Data Science Case
Lexplore, 2021

Machine Learning modeling

A subset of the dataset from a American research study consists of 1000 screening sessions, here with the aloud reading part of the session.

Each text has the following subset of features in the case dataset:

- fix Time mean
- fix Time sd
- progfix Time mean
- regfix Time mean
- fix BXstd mean
- sac BXdst sd
- sweep BXdst mean
- sweep BXdst sd
- fixToWordRatio
- pNoiseDur

The target variable is **WCPM**.

The task is to train a supervised machine learning model (maybe compare two different models?) and evaluate the model performance. Please use the **MldataCase.csv** file for this part of the case.

Statistical distribution

The dataset for building statistical distributions consists of 5491 read texts for oral reading, with two texts in each screening session for grade 1. Please fill in of number of screening sessions and students per grade:

Grade	1	2	3	4	5	6	7	8
Number of screening sessions								
Number of students								

The number of screening sessions can be found in the variable **screeningId** and the number of students can be found in variable **pupilId** in the **StatisticalData.csv** dataset.

The next part of the statistical distribution setup is to do parameter estimations ***mu*** and ***sigma*** for different time periods for the target variable predictions ***WCPM_aloud_pred***. Assume that the data is normally distributed. Please fill in the table:

Screening month	13-17	18-24	25-29	30-37	38-41	42-48
Mu						
Sigma						
Number of observations						

After the parameter estimation is made. Please use the Cumulative Distribution Function for a Normal distribution with inputs ***mu***, ***sigma*** and ***WCPM_aloud_pred*** to get the ***percentile*** for every observation in each time period. Save the results in a new column in the **StatisticalData.csv** dataset.