ESTIMATING THE CAUSAL EFFECTS OF REVIEWS ON SALES
AMAZON KINDLE EBOOK SALES

AMAZON LOGO

PROJECT INTERIM REPORT

PROJECT SUMMARY

REPORT DATE	PROJECT NAME	PREPARED BY
10/29/2018	ESTIMATING THE CAUSAL EFFECTS OF REVIEWS ON SALES	Seth Turnage

INTRODUCTION

It is no secret that Amazon Kindle is having sales problems. Along with formatting issues (resulting in poor reviews), Kindle eBooks often don’t inspire enough confidence for consumers to chance purchasing a Kindle eBook as opposed to a physical copy. For potential eBook publishers, having an analytical tool to help quantitatively gauge how their reviews affect the number of sales (downloads) of a Kindle eBook, in conjunction with the effect that the description has (whether it is interesting, misleading, reassuring, etc.) is vital to helping map go-to-market strategies for potential Kindle eBook creators.

PROBLEM STATEMENT

Our task is to create an analytical model which identifies consumer *needs*, specifically as they pertain to eBooks, and uses textual analysis in conjunction with a clustering algorithm to identify certain themes consistent between reviews and product descriptions. Then, a model which identifies consumer ‘concerns’ will be correlated to sales-ranking (with the proper coefficient applied) of an eBook.

METHOD

Textual analysis will be used to identify consumer ‘needs,’ A sample of reviews will be taken, and from this sample words with recurring importance will be identified. Based on the list of top-ranked words, certain ‘synonyms’ will be identified manually, and the model will be re-ran to test if these terms have potential statistical correlation. If so, they will be grouped together. Once this has been done sufficient iterations, we will have a clustering model describing the commonality of the recurrence of terms like: ‘’ or ‘’. These will be correlated with sales data to construct our model, using linear regression.

Amazon doesn’t publish its sales data directly. Instead, it uses an unpublished ‘sales rank’ algorithm in order to notify consumers of the relative position of a given product in the marketplace. According to market research, this is a temporal algorithm, and how recent an order is disproportionately distorts the algorithm. Therefore, market observations will have to occur over a particular time interval, say a month, and the date at which reviews are left should be correlated with observations over this interval.

An algorithm using available sales data and linear regression can also be developed to estimate actual sales effected (as opposed to ranking), but this will be left for last as it is well outside the range of a minimum viable product.

PROJECT OVERVIEW

TASK	% DONE	DUE DATE	PROJECT MEMBER	NOTES
AMAZON REVIEW TEXT ANALYZER	25	11/26/2018	Seth Turnage	Beautiful Soup and Requests Python libraries
REVIEWER ‘NEEDS’ CLUSTERS	3	11/26/2018	Seth Turnage	Classifying product descriptions using keywords
REVIEW - SALES REGRESSION MODEL	0	11/26/2018	Seth Turnage	Identifying trends within Kindle market
*ACTUAL SALES ESTIMATOR	0	11/26/2018	Seth Turnage	*beyond MVP, icing on the cake