Feature Engineering On Book Price Prediction

Abstract:

This study explores applying feature engineering techniques to enhance the accuracy of predicting book prices. By crafting informative features from various aspects of book metadata, such as ratings, reviews, publication year, and sentiment analysis of synopses, the research aims to improve model performance. encoding categorical variables are incorporated to capture nuanced factors influencing book pricing. The findings contribute to advancing book price prediction models, providing valuable insights for publishers.

Introduction:

Predicting book prices accurately is a critical challenge. As the market dynamics shift and reader preferences evolve, understanding the factors that influence book pricing becomes essential for publishers and consumers alike. In this context, feature engineering plays a powerful role, allowing us to extract meaningful insights from the diverse array of data associated with books.
In this study we tried improving a random forest model's prediction for prices of books only using feature engineering methods including feature crafting, sentiment analysis, feature transformation, encoding, etc. the goal og this study was lowering the mse for test and train data as much as possible.

Methods:

The first step was to load train and test data and create a dataframe by concating these two so all the feature transformation steps would be applied to both of them. First, let's see the shape of test and train data:
Test: (537, 9)
Train: (5699, 9)

They both have 9 features but the test includes an "Unnamed: 0" column indicating its index and train data contains a "Price" column. By concating them we'd have a df with the shape of :
(6236, 10)
Indicating there is 6236 data samples and 10 features in this dataset.

Now next let's check if there is any missing value in this dataset.
Title 0
Author 0
Edition 0
Reviews 0
Ratings 0
Synopsis 0
Genre 0
BookCategory 0
Price 537
Unnamed: 0 5699
dtype: int64

We only have missing values for Unnamed: 0 and Price which is expected. But the rest of the data is so far clean.
By looking at the first 5 elements of the data:

	Title	Author	Edition	Reviews	Ratings	Synopsis	Genre	BookCategory	Price	Unnamed: 0
0	The Prisoner's Gold (The Hunters 3)	Chris Kuzneski	Paperback,– 10 Mar 2016	4.0 out of 5 stars	8 customer reviews	THE HUNTERS return in their third brilliant no...	Action & Adventure (Books)	Action & Adventure	220.00	NaN
1	Guru Dutt: A Tragedy in Three Acts	Arun Khopkar	Paperback,– 7 Nov 2012	3.9 out of 5 stars	14 customer reviews	A layered portrait of a troubled genius for wh...	Cinema & Broadcast (Books)	Biographies, Diaries & True Accounts	202.93	NaN
2	Leviathan (Penguin Classics)	Thomas Hobbes	Paperback,– 25 Feb 1982	4.8 out of 5 stars	6 customer reviews	"During the time men live without a common Pow...	International Relations	Humour	299.00	NaN
3	A Pocket Full of Rye (Miss Marple)	Agatha Christie	Paperback,– 5 Oct 2017	4.1 out of 5 stars	13 customer reviews	A handful of grain is found in the pocket of a...	Contemporary Fiction (Books)	Crime, Thriller & Mystery	180.00	NaN
4	LIFE 70 Years of Extraordinary Photography	Editors of Life	Hardcover,– 10 Oct 2006	5.0 out of 5 stars	1 customer review	For seven decades, "Life" has been thrilling t...	Photography Textbooks	Arts, Film & Photography	965.62	NaN

We can see that two columns Ratings and Reviews have each others column name. So we'd replace these two by each other.

	Title	Author	Edition	Ratings	Reviews	Synopsis	Genre	BookCategory	Price	Unnamed: 0
0	The Prisoner's Gold (The Hunters 3)	Chris Kuzneski	Paperback,– 10 Mar 2016	4.0 out of 5 stars	8 customer reviews	THE HUNTERS return in their third brilliant no...	Action & Adventure (Books)	Action & Adventure	220.00	NaN
1	Guru Dutt: A Tragedy in Three Acts	Arun Khopkar	Paperback,– 7 Nov 2012	3.9 out of 5 stars	14 customer reviews	A layered portrait of a troubled genius for wh...	Cinema & Broadcast (Books)	Biographies, Diaries & True Accounts	202.93	NaN
2	Leviathan (Penguin Classics)	Thomas Hobbes	Paperback,– 25 Feb 1982	4.8 out of 5 stars	6 customer reviews	"During the time men live without a common Pow...	International Relations	Humour	299.00	NaN
3	A Pocket Full of Rye (Miss Marple)	Agatha Christie	Paperback,– 5 Oct 2017	4.1 out of 5 stars	13 customer reviews	A handful of grain is found in the pocket of a...	Contemporary Fiction (Books)	Crime, Thriller & Mystery	180.00	NaN
4	LIFE 70 Years of Extraordinary Photography	Editors of Life	Hardcover,– 10 Oct 2006	5.0 out of 5 stars	1 customer review	For seven decades, "Life" has been thrilling t...	Photography Textbooks	Arts, Film & Photography	965.62	NaN

Now to make the first steps of transforming all features to numerical values. Let's transform these two columns to only their numeric value.
First, we check if all of the values for example in the Ratings column follow the regex of "n out of 5 stars" or not. After getting True from this regex matching. We replaced both of the values in these columns with their numeric value.
Now let's focus on another feature which is Edition. This column contains data that can be separated into 3 features, EditionDate, EditionBinding, and EditionType.
After we created these three features we converted EditionDate to date type and now we can extract month and year as two new features. Now let's take a look at missing values:

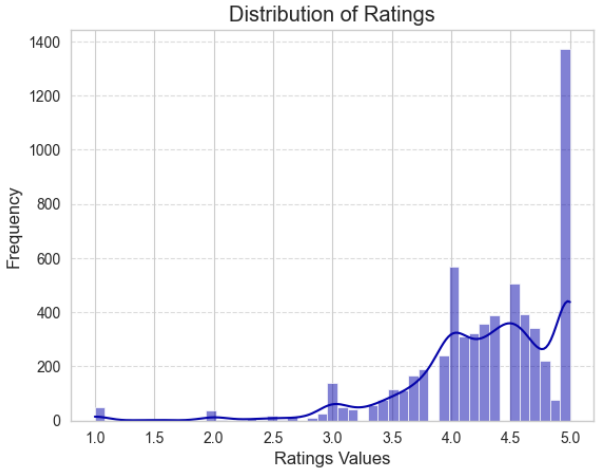
```
Title              0
Author             0
Edition            0
Ratings            0
Reviews            0
Synopsis           0
Genre              0
BookCategory       0
Price            537
Unnamed: 0      5699
EditionBinding     0
EditionType1       0
EditionType        0
EditionDate      452
Year             452
Month            452
dtype: int64
```

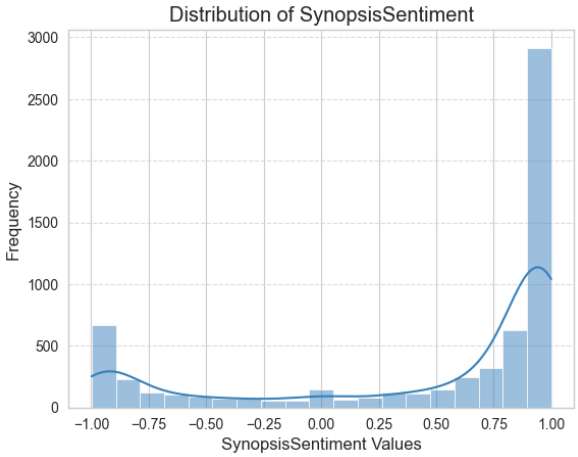
We can see month and year contain missing values. This indicated date data contained some missing information. For example some of the data samples we like 10 Mar or Feb 2010.
To reduce the number of missing values we can extract furthur data from these incomplete data. Meaning however we can't extract year data from 10 Mar but we can extract its Month.
After doing so we are still remained with some missing values which are less than before.

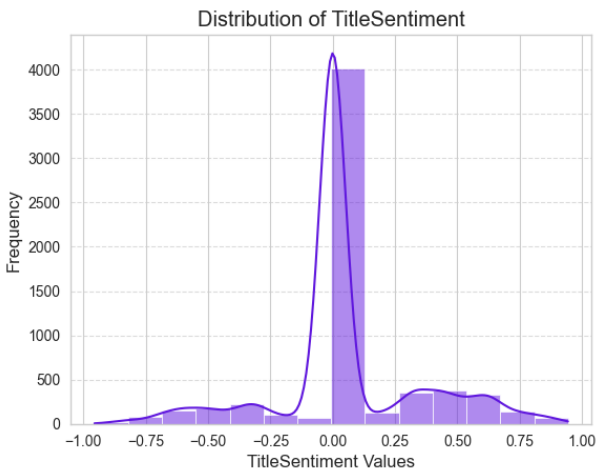
```
Title              0
Author             0
Edition            0
Ratings            0
Reviews            0
Synopsis           0
Genre              0
BookCategory       0
Price            537
Unnamed: 0      5699
EditionBinding     0
EditionType1       0
EditionType        0
EditionDate      452
Year             110
Month            369
dtype: int64
```

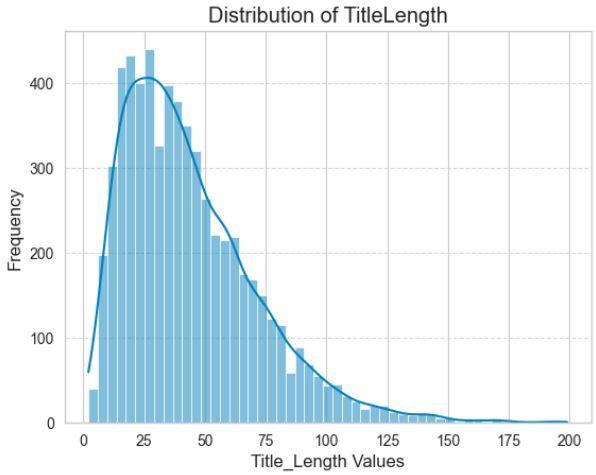
For these remaining data, we can simply use imputation methods. In this study, we imputed the mean value for the year column and the most frequent value for the month column.

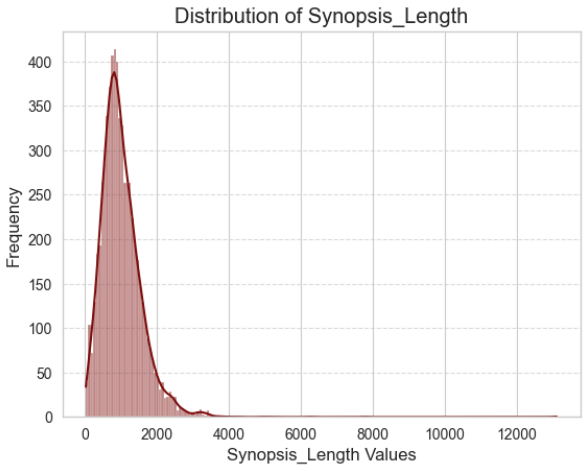
So now let's see if we can add some other features to this dataset.
Since there exist two Title and Synopsis columns. We can perform sentiment analysis on them and save their sentiment score as two new columns.
We used SentimentIntensityAnalyzer from nltk.sentiment for this means.
Also, we can add two features indicating Title_Length and Synopsis_Length.
There also can be another feature that determines if the author's name has initials in it or not. Determining by dot in their names.
Now let's investigate these features we created through visualization.
Let's first plot the distribution of our numerical features:

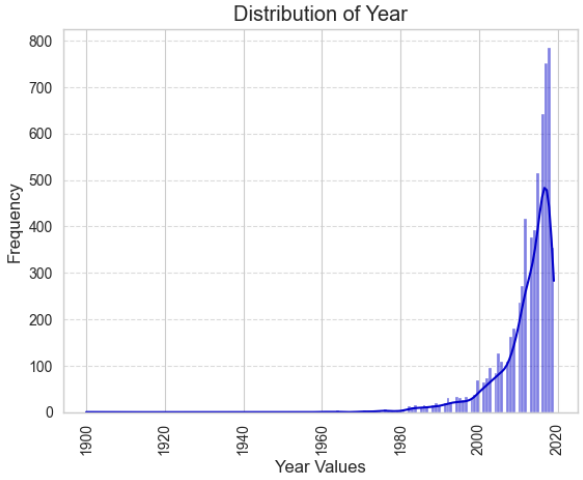


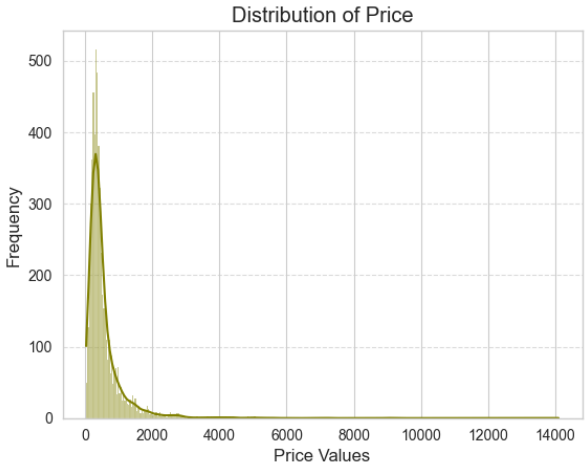


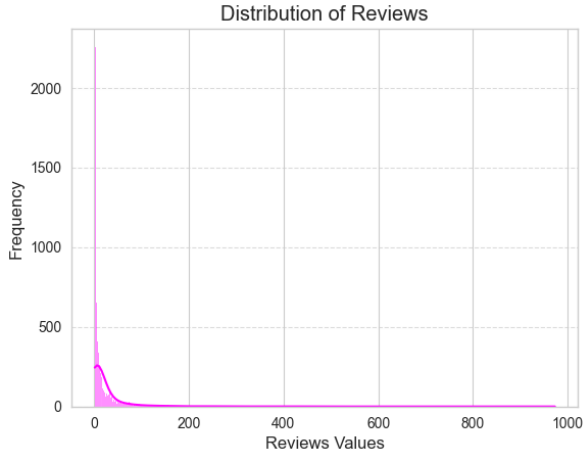




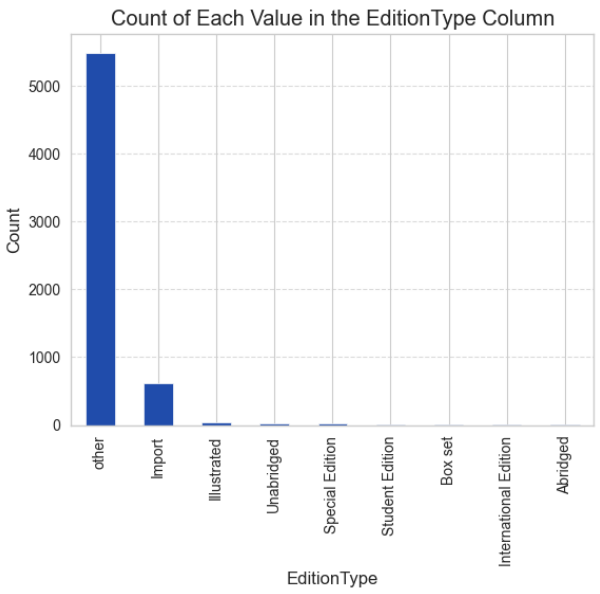


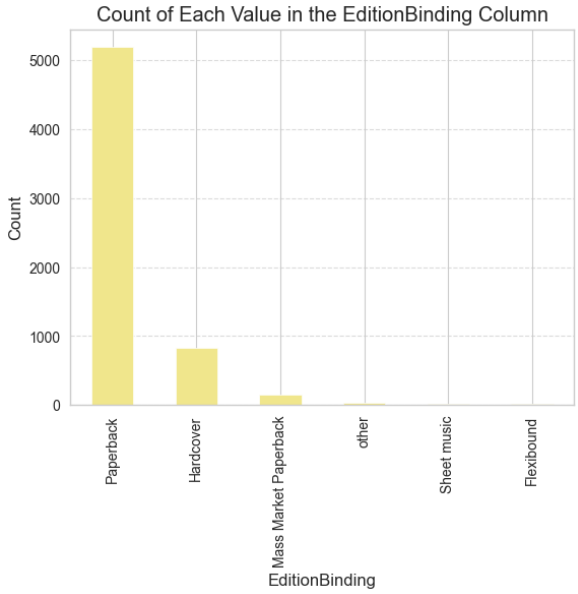


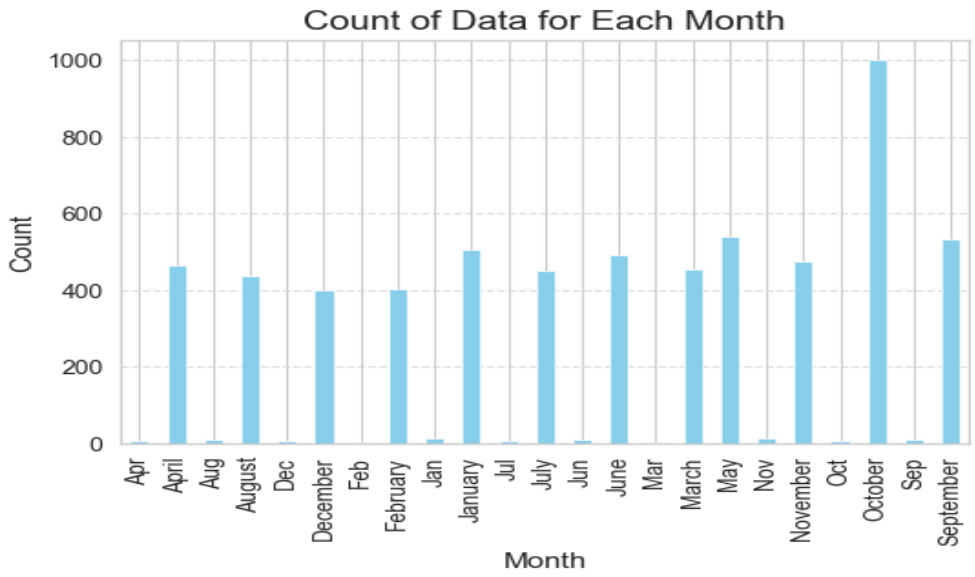




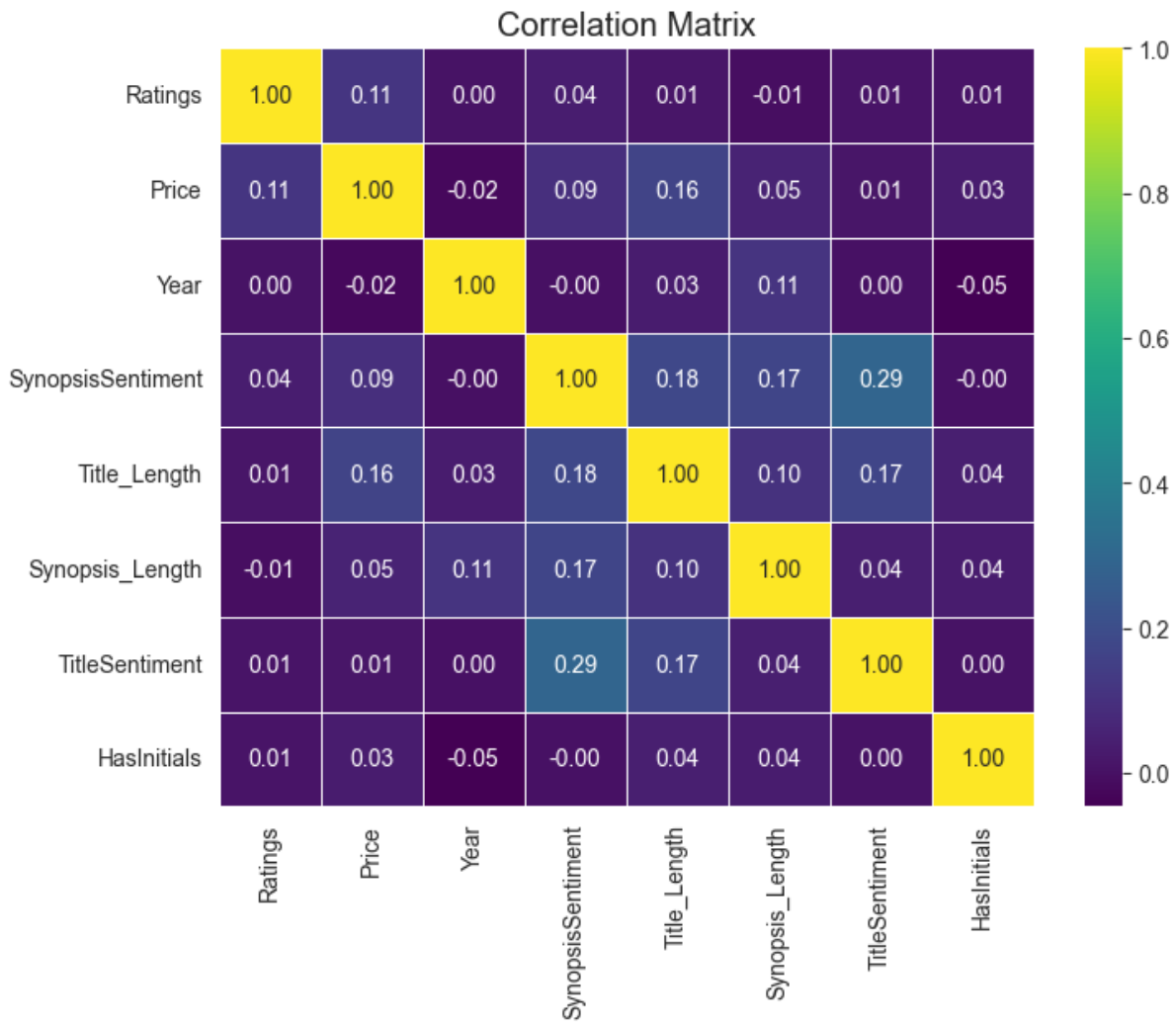
Now let’s plot count of some categorical values:







What we can understand from above plots:
titleLength, SynopsisLength, Price and reviews are left skewed, and distribution of year is right skewed, however titleSentiment also seems approximately normally distributed when SynopsisSentiment has a strong polarity on both sides.
Data distributed over months doesnt seems to differ alot, however most of the data is recorder on october.
Now that we've done some univariate data analysis let's examine the correlation of numerical data:



This plot indicated there is a strong positive correlation between title sentiment and synopsis sentiment which is logical. Also there is a strong correlation otherwise aswell, between title length and sentiment and synopsis's. Meaning as much the length of these two is longer, there is a higher chance the sentence has a positive sentiment.
Also the features which have the highest effect on price are title length and ratings.
Since random forest perform a feature selection whitin itself, there is no need for doing it ourselves.
Now the next step to getting closer to modeling this data is encoding the categorical data. In this study, one hot encoding was mainly used. For genre and book category first we get their dummies and set them as a new feature which can be either one or zero.
Also, cycle encoding was used for month column.
Now let's separate the test data we first added. We can do this by separating all the data which has a nan price.
Now with getting the shape of these two dataframes we get:
Train: (5699, 516)
Test: (537, 516)

Meaning there were 516 features left for us after encoding process
We also separated 20 percent of the train data for test and by running the provided function for getting the mse of this dataset through random forest. We obtained mse of:
Train mse is: 56264.211572963985 // Test mse is: 270643.60718845867
However by uploading the data on kaggle challenge and getting evaluated by the original test dataset. Result of mse : 512,115.4 was obtained.

Conclusion:

Train mse	Test mse	Kaggle mse
56264.211572963985	270643.60718845867	512,115.4