PREDICTING AMAZON BOOK REVIEW SCORES

Arnaldo Chacón Madrigal
arnaldochacon1996@gmail.com

Félix Jiménez Boza
felixjb13@gmail.com

November 14, 2023

ABSTRACT

In this data science project, we focused on predicting Amazon book review scores. After careful preprocessing, which involved the removal of unused variables, handling outliers, and filling missing values, we conducted feature engineering on the text column of the reviews, extracting valuable information. Transformations were applied to enhance the predictive power of our variables. Our comprehensive approach included the utilization of various machine learning models, among which the XGBoost model emerged as the most effective, reaching an Root Mean Squared Error (RMSE) of 0.99 and an R-squared score of 0.12. This report details our methodology, highlights key findings, and discusses the significance of our results in the context of predicting Amazon book review scores.

***K*eywords** Amazon books reviews · Regression · Data Analytics

1 INTRODUCTION

Amazon.com is a prominent e-commerce and web services platform. Initially focused on online book sales, Amazon faced the challenge of fostering customer trust in purchasing intangible goods. To address this issue, they implemented a customer review system, effectively resolving the problem and simultaneously establishing a feedback mechanism.

Over time, the reviews section evolved into a crucial aspect of the online shopping experience. Additionally, Amazon introduced a "Helpfulness" system, enabling users to upvote or downvote reviews, thus assessing their relevance based on the proportion of positive votes to total votes.

This study delves into the relationship between review content and ratings using various machine learning techniques. The aim is to determine whether a rating can be predicted primarily based on the characteristics of the text review.

1.1 MODELS

The model selection for a regression project is primarily focused on analyzing the strengths and weaknesses of a set of regression models that would best align with the distributions and characteristics of the selected set of variables from the dataset: Some of the models considered for the review prediction are:

- Linear Regression: Useful when the relationship between the characteristics and the target variable is linear. Requires that the characteristics have a significant correlation with the target variable to obtain good results.
- Decision Tree: Useful when the relationships between the characteristics and the target variable are nonlinear or complex. Can handle numerical and categorical features without the need for standardization.
- Random Forest Regressor: Handles non-linear relationships and interactions, robust to outliers and biases in features and Provides feature importance.
- XGBoost Regressor: Handles non-linear relationships well. Can be Robust to outliers and biases in features, and Internal feature selection for irrelevant features.
- Neural Network: After being trained, can be used to find highly complex relationships between variables otherwise hard to detect to the human eye.

1.2 HYPERPARAMETERIZATION

The hyperparameterization consists mainly of a subset of variables that can be used to adjust it to accomplish better performance and effectiveness, variables distinct to the main ones of the model. These hyper parameters are usually auto set by default, but an usual technique to better fit using these parameters is brute-forcing an user predefined set and testing which will let the model obtain the best results. [5]

1.3 TRANSFORMATIONS

It is a technique by which we can boost our model performance. Feature transformation is a mathematical transformation in which we apply a mathematical formula to a particular column(feature) and transform the values which are useful for our further analysis. It is also known as Feature Engineering, which is creating new features from existing features that may help in improving the model performance.[6]

2 DATASET DESCRIPTION

The dataset used in this this project is divided in to two datasets:
The first dataset, Amazon Books Reviews, contains about 3 Million Reviews for 212404 unique books available on Amazon Books Online Store. The data set is part of the Amazon review dataset that contains multiple product reviews and metadata from Amazon, including 142.8 million reviews from May 1996 to July 2014.

The second dataset, Books Data, contains detailed information of the 212404 unique books from the previous dataset. Books Data file is built by using Google Books API to get information about the books, including the published year, the authors, the publisher, and more.

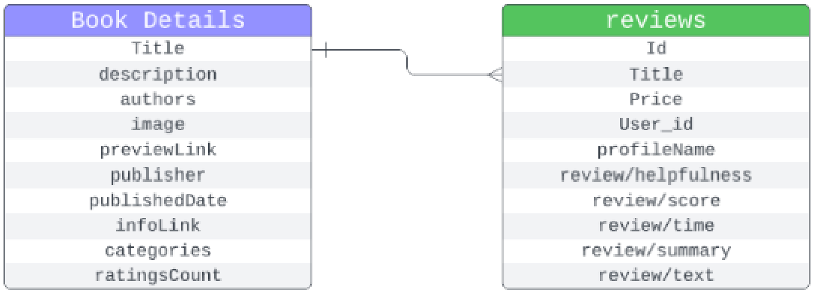


Figure 1: Datasets "Book Details" and "Reviews" (Book Reviews). Joined by book "Title". Source: [1]

Both Datasets are joined by book Tittle.

Source: Mohamed Bekheet. Amazon Books Reviews [1]

2.1 HYPOTHESIS GENERATION

The Hypothesis of the project is that there is a significant relationship between the information contained in the dataset and the review score of a book review. By applying regression techniques it is expected to model and demonstrate the relation and patterns between the dependent variable (review score) and the independent variables chosen from the dataset . That server as predictors of the review score. The validity of this hypothesis could provide an effective model for accurately anticipating book review scores based on the information provided in the dataset.

3 METHODOLOGY

The basic methodology is described by the following figure. Which is further elaborated on the following subsections.

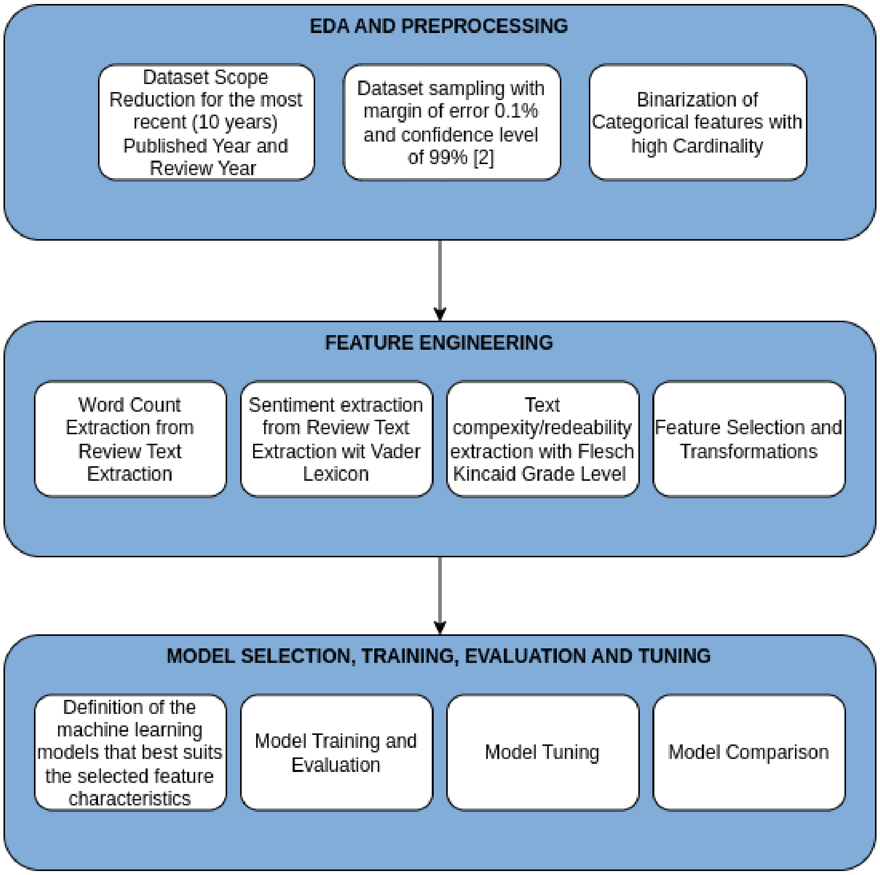

EDA AND PREPROCESSING
Dataset Scope Reduction for the most recent (10 years) Published Year and Review Year
Dataset sampling with margin of error 0.1% and confidence level of 99% [2]
Binarization of Categorical features with high Cardinality
FEATURE ENGINEERING
Word Count Extraction from Review Text Extraction
Sentiment extraction from Review Text Extraction wit Vader Lexicon
Text compexity/redeability extraction with Flesch Kincaid Grade Level
Feature Selection and Transformations
MODEL SELECTION, TRAINING, EVALUATION AND TUNING
Definition of the machine learning models that best suits the selected feature characteristics
Model Training and Evaluation
Model Tuning
Model Comparison

Figure 2: Project Methodology Overview. Elaborated by Arnaldo Chacón and Félix Jiménez

3.1 PREPROCESSING

The data was presented in the form of two csv files on the Kaggle Dataset web page. The preprocesing involved the removal of multiple variables because high cardinality and low predictive power.

Published Date variable was converted to year and multiple wrong values where handle, for example some years extracted had the format of "YYYY*" or "YYY?". For those cases, the non-numeric characters were removed or changed by the respective mean value, as applicable.

Review Date was converted from the original UTC format, to python datetime format, and then the year is extracted.

Helpfulness of the review is converted to percentage as it represents a proportional feature.

Categories list feature is exploted, Pareto principle is applied to reduce the cardinality, and then, as the cardinality is still high and main category "Fiction" is the most represented (42.7%), the categories feature is binarize as a new feature fiction Category (0 or 1).

For Publisher feature, Pareto principle is applied to reduce the cardinality, and then, as the cardinality is still high, the feature is binarized on a new feature called "top_publisher". Where 1 represents the 10 top publishers most reviewed in the dataset, and 0 all the other publishers.

From review text feature, text was converted to lower, special characters where removed using regular expressions, stop words where removed, words where lemmatized keeping the words with more than 3 letters. Outliers and duplicates where removed from the dataset.

3.2 FEATURE ENGINEERING

Three new variables where extracted from the review text feature:

3.3 Word Count

From the clean text, and preprocesing and lemmatizing the review text, the total number of tokens per review text were counted and stored on this new variable.

Box-Cox transformation was applied to word count in order to make the distribution of this variable more symmetric.

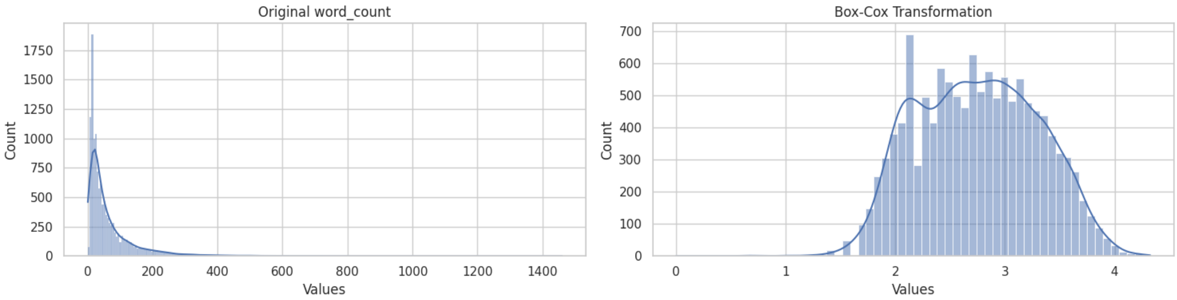


Figure 3: Word Count feature distribution. Box-Cox Transformation. Elaborated by Arnaldo Chacon and Félix Jimenez

3.4 Compound Sentiment

Vader lexicon was used to measure the review text sentiment. This lexicon provides an output variable with values between -1 and 1. Where values near or equal to 1 represent positive sentiment, values equal or near to -1 represent negative sentiment, and values equal or near to 0 represent neutral sentiment, from the review text.

Hyperbolic Arctan Transformation was applied to compound sentiment in order to make the distribution of this variable more symmetric.

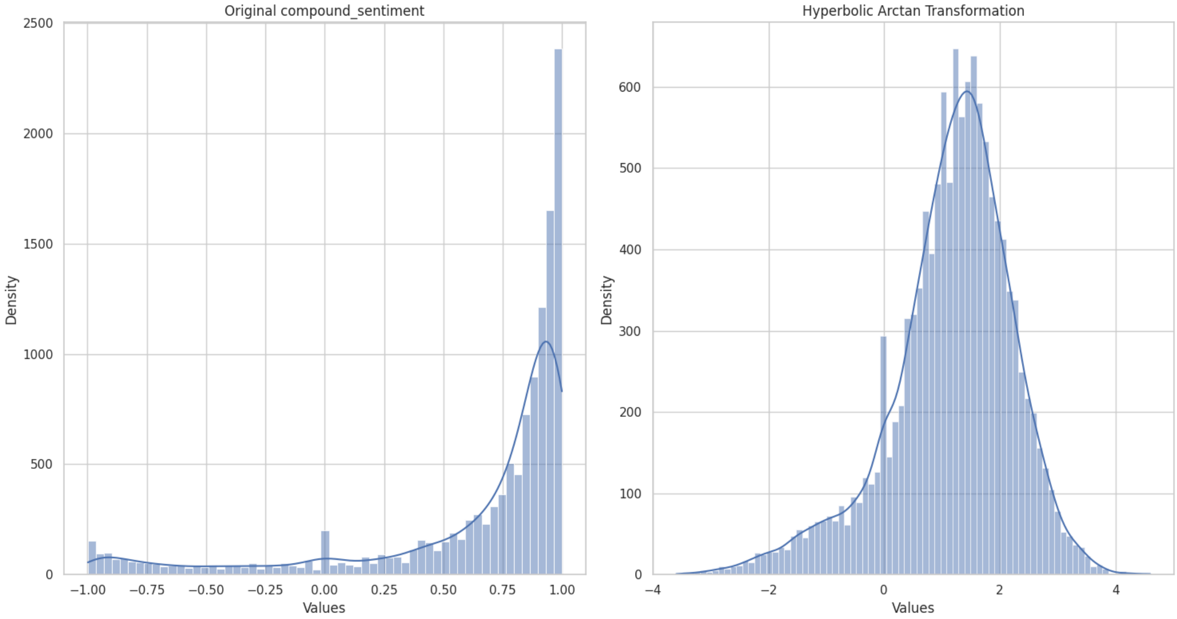


Figure 4: Compound Sentiment feature distribution. Hyperbolic Arctan Transformation. Elaborated by Arnaldo Chacon and Félix Jimenez

3.5 Text Complexity

Flesch-Kincaid Grade Level Score was used to measure the complexity/readability of the original text.

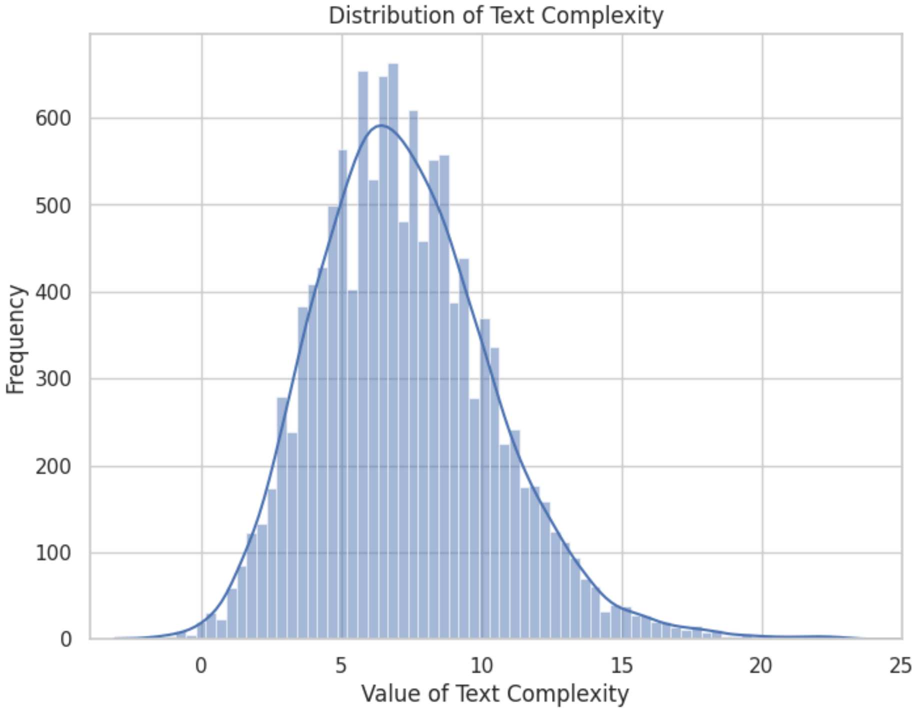


Figure 5: Text Complexity feature distribution.. Elaborated by Arnaldo Chacon and Félix Jimenez

The expected results for FKGL are as follows:

- 0-6: Easy to read, typically understandable by an average 11-year-old student or below.
- 6-8: Readable by a 12-14-year-old student (7th to 8th grade).
- 8-10: Readable by a 14-16-year-old student (9th to 10th grade).
- 10-12: Readable by a 16-18-year-old student (11th to 12th grade).
- 12-14: Readable by a college graduate.
- 14-16: Readable by someone with a post-graduate education.

3.6 Feature Selection. Final Feature Set

- fiction_category: Binary feature. It is 1 (True) when reviewed book category is "Fiction", and 0 when other.
- top_publisher: Binary feature. It is 1 (True) when reviewed book publisher is in the Top 10 most reviewed publishers, and 0 when not.
- helpfulness: Continuous Numerical feature. Represents a proportion of the amazon users reactions to the review that have identified it help full vs non-help full.
- compound_sentiment: Continuous Numerical feature. Represents the sentiment measured with Vader Lexicon on the review text.
- text_complexity: Continuous Numerical feature. Represents the text complexity/readability measured with Flesch-Kincaid Grade Level Score on the review text.
- word_count: Discrete Numerical feature. Represents the count of tokens quantified from processed and lemmatized review text.
- review: Review Score of the review, given by the Amazon Books User.

4 Exploratory Data Analysis

4.1 Numerical Features

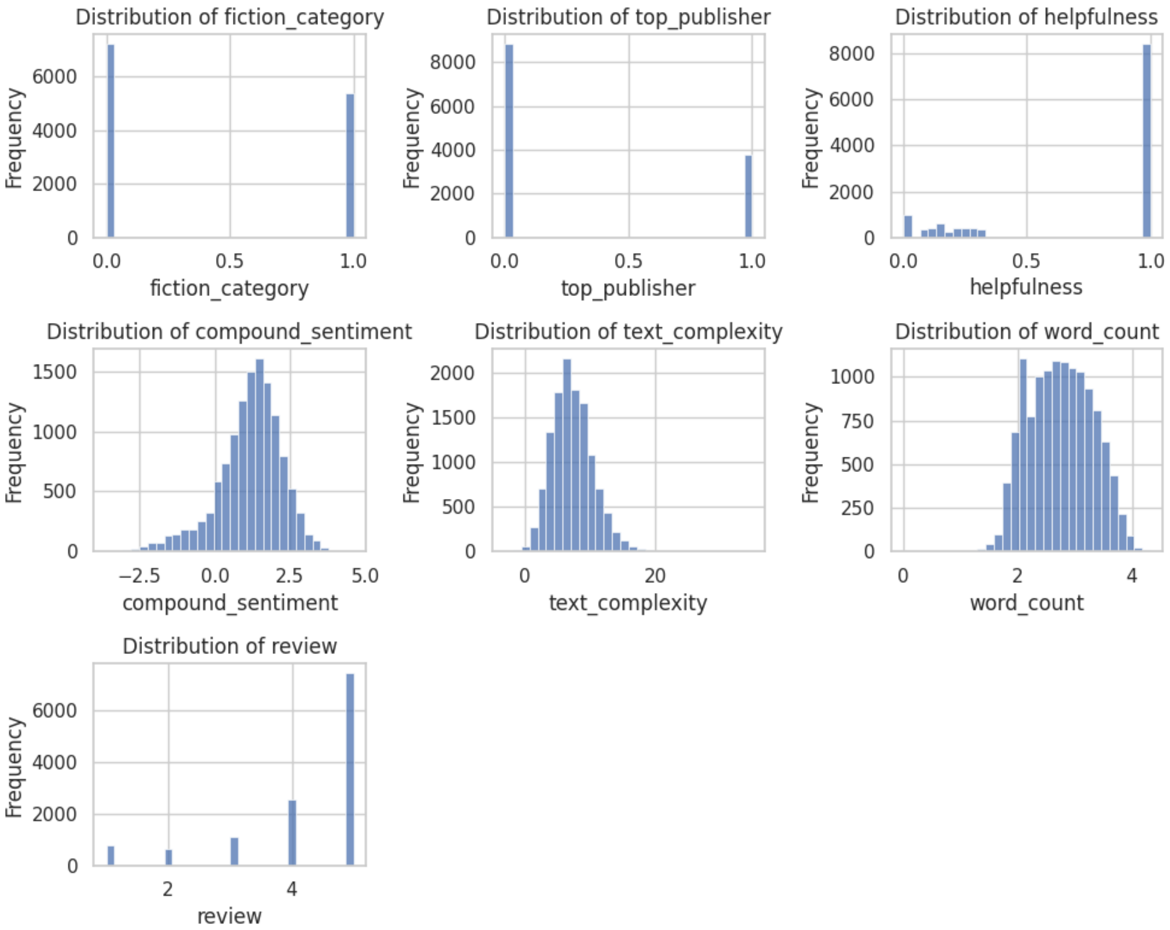


Figure 6: Distribution of selected features. Elaborated by Arnaldo Chacon and Félix Jimenez

We start our EDA with the frequency of all of the variables. The first thing to notice is the disparity there is in the categorical variables, with 'fiction_category' and 'top_publisher' balancing more towards lower values, while 'helpfullness' and 'review' showed greater frequency in their higher ends (1 for 'helpfullness' and 5 for 'review')

Moving to the discrete variables, 'compound_sentiment' is more skewed to the right, whereas 'text_complexity' trends positively. The variable 'word_count', as expected, has a normal behavior.

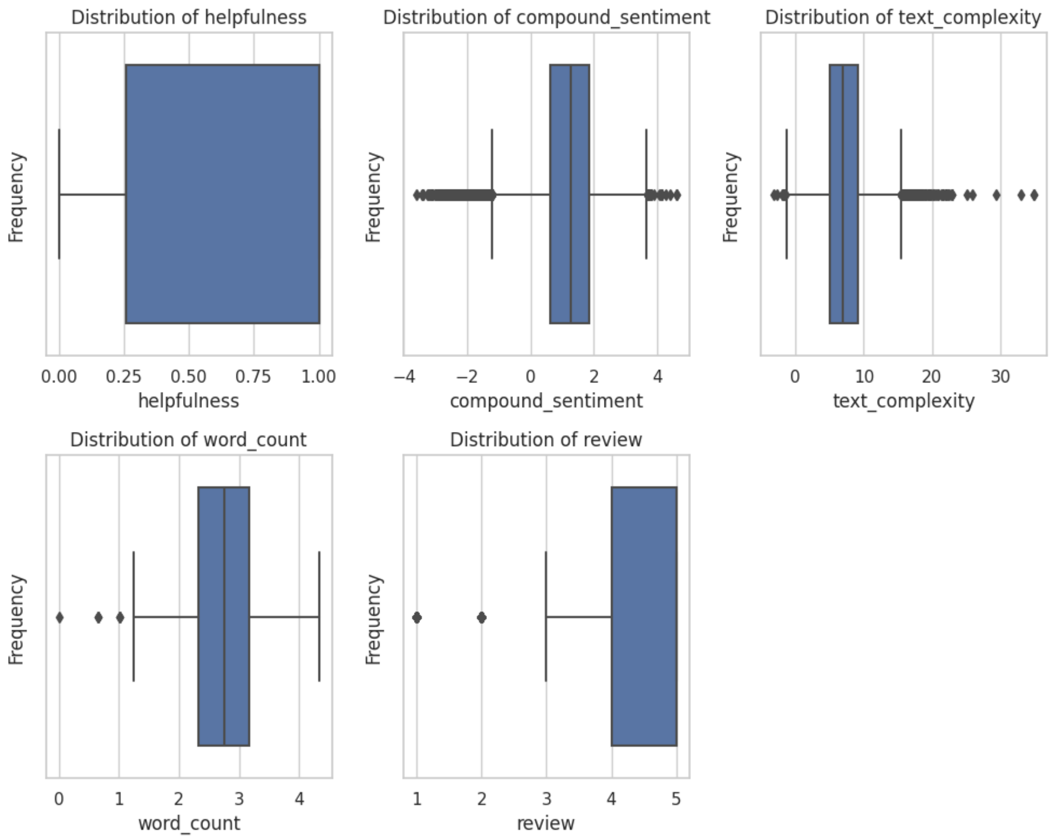


Figure 7: Box Plots of selected features. Elaborated by Arnaldo Chacon and Félix Jimenez

Boxplots showed us that helpfulness has a broad distribution mainly charged by the high quantity of 1 values set previously and a few elements of the lower end, creating a false sense of broadness. In consistency, this negative bias is also present in 'compound_sentiment' and 'review'. Main 'text_complexity' values are located in the 5-10 portion (ranging '*6-8: Readable by a 12-14-year-old student (7th to 8th grade)*' and '*8-10: Readable by a 14-16-year-old student (9th to 10th grade)*').

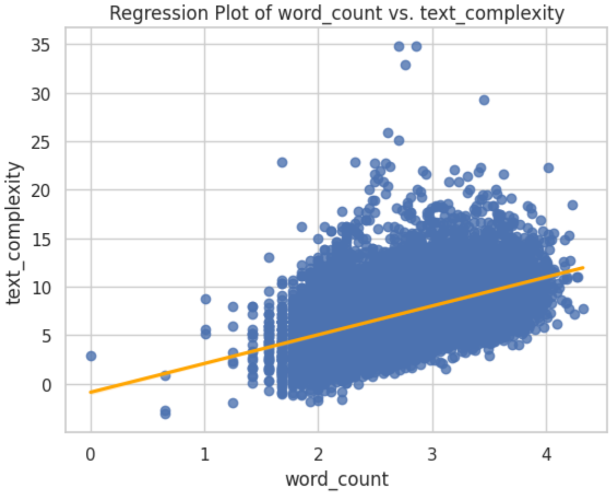


Figure 8: Word Count vs Text Complexity. Elaborated by Arnaldo Chacon and Félix Jimenez

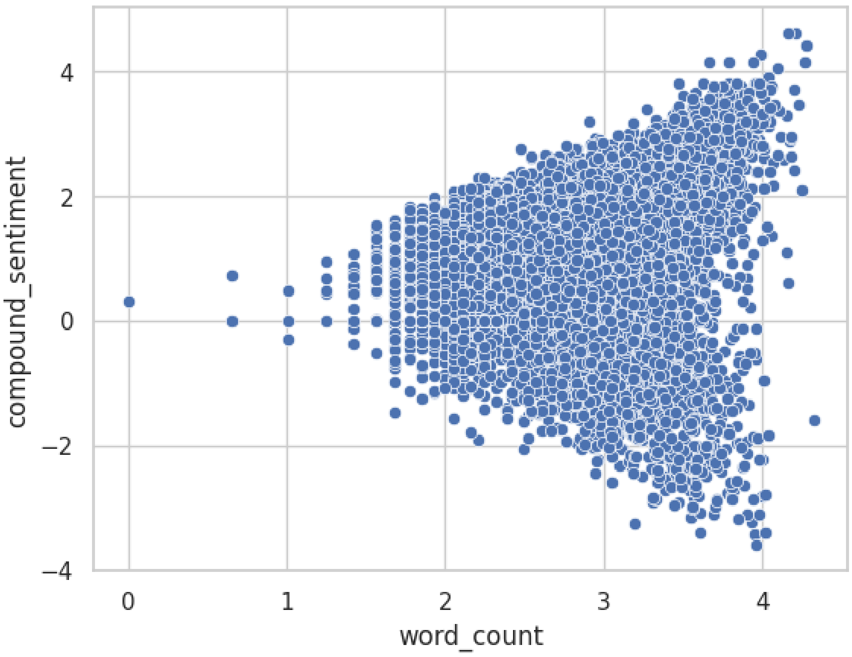
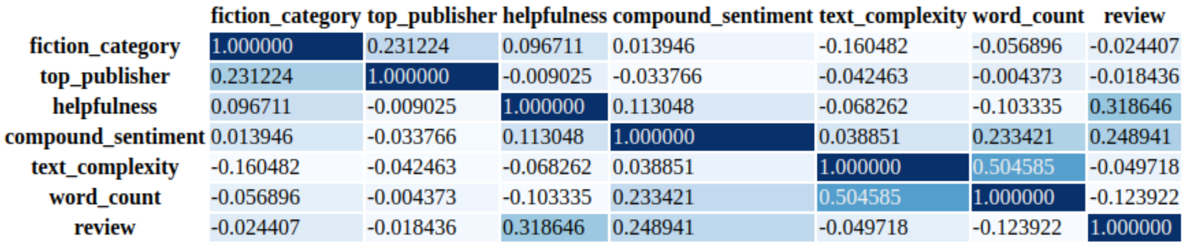


Figure 9: Word Count vs Compound Sentiment. Elaborated by Arnaldo Chacon and Félix Jimenez

Both 'text_complexity' and 'compound sentiment' when plotted against 'word_count' did not show any clear trend.



	fiction_category	top_publisher	helpfulness	compound_sentiment	text_complexity	word_count	review
fiction_category	1.000000	0.231224	0.096711	0.013946	-0.160482	-0.056896	-0.024407
top_publisher	0.231224	1.000000	-0.009025	-0.033766	-0.042463	-0.004373	-0.018436
helpfulness	0.096711	-0.009025	1.000000	0.113048	-0.068262	-0.103335	0.318646
compound_sentiment	0.013946	-0.033766	0.113048	1.000000	0.038851	0.233421	0.248941
text_complexity	-0.160482	-0.042463	-0.068262	0.038851	1.000000	0.504585	-0.049718
word_count	-0.056896	-0.004373	-0.103335	0.233421	0.504585	1.000000	-0.123922
review	-0.024407	-0.018436	0.318646	0.248941	-0.049718	-0.123922	1.000000

Figure 10: Correlation Matrix for Selected features. Elaborated by Arnaldo Chacon and Félix Jimenez

Interestingly enough, features with greater correlation are 'word_count' and 'text_complexity', probably coming from that a bigger text allows for a greater diversity in words, reaching 50.45% of correlation. In second place, 'helpfulness' with a correlation of 31.86% and 'compound_sentiment' in third with 24.89%, both against 'review'. No other features showed notable correlation.

4.2 Text Feature - NLP

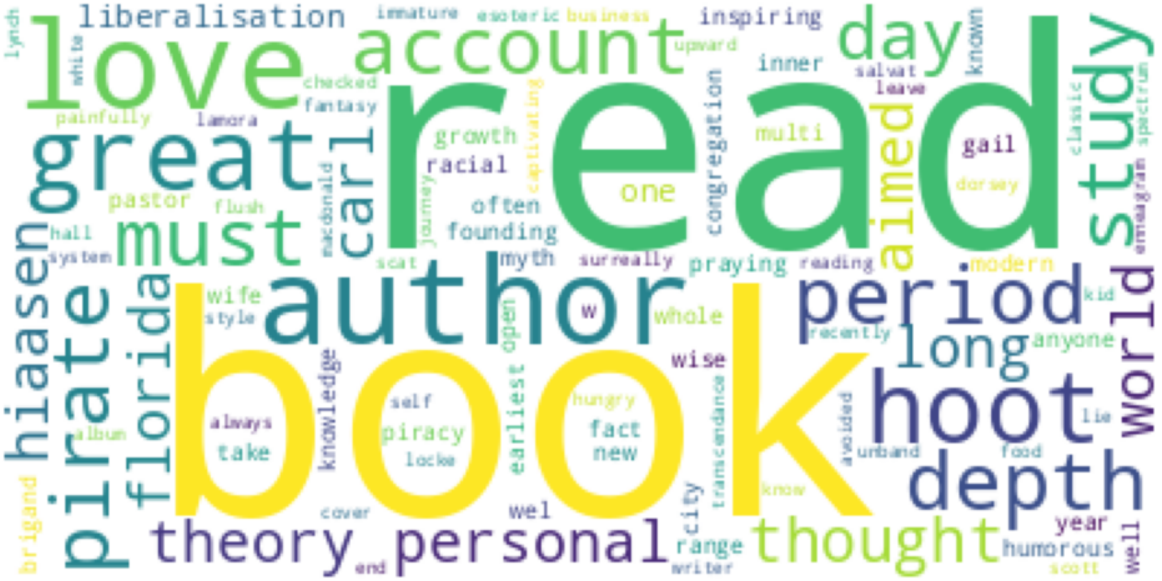


Figure 11: Word Cloud for Reviews with **Highest Score (5)**. Elaborated by Arnaldo Chacon and Félix Jimenez

Expected words to be found in this text: 'great', 'inspiring', 'growth', 'painfully'. Are words a human can be related to and may have positive sensation when giving the review.

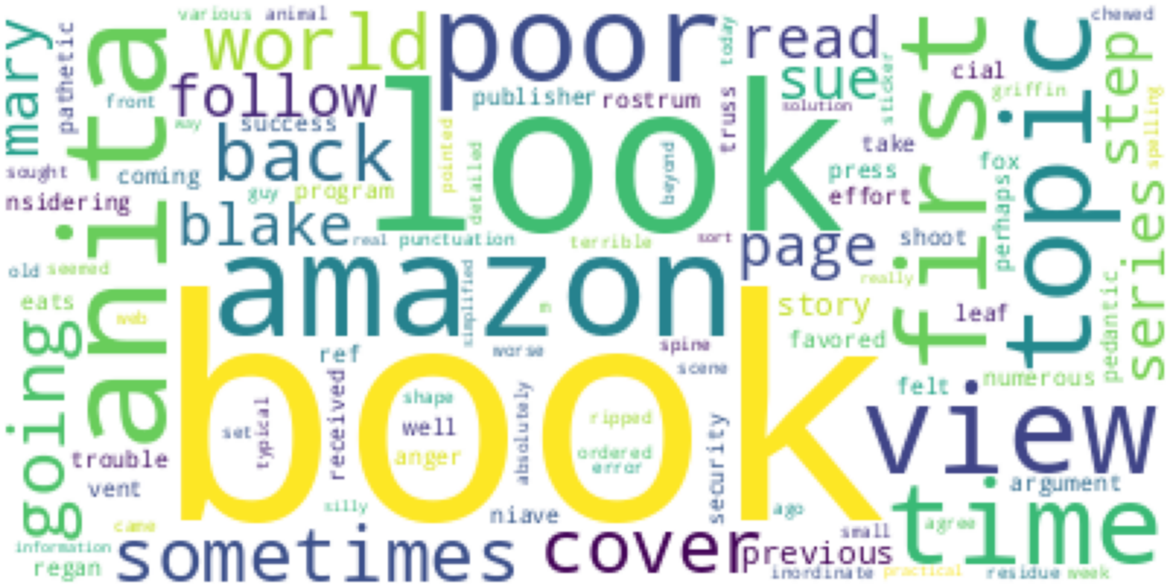


Figure 12: Word Cloud for Reviews with **Lowest Score (1)**. Elaborated by Arnaldo Chacon and Félix Jimenez

Expected words to be found in this text: 'poor', 'pathetic', 'trouble'. Are words a human can be related to and may have negative sensation when giving the review.

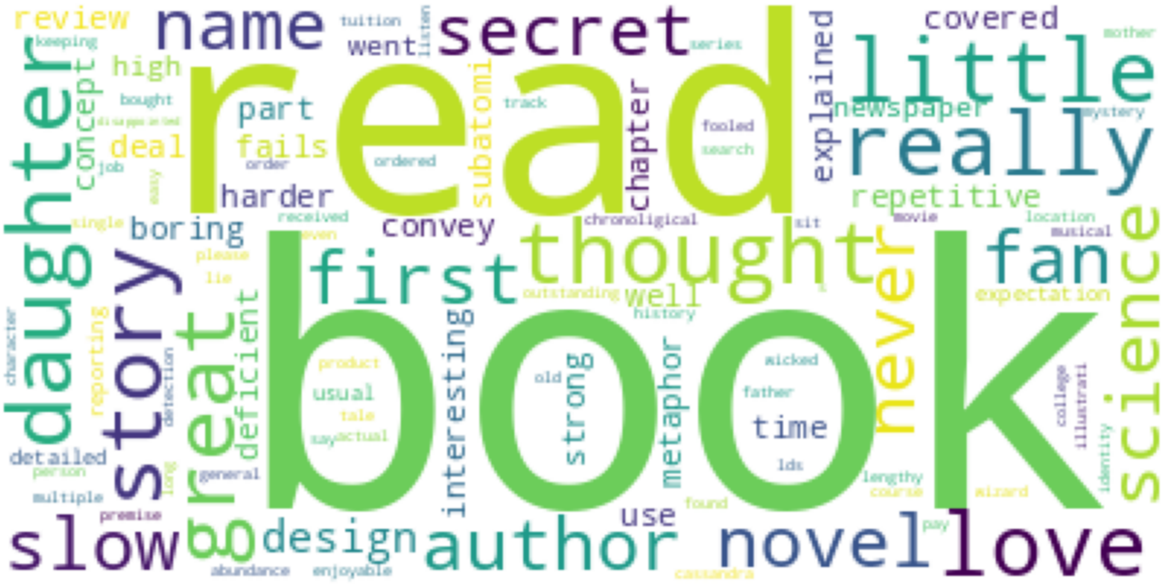


Figure 13: Word Cloud for Reviews with **Neutral Score (3)**. Elaborated by Arnaldo Chacon and Félix Jimenez

Neutral Review Score Word Clould from the figure 12 shows the main words for this score. As expected, on this wordcloud a mix of positive (i.e., love, fan, well, interesting, etc) and negative (i.e., deficient, slow, boring, etc) words can be found.

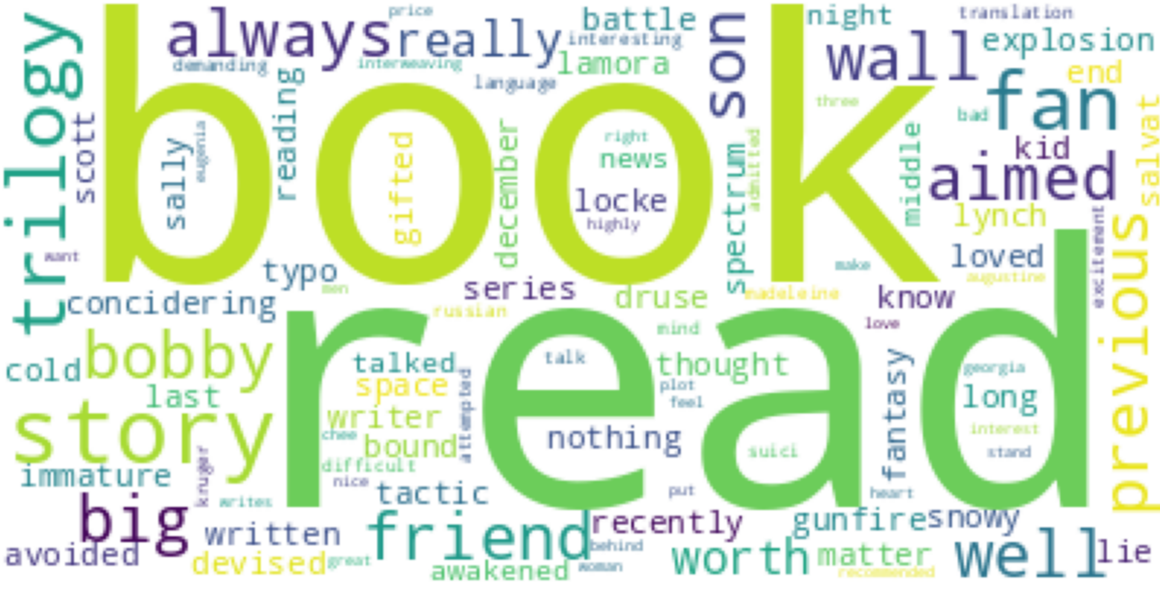


Figure 14: Word Cloud for Reviews of **Fiction Books**. Elaborated by Arnaldo Chacon and Félix Jimenez

For Fiction books reviews many words related with this category are shown with high importance on the wordcloud, some examples are fantasy, battle, triology, tactic and spectrum.

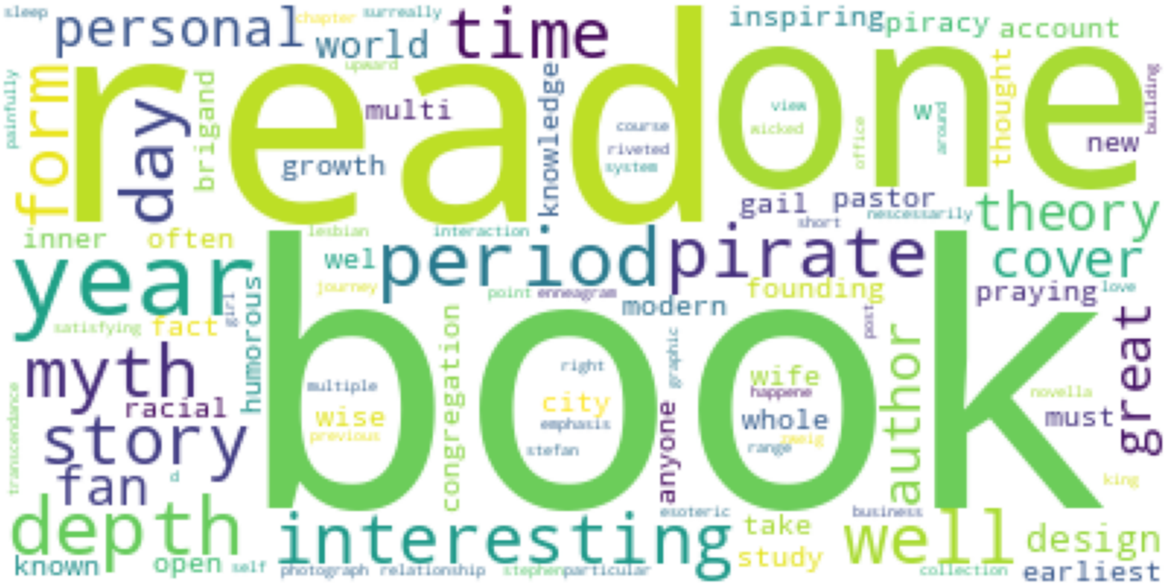


Figure 15: Word Cloud for Reviews of **Non Fiction Books**. Elaborated by Arnaldo Chacon and Félix Jimenez

Main words present on reviews related related with non-fiction categories are presented on the figure 14. Showing some words related with the real world or day to day living, like design, modern, myth, racial, inspiring, business and knowledge.

Some words are repeated across all the wordclouds, as read, book, author, and time. Some of these words will not make a huge difference in order to describe the relations or patterns between the review text and the review score. But specific words found only on specific wordclouds will make a big difference when using the vectorized text to make predictions.

5 DISCUSSION AND RESULTS

5.1 Model Selection

For creating a good predictive model, selecting the right machine learning algorithm is a very important decision that significantly influences the performance and accuracy of the model. This process involved a series of steps, including model selection, training, evaluation, and tuning.

The initial phase of our workflow involved meticulous model selection, where we considered a range of algorithms suited to the nature of our dataset, including Random Forest, XGBoost, Linear Regression, Decision Tree, and Neural Network, with the aim of comprehensively understanding their strengths and limitations. And based on the theoretical information from section 1.1.

Subsequently, the selected models where exposed to the training data, where the importance of features for each specific Model where identified, see figure 15.

Following training, each model was evaluated. Metrics such as Root Mean Squared Error (RMSE) and R2 Score were employed to gauge the accuracy and reliability of the models.

5.2 Model Hyperparameter Tuning

To enhance the models' performance, a crucial step involved hyperparameter tuning. This process fine-tuned the internal settings of each algorithm to optimize their predictive power. Grid search technique was applied to discover the most effective combination of hyperparameters for each model.

The XGBoost found out to be greatly benefited after tuning, having an improvement as follows:

Model	RMSE	R2 Score
Random Forest	1.0149	0.2776
Random Forest Tuned	1.0069	0.2888
XGBoost	1.0244	0.2640
XGBoost Tuned	**0.9859**	**0.3182**
Linear Regression	1.0368	0.2460
Decision Tree	1.4449	-0.4642
Neural Network	1.082	0.2586

Table 1: Machine Learning Models Root Mean Square Error and R2 Score Comparison

Results for the evaluation phase are summarized on the Table 1. Where Hyper parameterized XGBoost provided the best results.

5.3 Model Evaluation

Given that XGBoost was the model with the best adjustment, we proceed with an analysis of the feature weight after modeling.

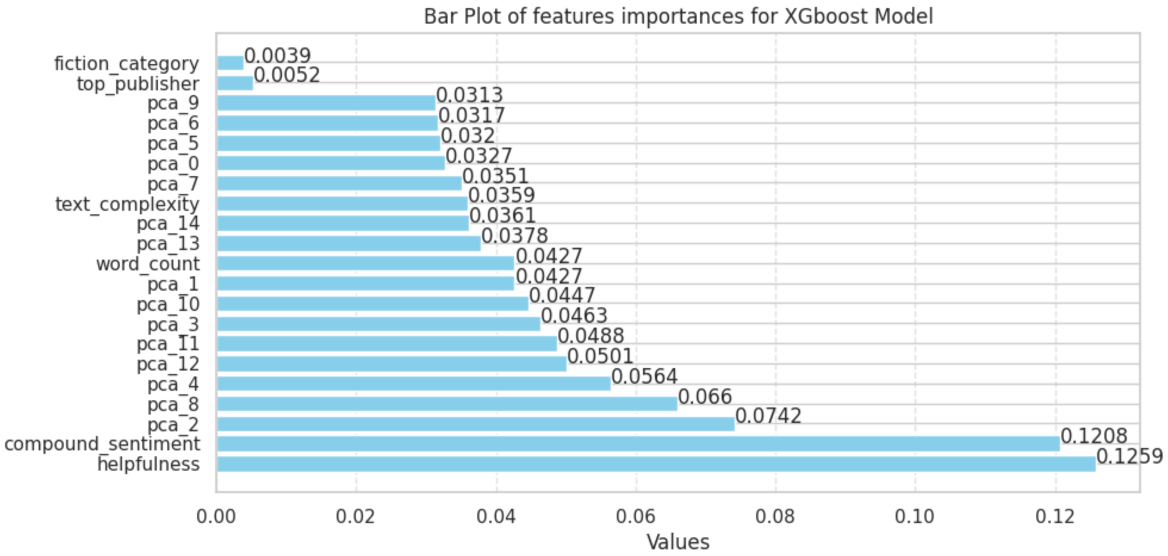


Figure 16: XGBoost Algorithm, parameter importance. Elaborated by Arnaldo Chacon and Félix Jimenez

Note that 8 of the top 10 most important features in the dataset are variables coming from the PCA application. Nonetheless, 'compound_sentiment' and 'helpfulness' were our main contributors to this model prediction.

6 CONCLUSIONS

As shown in the Table 1, XGBoost was the best model for predicting Amazon book review scores. Its ability to learn from mistakes and improve with each attempt, known as gradient boosting, helped it capture the relations and patterns in our data.

Results for features importance from Figure 15 show that what reviewers and readers find helpful and the overall sentiment in the reviews turned out to be the most important features. This suggests that community opinions and the emotional tone of reviews matter a lot in determining how books are rated.

Extracting information from the actual text of the reviews was crucial. Tools like TF Vectorizer and PCA helped us understand the importance of words and expressions in predicting review scores. From the figure 15, it shown that 8 of the Top 10 most important features where extracted from the text.

In summary, our testing with different machine learning models and features extraction from amazon reviews revealed a lot about what matters in predicting Amazon book review scores. Understanding what readers say and feel in their reviews, especially in the text, is crucial.

References

[1] Mohamed Bekheet. *Amazon Books Reviews [Dataset]*. 2022. https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

[2] https://www.surveymonkey.com/mp/sample-size-calculator/

[3] Readable. *Flesch Reading Ease and the Flesch Kincaid Grade Level*. https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

[4] Aryan Bajaj. *Can Python understand human feelings through words? – A brief intro to NLP and VADER Sentiment Analysis*. 2023. https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/

[5] 4 Geeks. *Hyperparameter optimization*.2023. Notesfromvirtualclassroom

[6] Analytics Vidhya *Feature Transformations in Data Science: A Detailed Walkthrough*. 2022. https://www.analyticsvidhya.com/blog/2021/05/feature-transformations-in-data-science-a-detailed-walkthrough/