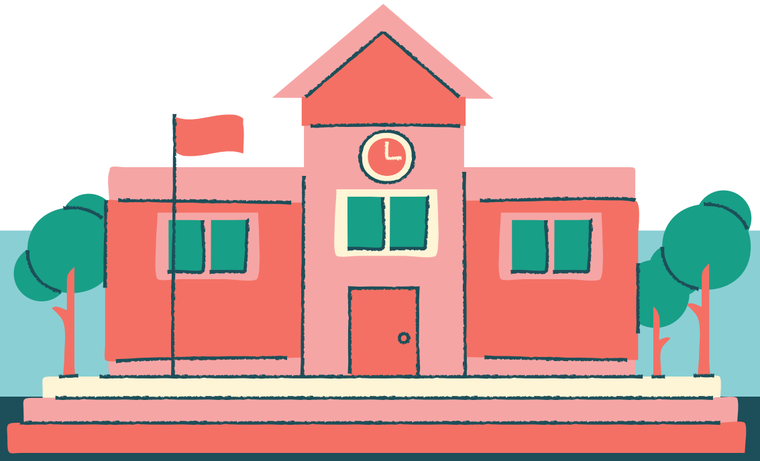
Predicting Readability in Student Literature

Christian Moya

Data Scientist



Welcome to class!

Today's Agenda

- About Us
- Problem Statement
- Methodology
- Data
- Results
- Recommendations
- Future WOrk

About Us

We are a group of data scientist, hoping to use machine learning to close the literacy gap for students across America.

Problem Statement

Can we predict the readability of student literature?

Null hypothesis: words in a text **do not** have an effect on its readability.

Alternate hypothesis: words in a text **do** have an effect on its readabiltity.

Methodology

Business Understanding

How can we ensure we are giving students texts within their zone of proximal development?

Data Understanding

2,800+ texts for students grades 3-12 ranging from -3.67 - 1.71 with an average of -0.9.

Data Preparation

Clean and tokenize the text data to prepare for modeling

Modeling

Linear Regression Model
Ridge Regression Model
Random Forest Classifier
GridSearch

Evaluation

How can we improve our RMSE score?

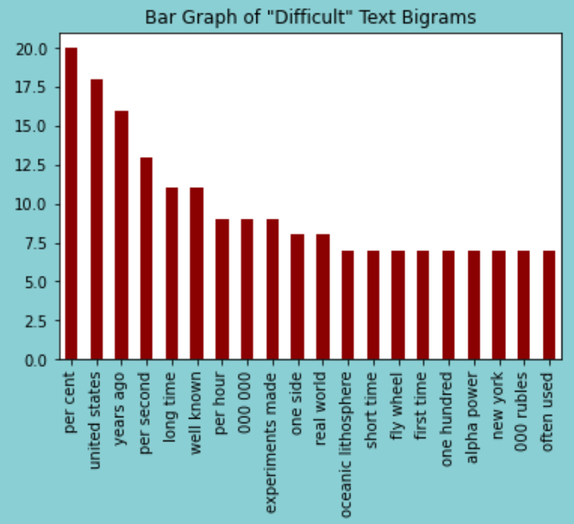
Deployment

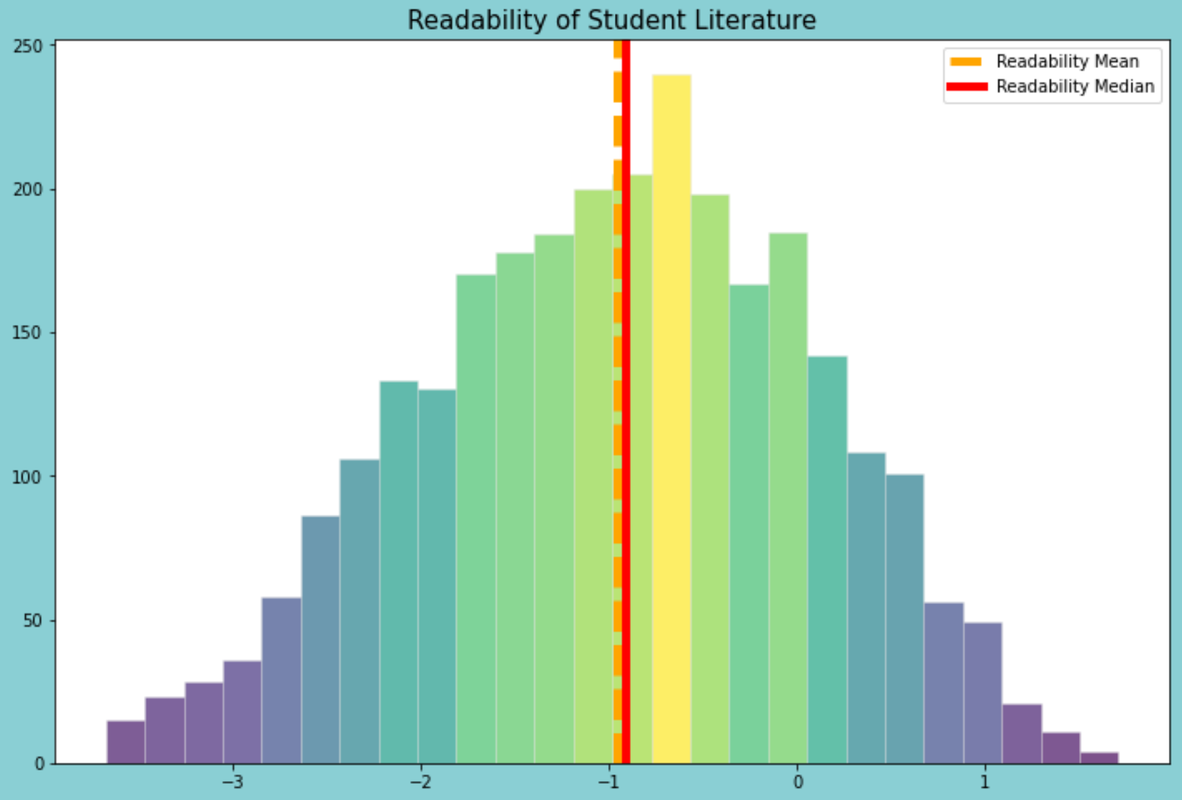
Recommendation to Follow

Pre-Model Exploratory Data Analysis

Readability: -3.68

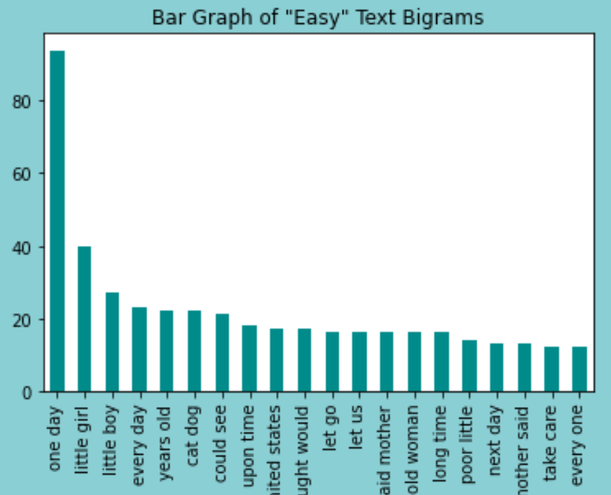
The commutator is peculiar, consisting of only three segments of a copper ring, while in the simplest of other continuous current generators several times that number exist, and frequently 120! segments are to be found. These three segments are made so as to be removable in a moment for cleaning or replacement.



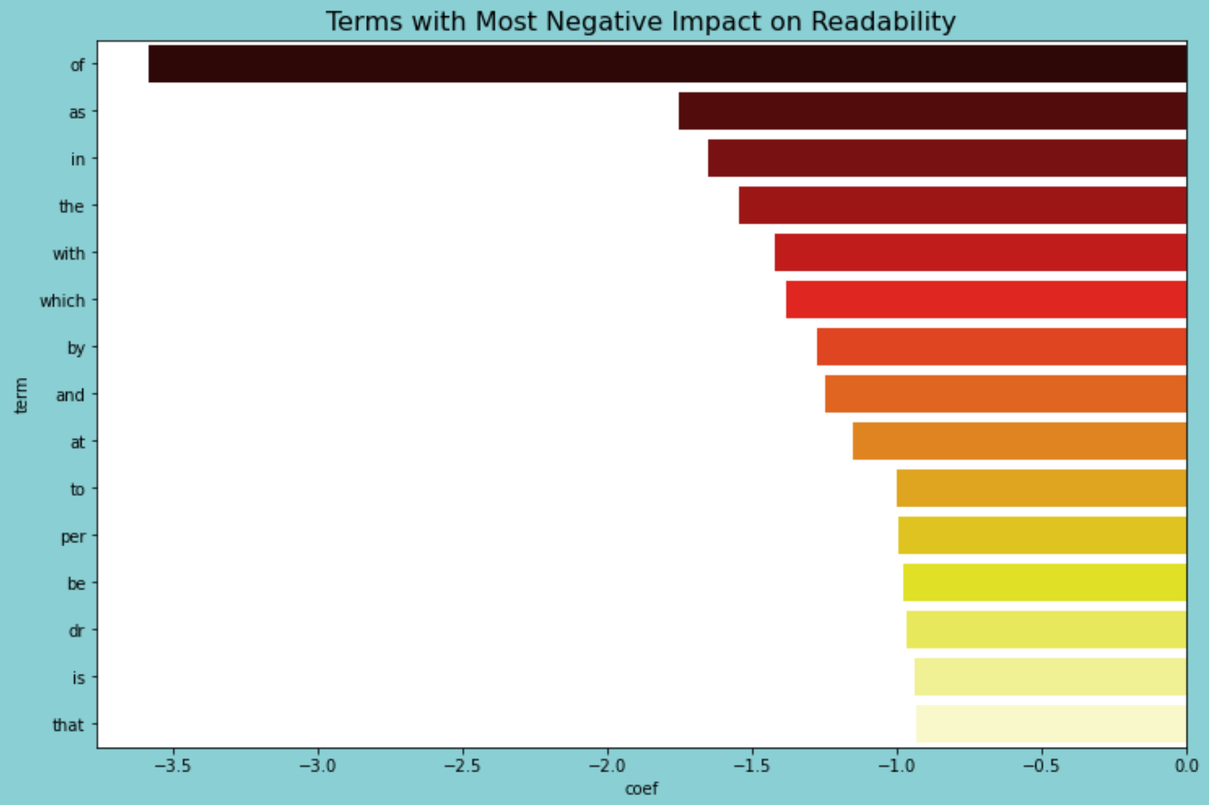


Readability: 1.71

When you think of dinosaurs and where they lived, what do you picture? Do you see hot, steamy swamps, thick jungles, or sunny plains? Dinosaurs lived in those places, yes. But did you know that some dinosaurs lived in the cold and the darkness near the North and South Poles?

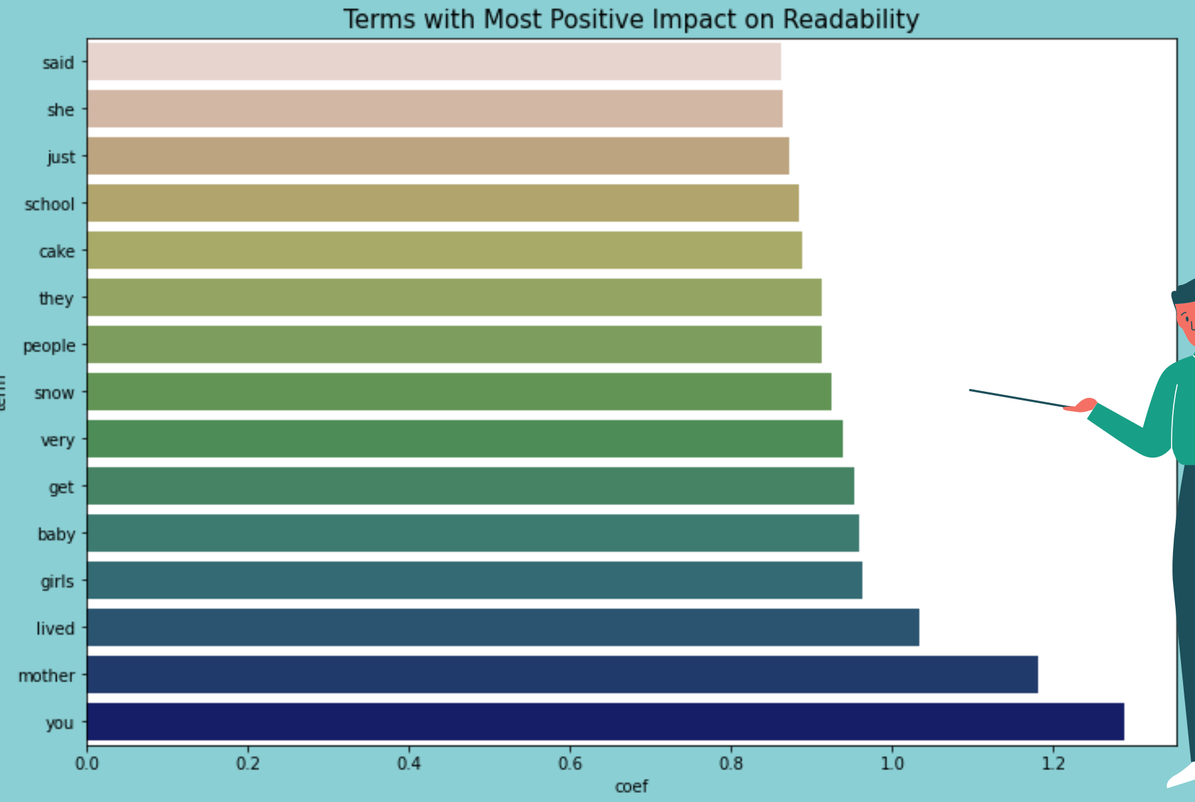


Post-Model Results : RMSE .698, 5% decrease

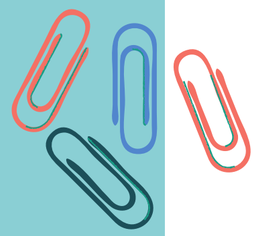


Terms with *negative* coefficients are more prepositional words.

Terms with *positive* coefficient are more story-based words.



Recommendations



For Struggling Readers:

- Introduce tougher subjects like science and math through storytelling to provide some sort of access point.
- If a text is too challenging, remove unnecessary prepositional phrases to make the text more direct and easy to comprehend.

For Excelling Readers:

- Provide texts a nudge more difficult than they're comfortable with to continue developing their literacy skills.
- Challenge more talented readers to write out summaries of the texts to provide access points to struggling readers as well.

Next Steps

Front-End

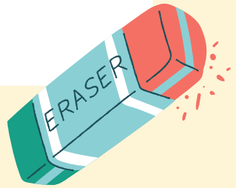
Build out a front-end so teachers can determine the readability of assigned texts and make adjustments accordingly.

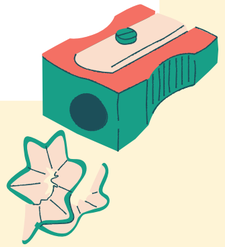
PCA

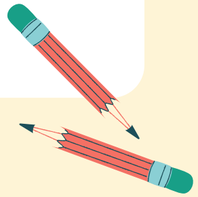
Use PCA to create a more accurate model, given that text data contains multiple zeros.

K-2

Include texts for kindergarteners to 2nd grade and explore readability for phonics and decoding.

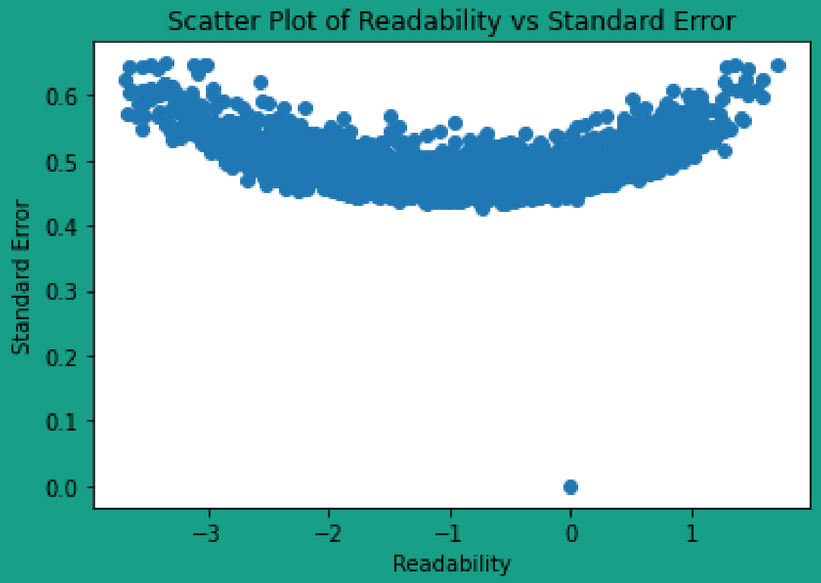


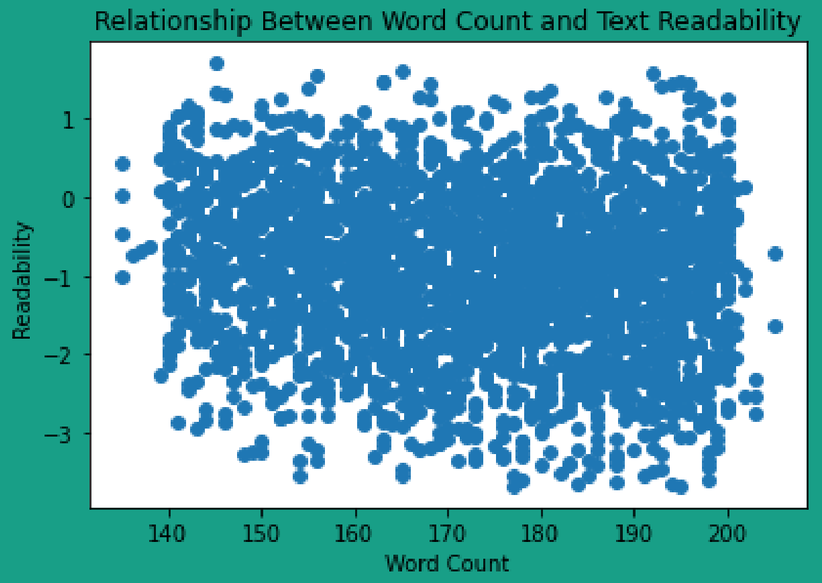


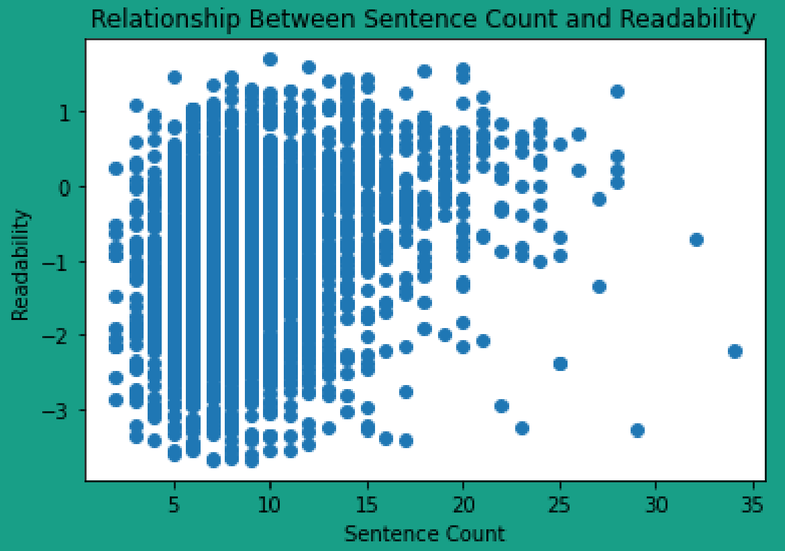


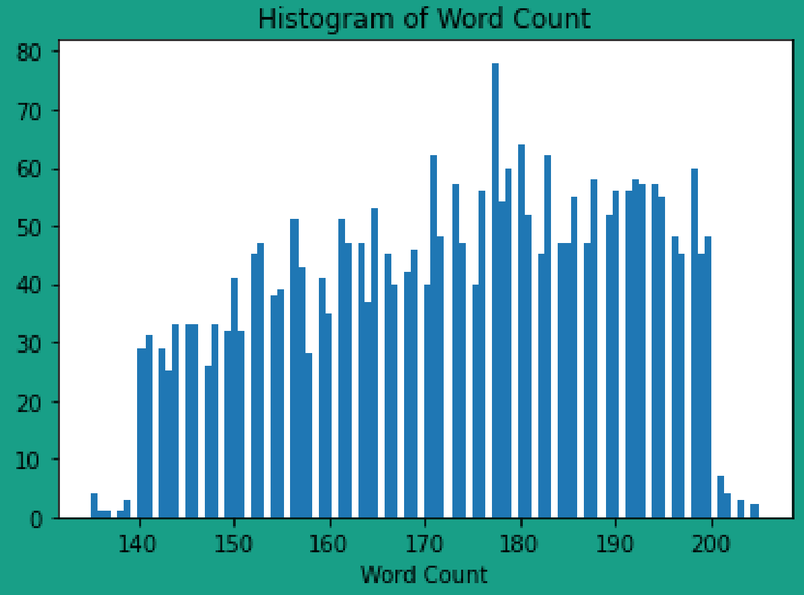
Thank you!

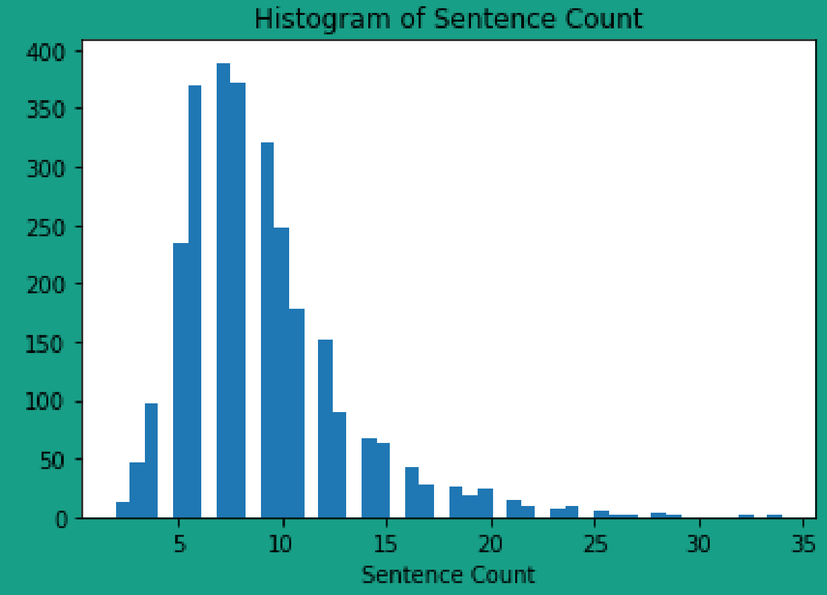
Use this space for announcements, homework,
or ways students can approach you if ever they have questions.


Scatter Plot of Readability vs Standard Error
Standard Error
Readability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1


Relationship Between Word Count and Text Readability
Readability
Word Count
1
0
−1
−2
−3
140
150
160
170
180
190
200


Relationship Between Sentence Count and Readability
Readability
Sentence Count
1
0
−1
−2
−3
5
10
15
20
25
30
35


Histogram of Word Count
Word Count
0
10
20
30
40
50
60
70
80
140
150
160
170
180
190
200


Histogram of Sentence Count
Sentence Count
0
50
100
150
200
250
300
350
400
5
10
15
20
25
30
35