Books Recommendation Based on Readability Score using Machine Learning

Business Case Report

Prepared by

Michael Pham
23/09/2021

Executive Summary

A business challenge was proposed for the automation of selecting appropriate books for the education levels from years 9 to 13. With an increasing focus on the use of Natural Language Processing technologies, a study was done to create a machine learning model as the solution to this challenge.

This report proposes a supervised neural network regressor that was trained in the Kaggle CommonLit Readability Challenge and was then later used to give recommendations between 100 books to the years group from the GutenBerg website. The model was able to make predictions of the readability score for each book. The books were then listed from the most complex to the simplest ranged from lowest readability score to highest readability score. Therefore, 5 books were then recommended to each year group from 9 to 13.

Introduction

This business case is based on an international company that specializes in developing English language curriculums for educational institutions. The company is looking to use NLP to optimize development for its curriculum. To keep their curriculums updated every year, the company must select appropriate levels of complexity for reading materials for the years from 9 to 13. Hence, the solution is to develop a robust machine learning model to score the material in terms of readability and make accurate recommendations.

The model discussed in this report was created to enter the CommonLit Readability Kaggle challenge and was ranked among others. Using a training dataset of 2834 unique text excerpts and a test dataset of 7 unique values, the models are ranked on the root mean squared error of the predicted readability scores on the test dataset. Although it wasn't ranked very high in the competition, the recommendations created from the model's predictions are relatively accurate.

This report will discuss how the methodology of the project as well as the model. It will also go over the learning gained in the process and how the results can be used to gain insights into the proposed business case.

Analysis

2.1. Methodology

The model used for this study from scikit-learn package, an open-source python package built to solve problems using machine learning models. The library provides an inbuilt supervised neural network model called MLPRegressor model. The model is a Multi-Layer Perceptron regressor that optimizes the squared loss using stochastic gradient descent. A neural network is a network or circuit of neurons composed of artificial neurons that essentially detects features and patterns in a problem.

A book is made up of paragraphs, a paragraph is made up of sentences, and sentences are made up of words. Therefore, at the simplest level, a pattern of complexity in an entire book can be recognized by assessing the difficulty of every single word within the book. In this case, the model needs to be able to analyses these vocabularies are broken down from paragraphs to predict the overall readability of a book.

Therefore, the model will first need to be trained using a set of paragraphs with their corresponding pre-predefined readability scores. For training, the training dataset provided from the Kaggle competition was used. This dataset serves as a sufficient dataset as it includes a good range of complexities in paragraphs. This will allow the model to pick up these patterns more easily. Another crucial step in this process is that the data needs to first be processed and cleaned so that the accuracy of the model can be improved.

After the model has been trained, it is then used to make predictions on a new dataset that is better simulates what paragraphs in actual books look like. For this, the test data was scraped from the website GutenBerg, which is a library of over 60,000 downloadable free eBooks. The test dataset consists of 100 paragraphs from 100 books. The longest paragraph was taken from each book under the assumption that it'll represent the highest complexity of the book.

2.2. Results

After using the trained model to make predictions for the test dataset, the result can be shown in appendices 1 and 2. The two graphs show the distribution of the readability score across the 100 books scraped from the GutenBerg website. Note that the lower the readability score represents an increase in complexity.

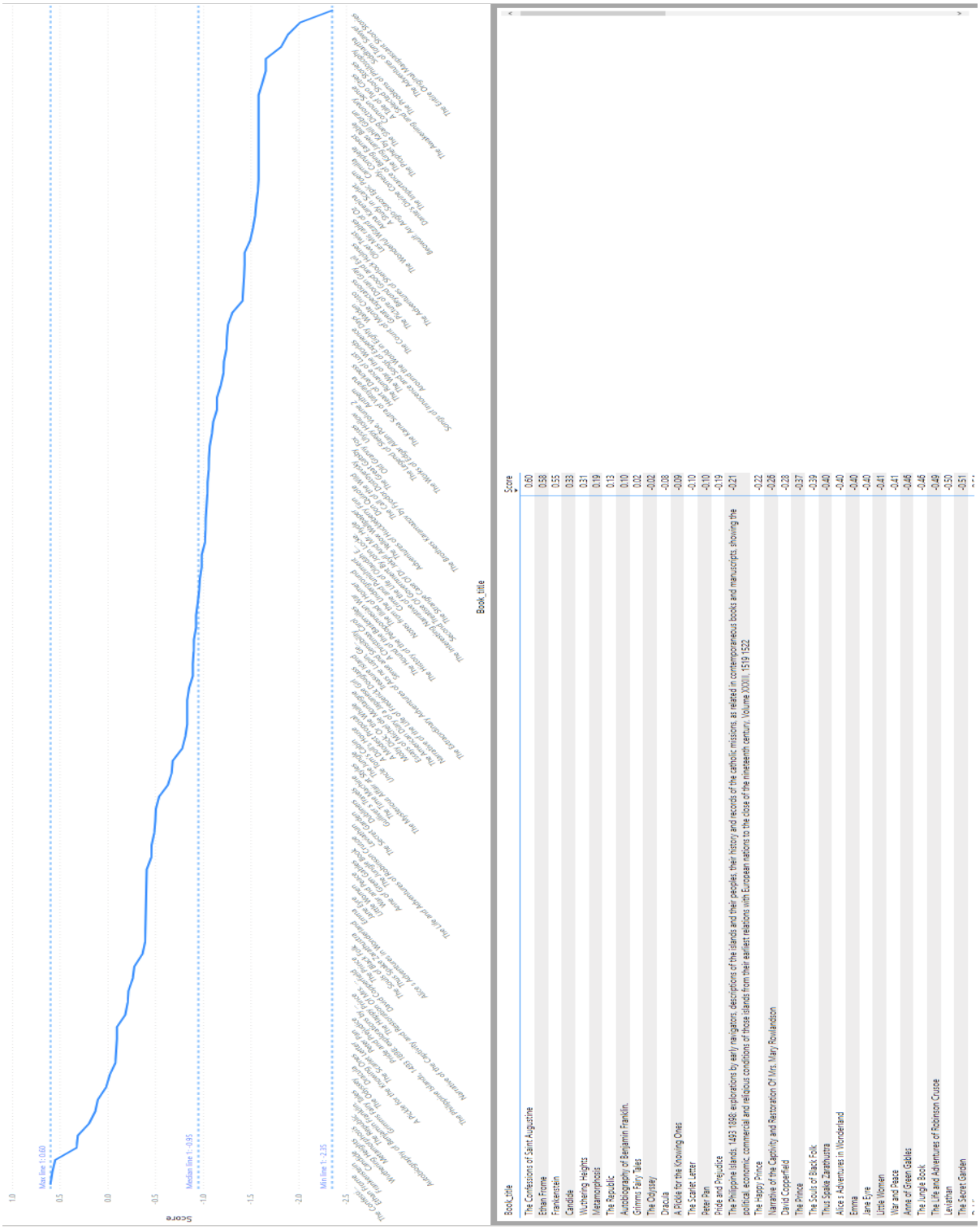
Conclusion

From the two graphs, we can see that the 100 books can be ranked from the lowest readability score to the highest, from the most complex to the simplest. Thus, we can make recommendations for the reading level from levels 9 to 13 according to this ranking, refer to Table 1.

	Year 9	Year 10	Year 11	Year 12	Year 13
Book 1	Narrative of the Captivity and Restoration Of Mrs. Mary Rowlandson	Uncle Tom's Cabin	The Strange Case Of Dr. Jekyll And Mr. Hyde	Great Expectations	The Entire Original Maupassant Short Stories
Book 2	The Happy Prince	The Jungle	Second Treatise Of Government By John Locke.	The Count of Monte Cristo	Siddhartha
Book 3	Pride and Prejudice	The Mysterious Affair at Styles	The Interesting Narrative of the Life of Olaudah Equiano, or Gustavus Vassa, The African	Walden	The Problems of Philosophy
Book 4	Peter Pan	The Time Machine	The Adventures of Tom Sawyer	Around the World in Eighty Days	The Awakening and Selected Short Stories
Book 5	The Scarlet Letter	Gulliver s Travels	Crime and Punishment	Songs of Innocence and Songs of Experience	A Tale of Two Cities

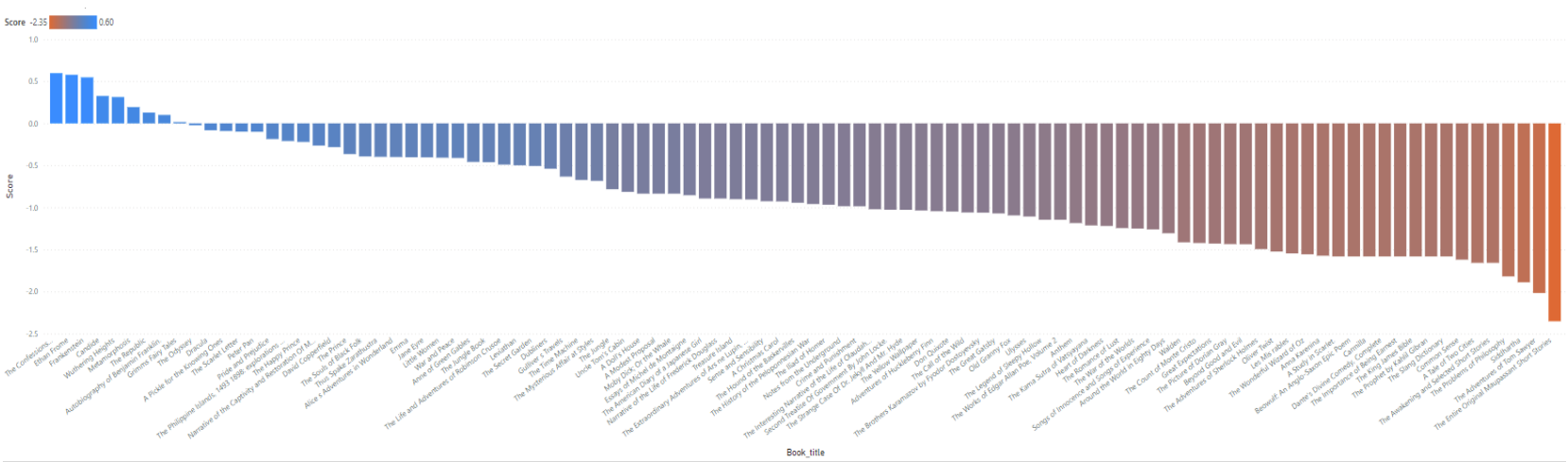
Table 1: Book recommendations for year levels 9 to 13

Appendices



Appendix 1: Line graph of readability score distribution of 100 books

Appendix 2: Column graph of readability score distribution of 100 books



Book_title	Score
The Confessions of Saint Augustine	0.60
Ethan Frome	0.58
Frankenstein	0.55
Candide	0.33
Wuthering Heights	0.31
Metamorphosis	0.19
The Republic	0.13
Autobiography of Benjamin Franklin.	0.10
Grimms Fairy Tales	0.02
The Odyssey	-0.02
Dracula	-0.08
A Pickle for the Knowing Ones	-0.09
The Scarlet Letter	-0.10
Peter Pan	-0.10
Pride and Prejudice	-0.19
The Philippine Islands, 1493 1898: explorations by early navigators, descriptions of the islands and their peoples, their history and records of the catholic missions, as related in contemporaneous books and manuscripts, showing the political, economic, commercial and religious conditions of those islands from their earliest relations with European nations to the close of the nineteenth century, Volume XXXIII, 1519 1522	-0.21
The Happy Prince	-0.22
Narrative of the Captivity and Restoration Of Mrs. Mary Rowlandson	-0.26
David Copperfield	-0.28
The Prince	-0.37
The Souls of Black Folk	-0.39
Thus Spake Zarathustra	-0.40
Alice s Adventures in Wonderland	-0.40
Emma	-0.40
Jane Eyre	-0.40
Little Women	-0.41
War and Peace	-0.41
Anne of Green Gables	-0.46
The Jungle Book	-0.46
The Life and Adventures of Robinson Crusoe	-0.49