# ONLINE TRAVEL SEARCH HISTORY ANALYSIS

**BY:**

**Alle Likhitha**

ID:QSML100420

# CONTENTS

## ABOUT DATASET

- The dataset is from one of the worlds largest online travel agency (OTA) which powers search results for millions of travel shoppers every day
- It consists of 54 features.
- The target features in the dataset are clicked, booked value, booked
- Here the clicked and booked are categorical data with ‘0’ and ‘1’s

# Given data

| | Column Name | Data Type | Description |
|---|---|---|---|
| 1 | search_id | Integer | The ID of the search |
| 2 | timestamp | Date/time | Date and time of the search |
| 3 | site_id | Integer | ID of the website point of sale (i.e..com, .co.uk, .co.jp, ..) |
| 4 | user_country_id | Integer | The ID of the country the customer is located |
| 5 | user_hist_stars | Float | The mean star rating of hotels the customer has previously purchased; null signifies there is no purchase history on the customer |
| 6 | user_hist_paid | Float | The mean price per night (in US$) of the hotels the customer has previously purchased; null signifies there is no purchase history on the customer |
| 7 | listing_country_id | Integer | The ID of the country the hotel is located in |
| 8 | listing_id | Integer | The ID of the hotel |
| 9 | listing_stars | Integer | The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized. |
| 10 | listing_review_score | Float | The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available. |
| 11 | is_brand | Integer | +1 if the hotel is part of a major hotel chain; 0 if it is an independent hotel |
| 12 | location_score1 | Float | A (first) score outlining the desirability of a hotel's location |
| 13 | location_score2 | Float | A (second) score outlining the desirability of the hotel's location |
| 14 | log_historical_price | Float | The logarithm of the mean price of the hotel over the last trading period. A 0 will occur if the hotel was not sold in that period. |
| 15 | listing_position | Integer | Hotel position on the search results page. This is only provided for the training data, but not the test data. |
| 16 | price_usd | Float | Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay |
| 17 | has_promotion | Integer | +1 if the hotel had a sale price promotion specifically displayed |
| 18 | booking_value | Float | Total value of the transaction. This can differ from the price_usd due to taxes, fees, conventions on multiple day bookings and purchase of a room type other than the one shown in the search |
| 19 | destination_id | Integer | ID of the destination where the hotel search was performed |
| 20 | length_of_stay | Integer | Number of nights stay that was searched |
| 21 | booking_window | Integer | Number of days in the future the hotel stay started from the search date |
| 22 | num_adults | Integer | The number of adults specified in the hotel room |
| 23 | num_kids | Integer | The number of (extra occupancy) children specified in the hotel room |
| 24 | num_rooms | Integer | Number of hotel rooms specified in the search |
| 25 | stay_on_saturday | Boolean | +1 if the stay includes a Saturday night, starts from Thursday with a length of stay is less than or equal to 4 nights (i.e. weekend); otherwise 0 |
| 26 | log_click_proportion | Float | The log of the probability a hotel will be clicked on in Internet searches (hence the values are negative) A null signifies there are no data (i.e. hotel did not register in any searches) |
| 27 | distance_to_dest | Float | Physical distance between the hotel and the customer at the time of search. A null means the distance could not be calculated. |
| 28 | random_sort | Boolean | +1 when the displayed sort was random, 0 when the normal sort order was displayed |
| 29 | competitor1_rate | Integer | +1 if agency has a lower price than competitor 1 for the hotel; 0 if the same; -1 if the agency's price is higher than competitor 1; null signifies there is no competitive data |
| 30 | competitor1_has_availability | Integer | +1 if competitor 1 does not have availability in the hotel; 0 if both agency and competitor 1 have availability; null signifies there is no competitive data |
| 31 | competitor1_price_percent_diff | Float | The absolute percentage difference (if one exists) between the agency and competitor 1's price (agency's price the denominator); null signifies there is no competitive data |
| 32 | competitor2_rate | Integer | |
| 33 | competitor2_has_availability | Integer | |
| 34 | competitor2_price_percent_diff | Float | |
| 35 | competitor3_rate | Integer | |
| 36 | competitor3_has_availability | Integer | |
| 37 | competitor3_price_percent_diff | Float | |
| 38 | competitor4_rate | Integer | |
| 39 | competitor4_has_availability | Integer | |
| 40 | competitor4_price_percent_diff | Float | |
| 41 | competitor5_rate | Integer | |
| 42 | competitor5_has_availability | Integer | (same, for competitor 2 through 8) |
| 43 | competitor5_price_percent_diff | Float | |
| 44 | competitor6_rate | Integer | |
| 45 | competitor6_has_availability | Integer | |
| 46 | competitor6_price_percent_diff | Float | |
| 47 | competitor7_rate | Integer | |
| 48 | competitor7_has_availability | Integer | |
| 49 | competitor7_price_percent_diff | Float | |
| 50 | competitor8_rate | Integer | |
| 51 | competitor8_has_availability | Integer | |
| 52 | competitor8_price_percent_diff | Float | |
| 53 | clicked | Boolean | if the listing is clicked 1, else 0 |
| 54 | booked | Boolean | if the listing is booked 1, else 0 |

1.Load the dataset into R or Python and identify the type of the dataset features and report them

- I have uploaded the dataset as zip file to my drive and then in google colab I have unzipped it. So now it is present in google colab.

2 .Perform summary statistics and explain what issues these statistics reveal.

- I read the dataset using pandas and performed the functions like shape, info and describe to know about the dataset
- As the data fame is huge I have not performed pandas profiling because it takes lot of time for large data frames
- Describe function gives all the statistics of the data frame.

3.Perform exploratory analysis to identify any collinearities and explain which issues collinear-ity causes.

-To perform collinearities we use data frame.corr() function. This function returns the collinearity between each feature.

- The features are not so related to each other some features like clicked, booking value , booked have high collinearity.

4. There is currently a problem in the data which will lead to inflation in the success of the metric of choice (i.e.inflated accuracy, or false reduction in loss). This will keep the model from generalizing to the test set. Please identify what it is and explain the problem.

- The problems I have identified in the data is
    - Huge amount of Null Values
    - Presence of Outliers
- Null values leads to bias in the data
- Due to presence of the outliers it leads lo wrong predictions, accuracy.

5. Propose and implement the solutions for the issues you have found. The issues listed here are not exhaustive. If you encounter any other issues, please propose and implement solutions for those as well

- To reduce these problems I have removed the features with null values > 80% , after removing them the outliers are checked by using box plot.
- The null values are filled with the mean, median, mode according to the type of data present in the features

  Ex: listing_review_score - The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available.

  I.e As the information is not available and data is continuous so it is filled with the mean values.
- Outliers are checked and removed by using IQR(Inter Quartile Range)

6.Are there any data privacy and security issues in this dataset? If so, what are they? How would you solve these problems?

- The data consists of the customer_id, site_id etc by using these values the hotel management can know about the bookings of the customers easily so the data is not so secured.

Q. This dataset has 3 possible outcome variables: clicked, booking_value and booked. Select one of those variables to model and train a machine learning model of your choice.

- The variables clicked and booked are categorical and are correlated on each other. As they are catogogorical I have applied the classification algorithms Decision Tree and SVM.

# Pandas-Profiling:

Generates profile reports from a pandas `DataFrame`. The pandas `df.describe()` function is great but a little basic for serious exploratory data analysis. `pandas_profiling` extends the pandas DataFrame with `df.profile_report()` for quick data analysis.

For each column the following statistics - if relevant for the column type - are presented in an interactive HTML report:

- **Type inference**: detect the types of columns in a dataframe.
- **Essentials**: type, unique values, missing values
- **Quantile statistics** like minimum value, Q1, median, Q3, maximum, range, interquartile range
- **Descriptive statistics** like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- **Most frequent values**
- **Histogram**
- **Correlations** highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
- **Missing values** matrix, count, heatmap and dendrogram of missing values
- **Text analysis** learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data.
- **File and Image analysis** extract file sizes, creation dates and dimensions and scan for truncated images or those containing EXIF information.

# Example snapshots of pandas profiling report

Pandas Profiling Report | Overview | Variables | Interactions | Correlations | Missing values

## Overview

Overview | Reproduction | Warnings 9

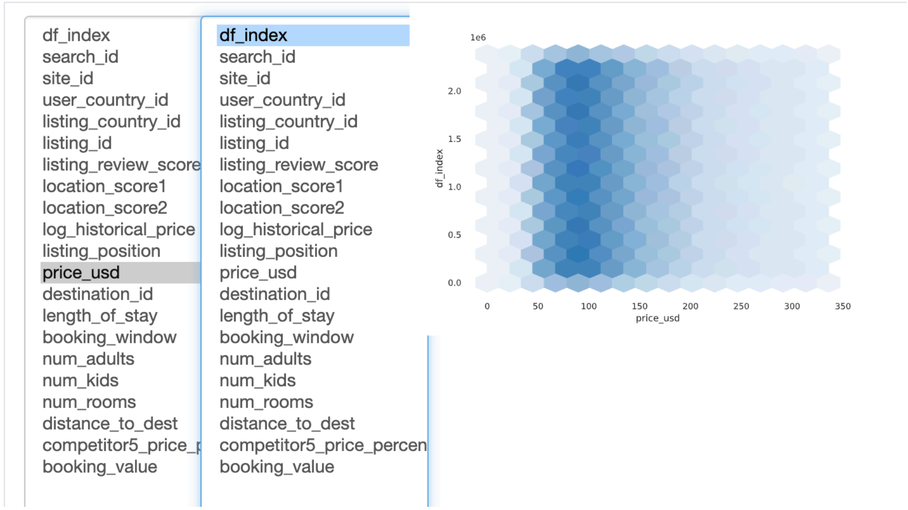
### Dataset statistics

| | |
|---|---|
| **Number of variables** | 36 |
| **Number of observations** | 1305747 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |
| **Total size in memory** | 358.6 MiB |
| **Average record size in memory** | 288.0 B |

### Variable types

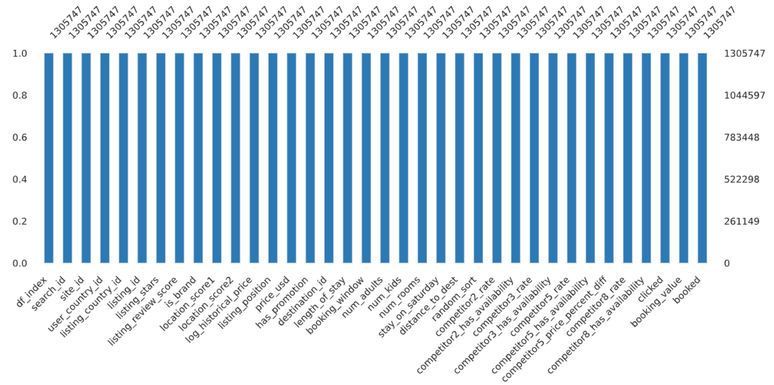
| | |
|---|---|
| **NUM** | 21 |
| **CAT** | 9 |
| **BOOL** | 6 |

## Interactions



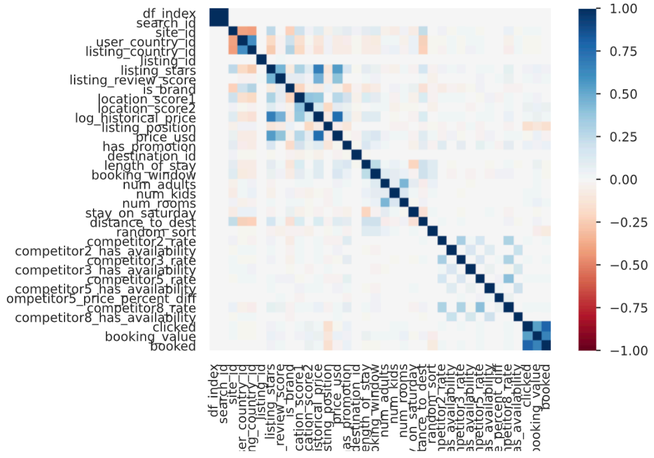

## Missing values

Count | Matrix




Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk) | Cramér's V (φc)

Toggle correlation descriptions



# SAMPLING:

## RANDOM OVERSAMPLING IMBALANCED DATASETS

- As the features clicked and booked have imbalanced data oversampl

ing is applied.

-Random oversampling involves randomly duplicating examples from

minority class and and adding them to the training dataset.

- This may increase the likelihood of overfitting, specially for higher over-sampling rates. Moreover, it may decrease the classifier performance and increase the computational effort.

```
[ ] print(sum(dataf_2['booked']==1))
    print(sum(dataf_2['booked']==0))
```

```
66388
2314169
```

```
[ ] print(sum(dataf_2['clicked']==1))
    print(sum(dataf_2['clicked']==0))
```

```
106094
2274463
```

```
[ ] print(sum(df_clean['booked']==1))
    print(sum(df_clean['booked']==0))
```

```
36378
1269369
```

```
[ ] print(sum(df_clean['clicked']==1))
    print(sum(df_clean['clicked']==0))
```

```
55346
1250401
```

## Algorithms Used:

### Decision tree classification

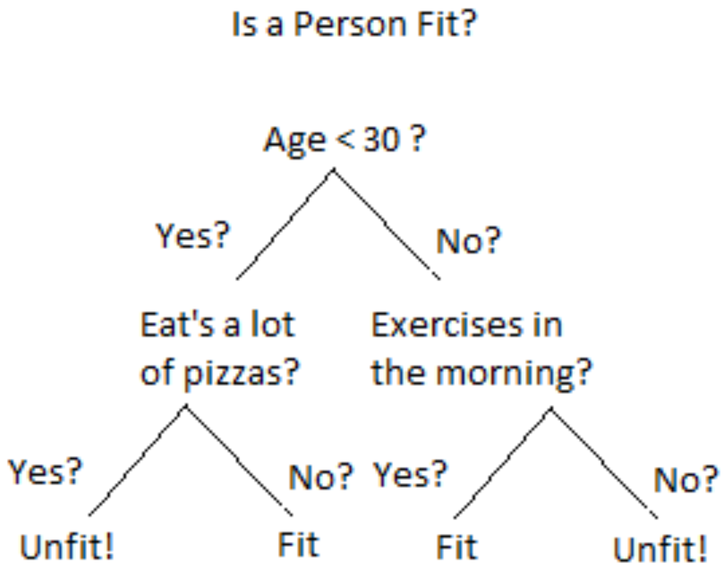
-A decision tree is a simple representation for classifying data. It is a Simple Supervised Machine Learning  where the data is continuously split according to certain parameter

- Nodes : Test for the value of a certain attribute.

- Edges/ Branch : Correspond to the outcome of a

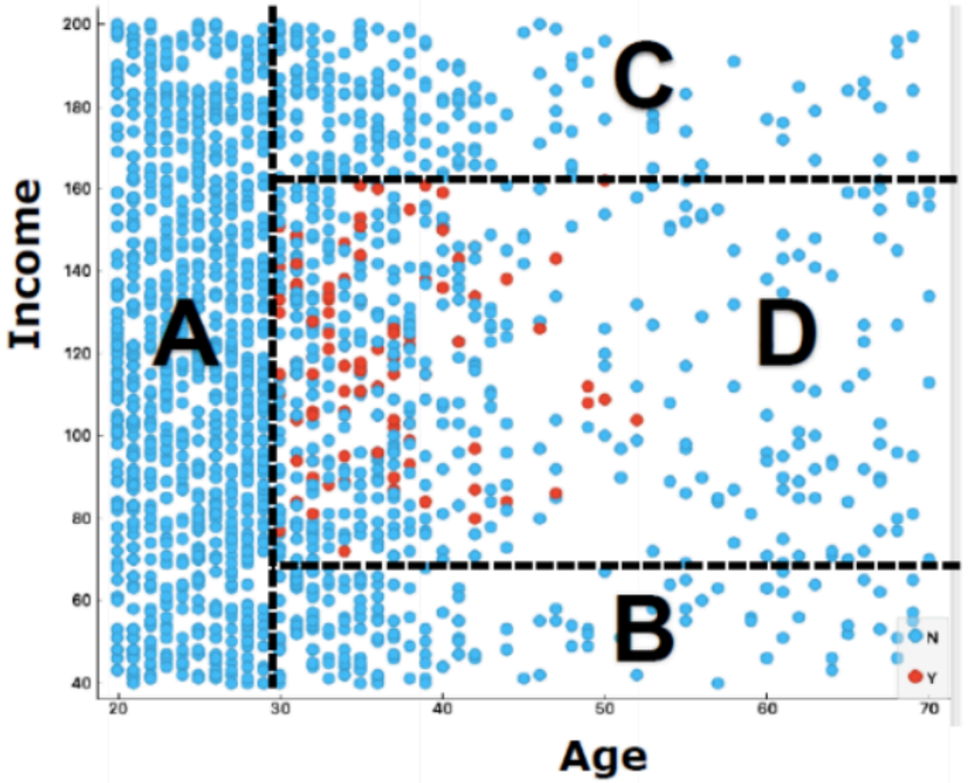
  test and connect to the next node or leaf.

-Leaf nodes : Terminal nodes that predict the outcome

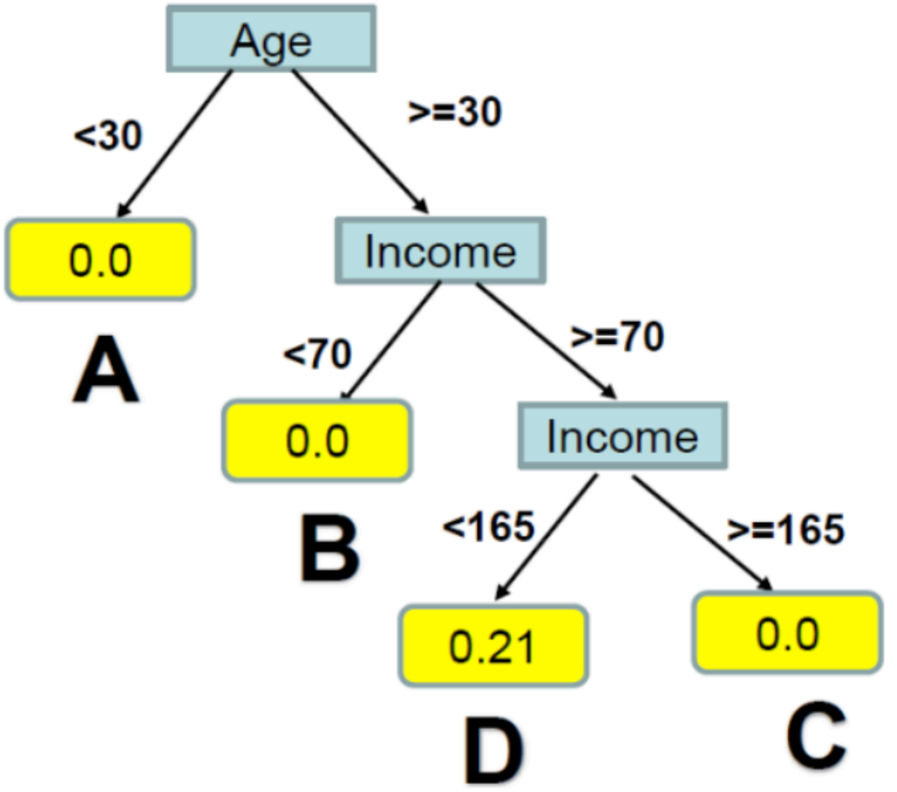



EXAMPLE DECISION  TREE

Lets say we have a decision tree partly built with 4 regions as shown




*Should the tree be grown further.*
*If so, how?*




Regions “A”, “B” and “C” are “pure”. They should not be split.
Region D should be considered for splitting.

```
F1 Score: 0.986
accuracy Score: 0.986
```

- **F1 Score: F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance. In other words, an F1-score (from 0 to 9, 0 being lowest and 9 being the highest) is a mean of an individual's performance, based on two factors i.e. precision and recall.**

- **Accuracy : Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:**

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- **For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:**

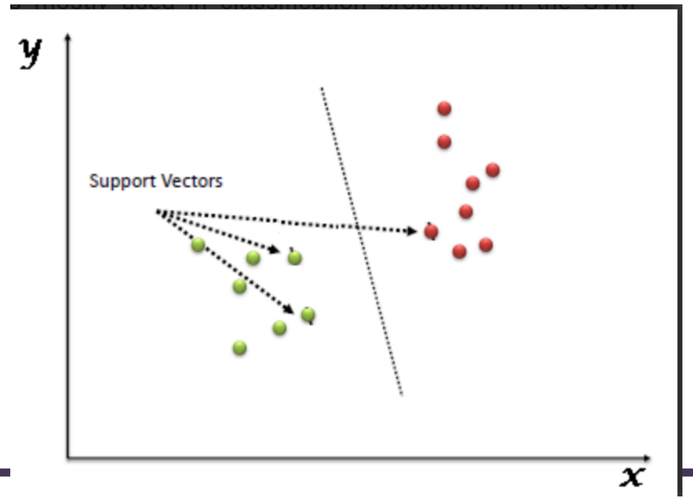
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.**

# SVM Algorithm

- Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

- Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes



# CONCLUSION

- The data has even more information to dig from as I have only performed  two algorithms for classification still we can do so much using the information present in the data set. The important steps we have to do are checking null values , if any then do the suitable methods to eliminate them , removing outliers , sampling the imbalanced data and applying  the regression and classification algorithms on the data.

# THANK YOU