**Final Report**

**Project Title**

**Author Earnings**

By-

**BI.002- Group 04**

**Ila Sharma**

**Arti Nanote**

**Dhruvesh Shah**

**Ashwin Gajendra**

**Goutham Kumar**

**Executive Summary**

|  |  |
| --- | --- |
| **Sub Stage** | **Text** |
| Overview and project aim | This project involved the analysis of a dataset of Top 7000 genre e-book bestsellers (publishers) - an excessively vast dataset with a chain of dependent and independent variables. The primary aim of the task was to study and deep dive into the data on SAS Enterprise Miner platform to deduce a model which functioned successfully and helped budding authors decide which publisher to approach based on daily units sold. |
| Outline of Process and Planning stage | Initially, individual variables of the data were studied separately. Team members brainstormed and integrated ideas for dependency and other parameters. Gradual implementation of techniques taught during lectures was done and discussions were held regarding the proposed model. |
| Outline of initial analysis stage and data pre-processing. | Following data pre-processing, we rejected the variables Children’s e-book,Comic\_Graphic\_novels and Foreign Language books as these variables had low variable worth in their contribution to data analysis. Following this we decided on the train and validation data in the ratio of 60:40 with no test data. |
| Outline of evolution of the data model and modifications | After data division, we had to deal with missing values so we implemented the replacement node. Mean was used as the imputation method as it turned out a winner in comparison to median imputation. Observing the high skewness of our data, we used Log10 transformation. Interactive binning was an important aspect of our decision making. Bins of input variables (average\_rating, sales\_price and total\_reviews) were created to reduce the number of unique levels as well as the attempt to improve the predictive power of each input. |
| Outline of construction and testing of final model (including issues and changes) | In construction of the final data mining model, we ran a model comparison on neural network, decision tree and logistic regression. Of the implemented models, on running the model comparison, Decision Tree emerged as the best model. |
| Summary of results | Inference from this analysis states that the model can help a budding author to a great extent to decide which publisher to approach for maximum profits on the daily units sold but this analysis can further be extended to provide better results. |

**Motivation**

One of the business analytic problem authors face is the earnings from the published items. The main motive of this project is to help the author make proper decision in choosing the publisher and get the daily units sold to earn more profit.

**Source of Data**

The data we acquired is from(source): <http://authorearnings.com>, which is a second-hand data gathered from various sources i.e. first hand data and also from the authors themselves.

The dataset we will be working on is a result of widespread sharing of data by the retailers, publishers, and authors.

How were books classified as “Indie-Published”, “Amazon Imprint Published,” “Big Five Published,” or “Uncategorized Single-Author Publisher”?

Here’s how:

1) The Big-5 Published books were easy to separate out, no matter what imprint they were published under, by checking the “Sold by” line in the Amazon Product Details, which listed one of: Random House, Penguin, Hachette, Macmillan, HarperCollins, or Simon & Shuster as seller.

2) Amazon’s Publishing imprints were also easy to separate out, as there are relatively few of them

3) A lot of these might indeed be Indie Publishers, but we wanted to be conservative and err on the side of understating–rather than overstating–Indie numbers.

4) If no Publisher at all was listed under Product Details, the book was considered Indie- Published.

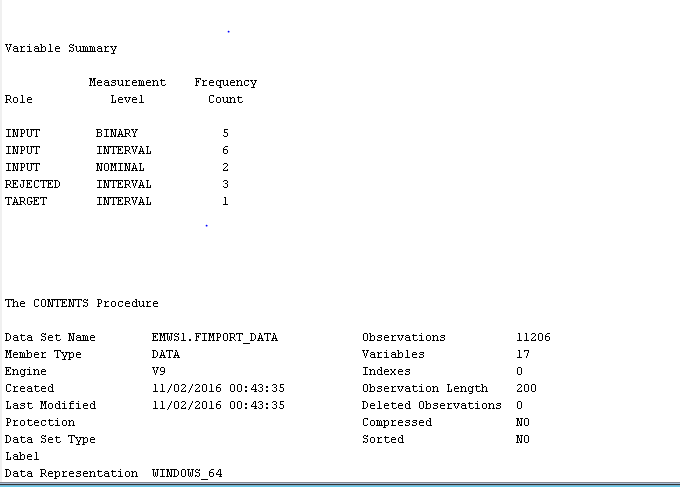
5) The remaining books, whose publishers represented only a single author name, were grouped under Uncategorized Single-Author Publisher, and sorted by revenue.

**Data Definition**

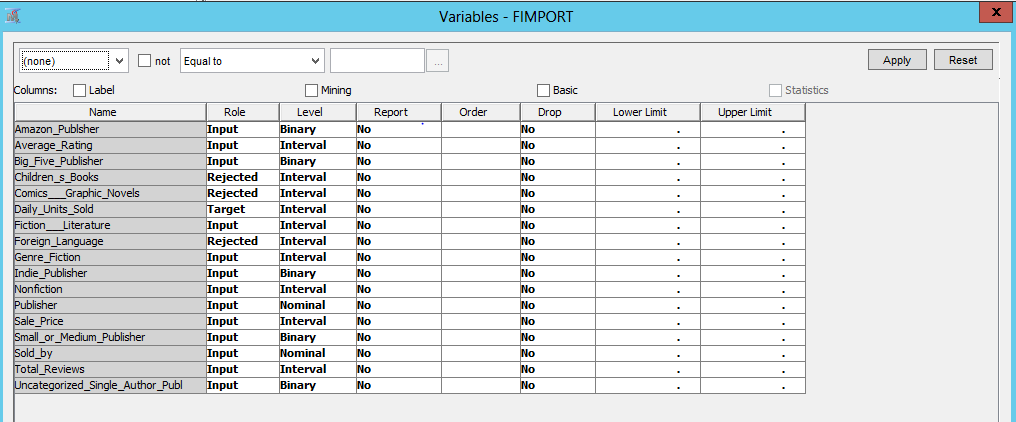
The data set is a CSV file which consists of 11206 observation. It is the result of widespread sharing of data by the retailers, publishers, agents, and authors. We have 16 input variables and 1 output variable.

1. **Data Import**

The CSV file will be the source for importing data to SAS using “File Import” node.

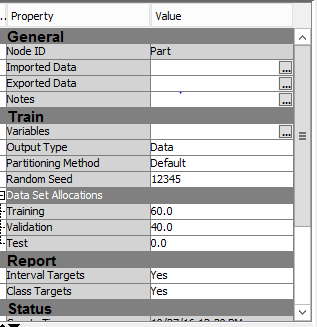


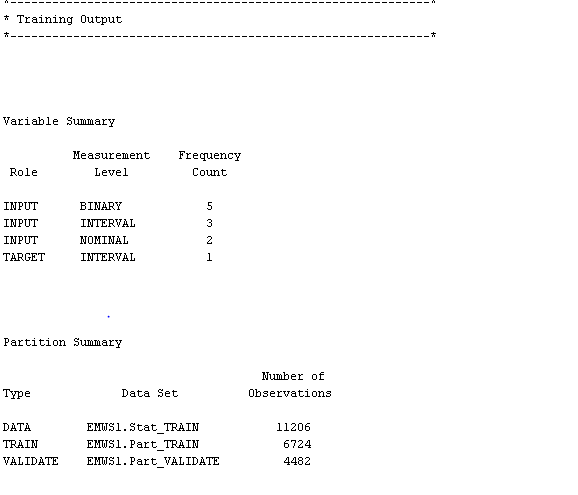
2) Once the data was successfully imported, we changed the roles of each of the variables as per our requirement. “**Daily units sold**” variable is our target variable. We manually rejected those variables which would not contribute to our analysis. Children's e-book, Comics \_Graphics \_novels and Foreign Language books are the variables that were rejected because the variable worth for these variable is very less as compared to other variables.



**3) Data Partition**

Data partition node is used to split the data into partitions. Train Data set is 60%, Validation is 40% and test is 0%. There is no generalize rule as to how much data each partition should contain but we usually have more data in the training to have better classifiers.



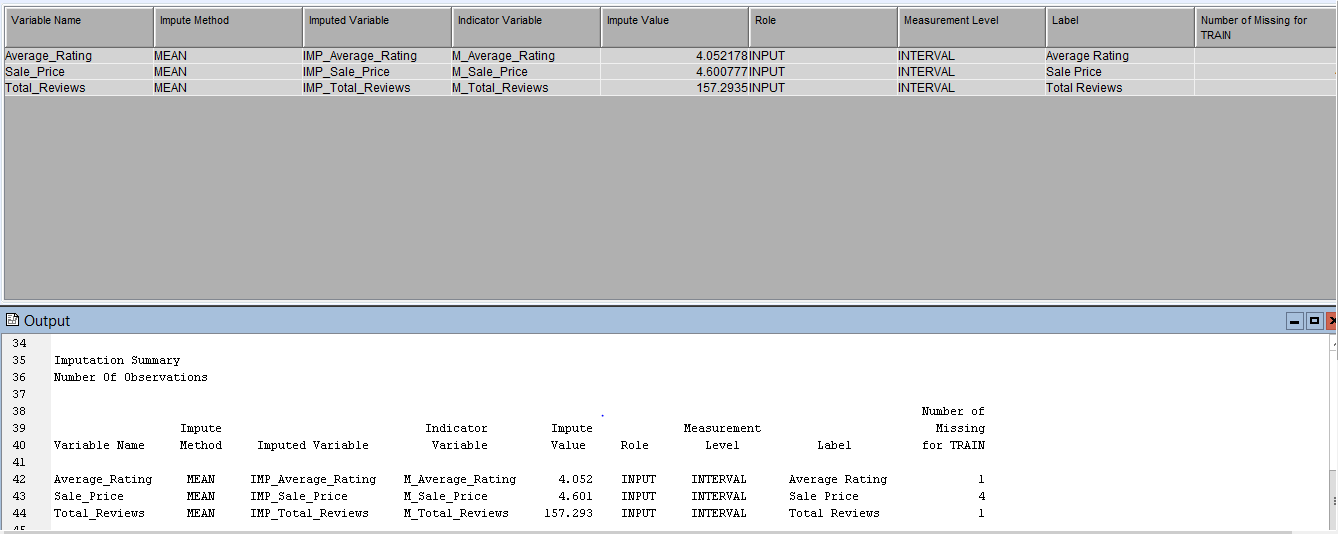


4) We will now **pre-process** the data to make it better suitable for models:

1. **Data Cleaning - Handling missing values.**

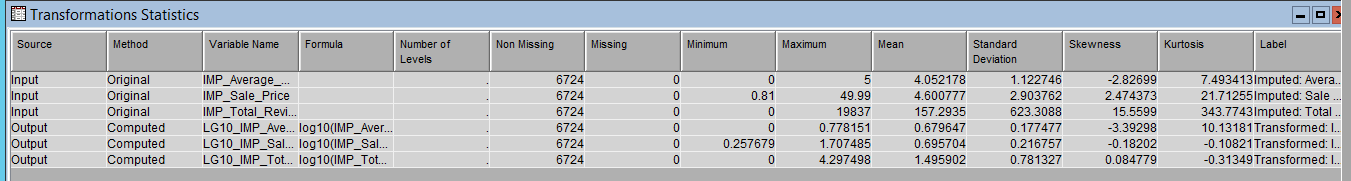
There are some missing values in the dataset which are represented by ‘.’, hence we used replacement node that identifies such values as missing. We have missing values for Average rating, Sales\_Price and Total reviews. If the average rating is 0 we have represented it by a ‘.’ since we are considering 1 as the lowest rating. Also, if sales price is less than 0.5 we have replaced it by a ‘.’ . Similarly if the total reviews value is less than 10 we have considered it as missing and replaced it by a ‘.’.

1. **Impute Node**  
   We have used impute node to impute the missing values for average\_rating, sales\_price and total\_reviews. The impute method we used here is mean. We tried with median but it did not give much difference hence we decide to go with mean.



1. **Data Transformation**:  
   Further step in data pre-processing is transforming the input variables for skewness.

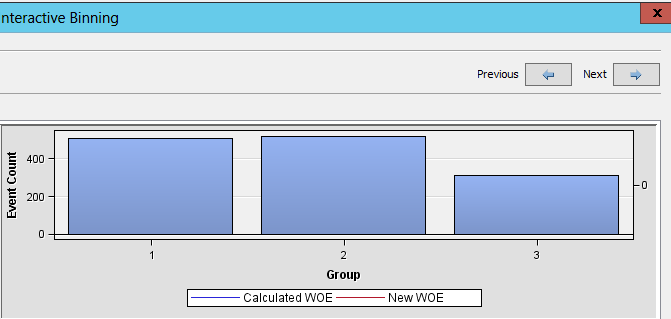
We have used transform variable node to reduce the standard deviation and skewness of the input variables (average\_rating,sales\_price and total\_reviews) and make it better suited for the models. We used log10 transformation because it makes highly skewed distributions less skewed.



1. **Interactive Binning Node:**

We used the Interactive Binning node to create bins of input variables (publisher, sold\_by, total reviews) to reduce the number of unique levels as well as attempt to improve the predictive power of each input.

Total Reviews:



**The SAS diagram is shown below**



**4) Next Steps**

Now, that the data has been pre-processed we build the predictive model.

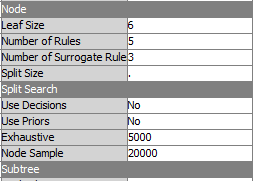
1. **Decision Tree**:  
   We built a model using decision tree as it enables SAS Enterprise Miner to automatically train a full decision tree and to automatically prune the tree to an optimal size.

Train properties:

Maximum Depth: 6  
Leaf Size= 5

Number of Surrogate Rules: 3

We tried different values but the above set of values gave the best result.



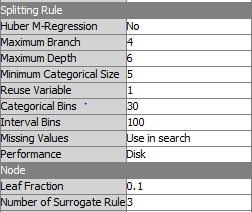
1. **Gradient Boosting**:

We also used gradient boosting model as boosting is less prone to overfit the data than a single decision tree. If a decision tree fits the data well, then boosting often improves the fit.

Train properties:  
Maximum Depth:6

Maximum Branch:4

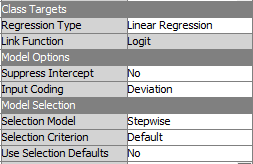
Number of Surrogate Rules:3



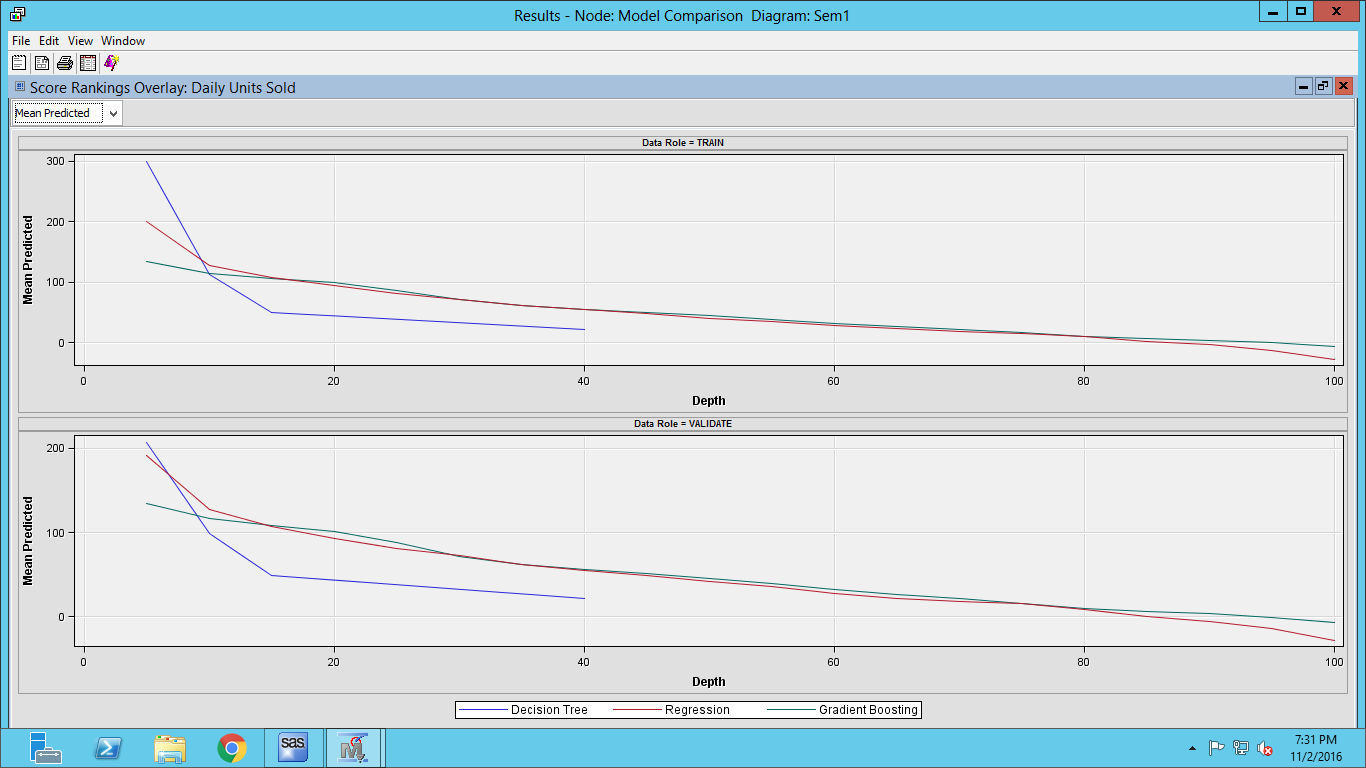
1. **Regression**:  
   Since our target variable is interval, we have used linear regression.

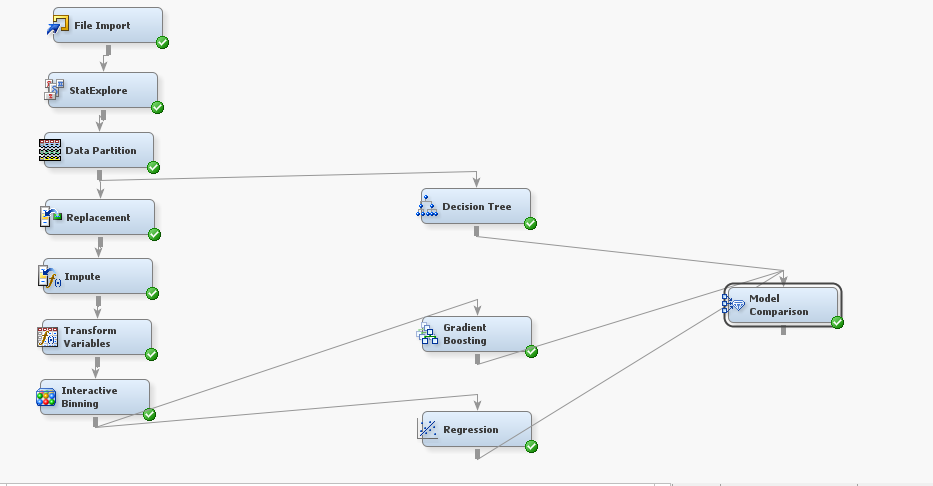
* Among the various automatic model-selection methods, we used stepwise regression. The stepwise approach is much faster, it's less prone to overfit the data. Stepwise regression presents us with a single model constructed using the [p-values](http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients) of the [predictor variables](http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/regression-models/what-are-response-and-predictor-variables/).
* Best subsets regression does not pick a final model but it does present with multiple models and information to help us choose the final model.

We have used stepwise regression due to its above-mentioned advantages.



5) Next, we implemented data mining techniques such as Decision Tree, Gradient boosting and Regression and compared them using Model comparison node and we found out that the decision tree is the best model. The model performed reasonably well.





**Conclusion**:

Decision tree comes out to be our best model. One of the major advantage of decision tree is that it can easily be converted to rules.

400-800

>800

<400

Selling Price

<3.5

>3.5

>3.5

<3.5

Selling Price

**R**

**Rules**:

If a new author wishes to sort out a publisher for his/her book, he will have to go through the reviews of the publishers first.

1. **IF (Reviews <400 and Publisher is Indie Publisher or Uncategorized Single Publisher and Selling Price >3.5) THEN Daily Units Sold <299.3.**
2. **IF (Reviews <400 and Publisher is Indie Publisher or Uncategorized Single Publisher and Selling Price <3.5) THEN Daily Units Sold >299.3.**
3. **IF (Reviews between 400 -800 and Publisher is Amazon) THEN Daily Units Sold >299.3.**
4. **IF (Reviews >800 and Publisher is Big Five Publisher and Selling Price >3.5) THEN Daily Units Sold <299.3.**
5. **IF (Reviews >800 and Publisher is Big Five Publisher and Selling Price <3.5) THEN Daily Units Sold >299.3.**

The above rules are as per the decision tree model.

On a more detailed note, Amazon is a standard publisher and its daily unit sold doesn’t depend on the selling price. On the other hand, Indie publisher, uncategorized single author publisher and Big Five are a combination of other publishers; so if a new author is unable to track its seller through Amazon Publisher, he/she will have to approach Indie Publisher/uncategorized single author publisher or the Big Five Publisher. Furthermore, if the author wants to start on a smaller scale, he/ she would have to preferably go to Indie Publisher i.e. self- publish the book. However, if the author wants to gamble and approach the large-scale publishers, he/ she must deal with Big Five publishers.

**Business Perspective**

Revenue plays a major role in deciding the publisher. Amazon and Big Five Publisher take up a large chunk of the profits generated through daily units sold but Indie publisher provides the major profit share to the author.

These two are the basic decision criterion for an author to decide on a publisher.

**References**

<http://authorearnings.com/report/the-report/>

<https://support.sas.com/documentation/cdl/en/emgsj/67981/PDF/default/emgsj.pdf>