# **Prediction of the Book Sales Volume**

**Team Non-Programmers**

Akanksha Gupta ([Gupta366@purdue.edu](mailto:Gupta366@purdue.edu))

Harpreet Kaur ([kaur66@purdue.edu](mailto:kaur66@purdue.edu))

Mohaneesh Khubchandani ([mkhubcha@purdue.edu](mailto:mkhubcha@purdue.edu))

Hee Sun (Sunny) Kim ([kim2397@purdue.edu](mailto:kim2397@purdue.edu))

Amer Nasrawi ([aalnasra@purdue.edu](mailto:aalnasra@purdue.edu))

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

## **Abstract**

There have been many changes in American book culture in terms of the number of readers and the format of the books. Books today can be accessed in a tangible – paperback or hardcover or an intangible – kindle or audible formats. According to Pew Research Center, a nonpartisan American think tank based in Washington, D.C., the number of book readers in the United States is decreasing, and American adults who had not read a single book account for nearly a 25% of the entire population. The decline in reading occurred regardless of the format of the books: print, digital, and audio. With Amazon leading the book market, by making it convenient for readers to access books through their online platform. Amazon, with its Kindle direct publishing has opened the industry to indie authors and fresh content. Because of the current market situation, Amazon is concerned about declining sales and would like to monitor the revenue as well as the platform market share being captured by the indie and uncategorized authors.

K**eywords:** #Amazon #Books #DailyUnitSales #Prediction #Platform #Dashboard #MLR

## **Business Problem**

Amazon has a huge database, representing books sold by it under various formats such as Audio, Kindle, paperback, hardcovers, with their respective sales prices, the type of publisher used to publish those books, the number of reviews received and the average rating. Also, the company would like to project its next month revenue and summarize its current platform market share. Amazon is concerned about declining sales and would like to monitor the revenue as well as the platform market share being captured by the indie and uncategorized authors. This would help the company monitor the market demand as well as measure a change in their estimated growth rate.

## **Analytics Problem**

Part 1 – Identify significant variables that impact the daily units sold for an individual book and predict the revenue value of a new addition to the platform.

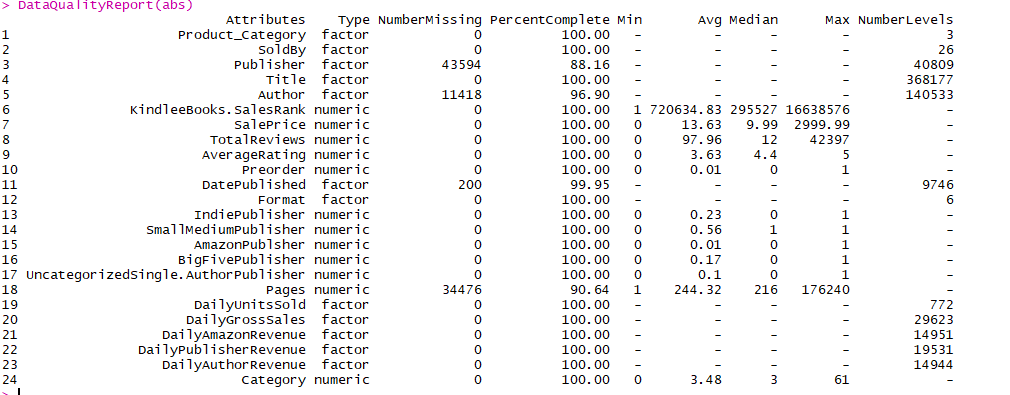
Part 2 – Summarize the current platform market share and predict the revenue for the next month.

**Data**

We referred to the February 2016 Author Earnings Report on Amazon’s e-book, Print, and Audio Sales for the data on sales and daily units sold per title.

<http://authorearnings.com/report/february-2016-author-earnings-report/>

* Variables/Attributes –

[Image - 1]

The data includes three types of books that are being sold in the market, namely, audio, printed and kindle e-books under the factor attribute: Format as well as Product Category.

* Attribute Descriptions

Product\_Category - Kindle, Audible or Print format of the book

SoldBy - Amazon or third party seller

Publisher - Character class variable specifying the publisher type

Title - Unique Title ID

Author - Unique Author ID

SalesRank - Rank based on sales revenue generated

SalePrice - Listed price in $

TotalReviews - Number of reviews

AverageRating - User rating 1 to 5

Preorder - Binart variable yes or no.

DatePublished - Published on Amazon platform

Format - Kindle, Audible, Paperback, hardback

IndiePublisher - Binary yes or no

SmallMediumPublisher - Binary yes or no

AmazonPublsher - Binary yes or no

BigFivePublisher - Binary yes or no

UncategorizedSingle.AuthorPublisher - Binary yes or no

Pages - Number of pages in the book

DailyUnitsSold - units of each book sold on an average in a day

DailyGrossSales - Total Revenue

DailyAmazonRevenue - Amazon Revenue share

DailyPublisherRevenue - Publisher revenue share

DailyAuthorRevenue - Author revenue share

Category - Number of book categories like fiction, non fiction, etc. the book falls under.

**Methodology Selection**

The problem at hand is a prediction problem where in we have to predict the daily units sold based on significant variables. Thus, discriminant analysis, logistic regression and classification trees which are majorly classification methods cannot be used.

Multiple Linear Regression or Artificial Neural Networks are two methods that can be used for the prediction analysis. In our analysis, the method that we chose to use was multiple linear step-wise regression using the adjusted r square method. Artificial Neural Networks was not tried at this point in time due to unfamiliarity of the method, though it might be a good analysis to see the functionality of the model developed using it.

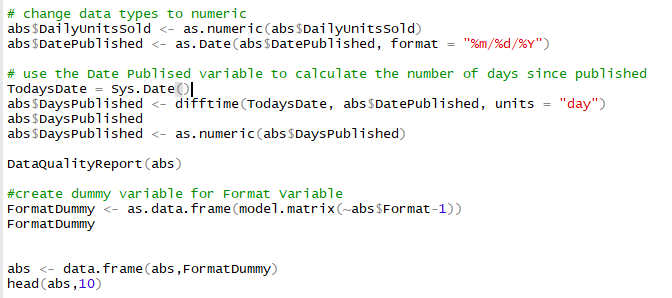
R proves to be a viable tool since other methods do not provide an effective platform to deal and interpret such a huge database. A dataset for approximately 300k entries cannot be manipulated using excel. R programming provides an easy and manageable way to develop a user-friendly application that can be used by any individual. Also, the processing of this big data is much faster in R programming as compared to in other analytical tools, we know.

## **Model Building**

A. Data Manipulation

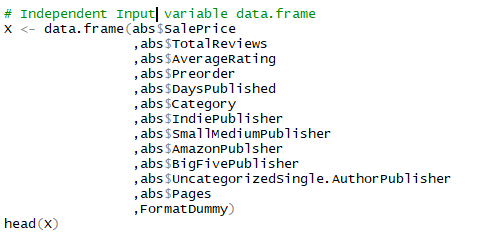
The [Image - 1] above shows the class of all the variables/attributes in the dataset. On analyzing the class of the variables, we make the following manipulations to conduct a multiple linear regression.

1. Change the datatype of Daily Units Sold from factor to numeric.
2. Use the date published variable to calculate the ‘Number of Days since published’ using the “difftime” function and save the variable as a numeric variable
3. Convert the character variable format to dummy variable.



[Image - 2]

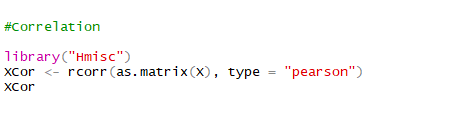
B. Regression Diagnostics



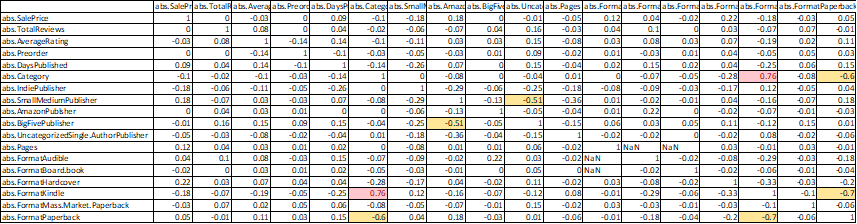
[Image - 3]

Based on the observation and logic, we can see that all input variables shown in [Image 3] are independent and the dataset represents all books on Amazon platform.

We check for multicollinearity and linearity of the variables by evaluating the correlations between them using the code shown in [Image - 4]

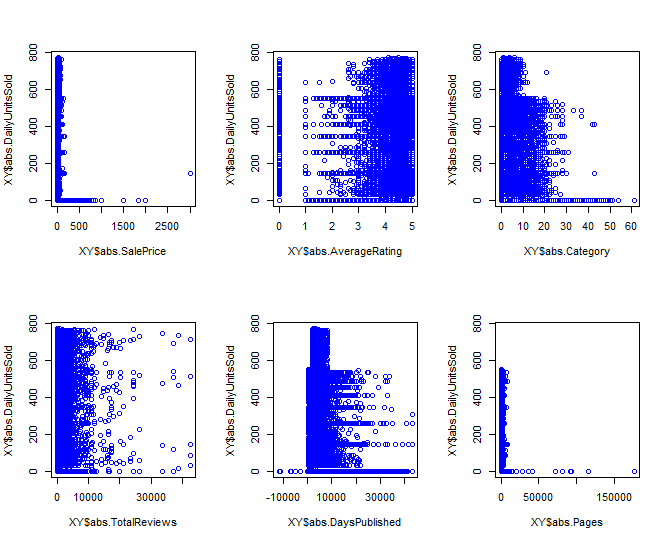


[Image - 4] Correlation Code



[Image - 5] Correlation Matrix

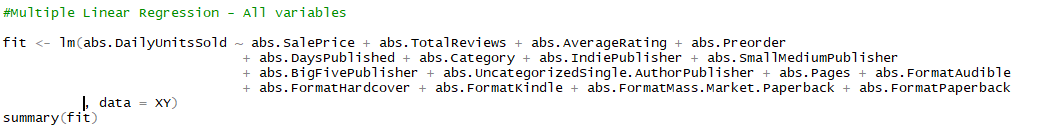
Next, we plotted all independent non-binary input variables against the target variable to do a preliminary check for constant variance or homogeneity in variance. We observed a fairly constant variance as shown in the [Image-6] below.



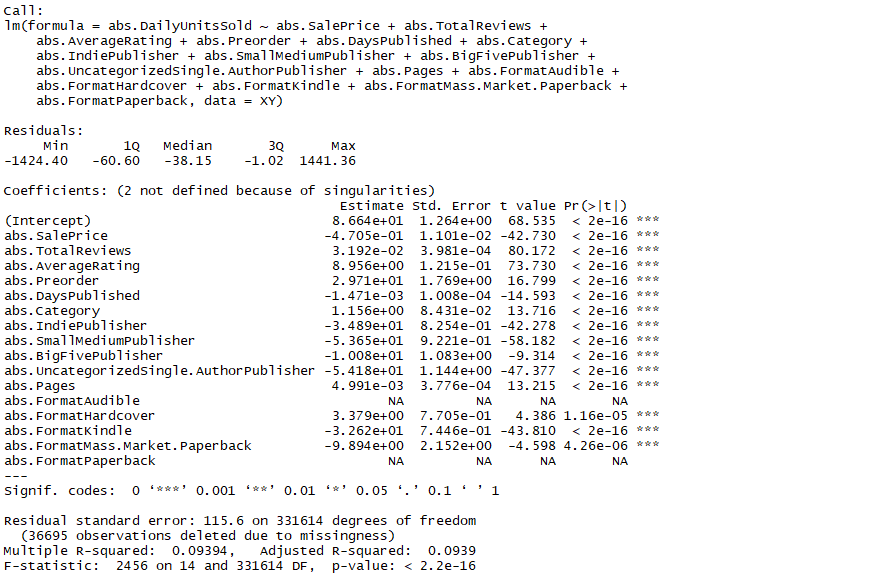
[Image - 6]

C. Multiple Linear Regression Modelling

We began by running a multiple regression with all variables in the model to identify the significant variables.



[Image - 7] Multiple Linear Regression all input variables



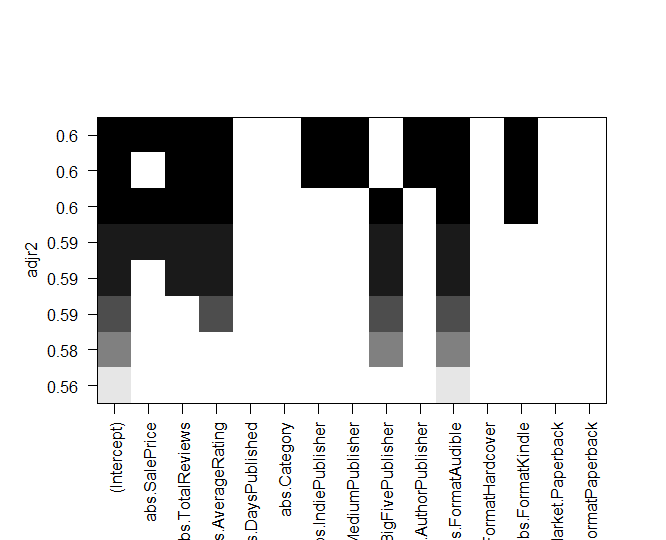
[Image - 8] MLR Output

We observe that though all variables are significant, the adj R-squared for the model is 0.0939 making the model extremely weak for predictions. We eliminate the highly correlated variables to avoid multicollinearity.

Then, we conducted a step-wise regression using the adjusted R square metric. We used the leap function [Image - 9] to conduct regressions by varying the input variables and plotted the Adj-R-Square shown below in the [Image - 10]



[Image - 9] Leap function to run multiple regression



[Image - 10] Adjusted R Squared Plot

We selected the significant variables giving the highest adjusted R squared for model to predict the daily units sold.

The Multiple Linear Regression Model Selected is -

[DailyUnitsSold] =

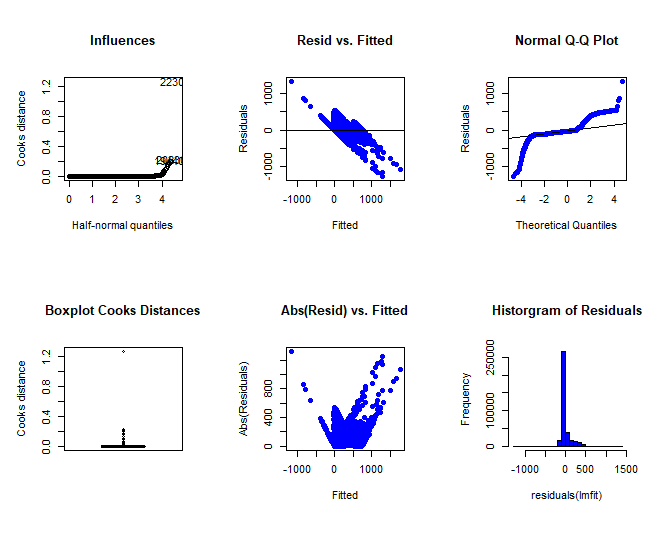
Intercept + [SalePrice] + [TotalReviews]+ [AverageRating] + [IndiePublisher] + [SmallMediumPublisher] +[UncategorizedSingle.AuthorPublisher] + [FormatAudible] + [abs.FormatKindle]

## **Functionality**

[Image - 11] Model Summary

As seen, the model selected explains approximately 60.12% of the variation in the dataset.

We further check the cooks distance for influencers, the residual vs fitted plots for Homoscedasticity along with the Q-Q plot and Residual histogram. [Image - 12] shows the plots.



[Image – 12]

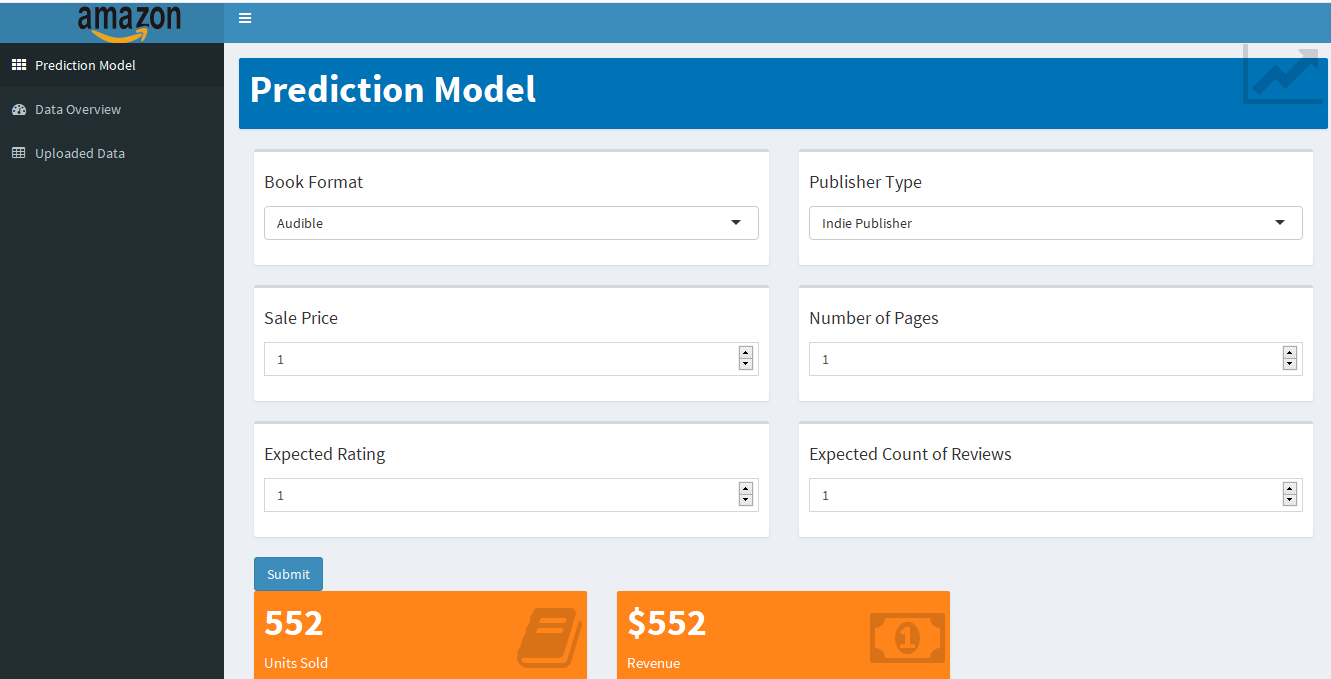
We observe the following,

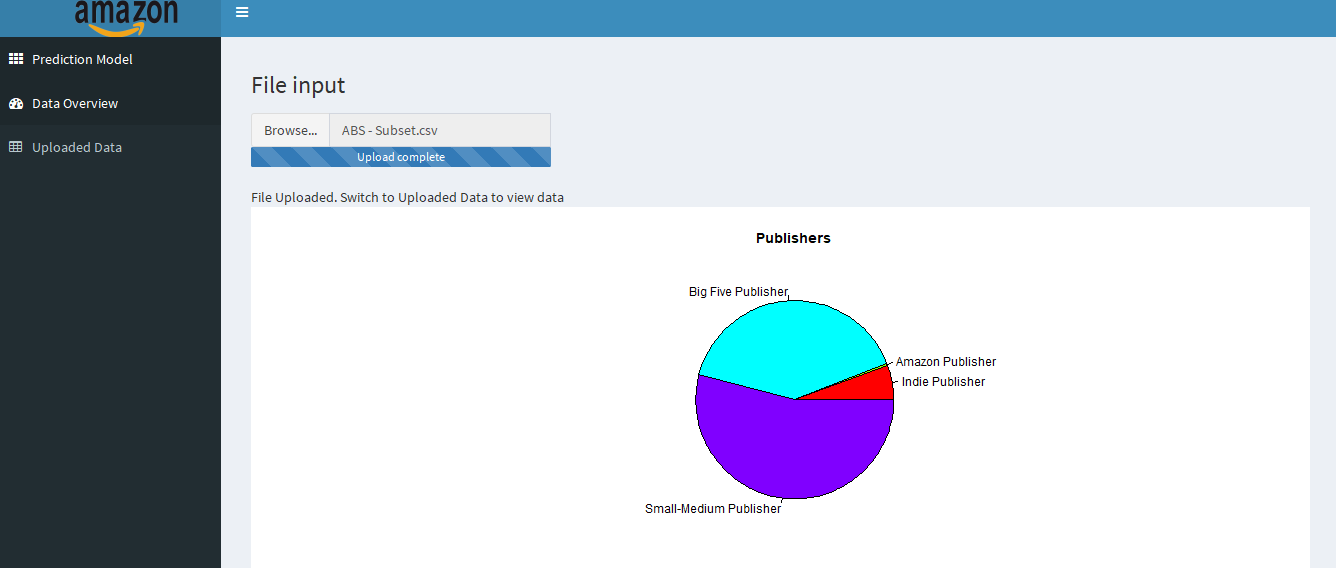
1. The cooks distance plots suggest at least one influential data point.
2. The residual vs fitted plot suggests heteroskedasticity. We can transform variables or increase the degree the of the model to try and address the issue.
3. The QQ plot and Histogram, show a left skewed distribution instead of normal. We could try to square root or log the input variables to try and achieve a normal distribution.

## **GUI Design and Functionality**

The GUI Design includes three components.

**Prediction Model**: This tab allows to key in the inputs for the attributes of a book. The values provided are taken as inputs for the regression model to calculate the expected units sold and the Expected Revenue from the sales.   
Currently, we’re unaware of Amazon’s margin for each unit sold. Realistically, this may vary on multiple factors like Genre of book, Selling price, Book format. Refer to image below to view the factors that are incorporated in the regression model.

  
  
**Dashboard**: The file upload control allows to upload the dataset that needs to be analyzed. After, a CSV file is uploaded, the dashboard shows a distribution of the Types of Publishers for various books available on Amazon. This dataset could be a monthly dataset or aggregated over months.  
Incrementally, it will enable periodic monitoring of trends and distributions.  
The distributions are updated as and when the new data set is uploaded.

  
  
**Uploaded Data:** This tab gives a quick snapshot of the uploaded raw data for the user to refer to view.



**Conclusions**

The model developed explains 60% of the variation. The heteroscedasticity can be rectified by attempting to transform the variables or weighted regression method. Moving ahead it would also be ideal to compare the functionality of the MLR model to the Artificial Neural Network Model. We would also like to gain data on more variables like the author popularity etc. Similarly, the current category variable represents the target market size in a weak form. We would like to evaluate it in a more granular manner, as an ordered multi-layer factor.

The shiny app tool developed is a great asset for the strategy team at Amazon. Once we receive the approval from the client regarding the usability of the data, we shall build additional functionalities to capture monthly data and help monitor a growth. Ideally, the app can evolve into a database interface, which would store data uploaded during each instance and refit the curve dynamically.

The individual unit daily sales projection can also be extended to act as a demand generation tool, allowing indie publishers and uncategorized authors to project their unit sales.

**Application Code base:** https://github.com/kaur66/Amazon-Books

**References (if any)**

[1] Jim Milliot, “[Pew Survey Shows Adult Reading in Decline](http://www.publishersweekly.com/pw/by-topic/industry-news/bookselling/article/68478-pew-survey-shows-adult-reading-in-decline.html),” *Publishers Weekly*, October 23, 2016.