Methodology Document

Technical Specifications

Physical

S.No.	Machine Type	Model	Processor	RAM	GPU
1.	Laptop	*hp spectre x360 convertible 15-bl0xx*	*i7 8th Gen.*	*16 GB*	

Software

S.No.	OS/Software	Version	Details (any specifics)	URL
1.	*Windows*	*10 Pro*		https://www.microsoft.com/en-us/p/windows-10-pro/df77x4d43rkt/48DN
2.	*Python*	3.7		https://www.python.org/
3.	*Anaconda*	2019.03		https://www.anaconda.com/distribution/

Feature Summary

	numberofadults	numberofchildren	roomnights	total_pax	advbook	stay
count	488189.000000	488189.000000	488189.000000	488189.000000	488189.000000	488189.000000
mean	3.275522	0.362573	3.736133	3.191893	46.119677	2.466323
std	1.764458	0.758078	2.479779	1.166638	38.693097	1.094641
min	0.000000	0.000000	0.000000	0.000000	-2219.000000	1.000000
25%	2.000000	0.000000	2.000000	2.000000	15.000000	2.000000
50%	3.000000	0.000000	3.000000	3.000000	33.000000	2.000000
75%	4.000000	0.000000	4.000000	4.000000	82.000000	3.000000
max	32.000000	13.000000	80.000000	24.000000	177.000000	26.000000

Data Cleaning

S.No.	Column Name	Treatment	Details
1.	season_holidayed_code	Missing value	0.0305% missing value. Performed feature imputation using Mode.
2.	state_code_residence	Missing value	Filled with 17 since all the number are present from 1 to 38 except 17.
3.	`reservation_id`	Removed	Since all are unique categories no importance.
4.	`memberid`	Removed	Since ~30% of data has unique values.

5.	`booking_date`
	`checkin_date`
	`checkout_date`
		Data type converted in to DateTime	By default, it is in string data type so converted into date time object.

6.	`'booking_date','checkin_date','checkout_date','channel_code','main_product_code','persontravellingid','resort_region_code','resort_type_code','room_type_booked_code','season_holidayed_code','state_code_residence','state_code_resort','member_age_buckets','booking_type_code','cluster_code','reservationstatusid_code','resort_id'`	Data Type Conversion	Into String since going to use CatBoost.
7.	`roomnights`	Replaced -45 to 45	Since room nights cannot be -ve replaced with 45.

Feature Engineering

Derived Variable

S.No.	New Column Name	Treatment	Details
1.	advbook	checkin_date - booking_date	Created new column advance booking by subtracting check in and booking dates.

2	stay	checkout_date-checkin_date	Created a new column stay by subtracting checkout and check in dates.
3	person_days	(numberofadults+ numberofchildren)* roomnights	Since the problem statement is to predict amount spent, I thought total persons times room nights will give good results.

Exploratory Data Analysis

EDA

1. Took care of data types, at one point had to convert some features in to one data type and change it back to default for feature generation and algorithms compatibility purposes.
2. Found out the number of unique values of each feature it helped to decide what feature to remove.
3. Individually checked the non-categorical values of a features and found one flaw and fixed it (from -45 to 45).
4. Histogram of output variable i.e.., amount_spent_per_room_night_scaled.
5. Since there are lot of categorical features and for features like dates have lot of categories decided to use ensemble algorithm CatBoost.

Model Run

Run No.	Model	Metric	Value	Hyperparameter values
1	CatBoost	RMSE	Train: Test:	border_count=225, l2_leaf_reg=2, depth=12, iterations=800, loss_function='RMSE' learn: 0.9112084, total: 34m 55s
2	Others	RMSE	Train: Test:	Tried other things like without features engineering, with and with out regularization term and tried averaging different model's outputs but the above one gave good results.

Coding Details

S.No	Programming Language	Package Used	Details
1.	Python	Scikit learn	General packages like cross_val_score, GridSearchCV
2.	Python	hyperparameter-hunter	GridSearchCV
3.	Python	CatBoost	CatBoost Regressor function.
4.	Python	Pandas, NumPy, SciPy, MatPlotlib.	For preprocessing and analysis purposes.

Platforms/Tools Used (if any)

S.No	Platform Tool	Details
1.	Anaconda	Just used pain Jupyter Notebook.

Note:

This is my first ever live hackathon thank you, very much for conducting this competition.

Things that I could do to improve my scores:

1. Staking of different models.
2. Can try Bayesian optimization.
3. Could use Category count for important features.
4. Member ID can be added (worth giving it a shot).
5. Should have used CV=10 instead of 5.
6. Frequency count of important features.