Amazon Top 50 Bestselling Books 2009 - 2019

Ekans

Ali Hadi Altungök
181180760
altungokalihadi@gmail.com

Ufuk Bakan
181180011
println.ufukbakan@gmail.com

Karim Emenov
181180401
icymoon03@gmail.com

ABSTRACT

We will examine “Amazon’s Top 50 best selling books from 2009 to 2019" dataset[1]. 550 books, data has been categorized into fiction and non-fiction.

CCS Concepts

Computing methodologies→Machine learning→Learning paradigms→Supervised learning→Supervised learning by regression
High Relevance

Computing methodologies→Machine learning→Learning paradigms→Supervised learning→Supervised learning by classification
Medium Relevance

Keywords

amazon; top 50; best selling; book; from 2009-2019; author; user rating; reviews; fiction book; non fiction book; book genre; amazon book price; regression; classification; best book; most expensive; best cheapest book; most selling;

1. INTRODUCTION

There are 7 columns in this data set:

1. Name column contains book titles and it is String.
2. Author column contains Author name and they are string values.
3. User Rating column includes user’s votes out of 5 and they are floating point values.
4. Reviews column includes how many comments have been made and they are integer values.
5. Price column contains the prices of the books and they are integer values.
6. Year column contains the year the book was published.
7. Genre column contains the genre of the books and it can be fiction or non fiction.

There are 11 years and 50 books per year so the total number of rows is 550. And in the data there isn’t any missing value. In the data, the total number of reviews is 6574305.

2. THE APPROACH

After the initial analysis, visualizations were carried out to find out whether there is a relationship between some columns in the data and to see how the data distribution was. And then we wanted to predict some features. So firstly we examined and found no missing value in the dataset.

2.1. Distribution of Genre

There are more nonfiction books than fiction books in the dataset. 310 of all 550 books are non fiction.

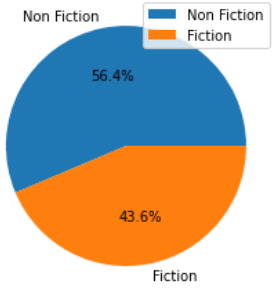

Figure 1. Pie Chart of Genre Distribution

2.2. Distribution of Genre per Year

Only in 2014 fiction books sold more than non fiction ones. Almost every year there are more non fiction books in the top 50 bestselling books list.

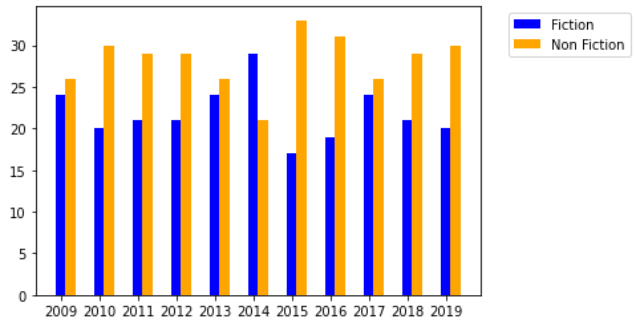

Figure 2. Pie Chart of Genre Distribution per Year

2.3. Correlation Between User Rating and Price

We have detected a negative correlation between user rating and price. To show that we drew a linear regression line:

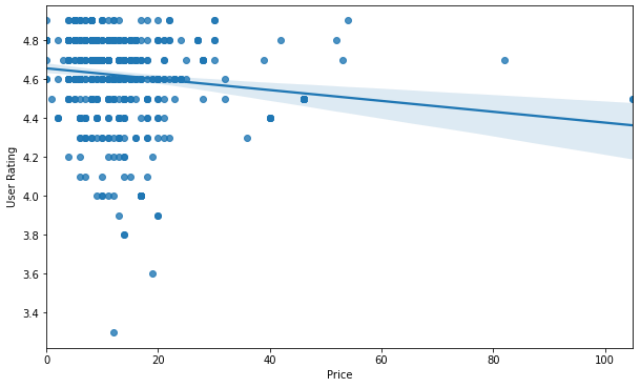

Figure 3. Correlation Between User Rating and Price

2.4 Predicting Features of a New Book

Then, we created our model with regression algorithms for some column values we determined. For example, to predict user rating; the regression was created using columns other than user rating. In the first stage of this, the data has been prepared by converting the string values into the integer values. As a result of the algorithm, we removed the Name column, which we thought would be ineffective or negatively affected. Then the result is visualized by linear and polynomial regressions.

2.4.1 Predicting User Rating

We used 5 of features to predict user rating but also we wanted to visualize prediction on a graph. Because of this, we converted 5 features into 1 PCA value and we separate data into train and test data so we can use PCA values from test data to make predictions.[2] Also we removed outliers where the price quantile below 5% and above 95%. We used 20% of data to train the models.

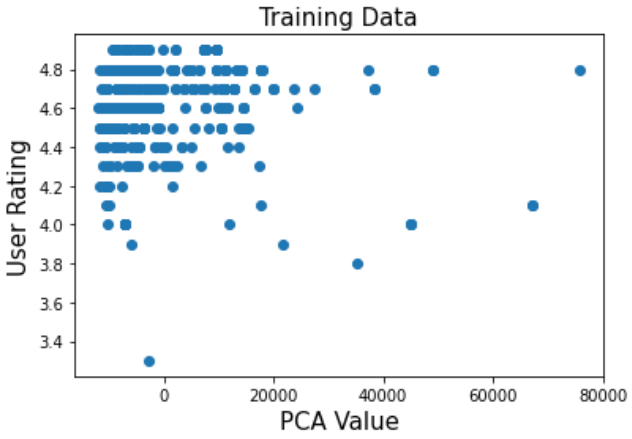

Figure 4. Training Data for User Rating

Then we predict test data by both a linear regression and a 3rd degree polynomial regression.

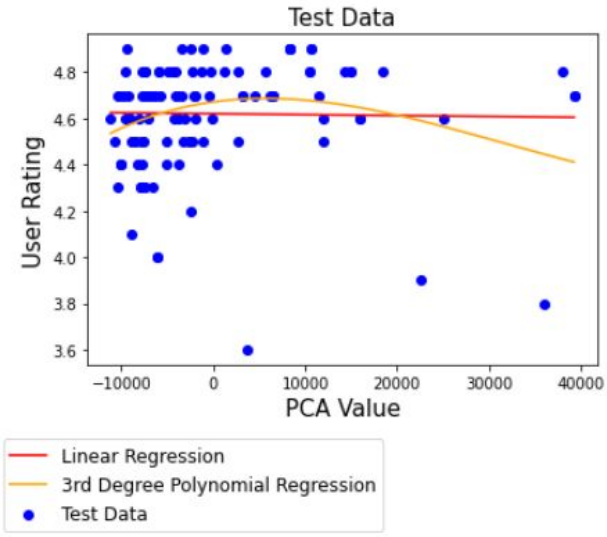

Figure 5. Test Data and Predictions for User Rating

2.4.2 Predicting Price

Again we used 5 features but this time User Rating included in training columns instead of Price. Also we increased the degree of polynomial regression to 6. We used 20% of data to train the models

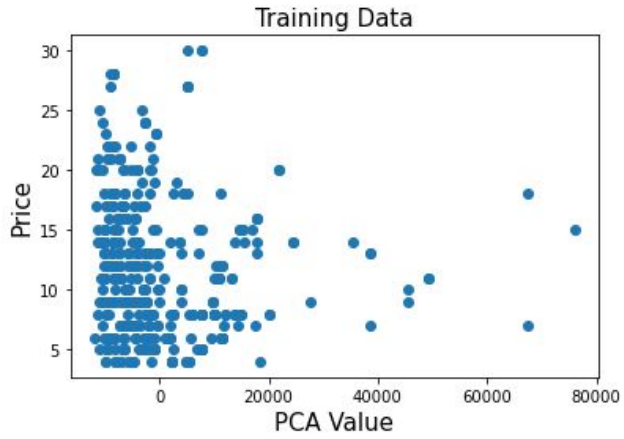

Figure 6. Training Data for Price

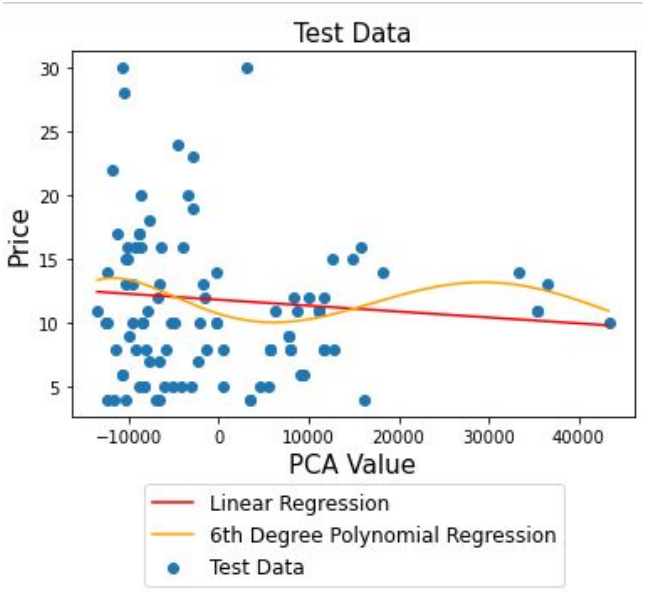

Figure 7. Test Data and Predictions for Price

2.4.3 Predicting Genre

We didn't drop the 'Name' column especially for this prediction. We thought that books with similar names might be of the same genre. We used k-nearest neighbors algorithm to classify and we chose 3 nearest neighbors.[3] There are 2 classes, one of them is fiction and the other one is non fiction. We shrinked 5 features into 2 as PCAX value and PCAY value to draw a 2D classification graph. Since we don't have a large dataset but still want to get acceptable results, we used 95% of data to train this model.

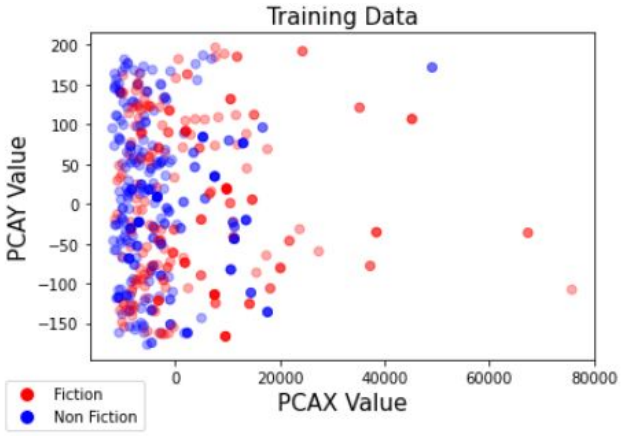

Figure 8. Training Data and Classes

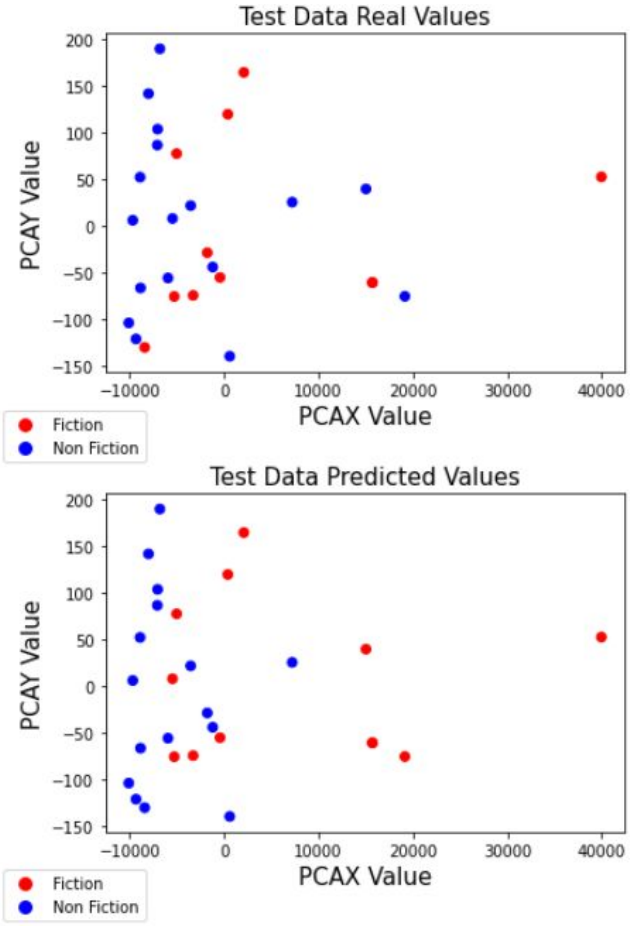

Figure 9. Test Data and Classification by KNN

As we can see in Figure 9 we predict the test data with KNN and we can see that the results are similar to the real points. Algorithm predicted truly 23 of 28 test data.

3. RESULT

As a result, we examined the structure of the data in this data set and examined the relationships between the columns. We showed that non fiction books sold more than fiction books and cheap books rated higher than expensive books. We analyzed some columns with regression algorithms and compared the results of polynomial regression and linear regression by visualizing them. Our data fitted accurately on polynomial regression at some points. We classified the data using the KNN classification algorithm according to the Genre column. We compared the forecast results with the real values and examined the match status.

4. CONCLUSION

It is not accurate to estimate by linear regression or polynomial regression between columns that do not have a direct relationship between them.

Although polynomial regression seems to be better than linear regression, it would be absurd to predict with algorithms in data that have such little data and do not have strictly separated proportions. Only statistical results that are easy to access like the most seller author or the most expensive book can be obtained from such data.

5. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

6. REFERENCES

[1] Sooter Saalu. Amazon Top 50 Bestselling Books 2009 - 2019. (October 2020). Retrieved January 19, 2021 from https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019bunu

[2] Usman Malik. 2019. Implementing PCA in Python with Scikit-Learn. Retrieved January 19, 2021 from https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/

[3] OnellHarrison.Machine Learning Basics with the K-Nearest Neighbors Algorithm Sep 10 2018 from

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761