Assignment 2: English Language Learning

Eirik M. Hauge, Submitted 18.11.2022

Disclaimer: The rapport has been writing during and after the projects end.

Contents

[The Problem 1](#_Toc119705971)

[Scope 1](#_Toc119705972)

[Business Objective 1](#_Toc119705973)

[Performance Metric 1](#_Toc119705974)

[Big picture/System 1](#_Toc119705975)

[Stakeholders 1](#_Toc119705976)

[Metrics 1](#_Toc119705977)

[Real World Metrics 1](#_Toc119705978)

[Feedback 1](#_Toc119705979)

[Data 1](#_Toc119705980)

[Modeling 1](#_Toc119705981)

[Deployment 1](#_Toc119705982)

[Proof of concept 1](#_Toc119705983)

[The Future 1](#_Toc119705984)

[References (or lack thereof) 1](#_Toc119705985)

# The Problem

When it comes to learning English, students that do not have English as their mother tongue need more and regular targeted practice. These students are referred to as ELL’s, English Language Learners.

Vanderbilt University is therefore hosting a Kaggle competition. Because in their words: “Existing tools are unable to provide feedback based on the language proficiency of the student, resulting in a final evaluation that may be skewed against the learner. Data science may be able to improve automated feedback tools to better support the unique needs of these learners”.

The feedback tools of today are lacklustre, the goal is to remedy this with the help of machine learning.

# Scope

## Business Objective

Vanderbilts describes their competition objective on Kaggle: “The resulting tools could serve teachers by alleviating the grading burden and support ELLs by ensuring their work is evaluated within the context of their current language level”. Automating the process of evaluating students could save human resources and encourage learning among ELLs.

## Performance Metric

The performance metric used in the competition is MCRSME (mean column wise root mean squared error). All solutions are to predict from 1 to 5 on six labels for each text. Namely: cohesion, syntax, vocabulary, phraseology, grammar, conventions. The scoring is the mean RSME of the columns.

## Big picture/System

If the model performs well enough, it could be integrated anywhere there are ELLs. It could become part of a pipeline for learning. Perhaps the predictions would influence another ML model that suggests learning material for students. One could use it on historical data to see how proficiency in writing affects proficiency in other subjects.

## Stakeholders

The university is this stakeholder of this project.

# Metrics

The performance metric used in the competition is RSME (root mean squared error). In the real world the cost benefit would perhaps be more fitting. There is no definite RSME which would fulfil the objective. It is merely a measure between humans and models.

## Real World Metrics

Looking at the objective, there are two metrics. To what degree does the model alleviate teachers? A model grading all texts a 3.0 would save teachers time, but ELLs would get no feedback. The metric is clear but may be difficult to measure.

## Feedback

The only way to test the metric of the model would be trying it. It would have to pass some threshold first, must work to a certain degree.

How can ELLs learn from their feedback? Does the feedback contribute to their learning? What do the teachers think of the solution? Long term it could be used in studies to measure difference in objective metrics like SATs or other nationwide tests.

# Data

The data is downloaded from the Kaggle competition. It consists of one feature and six labels. The feature ‘full\_text’ is an essay while the labels are floats between 0.0 and 5.0 by 0.5 increments.

It has been difficult to find ‘untrue’ labels. The labels usually deviate by less than 1.0. I have read some of the essay that have a wider spread but have not been able to see that is ‘bad’ or ‘corrupt’ data as determining writing quality of an essay is not straightforward. The same applies for ‘ground truth’. I would need someone involved in creating the data.

All student names or any other private data has been removed from the set, privacy is not an issue.

I wanted to have a go at this with the knowledge I had. I did not use a ML text analysis model. Rather an open-source language tool I found online was used. In short it uses a set of grammar and spelling rules to detect errors in text, see notebook for deeper explanation. The results have been used to calculate error frequencies for different categories/types of errors as new features. Features we’re created and scaled using a pipeline.

# Modeling

The task has been more demanding than first anticipated. I therefore decided to use pycaret for training, comparing, tuning and ensemble models. This was necessary for finishing the project on time. The models perform at an acceptable level, at least compared to the scores on Kaggle. Submissions there are scored at about 0.43, while my models average between 0.50 and 0.55.

Feature modeling has been the bulk of this task. Numerical data may be easier to differentiate and model after. Finding features here has been trial and error. Try new features, how do they matter to the model.

# Deployment

## Proof of concept

The models have been deployed using flask. I encountered a problem shortly before the deadline and had to switch solutions. The result is a virtually the same solution as the hospital deployment from the course. Had issues exporting/implementing the pipeline to flask, had to transfer all the code. Shows how it could work and the potential is there, all be it not looking pretty right now.

## The Future

Goals for the future would be to: store new texts and predictions, train the model if more data becomes available.

# References (or lack thereof)

Most sources I came across is based on entirely different models. Usually, it was about models ran by The Big Five on large datasets.

The personal goal for this assignment was to get a firm grip on the practical part of machine learning. The learning curve and new realisations has been of greater value than making the best model.