Methodology Document

# Technical Specifications

## Physical

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Machine Type** | **Model** | **Processor** | **RAM** | **GPU** |
| 1. | Laptop | *Asus R558UQ-DM513D* | *Upto 3.1GHz Intel Core i5-7200U 7th Gen processor* | *12 GB* | *2GB Nvidia 940MX* |
| 2. | Google Colab |  |  |  |  |

## Software

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **OS/Software** | **Version** | **Details (any specifics)** | **URL** |
| 1. | *WIndows 10* | *10* |  |  |
| 2. | *Python3* | 3.6 |  |  |

# Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Treatment** | **Details** |
| 1. | season\_holidayed\_code | Missing value | 3.33% missing value. Imputed 0 in place of missing values. |
| 2. | state\_code\_residence | Missing value | 1.39% missing value. Imputed 0 in place of missing values. |
| 3. | checkin-booking | Negative Values | Imputed 0 in place of negative values. |

# Feature Engineering

## Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Transformation** | **Details** |
| 1. | checkin\_date | Changed Date format | pd.to\_datetime(data.checkin\_date, dayfirst=True) |
| 2. | booking\_date | Changed Date format | pd.to\_datetime(data.booking\_date,dayfirst=True) |
| 3. | checkout\_date | Changed Date format | pd.to\_datetime(data.checkout\_date,dayfirst=True) |
| 4. | Column 15,17,18,19 | Label Encoding | Converted categorical data into numbers |

## Derived Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **New Column Name** | **Treatment** | **Details** |
| 1. | checkin-booking | Difference | (data['checkin\_date'] - data['booking\_date']).dt.days |
| 2 | checkout-checkin | Difference | (data['checkout\_date'] - data['checkin\_date']).dt.days |

# Model Run

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run No.** | **Model** | **Metric** | **Value** | **Hyperparameter values** | |
| 1 | XGBoost | rmse | Train:  Test: | | booster='gbtree', colsample\_bylevel=1,  colsample\_bytree=0.8, gamma=0.5, learning\_rate=0.1,  max\_delta\_step=0, max\_depth=5, min\_child\_weight=5, missing=None,  n\_estimators=250, n\_jobs=1, nthread=-1, objective='reg:linear',  random\_state=0, reg\_alpha=10, reg\_lambda=1, scale\_pos\_weight=1,  seed=None, silent=True, subsample=0.8 | |

# Coding Details

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Programming Language** | **Package Used** | **Details** |
| 1. | Python | Scikit learn | sklearn.model\_selection.train\_test\_split sklearn.preprocessing.LabelEncoder |
| 2. | Python | XGBoost | xgb.XGBRegressor |

# Platforms/Tools Used (if any)

|  |  |  |
| --- | --- | --- |
| **S.No** | **Platform Tool** | **Details** |
| 1. | Google Colab | Google Colab was also used for model training. |