AI511 GROUP 6 HACKATHON

Assignment 2 Classification Task : Safe To Surf

Manan Mehta IIITB (MT2021517)

Vishwajeet Deulkar IIITB (MT2021154)

PROBLEM STATEMENT AND DATASET INTRO

- The goal of this assignment is to, predict whether it is safe to surf the next day or not with dataset having 26 features.
- Dataset contains split train (99535 rows x 26 columns) and test (42658 rows x 26 columns)
- The problem has been solved using classification models (logistic regression without outlier, SVM without outlier, logistic regression with outlier, PCA + Logistic regression)
- Workflow : 1) Data Preprocessing on train-test data by merging (Dealing With Missing Values, One Hot Encoding, Label Encoding, Dropping least correlated columns) 2) Outlier Detection and removal 3) Feature Scaling 4)Train test split 5) Applying model and predicting output probabilities.

DATA PRE-PROCESSING AND FEATURE ENGINEERING

- 1) Dealing With Missing Values: In our dataset we have missing values for many columns we replaced it with mean of each column respectively as column values were continuous and mean & median was approx. Same.
- 2) Dropping least related columns: After checking in the correlation matrix and logically day, month, year columns were had no direct relationship with prediction.
- 3) Label and One-Hot encoding: Label encoding used for categorical column with yes-no values and for rest categorical columns we used one-hot encoding.
- 4) Outlier Detection and removal: From plots we got to know features had skewed distribution, hence we used Inter-Quartile Range (IQR) proximity rule for outlier detection and removal.
- 5) Feature Scaling: Robust scaling is used to handle outliers & skewness.
- 6) Train & Test split : Already train & test data split has been provided.

MODELS USED AND KAGGLE SCORE

Preprocessing	Model	Hyperparameters	Kaggle score (if submitted)
Dealing With Missing Values, One Hot Encoding, Label Encoding, Dropping least correlated columns and Feature Scaling	logistic regression (Finally submitted)	Tried with diff max_iter parameter which converges faster	Private score: 0.87284 Public score: 0.86766
Dealing With Missing Values, One Hot Encoding, Label Encoding, Dropping least correlated columns, Outlier detection and removal and Feature Scaling	logistic regression without outlier (Finally submitted and accepted)	Tried with diff max_iter parameter which converges faster	Private score: 0.87514 Public score: 0.87053
Same as above	PCA + logistic regression without outlier	Tried with diff max_iter parameter which converges faster	Private score: 0.87270 Public score: 0.86862
Same as above	SVM without outlier(Finally submitted)	Tried with diff C parameter which gave better prediction values	Private score: 0.87475 Public score: 0.87020
Same as above	logistic regression (Using less features taken from correlation matrix)	Tried with diff max_iter parameter which converges faster	Private score: 0.87270 Public score: 0.86862