Finding the Perfect Match: Identifying Text Maturity

Evan Simpson
DSR Batch 21: Demo Day
p.evansimpson@gmail.com

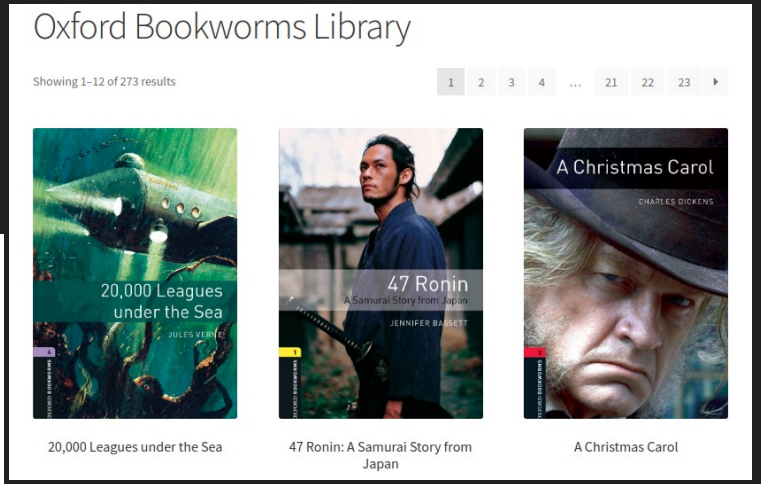
Outline

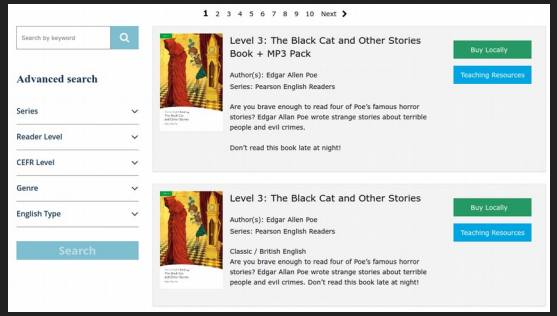
- Introduction
- Data Collection
- Models
  - Tabular - XGBoost
  - Sequences - LSTM
  - Computer Vision - 2D Convolutional NN
- Future Directions
- Questions

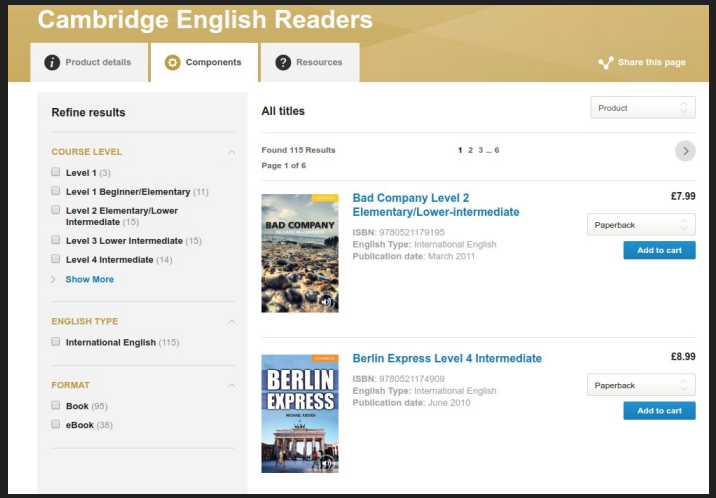
A long time ago, in a country far, far away….

I worked as a language teacher in South Korea.

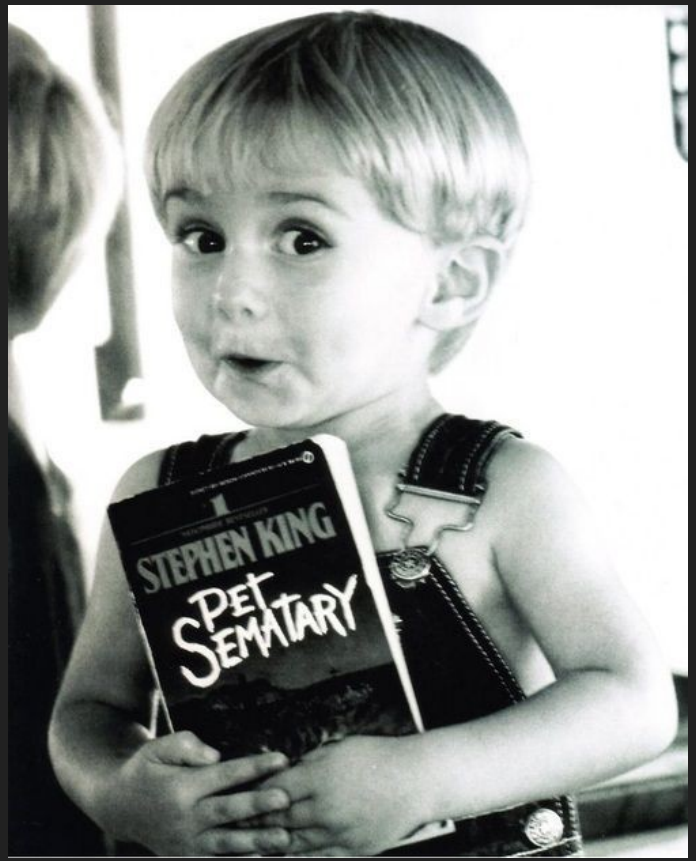
I realized there were several methods for identifying the complexity of the language of a book

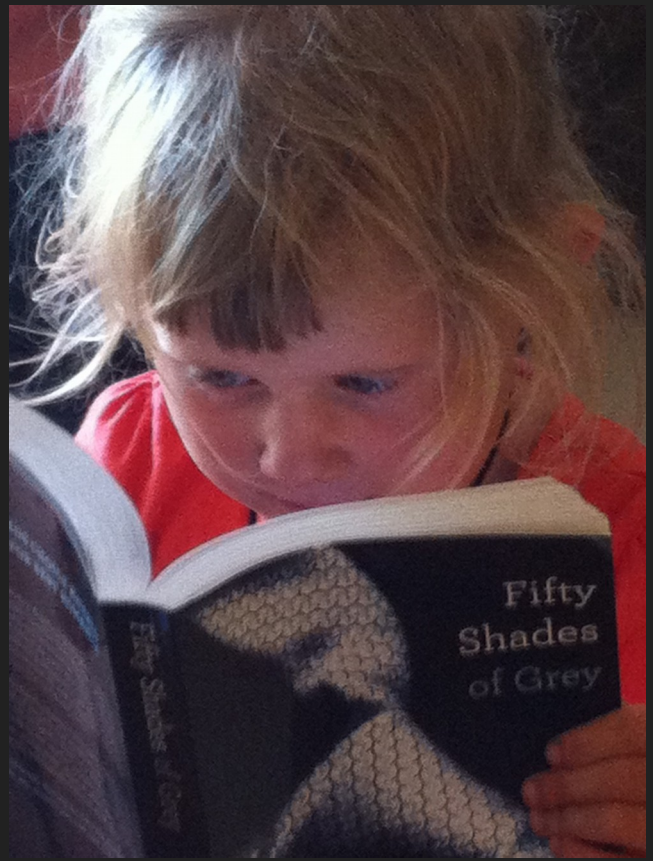






But what about maturity?





Challenge:

Create a program to identify the ***minimum*** maturity level required to read a text.

Where to get the data? Common Sense Media

common sense media

Find movies, books, and more ...

Sign in | Become a member

Movies & TV | Books | Apps & Games | Parents Need to Know | Latino | Research | About Us | Coronavirus Support

Narrow results

5,857 results

AGES

2 3 4 5 6 7
8 9 10 11 12 13
14 15 16 17 18

ENTERTAINMENT TYPE

Movies	10,094
Games	4,772
Apps	4,612
Websites	1,307
TV shows	6,776
Books	
Music	2,206

Book Reviews

Sort By Most Recent

Books

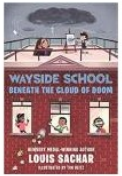
Deathless Divide: Dread Nation, Book 2

age 13+ ★★★☆☆

Engaging, occasionally uneven finale to a great story.

By Justina Ireland (2020)

Continue reading



Wayside School Beneath the Cloud of Doom: Wayside School, Book 4

age 8+ ★★★★★

Sweet absurdity abounds with normal kids in a wacky school.

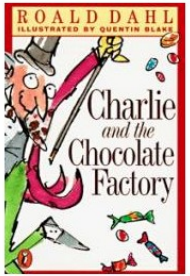
By Louis Sachar (2020)

Continue reading

Look at the data to validate the ratings

Charlie and the Chocolate Factory

Book review by Stephany Aulenback, Common Sense Media



Common Sense says

age 6+ ★★★★★

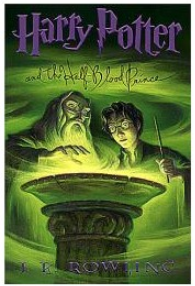
Classic morality tale is wildly entertaining.

Roald Dahl | Fantasy | 1964

Save | Rate book

Harry Potter and the Half-Blood Prince

Book review by Matt Berman, Common Sense Media



Common Sense says

age 11+ ★★★★★

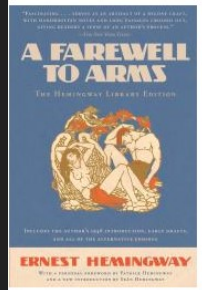
Powerful penultimate book mines Voldemort's past.

J. K. Rowling | Fantasy | 2005

Save | Rate book

A Farewell to Arms

Book review by Barbara Schultz, Common Sense Media



Common Sense says

age 16+ ★★★★★

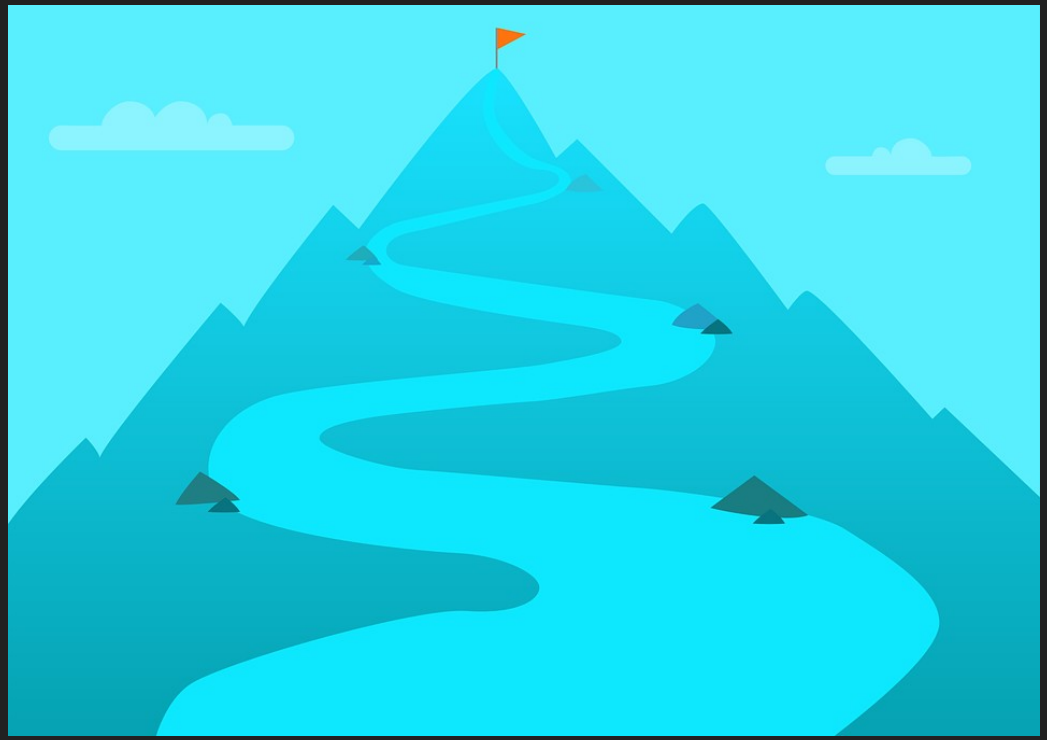
Classic war novel still speaks to today's readers.

Ernest Hemingway | History | 1929

Save | Rate book

Project Outline

- Collect the data
- Clean the data
- Create models
- Tests models
- Present the Results



Down the (coding) Rabbit Hole!

Get the Data: Scraping with Beautiful Soup

```python
titles = []
links = []
urls = []
file_names = []
```

```python
def get_titles(soup):
    for s in soup.findAll(class_="views-field views-field-field-reference-review-ent-prod result-title"):
        titles.append(s.get_text().strip())
    return titles
```

```python
def get_href(soup):
    for s in soup.findAll(class_="views-field views-field-field-reference-review-ent-prod result-title"):
        links.append(s.a)
    return links
```

```python
def make_urls(links):
    for l in links:
        l = str(l)
        urls.append(base_url + l.split('"',2)[1])
    return urls
```

```python
def make_f_names(links):
    for l in links:
        l = str(l)
        l = l.split('/')[2]
        file_names.append(l.split('"')[0])
    return file_names
```

```python
for file in files:
    with open(file, 'r') as f:
        soup = BeautifulSoup(f.read(), 'html.parser')
        get_titles(soup)
        get_href(soup)

urls = make_urls(links)
f_names = make_f_names(links)
```

```python
df = pd.DataFrame()
df['title'] = titles
df['f_name'] = f_names
df['url'] = urls
df.head()
```

Inspect the Data: 5,816 Complete Observations

Title	Text
Description	Text
Plot Summary	Text
Review	Text
What Parents Need to Know	Text
Genre	Category
Type	Category
Publication Date	Date
CSM Rating	Category

Get the Data: Scraping with Beautiful Soup

Get the Data: Scraping with Beautiful Soup

WHAT'S THE STORY?

What strange and marvelous creatures will Alice find down the rabbit hole, and what amazing thing will happen next? The inventive language and charming fantasy make this a classic that both adults and kids love. Older ones will appreciate the satire, but some younger children will be confused or bored. Updated illustrations are appealing to children.

WHAT PARENTS NEED TO KNOW

Parents need to know that constantly changing predicaments, strange creatures, and the watercolors are very child-friendly. But difficult language, Carroll's nonsense poems, and adult humor will leave some children bored or confused. Still, it's a classic well worth the trouble and particularly fun as a read-aloud.

BOOK DETAILS

Author: Lewis Carroll
Illustrator: Helen Oxenbury
Genre: Literary Fiction
Topics: Magic and Fantasy, Adventures, Misfits and Underdogs
Book type: Fiction
Publisher: Candlewick Press
Publication date: November 26, 1865
Publisher's recommended age(s): 9 - 12
Number of pages: 207
Last updated: November 15, 2019

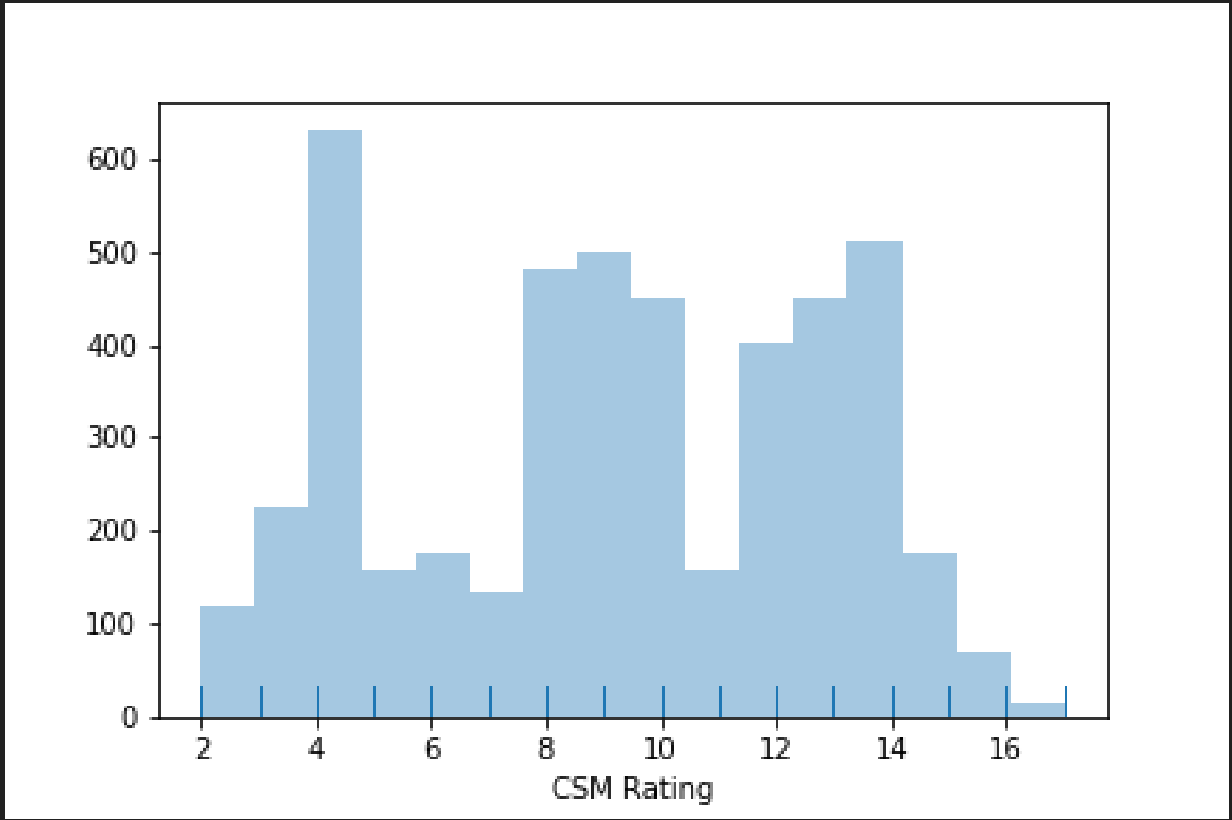
IS IT ANY GOOD?

Though there are many video versions and a lot of simplified retellings of this story, all kids deserve to know this wonderful adventure as Lewis Carroll wrote it. But it takes a particular kind of child to enjoy this: Complex language, nonsense, and the lack of a sensible plot are not to every child's taste. The book needs to be thoughtfully read aloud by an adult; few children will read this on their own. But, read aloud, the rhythmic poems can delight kids for their sounds and silly images.

The book works on two levels: as a delightful children's fantasy and as an impish poke in the eye to adults. Alice's strange new world remains just enough like the polite society of Victorian England that we can recognize it -- but it isn't terribly polite, allowing adults to understand much of the book as satire. Of course, kids usually don't see the satire; they simply enjoy the nonsense. If you've forgotten how to do that, Alice can help you remember.

What’s the target?

Distribution of the Target



Establish the Baseline

Mean = 9.00

Mean Absolute Error (MAE) = 3.27

Model 1: XGBoost - “King of Kaggle”

Bag of Words:

- Title
- Description
- Plot Summary
- Review
- What Parents Need to Know

One Hot Encode:

- Type
- Genre

Min Max Scale:

- Publication Date

Hyperparameters tuning:

- RandomizedSearchCV()

Model 1: XGBoost - Result

train = 1.50

test = 1.62

Model 2: Long Short Term Memory (LSTM)

Title
Description
Plot Summary
Review
What Parents Need to Know
Genre
Type
Publication Date
CSM Rating

Model 2: Long Short Term Memory (LSTM)

Title + Description + Plot Summary + Review + What Parents Need to Know + Genre + Type + Date

Example:

'Agent of Chaos: The X-Files Origins, Book 1 Set in 1979, AGENT OF CHAOS follows a 19. The plot revolves around a villain who kidnaps and murders young children, but the level of actual violence is low. A supporting character dies by having his neck snapped. […] Sexual content is minimal, limited to a few passionate kisses and a night spent in bed with clothes on. […] Science Fiction Fiction 2017'

Model 2A: LSTM with Custom Embeddings

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 979, 300)          13570500
_________________________________________________________________
lstm (LSTM)                  (None, 979, 128)          219648
_________________________________________________________________
global_max_pooling1d (Global (None, 128)               0
_________________________________________________________________
dense (Dense)                (None, 64)                8256
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 65
=================================================================
Total params: 13,798,469
Trainable params: 13,798,469
Non-trainable params: 0
_________________________________________________________________
None
```

ModeL 2B: LSTM with Pre-trained Embeddings (Glove 6B 300)

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 979, 300)          13570500
_________________________________________________________________
lstm_1 (LSTM)                (None, 979, 128)          219648
_________________________________________________________________
global_max_pooling1d_1 (Glob (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 64)                8256
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 65
=================================================================
Total params: 13,798,469
Trainable params: 227,969
Non-trainable params: 13,570,500
_________________________________________________________________
None
```

LSTM Results:

Custom Embeddings Train = 0.39

Custom Embeddings Test = 1.38

Pre-trained Embeddings Train = 0.66

Pre-trained Embeddings Test = 1.04

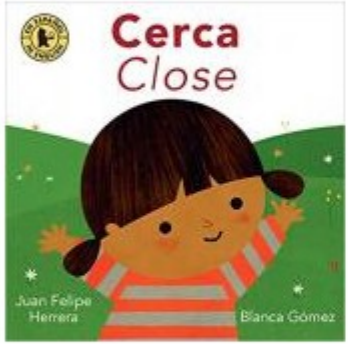
Model 3: Computer Vision

Can you judge a book by its cover?

Model 3: Computer Vision

Cerca / Close

Book review by Monica Encarnacion, Common Sense Media



Common Sense says

age 2+ ★★★★★

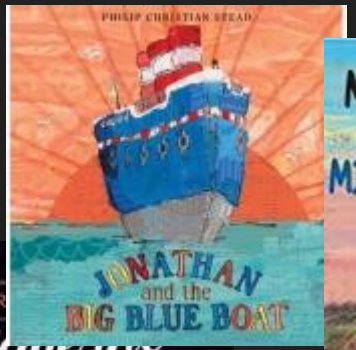
Engaging, simple bilingual book shows what "close" means.

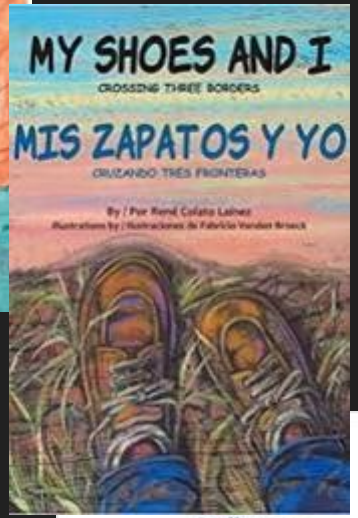
Juan Felipe Herrera | Board | 2019

Save | Rate book

One Issue:
While a book can have multiple covers, I only had one version for each book.

Issue 2: Book covers come in different sizes. I had to standardize them so they would fit into the model.





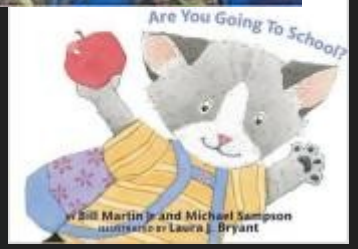
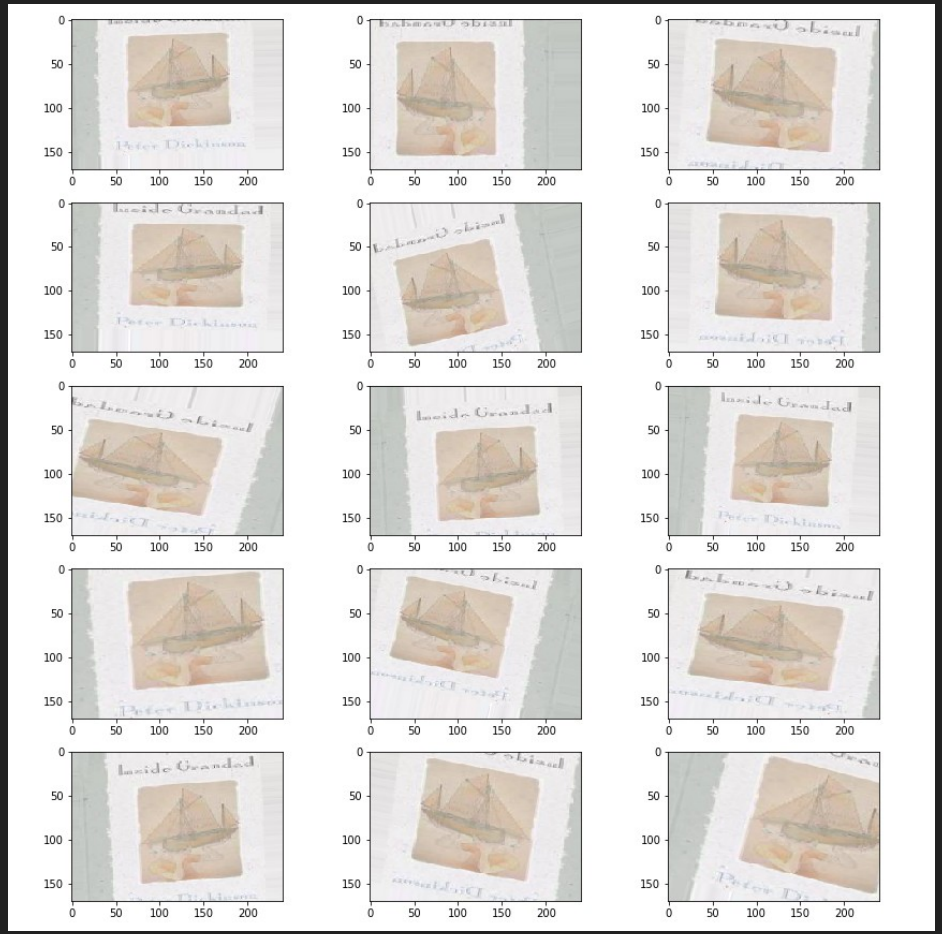


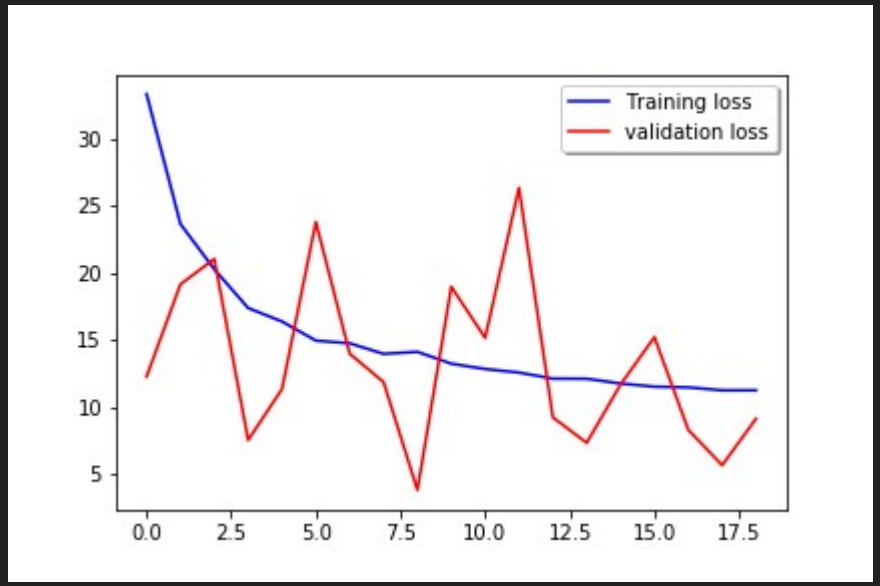
Image Data Augmentation

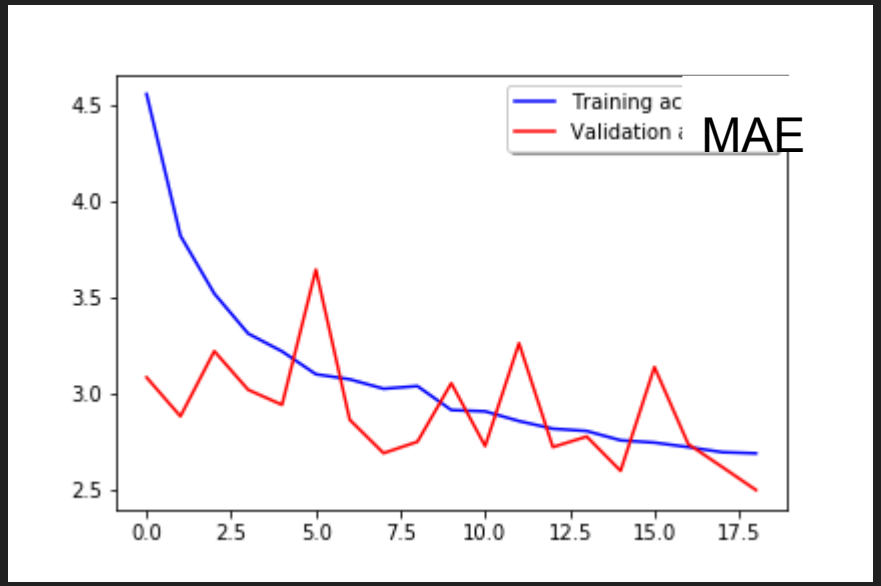
- Rotating
- Shearing
- Flipping
- Skewing
- Zooming



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 168, 238, 32)      896
_________________________________________________________________
batch_normalization_1 (Batch (None, 168, 238, 32)      128
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 84, 119, 32)       0
_________________________________________________________________
dropout_1 (Dropout)          (None, 84, 119, 32)       0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 82, 117, 64)       18496
_________________________________________________________________
batch_normalization_2 (Batch (None, 82, 117, 64)       256
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 41, 58, 64)        0
_________________________________________________________________
dropout_2 (Dropout)          (None, 41, 58, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 39, 56, 128)       73856
_________________________________________________________________
batch_normalization_3 (Batch (None, 39, 56, 128)       512
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 19, 28, 128)       0
_________________________________________________________________
dropout_3 (Dropout)          (None, 19, 28, 128)       0
_________________________________________________________________
flatten_1 (Flatten)          (None, 68096)             0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               34865664
_________________________________________________________________
batch_normalization_4 (Batch (None, 512)               2048
_________________________________________________________________
dropout_4 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 513
=================================================================
Total params: 34,962,369
Trainable params: 34,960,897
Non-trainable params: 1,472
```

Did it train?





Q: Why the bounce?

A: Image Data Augmentation

Model 3: Computer Vision Result

train = 2.69

test = 2.50

Results

Model	*MAE*
Naive Baseline (Mean)	3.27
Model 3: Computer Vision	2.50
Model 1: XGBoost	1.62
Model 2A: Concat LSTM	1.38
Model 2B: Concat LSTM (Pre-trained Embeddings)	1.04

“Life is not a pony show”

AKA, life is not a Kaggle competition.

Meaning, the point isn’t to find the ***best possible*** model, it is to find the ***best possible realistic*** model.

For instance:

If we had these features:
Title + Description + Plot Summary + Review + What Parents Need to Know + Genre + Type + Date
We would surely have the target as well
CSM Rating

So, create a realistic model:

Title + Genre + Type + Date
CSM Rating

LSTM Results: Title + Genre + Type + Date

Custom Embeddings train	= 0.74
Custom Embeddings test	= 1.79
Pre-trained Embeddings train	= 1.36
Pre-trained Embeddings test	= 1.65

Final Results

Model	*MAE*
Naive Baseline (Mean)	3.27
Model 3: Computer Vision	2.50
Model 2C: Title + Genre + Type + Date (PTE)	**1.65**
Model 1: XGBoost	1.62
Model 2A: Concat LSTM	1.38
Model 2B: Concat LSTM (Pre-trained Embeddings)	1.04

Future Plans

- Feature Engineering
  - Authors
  - Page Numbers
  - Awards
- Multiple Inputs

Sources

Introduction to XGBoost

Brownlee, J (2019) How to Configure Image Data Augmentation in Keras. Retrieved from

https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/

Olah, C (2015) Understanding LSTM Networks. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Pennington, Socher, & Manning (2014) GloVe: Global Vectors for Word Representation.

Keep in touch!!!!!!!!

Blog: Educators R Learners

Email: p.evansimpson@gmail.com

LinkedIn: linkedin.com/in/evansimpson1/

Twitter: @pevansimpson

Repo: https://github.com/educatorsRlearners/book-maturity