Methodology Document

# Technical Specifications

## Physical

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Machine Type** | **Model** | **Processor** | **RAM** | **GPU** |
| 1. | Server |  | Intel(R) Xeon(R) CPU @  2.20 GHzR) Xeon( | *16 GB* | *14 GB* |

## Software

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **OS/Software** | **Version** | **Details (any specifics)** | **URL** |
| 1. | *Linux* SMP Debian | 4.9.144-3.1 |  |  |

# Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Treatment** | **Details** |
| 1. | season\_holidayed\_code | Missing values | Treated as new category and filled empty places with 5. |
| 2. | state\_code\_residence | Missing values | Treated as new category and filled empty places with 39. |
| 3. | total\_pax | Removed Outliers | Removed rows whose total-pax count was less than 15. |
|  |  |  |  |

# Feature Engineering

## Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Transformation** | **Details** |
|  |  |  |  |

## Derived Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **New Column Name** | **Treatment** | **Details** |
| 1. | checkin\_date\_weekday | Extraction | Created new column, the weekday number from checkin\_date |
| 2 | total\_people\*night | Combination | Created new column, total\_pax multiply room nights |
| 3 | difference\_booking\_checkin | Difference | Created new column, number of days between booking\_date and checkin\_date |
| 4 | difference\_checkout\_checkin | Difference | Created new column, number of days between checkout\_date and checkin\_date |
| 5 | difference\_checkout\_booking | Difference | Created new column, number of days between checkout\_date and booking\_date |
| 6 | total\_people | Combination | Created new column, total number of people travelling adults and children |
| 7 | total\_people\*diff | Combination | Created new column, total\_people and number of days between checkin and checkout date |
| 8 | year | Extraction | Created new column, year in which the person is travelling, extracted from checkin date |
| 9 | month | Extraction | Created new column, month in which the person is travelling, extracted from checkin date |
| 10 | checkout\_yearday | Extraction | Created new column, number of day in a year ,extracted from checkout date |
| 11 | checkin\_yearday | Extraction | Created new column, number of day in a year ,extracted from checkin date |
| 12 | mix45 | Combination | Created new column using main product code and difference between checkin and checkout dates |

# Exploratory Data Analysis

EDA

Many of the features consists of categorical features.

1.So my first approach was to find what kind of categories exists in different columns, i.e the different value counts present in different features.

2.Many of the numerical features were actually categorical. So, the first thought was to convert them using a mix of label encoder and one-hot encoder.

3.During my analysis for date features, I found out that many booking dates and check-in dates were from 2012 but that was strange, So I assumed it was the default value for the dates or it may be a system generated error.

4.In my further analysis I found out that total-pax > 15 are actually outliers and to remove them.

5.Also there were missing values in season-holidayed-code and state-code-residence, initially I didn’t have much information about these missing values, so I assumed it to be a different category.

6.There was correlation between number of person travelling and hotel type, so I thought to generate different features depending on them.

7.There was also a Strong relation between The type of traveler and the target.

8.Also the data for the number of children was having outliers.

9.There was also some strange patterns in which room-nights value was in negative, I didn’t got the time to further investigate in this.

10.As the large number of features were categorical and I thought it would be good to use target encoding, so my first choice was CatBoost Regressor as it implements both the handling of categorical features as well as target encoding.

# Model Run

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run No.** | **Model** | **Metric** | **Value** | **Hyperparameter values** | |
| 1 | CatBoost | RMSE | Train:  Test: | | eta: 0.02999999933, max\_depth:10, nthread:1, random\_state: 42,  task\_type: ‘GPU’,  eval\_metric: ‘RMSE',  loss\_function: 'RMSE',  depth: 10,  iterations: 2200 | |

# Coding Details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Programming Language** | | **Package Used** | **Details** |
| 1. | Python | | CatBoost | catboost.CatBoostRegressor |
| 2. | Python | | Pandas | pandas |
|  | |  | | |

# Platforms/Tools Used (if any)

|  |  |  |
| --- | --- | --- |
| **S.No** | **Platform Tool** | **Details** |