CS4622 - Machine Learning
Lab 01 - Feature Engineering

Dr.R.T.Uthayasanker

August 5, 2023

This is an individual assignment!
Due Date: 15 August 2023 by 12.15 PM

Data-set Description

1. For this lab, 2 CSV files have been provided.
   - **train.csv :** Training data set with 28,520 rows and columns with 256 features and 4 target labels
   - **valid.csv :** Validation data set with 750 rows and columns with 256 features and 4 target labels
2. Both CSV files are generated using the dataset **AudioMNIST**.
3. The first 256 columns are 256 values of the speaker embedding vector of each audio file in the data set AudioMNIST created using **wav2vec-base**. The last 4 columns are speaker-related labels corresponding to each speaker embedding vector.
   - Label 1 - Speaker ID
   - Label 2 - Speaker age
   - Label 3 - Speaker gender
   - Label 4 - Speaker accent
4. Both the train and validation data sets can be downloaded from the link given below
   - train.csv
   - valid.csv

Assignment Tasks

- Your task is to apply all that you learned about feature selection & engineering for each target label.
  1. Feature selection/removal: Eg. using data cleaning/feature scoring techniques (SHAP values)
  2. Feature engineering
  3. Feature crossing

4. Any other advanced feature engineering techniques
5. Dimensionality Reduction
6. Etc. . .

- Finally, you should give the reduced set of features enough to predict each target label.

Note : There are some **missing values** in the label 2 column and the label 4 column is not equally distributed. Consider these things when you are applying feature engineering techniques.

Evaluation

- We have another CSV file called test.csv with 750 rows. That will be given on **15th August 2023** during the lecture at **Level 01 lab**.
- On that day you should be able to give these details of your model for each label in that dataset

1. Number of features
2. Accuracy
3. Precision
4. Recall

- In addition, you should submit a **report** comprising the illustrations and short descriptions of the tasks you performed

References

1. Feature selection techniques in machine learning
2. Machine Learning Explainability - A kaggle short course