Classification, Regression or Ordinal Regression?
Assessing oral proficiency of non-native English speakers and interpreting results.

Pakhi Bamdev [1] Manraj Singh Grover [1] Yaman Kumar [1] Payman Vafayee [2,3] Mika Hama [2] Rajiv Ratn Shah [1]

[1]MIDAS Lab, Indraprastha Institute of Information Technology Delhi [2]Second Language Testing Inc. [3]Columbia University

Introduction

With the rise in demand for English proficiency assessments for both the academia and the industry, it has become increasingly necessary to have the human-level interpretation of the results to prevent bias and ensure meaningful feedback to the second language learners.

In this ***on-going work***, we:

- analyze and assess multiple classical models to choose the **best formulation of the spontaneous speech scoring task** among regression, classification and ordinal regression.
- identify the **feature groups that correlate strongly with the proficiency levels** via an ablation study.
- **explore model-agnostic interpretation methods** to gain insight on about the linguistic features that the model learns from the given set of linguistic features.

Dataset

- **Source:** We use the data collected by SLTI through Simulated Oral Proficiency Interview (SOPI) for L2 English speakers, primarily from the **Philippines**.
- **Test Format:** Each candidate is presented with a form containing six prompts of varying difficulty levels. These prompts demand opinions, reasoning, and narration in the form of spontaneous response.
- **Scoring Rubrics:** The prompts, as well as rubrics, are aligned with the guidelines of the Common European Framework of Reference (CEFR).
- **Scoring Responses:** The recorded monologue responses are scored independently by two human-expert raters. In case of any disagreement between two scores, a third expert is brought in to resolve the conflict.

Table 1:Statistics of the dataset. **P**: Prompt Number, **#R**: Number of responses, **D**: Difficulty level, **Sz**: Average Response Size (Duration in seconds, Length in number of word tokens), and **DS**: Distribution of Scores (prefixes 'L' means Low and 'H' means High).

P	#R	D	Sz		DS				
			Duration	Length	A2	LB1	HB1	LB2	HB2
1	7877	B1	57.67	100.69	275	1557	6045	-	-
2	7432	B1	58.72	110.03	465	2824	4143	-	-
3	8042	B2	81.43	148.96	117	664	3493	3666	102
4	8020	C1	104.15	180.73	121	720	3536	3534	109
5	7936	C1	105.95	196.55	110	551	3004	4120	151
6	8002	B1	55.87	109.38	119	1028	6855	-	-

Why study classification, regression and ordinal regression for the speech scoring task?

Speech Scoring as a Classification Task: The multi-class approach of classification does not consider the order of the classes during the training process. This becomes a disadvantage as the order of classes is an intrinsic property of speech scoring task. [1]
Speech Scoring as a Regression Task: The regression approach takes the order of a class into consideration but the distances between adjacent classes may not always be the same. This calls for a careful transformation of the numeric labels to the values for regression. [1]
Speech Scoring as an Ordinal Regression Task: Ordinal Regression is a regression analysis that preserves the order of the classes. This solves the shortcomings of treating speech scoring as a classification or a regression task. [2]

Experimentation and Results

- **Regression vs Classification vs Ordinal Regression*:** We trained Logistic Regression and Linear Regression for classification and regression analysis respectively along with Random Forest, Gradient Boosted Trees and XGBoost for both variants of speech scoring formulation. We found that **regression performed better for majority of prompts** (Table 2) for the number of models we trained and the **best performing model being XGBoost**. Figure 1 shares the scoring pipeline architecture.

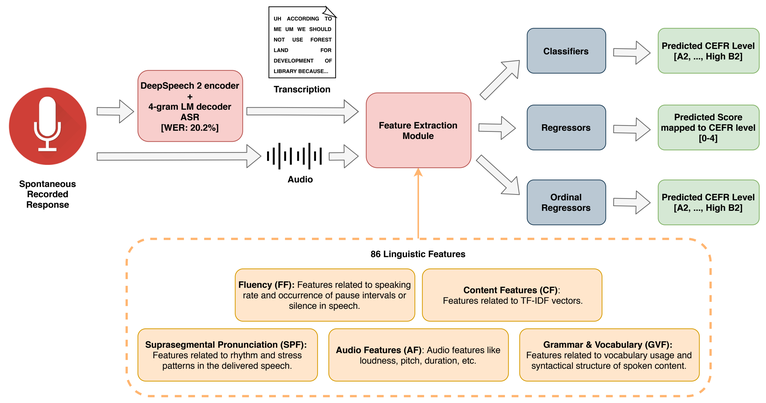

Figure 1:End-to-end Speech Scoring Pipeline.

- **Which Feature Category is more important?** After performing drop ablation study—retraining XGBoost regressor after removal of a feature category one by one from the dataset while keeping others intact—we found that the percentage drop in Quadratic Weighted Kappa (QWK) after the ablation study varies across all prompts (Figure 2). This reveals certain characteristics of the individual prompts as discussed below:
  - **Removal of Content Features (CF):** Prompts 1, 2 and 4 display drop of 6% in their QWK, while it was only approximately 4% drop for prompts 3 and 6. Prompt 5 seems to have been affected the least by removing the content features. We believe this is because the prompt is relatively more open-ended and highly dependent on the individual's opinion.
  - **Removal of Grammar and vocabulary features (GVF):** The drop percentage ranges between 7-13% for each prompt, thus confirming the importance of capturing grammatical construct and vocabulary usage.
  - **Removal of Fluency features (FF) and suprasegmental pronunciation features (SPF)** tend to show have almost similar drop percentage range. The impact of dropping suprasegmental pronunciation features shows a maximum drop in QWK for prompts 2 and 6.
  - **Removal of Audio features (AF)** have none to least impact on QWK for every prompt except for prompts 2 and 6. This infers that audio features and speech delivery metrics like rhythm and stress patterns are one of the key features to score prompts 2 and 6.

Table 2:Results for classification and regression analysis.** **QWK**: Quadratic Weighted Kappa, **MSE**: Mean Squared Error, **RF**: Random Forest, and **HH**: Human-Human.

Model	Prompt 1		Prompt 2		Prompt 3		Prompt 4		Prompt 5		Prompt 6	
	QWK	MSE	QWK	MSE	QWK	MSE	QWK	MSE	QWK	MSE	QWK	MSE
XGB Reg	**0.520**	**0.211**	**0.298**	**0.434**	0.498	0.430	**0.557**	**0.371**	**0.536**	**0.371**	**0.443**	**0.136**
XGB Cls	0.473	0.230	0.254	0.472	**0.509**	**0.432**	0.556	0.377	0.529	0.374	0.427	0.133
RF Reg	**0.517**	**0.210**	**0.289**	**0.441**	**0.479**	**0.447**	**0.550**	**0.370**	**0.530**	**0.374**	0.396	0.142
RF Cls	0.401	0.244	0.235	0.474	0.452	0.472	0.488	0.418	0.456	0.416	**0.456**	**0.416**
H-H	0.685	0.156	0.560	0.319	0.781	0.216	0.808	0.195	0.826	0.160	0.683	0.094

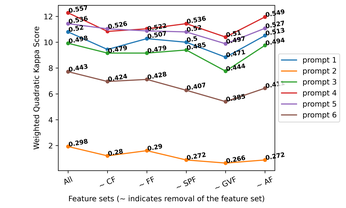

Figure 2:Plotting results for the ablation study. ~ denotes removal of a feature set.

Model Interpretation

We employ the use of model agnostic methodologies of interpretability, specifically Partial Dependence Plots** and SHAP values. In this poster, we explain XGBoost Regressor using SHAP summary plot only. It computes the contribution of each feature for making a certain prediction given an data point from the dataset and thus explains why a certain prediction was made. Additionally, it also gives an average impact of features on the models' prediction ability.

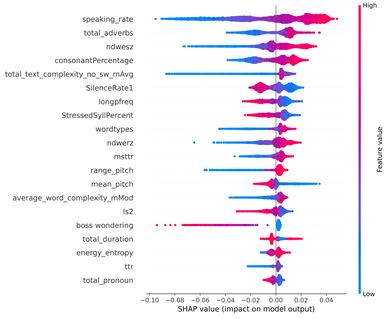

Figure 3:SHAP Summary Plot: Impact of each feature on the predictions generated by the model. (Prompt 4)

Future Work

- We are working towards the ordinal regression formulation of speech scoring task and also extending the work towards deep learning solutions.
- We plan to improve the existing feature set by adding more feature categories like segmental pronunciation.

References

[1] Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92--102, Florence, Italy, August 2019. Association for Computational Linguistics.

[2] Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 2071–2080, New York, NY, USA, 2017. Association for Computing Machinery.

∗ Work in progress. ∗∗ Unable to present detailed results due to lack of space.