Methodology Document

# Technical Specifications

## Physical

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Machine Type** | **Model** | **Processor** | **RAM** | **GPU** |
| 1. | Desktop | *Dell Inspiron 13 5000* | *intel core i5* | *16GB* |  |
|  |  |  |  |  |  |

## Software

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **OS/Software** | **Version** | **Details (any specifics)** | **URL** |
| 1. | *R version 3.5.1* |  |  |  |

# Feature Summary

# Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Treatment** | **Details** |
| 1. | season\_holidayed\_cod | Missing value | 0.03% missing value. Performed feature imputation using 0 (custom value to retain uniqueness of missing data) |
| 2. | state\_code\_residence | Missing Value | 1.43% missing value. Performed feature imputation using mode value 8 |
| 3. | booking\_date | Date format | converted date to yyyy-mm-dd from dd-mm-yy |
| 4. | checkin\_date | Date format | converted date to yyyy-mm-dd from dd-mm-yy |
| 5. | checkout\_date | Date format | converted date to yyyy-mm-dd from dd-mm-yy |

# Feature Engineering

## Derived Variable

|  |  |  |
| --- | --- | --- |
| **S.No.** | **New Column Name** | **Details** |
| 1. | season | four values based on the month of checkin date |
| 2 | diff\_book\_checkin | difference of booking\_date and check in date |
| 3. | no\_of\_days\_booked | difference between checkout date and checkin date |
| 4. | extra\_stay | difference between roomnights and no\_of\_days\_booked |
| 5. | extra\_pax | difference between (noofadults+noof children) and total\_pax |

# Exploratory Data Analysis

EDA

1. Missing value treatment
2. Outlier analysis - 14 rows where where checkin date < booking\_date, roomnights having negative value of -45.
3. Univariate and Bivariate analysis
4. Correlation between numeric variables
5. One hot encoding for 05 categorical variables

# Model Run

|  |  |  |  |
| --- | --- | --- | --- |
| **Run No.** | **Model** | **Metric** | **Value on Public leaderboard** |
| 1 | Linear Regression | RMSE | 99.922755342454 | |
| 2 | PCR | RMSE | 99.9203059391169 | |
| 3 | Ridge | RMSE | 99.9203689052875 | |
| 4 | StackedEnsemble  (using H2O automl) | RMSE | 97.2235878563442 | |
| 5 | deeplearning(using H2O) | RMSE | 98.2911017636142 | |

# Coding Details

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Programming Language** | **Package Used** | **Details** |
| 1. | R | data.table  lubridate  h2o  car  MASS  glmnet  pls  dummies  ggplot2 | Well commented code file is attached |
|  |  |  |  |

# Platforms/Tools Used (if any)