In-depth analysis

The following report outlines two standard approaches for predicting book sales: linear regression and tree-based regression. For each approach, two algorithms will be compared using GridsearchCV on training data. The better performer is chosen by having the lower root mean squared error (RMSE). Afterward, the whole dataset is refit for interpretation.

Ridge and lasso are the linear algorithms of choice. Both run fast and perform decently well with easily interpretable coefficients. For tree-based regression, gradient boosting and random forest offer strong performance improvement but at the cost of training time and no interpretable coefficients.

The structure of the fitting process is as followed:

1. Features cleaning and engineering
2. GridsearchCV ridge and lasso regression => refit better algorithm and interpretation
3. GridsearchCV boosting and random forest => refit better algorithm and interpretation

Feature cleaning and engineering

The book sales data is clean and tidy. However, some features must be transformed for analysis:

1. publisher type, seller, and genre column will be each converted to multiple columns as dummy variables. One column is dropped from each feature to prevent collinearity. This one column also serves as base line when coefficient is interpreted.
    a. children genre is dropped from genre dummy columns.
    b. amazon published book is dropped from publisher type dummy columns.
    c. Amazon digital service is dropped from seller dummy columns.
2. sale price and unit sold are log transformed to minimize the effect of the long tail.
3. publisher type is transformed to a vector matrix with hashing trick.

In the end, our feature list will include: price, 5 sub-genre columns, number of reviews, average rating, 4 sub-publisher type, 10 sub-sold by columns, and 309 hashing vectors. We will use these to model daily sales variable.

Lasso

The only hyperparameters here is alpha with ridge using L2 regularization and lasso suing L1. Both algorithms give near identical results with the test RMSE of 1.12. Without any preference, we refit the whole dataset on lasso using alpha = 0.005.

Model:

$$\ln(units\_sold) = \beta_1 \ln(price) + \sum_{k=2}^{n} \beta_k x_k + \lambda \sum_{j=1}^{n} \left|\beta_j\right|$$

where lambda is alpha.

The coefficients are as follow:

- average_rating 0.0676
- sale_price -0.1557
- total_reviews 0.3436
- genre_comics 0.0091
- genre_fiction 0.1952
- genre_foreign language -0.0271
- genre_genre fiction 0.5271
- genre_nonfiction 0.1252
- sold_by_Cengage Learning -0.0037
- sold_by_DC Comics 0.0185
- sold_by_Hachette Book Group 0.0000
- sold_by_HarperCollins 0.0041
- sold_by_Idea & Design Works 0.0036
- sold_by_Macmillan -0.0353
- sold_by_Penguin Group (USA) LLC -0.0000
- sold_by_RCS MediaGroup S.p.A. -0.0000
- sold_by_Random House -0.0764
- sold_by_Simon and Schuster Digital Sales Inc -0.0196
- publisher_type_big five -0.0311
- publisher_type_indie 0.0000
- publisher_type_single author -0.0000
- publisher_type_small/medium -0.0584

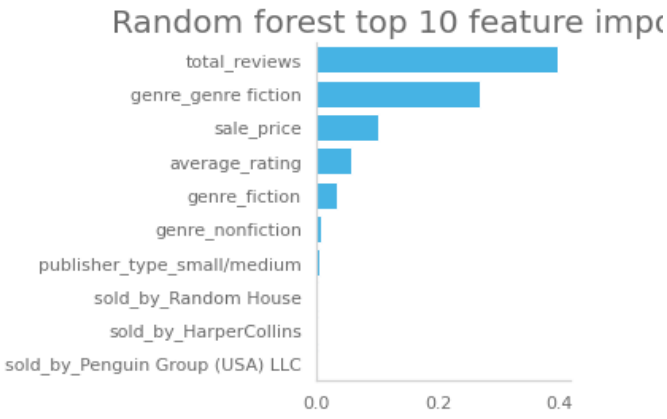
Observations:

- Price has a negative relationship with sales. On average, a 100% increase in ebook price decrease sales by 15.57%.
- Increasing rating by 1 star only generate 6.76% more sales. In contrast, 1 more review can be as impactful as increasing sales by 34.36%.
- Genre:
    - genre fiction books generate 52.71% more sales on average than children books.
    - fiction books generate 19.52% more sales on average than children books.
    - Non-fiction books generate 12.52% more sales on average than children books.
- Books sell by Random House generates the least sales compared to Amazon digital.
- Publishers:
    - big five publishers generate 3.11% less sales than amazon published books
    - small/medium publishers generate 5.85% less sales than amazon published books

Random Forest

For this Gridsearch, the only hyper parameter that is tuned is max depth. Even adding one or two more hyper parameters can increase the training time significantly. Nevertheless, both tree-based regression performs better than linear models with 19% lower RMSE. The cost of training time is also significant even on this small dataset. With 3+ hyper parameters tuning, tree-based regression can take hours instead of minutes compared to linear methods.

Since random forest gives almost identical RMSE to boosting, we will refit the whole dataset using random forest to obtain the feature importance.



The top 10 feature importance confirms our analysis with lasso above. One difference is the order of total number of reviews which is the most impactful variable in random forest versus genre fiction in lasso.