Predicting the next bestseller, a predictive model on sales of books

Anuashka Sharma[1] and Prateek Khare[2]
[1]Management Information System, Oklahoma State University, Stillwater, Oklahoma 74075
[2]Management Information System, Oklahoma State University, Stillwater, Oklahoma 74075

Introduction

Despite attrition in Number of Book Readers over the past few years, there are many books which managed to remain unaffected by it. The authors and publishers thus may find it helpful to analyze the factors that mainly affect the sale of any book. This kind of analysis would help them in strategizing their publishing and promotion of any book.
We have used JMP ® to analyze the data and build predictive model to predict the sales of books. We have used the sales data of bestselling books of last three years along with the different attributes of the books which seemed to be affecting the sales. The dataset was prepared using the newspaper and magazines report of 100 bestselling books for each of last three years. Since the data had too many nominal variables, we had to create dummy variables so that they could be used in analysis. Using this data we have built predictive model that predict the sales of books. Our target variable is the total volume of books sold. The explanatory variables used are: Price of the Book , Audience, Series or Individual title, Binding Type, Month of Publication, Genre Related variable, Number of Pages.

Methods

Data Preparation and Initial Analysis:

To have a better understanding of the data, we did some initial distribution analysis across different categorical variables.

- Multivariate Analysis to find out instances of multicollinearity (Fig. 1)
- ANOVA Testing to compare the volume of books sold month wise (Fig. 2).
- Comparing the distribution of Volume of books sold over different categorical variables (Fig. 3,4 and 5)

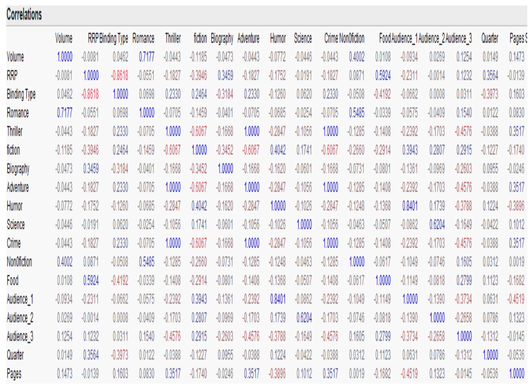


Fig. 1 Multivariate analysis for colinearity

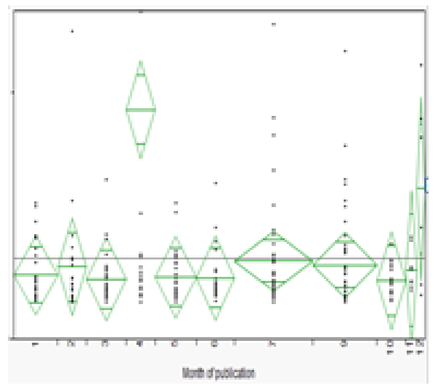


Fig. 2 ANOVA Analysis of Volume by Month of Publication

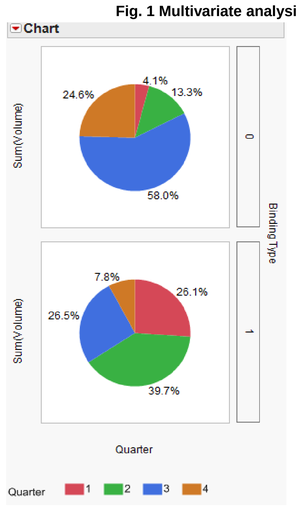


Fig. 3 Volume by quarter, grouped by binding type

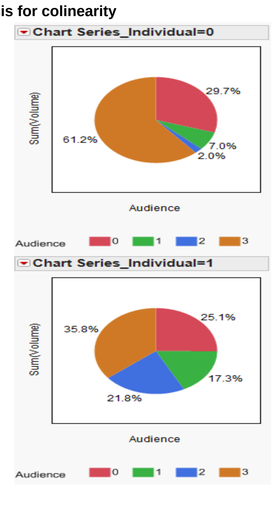


Fig. 4 Volume by Audience Type, grouped by Series/ Individual

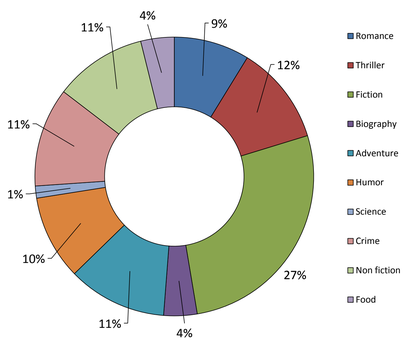


Fig. 5 Volume by Genre

Profiling and Predictive Modeling

For predictive modeling, we built different models taking 'Volume' as the target variable. We used the R-squared and Adj R- Squared values (which indicate the percent of variation in target variable explained by the model) as performance metrics to compare and select the bets models from following:.

- Decision Tree model
- Neural Network model with 3 hidden nodes
- Multiple regression model

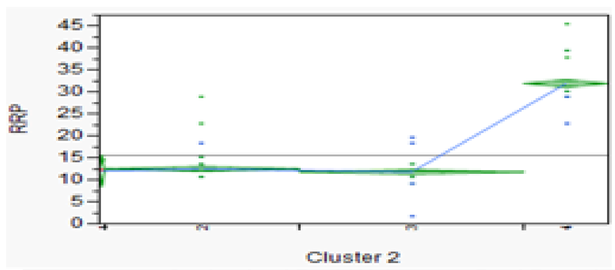
Apart from building predictive models, we also performed clustering of books forming groups of books similar in some fashion. Four clusters were formed which differ from each other in terms of the attributes. The clusters and their description are shown further in the poster.

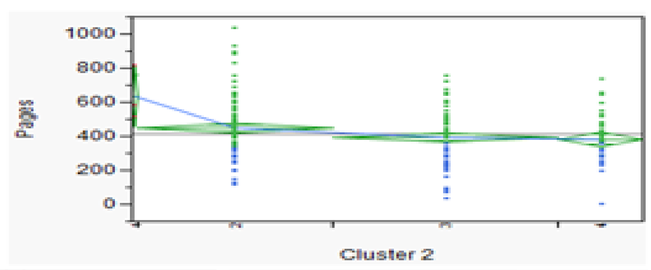
Results

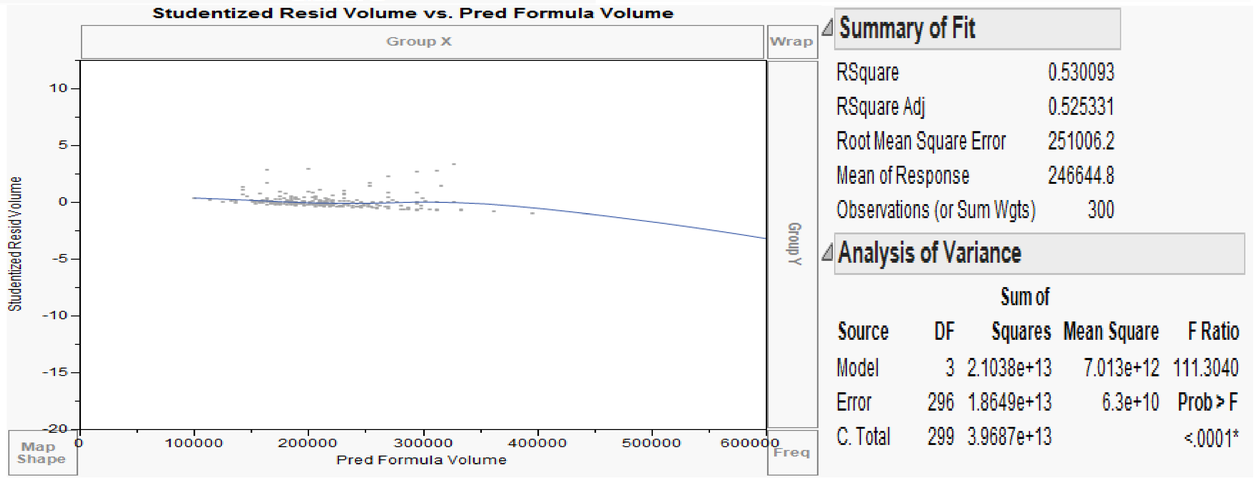
On comparison of the built models, we found Multiple Regression model to be the best of all built models. The regression model represents target variable in terms of an equation containing the significantly important independent variables. The resulting predictive model is represented as:

Volume = 1131980.21 – 998595.47 * Romance + 221.4462 * Pages – 33426.73* Series_Individual

- Through the model above we can infer that the predicted sales of bestseller books which are part of a series tend to be lower than those which are not.
- Romantic genre seems to decrease the sales of bestseller books.
- To test our assumptions in regression model, we plot the studentized residual volume against the predicted volume. There is a pattern in the plot that suggest presence of non-constant variance that may be handled via transformation of the target variable. However, such transformation will make the results less understandable from a managerial standpoint and therefore we chose not to do it in this poster.







Cluster 1	Books with highest number of pages, published in Q2, Main audience are children
Cluster 2	Paperback books published in Q1, Main audience are children and youth
Cluster 3	Main audience are adult, most of the books are part of a series
Cluster 4	Most expensive books, least number of pages published in Q4

Discussion

- Clustering of books helped in understanding the reading pattern and preferences of readers.
- The model suggests that the most important variables highly affecting the sales of any book are the genre 'Romance', Number of pages and if the book is part of a series or an individual one where Series_Individual = 1 indicates that the book is part of a series and 0 indicates book being an individual one.
- The model explains reasonably variation in the target variable (Adj R-Squared value = .5253)
- Our analysis was however limited by the small volume of available data. By increasing the number of observations as well as attributes, better results can be obtained.
- This concept can be extended to predict sales of products of same category that differ in some other attributes which would prove to be a great asset to the retail industry and marketers in formulating their important business and marketing strategies.

Reference

- Sharda and Delen, 2005, Predicting Box office success of movies with neural networks
- Marion and Walker, 1978, Short Run predictive models for retail meat sales

Acknowledgement

- Dr Goutam Chakraborthy**,** Professor, Department of Marketing, Oklahoma State University
- Gaurav Pathak, Management Information System, Oklahoma State University