Project proposals

Q. What is the problem I want to solve.

Write a bestseller is hard. In fact, there are thousands of title being published every year and most of them goes unnoticed. However, a bestseller can create a generous income stream that trickles from the author to publishers to movies and to many retailers like Amazon. Taking a look at what popular with readers can provide valuable insights on how to tackle/chose the next book project that can be a financial success!

And even before it started to invade the prestigious New York Time list, what are the best ideas to get a successful book?

How to get the “generally” good sales? What genre should it focuses on?

How to get the best rating possible? Below is the objective analysis on this very subjective discipline.

Q. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

While this analysis would provide a good starting point for the future machine-writer who wants to challenge the human writer, there is no reason why this cannot apply to the many dorm room authors out there.

This paper will not show you how to write a good book – English classes teach you that. However, it will provide honest facts on building a popular book.

Q. What data are you using? How will you acquire the data?

I am using multiples of dataset:

1. [publishers](https://corgis-edu.github.io/corgis/csv/publishers/)
    * Ebook sales data from Amazon for 27k titles in 2015
2. [BX-Book-Rating](http://www2.informatik.uni-freiburg.de/~cziegler/BX/)
    * Rating info on over 270k titles
3. [BX-Books](http://www2.informatik.uni-freiburg.de/~cziegler/BX/)
    * Books info on over 270k title above. Lacking isbn!
4. [kindle](https://bigml.com/dashboard/dataset/5e7999ae59f5c368a40037e0)
    * Books info on 45k kindle books. Including price
5. [goodreads](https://www.kaggle.com/jealousleopard/goodreadsbooks#books.csv)
    * Goodreads book dataset including rating and reviews
6. [nyt_fiction](https://www.kaggle.com/cmenca/new-york-times-hardcover-fiction-best-sellers)

Q. Briefly outline how you’ll solve this problem

1. Is this a supervised or unsupervised problem?

    This is a supervised problem.

2. If supervised is it a classification or regression problem?

    Regression.

3. What variable is it you are trying to predict?

    Sales volume / Revenue

4. What variables will you use as predictors?

genre, sale_price, publisher_type, statistics_total_reviews, sold_by

5. What will be your training data?

I will use K-fold for this problem.