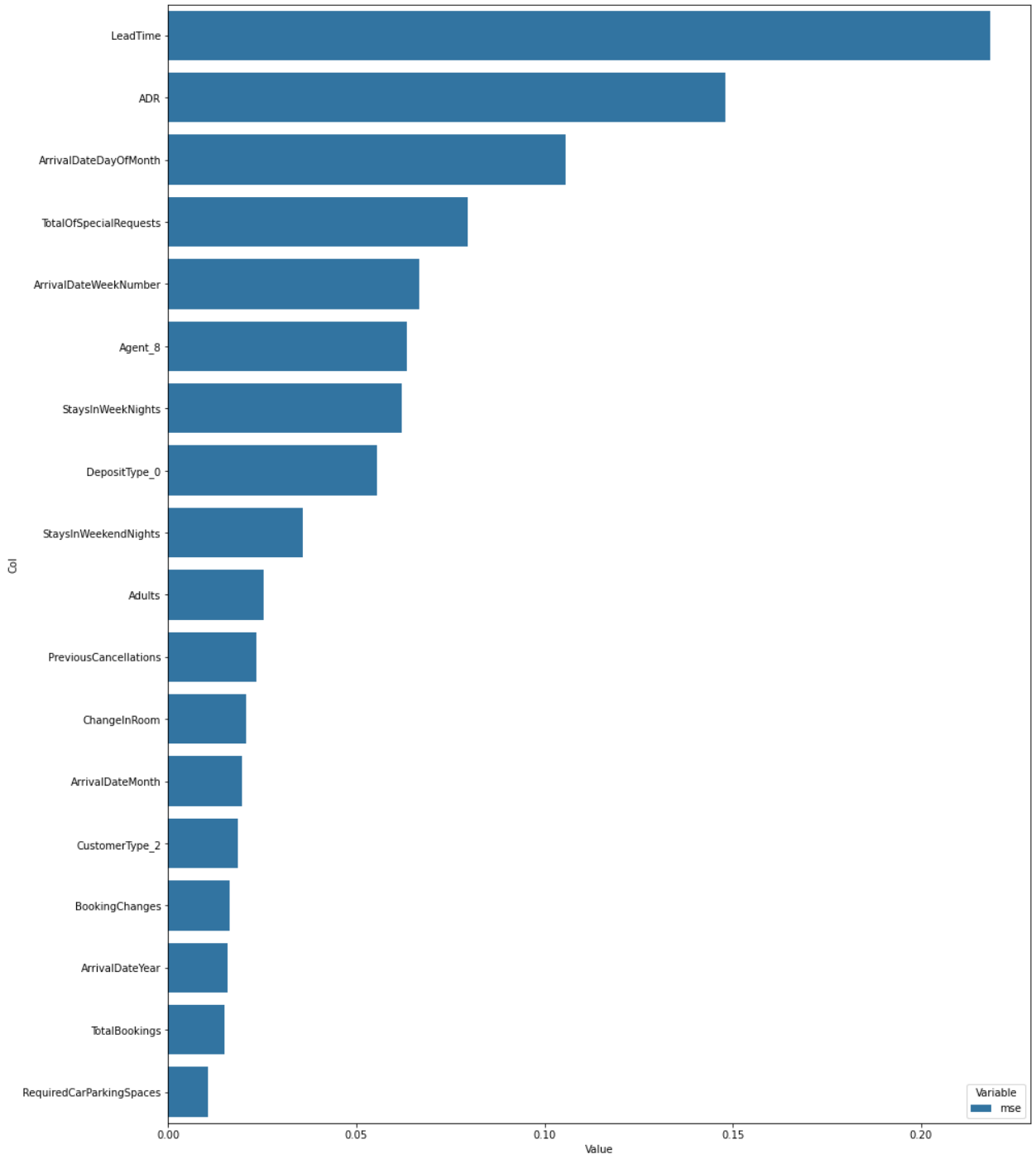
G. Correlation



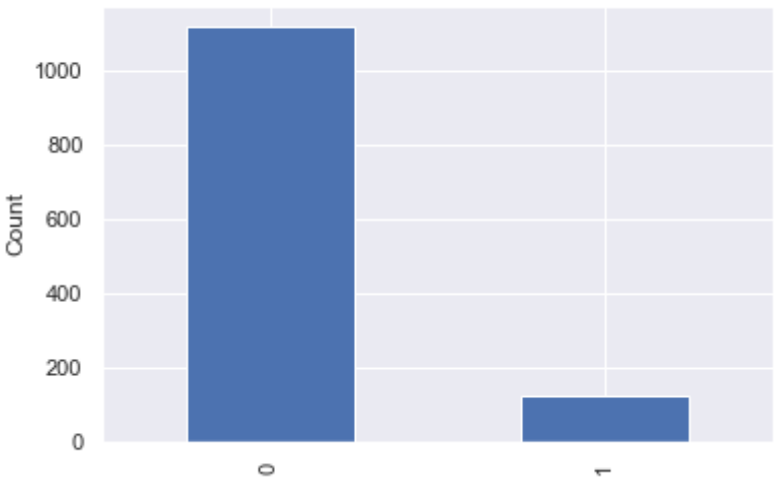
	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	StaysInWeekendNights	StaysInWeekNights	Adults	Children	Babies	IsRepeatedGuest	PreviousCancellations	PreviousBookingsNotCanceled	BookingChanges	DaysInWaitingList	ADR	Tota
IsCanceled	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
LeadTime	0.191432	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ArrivalDateYear	0.083379	0.175354	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ateWeekNumber	-0.009620	0.088522	-0.512430	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
DateDayOfMonth	0.003982	0.029436	-0.015622	0.096437	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
nWeekendNights	0.062757	0.092354	0.024292	0.000189	-0.007878	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
ysInWeekNights	0.110565	0.168613	0.045186	-0.008582	-0.013091	0.307756	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
Adults	0.085131	0.136089	0.096573	-0.002223	0.000792	0.077796	0.096145	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
Children	0.035441	0.026670	0.035068	0.010197	0.015893	0.018749	0.024419	-0.020817	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
Babies	-0.021366	-0.022094	-0.019530	0.009406	0.000676	0.001621	0.001966	0.010535	0.017857	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
sRepeatedGuest	-0.075188	-0.146908	-0.013239	-0.002471	-0.010050	-0.095104	-0.115246	-0.214319	-0.041925	-0.004699	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
usCancellations	0.039451	-0.011443	-0.062736	0.001004	-0.006280	-0.025262	-0.026572	-0.063662	-0.024073	-0.003435	0.325057	1.000000	0.000000	0.000000	0.000000	0.000000	
ngsNotCanceled	-0.048497	-0.073661	0.016549	-0.009381	-0.000918	-0.040927	-0.048800	-0.130108	-0.026039	-0.005372	0.493939	0.499259	1.000000	0.000000	0.000000	0.000000	
BookingChanges	-0.092845	0.053406	0.001458	0.010555	0.010349	0.031166	0.073205	-0.089927	0.026951	0.063806	0.009841	-0.009335	0.010796	1.000000	0.000000	0.000000	
aysInWaitingList	0.008490	0.161978	-0.042650	0.011698	0.004753	-0.040985	0.005341	-0.018306	-0.022247	-0.006112	-0.014741	0.005163	-0.006704	0.025909	1.000000	0.000000	
ADR	0.086211	-0.025202	0.212477	0.043265	0.020138	0.006271	0.024335	0.310758	0.311411	0.006687	-0.176407	-0.076450	-0.093362	-0.004659	-0.048516	1.000000	
SpecialRequests	-0.164444	0.019369	0.069592	0.046900	-0.005600	0.020407	0.035867	0.146541	0.070322	0.068744	0.007351	0.000256	0.035628	0.015148	-0.057068	0.118645	
rParkingSpaces	-0.125299	-0.053122	-0.023872	0.002128	0.005175	-0.046167	-0.051497	0.005017	0.040017	0.015697	0.094349	0.018580	0.063860	0.028987	-0.016313	0.045101	

H. Feature Importance

	features	feature_importance
0	LeadTime	0.200967
1	ArrivalDateYear	0.016222
2	ArrivalDateMonth	0.020719
3	ArrivalDateWeekNumber	0.073004
4	ArrivalDateDayOfMonth	0.096832
5	StaysInWeekendNights	0.034904
6	StaysInWeekNights	0.058606
7	Adults	0.016705
8	Children	0.005239
11	PreviousCancellations	0.017441
12	PreviousBookingsNotCanceled	0.011819
13	BookingChanges	0.016443
15	ADR	0.128362
16	RequiredCarParkingSpaces	0.013494
17	TotalOfSpecialRequests	0.082620
18	ChangeInRoom	0.021610
19	Dependants	0.006183
236	Agent_8	0.057006
454	CustomerType_2	0.014645
457	DepositType_1	0.049416



I. Example of output on platform



J. Example of output on platform for Net Demand