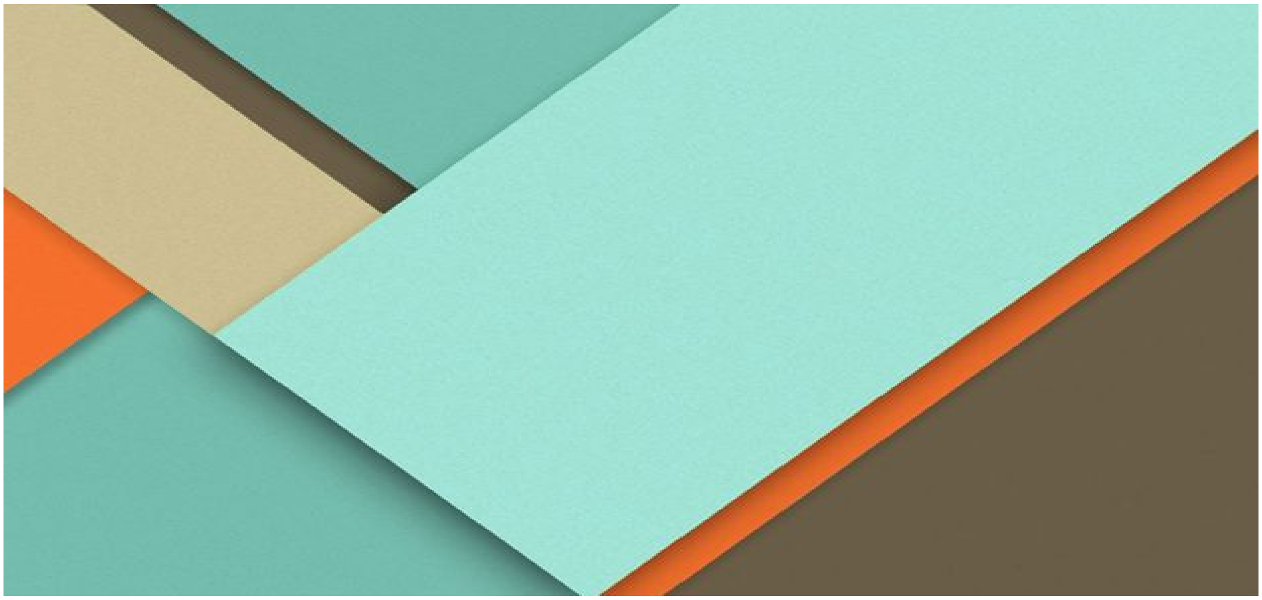


BAN 620 Data Mining Project Report

Analysis and prediction of the characteristics of Amazon top 50 bestselling books 2009 - 2019

Group 5
Abhishek Yadav
Devi Nadimpally
Priyanka Shah
Sayali Peshwe
Uttara Dabbiru
Victor Antony

1) Introduction

Amazon is known as the world's largest online marketplace and Amazon's bestsellers list is a good indicator of how well the product is selling overall. Books are one of the most popular products sold on Amazon. Recently Amazon released the list of bestseller books for the time period 2009-2019. The intent of this study is to analyze and predict the characteristics of Amazon's bestsellers books based on this historical data.

2) Motivation

Reading is one of the common recreational activities. But with modern modes of entertainment developing over the last few decades, the traditional forms of recreation like book reading seem to be on a decline. This could perhaps impact the publishers/authors. Millions of books are published in a year, yet just few are read widely. The publisher's market is hence an unpredictable one. These facts encouraged us to look into the Amazon's top best-selling books dataset (2009-2019). In our analysis we are primarily trying to analyze the top 50 best selling books of Genres, Fiction and Nonfiction, from the years 2009 to 2019, and predict some of the characteristics of these books. We intend to analyze the factors contributing towards making a book a bestseller. We wish to see if one or all of these factors influence each other, and if so how can we use them to optimize or maximize the other. The proposed analysis can be helpful to gain insights into the effectiveness of the marketing campaigns of the book sellers/authors and also speculate the book's overall success in terms of reviews, ratings etc.

3) Objective

The main objective of this study is to predict the attributes of the Bestsellers of Amazon for the next year. This study is conducted to predict for the year 2020. Although we are at the end of 2020 this study can be extended for the year 2021. The main objectives of this study are as follows:

1. **Predict the minimum number of bestseller Fiction or Non-Fiction books that would be bought by the users** - The column 'Review' is the count of unique users' reviews for the book. Since each count is unique to one user who has bought the book and provided his/her reviews, we believed this count translates to the minimum number of books sold or bought by users. **Note**: This does not indicate the sales of the book. We intend to analyse the factors that affect Reviews and predict it. When we combine the number of books sold with the Price of the book, we can predict the profits too. As such achieving a higher accuracy is important here. As such, we intend to use the KNN and the Classification tree algorithm to compare and arrive at the best prediction accuracy. If we encounter a lesser accuracy because of the lower number of records in the dataset, we can perhaps improve it by using the Cross-validation, Boosted Trees and Random forests techniques.
2. **Predict the average Price of a Fiction or Non-Fiction book** - The idea is to help gain an insight on what the average cost of a bestseller Fiction or NonFiction book would be for the next year. With the information gained from 1) the predicted Price information will help the publisher/seller base their price to maximize their profits for future. We intend to use Multiple Linear Regression technique here with input variables as User Rating, Reviews, Genre and Author.
3. **Predict the more popular Genre for the future year**- The books are primarily classified as Fiction or Nonfiction. We intend to predict the more popular genre using the logistic regression algorithm with predictors such as Author, User Rating, Reviews and Price of

previous years' data. With this information new Authors/Publishers will have an idea on the user's interest and create a book for a particular Genre.

4) Out of Scope

1. The project is not intended to prescribe what types of titles/names should be used for the books sold by a seller/publisher.
2. The Bestsellers is based on Amazon.com US. Other geographies are not part of this analysis.

5) Data Source

The data was taken from www.kaggle.com. Name of the Data set **-** Amazon Top 50 Bestselling Books 2009 - 2019

Link to data Source:
https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019

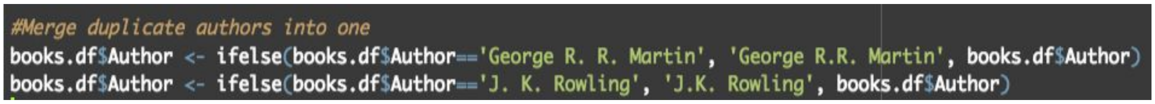
6) Variable Description

The data set contains a total of 550 rows, where each row corresponds to a book, written by various authors from the timeframe 2009-2019 along with price. There are Reviews and User Ratings given by users to the books.

Name of variable	**Description**	**DataType**
Name	Name/Title of the book	Categorical
Author	Author of the book	Categorical
User Rating	Amazon user Rating for the book(not author or seller)	Categorical
Reviews	Number of written reviews for book on amazon	Continuous
Price	Price of the book	Continuous
Year	The year when the book is ranked as bestseller	Categorical
Genre	Whether book is Fiction or Nonfiction	Categorical

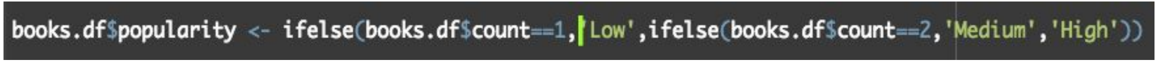
7) Preprocessing

1. Our initial observation showed there were duplicate authors because of spacing issues. For example 'George R. R. Martin' and 'George R.R. Martin'. They were combined to one value



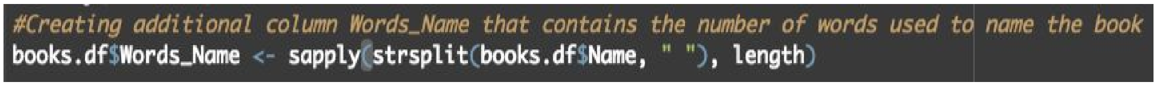
```
#Merge duplicate authors into one
books.df$Author <- ifelse(books.df$Author=='George R. R. Martin', 'George R.R. Martin', books.df$Author)
books.df$Author <- ifelse(books.df$Author=='J. K. Rowling', 'J.K. Rowling', books.df$Author)
```

2. ‘Author’ column was a categorical variable. Instead of omitting it, we wanted to make use of it as Author in reality essentially provides life to a book. We converted the column ‘Author’ into another column called ‘popularity’, that contained three levels, ‘Low’, ‘Medium’ and ‘High’, based on the number of times an author appeared in the best sellers list. The counts were : Low=1, Medium=2 and High>=3



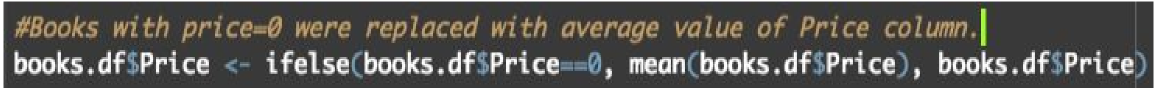
```
books.df$popularity <- ifelse(books.df$count==1,'Low',ifelse(books.df$count==2,'Medium','High'))
```

3. We believed the number of words chosen for the Name or Title of the book has an impact on the reader. Therefore we replaced the column ‘Name’ with another column called ‘Words_Name’, which contained the number of words used for the Name of the book.



```
#Creating additional column Words_Name that contains the number of words used to name the book
books.df$Words_Name <- sapply(strsplit(books.df$Name, " "), length)
```

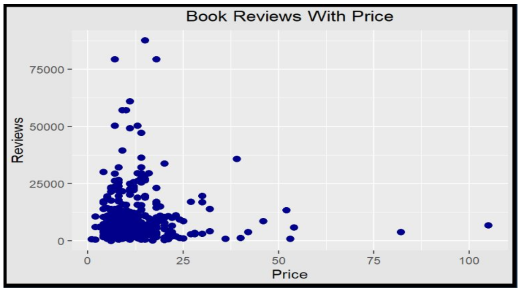
4. There were some books with Price as ‘0’. We replaced the values with 0 in the ‘Price’ column with the average price of the book.

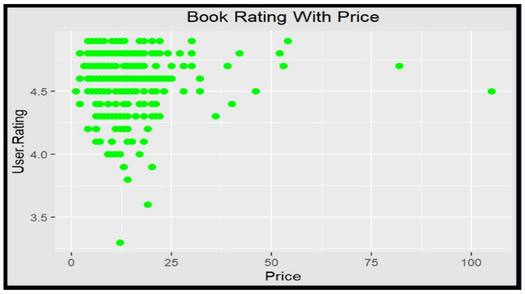


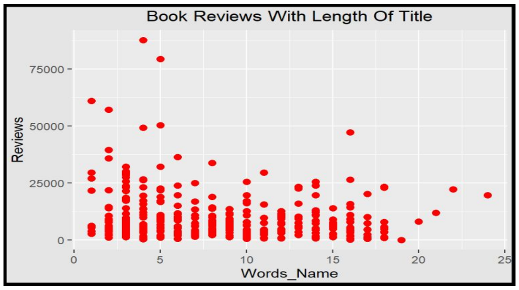
```
#Books with price=0 were replaced with average value of Price column.
books.df$Price <- ifelse(books.df$Price==0, mean(books.df$Price), books.df$Price)
```

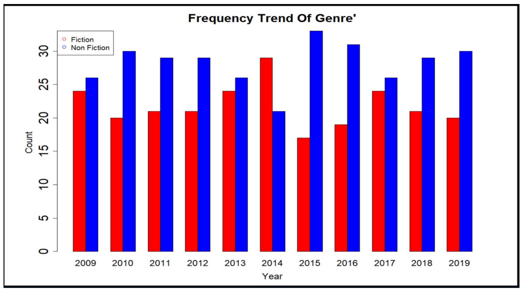
5. The column ‘Review’ is the count of unique users’ reviews for the book. Since each count is unique to one user who has bought the book and provided his/her reviews, we believed this count translates to the minimum number of books sold or bought by users. **Note**: This does not indicate the sales of the book.

8) Initial Analysis





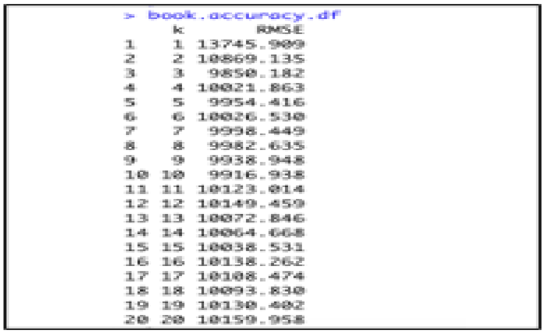




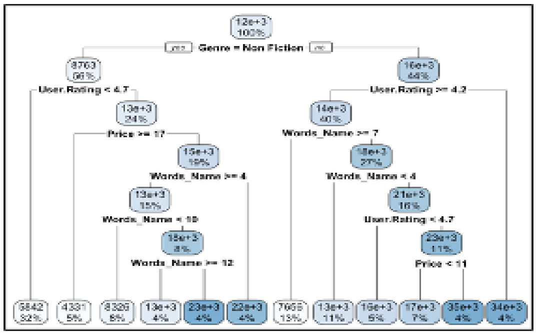
1. Books with most reviews price less than $25.
2. Books priced more than $40 received fewer reviews.
3. Most of the books with a rating greater than 4.0 are priced within $25.
4. Books with a title less than 30 , words received more reviews.
5. The number of reviews is decreasing significantly for bigger titles.
6. Non- fiction books are more popular than fiction books.
7. Fiction books were more popular in the year 2014.

9) Analysis and Prediction for Reviews(Number of books sold)

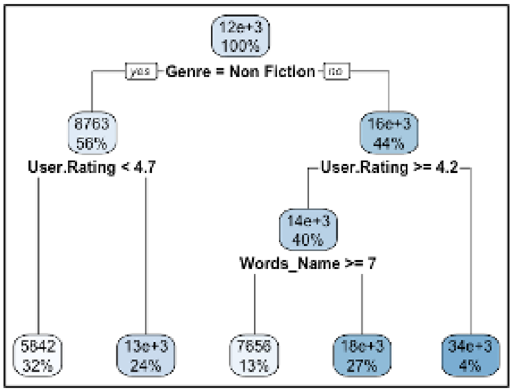
Knn Algorithm gave highest accuracy for k = 3 (Lowest RMSE for k=3). In order to achieve an improved accuracy Regression tree model was used for prediction of reviews.



```
> book.accuracy.df
    k      RMSE
1   1 13745.909
2   2 10869.135
3   3  9850.182
4   4 10021.863
5   5  9954.416
6   6 10026.530
7   7  9998.449
8   8  9982.635
9   9  9938.948
10 10  9916.938
11 11 10123.014
12 12 10149.459
13 13 10072.846
14 14 10064.668
15 15 10038.531
16 16 10138.262
17 17 10108.474
18 18 10093.830
19 19 10130.402
20 20 10159.958
```



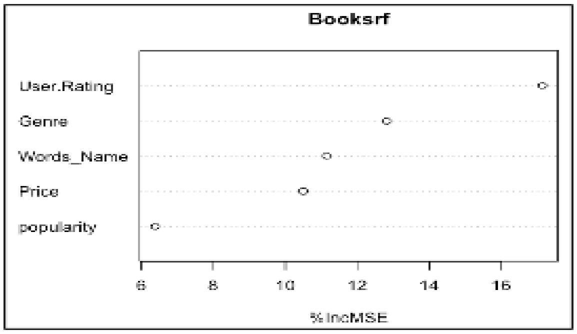
1. We considered the following variables for predicting no of Reviews: User ratings, Price, Genre (Fiction/Non-fiction), word count in title and Popularity
2. Lowest RMSE (9850.182) is achieved for k=3, hence we choose k=3 for our knn model.
3. The prediction of average no. of Reviews obtained, for the test set is 13283.24
   For this model we obtain an RMSE= 13933.44.
4. In order to achieve a better accuracy than the kNN model ,we used a Regression tree model for predicting the average no. of reviews and also to observe some user rules.
5. The default tree rendered an RMSE = 10036.91, which was lower than the RMSE for kNN Model.
6. To achieve an improved predictive accuracy, we further attempted predictions using the Decision (Regression) tree. The tree was optimized using techniques like cross validation & Random Forest.Pruning the default tree for lowest cp, we obtained a pruned tree having length = 5 and RMSE= 9668.707 (lower than the default tree).



Following rules can be interpreted from the tree above:

1. If the Genre is non-Fiction, and the user rating is less than 4.7, the book will receive 5800 reviews on an average. And if the rating is more than 4.7, the book will receive around 13,000 reviews.
2. If the genre is Fiction, and user rating is greater than 4.2, and Words in the title are greater than or equal to 7, the book is likely to revive around 7500 reviews whereas, if the word count is less than 7, the book is likely to receive approx. 18,000 reviews

3. Lowest RMSE = 8457.916, was obtained for Random forest. Based on the Random Forest variable importance plot, the User Rating, Genre and word count in title (word_name) turned out to be main influential variables for predicting Reviews.



Prediction result for the test set:

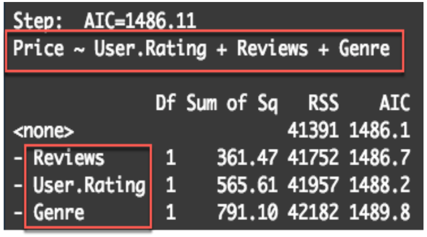
Prediction results on test set:	**RMSE**
Random Forest	**11249.4**

1. As the best RMSE was obtained for Random Forest, we selected the Random forest as the final model to predict the number of Reviews.
2. Popularity of the author does not seem to have much influence on the number of reviews
3. The number of words in a title can possibly influence the number of reviews was an interesting finding in this particular analysis. And as per our assumption, no. of Reviews translates to no. of books sold, the Genre and Word_name turn out to substantially influence the volume of books sold.

10) Analysis and Prediction for Price

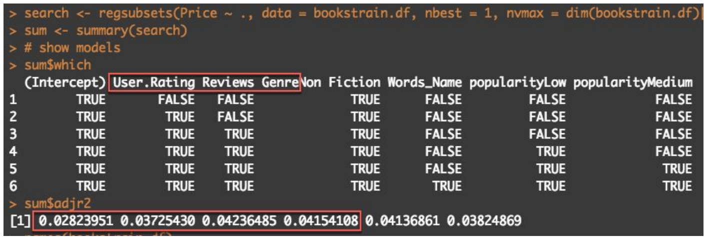
Multiple Linear Regression algorithm was used to predict the Price variable. The following observations were noticed :

1. In the first run of the algorithm we noticed User.Rating and Genre are strong predictors of price based on the p values.
2. We further ran the Backward Search model and Exhaustive Search model which showed the main predictors were User.Rating, Genre and Reviews.



```
Step:  AIC=1486.11
Price ~ User.Rating + Reviews + Genre

              Df Sum of Sq   RSS    AIC
<none>                     41391 1486.1
- Reviews      1    361.47 41752 1486.7
- User.Rating  1    565.61 41957 1488.2
- Genre        1    791.10 42182 1489.8
```



```
> search <- regsubsets(Price ~ ., data = bookstrain.df, nbest = 1, nvmax = dim(bookstrain.df)
> sum <- summary(search)
> # show models
> sum$which
  (Intercept) User.Rating Reviews GenreNon Fiction Words_Name popularityLow popularityMedium
1        TRUE       FALSE   FALSE             TRUE      FALSE         FALSE            FALSE
2        TRUE        TRUE   FALSE             TRUE      FALSE         FALSE            FALSE
3        TRUE        TRUE    TRUE             TRUE      FALSE         FALSE            FALSE
4        TRUE        TRUE    TRUE             TRUE      FALSE          TRUE            FALSE
5        TRUE        TRUE    TRUE             TRUE      FALSE          TRUE             TRUE
6        TRUE        TRUE    TRUE             TRUE       TRUE          TRUE             TRUE
> sum$adjr2
[1] 0.02823951 0.03725430 0.04236485 0.04154108 0.04136861 0.03824869
```

Final Model : (lm(Price ~ User.Rating, Genre, Reviews), data=bookstrain.df)
Our final model included User.Rating, Genre and Reviews as the predictors to predict Price.

Observations

1. Best variables suggested : User.Rating , Reviews & Genre
2. Accuracy was slightly less : RMSE: 8.65, MPE: -49.85

Train Data	Validation Data
Significant variables(at 95% Confidence): User.Rating, Genre Residual standard error: 11.85 Adjusted R-squared: 0.04	RMSE : 8.6 MPE: -49.8

Prediction Results(on Test Data)

- Accuracy : RMSE: 8.65, MPE: **-62.82**
- Mean Predicted price : **$14.06**

11) Analysis and Prediction for Genre

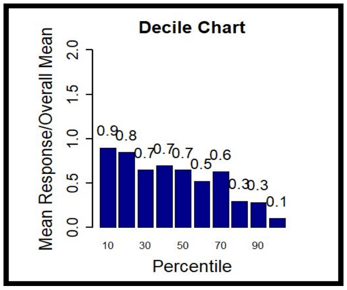
To predict the most popular Genre for future years, we trained our model using two algorithms such as Logit and CART. We compared the accuracy of models on validation data and ran the finalized model on test data.

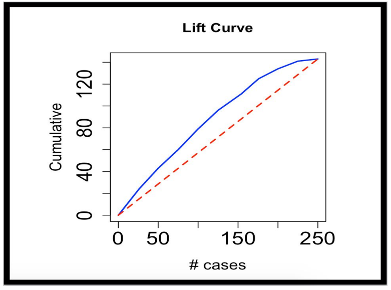
Logistic Regression

Logistic regression is very much like **linear regression**. Predictors are combined linearly using coefficient values to predict outcome variables using maximum-likelihood estimation.

For this dataset, The Logistic regression is used to explain the relationship between independent variables and one dependent categorical variable(Fiction and NonFiction).

1. Considering the variables User.Rating, Reviews, Price, Words_Name, and Popularity. The Accuracy is around 0.7, with low sensitivity and specificity.
2. **Lift Chart**: A lift chart graphically represents the improvement that a model provides when compared against a random guess
3. **Decile Chart:** Decile analysis is created to test the model's ability to predict the intended outcome.
4. We tried to assess the decile chart and also the lift chart to see how well the model can predict the response and observed there is no consistent decline in the number of cases predicted.
5. In both cases, it says the model is not performing well





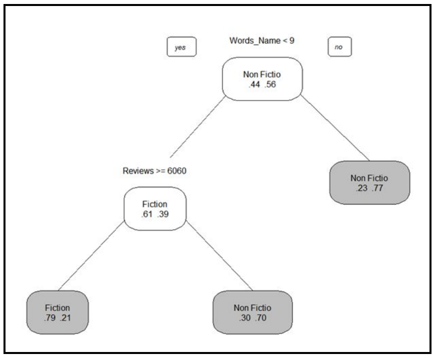
Classification trees:

We used classification trees for predicting Genre. Repeatedly splitting the records into two parts to achieve maximum homogeneity of the outcome within each new part to classify all the top 50 best selling books as Fiction and Non-Fiction. We are Considering Default, Prune, and Random Forest trees.

Default Tree:

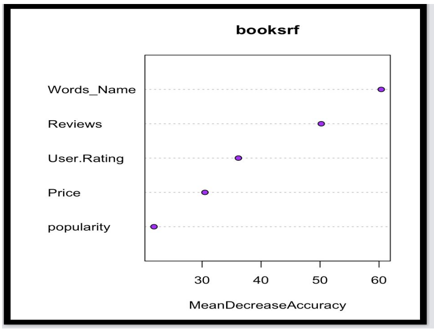
The tree is very complex and overfits the data which creates noise in the model.

The accuracy of a prediction in Default is better than the Logistic regression. The Decision Tree algorithm is ideal for this type of application. The below tree shows a rule generated by a Decision Tree.



Random Forest:

Random Forest is an ensemble of decision trees, which combine the mean prediction from many trees. One big advantage of the random forest is that it can be used for both classification and regression problems. It ranked the following predictors in the order of importance.



Observations:

Types of Models	**Logit**	**Default tree**	**Random Forest**
Accuracy	0.7	0.73	0.81
Sensitivity	0.59	0.60	0.77
Specificity	0.77	0.82	0.84

Prediction Results(on Test Data)

- Accuracy : 0.88
- Predicted Genre : **NonFiction**

Use case of the model

Predicting the more popular Genre will give an idea of which genre books can be expected to make it to the top sellers. Genre affects the price. Price affects sales. Hence it becomes important to give an idea to the sellers or publishers on what they can expect to maximize profits.

12) Comparisons and Lessons Learnt

Predictor	Comparison	Lessons Learnt
Reviews (Number of books sold)	1. We first started with knn as it is a simple nonparametric prediction algorithm. 2. We went ahead and used Classification Trees to compare and choose the best accuracy. 3. Choosing a better accuracy is important here because a small percentage improvement in accuracy means thousands of Reviews(no. of books sold). When this number is multiplied by Price of the book, it results in a significant amount of gain. 4. The best pruned tree and the Random forest showed an improved accuracy as compared to kNN. 5. Random Forest indicated User Ratings, Genre & Title word count are influential variables	1. Our initial decision of using the Number of words used for the title of the book turned out to be an important predictor in determining the number of books sold.
Price	1. We first ran the multiple linear regression(MLR) with all the independent variable to find the significant variables based on their p values 2. User.Rating & Genre were found to be significant 3. We further ran the Exhaustive Search and Backward Search methods of MLR and concluded that main variables for the final model are User.Rating, Genre and Reviews. 4. The accuracy did not improve despite adding the Reviews variable.	1. Since we can predict the Reviews(no. of books sold) we decided to use Reviews in the final model.
Genre	1. Our initial preference was employing the	1. Again the decision to use the

	Logit algorithm since the variable we are predicting is Categorical. 2. Although we obtained a good accuracy we noticed that the Decile chart bars did not consistently decline, indicating the model is perhaps not doing a good job of predicting the responses correctly. 3. We further used the Classification Trees algorithm which not only provided us with better accuracy but also showed us the important variables on the top of the tree. 4. Surprisingly the accuracy on Test data was higher compared to Validation data 5. Further, Random Forest algorithm showed the main variables were Title word count and Reviews	Number of words used for the title of the book turned out to be useful. 2. The higher accuracy on Test Data was probably because our Training Data has seen similar data already. But when we increased the records in the test data the accuracy decreased for the test data than validation. A different Test data may probably provide a lower accuracy. 3. Generally Random Forest always gave better accuracy when used on this data, both for predicting Reviews and Genre.

13) Conclusion

1. The main variables that affect the Reviews(number of books sold) are Amazon User Ratings, Genre and the Number of words used for the title of the book.
2. Genre and User Rating are the main predictors for Price. Reviews slightly affect the Price of a book, and hence we decided to use it in our final model.
3. The more popular Genre for the future year is NonFiction.
4. Title of the book has an affect on Genre and Reviews(number of books sold)

14) Recommendations

Based on the models we have developed we believe the below recommendations would be useful for the publisher or seller of the book in order to maximize their profits, provide an idea on what’s the more popular genre in future for new Authors hoping to make it to the Best Sellers list.

1. Choose a Name/Title of the book with less than 10-12 words to sell more books.
2. New Authors should try to produce NonFiction books as they are predicted to be the more popular Genre and they will also be priced higher.
3. A Name/Title of 5-13 words are usually used for NonFiction books
4. Since we can predict Reviews(number of books sold), we can make use of this to determine the best price.