Regression Analysis for Book Price
Machine learning for statistical NLP: Advanced LT2326

Lívia Vicente Dutra
University of Gothenburg

1. Introduction

Books have been part of the world for milenios and its price - not only monetary, but also regarding the owner's status - has varied through the years. Some centuries ago, to own books or even to be able to read them was a privilege for few people, but that changed over the years and books now are much more accessible. To be more accessible doesn't mean, though, that they aren't expensive.

According to some surveys as the one presented in Lectupedia, Canada has an average of 17 books read per year while other countries vary this number being Argentina with the lower number 1.6 books per year. This number tends to increase though, because recent surveys shows for exemple that the new generation has increased this number. As a reader myself, I wonder constantly about people's motivations to read and I seek to stimulate the most people I can to get a book. While technology is a great tool and improves our life exponentially, it can be a setback to stimulating reading, since it seems to be more appealing than a book.

But apart from technology, could the price of books also create a distance between people and books? Recently, searching for an academic book, I was surprised to find out that it cost over 500 pounds. What features determined the price of that book, or of any book in fact? Could that be, that once again, books will be a privilege?

The question that motivated this project was " What can impact a book price and can we predict it?". Beside my own motivation, along with some research I came to understand that it could also be of value to answer this question when it comes to the publisher market.

This project aims to analyze if regression models can be utilized to predict book prices and what possible factors would have a direct impact on the final price. I will go through a short background on some similar research, an overview on the methodology used on this project and the final results obtained along with the analyses and my final conclusions.

2. Background

When it comes to predicting prices, regression models are very much up for the task. Those models aim to capture the underlying patterns and dependencies within the data, allowing for the estimation of numerical outcomes. Because of that, we can find a lot of research on the topic, but from a general point of view. Reddy and Sriramya (2022), for example, published a research using regression models to predict real estate prices. Stock predictions are also a hot-topic when it comes to regression models.

When talking about book predictions, however, there was very little material to serve as support. Fathalla, Salah and Ali (2023) published a paper on Novel Prediction for E- commerce that gave me a little more understanding on the topic, but I only found it by the end of the project.

Surprisingly, my best material for support for this project was challenges posted on blogs that incentivize people to try to accomplish goals similar to mine. Particularly, the challenge along with the tutorial on the website Analytics India Mag outstanded from others available.

3. Methodology

The Methodology for this project consisted of dataset collection, dataset preprocessing, the study and test of different regression models, the evaluation and analysis.Each stage will be discussed below:

3.1. Dataset Collection

The first question regarding the project was which corpus would be the best fit for this study. After some research, I came to the conclusion that to be able to analyze the features and factors I was interested in observing the behavior, I wouldn't be able to use an already done corpus as the ones available, for example, in Kagle. Among other factors, that many times were not sufficient, I needed to have a price for all the books and that meant for the specific versions presented in the corpus and that could not only be extremely hard, but also very time consuming.

Therefore, I decided to construct my own corpus. There are many down points on following this path. First, it is also very time consuming and because of that I ended up with a very limited size corpus - 900 books -, especially due to the time frame we had to accomplish the task. Bias could also be a problem, since we tend to pick genres and authors already know, which would kill the diversity of the corpus. Along with that, we also have the potential for errors, because it is very easy to have typos, mistake, miss or duplicate some information.

However, there are also many relevant advantages on building your one corpus as well. First, I got to have control of my corpus and its quality, that means I got the chance for example to revise errors and fix them, I knew exactly my sources and why they had been targets and more. I was able to pick a specific domain, which was Amazon United Kingdom, that means I had a balanced level of information, all books with the same currency for prices and with the same logic for calculating them and also the possibility of working with kindle and paperback versions. I was also able to try to minimize the bias by using tools such as the ChatGPT - I would ask for 7 books of 10 different authors of a given genre. It didn't eliminate the bias issue, but it expanded the diversity. But, maybe the most important, pro factor on building this corpus, was the possibility of customization.

As previously mentioned, my focus on this project was to build a suitable corpus for the purpose of this project. That means I had specific features in mind that I believed to be important when predicting a price. I couldn't take all the features, but I was able to select nine that I considered as main to compose the corpus (table 1): Authors, Title, Genre, Number of pages, Year, Publisher, Languages, Version and Price. At first, my intention was to have more than one language, but I decided to have only english, but keep the feature to keep the information of language and the possibility of future addition of other languages.

Authors	Title	Genre	Number of pages	Year	Publisher	Languages	Version	Price
~50	unique	~30	diverse	-----	~120	English	Kindle Paper	X

Table 1 - Corpus Information

One final consideration on the dataset collection step, is the fact I decided to create the corpus manually, instead of automating the process. To decide on doing it manually, I balanced the time factor and the experience - both on doing it manually and on creating something similar as a "web scrap".

The fact that I didn't have any experience on either, made me choose to proceed using the manual approach.

3.2. Dataset Preprocessing

Given the fact the corpus was manually created, the preprocessing of the data started by a very thorough inspection and cleaning. As mentioned above, I was able to come back and fix all the mistakes as: duplicate or missing information or types. So that part of the process helped to improve the quality of the final product. The inspection also consisted of looking for possible relations between the selected features. Unfortunately, maybe given the small size of the corpus, it wasn't possible to draw any conclusions, as for example the number of pages, genre or version had a direct impact on the final price.

I used the one-hot encoding to transform the information of the corpus and assign a unique binary value to each category by creating a binary vector representation. This method not only preserves the categorical nature of the data but also facilitates efficient computation and analysis.

Beyond that, some of the features were not bringing relevant information to the models as the books titles - they are unique and any kind of semantic analyses were conducted - and the language - it was only English. The authors names were also disconsidered, but after we had performed the split of of the corpus in train and test, since the division was done by author so we would be able to predict not only seen authors, but also we wouldn't be biased by that repetitive information that could confuse the model.

3.3. Regression Models

At the beginning of the project, my idea was to build my own regression model, so aiming to better understand the behavior of the features of the corpus and its possible impacts on price prediction, I used Lazy Prediction to try to see what kind of models would better work for this dataset. From the results I gathered - and that will be discussed - it was possible to see that I couldn't get good results given my dataset, so I moved on to compare the results I got from the Lazy Prediction.

From the results obtained, I picked six models to compare the results and proceed with the evaluation: Hubber Model, MLP (Multi-Layer Perceptron)Regressor, Adaboost, SGD (Stochastic Gradient Descent), Linear SVR (Support Vector Regressor) and Random Forest Regressor.

The variation in predictive performance among machine learning models can happen given several factors. The characteristics of the dataset, such as its size, complexity, and presence of outliers, play a crucial role. The models with a better performance were Hubber and Random Forest (table 2) . Decision tree-based models like Random Forest are adept at capturing complex relationships, especially in the presence of non-linear patterns, while HuberRegressor's robustness to outliers might contribute to its effectiveness. Neural networks like MLPRegressor may require careful hyperparameter tuning and benefit from scaled features, and their performance often improves with larger datasets, which isn't the case of the dataset used. The ensemble nature of Random Forest and AdaBoost helps mitigate overfitting and the sensitivity of other models, like Linear SVR, to hyperparameters necessitates thorough tuning for optimal results.

3.4. Evaluation

When evaluating the performance of machine learning regression models, the selection of appropriate evaluation metrics is very important. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are quite good indicators of a model's predictive accuracy. MSE quantifies the average

squared discrepancies between predicted and actual values, assigning greater significance to larger errors and proving particularly sensitive to outliers. On the other hand, MAE calculates the average absolute differences, which gives a more resilient measure against extreme values. When evaluating the models pointed in the previous section with both metrics , I was able to get to some conclusions.

By analyzing the MSE results, it is possible to see that HuberRegressor and Random Forest have better performance with MSE values of 29.27 and 30.95, respectively, which indicates a lower predictive error. The fact that the HuberRegressor's deals better with outliers contributes to its accuracy, while Random Forest can capture complex relationships within the data. MLPRegressor follows with a higher MSE of 35.98, suggesting potential sensitivity to outliers and a need for hyperparameter tuning. AdaBoost has the worst performance at 47.11, indicating challenges in capturing the underlying patterns, and room for more improvement. The same pattern can be captured when observing the results from the evaluation using MAE. HuberRegressor stands out with the lowest value at 3.57 and Random Forest follows closely with 3.70. MLPRegressor exhibits a slightly higher MAE at 4.51, while AdaBoost records a higher MAE of 5.72, signifying challenges in accurately predicting book prices.

The results suggest that HuberRegressor and Random Forest can be good choices when it comes to minimizing absolute prediction errors, and the other models still need some work to better perform as improving the parameter tuning for instance, or that they might not be the best candidate given factors as the corpus size. But even though we had better results with the Hubble Regression and Random Forest, they are still not great and give us a poor perspective when it comes to book prices prediction. With that in mind, I tried to use other strategies aiming to get some improvement on the results.

Several strategies can be used to improve the models, but I chose three to test on the model. The first was to eliminate low correlations between features and the target variable improves model efficiency by focusing on the most relevant information. Next, I tried to use the Principal Component Analysis (PCA) which has a dimensionality reduction technique that captures essential features while eliminating irrelevant information. This helps to improve the models' ability to detect meaningful patterns. Both of them, although, are more used for larger corpus, and once the dimensionality wasn’t a problem for this project the results were not better. At last, I tried to adjust the hyperparameters, such as learning rates or number of estimators. Unfortunately, that also didn’t give any considerable improvement.

4. Discussion

Based on the facts and results presented, it's clear that it’s not possible to come to any conclusions on regression models and its use for book price predictions. Some points need to be addressed and taken into consideration in order to have a better comprehension of the task.

First, the size of the corpus was a big set back. I believe that a bigger corpus could provide a better result or at least a clearer direction on what the current information presented on the corpus can lead to. To come back to the idea of automating the corpus creation, so we are still able to use the same domain and customization is a good approach and worth the work.

Second, the features are not enough. I tried to gather the most straight points that could be analyzed, but there are many other features that can have a quite big impact on a book's price. We could try to implement a semantic analysis on book description , maybe through key words, combined with the analyses and considering reviews and ratings. Social aspects as market information and trends and social media can also be very influential on the price. And given the fact we are working with a

regression model, to have access to the history of prices, might also help to get a better outcome. So, to study those features and include them in the corpus seems essential.

Finally, we might not be able to accomplish the book price prediction using only a regression model. A combination of other machine learning techniques might be necessary to fully capture all the nuances necessary when it comes to predicting prices of books.

5. Conclusion

To conclude, this project aimed to understand and try to perform a satisfying result for book price prediction, focusing on capturing the most important features for accomplishing such a task. The main part consisted of creating and treating the corpus manually and analyzing the best options for regression models that could be used. It includes an evaluation and analysis of the models.

It wasn't possible to achieve the main goal of the project, but satisfying conclusions were drawn from the work that could be done. Also, next steps such corpus improvement related not only to size but features and the possible expansion of machine learning techniques

References: