Udacity Capstone Project Proposal

Background:

This project is based on Kaggle competition as suggested in the proposal page of Udacity.

Writing is a foundational skill. Sadly, it's one few students are able to hone, often because writing tasks are infrequently assigned in school. A rapidly growing student population, students learning English as a second language, known as English Language Learners (ELLs), are especially affected by the lack of practice. While automated feedback tools make it easier for teachers to assign more writing tasks, they are not designed with ELLs in mind.

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.

Competition host Vanderbilt University is a private research university in Nashville, Tennessee. It offers 70 undergraduate majors and a full range of graduate and professional degrees across 10 schools and colleges, all on a beautiful campus—an accredited arboretum—complete with athletic facilities and state-of-the-art laboratories. Vanderbilt is optimized to inspire and nurture cross-disciplinary research that fosters discoveries that have global impact. Vanderbilt and co-host, The Learning Agency Lab, an independent nonprofit based in Arizona, are focused on developing science of learning-based tools and programs for social good.

Vanderbilt and The Learning Agency Lab have partnered together to offer data scientists the opportunity to support ELLs using data science skills in machine learning, natural language processing, and educational data analytics. You can improve automated feedback tools for ELLs by sensitizing them to language proficiency. 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.

Problem Statement:

The goal of this project is to assess the language proficiency of 8th-12th grade English Language Learners (ELLs). Utilizing a dataset of essays written by ELLs will help to develop proficiency models that better supports all students. This work will help ELLs receive more accurate feedback on their language development and expedite the grading cycle for teachers. These outcomes could enable ELLs to receive more appropriate learning tasks that will help them improve their English language proficiency.

Datasets and Inputs:

The dataset presented here (the ELLIPSE corpus) comprises argumentative essays written by 8th-12th grade English Language Learners (ELLs). The essays have been scored according to six analytic measures: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Each measure represents a component of proficiency in essay writing, with greater scores corresponding to greater proficiency in that measure. The scores range from 1.0 to 5.0 in increments of 0.5. The goal is to predict the score of each of the six measures for the essays given in the test set.

File and Field Information:

train.csv - The training set, comprising the full text of each essay, identified by a unique text_id. The essays are also given a score for each of the seven analytic measures above: cohesion, etc. These analytic measures comprise the target for the competition.

test.csv - For the test data we give only the full text of an essay together with its text_id.

Solution Statement

Convert the essay to tf-idf form by tokenization, lemmatization and perform regression. Test different regression methods and build a model with the best method. Also, test different normalization methods.

Benchmark model

The current leaderboard has the score of 0.3

Evaluation metrics

$$\mathrm{MCRMSE} = \frac{1}{N_t}\sum_{j=1}^{N_t}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}$$

Mean Column wise Root Mean Squared Error:

where N_t is the number of scored ground truth target columns

y and ŷ are the actual and predicted values, respectively.

Project Outline

Data Exploration - Tokenization – Lemmatization – tf-idf transformation – regression methods – evaluate the models – deploy the best model

Data exploration – Explore the predictor and target variables. Here the predictor variable is the essay and there are 6 target variables whose scores range from 1 through 5. Determine the number of words in each essay to further establish the pre processing steps

Perform tokenization, Lemmatization (Explore stemming but not likely implementable because the problem is to feedback English proficiency, so lemmatization is probably better) and then perform tf-idf transformation (or simple numeric transformation) using ‘nltk’ libraries in Python

Modeling – Looking at the data in first glance, it looks classification algorithms best fit since the scores are from 1 to 5. But inspecting the data, there are real numbers such as 3.5, 4.5 etc. So, regression might be the best fit to model. However, experiments might be done with the data such as convert the real numbers to integers to test classification algorithms best model the data. Supervised classification algorithms such as knn, decision tree and other ensemble methods will be tested. Regression models such as multi linear regression, lasso regression and other ensemble algorithms will be tested. Based on the evaluation, best model will be selected. ‘sklearn’ and relevant other python libraries will be used for coding.