Explaining Character Stability Index (CSI) and Population Stability Index (PSI)

# Introduction

This document explains the concepts of Population Stability Index (PSI) and Characteristic Stability Index (CSI), using a Python script that computes these indices for both continuous and categorical variables across training, testing, and out-of-time (OOT) datasets. These indices are vital for detecting data shift and ensuring the stability of machine learning models over time.

# Characteristic Stability Index (CSI)

CSI evaluates the stability of input features between two datasets. It's useful for pinpointing which specific features have changed in distribution, potentially leading to reduced model performance.

CSI for Bin of Variable:

CSI for Variable:

Interpretation of PSI/CSI values:

- CSI < 0.1: No significant change; model is stable.

- 0.1 ≤ CSI < 0.2: Moderate change; monitor the model.

- CSI ≥ 0.2: Significant change; model retraining may be necessary.

# Population Stability Index (PSI)

PSI quantifies the shift in the distribution of dependent variable(probability) between two datasets, typically between training and a subsequent dataset like testing or OOT. It's instrumental in identifying data drift that may impact model performance.

# Overview of the Python Script

## 1. Data Loading and Preprocessing:

- Loads training, testing, and OOT datasets.  
- Handles missing/negative values by replacing them with -99.

## 2. Variable Segregation:

- Segregates variables into continuous and categorical types. In the given script, all are treated as continuous.

## 3. PSI Calculation Functions:

- `calculate\_psi\_continuous`: Uses quantile-based binning for PSI calculation of continuous variables.

* Uses qcut to create quantile-based bins.
* Calculates frequency distributions for both datasets.
* Applies the PSI formula using relative proportions and log-ratios.

- `calculate\_psi\_categorical`: Compares frequency distributions for categorical variables.

## 4. PSI Computation:

- Computes PSI for both test and OOT datasets relative to the training dataset.

## 5. Result Compilation:

- Combines PSI results into a final DataFrame.

# Code:



# Conclusion

Computing PSI and CSI regularly helps ensure model reliability by detecting when input data distributions shift. This allows for timely intervention, such as retraining models, thereby maintaining consistent model performance.

# Reference

For more information, refer to the article on Medium: <https://parthaps77.medium.com/population-stability-index-psi-and-characteristic-stability-index-csi-in-machine-learning-6312bc52159d>