

Yumi Kim, Aluko Ademola, Jeong Hyeun Ko and Heesop Kim*

Department of Library & Information Science, Kyungpook National University (KNU), 80 Daehak-ro, Buk-gu, Daegu, Republic of Korea

*Corresponding author: Heesop@knu.ac.kr

ABSTRACT

In order to achieve the goal of the CET sub-task of the NTCIR-16 RCIR, we applied and compared five regression models: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, eXtreme Gradient Boosting (XGB) Regressor, and Voting Regressor. As the results of the analysis, we found that Gradient Boosting and Random Forest Regressor generally show better performance with Spearman's ρ of 0.53 and 0.57, respectively. And the feature importance analysis indicated that each participant shows different eye-tracking tendencies for their reading comprehension.

Findings in our work may bring insight into the understanding of human reading and information seeking processes with the help of eye-tracking systems by applying various regression models.

Methods



Implementation

- Python
- Scikit-learn packages
- Pandas
- Numpy and so on.



Five regression model

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- XGB Regressor
- Voting Regressor



Dataset

- Obtained in the NTCIR-16 RCIR Task

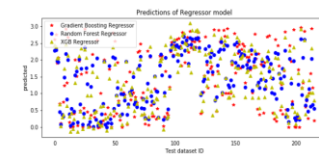
RESULTS

Table 3: Running Result from NTCIR-16 RCIR *

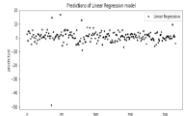
	ρ	p -value
Gradient Boost Regressor	0.53186	< 0.00000
Random Forest Regressor	0.57061	< 0.00000
Linear Regression	0.05021	< 0.46292
Voting Regressor	0.31124	< 0.00000

* The result of XGB Regressor is not included due to late submission.

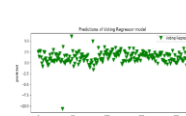
- ❑ The actual results of the submitted data from NTCIR-16
- ❑ Gradient Boosting and Random Forest Regressor show better performance.
- ❑ Gradient Boosting Regressor, Random Forest Regressor, and XGB Regressor perform better with the prediction values reside from 0 to 3.



(a) Gradient Boost Regressor, Random Forest Regressor, and XGB Regressor



(b) Linear Regression



(c) Voting Regressor

Figure 1: Scatter plots for prediction with test data.

Feature Importance Analysis

- RATE_X_BWD
- RATE_BLINK
- FIXA_X_FWD_tr_max
- FIXA_X_FWD_maxmin

SUMMARY

- ❑ We applied and compared five regression models in order to predict the comprehension score based on participants' reading data and eye-tracking metadata
- ❑ Gradient Boosting and Random Forest Regressor showed better performance with the Spearman's ρ values of 0.53 and 0.57.
- ❑ [RATE_X_BWD, RATE_BLINK, FIXA_X_FWD_tr_max, and FIXA_X_FWD_maxmin] affected to the prediction result highly as the important features.