Methodology Document

# Technical Specifications

## Physical

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Machine Type** | **Model** | **Processor** | **RAM** | **GPU** |
| 1. | Server | *Dell Precision T7810* | *Dual Xeon e5-2630 v3 (16 cores, 32 threads)* | *128 GB* |  |
| 2. | Google Cloud |  |  |  |  |
| 3. | [Intel Movidius Neural Compute Stick](https://software.intel.com/en-us/movidius-ncs) |  |  |  |  |
|  |  |  |  |  |  |

## Software

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **OS/Software** | **Version** | **Details (any specifics)** | **URL** |
| 1. | *Ubuntu 16.04 LTS* | *16.04 LTS* |  |  |
| 2. | *R version 3.5.1* |  |  |  |
| 3. | *data.table 1.11.8* |  |  |  |
|  |  |  |  |  |

# Feature Summary



# Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Treatment** | **Details** |
| 1. | V2 | Missing value | 20% missing value. Performed feature imputation using knn |
| 2. | V3 | Data Type mismatch | Mixed type column. Removed the character type |
|  |  |  |  |
|  |  |  |  |

# Feature Engineering

## Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Column Name** | **Transformation** | **Details** |
| 1. | V2 | Natural Log | Found skewness in data |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Derived Variable

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **New Column Name** | **Treatment** | **Details** |
| 1. | new\_V1 | Discretization | Created five buckets for column V1 |
| 2 | new\_V1 | Quartile | Created a new column with quartiles of the column V1 |
|  |  |  |  |
|  |  |  |  |

# Exploratory Data Analysis

EDA

1. Suggest Hypothesis
2. Assess assumptions
3. Univariate and Bivariate analysis
4. Correlations, Skewness, Kurtosis etc.
5. Support for model selection
6. etc

# Model Run

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run No.** | **Model** | **Metric** | **Value** | **Hyperparameter values** | |
| 1 | XGBoost | AUC | Train:  Test: | | eta:0.835185305133136, max\_depth:12, min\_child\_weight:3.12866778927418, subsample:0.55903795955237, colsample\_bytree:0.729171768995002, colsample\_bylevel:0.263564719818532, lambda:13.9559372190669, alpha:0.00924466163302066, nthread:1, nrounds:2824, seed:1, | |
| 2 | Ensemble | AUC | Train:  Test: | | eta:0.789194774407523, min\_child\_weight:3.01012772712698, subsample:0.577084011211991, colsample\_bytree:0.667837589979172, colsample\_bylevel:0.642051564063877, lambda:0.0052829360042111, alpha:0.0173531776151954, nthread:1, nrounds:4390, openml.seed:1, | |

# Coding Details

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Programming Language** | **Package Used** | **Details** |
| 1. | Python | Scikit learn | [sklearn.linear\_model](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model).LogisticRegression |
| 2. | Python | hyperparameter-hunter | opt.BayesianOptimization |
|  |  |  |  |

# Platforms/Tools Used (if any)

|  |  |  |
| --- | --- | --- |
| **S.No** | **Platform Tool** | **Details** |
| 1. | H2O driveless AI | Details about setup, experiments, model flow. Provide screenshot of the model flows |
| 2. | Google Auto ML | Details about setup, experiments, model flow. Provide screenshot of the model flows |
| 2. | Microsoft Azure ML Studio | Details about setup, experiments, model flow. Provide screenshot of the model flows |