Project Documentation:

Data analysis of books placed on Flip.kz online store

Altemir Omar
ID: 210107119

Goal

The goal of this project is to gain insights about books on such online market as Flip kz. The project focuses on examining the proportion, distribution, and trends related to prices and publishing houses of the books within the dataset collected from Flip kz.

Steps

1. Data collection

First of all, I started analyzing the structure of the site by inspecting the source code of the page to find the tag that holds a link to the book page. After recognition of tags, I wrote code for taking all the links using “requests” and “BeautifulSoap” libraries and saved them to a separate list.

Further, I allocated features that will have a book in the data set and found all the tags on the book page that held all the information according to those features. Then using for loop I scraped each link in the list that was saved earlier. The data from each link I saved in the dictionary and appended it to the new list that kept information about each book.

Finally, after finishing the scraping process I transformed the list of books to the data frame and turned it into CSV file format to examine it later.

2. Data cleaning

Taking the previous CSV file for examination, I noticed that I scraped one wrong product for the data set. Among books, some kind of device was scraped, and added additional 40 features for each book with null values. To get rid of these unwanted columns I used “dropna” function with a threshold of 1000, to remove columns that contained 1000 or more null values. After

keeping the necessary information, the shape of the data frame decreased significantly.

At the filling of null values, I examined the data set for features that had null values. I divided them into two categories: numeric and object. Numeric features were filled with the mean values and the object features were filled with the most frequent value on the column. Then I replaced all values that were the same but written differently creating new values in the column.

3. Data visualization

Using “seaborn” and “matplotlib” libraries I plotted 8 figures according to the features of the data set.

3.1 Rating distribuiton.

In this visual set, there’s such a feature as rating. Rating range from 1 to 5. The majority of rating goes from 4 to 5 thus making the graph look exponential.

3.2 Scatterplot of price and rating

In the scatterplot, I noticed that the relationship could be called linear if we divide the graph diagonally we can see that all entries go to the upper half of the graph.

3.3 Count of books by binding type

Implementing a bar chart to see a distribution of bookbinding by count of each value. There are 5 unique values. But mostly distributed are only 2 types of binding among books, it’s hard and soft. Approaxmitaley, each takes half of the books

3.4 Correlation matrix

Numeric features were taken into consideration in this matrix. Basically, price is significantly influenced by its format. Thickness, width, height and number of pages.

3.5 Pairplot

Then to examine the price, I pair-plotted rating, review count, number of pages, and price. All the relations to the price showed exponential growth, except the number of pages which the graph slightly shows linear relation.

3.6 Count of books released each year

Starting from the earliest year to the present, the highest number of books were published last year. This year have one more month to take over the last month.

3.7 Average thickness of books by paper type

Books with the majority of paper types have an average thickness from 20mm to 25mm. Only books with laminated and matte paper have an average thickness of 15mm.

3.8 Publishing house by number of books

The main part of the market takes 2 publishing houses, which is “АСТ” and “Эксмо”, the approximate number for them is 50% and 30% respectively.

4. ML models

4.1 Regression models for price prediction

To predict continuous value as price I used two regression models: Linear regression and random forest regression. But before training the data, there is one step left - Data encoding. We need to convert categorical data to numerical, so our model can work properly. Using label encoding, I encoded all categorical features across the data set. Then I split the data to train and test with 0.2 proportion. After training our models I measured accuracy with the r2 score. Linear regression showed 0.54 score and Random forest showed 0.49. The accuracy leaves much to be desired. I used a sample to see our predicted value. Linear regression showed results for 2030tg for a book, but

random forest regression showed for 4648tg. I took the average value between the results of the two models and the average is 3339tg for the book.

4.2 Classification models for publisher house prediction

For classification purposes, I used the KNN classifier and Random forest classifier. Firstly, I changed the features and the target value of previously split columns to current columns. Then I did the same process to split and encode data. Secondly, I fit the train data into the models. With accuracy test, KNN showed 0.33 and Random forest 0.57 accuracy value. I tried to increase KNN accuracy by increasing number of k-neighbours, but accuracy stayed the same. Putting sample to predict value KNN model showed “ACT” publishing house and Random Forest showed “Попурри” pubishing house.

Conclusion

In pursuit of the project's goal to gain insights about world books within the online marketplace Flip kz, a comprehensive analysis was conducted, focusing on the distribution, and relation associated with book prices and publishing houses. The implementation of classification and regression models revealed results of moderate efficacy. While these models provided valuable insights, there is room for further refinement and exploration of advanced techniques.