Introduction

In this analysis, I use Excel, MySQL, PowerBI, Tableau, Orange (Python), and JASP (R) to analyze the hotel booking dataset. The data for this project can be found here https://www.sciencedirect.com/science/article/pii/S2352340918315191 - The data consists of hotel reservations for two hotels found in Portugal. I imported the data to MySQL to join tables, adjust columns, and create new columns. As the data had a column for the country of origin (as an Alpha-3 Code) of the guest, I got the idea to find data on countries from various sources and join them to the main dataset. I obtained latitude and longitude data from Github and currency, region, and the full name of the country from the World Bank. Then, I used PowerBI and Tableau to extract the data from MySQL to create dashboards for data visualization purposes. Afterward, I took a sample of the data and conducted statistical analysis in JASP and predictive analytics (machine learning) in Orange. I used classification methods to predict hotel booking cancellations, regression methods to predict average daily rate, clustering to create new market segments, and dimensionality reduction to simplify the dataset.

Dataset Information

Main Dataset: https://www.kaggle.com/jessemostipak/hotel-booking-demand

Global Holiday Dataset: https://www.worldpop.org/doi/10.5258/SOTON/WP00689

World Happiness Indices: https://worldhappiness.report/archive/

Latitude and Longitude of Countries:

https://gist.github.com/gblmarquez/d398d3536252a149b58879f91084f5bf

Country Metadata: https://genderdata.worldbank.org/

Value of the data

- Descriptive analytics can be employed to further understand patterns, trends, and anomalies in data;
- Used to perform research in different problems like: bookings cancellation prediction, customer segmentation, customer satiation, seasonality, among others;
- Researchers can use the datasets to benchmark bookings' prediction cancellation models against results already known (e.g. [1]);
- Machine learning researchers can use the datasets for benchmarking the performance of different algorithms for solving the same type of problem (classification, segmentation, or other);
- Educators can use the datasets for machine learning classification or segmentation problems;
- Educators can use the datasets to obtain either statistics or data mining training.

A B S T R A C T

This data article describes two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

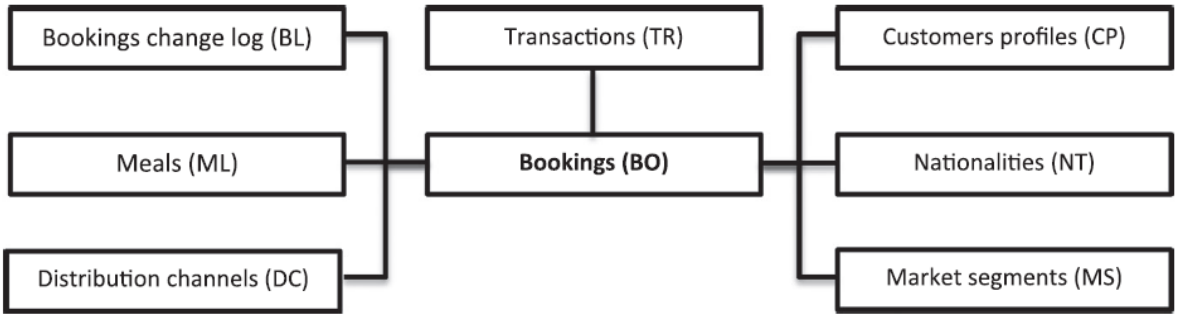


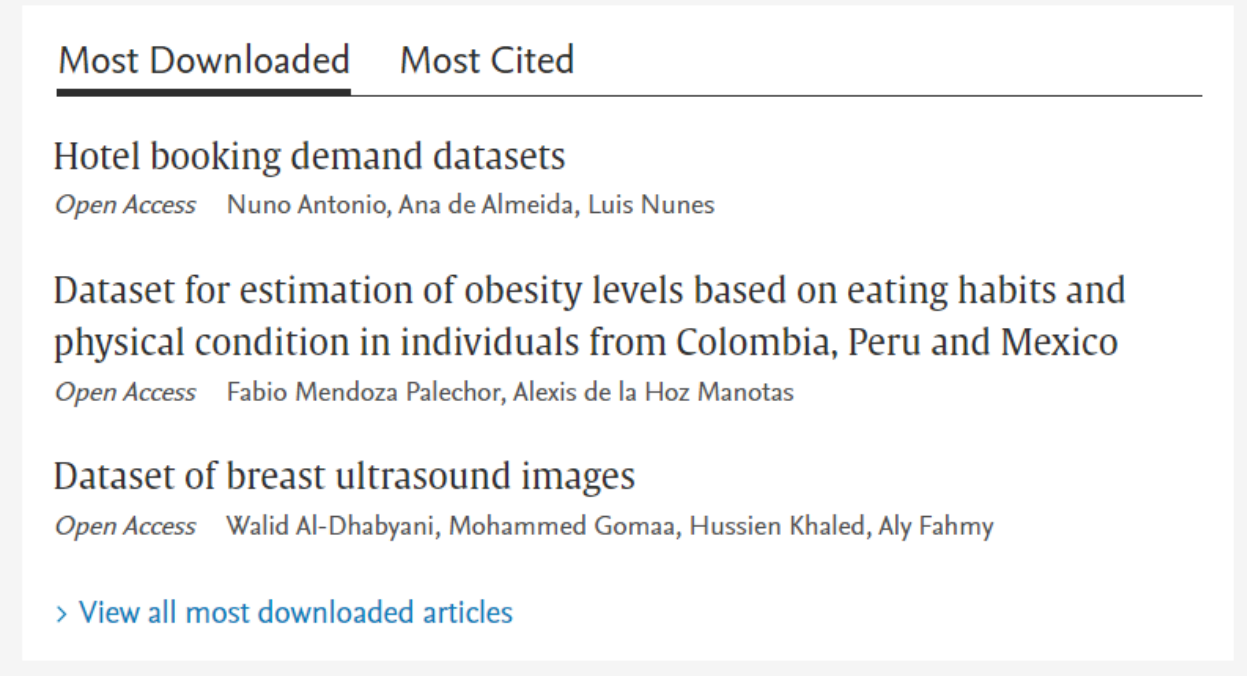
Fig. 1. Diagram of PMS database tables where variables where extracted from.

Why I Chose This Dataset

I really enjoy the hospitality industry and found customer behavior interesting. This dataset has also been used to conduct academic research and has led to many different published articles from many different authors. It is also the most downloaded dataset from the "Data in Brief" academic journal which is a huge feat as this journal consists of datasets from virtually every discipline including chemistry and biology.

General Steps for Data Projects

1. Create a data dictionary by giving context and labeling the type of data it is (continuous, discrete, etc.) 2. List the metadata such as number of rows, columns, etc.; how many missing values and what do we do about the missing values? 3. What business problems can I solve/research or questions I can ask? 4. Conduct exploratory data analysis; descriptive statistics 5. Data transformations (imputation, binning, missing values, normalizing, transformation, flags, standardization, attribute selection/weighting/generation, dimensionality reduction) 6. Modeling (training, testing, validation sets, hyperparameter tuning) 7. Compare models (visualization through ROC curves, confusion matrix, etc..)



Data Dictionary

Variable	Type	Description	Source/Engineering
		Main Hotel Dataset	
ADR	Numeric	Average Daily Rate	BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights
Adults	Integer	Number of Adults	BO and BL
Agent	Categorical	ID of the travel agency that made the booking	BO and BL
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date	BO and BL

ArrivalDateMonth	Categorical	Month of arrival date with 12 categories: "January" to "December"	BO and BL
ArrivalDateWeekNumber	Integer	Week number of the arrival date	BO and BL
ArrivalDateYear	Integer	Year of arrival date	BO and BL
AssignedRoomType	Categorical	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 for anonymity reasons presented instead of designation	BO and BL
Babies	Integer	Number of Babies	BO and BL
BookingChanges	Integer	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	BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal
Children	Integer	Number of Children	BO and BL/Sum of both payable and non-payable children

Company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons	BO and BL
Country	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format	BO, BL and NT
CustomerType	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to 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 other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking	BO and BL

DaysInWaitingList	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer	BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – 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 stay.	BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking's arrival or cancellation date. In case no payments were found the value is “No Deposit”. If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”. Otherwise the value is set as “Refundable
DistributionChannel	Categorical	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”	BO, BL and DC
IsCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)	BO

IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)	BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	BO and BL/ Subtraction of the entering date from the arrival date
MarketSegment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”	BO, BL, and MS

Meal	Categorical	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)	BO, BL, and ML
PreviousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking	BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled.

PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking	BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled.
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer	BO and BL
ReservationStatus	Categorical	Reservation 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 – customer did not check-in and did inform the hotel of the reason why	BO
ReservationStatusDate	Date	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel	BO

ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons	BO and BL
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	BO and BL/ Calculated by counting the number of weekend nights from the total number of nights
StaysInWeekNights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	BO and BL/Calculated by counting the number of week nights from the total number of nights
TotalOfSpecialRequests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)	BO and BL/Sum of all special requests

* Note: BO = Bookings; BL = Booking Change Log; ML = Meals; DC = Distribution Channels; TR = Transactions; CP = Customer Profiles; NT = Nationalities; MS = Market Segments

SQL Query

```
CREATE TABLE hotel_bookings_final (
SELECT
Hotel,
Agent,
Reserved_Room_Type,
Assigned_Room_Type,
CONCAT(Reserved_Room_Type,"-->",Assigned_Room_Type) as Reserved_Assigned,
Company,
case when Country="CN" then "CHN" else Country end as Country_Guest,
c.short_name as Country_Name_Short,
c.long_name as Country_Name_Long,
c.region as Country_Region,
```

```
c.Income_group as Country_Income_Group,
c.currency_unit as Country_Currency,
b.Latitude as Country_Latitude,
b.Longitude as Country_Longitude,
Customer_Type,
Deposit_Type,
Distribution_Channel,
Market_Segment,
Meal,
case when meal="BB" then "Bed & Breakfast"
when meal="HB" then "Breakfast and One Other"
when meal="FB" then "Breakfast, Lunch, Dinner"
when meal in ("SC", "Undefined") then "No Meal"
else "no meal" end as meal_text,
Reservation_Status,
Reservation_Status_Date,
Arrival_Date_Day_Of_Month,
Arrival_Date_Week_Number,
Arrival_Date_Month,
case when arrival_date_month = "January" then 1
when arrival_date_month = "February" then 2
when arrival_date_month = "March" then 3
when arrival_date_month = "April" then 4
when arrival_date_month = "May" then 5
when arrival_date_month = "June" then 6
when arrival_date_month = "July" then 7
when arrival_date_month = "August" then 8
when arrival_date_month = "September" then 9
when arrival_date_month = "October" then 10
when arrival_date_month = "November" then 11
when arrival_date_month = "December" then 12
else null end as arrival_date_month_adj,
Arrival_Date_Year,
Stays_In_Weekend_Nights,
Stays_In_Week_Nights,
(stays_in_weekend_nights + stays_in_week_nights) as total_roomnights,
Lead_Time,
case when lead_time=0 then "0 days"
when lead_time between 1 and 99 then "1 day to 99 days"
when lead_time between 100 and 199 then "100 days to 199 days"
```

```
when lead_time between 200 and 299 then "200 days to 299 days"
when lead_time between 300 and 399 then "300 days to 399 days"
when lead_time between 400 and 499 then "400 days to 499 days"
when lead_time between 500 and 599 then "500 days to 599 days"
when lead_time between 600 and 699 then "600 days to 699 days"
when lead_time between 700 and 799 then "700 days to 799 days"
else null end as lead_time_bin,
Days_In_Waiting_List,
Total_Of_Special_Requests,
ADR,
case when adr=0 then "$0"
when adr between .001 and 49.99 then "$1 to $49"
when adr between 50 and 99.99 then "$50 to $99"
when adr between 100 and 149.99 then "$100 to $149"
when adr between 150 and 199.99 then "$150 to $199"
when adr between 200 and 249.99 then "$200 to $249"
when adr between 250 and 299.99 then "$250 to $299"
when adr between 300 and 349.99 then "$300 to $349"
when adr between 350 and 399.99 then "$350 to $399"
when adr between 400 and 449.99 then "$400 to $449"
when adr between 450 and 499.99 then "$450 to $499"
when adr > 500 then "$500+"
else null end as adr_bin,
Adults,
Children,
Babies,
(adults+children+babies) as Total_People,
Is_Canceled,
Is_Repeated_Guest,
Booking_Changes,
Previous_Bookings_Not_Canceled,
Previous_Cancellations,
Required_Car_Parking_Spaces
FROM hotelbooking.hotel_bookings a
LEFT JOIN hotelbooking.countries b ON a.country=b.Three_Letter_Code
LEFT JOIN world_bank.wb_detailedstats c ON a.country=c.country_code
);

SELECT a.*,
```

```
concat(arrival_date_month_adj, "/",Arrival_Date_Day_Of_Month,"/",arrival_date_year) as
Arrival_Date,
concat(arrival_date_year,"-",arrival_date_month_adj) as Arrival_Year_Month
 FROM hotelbooking.hotel_bookings_final a;
```

Created Columns

Arrival_date_month_adj

Total_roomnights

Total_people

Meal_adj

Lead_time_bin

ADR_bin

Country_Name_Short

Country_Currency

Country_Name_Long

Cuontry_Region

Country_Income_Group

YearMonth

Predictive Analytics

Which Variables We Will Predict and Why

For our classification problem, we will predict cancellations (0 for not cancelled, 1 for cancelled). It is important to predict cancellations so that the hotel can better manage overbookings and overall revenue management. For our regression problem, we will predict ADR (average daily rate). This metric is standard in hotel revenue management as it will help the hotel better gauge where the most revenue will be derived from.

Variable Inclusion for Predictions

Just because we have the data, does not mean we should include it in our model. More data is not always better due to the added dimensionality and computational resources needed. Remember, we are training a model to predict the value of one column on new rows of data. We need to ask ourselves, which pieces of data will help us achieve the most accurate predictions?

Cancellation Prediction: hotel, lead_time, arrival_date_month, stays_in_weekend_nights, stays_in_week_nights, adults, market_segment, distribution_channel, is_repeated_guest, previous_cancellations, reserved_room_type, booking_changes, deposit_type, customer_type, adr, total_of_special_requests

ADR Prediction: hotel, arrival_date_month, stays_in_weekend_nights, stays_in_week_nights, adults, meal, market_segment, distribution_channel, reserved_room_type, customer_type, total_of_special_requests, lead_time, deposit_type

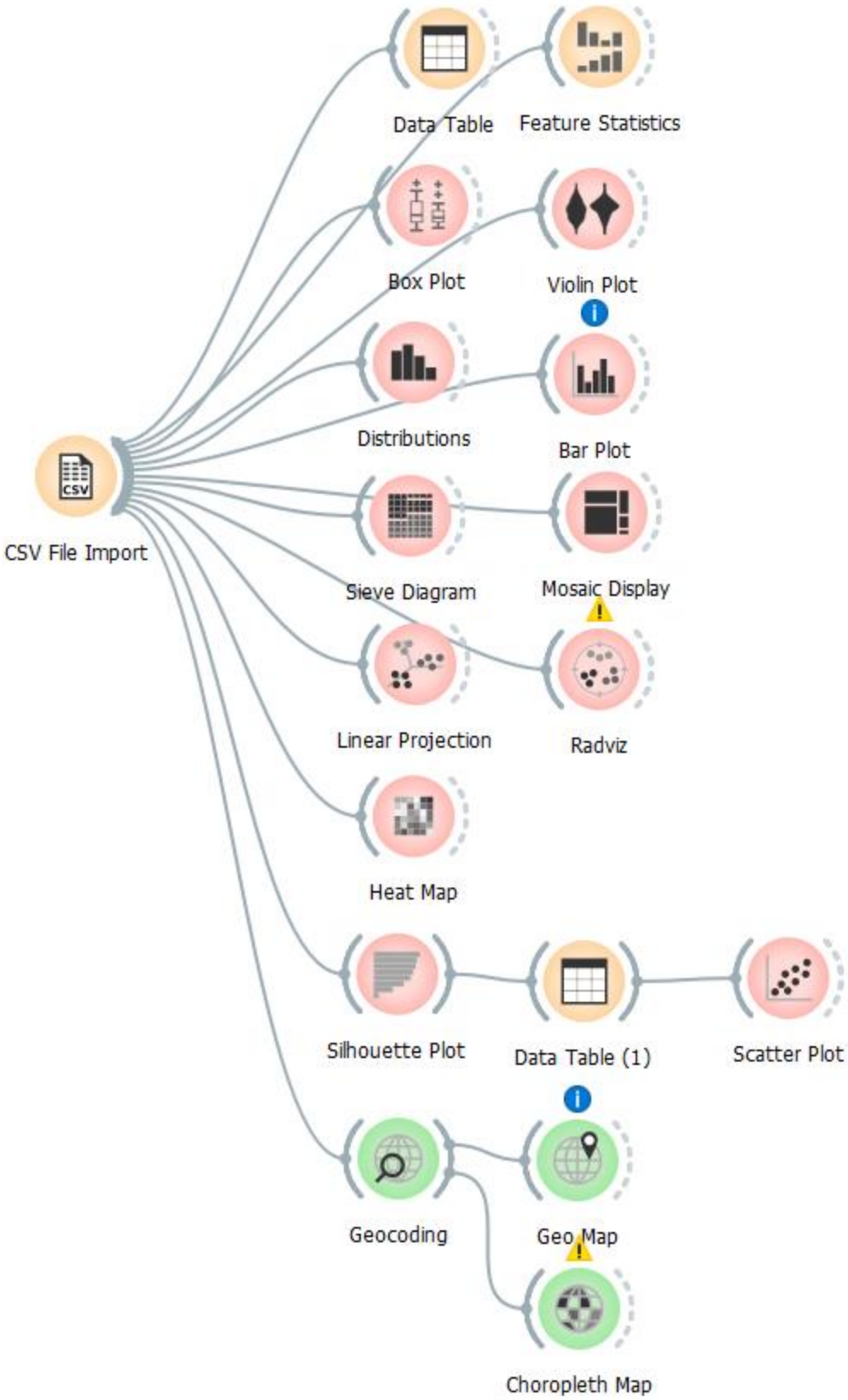
Clustering: hotel, customer_type, market_segment, distribution_channel, reserved_room_type, lead_time, reservation_status, stays_in_weekend_nights, stays_in_week_nights, total_of_special_requests, adults, adr

Predictive Analytics Results

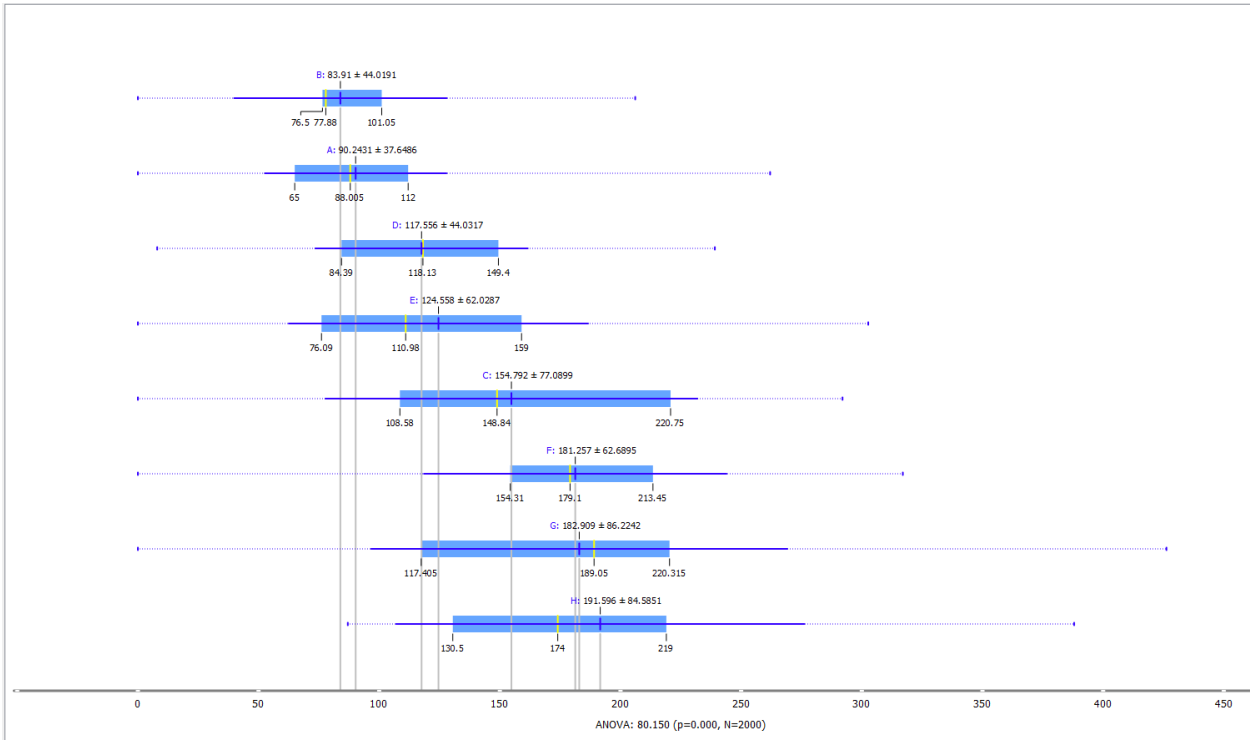
Exploratory Data Analysis

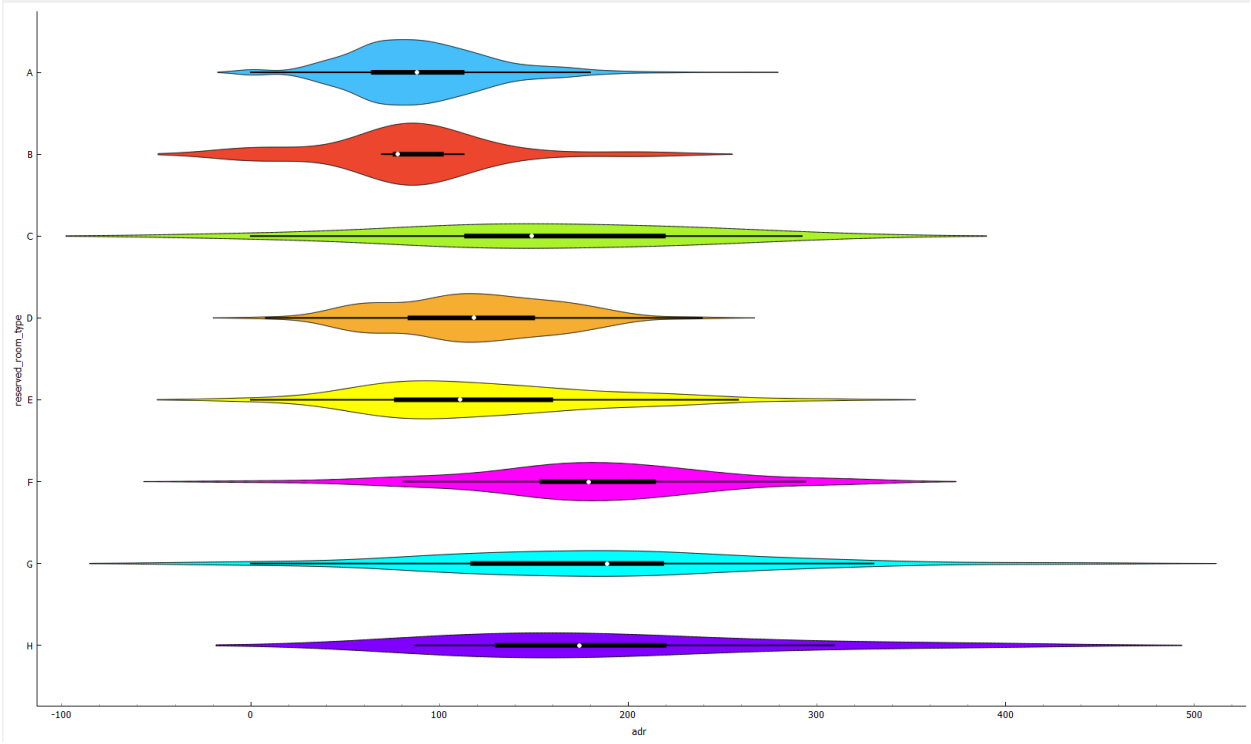
Exploratory data analysis was conducted to create summary statistics, analyze distributions, find outliers, and visualize the data in various ways.

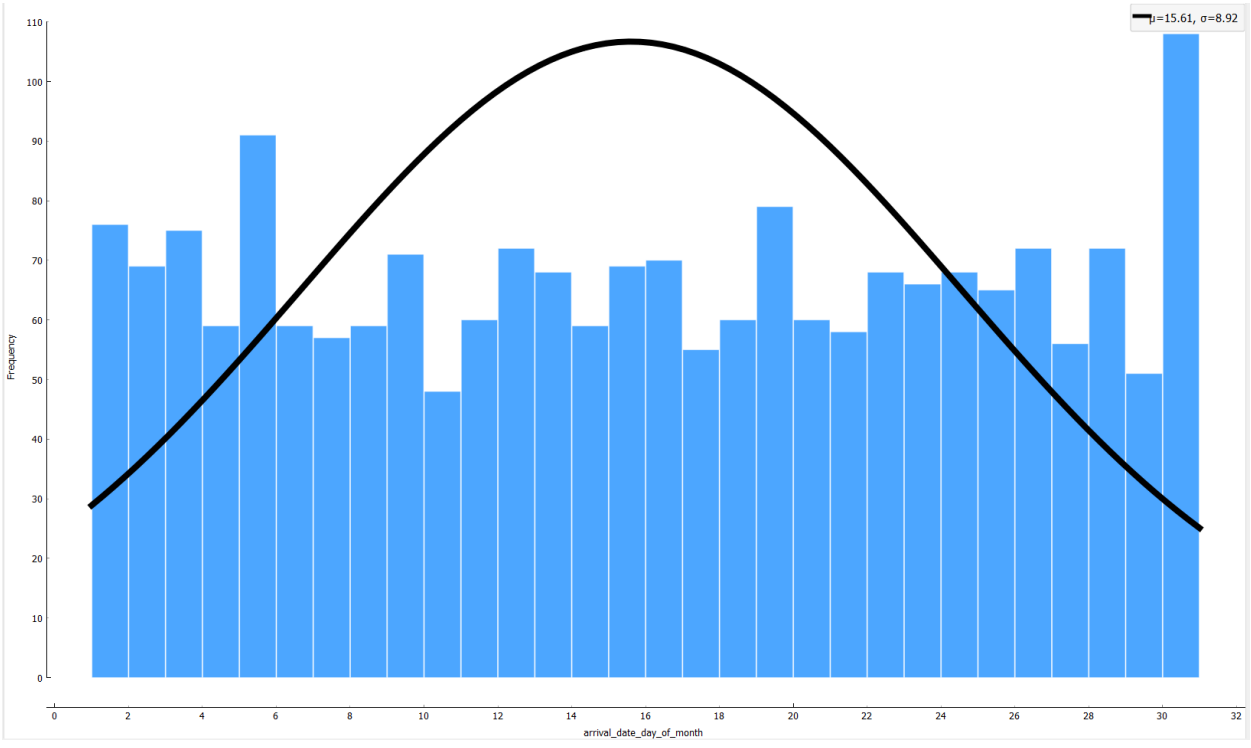
Exploratory Data Analysis

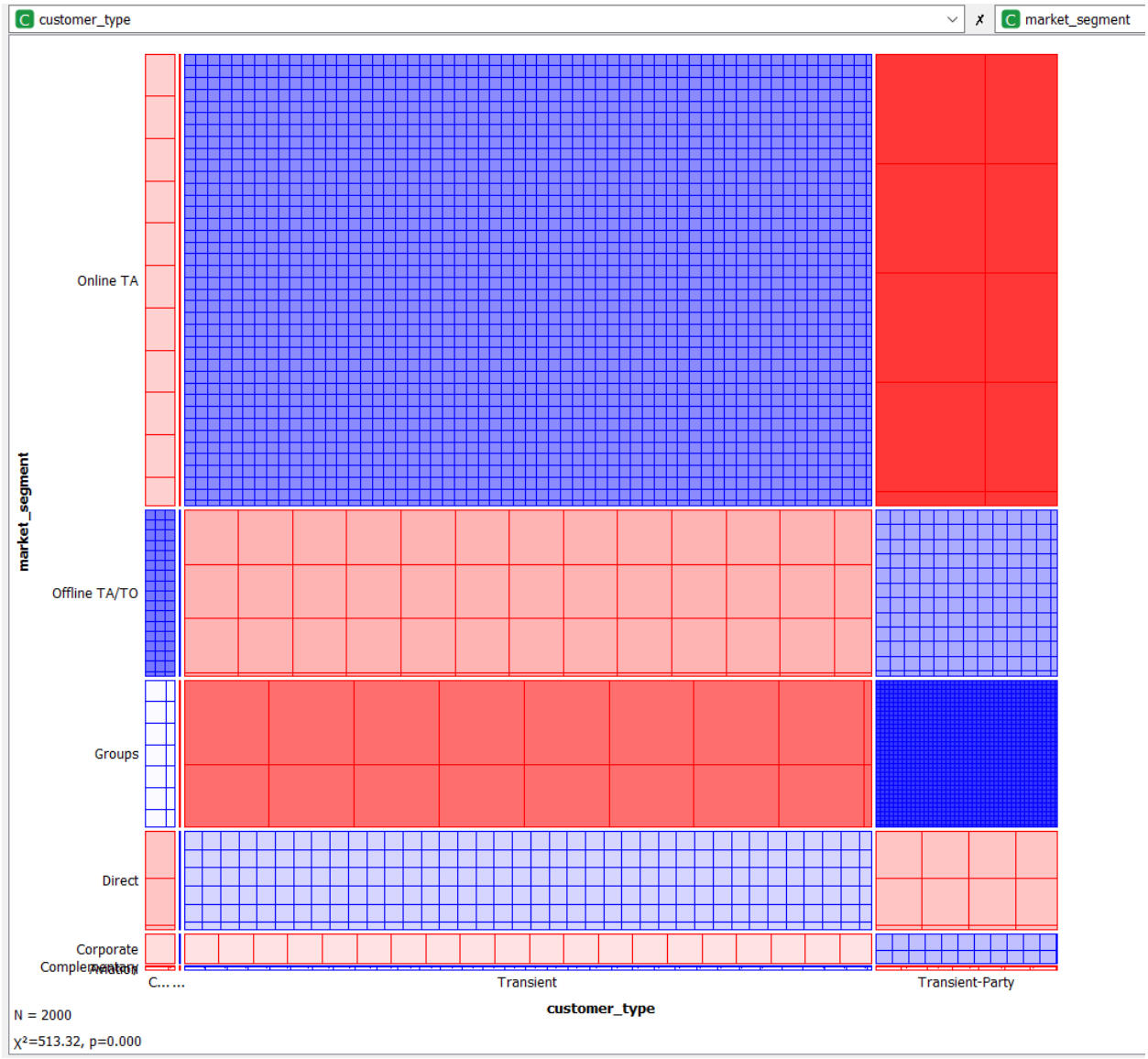


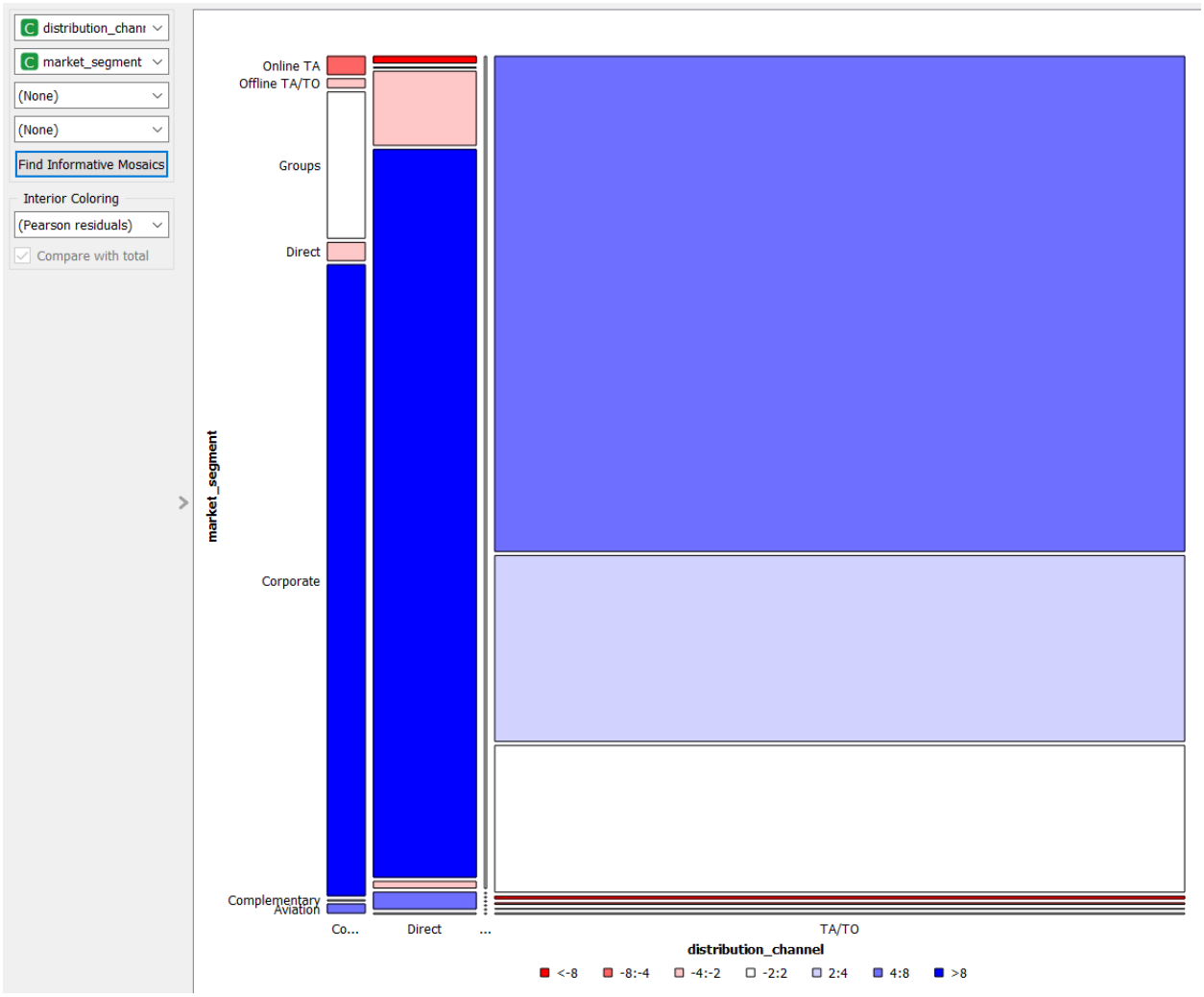
	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
N	adr		101.329	94.65	0.482723	0.00	426.25	0 (0%)
N	adults		1.85	2	0.27	0	4	0 (0%)
N	arrival_date...		15.61	16	0.57	1	31	0 (0%)
C	arrival_date...			August	2.45			0 (0%)
N	arrival_date...		26.92	27	0.51	1	53	0 (0%)
N	arrival_date...		2016.16	2016	0	2015	2017	0 (0%)
C	assigned_ro...			A	1.23			0 (0%)

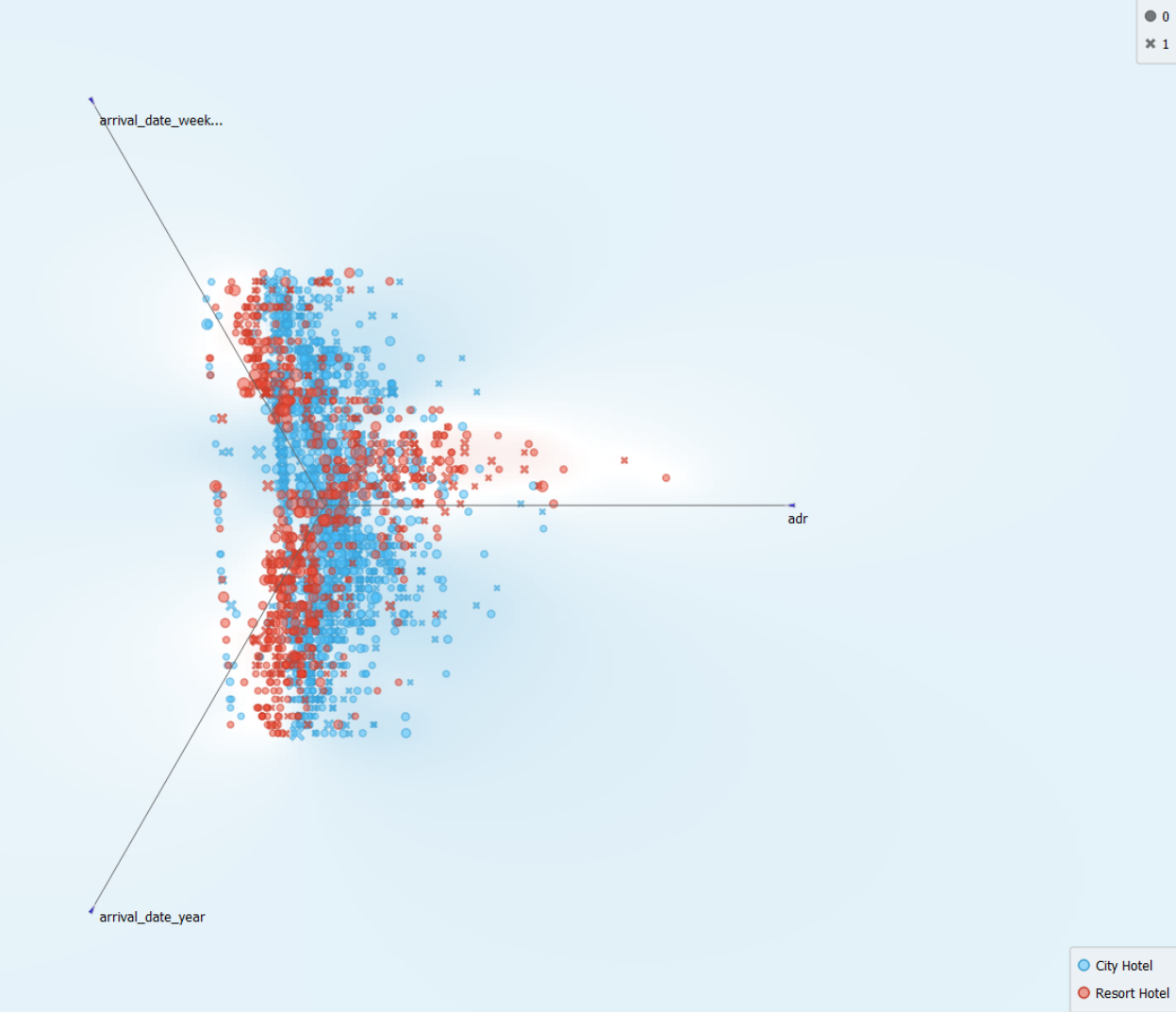


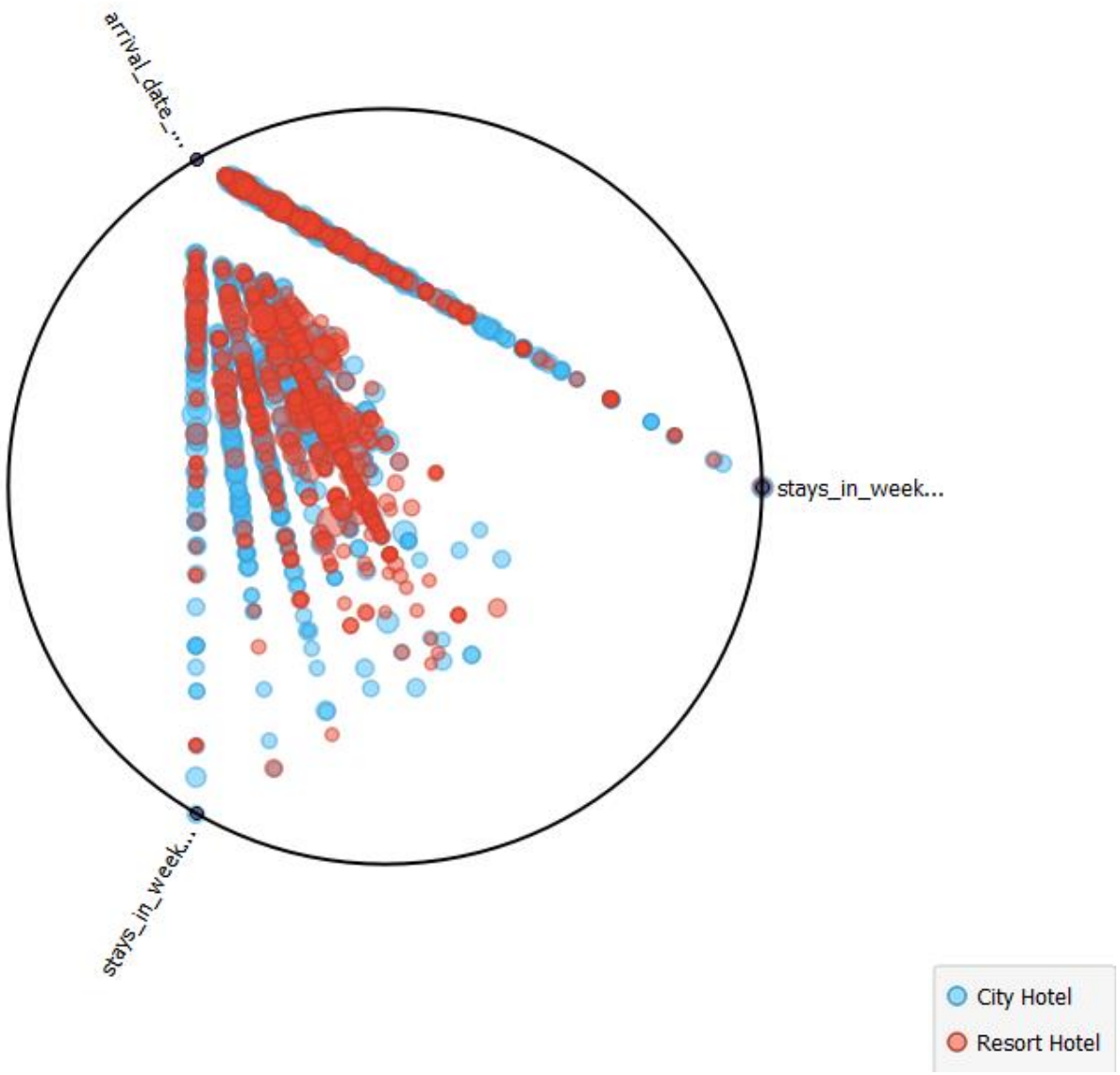

reserved_room_type
A
B
C
D
E
F
G
H
-100
0
100
200
300
400
500
adr

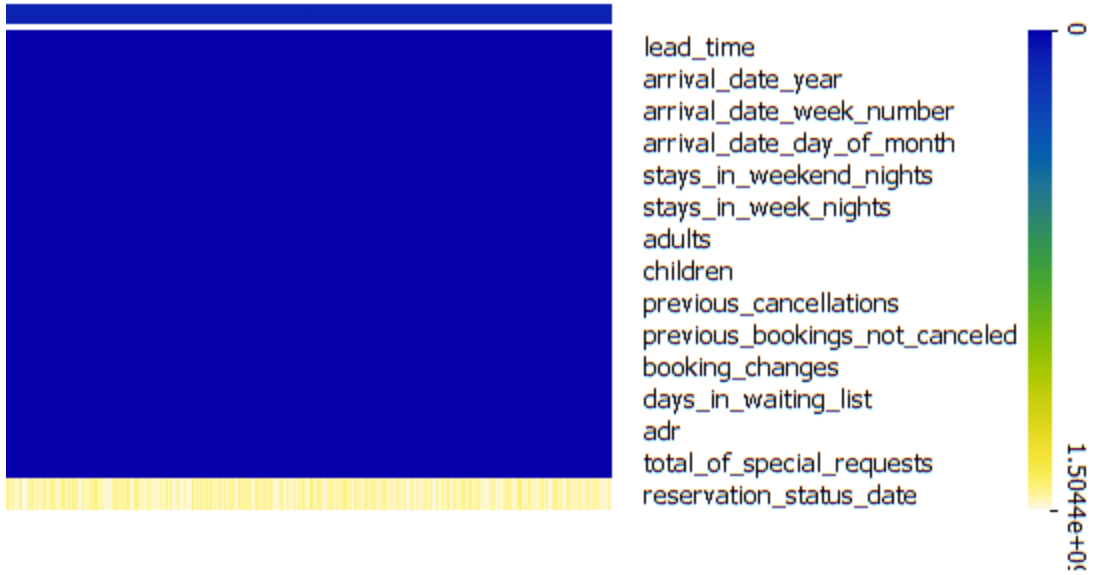

μ=15.61, σ=8.92
Frequency
0
10
20
30
40
50
60
70
80
90
100
110
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
arrival_date_day_of_month


customer_type
market_segment
Online TA
Offline TA/TO
Groups
Direct
Corporate
Complementary
Aviation
market_segment
C... ...
Transient
Transient-Party
customer_type
N = 2000
X²=513.32, p=0.000

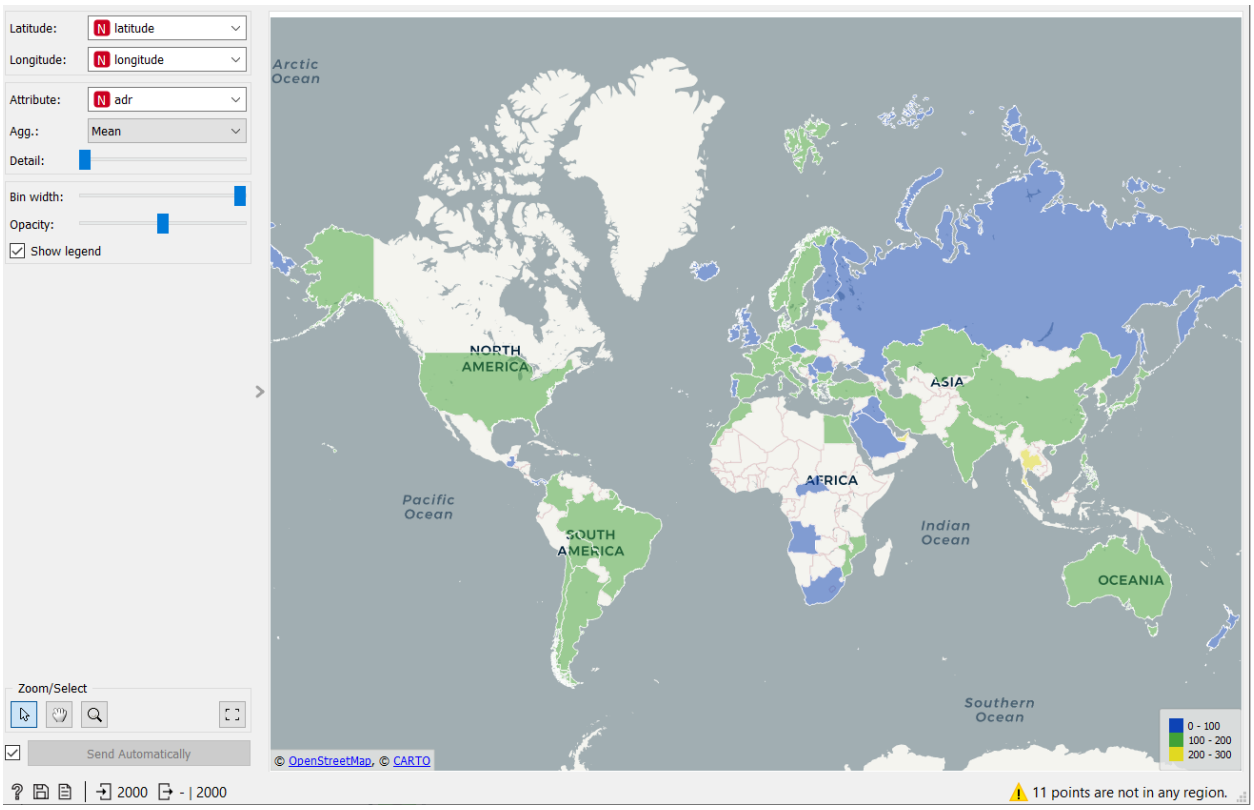

distribution_chan
market_segment
(None)
(None)
Find Informative Mosaics
Interior Coloring
(Pearson residuals)
Compare with total
Online TA
Offline TA/TO
Groups
Direct
Corporate
Complementary
Aviation
market_segment
Co...
Direct
...
TA/TO
distribution_channel
<-8
-8:-4
-4:-2
-2:2
2:4
4:8
>8


0
1
arrival_date_week...
adr
arrival_date_year
City Hotel
Resort Hotel


arrival_date_...
stays_in_week...
stays_in_week...
City Hotel
Resort Hotel

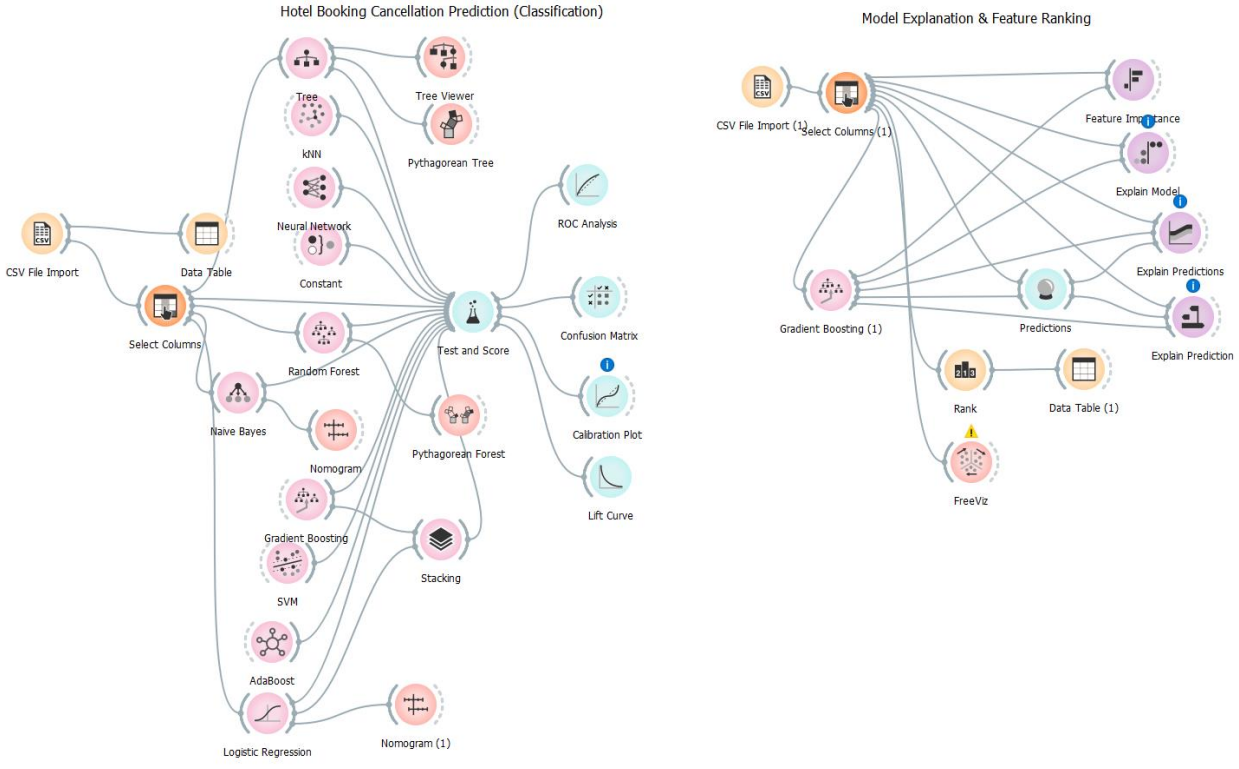

lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
previous_cancellations
previous_bookings_not_canceled
booking_changes
days_in_waiting_list
adr
total_of_special_requests
reservation_status_date
0
1.5044e+0

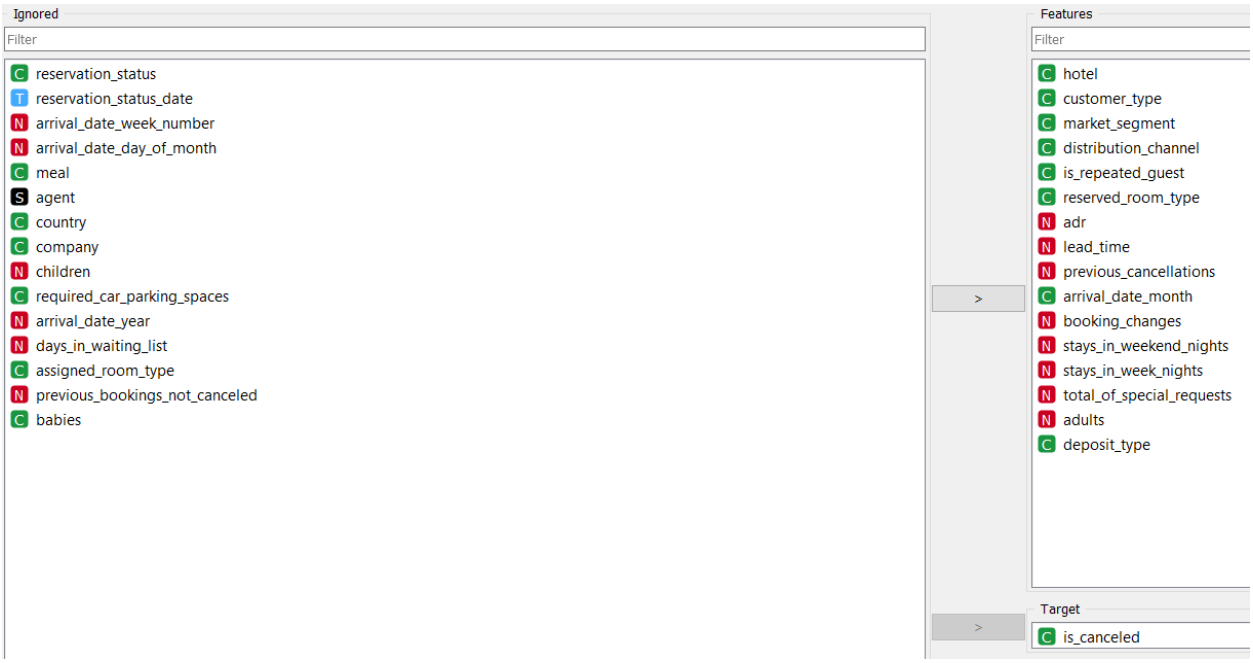

Baffin Bay
City Hotel
Resort Hotel
© OpenStreetMap contributors


Latitude:
latitude
Longitude:
longitude
Attribute:
adr
Agg.:
Mean
Detail:
Bin width:
Opacity:
Show legend
Zoom/Select
Send Automatically
2000
- | 2000
Arctic Ocean
NORTH AMERICA
ASIA
AFRICA
Pacific Ocean
SOUTH AMERICA
Indian Ocean
OCEANIA
Southern Ocean
0 - 100
100 - 200
200 - 300
© OpenStreetMap, © CARTO
11 points are not in any region.

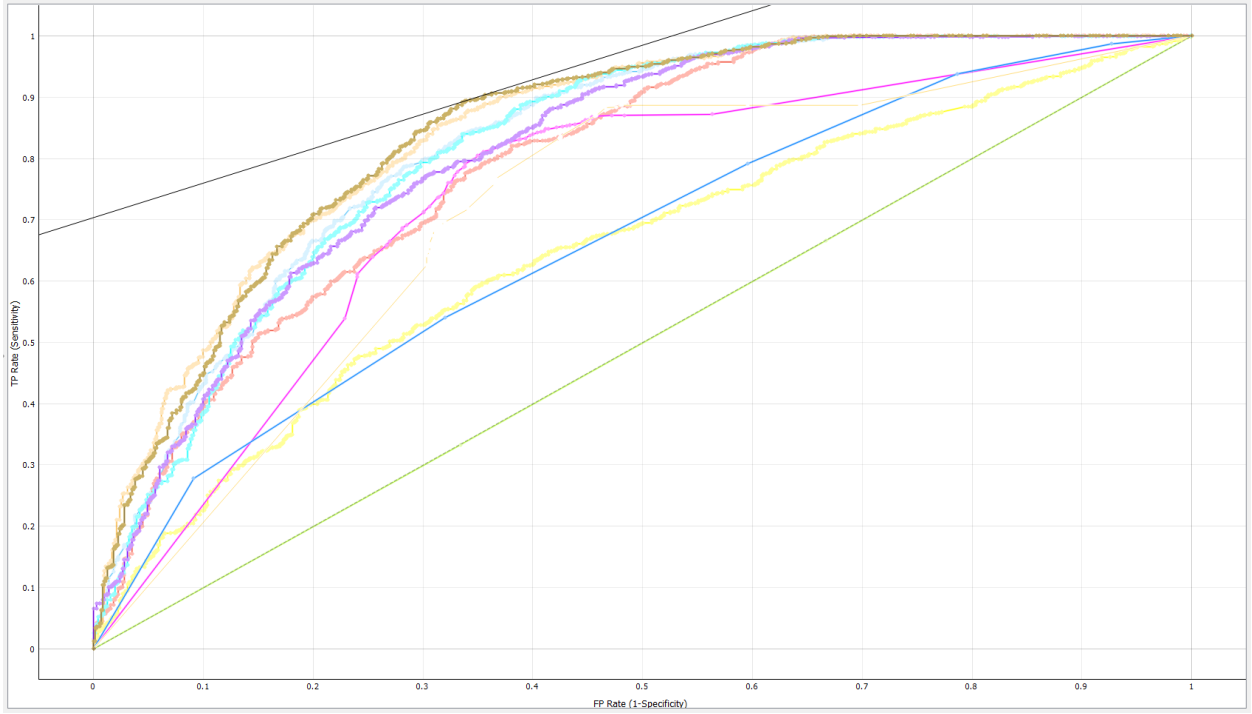
Classification

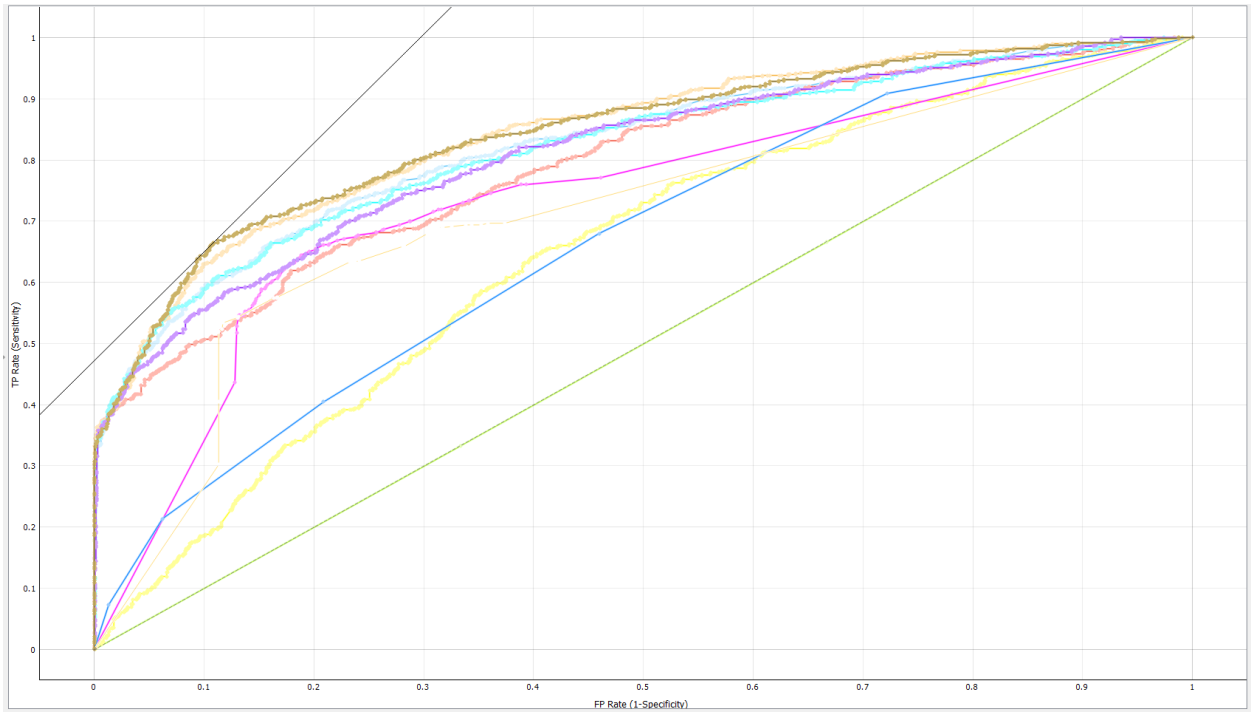
Various classification models were built to predict whether the guest would cancel a hotel booking. Not all of the variables in the dataset were used as sometimes less is more in terms of model training. The data was selected based on feature importance and what made sense at face value. The performance of eleven models were compared to one another based on AUC, accuracy, precision, recall and F1 score. The stacked model (gradient boosted trees + logistic regression) performed the best based on all metrics except for AUC. These to methods were used as gradient trees are more black-box and logistic regression is more interpretable which covers each method's strengths and weaknesses. The ROC curve visualizes the comparison of model performance. Deposit type, booking lead time, total number of special requests, and number of previous cancellations are the most important features for cancellation prediction.




Ignored
Filter
reservation_status
reservation_status_date
arrival_date_week_number
arrival_date_day_of_month
meal
agent
country
company
children
required_car_parking_spaces
arrival_date_year
days_in_waiting_list
assigned_room_type
previous_bookings_not_canceled
babies
>
>
Features
Filter
hotel
customer_type
market_segment
distribution_channel
is_repeated_guest
reserved_room_type
adr
lead_time
previous_cancellations
arrival_date_month
booking_changes
stays_in_weekend_nights
stays_in_week_nights
total_of_special_requests
adults
deposit_type
Target
is_canceled

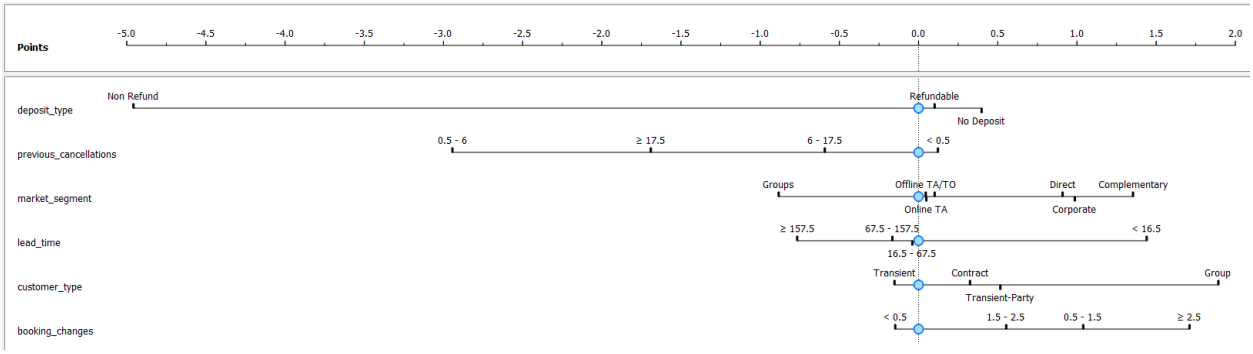
Model	AUC	CA	F1	Precision	Recall
Stack	0.844	0.806	0.800	0.806	0.806
Gradient Boosting	0.845	0.801	0.793	0.801	0.801
Logistic Regression	0.820	0.795	0.784	0.798	0.795
Random Forest	0.825	0.789	0.780	0.787	0.789
Neural Network	0.812	0.759	0.756	0.755	0.759
AdaBoost	0.737	0.752	0.751	0.750	0.752
Naive Bayes	0.792	0.750	0.742	0.744	0.750
Tree	0.712	0.742	0.737	0.736	0.742
kNN	0.662	0.653	0.642	0.639	0.653
SVM	0.647	0.630	0.630	0.629	0.630
Constant	0.499	0.644	0.505	0.415	0.644

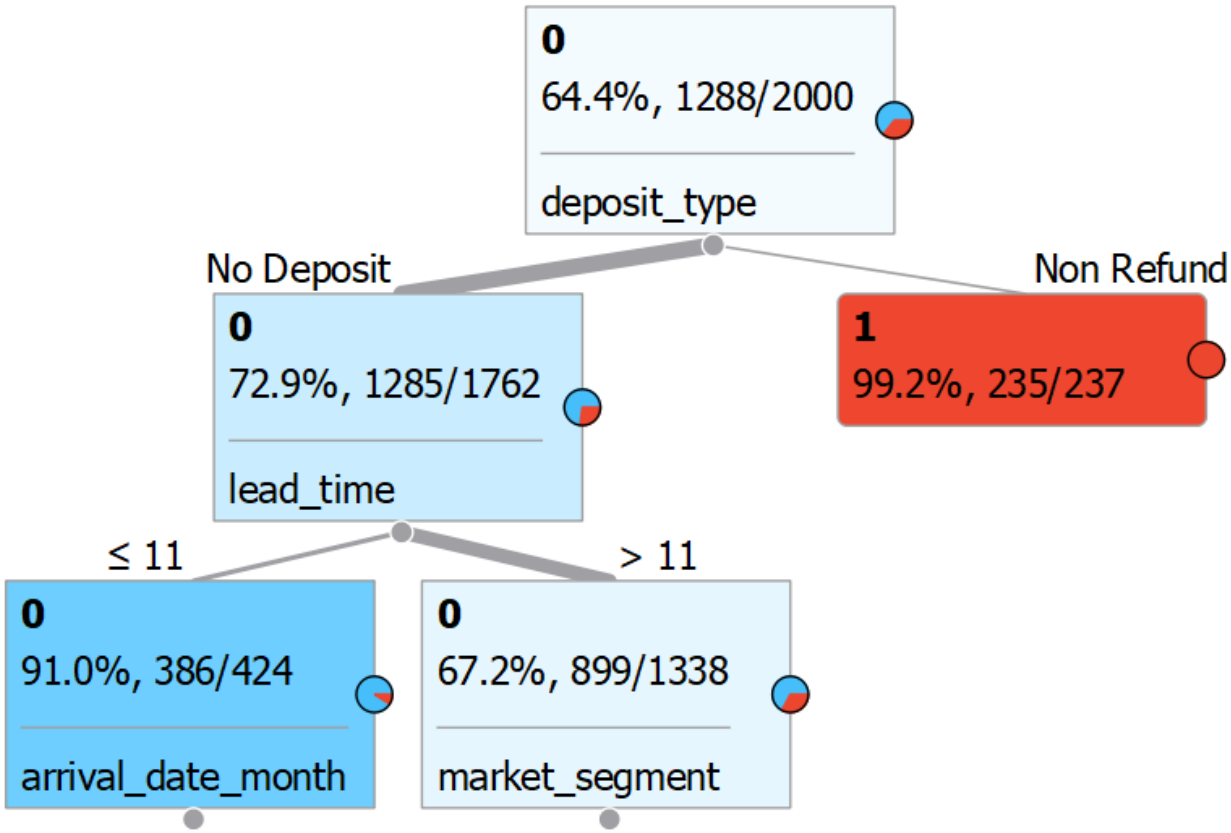


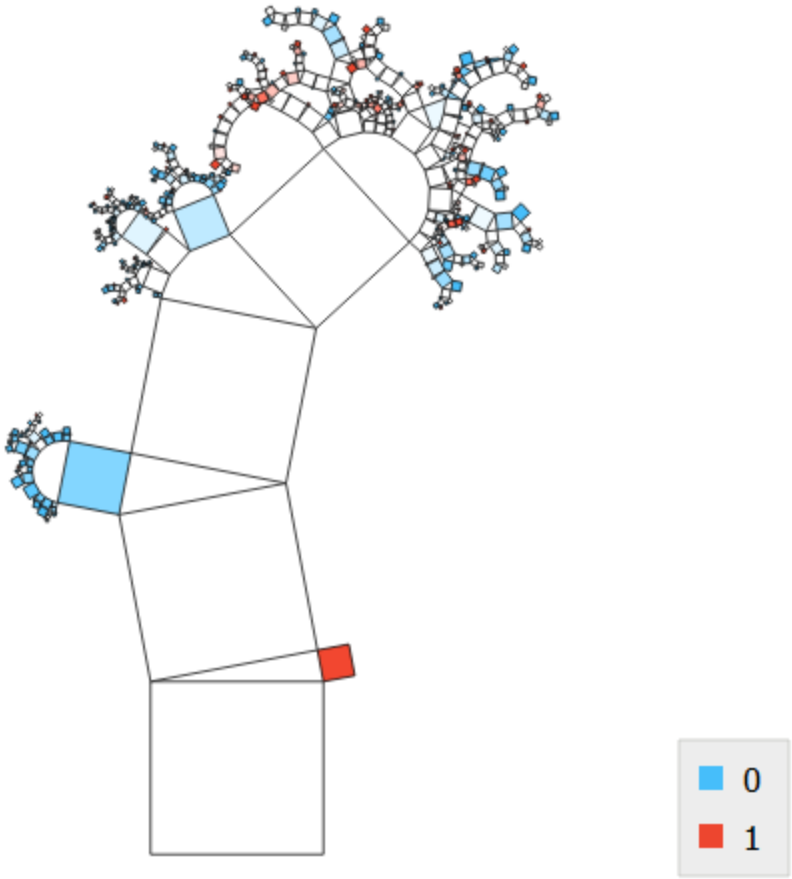


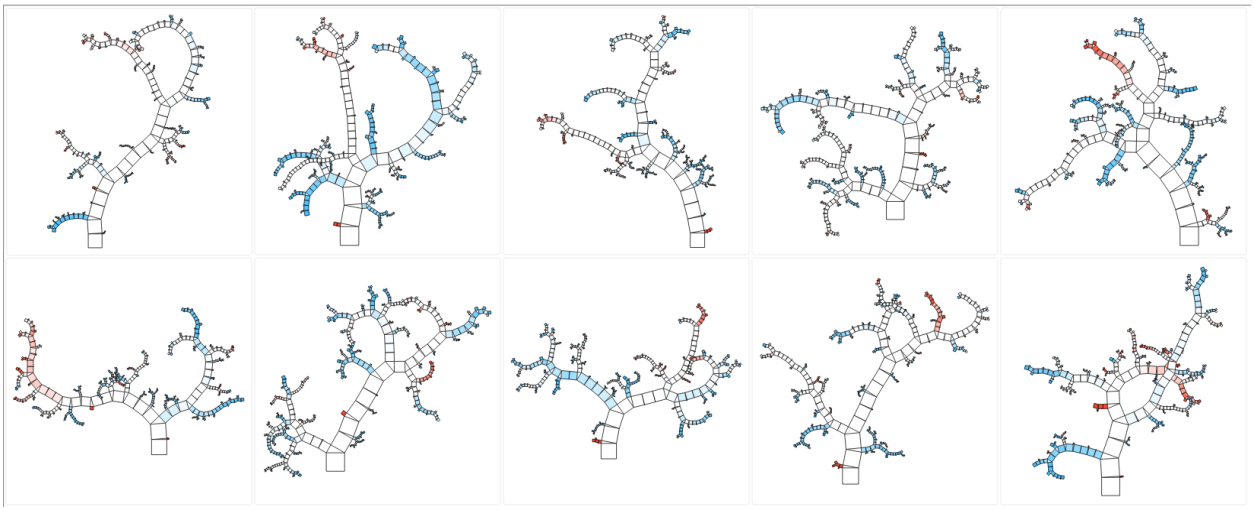
Predicted

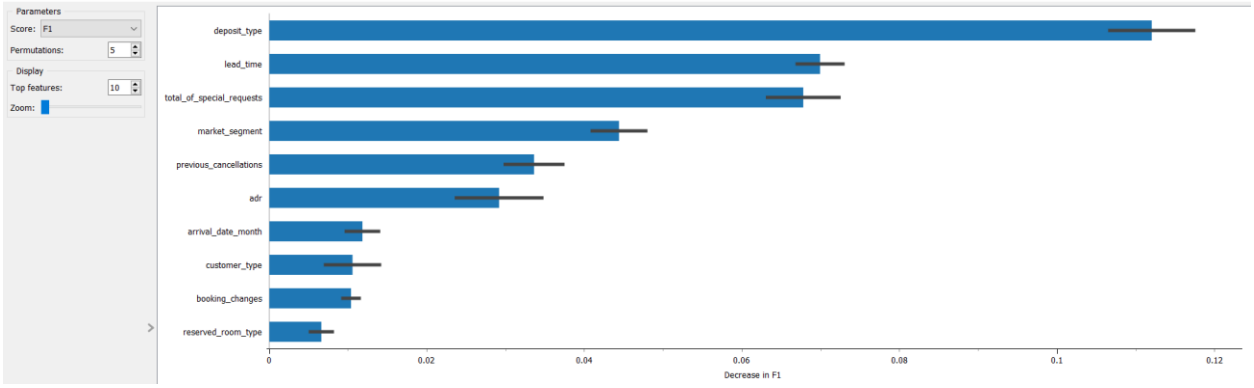
Actual	0	1	Σ
0	81.0 %	20.2 %	1288
1	19.0 %	79.8 %	712
Σ	1455	545	2000



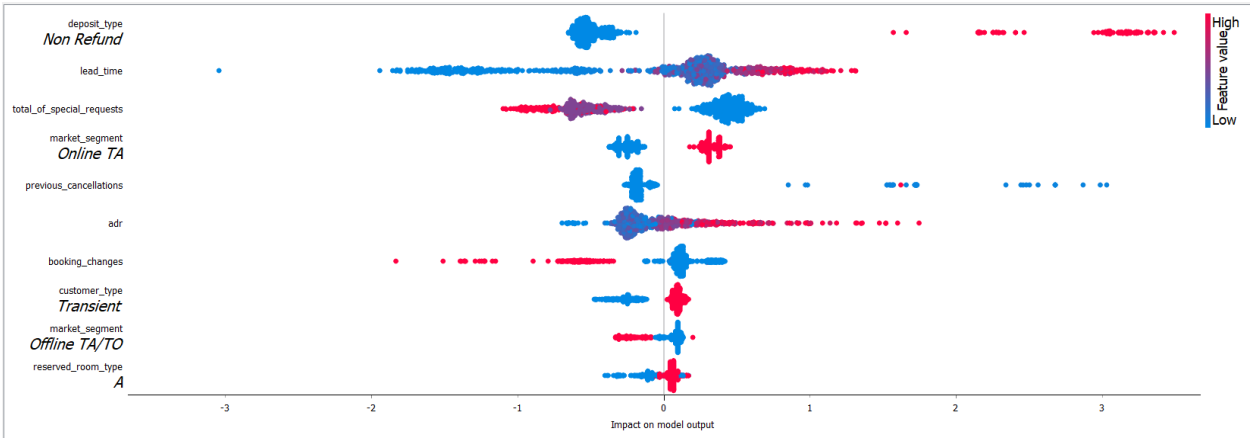

0
64.4%, 1288/2000
deposit_type
No Deposit
Non Refund
0
72.9%, 1285/1762
lead_time
1
99.2%, 235/237
≤ 11
> 11
0
91.0%, 386/424
arrival_date_month
0
67.2%, 899/1338
market_segment

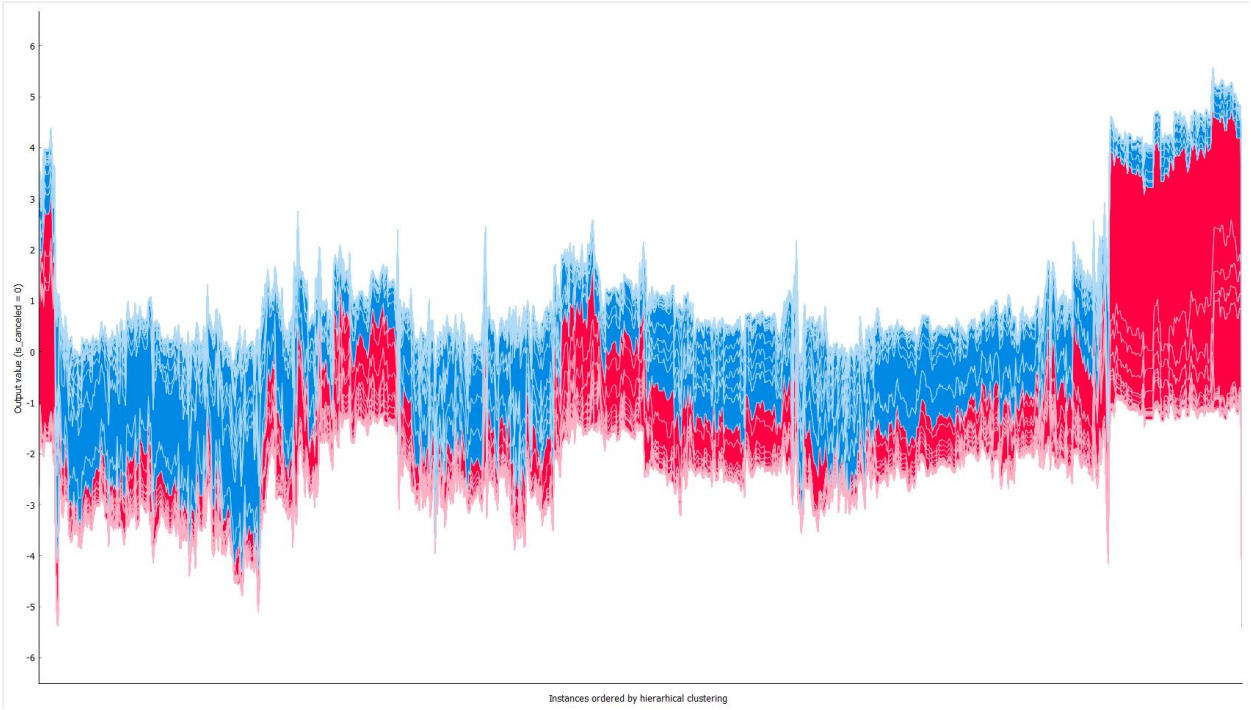

0
1

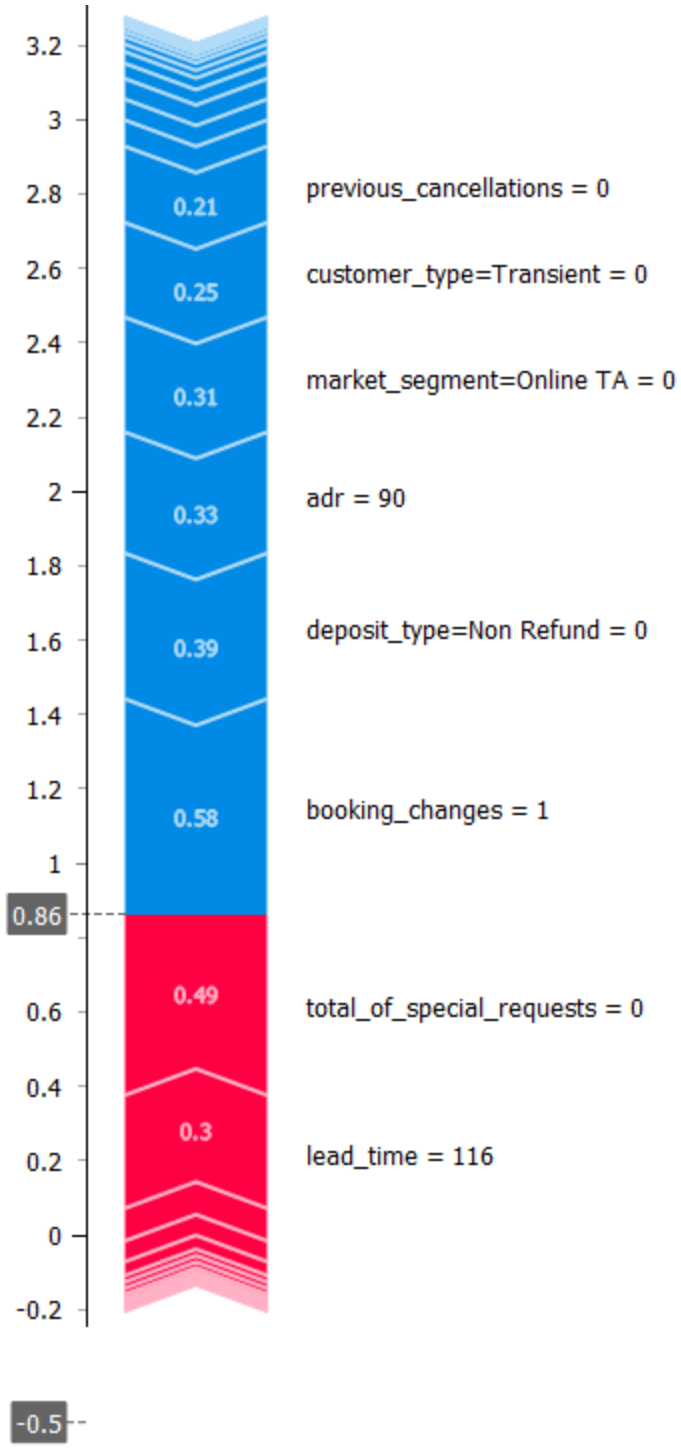



Parameters
Score: F1
Permutations: 5
Display
Top features: 10
Zoom:
deposit_type
lead_time
total_of_special_requests
market_segment
previous_cancellations
adr
arrival_date_month
customer_type
booking_changes
reserved_room_type
0
0.02
0.04
0.06
0.08
0.1
0.12
Decrease in F1

		#	Inf...ain	Gai...tio	Gini	χ^2	ReliefF	FCBF
1	C deposit_type	3	0.189	0.355	0.109	410.168	0.106	0.345
2	N lead_time		0.082	0.041	0.047	154.277	0.011	0.000
3	N total_of_special_requests		0.060	0.041	0.036	142.652	0.019	0.000
4	N previous_cancellations		0.050	0.175	0.031	133.870	0.002	0.000
5	C market_segment	7	0.043	0.022	0.026	1.068	0.098	0.000
6	N booking_changes		0.025	0.031	0.014	72.046	0.003	0.000
7	C distribution_channel	4	0.021	0.025	0.012	14.943	0.006	0.024
8	C customer_type	4	0.015	0.015	0.009	0.856	0.044	0.000
9	N stays_in_week_nights		0.009	0.005	0.006	4.235	0.004	0.000
10	C hotel	2	0.009	0.010	0.006	16.528	0.020	0.000
11	N adults		0.007	0.007	0.004	2.318	0.011	0.000
12	C is_repeated_guest	2	0.007	0.036	0.004	15.981	0.016	0.000
13	C arrival_date_month	12	0.006	0.002	0.004	7.306	0.004	0.000
14	C reserved_room_type	8	0.005	0.004	0.003	34.115	0.006	0.000
15	N stays_in_weekend_nights		0.002	0.001	0.001	2.169	-0.004	0.000
16	N adr		0.000	0.000	0.000	0.438	0.006	0.000



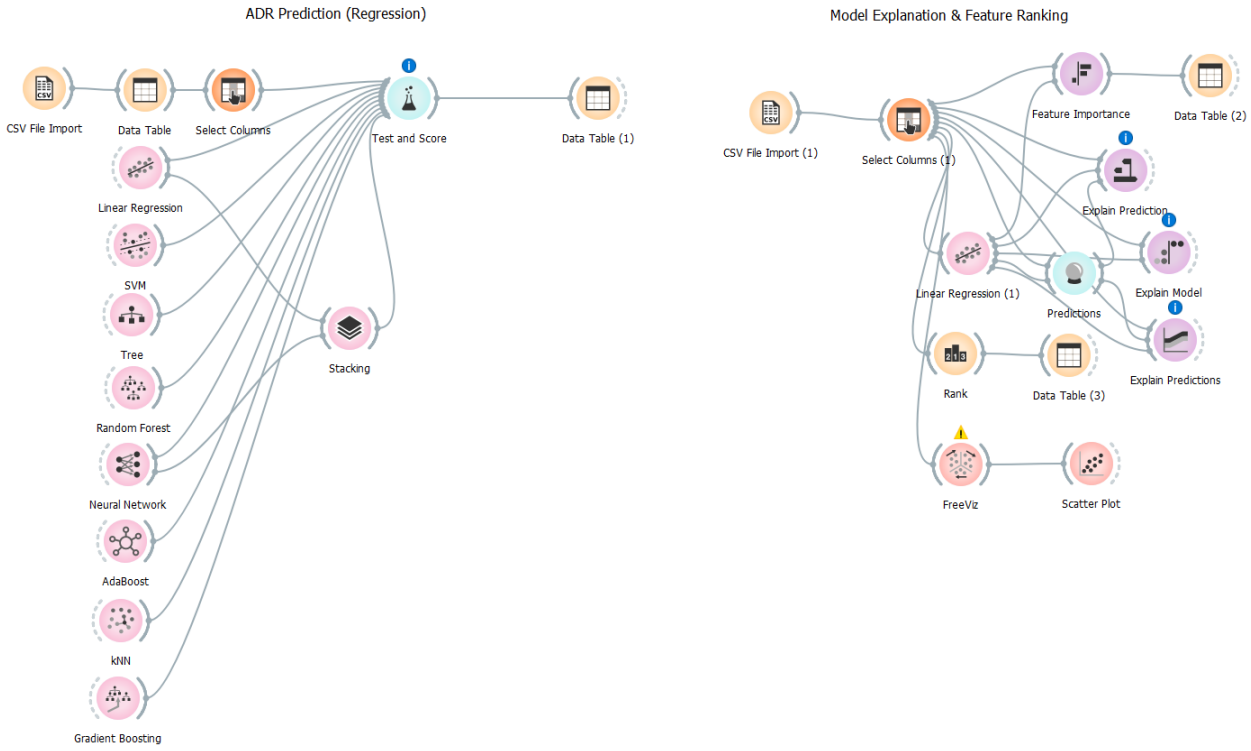

Output value (is_canceled = 0)
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
Instances ordered by hierarhical clustering

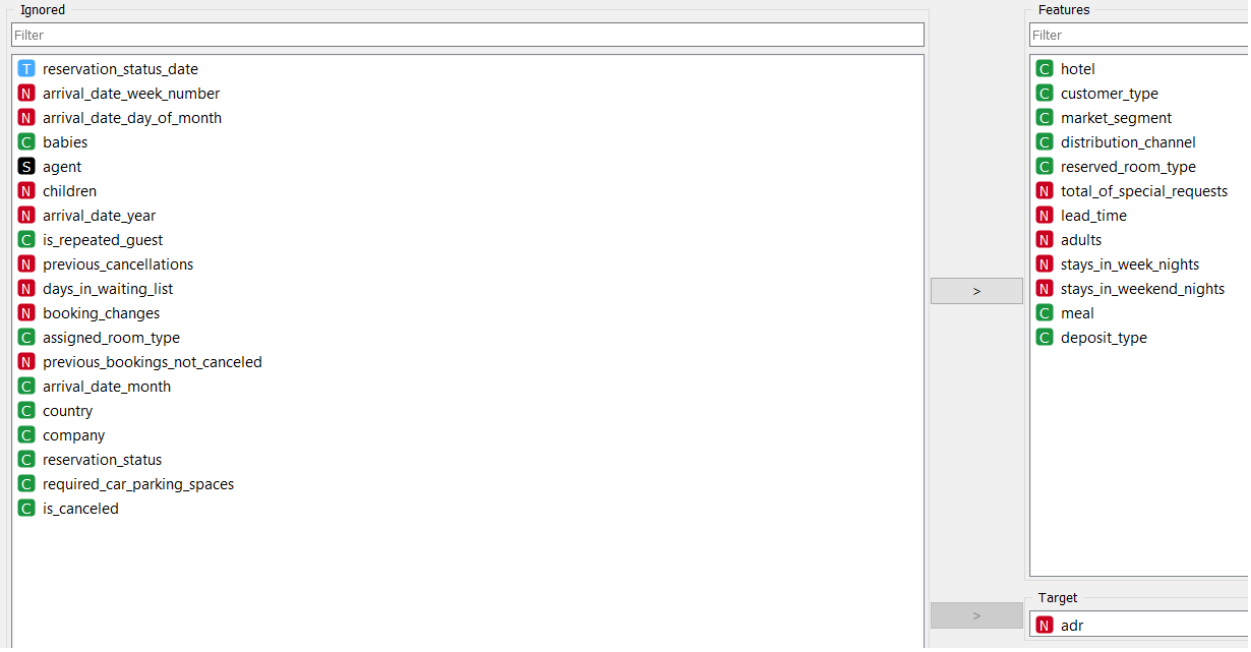


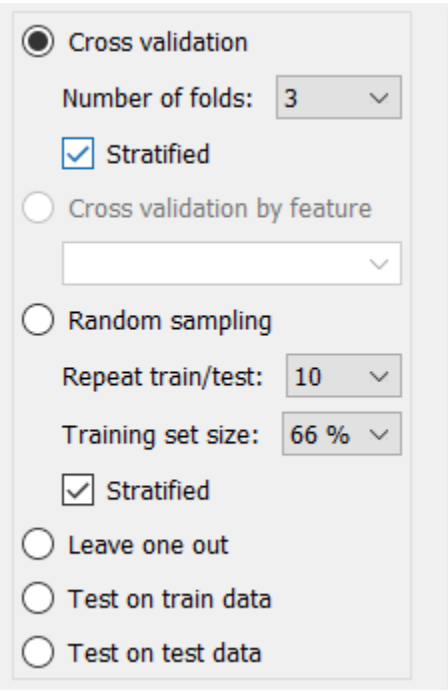
Regression

Regression models were built to predict the average daily rate of the hotel at the time of the reservations. MSE, RMSE, MAE, and R-squared were used to evaluate the performance of the

models. A lower value is best for all metrics except for r-squared where a higher value is a better indicator for predictive performance. Surprisingly, linear regression performed better than the other more complex models which show that complex does not always mean better. Arrival month, room type, the hotel type, market segment, booking lead time, and number of adults were the most important features for the prediction which make sense at face value.

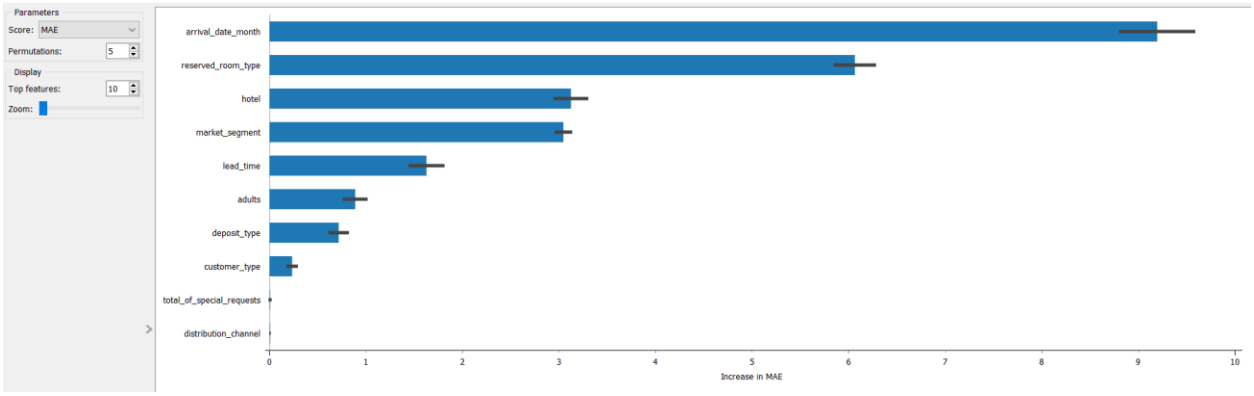


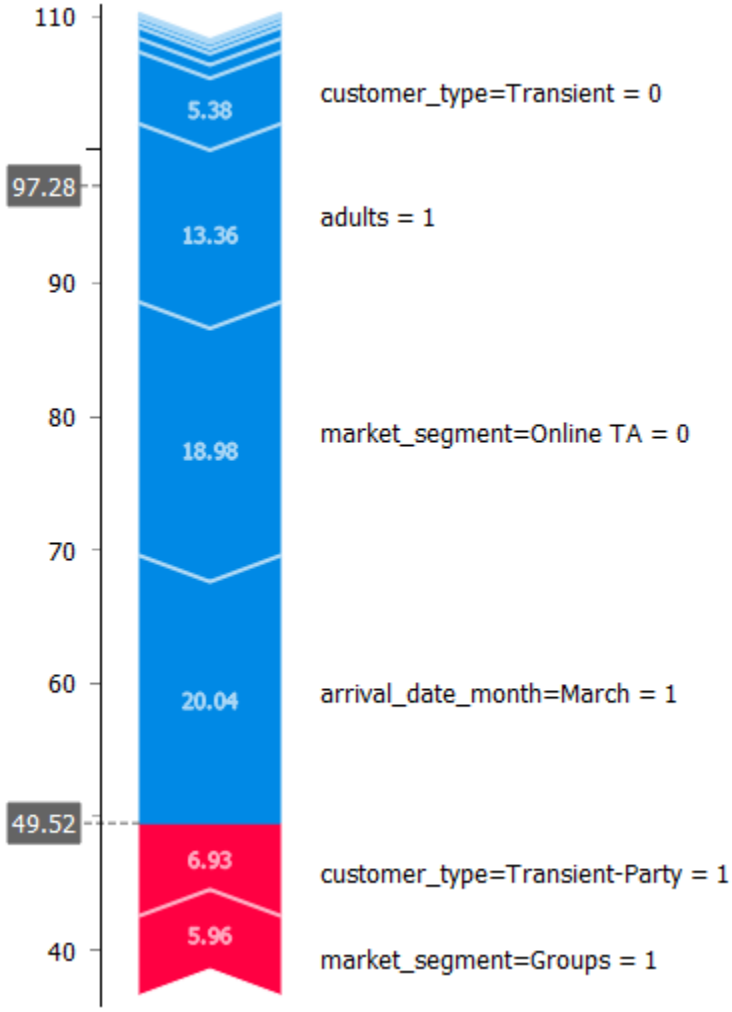


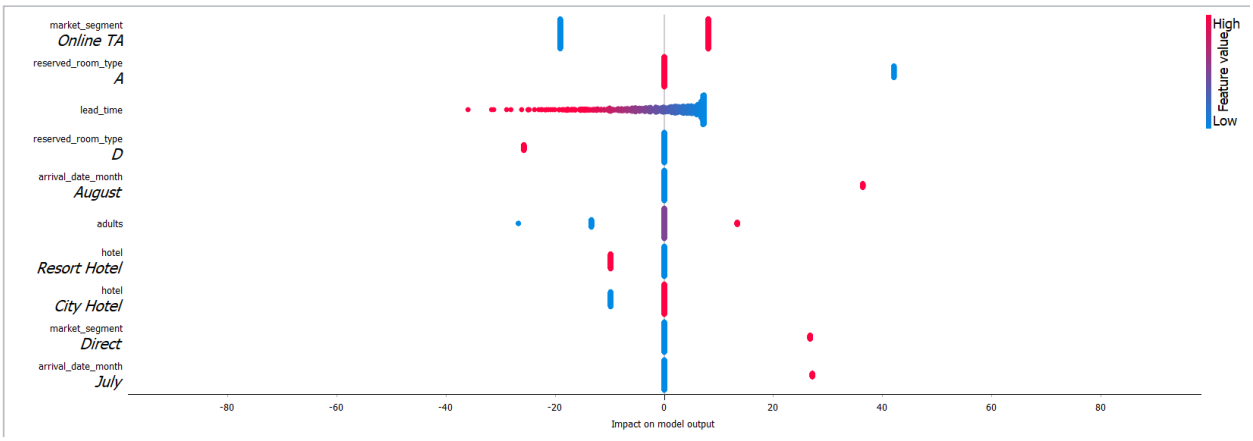


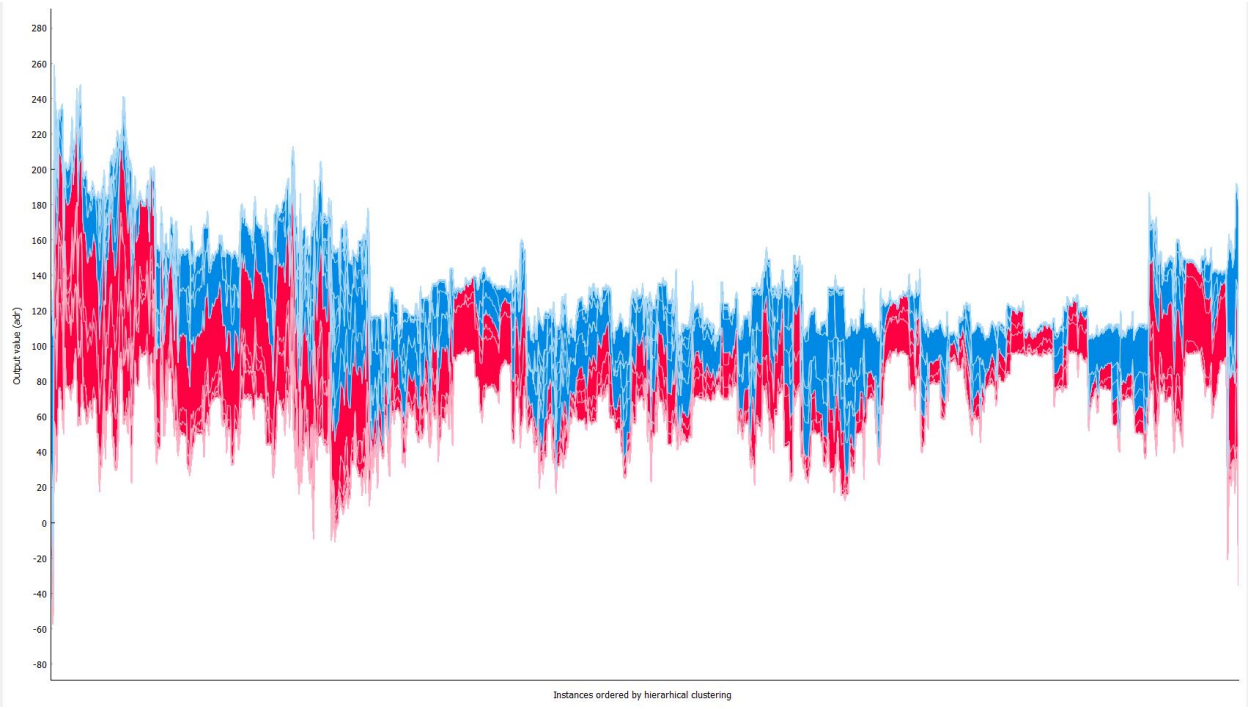
Model	MSE	RMSE	MAE	R2
Linear Regression	1647.981	40.595	29.697	0.311
Neural Network	1735.415	41.658	29.949	0.275
Random Forest	1886.495	43.434	30.480	0.212
Tree	2246.119	47.393	32.942	0.061
SVM	2667.299	51.646	37.150	-0.115

Model	MSE	RMSE	MAE	R2
Gradient Boosting	1477.564	38.439	27.400	0.382
Random Forest	1602.358	40.029	27.482	0.330
Stack	1424.048	37.737	27.498	0.405
Neural Network	1448.395	38.058	27.627	0.395
Linear Regression	1448.185	38.055	27.810	0.395
AdaBoost	1721.374	41.489	27.865	0.281
Tree	2365.176	48.633	32.683	0.011
kNN	2306.793	48.029	35.310	0.036
SVM	2392.658	48.915	36.020	-0.000


Parameters
Score: MAE
Permutations: 5
Display
Top features: 10
Zoom:
arrival_date_month
reserved_room_type
hotel
market_segment
lead_time
adults
deposit_type
customer_type
total_of_special_requests
distribution_channel
0
1
2
3
4
5
6
7
8
9
10
Increase in MAE


110
97.28
90
80
70
60
49.52
40
5.38
13.36
18.98
20.04
6.93
5.96
customer_type=Transient = 0
adults = 1
market_segment=Online TA = 0
arrival_date_month=March = 1
customer_type=Transient-Party = 1
market_segment=Groups = 1

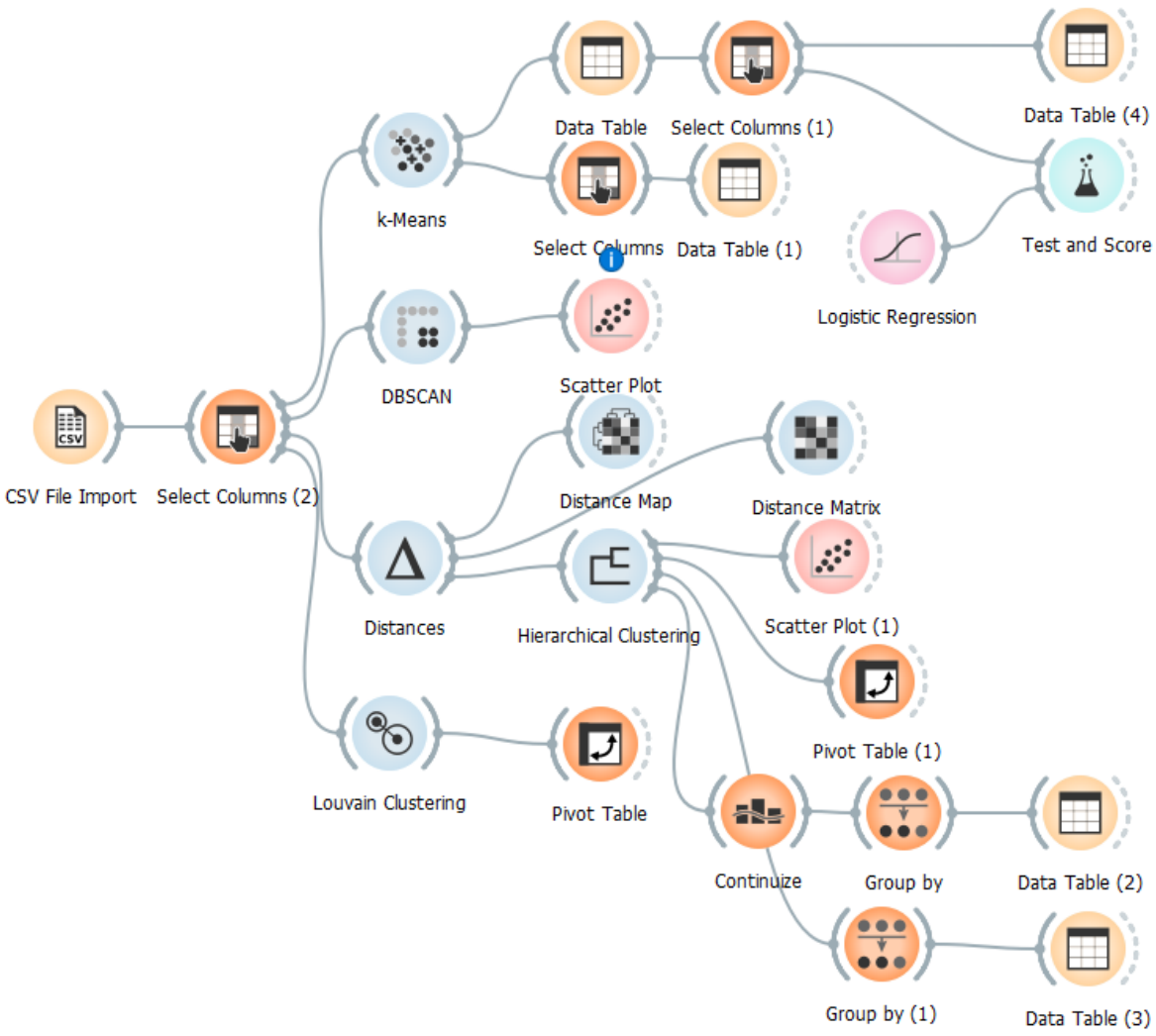


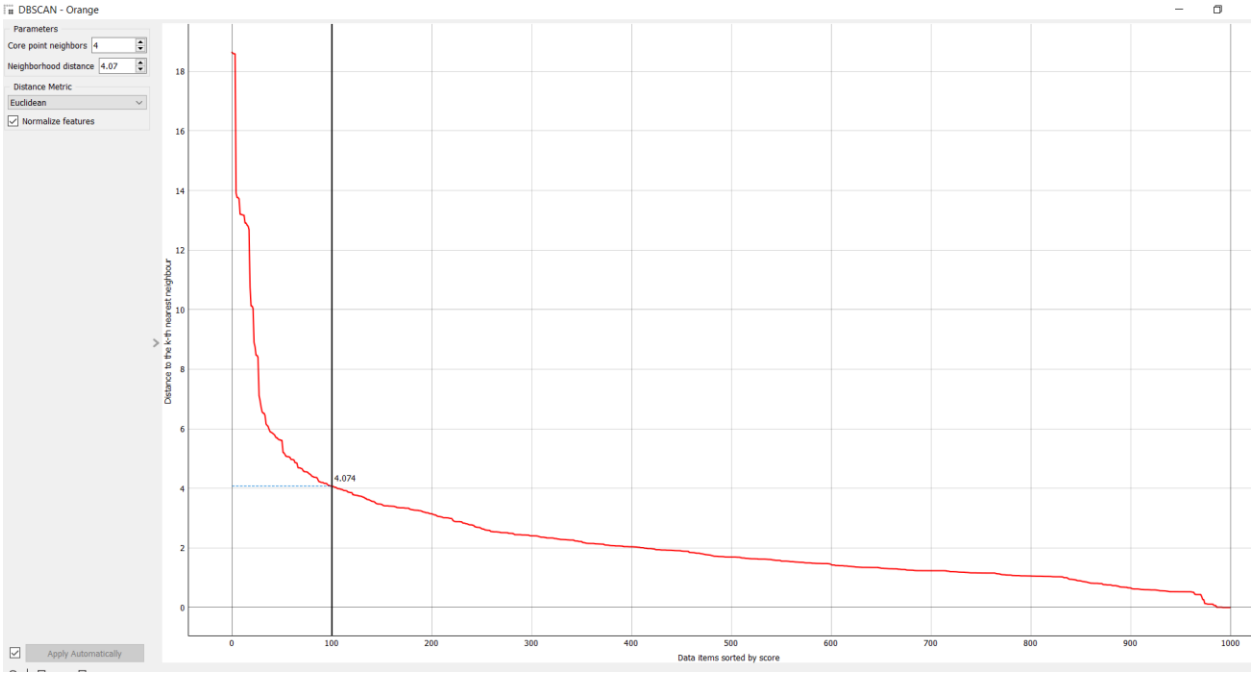


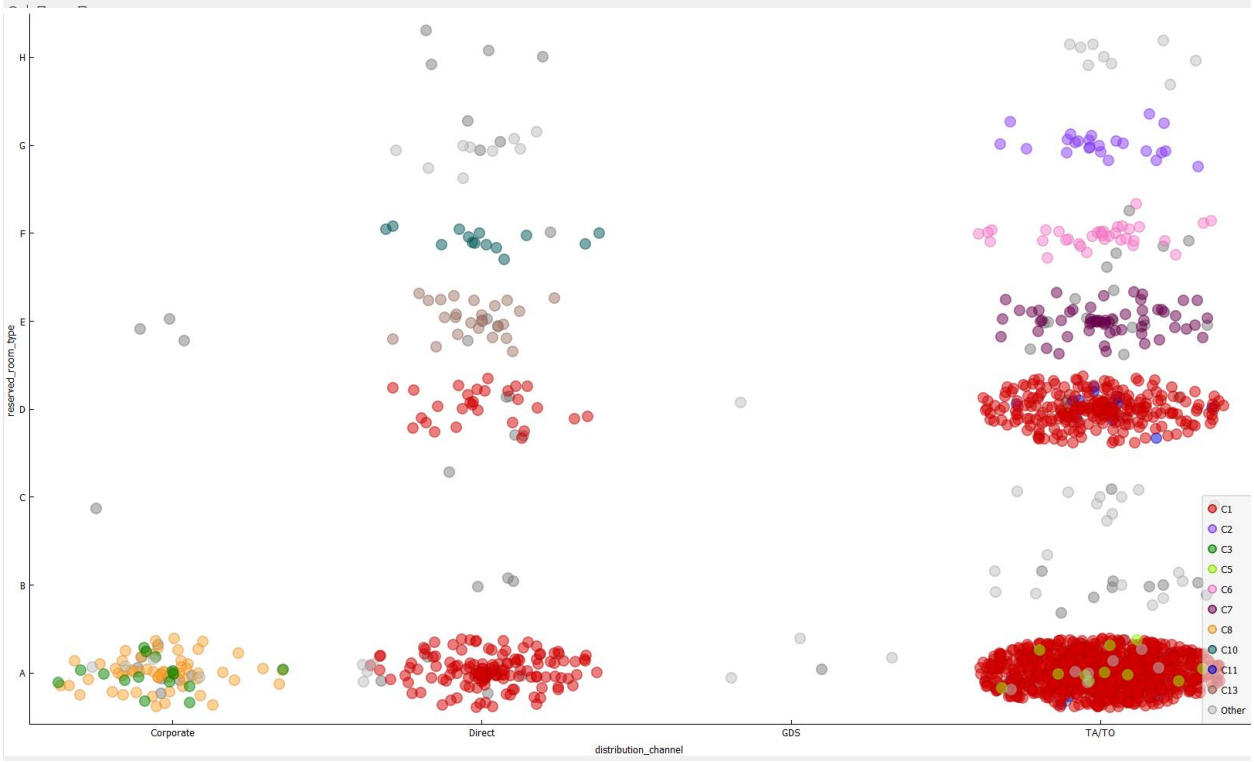
Clustering

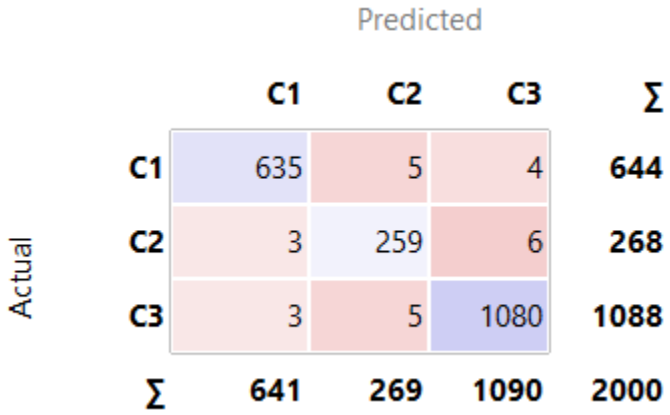
Market segmentation is very important in hospitality revenue management. Although there are industry-specific market segments that can be included in the data - there may be more specialized segments that can be derived from the hotel's unique sample of guests relative to the competition. Therefore, clustering methods were used to see if we can group together certain rows to create meaningful segments.

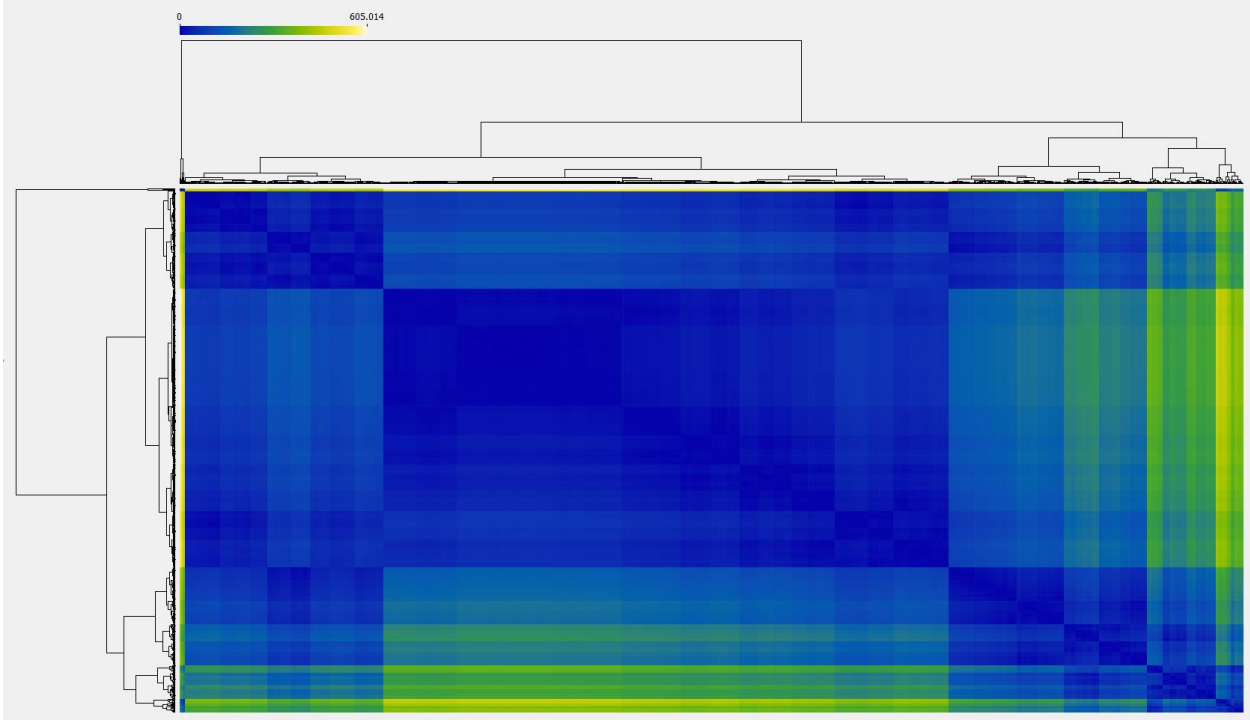
Clustering


Data Table
Select Columns (1)
Data Table (4)
k-Means
Select Columns
Data Table (1)
Logistic Regression
Test and Score
DBSCAN
Scatter Plot
CSV File Import
Select Columns (2)
Distance Map
Distance Matrix
Distances
Hierarchical Clustering
Scatter Plot (1)
Pivot Table (1)
Louvain Clustering
Pivot Table
Continuize
Group by
Data Table (2)
Group by (1)
Data Table (3)

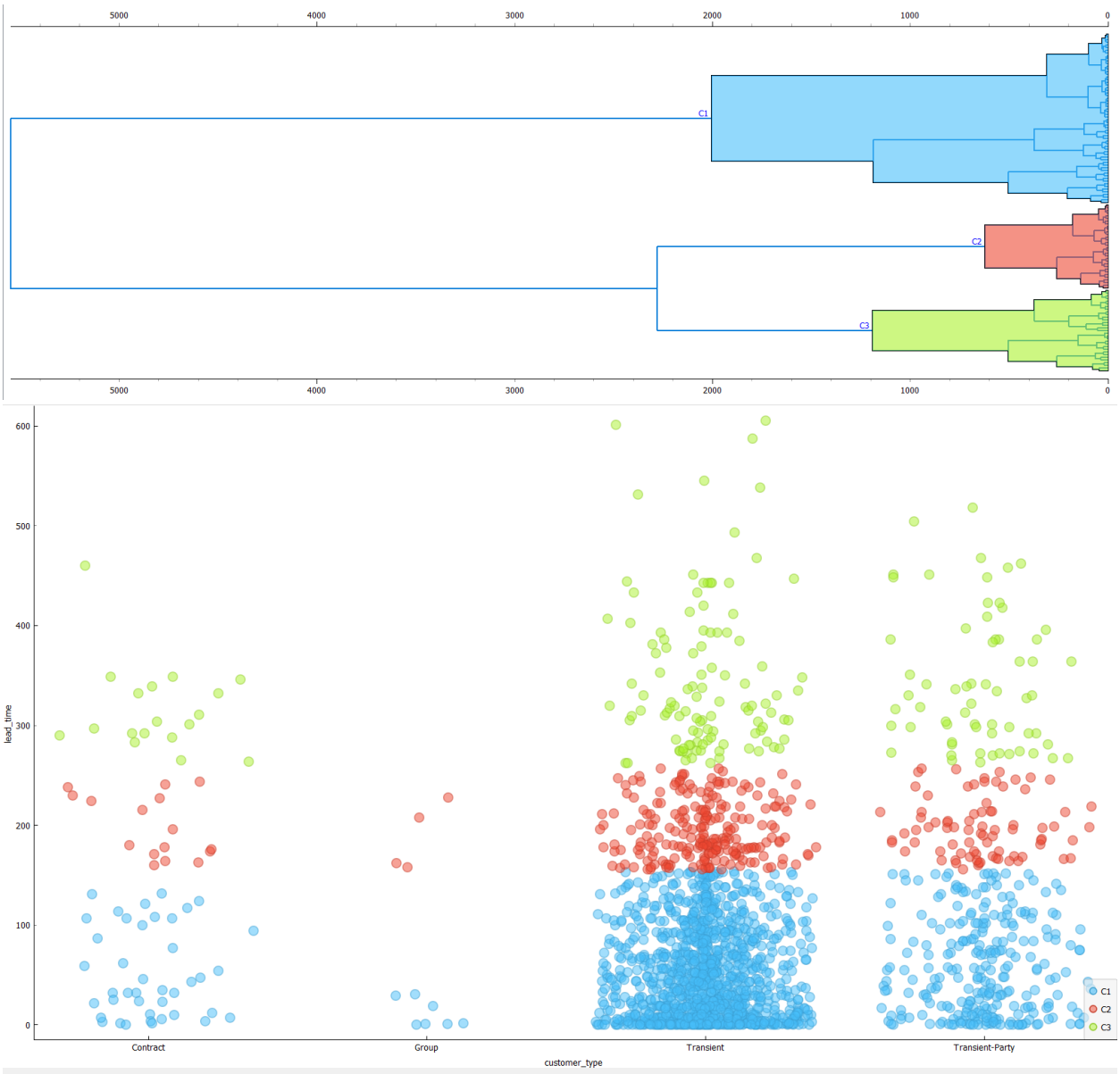

DBSCAN - Orange
Parameters
Core point neighbors 4
Neighborhood distance 4.07
Distance Metric
Euclidean
Normalize features
Apply Automatically
Distance to the k-th nearest neighbour
Data items sorted by score
4.074


reserved_room_type
distribution_channel
Corporate
Direct
GDS
TA/TO
C1
C2
C3
C5
C6
C7
C8
C10
C11
C13
Other


Predicted
Actual
C1
C2
C3
∑
C1
635
5
4
644
C2
3
259
6
268
C3
3
5
1080
1088
∑
641
269
1090
2000

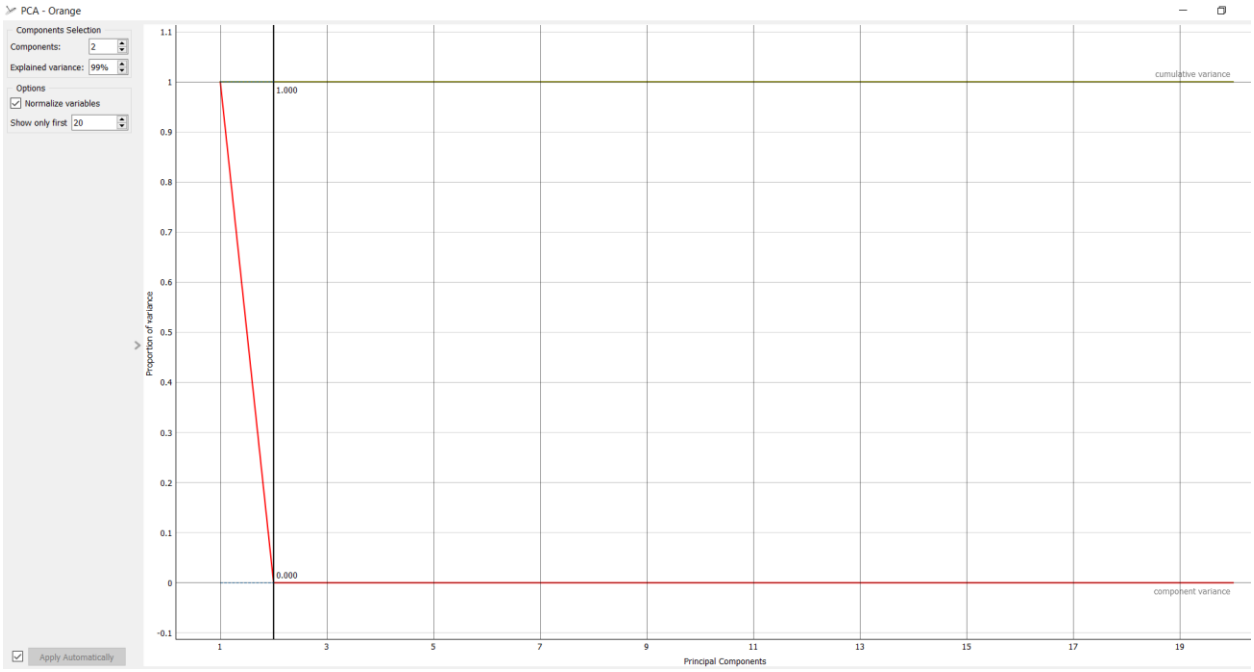


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1		3.742	50.120	99.040	87.046	114.018	15.297	32.156	81.019	83.042	58.052	70.021	122.020	51.069	108.032	108.023	42.059	5.477	61.041	50.249	113.040	51.098	105.024	25.417	92.033	60.033	57.114	13.191	155.013	116.039	67.067	110.050
2	3.742		47.053	96.016	90.039	111.032	18.221	35.071	84.036	80.012	61.025	67.052	125.032	48.042	105.043	105.010	39.051	8.124	58.000	53.151	110.018	48.062	102.020	28.213	89.017	57.026	54.056	10.198	158.016	113.027	64.016	113.018
3	50.120	47.053		49.020	137.047	64.109	65.146	82.037	131.065	33.045	108.046	20.396	172.035	3.162	58.086	58.078	8.660	55.082	11.225	100.065	63.103	3.742	55.100	75.067	42.107	10.392	7.937	37.148	205.020	66.129	17.321	160.012
4	99.040	96.016	49.020		186.032	15.395	114.075	131.015	180.036	16.031	157.025	29.206	221.020	48.073	9.434	9.381	57.079	104.034	38.039	149.037	14.387	48.114	6.708	124.044	7.550	39.077	42.178	86.052	254.016	17.436	32.140	209.012
5	87.046	90.039	137.047	186.032		201.025	72.049	55.055	6.557	170.038	29.069	157.035	35.100	138.018	195.023	195.018	129.019	82.043	148.024	37.202	200.037	138.014	192.029	62.073	179.017	147.034	144.031	100.065	68.059	203.032	154.029	23.130
6	114.018	111.032	64.109	15.395	201.025		129.019	146.048	195.008	31.145	172.023	44.034	236.019	63.071	6.928	6.557	72.035	119.029	53.066	164.094	2.828	63.087	9.165	139.097	22.181	54.056	57.096	101.025	269.007	3.000	47.096	224.025
7	15.297	18.221	65.146	114.075	72.049	129.019		17.436	66.030	98.071	43.058	85.047	107.056	66.061	123.069	123.024	57.035	10.296	76.053	35.398	128.031	66.061	120.033	11.136	107.033	75.060	72.042	28.142	140.032	131.023	82.037	95.063
8	32.156	35.071	82.037	131.015	55.055	146.048	17.436		49.132	115.030	26.153	102.069	90.039	83.018	140.021	140.025	74.047	27.129	93.027	18.193	145.059	83.030	137.040	7.348	124.032	92.043	89.084	45.144	123.041	148.057	99.056	78.019
9	81.019	84.036	131.065	180.036	6.557	195.008	66.030	49.132		164.030	23.108	151.007	41.097	132.030	189.037	189.016	123.016	76.039	142.021	31.448	194.015	132.034	186.011	56.232	173.020	141.025	138.036	94.032	74.047	197.010	148.020	29.206
10	83.042	80.012	33.045	16.031	170.038	31.145	98.071	115.030	164.030		141.025	13.342	205.024	32.125	25.219	25.120	41.097	88.034	22.045	133.060	30.100	32.187	22.136	108.074	9.381	23.108	26.230	70.043	238.015	33.151	16.186	193.016
11	58.052	61.025	108.046	157.025	29.069	172.023	43.058	26.153	23.108	141.025		128.035	64.086	109.032	166.033	166.015	100.025	53.047	119.013	8.888	171.020	109.028	163.028	33.196	150.020	118.034	115.030	71.063	97.031	174.026	125.012	52.067
12	70.021	67.052	20.396	29.206	157.035	44.034	85.047	102.069	151.007	13.342	128.035		192.021	19.235	38.171	38.092	28.107	75.047	9.381	120.129	43.081	19.313	35.043	95.153	22.181	10.296	13.491	57.044	225.016	46.054	4.243	180.036
13	122.020	125.032	172.035	221.020	35.100	236.019	107.056	90.039	41.097	205.024	64.086	192.021		173.012	230.017	230.013	164.012	117.026	183.022	72.153	235.034	173.026	227.018	97.082	214.012	182.014	179.039	135.030	33.136	238.029	189.037	12.369
14	51.069	48.042	3.162	48.073	138.018	63.071	66.061	83.018	132.030	32.125	109.032	19.235	173.012		57.061	57.018	9.110	56.018	10.198	101.079	62.097	2.000	54.056	76.066	41.037	9.165	6.633	38.079	206.017	65.077	16.248	161.012
15	108.032	105.043	58.086	9.434	195.023	6.928	123.069	140.021	189.037	25.219	166.033	38.171	230.017	57.061		3.000	66.076	113.040	47.096	158.054	6.782	57.061	4.690	133.041	16.310	48.083	51.186	95.068	263.013	9.220	41.207	218.021
16	108.023	105.010	58.078	9.381	195.018	6.557	123.024	140.025	189.016	25.120	166.015	38.092	230.013	57.018	3.000		66.015	113.000	47.021	158.070	5.568	57.053	3.606	133.053	16.031	48.010	51.039	95.011	263.010	8.367	41.049	218.014
17	42.059	39.051	8.660	57.079	129.019	72.035	57.035	74.047	123.016	41.097	100.025	28.107	164.012	9.110	66.076	66.015		47.021	19.105	92.119	71.056	9.274	63.032	67.119	50.030	18.083	15.199	29.069	197.013	74.041	25.120	152.020
18	5.477	8.124	55.082	104.034	82.043	119.029	10.296	27.129	76.039	88.034	53.047	75.047	117.026	56.018	113.040	113.000	47.021		66.015	45.244	118.025	56.054	110.018	20.347	97.005	65.008	62.032	18.055	150.017	121.025	72.028	105.029
19	61.041	58.000	11.225	38.039	148.024	53.066	76.053	93.027	142.021	22.045	119.013	9.381	183.022	10.198	47.096	47.021	19.105	66.015		111.072	52.038	10.296	44.045	86.070	31.048	2.000	4.690	48.042	216.012	55.055	6.164	171.012
20	50.249	53.151	100.065	149.037	37.202	164.094	35.398	18.193	31.448	133.060	8.888	120.129	72.153	101.079	158.054	158.070	92.119	45.244	111.072		163.098	101.079	155.097	25.159	142.081	110.114	107.140	63.253	105.119	166.108	117.094	60.100
21	113.040	110.018	63.103	14.387	200.037	2.828	128.031	145.059	194.015	30.100	171.020	43.081	235.034	62.097	6.782	5.568	71.056	118.025	52.038	163.098		62.113	8.246	138.101	21.166	53.066	56.054	100.020	268.017	3.317	46.022	223.027
22	51.098	48.062	3.742	48.114	138.014	63.087	66.061	83.030	132.034	32.187	109.028	19.313	173.026	2.000	57.061	57.053	9.274	56.054	10.296	101.079	62.113		54.093	76.066	41.085	9.487	6.782	38.158	206.024	65.092	16.248	161.019
23	105.024	102.020	55.100	6.708	192.029	9.165	120.033	137.040	186.011	22.136	163.028	35.043	227.018	54.056	4.690	3.606	63.032	110.018	44.045	155.097	8.246	54.093		130.092	13.191	45.033	48.083	92.011	260.010	11.091	38.079	215.019
24	25.417	28.213	75.067	124.044	62.073	139.097	11.136	7.348	56.232	108.074	33.196	95.153	97.082	76.066	133.041	133.053	67.119	20.347	86.070	25.159	138.101	76.066	130.092		117.056	85.100	82.110	38.314	130.065	141.106	92.098	85.041
25	92.033	89.017	42.107	7.550	179.017	22.181	107.033	124.032	173.020	9.381	150.020	22.181	214.012	41.037	16.310	16.031	50.030	97.005	31.048	142.081	21.166	41.085	13.191	117.056		32.031	35.043	79.019	247.012	24.145	25.100	202.015
26	60.033	57.026	10.392	39.077	147.034	54.056	75.060	92.043	141.025	23.108	118.034	10.296	182.014	9.165	48.083	48.010	18.083	65.008	2.000	110.114	53.066	9.487	45.033	85.100	32.031		4.000	47.011	215.009	56.062	7.483	170.021
27	57.114	54.056	7.937	42.178	144.031	57.096	72.042	89.084	138.036	26.230	115.030	13.491	179.039	6.633	51.186	51.039	15.199	62.032	4.690	107.140	56.054	6.782	48.083	82.110	35.043	4.000		44.068	212.021	59.051	10.198	167.027
28	13.191	10.198	37.148	86.052	100.065	101.025	28.142	45.144	94.032	70.043	71.063	57.044	135.030	38.079	95.068	95.011	29.069	18.055	48.042	63.253	100.020	38.158	92.011	38.314	79.019	47.011	44.068		168.015	103.019	54.056	123.041
29	155.013	158.016	205.020	254.016	68.059	269.007	140.032	123.041	74.047	238.015	97.031	225.016	33.136	206.017	263.013	263.010	197.013	150.017	216.012	105.119	268.017	206.024	260.010	130.065	247.012	215.009	212.021	168.015		271.017	222.020	45.078

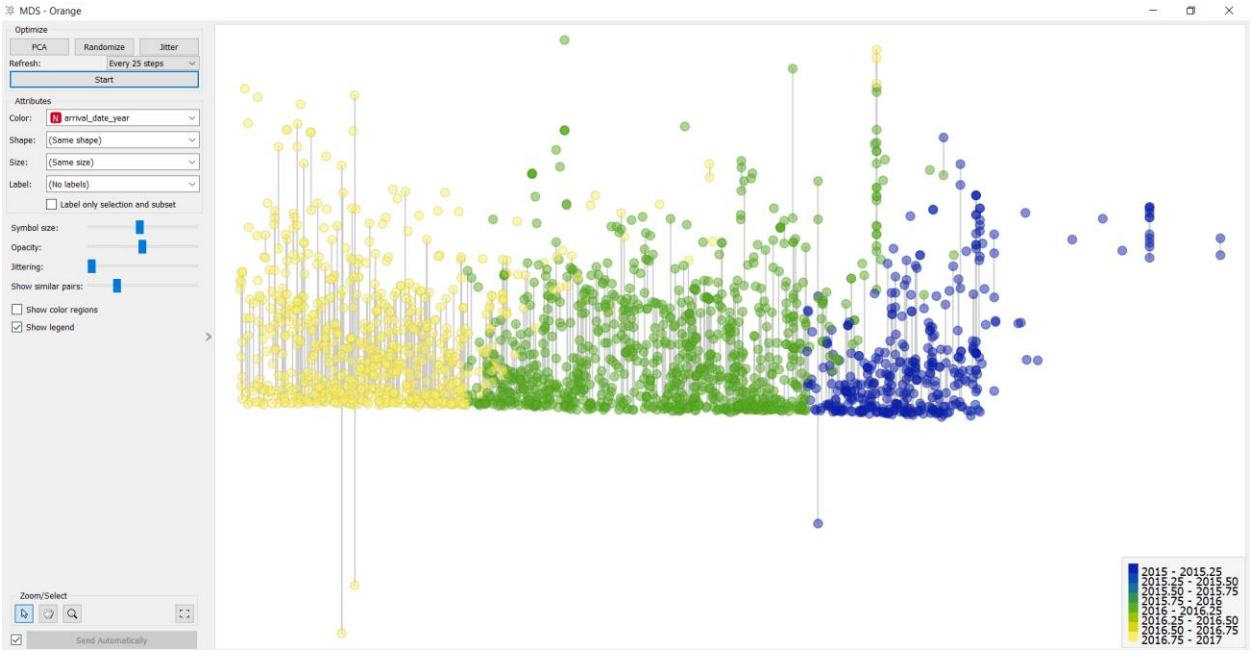


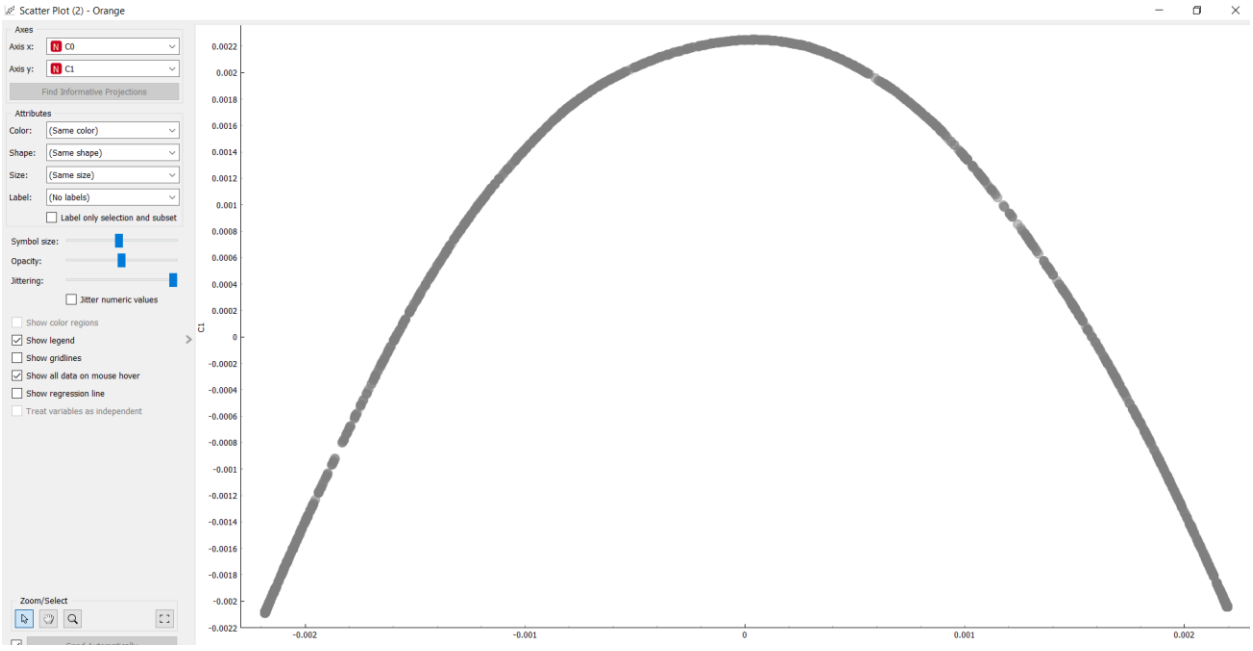
Dimensionality Reduction

Similar to how clustering groups together statistically similar rows, dimensionality reduction does the same for columns. Dimensionality reduction is absolutely necessary for unique datasets that have more columns than rows, however, these methods were used for this dataset as a demonstration.


PCA - Orange
Components Selection
Components: 2
Explained variance: 99%
Options
Normalize variables
Show only first 20
Apply Automatically
cumulative variance
component variance
1.000
0.000
Proportion of variance
Principal Components


t-SNE - Orange
Optimize
Perplexity: 30
Preserve global structure
Exaggeration: 1
PCA components: 20
Normalize data
Start
Attributes
Color: arrival_date_year
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Show color regions
Show legend
Zoom/Select
Send Automatically
2015 - 2015.25
2015.25 - 2015.50
2015.50 - 2015.75
2015.75 - 2016
2016 - 2016.25
2016.25 - 2016.50
2016.50 - 2016.75
2016.75 - 2017


MDS - Orange
Optimize
PCA
Randomize
Jitter
Refresh:
Every 25 steps
Start
Attributes
Color:
arrival_date_year
Shape:
(Same shape)
Size:
(Same size)
Label:
(No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Show similar pairs:
Show color regions
Show legend
Zoom/Select
Send Automatically
2015 - 2015.25
2015.25 - 2015.50
2015.50 - 2015.75
2015.75 - 2016
2016 - 2016.25
2016.25 - 2016.50
2016.50 - 2016.75
2016.75 - 2017


Scatter Plot (2) - Orange
Axes
Axis x:
C0
Axis y:
C1
Find Informative Projections
Attributes
Color:
(Same color)
Shape:
(Same shape)
Size:
(Same size)
Label:
(No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Zoom/Select
C1
0.0022
0.002
0.0018
0.0016
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
-0.0002
-0.0004
-0.0006
-0.0008
-0.001
-0.0012
-0.0014
-0.0016
-0.0018
-0.002
-0.0022
-0.002
-0.001
0
0.001
0.002

	PC1 1.0	PC2 0.0	hotel	customer_type	market_segment	istribution_channe	:served_room_typ	lead_time	eservation_statu:	s_in_weekend_ni	al_date_day_of_m
1	-2.0447e+07	-0.372033	City Hotel	Transient-Party	Groups	TA/TO	A	116	Check-Out	0	24
2	7.63301e+06	0.190472	City Hotel	Transient	Online TA	TA/TO	A	113	Check-Out	1	2
3	-2.20886e+07	1.29563	City Hotel	Transient	Online TA	TA/TO	G	66	Check-Out	0	11
4	-1.0079e+07	0.713748	City Hotel	Transient	Online TA	TA/TO	A	17	Check-Out	0	23
5	-2.6063e+07	3.31143	Resort Hotel	Transient-Party	Groups	TA/TO	D	203	Check-Out	2	25
6	-1.56086e+07	-0.0240025	Resort Hotel	Transient-Party	Groups	Corporate	A	2	Check-Out	0	28
7	-4.54939e+06	-1.45946	Resort Hotel	Transient-Party	Groups	Corporate	A	131	Canceled	2	9
8	1.03114e+07	1.82066	Resort Hotel	Transient	Online TA	TA/TO	A	148	Check-Out	1	30
9	7.89221e+06	0.432509	Resort Hotel	Transient-Party	Groups	TA/TO	A	197	Check-Out	1	30
10	5.81861e+06	1.46813	City Hotel	Transient	Online TA	TA/TO	A	33	Check-Out	0	24
11	-1.89782e+07	-0.536972	City Hotel	Transient-Party	Groups	TA/TO	A	174	Check-Out	2	5
12	1.22986e+07	0.972498	Resort Hotel	Transient-Party	Offline TA/TO	TA/TO	A	46	Check-Out	0	11
13	1.50634e+07	-2.89066	Resort Hotel	Transient	Groups	TA/TO	D	238	Canceled	0	7
14	-7.14139e+06	-1.82616	Resort Hotel	Transient	Online TA	TA/TO	A	65	Canceled	1	26
15	-2.39894e+07	0.828499	City Hotel	Transient-Party	Direct	Direct	A	8	Check-Out	0	2
16	1.18666e+07	-2.0655	City Hotel	Transient	Online TA	TA/TO	A	8	Canceled	1	25
17	1.6705e+07	-3.77415	Resort Hotel	Transient	Groups	TA/TO	A	74	Canceled	1	26
18	-1.64726e+07	-2.88775	City Hotel	Transient	Online TA	TA/TO	A	121	Canceled	1	9
19	-2.99419e+06	0.741765	City Hotel	Transient	Online TA	TA/TO	A	55	Check-Out	1	2
20	-7.48699e+06	1.08529	Resort Hotel	Transient-Party	Online TA	TA/TO	A	166	Check-Out	2	20
21	-1.21526e+07	0.113599	City Hotel	Transient	Online TA	TA/TO	A	3	Check-Out	1	19
22	2.8801e+07	3.2433	Resort Hotel	Transient	Direct	Direct	A	65	Check-Out	2	29
23	1.29034e+07	1.86628	Resort Hotel	Transient	Online TA	TA/TO	A	11	Check-Out	0	4
24	-3.34934e+07	0.799065	City Hotel	Transient	Online TA	TA/TO	D	141	Check-Out	2	17
25	-1.45718e+07	-1.88769	City Hotel	Transient	Online TA	TA/TO	D	24	Canceled	1	10
26	1.81738e+07	-1.97926	City Hotel	Transient	Online TA	TA/TO	A	56	Canceled	0	26
27	-8.35099e+06	-1.55901	City Hotel	Transient	Online TA	TA/TO	D	59	Canceled	2	11
28	1.76554e+07	-2.43786	City Hotel	Transient	Online TA	TA/TO	A	103	Canceled	0	21
29	3.29482e+07	0.31003	City Hotel	Transient-Party	Groups	TA/TO	A	271	Check-Out	0	15
30	1.44586e+07	3.34605	Resort Hotel	Transient	Online TA	TA/TO	A	0	Check-Out	1	15
31	-2.96918e+07	-0.32333	City Hotel	Transient	Online TA	TA/TO	A	49	Check-Out	2	8
32	-3.11606e+07	0.738505	Resort Hotel	Transient	Online TA	TA/TO	H	226	Check-Out	1	24
33	2.46538e+07	-4.8493	City Hotel	Transient	Groups	TA/TO	A	468	Canceled	0	20
34	-1.8719e+07	-2.97948	City Hotel	Transient	Online TA	TA/TO	A	133	Canceled	0	7
35	2.08522e+07	3.57654	Resort Hotel	Transient-Party	Groups	Corporate	A	19	Check-Out	0	2
36	1.4977e+07	0.376244	City Hotel	Transient-Party	Groups	TA/TO	A	59	Check-Out	2	6
37	-9.64699e+06	0.880813	Resort Hotel	Transient	Online TA	TA/TO	A	116	Check-Out	1	17

	Selected	t-SNE-x	t-SNE-y	hotel	customer_type	market_segment	distribution_channel	reserved_room_type	lead_time	reservation_status	ıys_in_weekend_nigl
1	No	-26.0342	15.3454	City Hotel	Transient-Party	Groups	TA/TO	A	116	Check-Out	0
2	No	9.15642	7.68099	City Hotel	Transient	Online TA	TA/TO	A	113	Check-Out	1
3	No	-28.2147	-5.74179	City Hotel	Transient	Online TA	TA/TO	G	66	Check-Out	0
4	No	-9.86292	22.8399	City Hotel	Transient	Online TA	TA/TO	A	17	Check-Out	0
5	No	-39.5549	-5.25496	Resort Hotel	Transient-Party	Groups	TA/TO	D	203	Check-Out	2
6	No	-18.4058	-1.90497	Resort Hotel	Transient-Party	Groups	Corporate	A	2	Check-Out	0
7	No	-8.23507	-5.19204	Resort Hotel	Transient-Party	Groups	Corporate	A	131	Canceled	2
8	No	11.4108	-8.35746	Resort Hotel	Transient	Online TA	TA/TO	A	148	Check-Out	1
9	No	9.55754	5.4126	Resort Hotel	Transient-Party	Groups	TA/TO	A	197	Check-Out	1
10	No	0.120335	13.769	City Hotel	Transient	Online TA	TA/TO	A	33	Check-Out	0
11	No	-17.3457	16.829	City Hotel	Transient-Party	Groups	TA/TO	A	174	Check-Out	2
12	No	13.6059	-21.5181	Resort Hotel	Transient-Party	Offline TA/TO	TA/TO	A	46	Check-Out	0
13	No	21.2495	3.80003	Resort Hotel	Transient	Groups	TA/TO	D	238	Canceled	0
14	No	0.30853	34.6998	Resort Hotel	Transient	Online TA	TA/TO	A	65	Canceled	1
15	No	-32.9234	-15.7411	City Hotel	Transient-Party	Direct	Direct	A	8	Check-Out	0
16	No	13.3052	-18.3485	City Hotel	Transient	Online TA	TA/TO	A	8	Canceled	1
17	No	20.3518	11.5241	Resort Hotel	Transient	Groups	TA/TO	A	74	Canceled	1
18	No	-17.7537	2.96696	City Hotel	Transient	Online TA	TA/TO	A	121	Canceled	1
19	No	-9.5095	-13.937	City Hotel	Transient	Online TA	TA/TO	A	55	Check-Out	1
20	No	-1.53917	33.8847	Resort Hotel	Transient-Party	Online TA	TA/TO	A	166	Check-Out	2
21	No	-12.2924	-29.7543	City Hotel	Transient	Online TA	TA/TO	A	3	Check-Out	1
22	No	42.4692	3.22559	Resort Hotel	Transient	Direct	Direct	A	65	Check-Out	2
23	No	20.3742	-6.66488	Resort Hotel	Transient	Online TA	TA/TO	A	11	Check-Out	0
24	No	-52.5602	-4.78871	City Hotel	Transient	Online TA	TA/TO	D	141	Check-Out	2
25	No	-18.7257	-7.02822	City Hotel	Transient	Online TA	TA/TO	D	24	Canceled	1
26	No	18.0523	19.3769	City Hotel	Transient	Online TA	TA/TO	A	56	Canceled	0
27	No	-5.3756	31.2306	City Hotel	Transient	Online TA	TA/TO	D	59	Canceled	2
28	No	18.0806	17.0049	City Hotel	Transient	Online TA	TA/TO	A	103	Canceled	0
29	No	47.3208	11.8892	City Hotel	Transient-Party	Groups	TA/TO	A	271	Check-Out	0
30	No	21.1423	1.10299	Resort Hotel	Transient	Online TA	TA/TO	A	0	Check-Out	1
31	No	-46.15	8.75467	City Hotel	Transient	Online TA	TA/TO	A	49	Check-Out	2
32	No	-51.2372	3.02549	Resort Hotel	Transient	Online TA	TA/TO	H	226	Check-Out	1
33	No	27.185	-17.5354	City Hotel	Transient	Groups	TA/TO	A	468	Canceled	0
34	No	-16.2285	15.9108	City Hotel	Transient	Online TA	TA/TO	A	133	Canceled	0
35	No	32.1348	2.46215	Resort Hotel	Transient-Party	Groups	Corporate	A	19	Check-Out	0
36	No	21.2455	3.40384	City Hotel	Transient-Party	Groups	TA/TO	A	59	Check-Out	2
37	No	-9.22596	26.3822	Resort Hotel	Transient	Online TA	TA/TO	A	116	Check-Out	1
38	No	-28.3524	-9.56532	City Hotel	Transient	Online TA	TA/TO	A	26	Canceled	1

	C0	C1
1	0.0018177	-0.000713263
2	-0.000809335	0.00172846
3	0.00192523	-0.00106882
4	0.000900189	0.00154616
5	0.00211833	-0.0017578
6	0.0014687	0.000301622
7	0.000401136	0.00212747
8	-0.00107857	0.00127214
9	-0.000831147	0.00169922
10	-0.000642363	0.00192219
11	0.00171628	-0.000397824
12	-0.00127368	0.000858658
13	-0.00148471	0.00030738
14	0.00065664	0.00189035
15	0.0020275	-0.00142688
16	-0.00122487	0.000970751
17	-0.00158234	1.8604e-05
18	0.00150778	0.000198244
19	0.000250637	0.00220955
20	0.000684561	0.00185767
21	0.00112035	0.00111709
22	-0.00216038	-0.00199618
23	-0.00131754	0.000753094
24	0.00219411	-0.00204246
25	0.00135894	0.000575856
26	-0.00168438	-0.000305572
27	0.000779542	0.00173061
28	-0.00165005	-0.000193875
29	-0.00218293	-0.00208412
30	-0.00143227	0.00045203
31	0.0021867	-0.00201415
32	0.0021931	-0.00203858
33	-0.00205687	-0.00160273
34	0.00169327	-0.000328695
35	-0.00186662	-0.000922052

	Selected	mds-x	mds-y	hotel	customer_type	market_segment	distribution_channel	reserved_room_type	lead_time	reservation_status	ys_in_weekend_nigl
1	No	-2.0447e+07	21.633	City Hotel	Transient-Party	Groups	TA/TO	A	116	Check-Out	0
2	No	7.63301e+06	8.35555	City Hotel	Transient	Online TA	TA/TO	A	113	Check-Out	1
3	No	-2.20886e+07	-34.0394	City Hotel	Transient	Online TA	TA/TO	G	66	Check-Out	0
4	No	-1.0079e+07	-81.0166	City Hotel	Transient	Online TA	TA/TO	A	17	Check-Out	0
5	No	-2.6063e+07	111.325	Resort Hotel	Transient-Party	Groups	TA/TO	D	203	Check-Out	2
6	No	-1.56086e+07	-92.0979	Resort Hotel	Transient-Party	Groups	Corporate	A	2	Check-Out	0
7	No	-4.54939e+06	32.6269	Resort Hotel	Transient-Party	Groups	Corporate	A	131	Canceled	2
8	No	1.03114e+07	44.1905	Resort Hotel	Transient	Online TA	TA/TO	A	148	Check-Out	1
9	No	7.89221e+06	94.4083	Resort Hotel	Transient-Party	Groups	TA/TO	A	197	Check-Out	1
10	No	5.81861e+06	-71.7419	City Hotel	Transient	Online TA	TA/TO	A	33	Check-Out	0
11	No	-1.89782e+07	78.9736	City Hotel	Transient-Party	Groups	TA/TO	A	174	Check-Out	2
12	No	1.22986e+07	-57.3145	Resort Hotel	Transient-Party	Offline TA/TO	TA/TO	A	46	Check-Out	0
13	No	1.50634e+07	131.43	Resort Hotel	Transient	Groups	TA/TO	D	238	Canceled	0
14	No	-7.14139e+06	-31.9619	Resort Hotel	Transient	Online TA	TA/TO	A	65	Canceled	1
15	No	-2.39894e+07	-87.7379	City Hotel	Transient-Party	Direct	Direct	A	8	Check-Out	0
16	No	1.18666e+07	-101.379	City Hotel	Transient	Online TA	TA/TO	A	8	Canceled	1
17	No	1.6705e+07	-32.6306	Resort Hotel	Transient	Groups	TA/TO	A	74	Canceled	1
18	No	-1.64726e+07	22.5249	City Hotel	Transient	Online TA	TA/TO	A	121	Canceled	1
19	No	-2.99419e+06	-45.8393	City Hotel	Transient	Online TA	TA/TO	A	55	Check-Out	1
20	No	-7.48699e+06	68.9512	Resort Hotel	Transient-Party	Online TA	TA/TO	A	166	Check-Out	2
21	No	-1.21526e+07	-94.1096	City Hotel	Transient	Online TA	TA/TO	A	3	Check-Out	1
22	No	2.8801e+07	-46.1437	Resort Hotel	Transient	Direct	Direct	A	65	Check-Out	2
23	No	1.29034e+07	-93.2241	Resort Hotel	Transient	Online TA	TA/TO	A	11	Check-Out	0
24	No	-3.34934e+07	47.2717	City Hotel	Transient	Online TA	TA/TO	D	141	Check-Out	2
25	No	-1.45718e+07	-72.4575	City Hotel	Transient	Online TA	TA/TO	D	24	Canceled	1
26	No	1.81738e+07	-51.0692	City Hotel	Transient	Online TA	TA/TO	A	56	Canceled	0
27	No	-8.35099e+06	-40.2549	City Hotel	Transient	Online TA	TA/TO	D	59	Canceled	2
28	No	1.76554e+07	-3.91323	City Hotel	Transient	Online TA	TA/TO	A	103	Canceled	0
29	No	3.29482e+07	158.209	City Hotel	Transient-Party	Groups	TA/TO	A	271	Check-Out	0
30	No	1.44586e+07	-105.226	Resort Hotel	Transient	Online TA	TA/TO	A	0	Check-Out	1
31	No	-2.96918e+07	-43.9204	City Hotel	Transient	Online TA	TA/TO	A	49	Check-Out	2
32	No	-3.11606e+07	126.93	Resort Hotel	Transient	Online TA	TA/TO	H	226	Check-Out	1
33	No	2.46538e+07	358.912	City Hotel	Transient	Groups	TA/TO	A	468	Canceled	0
34	No	-1.8719e+07	36.1742	City Hotel	Transient	Online TA	TA/TO	A	133	Canceled	0
35	No	2.08522e+07	-87.0474	Resort Hotel	Transient-Party	Groups	Corporate	A	19	Check-Out	0
36	No	1.4977e+07	-46.9037	City Hotel	Transient-Party	Groups	TA/TO	A	59	Check-Out	2
37	No	-9.64699e+06	19.9351	Resort Hotel	Transient	Online TA	TA/TO	A	116	Check-Out	1
38	No	-2.26934e+07	-68.3948	City Hotel	Transient	Online TA	TA/TO	A	26	Canceled	1

JASP

Support Vector Machine Regression

Support Vector Machine Regression

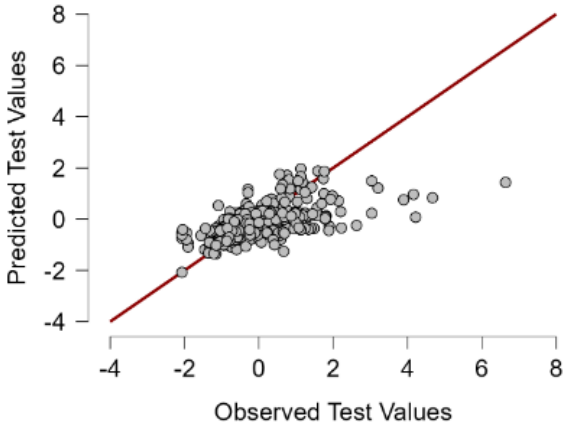
Support Vectors	n(Train)	n(Test)	Test MSE
1573	1600	400	0.739

Data Split

Evaluation Metrics

	Value
MSE	0.739
RMSE	0.86
MAE	0.582
MAPE	1921.43%
R^2	0.348

Predictive Performance Plot



Random Forest Classification

Random Forest Classification

Trees	Predictors per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
71	5	1280	320	400	0.753	0.795	0.568

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	246	7
	1	75	72

Class Proportions

	Data Set	Training Set	Validation Set	Test Set
0	0.644	0.651	0.631	0.632
1	0.356	0.349	0.369	0.367

Evaluation Metrics

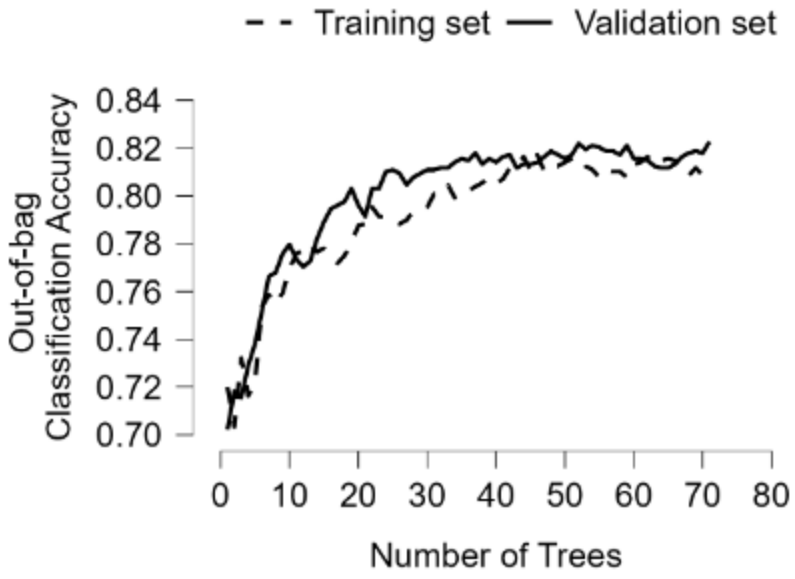
	0	1	Average / Total
Support	253	147	400
Accuracy	0.795	0.795	0.795
Precision (Positive Predictive Value)	0.766	0.911	0.820
Recall (True Positive Rate)	0.972	0.490	0.795
False Positive Rate	0.510	0.028	0.269
False Discovery Rate	0.234	0.089	0.161
F1 Score	0.857	0.637	0.776
Area Under Curve (AUC)	0.857	0.861	0.859
Negative Predictive Value	0.911	0.766	0.839
True Negative Rate	0.490	0.972	0.731
False Negative Rate	0.028	0.510	0.269
False Omission Rate	0.089	0.234	0.161
Threat Score	1.567	0.809	1.188
Statistical Parity	0.802	0.198	1.000

Note. All metrics are calculated for every class against all other classes.

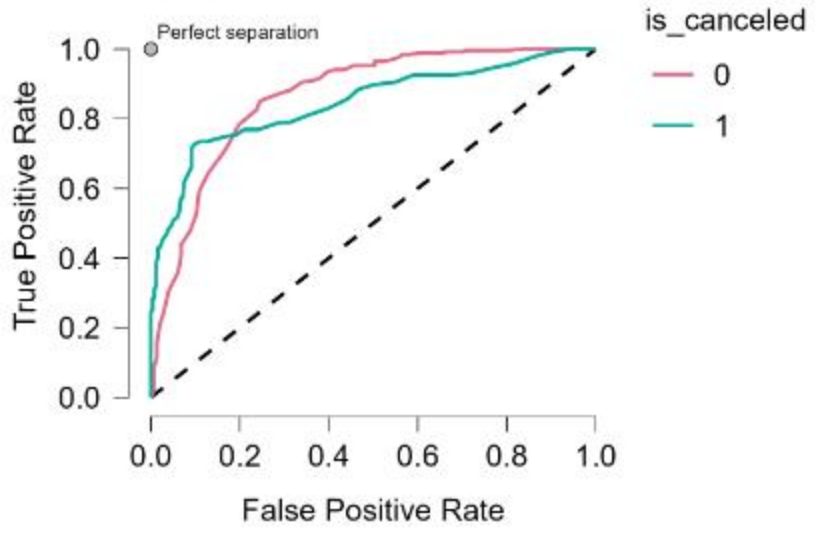
Variable Importance

	Mean decrease in accuracy	Total increase in node purity
deposit_type	0.049	0.090
total_of_special_requests	0.028	0.052
lead_time	0.026	0.045
assigned_room_type	5.381e−4	0.027
market_segment	0.014	0.015
reserved_room_type	−0.004	0.012
adr	0.007	0.009
previous_cancellations	0.013	0.008
customer_type	0.008	0.007
booking_changes	0.003	0.006
arrival_date_year	0.002	0.005
stays_in_week_nights	0.002	0.005
distribution_channel	0.002	0.005
required_car_parking_spaces	7.538e−4	0.004
meal	1.553e−4	0.002
hotel	0.001	0.001
stays_in_weekend_nights	0.002	0.001
children	5.902e−4	9.250e−4
previous_bookings_not_canceled	6.386e−4	4.727e−4
arrival_date_week_number	0.009	7.090e−5
babies	−1.127e−4	0.000
is_repeated_guest	6.969e−4	−1.442e−4
adults	0.001	−2.707e−4
days_in_waiting_list	0.003	−5.585e−4
arrival_date_day_of_month	−0.003	−8.388e−4
arrival_date_month	0.013	−0.003

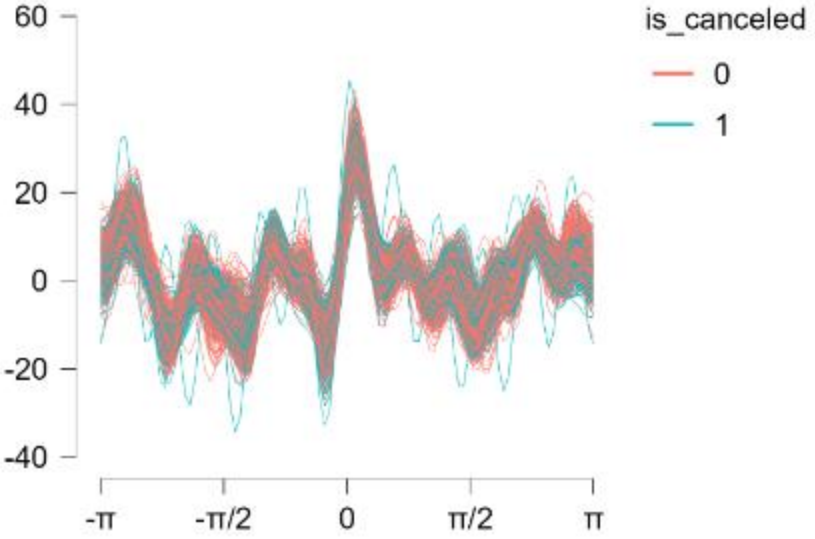
Out-of-bag Classification Accuracy Plot



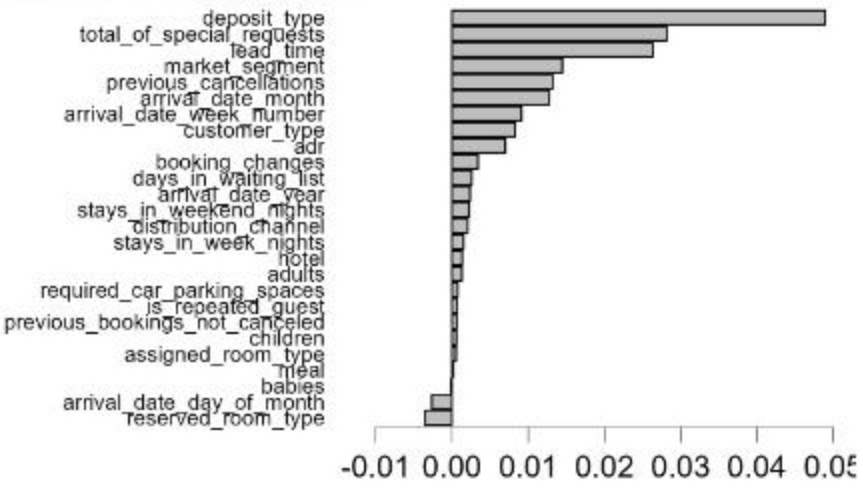
ROC Curves Plot

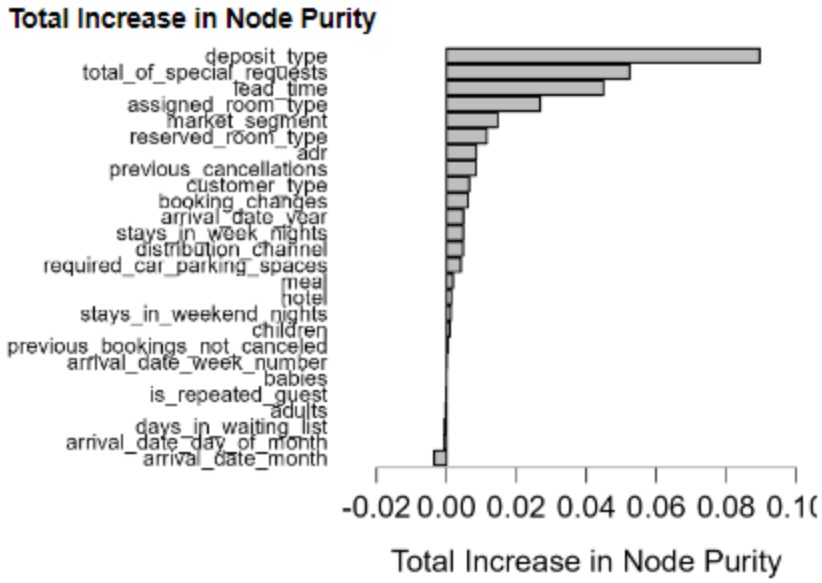


Andrews Curves Plot



Mean Decrease in Accuracy




Total Increase in Node Purity
deposit_type
total_of_special_requests
lead_time
assigned_room_type
market_segment
reserved_room_type
adr
previous_cancellations
customer_type
booking_changes
arrival_date_year
stays_in_week_nights
distribution_channel
required_car_parking_spaces
meal
hotel
stays_in_weekend_nights
children
previous_bookings_not_canceled
arrival_date_week_number
babies
is_repeated_guest
adults
days_in_waiting_list
arrival_date_day_of_month
arrival_date_month
-0.02 0.00 0.02 0.04 0.06 0.08 0.1(
Total Increase in Node Purity


lead_time
arrival_date_week_numb
arrival_date_day_of_mon
stays_in_week_nights
evious_bookings_not_can
days_in_waiting_list
arrival_date_week_nun
arrival_date_day_of_m
stays_in_week_nigh
previous_bookings_not_c
days_in_waiting_lis
adr
is_canceled
0
1

Neighborhood-Based Clustering

K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
10	2000	0.604	5681.690	6073.750	0.200

Note. The model is optimized with respect to the *BIC* value.
Note. The optimum number of clusters is the maximum number of clusters. You might want to adjust the range of optimization.

Cluster Information

Cluster	1	2	3	4	5	6	7	8	9	10
Size	299	305	15	192	314	45	344	7	219	260
Explained proportion within-cluster heterogeneity	0.128	0.151	0.094	0.111	0.104	0.051	0.150	0.031	0.092	0.089
Within sum of squares	708.622	834.541	520.686	614.203	576.006	280.139	831.750	170.239	510.662	494.840
Silhouette score	0.198	0.166	0.253	0.186	0.254	0.123	0.174	0.514	0.155	0.244
Center lead_time	−0.284	−0.416	−0.691	2.130	−0.595	0.570	−0.207	2.534	0.565	−0.369
Center arrival_date_week_number	1.152	0.068	−0.347	0.345	−1.170	0.075	−0.785	0.300	−0.253	1.004
Center arrival_date_day_of_month	0.946	0.142	−0.195	0.090	−0.775	−0.145	0.948	−0.196	−0.758	−0.958
Center stays_in_week_nights	−0.063	−0.109	−0.536	−0.239	−0.391	3.867	−0.273	0.082	1.082	−0.343
Center previous_bookings_not_canceled	−0.094	−0.087	9.218	−0.115	−0.014	−0.115	−0.035	−0.115	−0.115	−0.054
Center days_in_waiting_list	0.006	−0.122	−0.122	0.012	−0.116	−0.122	0.062	13.883	−0.103	−0.073
Center adr	−0.275	1.668	−0.779	−0.283	−0.476	−0.248	−0.242	−0.501	−0.155	−0.305

Note. The Between Sum of Squares of the 10 cluster model is 8451.31
Note. The Total Sum of Squares of the 10 cluster model is 13993

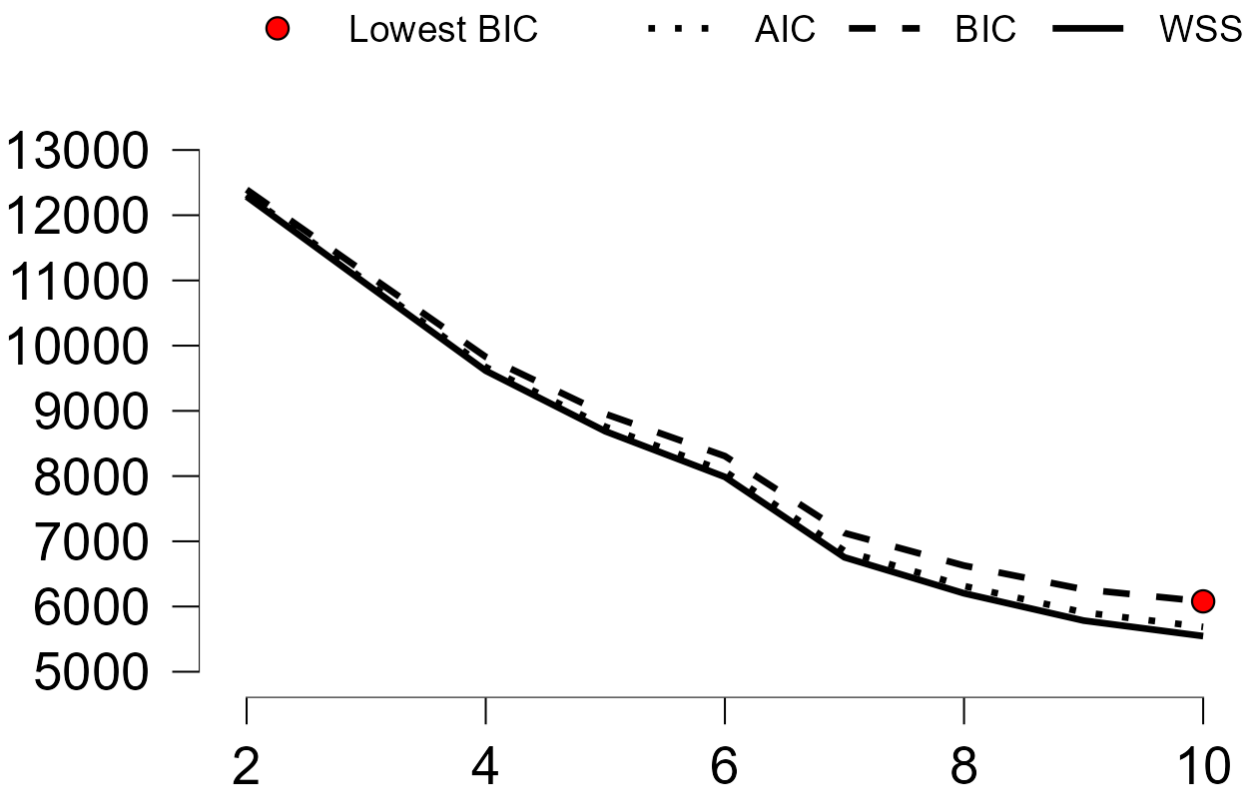
Cluster Means

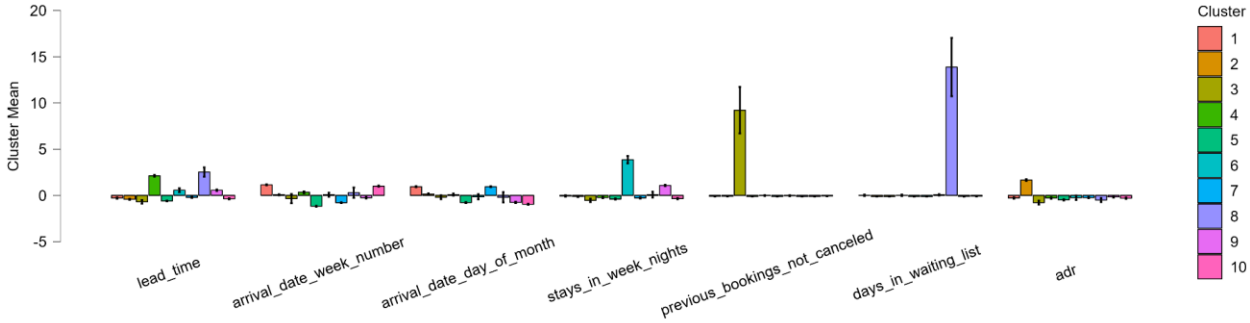
	lead_time	arrival_date_week_number	arrival_date_day_of_month	stays_in_week_nights	previous_bookings_not_canceled	days_in_waiting_list	adr
Cluster 1	−0.284	1.152	0.946	−0.063	−0.094	0.006	−0.275
Cluster 2	−0.416	0.068	0.142	−0.109	−0.087	−0.122	1.668
Cluster 3	−0.691	−0.347	−0.195	−0.536	9.218	−0.122	−0.779
Cluster 4	2.130	0.345	0.090	−0.239	−0.115	0.012	−0.283
Cluster 5	−0.595	−1.170	−0.775	−0.391	−0.014	−0.116	−0.476
Cluster 6	0.570	0.075	−0.145	3.867	−0.115	−0.122	−0.248
Cluster 7	−0.207	−0.785	0.948	−0.273	−0.035	0.062	−0.242
Cluster 8	2.534	0.300	−0.196	0.082	−0.115	13.883	−0.501
Cluster 9	0.565	−0.253	−0.758	1.082	−0.115	−0.103	−0.155
Cluster 10	−0.369	1.004	−0.958	−0.343	−0.054	−0.073	−0.305

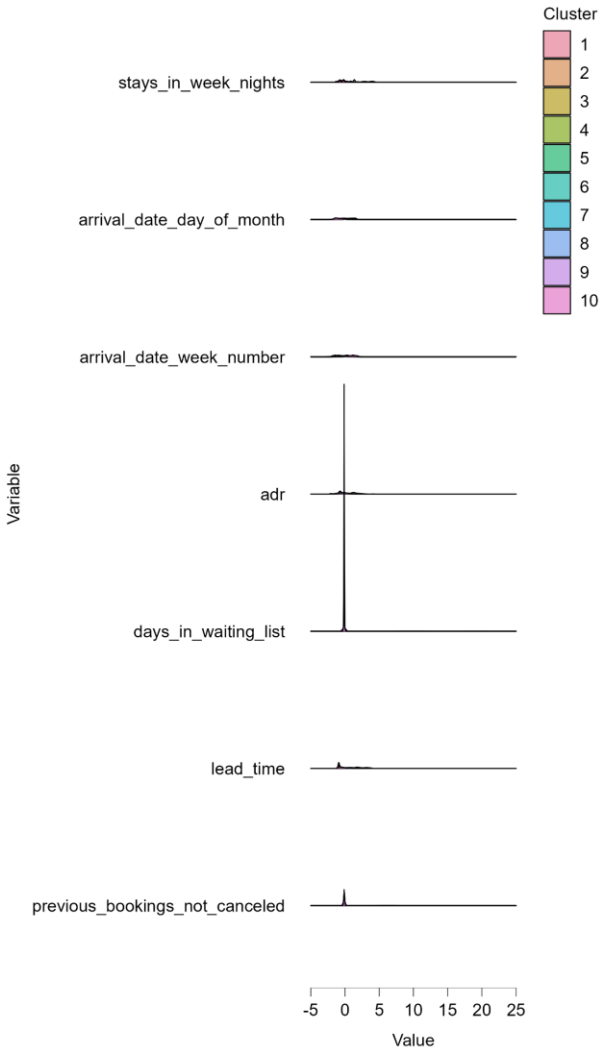
Evaluation Metrics

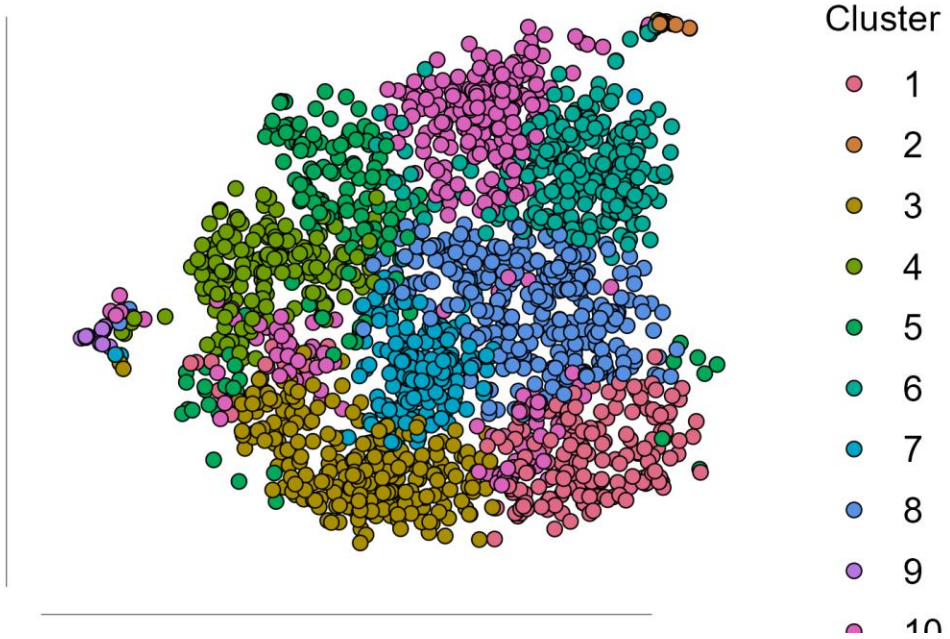
	Value
Maximum diameter	18.158
Minimum separation	0.164
Pearson's γ	0.265
Dunn index	0.009
Entropy	2.039
Calinski-Harabasz index	337.204

Note. All metrics are based on the *euclidean* distance.


Lowest BIC
AIC
BIC
WSS
13000
12000
11000
10000
9000
8000
7000
6000
5000
2
4
6
8
10
Number of Clusters


Cluster Mean
20
15
10
5
0
-5
lead_time
arrival_date_week_number
arrival_date_day_of_month
stays_in_week_nights
previous_bookings_not_canceled
days_in_waiting_list
adr
Cluster
1
2
3
4
5
6
7
8
9
10


Cluster
1
2
3
4
5
6
7
8
9
10
stays_in_week_nights
arrival_date_day_of_month
arrival_date_week_number
adr
days_in_waiting_list
lead_time
previous_bookings_not_canceled
Variable
-5
0
5
10
15
20
25
Value


Cluster
1
2
3
4
5
6
7
8
9
10

RapidMiner AutoML

x-Means - Summary

Number of Clusters: 5

Cluster 0 638

total_of_special_requests is on average **74.28%** smaller, **lead_time** is on average **60.96%** smaller, **stays_in_weekend_nights** is on average **32.98%** smaller

Cluster 1 476

total_of_special_requests is on average **170.06%** larger, **lead_time** is on average **28.82%** smaller, **stays_in_weekend_nights** is on average **17.64%** smaller

Cluster 2 388

lead_time is on average **133.30%** larger, **total_of_special_requests** is on average **83.70%** smaller, **stays_in_week_nights** is on average **19.16%** smaller

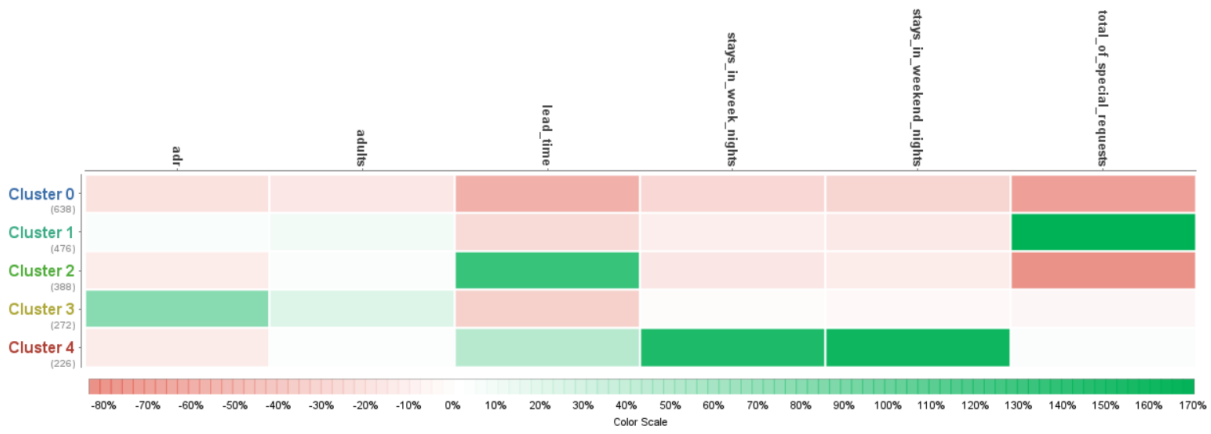
Cluster 3 272

adr is on average **79.07%** larger, **lead_time** is on average **36.67%** smaller, **adults** is on average **22.45%** larger

Cluster 4 226

stays_in_weekend_nights is on average **161.15%** larger, **stays_in_week_nights** is on average **151.04%** larger, **lead_time** is on average **48.06%** larger

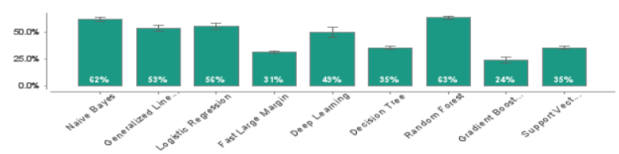
x-Means - Heat Map



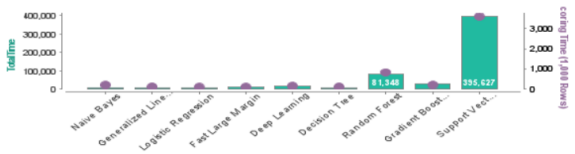
Overview

Number of Models: 205

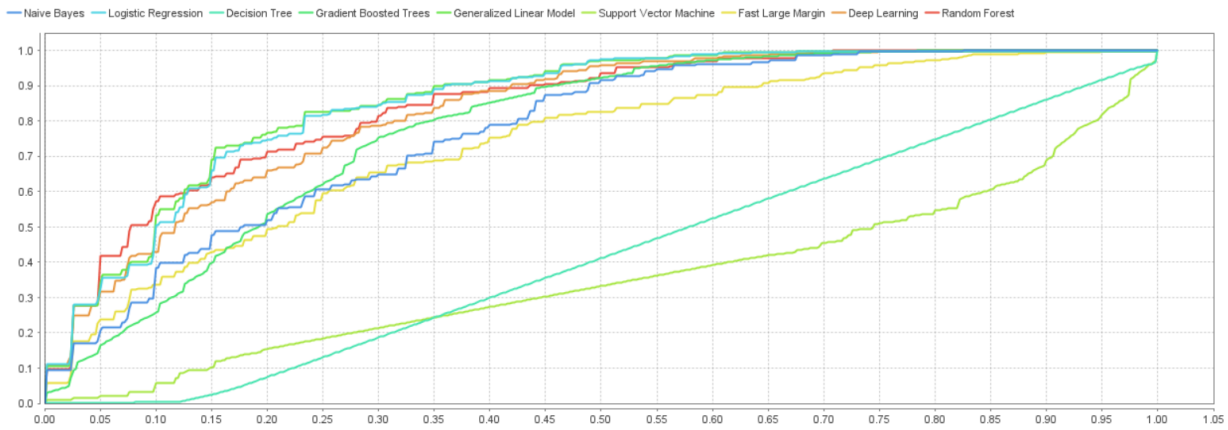
Classification Error



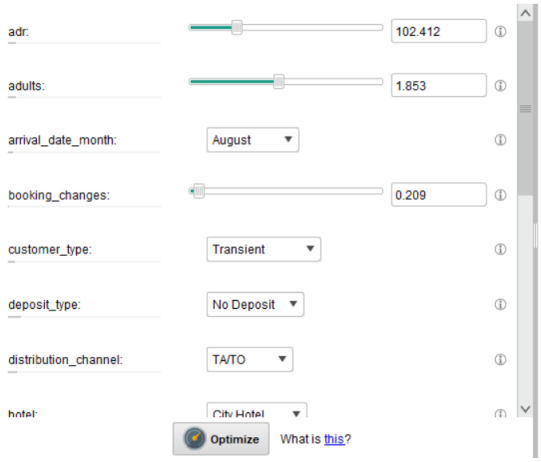
Runtimes (ms)

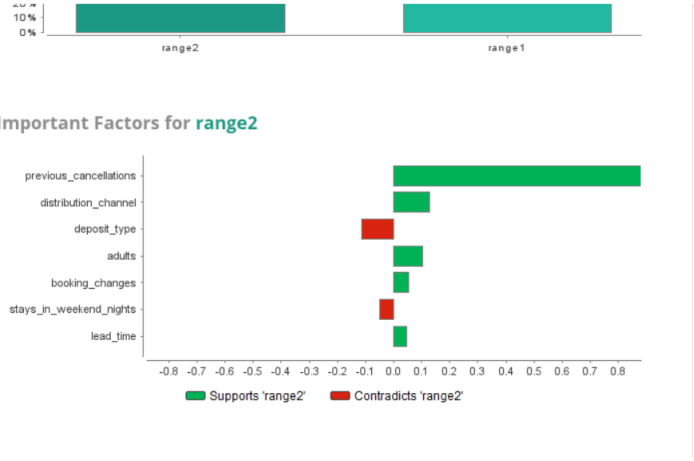


ROC Comparison

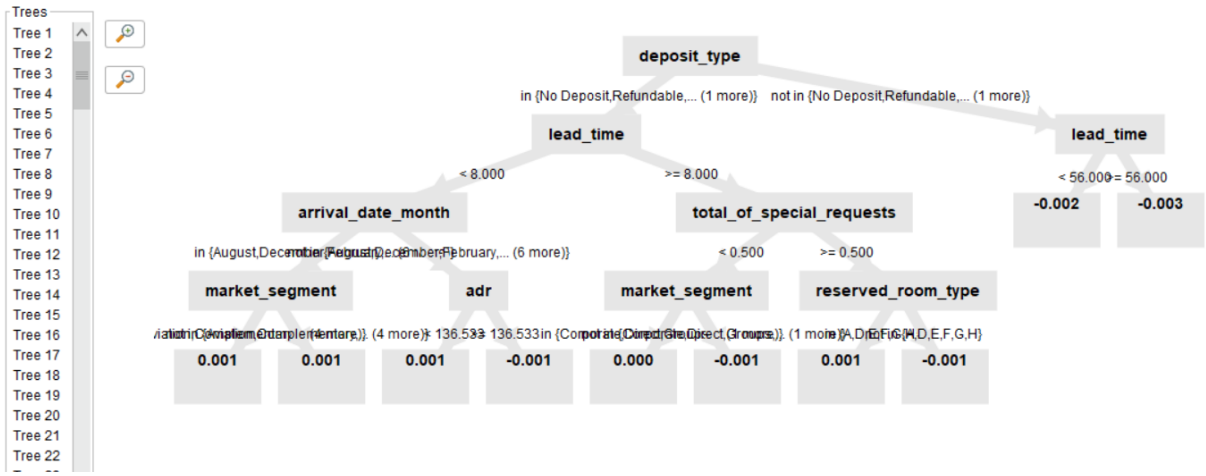


Deep Learning - Simulator





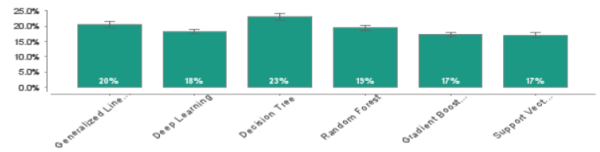
Gradient Boosted Trees - Model



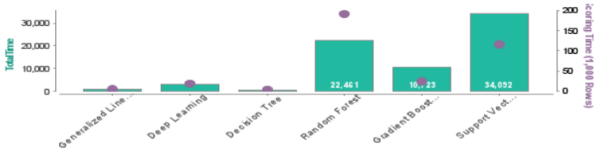
Overview

Number of Models: 8

Relative Error



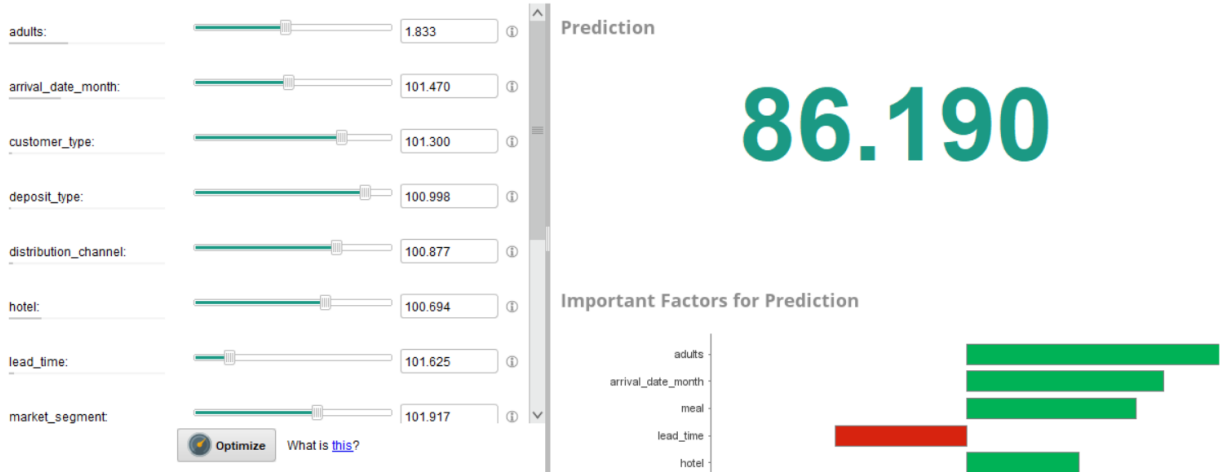
Runtimes (ms)



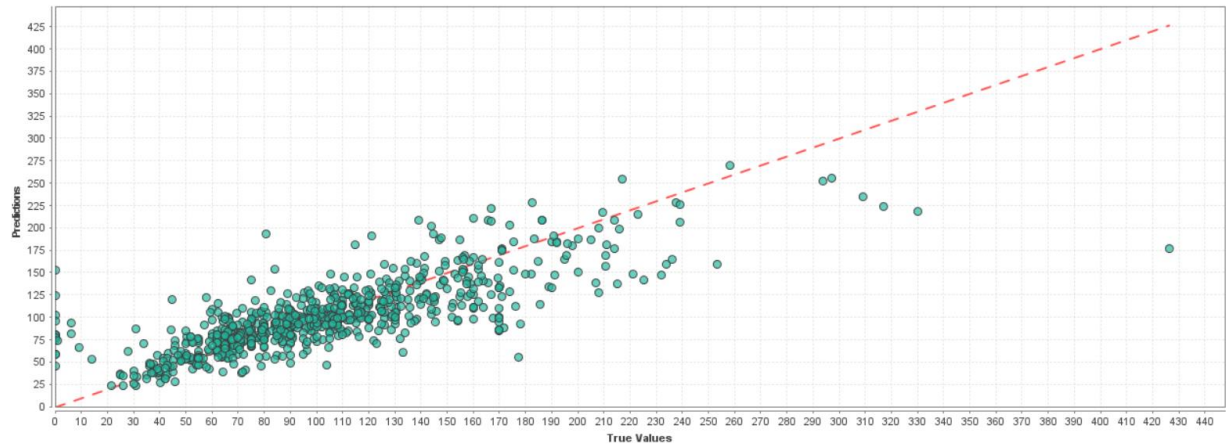
Support Vector Machine - Weights

Attribute	Weight
adults	0.381
arrival_date_month	0.333
market_segment	0.279
meal	0.237
hotel	0.212
stays_in_week_nights	0.145
reserved_room_type	0.091
stays_in_weekend_nights	0.072
total_of_special_requests	0.042
lead_time	0.029
distribution_channel	0.025
customer_type	0.023

Support Vector Machine - Simulator



Support Vector Machine - Predictions Chart



Possible Future Projects

- Classification, clustering, regression, dimensionality reduction, association rules, time series
- Build a model that predicts cancellations (classification)
- Time series analysis of ADR (regression)
- Customer segmentation based on various categorical and discrete variables (clustering)
- What type of meal do people order in what type of room (association rules mining)