National Research University
Higher School of Economics

Faculty of Computer Science
Degree Programme in Data Science and Business Analytics

BACHELOR'S THESIS

Research project

Predicting the English Proficiency Level of Non-native Speakers Using Artificial Intelligence and Eye Movement Data

Submitted by: Baminiwatte, Arachchiralalage Ranga Nayanakantha
student of Group 191, Year 4

Approved by Supervisor: Dr. Soroosh Shalileh,
Laboratory Head: Vision Modelling Laboratory,
Center for Language and Brain.

Abstract

Majority of the people in our world speaks more than one language. English is the most common second language in the world and the prominent lingua franca. There has been relatively low formal research on the topic of utilizing Artificial Intelligence (AI) in the domain of second language(L2) acquisition. The main objective behind this work is to explore the success of utilizing eye movements data as a predictor of English language proficiency of a non-native speaker. We applied 7 machine learning methods to predict the proficiency level of English learners based on their eye movements data while reading a set of texts and their demographics, under the hypothesis that eye movement patterns are related to one's language proficiency. All the models were fine tuned in order to figure out the best versions of each model and their performances were compared. Our experiments produced impressive results, with best performing models producing less than 2% mean absolute error. Among the tested models, Random Forest regression and K nearest neighbors' models produced the best results.

Аннотация

Большинство людей в нашем мире говорит более чем на одном языке. Английский является самым распространенным вторым языком в мире и важным лингва-франка. Было проведено относительно мало официальных исследований по теме использования искусственного интеллекта (ИИ) в области изучения второго языка (L2). Основная цель этой работы — изучить успешность использования данных о движениях глаз в качестве предиктора уровня владения английским языком для человека, для которого английский язык не является родным. Мы применили 7 методов машинного обучения для прогнозирования уровня владения английским языком на основе данных о движениях их глаз при чтении набора текстов и их демографических данных, исходя из гипотезы, что модели движения глаз связаны со знанием языка. Все модели были точно настроены, чтобы определить лучшие версии каждой модели и сравнить их характеристики. Наши эксперименты дали впечатляющие результаты: лучшие модели дают среднюю абсолютную ошибку менее 2%. Среди протестированных моделей наилучшие результаты дали регрессия случайного леса и модели K ближайших соседей.

Contents

1 Introduction

English language is the most widely used lingua franca in the world [23] to communicate with people from different native language backgrounds in various endeavors including international trade, scientific research, academia and diplomatic relations. With 1.080 billion of active English second language speakers in comparison to 450 million native speakers[24] the number of L2 learners are increasing in a rapid pace. The advent of internet and the advancement of technologies have made acquisition of language skills more accessible. Studies that had been done on L2 learning patterns in a global context indicate that the purpose and perception of learning English has changed over the years. Nowadays, integrating into native English-speaking culture is no longer a major concern for language learners, whereas constructing a bicultural or a world citizen identity to integrate into a global culture and to identify themselves in the international community is far more prominent [16].

Numerous research studies have been conducted in the past to better understand the language acquisition process, challenges and effective teaching approaches of L2 English learners. In the recent past, there has been much emphasis on linguistic cognitive ability and its' relationship with the indicators of language proficiency, comprehension. One's cognitive ability is the single most important factor in language acquisition as it helps to articulate and communicate effetely with the help of cognitive functions such as thinking, learning, remembering, reasoning as well as visual and spatial abilities. Previous research has suggested that cognition functions and eye movement patterns are closely related [6]. Therefore, eye-movements may prove to be an excellent indicator of one's linguistic ability and possibly can be used as a marker to classify proficiency level in a language.

Particularly in this work, we are interested in how visual and spatial abilities, constituted through eye movements correspond to the reading ability of a non-native English learner. We applied machine learning methods to predict the proficiency level of English learners based on their eye movements data while reading a set of texts under the hypothesis that eye movement patterns are related to one's language proficiency. Looking beyond the scope of traditional language proficiency tests to predicts one's proficiency in a different language is an interesting yet somewhat questionable approach given the fact that external factors may contribute to one's bodily mechanisms such as eye-movements (For example underlying diseases or cognitive learning disordered such as dyslexia). However, in the scope of our work, we assume that participants were not affected by these external conditions and their functionality of eyes are up to it's full potential.

The performance of the applied Machine learning methods on eye movements data and an in-depth analysis of the results are given on this thesis report.

1.1 Research Objectives and Motivation

In the beginning of this research project following objectives were established:

- Exploring the success of utilizing eye movements data as a predictor of English language proficiency of a non-native speaker.
- Training multiple machines learning models, using a combination of eye-movements data and demographics data as well as using only eye-movements data to compare their performances.
- Tuning the hyperparameters of each individual model and record the settings of best performing version in each model.
- Based on the results, find out the best performing models and conduct further research on improving them.

The motivation behind this project is the realization of possibilities in utilizing neurolinguistic data to predict neurolinguistic conditions and skills using machine learning methods and Artificial Intelligence, especially with the recent advancements in the technologies of acquiring neurolinguistic data.

Previously, Identification of dyslexia among school pupils with the help of Eye movements and demographics data using a similar methodology and experimental test-bed as this project was researched by Dr. Soroosh Shalileh and his colleagues at Center for language and brain, Higher School of Economics, Moscow. Based on its successful results, exploring further possibilities of using eye-movements to predict extent of neurolinguistic abilities such language proficiency was the main incentive behind this project.

1.2 Task formulation

We formulated the process of predicting average language proficiency score as a task of regression analysis[27].

In general, for a given set of eye-movements and/or demographics features X_i and the corresponding average language proficiency score Y_i, the regression model can be expressed using the following formula:

$$Y_i = f(X_i, \beta) + e_i \tag{1}$$

Where our goal is to estimate the function $f(X_i, \beta)$ that fits the input data using both linear and non-linear regression methods. Here β denotes the unknown parameters of the model and e_i is the random statistical noise of the model. We believe regressing the proficiency score is much convenient than classifying proficiency level, as we can always simply classify participants into proficiency levels based on their score, but not vice versa.

2 Literature Review

The utilization of Artificial Intelligence (AI) in the domain of second language(L2) analysis has been rather niche area in terms of formal research, yet we believe it is an area full of potential novel research. (Dodigovic., 2007) presented an AI based error remediation application in L2 called Intelligent tutor which detected typical writing mistakes made by university students who were taking English as a second language. Related work in L2 proficiency predictions based on cognitive and linguistic ability includes (Yang et al., 2015) which proposed a L2 language proficiency predictive machine learning (ML) model trained on learner's linguistic cognition aptitude ability under an assumption of a correlation between linguistic cognitive ability and scores in L2 aptitude tests. It has been a common practice in predictive modelling to acquire parametric variable data and feed them into multiple ML to access and compare accuracies. (Yang et al., 2015) for example had used Lexical decision tasks (LDT) in reading and listening as well as semantic and translation recognitions to measure linguistic cognitive ability. The ML models which had been used in the classification task included logistic regression, Naive Bayes, multi-layer perceptron and random forest model, which the latter resulted over a 70% prediction accuracy. On the other hand, natural language processing (NLP) based language proficiency evaluation methods were actively explored even a decade ago. A method of English L2 proficiency level prediction based on a writer's lexical ability was explored in (Crossley et al,2011)[4]. In their work, 100 writing samples of L2 learners from 3 different proficiency levels (beginner, intermediate, or advanced) based on their IRLTS/ACT scores were analyzed to identify the extent of each individual writer's core lexical skills. Their experiments resulted in 70% accuracy in classification of texts. Another aspect of automated language proficiency prediction is predictions based on oral data. (de Wet et al.,2009) [20] had attempted to asses language speaking proficiency based on measures such as rate of speech (ROS), repetitions and goodness of pronunciation (GOP) and compared it with the actual ratings given by human judges. (van der Walt et al.,2008) [18]. Uses a similar technique to asses both listening and speaking proficiency via Automatic Speech Recognitions systems (ASR).

There are multiple cognitive, psychological and linguistic factors impacting reading skills. Furthermore, there is systematic influence of linguistic characteristics of the texts on eye-movements patterns of reading (Rayner et al., 2012) [12]. It can be expected that cultural and linguistic roots of a person's native language will substantially influence one's L2 acquisition skill. Previous comparison studies made on L1 and L2 reading behaviors suggest that the extent of linguistic distance between two languages has a significant impact on reading speeds. For example, German L2 English readers show almost same reading speeds as Canadian L1 speakers of English unlike Finnish L2 readers with similar component skills due to smaller linguistic distance between German and English. (Nisbet et al.,2022) [9]. L2 reading is sometimes interpreted as

linguistic threshold i.e there must be minimum threshold in L1 proficiency in order to reap benefits from any L2. (Bernhardt, E. B., Kamil, M. L., 1995) [2] states that there is a substantial consistency in the variances accounted by first language literacy (20% upwards) and linguistic knowledge (30% upwards) on second language reading process.

Previous research on reading has showed that gaze information gives valuable insights about various attributes of the reader's cognitive state. Research done on eye movement patterns in bilingual reading suggests that L2 readers tend to show patterns of longer sentence reading times, shorter saccades, more fixations and lesser word skipping in comparison to L1 readers. Majority of the studies based on eye-movements in bilingual reading has been examining the fixations directed on embedded target words (or a critical target area), without considering the global eye movement behaviors of L2 reading. (Cop et al., 2015) [3]. However, (Titone, Debra et al., 2011) [17] and (Altarriba et al., 1996) [1] uses some basic word-level eye movement measures of paragraph readings to measure L2 reading proficiency. Eye movements benchmarks in reading depends on several factors including language proficiency and cultural language roots of the reader. (Parshina et al., 2020) [11] compares basic eye movements benchmarks of Russian heritage language speakers (HS) of varied proficiency levels with L2 readers, 8-year-old children and competent adult readers. Their findings indicated that lower proficiency HS have lower skipping probabilities, higher fixation counts than higher proficiency HS and high proficient HS had similar benchmarks to monolingual children whereas low proficient heritage speakers were on par with L2 learners.

(Kuperman et al., 2022) [19] compares reading performances in first language (L1) and L2 using a test battery of component reading skills and performs cross sample analysis on predictors of fluency and comprehension in L2 reading based on MECO L2, a novel data resource presented by them. In their findings it was evident that there is an underlying distinction between comprehension accuracy and L2 reading fluency. Furthermore, real-time strategy of reading in L1 and L2 was highly consistent within the participants. However, their analysis did not incorporate demographic data nor account for the nature and amount of exposure to L1 and L2 readers. To our knowledge, the use of behavioral traces in eye movements while reading to asses English language proficiency of non-natives was previously proposed in (Berzak et al., 2018) [28], which demonstrated the utilization of gaze information to predict outcomes of standardized English proficiency tests. This was a continuation of their previous work in predicting the first language of non-native English readers based on gaze information (Berzak et al., 2017) [29]. In contrast, our work focuses on a more general prediction of proficiency score and level based on eye movements data.

3 Experiments setting

3.1 Data Description

MECO L2 is a novel openly available data resource which contains behavioral records of eye-movements in text reading in English as a second language among 543 university students with different native language backgrounds. It also contains, a test battery of component skills of each individual reader, which would enable cross sample analysis of skills that may influence the English proficiency of nonnative speakers and identify the predictors of language fluency and comprehension. The data had been collected by 12 eye-tracking labs across Europe and Americas. However, since we are interested in only non-natives whose dominant language is not English, we do not use data from Canadian participants in our calculations. The remaining participants were from Belgium, Estonia, Finland, Germany, Cyprus, Israel, Italy, Norway, Russia, Argentina and Turkey. In the original study only typical university students without special exposure to English language had been selected. Each participant was asked to read 12 English texts in a given fixed order which consisted of texts that are commonly used for course placements tests in North American colleges. At the end of each text, few fixed comprehension questions regarding the text were presented for participants to answer. Furthermore, all participants had been given a battery of tests and questionnaires which captured participant's IQ, basic demographic and education information, spelling and vocabulary knowledge, reading efficiency as well motivation to excel in the tests. While the participants were reading the 12 texts, movements of the dominant eye including fixation locations and fixation times had been recorded.

The original data contains eye movement and proficiency data related to both L1 and L2. However, since our goals is to predict L2 proficiency of participants, the following data features related to L2 were extracted to be included in our predictive models.

Fixation data:
Fix_X – Horizontal fixation
Fix_Y – Vertical fixation
Fix_Duration – Total fixation time over a word.

L2 data:
L2_spelling_skill - Spelling Recognition Test score
L2_vocabulary_size - Vocabulary Knowledge test score
vocab.t2.5- LexTALE Vocabulary Knowledge score
L2_lexical_skill- lexical knowledge test score
TOWRE_word - Sight Word Efficiency score
TOWRE_nonword -Phonemic Decoding Efficiency score

Demographics data:
IQ – IQ level of the participant

Motiv – Motivation level
Age – Age of the participant
Sex – Gender of the participant

As we are interested in predicting L2 proficiency levels, we take rescaled the features in L2 data in order to bring all features to the same scale and then mean of the all 6 features is taken to establish the target variable.

Target_Ave - Proficiency level (Target variable)

3.2 Exploratory data analysis

We applied the chosen machine learning models to two specific data sets. In the first phase of the project, data entries related to Russian natives were extracted (87 unique participants) and 'Russian data (*Ru_data*)' set was fixed. Then the given ML methods were applied to this specific data set. In the second phase, computations were done on the whole data set (*All_data*) which contained eye movements and demographics data of 476 unique participants from all 11 countries.

Figure 1 and Figure 2 corresponds to correlation heat maps of Russian data and All nations data sets which illustrates how each feature correlates with each other. The purpose behind exploring these heat maps is to identify the features that may cause multicollinearity, which would affect the performance of the models [15].

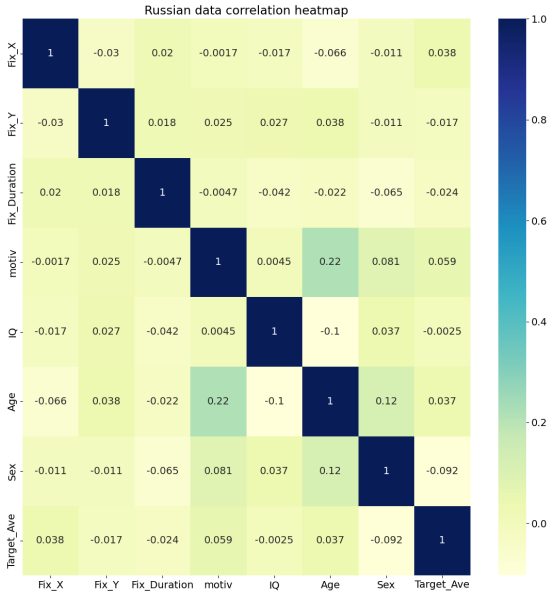

Figure 1: Correlation heat map of Russian data

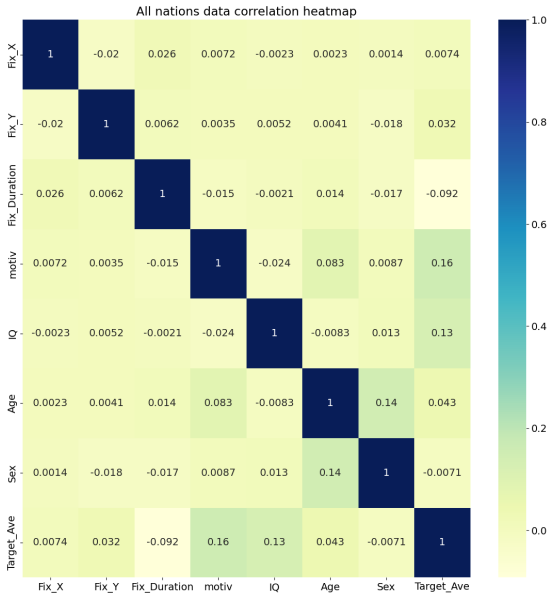

Figure 2: Correlation heat map of All-nations data

Looking at the given heat maps for both data sets, we can see that there is a relatively higher correlation between age and motivation in the Russian data set. However, there are no significant correlations between features in a way that we can declare possible multicollinearity.

Figure 3 and Figure 4 corresponds to data distribution histograms of demographic features of Russian data and All nations data sets.

In general, demographic features of both data sets have similar distributions. More male participants had taken part in the experiment. Despite, participants being university students, there seems to be some participants who were in their late thirties. However, in general majority of the participants were in their early to mid-twenties. IQ of Russian L2 learners seems to be almost normally distributed, but for all nations data, IQ distribution is slightly skewed to the right.

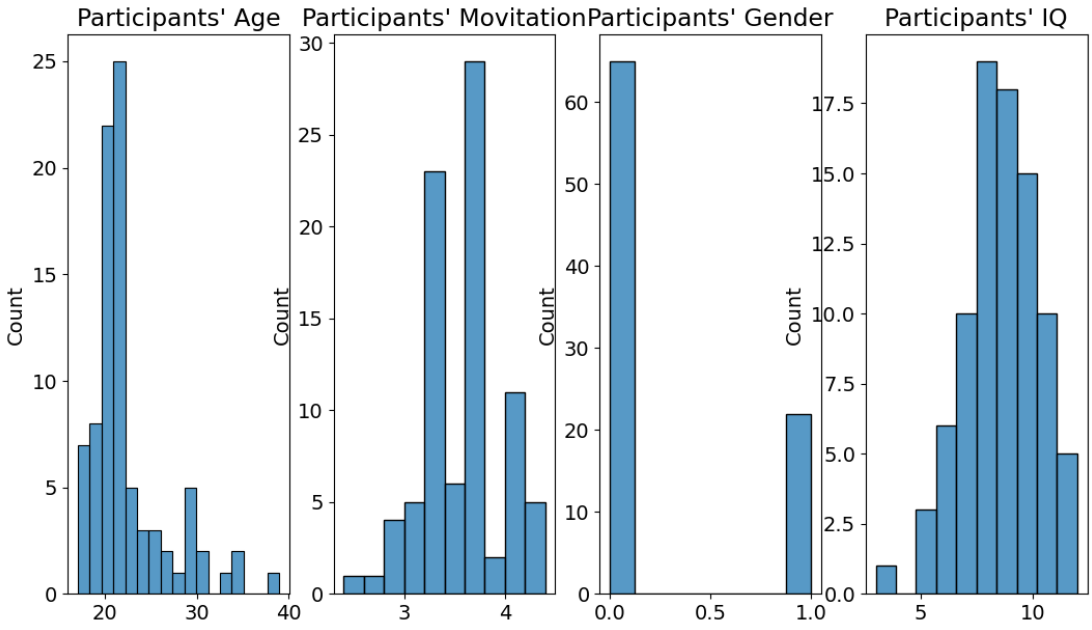

Figure 3: Demographic data distribution of Russian data

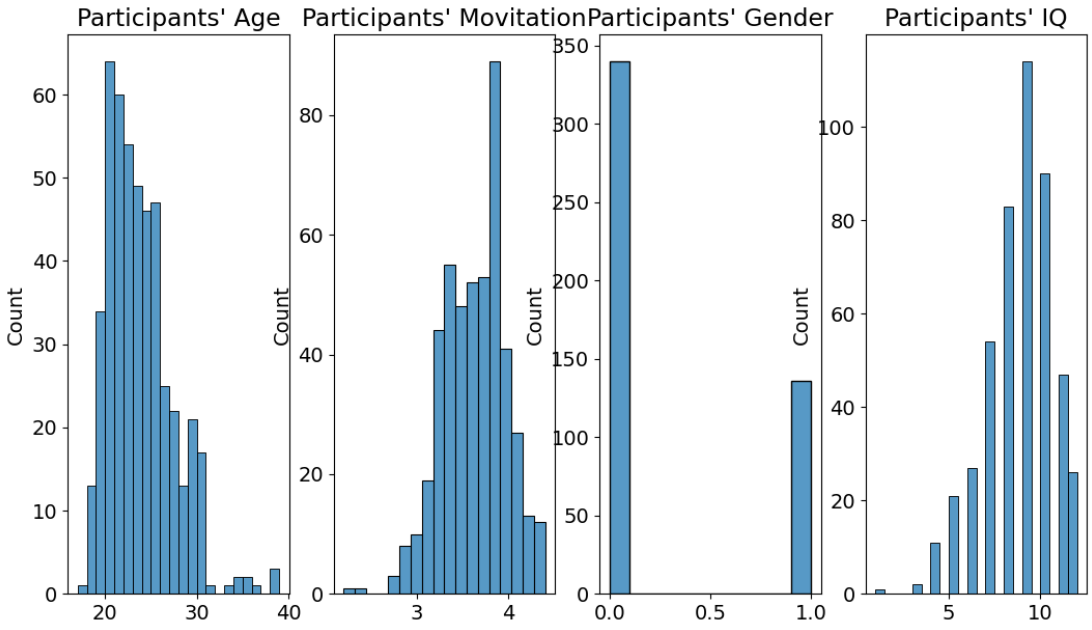

Figure 4: Demographic data distribution of All-nations data

3.3 Preprocessing techniques

Since gender of the participants is the only categorical feature, we convert it to its one-hot encoded version. After such a conversion, all the data sets and their corresponding independent variables were standardized using the MinMax technique –the target variables are intact.

3.4 Hyperparameter tuning strategy

We exploited the Bayesian optimization [8] to fine-tune the hyperparameters of the algorithms under consideration. Typically, Bayesian optimization (BO) assumes a prior probability distribution over the observations and maintains a posterior distribution using the Bayes principles. This is done by using a probabilistic model which maps hyperparameters to a probability given to the score of the objective function.

$$P(score|hyperparameters) \tag{2}$$

Bayesian optimization process takes several stages. First a surrogate probability model for objective function is created. Then hyperparameters are tuned for that surrogate. Then those tuned parameters are applied on the real objective function. While updating the surrogate model with newer results the process is repeated until the iteration limit is reached. Due to its self-learning nature, Bayesian opts are a much more efficient hyperparameter tuning strategy compared to traditional random and grid search.

We used the Random Forest and the expected improvement as our surrogate and acquisition functions respectively. In our computations, we use Scikit-optimize [7] python library to tune the hyperparameters. The algorithms' search spaces and the corresponding tuned hyperparameters are explained in the methods section.

3.5 Computational setting

Our computations consist of two constituents. First fine-tuning the hyperparameters of the methods under consideration, and second, comprehensive and fair evaluation of the fine-tuned methods.

To tune the hyperparameters, we follow the stratified k-fold cross-validation with $k = 5$ to split the data into train and validation sets randomly. We exploit BO to optimize the hyperparameters.

Once an algorithm is fine-tuned, for the sake of comprehensive and fair evaluation and to examine the possible occurrence of overfitting or underfitting, again, we apply the stratified k-fold cross-validation: this time with $k = 10$: to randomly split the data into ten disjoint pairs of train and test sets. At each fold, we train an algorithm on the train split (90% of data) and evaluate it using the unseen test split (10% of data). Finally, we report the average and the standard deviation of evaluation metrics (over these ten repeats).

3.6 Evaluation metrics

Following metrics were used to monitor and evaluate the performance of each model.

3.6.1 Mean Absolute Error(MAE)[25]

For given set of predicted values $Y = (y_1 \dots y_p)$ vs observed vales $X = (x_1 \dots x_p)$, MAE is calculated by dividing the arithmetic sum of absolute errors by the sample size (n)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3}$$

3.6.2 Mean Relative Absolute Error(MRAE)[26]

Mean Relative Absolute Error measures the prediction accuracy as a ratio of observed value. MRAE is sensitive to extreme values and outliers.

$$MRAE = \frac{1}{n} \sum_{i=1}^{n} |\frac{x_i - y_i}{x_i}| \tag{4}$$

3.6.3 R^2-SCORE [21]

R^2 score or coefficient of determination is a measure of proportion of variation of the predicted dependent variable from the observed dependent variable. For a given set of predicted values $Y = (y_1 \dots y_p)$ vs observed vales $X = (x_1 \dots x_p)$, R^2 is calculated by following formula.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

Here $SS_{res} = \sum_{i=1}^{n} (x_i - y_i)^2$ and $SS_{tot} = (x_i - \bar{x})^2$ and $\bar{x}$ - mean of observed data.

4 Methodology

4.1 Random prediction

To provide intuitive and empirical insight into the relationship between the algorithms' prediction and the evaluation metrics, we predict the target values uniformly at random by bounding the low and high range of the uniform distribution with the minimum and maximum values of the targets, and we call this case "Random prediction."

4.2 Linear models

4.2.1 Linear Regression

OLS Linear regression is a widely used statistical model in forecasting and prediction. It can be used to determine the acquaintance between the dependent variable and independent variables in a given data set. Furthermore, regression can be used for causal relationship inferences between variables. The unknown slope coefficients in the model are estimated via ordinary least squares (OLS) method [13].

The most common multiple linear regression model depicting the functional relationship between dependent variable (Y_i) and independent variable ($X_i, i = 1, \ldots, p$) is as follows.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i \tag{6}$$

Here, β_0 - constant term, $\beta_1, \beta_2, \ldots, \beta_p$ - slope coefficients and ϵ_i - residual error term.

The ordinary least squares estimator($\hat{\beta}$) in a least-squares hyperplane is expressed as:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{7}$$

Here, y - vector of the values of the target variable and X - regressor variables matrix.

Table 1: Ordinary Least Square (OLS): hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $C, N_i, l_1 ratio$, in respect, represents the inverse of regularisation strength, the number of iterations, and the Elastic-Net mixing parameter.

method / data set	parameters
	intercept
OLS	$\{False, True\}$
OLS at Ru-Demo-Fixation	$True$
OLS at All-Demo-Fixation	$False$

Table 2: Ordinary Least Square (OLS) and Logistic Regression (LR): hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $C, N_i, l_1 ratio$, in respect, represents the inverse of regularisation strength, the number of iterations, and the Elastic-Net mixing parameter.

method / data set	parameters
	intercept
OLS	$\{False, True\}$
OLS at Ru-Fixation	$True$
OLS at All-Fixation	$False$

Table 3: Ordinary Least Square (OLS) and Logistic Regression (LR): hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $C, N_i, l_1 ratio$, in respect, represents the inverse of regularisation strength, the number of iterations, and the Elastic-Net mixing parameter.

method / data set	parameters
	intercept
OLS	$\{False, True\}$
OLS at Ru-Demo	$True$
OLS at All-Demo	$False$

4.3 K-Nearest Neighbours

K-Nearest Neighbors is a popular non – parametric supervised algorithm which is used for prediction and classification for the given dataset or a problem. The number 'K' stands for the number of nearest neighbor data points. The regression prediction or classification decision for an instance is made based on the number of these nearest neighbors for the given dataset.

The algorithm directly classifies the training data set into classes of similar instances for a given K-value and makes predictions on test data set instances by searching similar 'K' neighbor instances in the training data set. The similar instances are identified via calculating the Minkowski distance between the existing instances(Y) and the new instance(X). The Minkowski distance (D) of p^{th} order is calculated using the following formula.

$$d(X,Y) = (\sum_{i=1}^{n} (x_i - y_i)^p)^{\frac{1}{p}} \tag{8}$$

Once the nearest neighbor data points are calculated, the regression prediction value of the new instance(X) is calculated by taking the local average of K-nearest neighbors.

$$X = \frac{1}{K} \sum_{i=1}^{n} x_i \tag{9}$$

where xi is the ith case of the training sample.

Table 4: K-Nearest Neighbours regression (KNN) method: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $K, p-values$, in respect, represents the number of nearest neighbours and P-value in the Minkowski distance metrics.

method / data set	parameters	
	K	$p-values$
KNN	$\{1, 2, ..., 10\}$	[1, 5]
KNN at Ru-Demo-Fixation	2	1.038
KNN at All-Demo-Fixation	4	1.039

Table 5: K-Nearest Neighbours regression (KNN) method: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $K, p-values$, in respect, represents the number of nearest neighbours and P-value in the Minkowski distance metrics.

method / data set	parameters	
	K	$p-values$
KNN	$\{1, 2, ..., 10\}$	[1, 5]
KNN at Ru-Fixation	10	1.330
KNN at All-Fixation	10	2.632

Table 6: K-Nearest Neighbours regression (KNN) method: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $K, p-values$, in respect, represents the number of nearest neighbours and P-value in the Minkowski distance metrics.

method / data set	parameters	
	K	$p-values$
KNN	$\{1, 2, ..., 10\}$	[1, 5]
KNN at Ru-Demo	8	1.204
KNN at All-Demo	9	1.504

4.4 Ensemble learning

4.4.1 Random Forests

$f(x) = E(Y|X = x)$ Random forests is a tree based ensemble learning method, which uses multitude of decision trees on a collection random variables. For a p dimensional random vector $X = (X_1 \dots X_p)^T$, and a random response variable Y, the goal of a random forest method is to find a prediction function $f(X)$ to predict Y for an unknown distribution $P_{XY}(X, Y)$. .

$f(X)$ is determined via minimizing the expected value of the loss function $L(Y, f(X))$

$$E_{XY}(L(Y, f(X))) \tag{10}$$

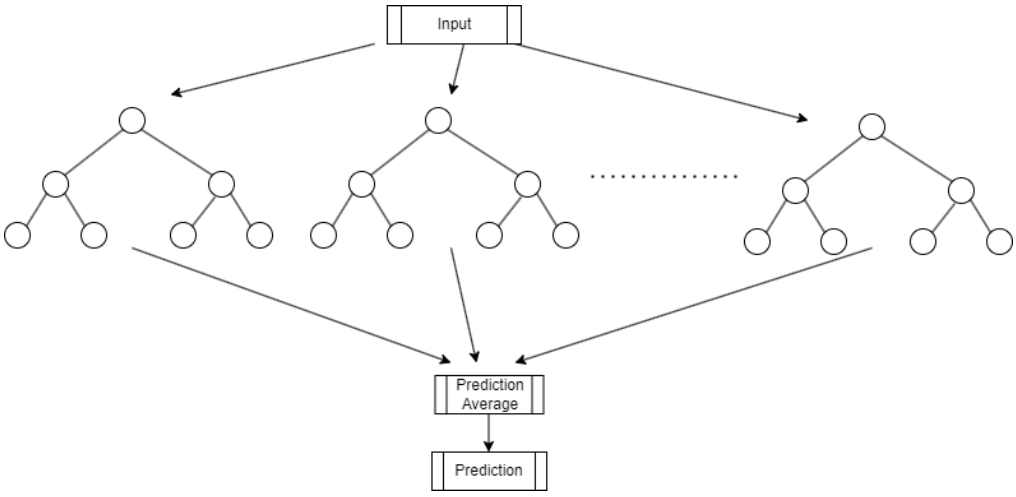

Figure 5: Random Forest regression model structure

In random forest regression, for a single 'base learner tree' the prediction

function is obtained via minimizing the expected value of the loss for squared loss error is equivalent to conditional expectation [5].

$$f(x) = E(Y|X = x)) \tag{11}$$

Set of 'base learner trees' $h_1(x), \ldots h_J(x)$ are combined to get the ensembled prediction function $f(x)$ and final prediction is the average of base learner predictions.

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \tag{12}$$

Figure 5 illustrates the ensemble of base learner trees and prediction process of Random Forest regression.

4.4.2 AdaBoost Regression

Similar to Gradient Boosting Regression, Adaptive boosting or AdaBoost regression combines weak learner regression models into a robust ensemble model.

For a given N-sample training sequence, $< (x_1, y_1), \ldots (x_N, y_N) >$ where $(x_k, y_k) \in X \times Y$. Here x_i is a input feature vector and y_i is corresponding output variable of interest.

The algorithm iteratively, trains and updates weak models by allocating weights to each data point according to model accuracy. In the process weights (W_i) from previous model is used to train the next weak model until the desired accuracy is obtained.

$$W_i = W_{i-1} * e^{\pm\alpha_t} \tag{13}$$

Here, α_t is the influence factor of t^{th} model on the final ensemble model.

$$\alpha_t = \frac{1}{2} ln \frac{1 - TotalError}{TotalError} \tag{14}$$

The final ensemble is created by combining the final weighted prediction of each weak model [14].

4.4.3 Gradient Boosting Regression

Gradient Boosting is a widely used ensemble technique in prediction and classifications, which uses a forward stage-wise fashion additive model to optimize arbitrary differentiable loss functions.

Gradient Boosting Regression merges weak prediction models such as decision trees to build a better and robust model. A Gradient Boosting regressor with M number of trees can be denoted as:

$$f_M(x_j) = \sum_{m=1}^{M} \gamma_m h_m(x_j) \tag{15}$$

Here h_m - Poorly performing weak learner (tree) and γ_m – scaling factor contribution from the m^{th} tree to the model. The gradient descent loss function is used to minimize errors in updating the estimations iteratively. The final model is an assembly of all preliminary estimations and their weights [10].

Table 7: Random Forest for regression (RF), Gradient Boosting for regression (GB), and AdaBoost for regression (AB) : hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $N_e, M_{ss}, M_{sl}, lr, \alpha$, in respect, represents the number of estimators, minimum number of samples per split, minimum number of samples per leaf, learning rate, and alpha-quantile of the Huber loss function.

method / data set	parameters				
	N_e	M_{ss}	M_{sl}	lr	α
AB	$\{10, 11, ..., 10000\}$	-	-	[1e-3, 5e-1]	-
RF	$\{10, 11, ..., 10000\}$	$\{2, 3, ..., 10\}$	$\{1, 2, ..., 10\}$	-	-
GB	$\{10, 11, ..., 10000\}$	$\{2, 3, ..., 10\}$	$\{1, 2, ..., 10\}$	[1e-3, 5e-1]	[1e-1, 9e-1]
AB at Ru-Demo-Fixation	7867	-	-	0.463	-
RF at Ru-Demo-Fixation	4245	7	10	-	-
GB at Ru-Demo-Fixation	33	6	2	0.006	0.196
AB at All-Demo-Fixation	6300	-	-	0.346	-
RF at All-Demo-Fixation	459	3	7	-	-
GB at All-Demo-Fixation	27	4	9	0.070	0.605

Table 8: Random Forest for regression (RF), Gradient Boosting for regression (GB), and AdaBoost for regression (AB): hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $N_e, M_{ss}, M_{sl}, lr, \alpha$, in respect, represents the number of estimators, minimum number of samples per split, minimum number of samples per leaf, learning rate, and alpha-quantile of the Huber loss function.

method / data set	parameters				
	N_e	M_{ss}	M_{sl}	lr	α
AB	$\{10, 11, ..., 10000\}$	-	-	[1e-3, 5e-1]	-
RF	$\{10, 11, ..., 10000\}$	$\{2, 3, ..., 10\}$	$\{1, 2, ..., 10\}$	-	-
GB	$\{10, 11, ..., 10000\}$	$\{2, 3, ..., 10\}$	$\{1, 2, ..., 10\}$	[1e-3, 5e-1]	[1e-1, 9e-1]
AB at Ru-Fixation	21	-	-	0.041	-
RF at Ru-Fixation	4895	3	10	-	-
GB at Ru-Fixation	2165	6	6	0.001	0.426
AB at All-Fixation	4441	-	-	0.363	-
RF at All-Fixation	3106	9	10	-	-
GB at All-Fixation	9509	2	8	0.002	0.102

Table 9: Random Forest for regression (RF), : hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The $N_e, M_{ss}, M_{sl}, lr, \alpha$, in respect, represents the number of estimators, minimum number of samples per split, minimum number of samples per leaf, learning rate, and alpha-quantile of the Huber loss function.

method / data set	parameters				
	N_e	M_{ss}	M_{sl}	lr	α
RF	$\{10, 11, ..., 10000\}$	$\{2, 3, ..., 10\}$	$\{1, 2, ..., 10\}$	-	-
RF at Ru-Demo	244	3	10	-	-
RF at All-Demo	13	7	10	-	-

4.5 Artificial neural networks

4.5.1 Multilayer Perceptrons (MLP)

Multilayer Perceptrons (MLP) are a type of feedforwarding neural networks which consists of multiple layers including an input layer, several hidden layers and an output layer. Figure 1 depicts a basic 3-layer MLP network. The given input values are fed to the hidden layers to result the the output values. In MLP's there are no inter connections between nuerons in the same layer.

Each node(neuron) in the network except input nodes have a nonlinear activation function (Φ). For a single neuron in the network, unit activation (a) is given as

$$A = \Phi(\sum w_i x_i + b) \tag{16}$$

Where x_i – inputs to the unit, w_i - weights and b- bias.

Weights and bias values are initially set randomly and are updated during the training using backpropagation method for a given loss function. The algorithm computes the gradient of the loss function with respect to weights and biases, and adjust them in the direction of a negative gradient for a given learning rate parameter. MLPs are helpful in distinguishing data samples that are not linearly separable.

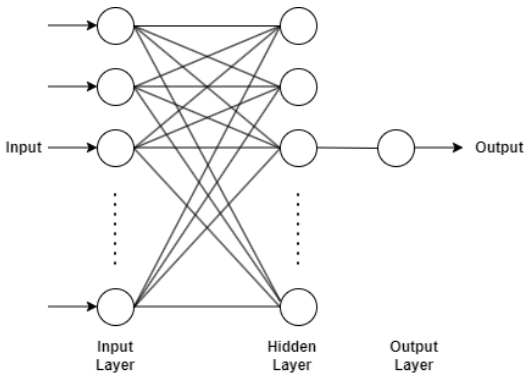

Figure 6: A basic three-layer MLP network architecture

Table 10: Multi-Layer Perceptron for regression (MLP) methods: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The N_n, N_e, lr, in respect, represents the number of neurons of the hidden layer, the number of epochs, and the learning rate.

method / data set	parameters				
	N_n	N_e	lr	activation	optimiser
MLP	{2, 3, ...200}	[100, 50000]	[1e-6, 1e-2]	{Identity, Logistic, Tanh, ReLu }	{ LBFGS, SGD, ADAM }
MLP at Ru-Demo-Fixation	97	24032	–	*RELU*	*ADAM*
MLP at All-Demo-Fixation	153	8944	–	*Identity*	*LBFGS*

Table 11: Multi-Layer Perceptron for regression (MLP) method: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The N_n, N_e, lr, in respect, represents the number of neurons of the hidden layer, the number of epochs, and the learning rate.

method / data set	parameters				
	N_n	N_e	lr	activation	optimiser
MLP	{2, 3, ...200}	[100, 50000]	[1e-6, 1e-2]	{Identity, Logistic, Tanh, ReLu }	{ LBFGS, SGD, ADAM }
MLP at Ru-Fixation	108	24629	–	*Logistic*	*SGD*
MLP at All-Fixation	39	48476	–	*RELU*	*LBFGS*

Table 12: Multi-Layer Perceptron for regression (MLP) method: hyperparameters' domain and the corresponding tuned values at the data sets under consideration. The N_n, N_e, lr, in respect, represents the number of neurons of the hidden layer, the number of epochs, and the learning rate.

method / data set	parameters				
	N_n	N_e	lr	activation	optimiser
MLP	{2, 3, ...200}	[100, 50000]	[1e-6, 1e-2]	{Identity, Logistic, Tanh, ReLu }	{ LBFGS, SGD, ADAM }
MLP at Ru-Demo	103	37145	–	*Logistic*	*ADAM*
MLP at All-Demo	118	2050	–	*RELU*	*SGD*

5 Experimental results and analysis

Computations on the two data sets were done in two approaches. First with the combination of both demographics data and eye movements fixation data. Secondly, with only using fixation data.

Over the conducted experiments, Language proficiency levels were predicted using 7 different models, for both Russian natives as well as the combined data set non-native speakers from 11 countries. Bayesian optimization method allowed us to fine tune the best parameters for given models so that we could compare models only in their best configurations.

The complete code of the experimental test bed, files of results and tuned parameters can be found in the projects' GitHub repository

5.1 Computations using fixation and demographics data

5.1.1 Results on Russian Native speakers

Table 13: Prediction results for Russian native speakers using the combination of demographic and fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Random prediction	1.444 ± 0.000	0.383 ± 0.000	-8.710 ± 0.000
Linear	0.415 ± 0.001	0.124 ± 0.000	0.036 ± 0.002
K-Nearest Neighbour	**0.010 ± 0.000**	**0.003 ± 0.000**	0.985 ± 0.001
Random Forest	**0.010 ± 0.000**	**0.003 ± 0.000**	**0.990 ± 0.000**
Gradient Boosting	0.399 ± 0.002	0.120 ± 0.001	0.087 ± 0.002
AdaBoost	0.384 ± 0.003	0.113 ± 0.001	0.219 ± 0.005
Multi-Layer Perceptron	0.066 ± 0.022	0.019 ± 0.007	0.945 ± 0.040
Average of ave. (std.)	0.003	0.001	0.006

5.1.2 Results on combined all nations data

Table 14: Prediction results for all nations using the combination of demographic and fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Random prediction	1.622 ± 0.000	0.445 ± 0.000	-8.122 ± 0.000
Linear	0.466 ± 0.001	0.145 ± 0.000	0.063 ± 0.001
K-Nearest Neighbour	0.069 ± 0.001	0.021 ± 0.000	0.875 ± 0.002
Random Forest	**0.065 ± 0.000**	**0.020 ± 0.000**	**0.899 ± 0.001**
Gradient Boosting	0.441 ± 0.001	0.137 ± 0.000	0.159 ± 0.002
AdaBoost	0.456 ± 0.001	0.140 ± 0.000	0.120 ± 0.003
Multi-Layer Perceptron	0.466 ± 0.001	0.145 ± 0.000	0.063 ± 0.001
Average of ave. (std.)	0.0007	0.0000	0.0025

After comparing all 7 models for Russian native speakers and All nations speakers with both demographic and fixations data, Random Forest regression models seem to be the model that outperforms every other model, in this context. K-nearest Neighbor model also seems to deliver promising results and falls only slightly short of Random Forest model in terms of metrics.

Comparing coefficients of determination($R^2 - score$), the performance of boosting methods is mediocre yet better than linear regression. A surprising result here was MLP regression models' performance in the all-nations data set. However, looking at the higher standard deviation in the average results of the Russian data it can be justified as we can expect MLP models to be highly volatile and might not be an ideal method in such predictive models.

5.2 Computations using fixation only data

5.2.1 Results on Russian native speakers

Table 15: Regression results for Russian native speakers using the fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Random prediction	1.367 ± 0.000	0.367 ± 0.000	-6.790 ± 0.000
Linear	0.415 ± 0.002	0.125 ± 0.001	0.005 ± 0.001
K-Nearest Neighbour	0.435 ± 0.002	0.130 ± 0.000	-0.083 ± 0.005
Random Forest	0.421 ± 0.002	0.126 ± 0.001	-0.013 ± 0.004
Gradient Boosting	**0.414 ± 0.002**	**0.125 ± 0.000**	**0.011 ± 0.001**
AdaBoost	0.416 ± 0.002	0.125 ± 0.001	0.010 ± 0.001
Multi-Layer Perceptron	**0.414 ± 0.002**	**0.125 ± 0.001**	0.000 ± 0.001
Average of ave. (std.)	0.001	0.0005	0.001

5.2.2 Results on combined all nations data

Table 16: Regression results for all nations using the fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Random prediction	1.623 ± 0.000	0.447 ± 0.000	-8.623 ± 0.000
Linear	0.481 ± 0.001	0.150 ± 0.000	0.011 ± 0.001
K-Nearest Neighbour	0.494 ± 0.001	0.154 ± 0.000	-0.079 ± 0.001
Random Forest	0.485 ± 0.001	0.150 ± 0.000	-0.012 ± 0.001
Gradient Boosting	0.479 ± 0.001	0.149 ± 0.000	0.016 ± 0.000
AdaBoost	0.488 ± 0.001	0.149 ± 0.000	0.003 ± 0.001
Multi-Layer Perceptron	**0.466 ± 0.001**	**0.145 ± 0.000**	**0.063 ± 0.001**
Average of ave. (std.)	0.0008	0.0007	0.022

When computed using only fixation features, MAE values tend be in the range of 40-50% for majority of the methods. However, R^2-scores are very low for all the methods in both Russian and All nations data sets indicating poor fit of the models.

Generally, computation results with only fixation data are worse in comparison to computations with both demographics and fixation data.

6 Conclusion and further research

In conclusion, this project presents a language proficiency prediction method which uses eye movements data to predict the average language proficiency score among L2 learners of English as opposed to conventional language proficiency tests.

We implemented and fine-tuned 7 machine learning models to fit the given two data sets which contained both a combination of eye-movements and demographics data as well as eye-movements data only. It was identified that random forest and K-nearest neighbours machine learning models produce the best results among all the models. Both these models resulted in a coefficient of determination (R^2) of more than 0.90 and less than 1% of mean absolute error for Russian natives' data set and a R^2 value of 0.85 and less than 6% of mean absolute error for all nations data set when computed with demographics and fixation data. Furthermore, it was realized that the model performance is vastly better when demographic data is incorporated in the model trainings. However, it is a concern, given the fact that our principal goal was to predict language proficiency based on eye movements fixation data and demographic data was additional information.

Table 17 and Table 18 in the appendix correspond to results when language proficiency was predicted using three of the best performing models in pervious experiments with using only demographics data and without data augmentation [22]. The results are not very impressive just like with fixation only computations. Therefore, we could conclude that neither fixation data nor demographics data by themselves are not very good predictors of English proficiency. It works the best when data sources are combined and fed into the model training process.

One of the more interesting questions that arose from analysing these results was the reason behind the success of models when both fixation and demographics features were combined and the significantly poor performance when those feature groups were used separately. A possible explanation is that incorporating demographics features increased the representativeness of data which made the model represent underlying patterns more accurately so that it could make more accurate predictions. Also, in some cases, combination of features can help to understand the relationship between target variable and the independent variables which could also improve model performance. Furthermore, we think fixation data can be noisy and augmenting demographic features cancelled out the noise created by fixation data which led to better results.

The work presented in this thesis is a part of the 'Machine Learning on Neurolinguistic Data' project of Vision Modelling Laboratory at Center for Language and Brain, Higher School of Economics. As a part of this project, a classification based on English language proficiency level was also conducted using

the same approach as the regression task which was explained above. However, since it was out of the scope of this thesis, it was not presented explicitly. Tables 13,14,15 and 16 correspond to some of the results obtained in the classification task. Looking the results in general, it can be concluded that the model that uses Random Forests classification method being the best performing model out of the bunch.

For further research, more advanced machine learning techniques such as convolutional Neural Networks (CNN), Temporal Neural Networks (TNN) and Generative Adversarial Networks (GANS) will be used to train models and will be tested for performance. The goal will be to increase the causal effect of eye movement fixation data on the language proficiency prediction while maintaining the performance of the models.

7 Appendix

7.0.1 Classification on Russian native speakers with demographics and fixations data

Table 17: Classification results for Russian native speakers using the combination of demographic and fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Classification	Metrics				
	Precision	Recall	F1-score	ROC-AUC	TNR
Random prediction	0.531 ± 0.061	0.253 ± 0.038	0.321 ± 0.044	-	0.759 ± 0.038
Logistic	0.669 ± 0.004	0.667 ± 0.002	0.609 ± 0.003	0.663 ± 0.003	0.460 ± 0.002
K-Nearest Neighbour	0.989 ± 0.001	0.989 ± 0.001	0.989 ± 0.001	**1.000 ± 0.000**	0.989 ± 0.001
Gaussian Naive Bayes	0.592 ± 0.003	0.510 ± 0.003	0.537 ± 0.003	0.675 ± 0.003	0.689 ± 0.004
Multinomial Naive Bayes	0.416 ± 0.000	0.645 ± 0.000	0.506 ± 0.000	0.627 ± 0.003	0.355 ± 0.000
Complement Naive Bayes	0.548 ± 0.003	0.578 ± 0.002	0.544 ± 0.002	0.617 0.003±	0.521 ± 0.003
Random Forest	**0.990 ± 0.000**	**0.990 ± 0.000**	**0.990 ± 0.000**	**1.000 ± 0.000**	0.987 ± 0.000
Gradient Boosting	**0.990 ± 0.001**	**0.990 ± 0.001**	**0.990 ± 0.001**	**1.000 ± 0.000**	0.987 ± 0.002
AdaBoost	0.470 ± 0.001	0.667 ± 0.001	0.547 ± 0.001	0.685 ± 0.002	0.414 ± 0.001
Multi-Layer Perceptron	0.989 ± 0.001	0.989 ± 0.001	0.989 ± 0.001	**1.000 ± 0.000**	**0.990 ± 0.001**
Average of ave. (std.)	0.007	0.004	0.005	0.001	0.005

7.0.2 Classification on all nations with demographics and fixations data

Table 18: Classification results for all nations using the combination of demographic and fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Classification	Metrics				
	Precision	Recall	F1-score	ROC-AUC	TNR
Random prediction	0.449 ± 0.027	0.251 ± 0.028	0.307 ± 0.029	-	0.752 ± 0.022
Logistic	0.506 ± 0.002	0.551 ± 0.002	0.498 ± 0.002	0.608 ± 0.002	0.541 ± 0.002
K-Nearest Neighbour	0.951 ± 0.001	0.951 ± 0.001	0.951 ± 0.001	0.994 ± 0.000	0.962 ± 0.001
Gaussian Naive Bayes	0.517 ± 0.001	0.553 ± 0.001	0.520 ± 0.001	0.619 ± 0.002	0.566 ± 0.001
Multinomial Naive Bayes	0.284 ± 0.000	0.533 ± 0.000	0.370 ± 0.000	0.572 ± 0.002	0.462 ± 0.000
Complement Naive Bayes	0.489 ± 0.015	0.515 ± 0.001	0.478 ± 0.001	0.572 ± 0.001	0.520 ± 0.001
Random Forest	**0.957 ± 0.001**	**0.957 ± 0.001**	**0.957 ± 0.001**	**0.996 ± 0.000**	**0.964 ± 0.000**
Gradient Boosting	0.955 ± 0.000	0.955 ± 0.000	0.955 ± 0.000	**0.996 ± 0.000**	**0.964 ± 0.000**
AdaBoost	0.524 ± 0.008	0.561 ± 0.004	0.468 ± 0.009	0.548 ± 0.001	0.521 ± 0.006
Average of ave. (std.)	0.006	0.004	0.004	0.0008	0.003

7.0.3 Classification on Russian native speakers with fixations data

Table 19: Classification results for Russian native speakers using the fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Classification	Metrics				
	Precision	Recall	F1-score	ROC-AUC	TNR
Random prediction	0.458 ± 0.057	0.230 ± 0.041	0.287 ± 0.046	-	0.730 ± 0.030
Logistic	0.420 ± 0.011	0.645 ± 0.000	0.506 ± 0.000	0.519 ± 0.004	0.355 ± 0.000
K-Nearest Neighbour	0.519 ± 0.004	0.595 ± 0.003	0.542 ± 0.003	0.519 ± 0.004	0.422 ± 0.003
Gaussian Naive Bayes	0.421 ± 0.001	0.640 ± 0.001	0.506 ± 0.001	0.535 ± 0.005	0.361 ± 0.001
Multinomial Naive Bayes	0.416 ± 0.000	0.645 ± 0.000	0.506 ± 0.000	0.530 ± 0.003	0.355 ± 0.000
Complement Naive Bayes	0.524 ± 0.005	0.429 ± 0.004	0.463 ± 0.004	0.528 ± 0.006	0.604 ± 0.007
Random Forest	0.542 ± 0.004	0.644 ± 0.000	0.515 ± 0.001	0.542 ± 0.003	0.362 ± 0.001
Gradient Boosting	**0.605 ± 0.014**	**0.646 ± 0.000**	**0.510 ± 0.001**	**0.557 ± 0.006**	0.358 ± 0.000
AdaBoost	0.545 ± 0.027	0.645 ± 0.000	0.507 ± 0.000	0.529 ± 0.001	0.356 ± 0.000
Multi-Layer Perceptron	0.418 ± 0.006	0.645 ± 0.000	0.506 ± 0.000	0.533 ± 0.000	0.355 ± 0.000
Average of ave. (std.)	0.019	0.004	0.005	0.004	0.003

7.0.4 Classification on all nations with fixations data

Table 20: Classification results for all nations using the fixation data set: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Classification	Metrics				
	Precision	Recall	F1-score	ROC-AUC	TNR
Random prediction	0.458 ± 0.041	0.235 ± 0.028	0.298 ± 0.028	-	0.751 ± 0.027
Logistic	0.485 ± 0.003	0.535 ± 0.001	0.405 ± 0.001	0.544 ± 0.002	0.481 ± 0.001
K-Nearest Neighbour	0.460 ± 0.002	0.491 ± 0.001	0.473 ± 0.001	0.516 ± 0.001	0.530 ± 0.001
Gaussian Naive Bayes	**0.486 ± 0.002**	0.533 ± 0.001	0.420 ± 0.001	0.543 ± 0.001	0.492 ± 0.001
Multinomial Naive Bayes	0.284 ± 0.000	**0.533 ± 0.000**	0.370 ± 0.000	0.543 ± 0.002	0.467 ± 0.000
Complement Naive Bayes	0.477 ± 0.002	0.479 ± 0.002	**0.477 ± 0.002**	0.541 ± 0.002	0.575 ± 0.002
AdaBoost	0.480 ± 0.002	0.536 ± 0.001	0.432 ± 0.001	0.531 ± 0.001	0.491 ± 0.001
Multi-Layer Perceptron	0.480 ± 0.002	0.536 ± 0.000	0.418 ± 0.000	**0.546 ± 0.002**	**0.486 ± 0.002**
Average of ave. (std.)	0.006	0.004	0.004	0.001	0.004

7.0.5 Regression on Russian native speakers with demographic data

Table 21: Regression results for Russian native speakers using demographic data only: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Linear	0.443 ± 0.059	0.131 ± 0.024	-0.147 ± 0.093
K-Nearest Neighbour	0.449 ± 0.105	0.133 ± 0.040	-0.153 ± 0.270
Random Forest	0.445 ± 0.052	0.132 ± 0.015	-0.161 ± 0.148
Multi-Layer Perceptron	**0.424 ± 0.115**	**0.126 ± 0.036**	**-0.091 ± 0.157**
Average of ave. (std.)	0.083	0.028	0.167

7.0.6 Regression on all nations speakers with demographic data

Table 22: Regression results for all nations speakers using demographic data only: the average and standard deviation of evaluation metrics over 10 different data splits. The best results are bold-faced.

Regression	Metrics		
	MAE	MRAE	R^2-score
Linear	**0.470 ± 0.039**	0.144 ± 0.012	**0.025 ± 0.057**
K-Nearest Neighbour	0.486 ± 0.049	0.149 ± 0.014	-0.075 ± 0.094
Random Forest	0.477 ± 0.039	0.145 ± 0.013	0.006 ± 0.099
Multi-Layer Perceptron	0.468 ± 0.042	**0.142 ± 0.014**	0.018 ± 0.096
Average of ave. (std.)	0.042	0.013	0.086

References

[1] Kroll-J.F. Sholl A. et al Altarriba, J. The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Mem Cogn*, 24, 1996.

[2] Elizabeth B. Bernhardt and Michael L. Kamil. Interpreting relationships between l1 and l2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16:15–34, 1995.

[3] Duyck W Cop U, Drieghe D. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLOS One 10(8): e0134008*, 2015.

[4] Scott A. Crossley, Tom Salsbury, and Danielle S. McNamara. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2):243–263, 2012.

[5] Adele Cutler, David Cutler, and John Stevens. *Random Forests*, volume 45, pages 157–176. 01 2011.

[6] John M. Henderson, Svetlana V. Shinkareva, Jing Wang, Steven G. Luke, and Jenn Olejarczyk. Predicting cognitive state from eye movements. *PLOS ONE*, 8:1–6, 05 2013.

[7] Gilles Louppe, Manoj Kuma, and Holger Nahrstaedt. Bayesian optimization with skopt.

[8] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

[9] Kelly Nisbet, Raymond Bertram, Charlotte Erlinghagen, Aleks Pieczykolan, and Victor Kuperman. Quantifying the difference in reading fluency between l1 and l2 readers of english. *Studies in Second Language Acquisition*, 44(2):407–434, 2022.

[10] Daniel Asante Otchere, Tarek Omar Arbi Ganat, Jude Oghenerurie Ojero, Bennet Nii Tackie-Otoo, and Mohamed Yassir Taki. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, 208:109244, 2022.

[11] Sekerina IA Parshina O, Laurinavichyute AK. Eye-movement benchmarks in heritage language reading. *Bilingualism: Language and Cognition 1–14.*, 2020.

[12] Pollatsek A. Ashby J. Clifton Jr. Rayner, K. *Psychology of Reading: 2nd Edition (1st ed.)*. Psychology Press, 2012.

[13] Rui Sarmento and Vera Costa. *Introduction to Linear Regression*. 01 2017.

[14] Dimitri Solomatine and Durga Shrestha. Adaboost.rt: A boosting algorithm for regression problems. volume 2, pages 1163 – 1168 vol.2, 08 2004.

[15] Aya Spencer. 5 Minute EDA: Correlation Heatmap. — Wmedium, the free encyclopedia. `medium.com/5-minute-eda/5-minute-eda-correlation-heatmap-b57bbb7bae14?title=Lingua%20franca&oldid=1153495050`, 2023. [9 Apr. 2022].

[16] Chit Sung. Learning english as an l2 in the global context: Changing english, changing motivation. *Changing English*, 20, 12 2013.

[17] Mercier J Whitford V Pivneva I Titone D, Libben M. Bilingual lexical access during l1 sentence reading: The effects of l2 knowledge, semantic constraint, and l1-l2 intermixing. *J Exp Psychol Learn Mem Cogn*, 37, Nov 2011.

[18] Christa Van der Walt, Febe Wet, and Thomas Niesler. Oral proficiency assessment: The use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies - SOUTH AFR LINGUIST APPL LANG*, 26:135–146, 06 2008.

[19] Sascha Schroeder Cengiz Acarturk Victor Kuperman, Noam Siegelman. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, March 2022.

[20] Febe Wet, Christa Van der Walt, and Thomas Niesler. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, 51:864–874, 10 2009.

[21] Wikipedia. Coefficient of determination — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Coefficient%20of%20determination&oldid=1148160373`, 2023. [Online; accessed 02-May-2023].

[22] Wikipedia. Data augmentation — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Data%20augmentation&oldid=1154785727`, 2023. [Online; accessed 17-May-2023].

[23] Wikipedia. Lingua franca — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Lingua%20franca&oldid=1153495050`, 2023. [Online; accessed 08-May-2023].

[24] Wikipedia. List of languages by total number of speakers — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=List%20of%20languages%20by%20total%20number%20of%20speakers&oldid=1153344805`, 2023. [Online; accessed 08-May-2023].

[25] Wikipedia. Mean absolute error — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Mean%20absolute%20error&oldid=1146984739, 2023. [Online; accessed 02-May-2023].

[26] Wikipedia. Mean absolute percentage error — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Mean%20absolute%20percentage%20error&oldid=1146010149, 2023. [Online; accessed 02-May-2023].

[27] Wikipedia. Regression analysis — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Regression%20analysis&oldid=1153128912, 2023. [Online; accessed 22-May-2023].

[28] Roger Levy Yevgeni Berzak, Boris Katz. Assessing language proficiency from eye movements in reading. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers):1986–1996, 2018.

[29] Suzanne Flynn Yevgeni Berzak, Chie Nakamura and Boris Katz. Predicting native language from gaze. *ACL*, page 541–551, 2017.