# Online travel agency: Data analysis and Machine Learning project report

|  |  |
| --- | --- |
| [by Dimitar Petrov] | [23/04/2023] |
|  |  |

# Introduction

In this report, I present the findings of a data science project using the Online travel agency (OTA) dataset. The dataset contained various features, including categorical and numerical data. My first step was to load the dataset into Python and also passing the column named ‘timestamp’ as dates. This means that pandas will try to convert the values in that column to a datetime object. I found that most features were categorical, while some were numerical and one of them was converted to a datetime object. The dataset has 2,380,557 rows and 54 columns.

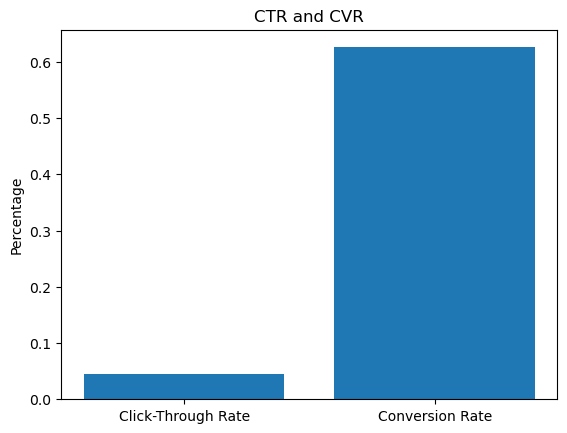
## Data cleaning and Analysis

I performed some basic statistics such as mean, standard deviation, and quartile range. Most of the features exhibited a high degree of variability and had many outliers indicating a possible presence of noise or errors in the data collection process. The majority of the features had a positive skewness, meaning that they had a long tail to the right and a mode that was lower than the mean and median. This suggested that the data was not normally distributed and might require some transformation or normalization techniques. The features ‘destination\_id’, ‘listing\_id’ and ‘search\_id’ had a nearly uniform distribution, implying that they were likely to be unique identifiers for each booking and not informative for the analysis. Some features had a constant value for all the bookings, indicating that they were **redundant** and could be dropped from the dataset. A few features had a negative skewness, meaning that they had a long tail to the left and a mode that was higher than the mean and median. This indicated that the data has some outliers or extreme values that could affect the analysis.

Then I calculated and plotted a heatmap of the correlation between the features in the dataset. It revealed that the feature ‘booking\_value’ had a correlation coefficient of NaN with both clicked and booked features, indicating that there was no linear relationship between them. This could be due to the presence of collinearity, which occurs when two or more features are highly correlated with each other and can cause problems in the analysis. The features ‘clicked’ and ‘booked’ had a very high correlation coefficient of 0.98, indicating that they were almost perfectly linearly related.

Part of the analysis of the dataset involved dealing with missing values. I applied different strategies based on the percentage of missing values in each column and ended up dropping 53% of the features. The reason for that was that the dataset had some columns with more than 30% missing values. These columns were unlikely to have much predictive power, so I removed them for simplicity and efficiency of the model training. The dataset also had some columns with less than 30% missing values. These columns were potentially useful, so I tried to impute the missing values using statistical measures, such as mean, median, mode, etc. I chose the appropriate measure based on the distribution and type of each column. As a result of the data cleaning process, 53% of the features of the original dataset were lost.

After cleaning and analyzing the dataset, I calculated the click-through rate (CTR) and conversion rate (CVR). The CTR is 4.46%, meaning that out of all the users who viewed the hotel listing, only 4.46% of them clicked on at least one hotel listing. The CVR is 62.57% which means that out of all the users who clicked on a hotel listing, 62.57% of them eventually booked the hotel.



## Data privacy and security issues

There are no data privacy and security issues in this dataset as it does not contain any personal information about the users, but it is recommended implementing security measures such as encryption and access controls to prevent unauthorized access and protect the data from any potential threats.

## Model Training

For the model training I split the dataset into training and test sets with a ratio of 80:20, respectively and normalized the features with StandardScaler. I chose to predict the variable “booked” because it is more important for the profit of the travel agency to predict completed bookings than the clicks.

It was convenient to use Logistic regression because it is computationally efficient and has low variance, which makes it less prone to overfitting, but it may not work well with complex non-linear relationships between the input features and the output class.

That is why I had to choose an algorithm that can handle missing values, non-linearity and be robust to overfitting. I decided to use CatBoost because it is also particularly useful for high-dimensional datasets with complex non-linear relationships and uses a technique which orders the categorical variables by their target and uses this information to improve the model’s performance.

I used grid search to tune the hyperparameters of the model, including the depth, the regularization strength and the learning rate. The best hyperparameters are as follows: Best hyperparameters: {'depth': 6, 'l2\_leaf\_reg': 1, 'learning\_rate': 0.1}.

The classification report generated for the trained CatBoost model indicates that the model is performing well in predicting. The precision and recall for class 0 (not booked) are both 1.00, which means that the model is able to correctly identify all instances of the not booked class. The precision for class 1 (booked) is 0.86, which indicates that out of all instances predicted as booked, only 86% are actually booked. The recall for class 1 is 0.94, indicating that the model can correctly identify 94% of instances of the booked class. The f1-score is 1.00 for class 0 and 0.90 for class 1. The high accuracy of 0.9939 for the test set indicates that the correctly predicted the class of 99.39% of the instances in the test set. Overall, these results tell us that the model is performing well on this data set.