We will apply two machine learning algorithms to the book sales dataset. Two methods are Ridge Regression and Random Forest. Ridge regression will give us a concrete coefficient to interpret the results while Random forest will provide feature importance as well as a modern take on the data.

Two objects of this analysis are:

1. How other features effecting the daily book sales values?
2. How can this be applied?

The dataset is book sales data that has been cleaned and tidied. However, some features can be transformed for analysis.

1. Publisher_type, seller, and genre will be converted to integer values.
2. Publisher_name is converted to multiple vectors using hashing tricks.
3. Price_range is omitted as this correlates with price variables.

In the end, our feature list will include: price, genre, number of reviews, average rating, publisher type, sold by, and 8 hashing vectors. We will use these to model daily sales variable.

Ridge Regression

We apply GridsearchCV to compare Ridge and Lasso across multiple alphas. Both gives near identical results with the RMSE of 148. This is quite a good result given little tuning of variables and hyperparameters. Moreover, the classical methods also provide practical interpretations.

The coefficients are as follow:

- genre 0.773262
- sold_by -17.106463
- publisher_type -31.967857
- average_rating -0.161827
- sale_price -6.352419
- total_reviews 51.564657

Some interesting observations:

- genre has a smaller effect than expected. However, the genre fiction (code 4) and fiction (code 3) still out selling any other genres.
- Large negative on sold_by / publisher_type indicates that big publishers generally have better sales. This is expected since they have more resources from quality control to advertisement. It also means that one should get the big publishers to publish their books since they have higher profit margins and popularity.
    - Big publishers is code as 1 vs 2+ for other publisher types.
- Average rating is negatively correlated with unit_sold. This needs to be studied more with plots since higher rating should return higher sales.
- On average, an increase of $1 will decrease sales by 6.35 units.

- Total number of reviews become the most important factor in how well a book sell. This is not surprising given a popular book will get more sales and more reviews. However, a big distinction here is that a popular book does not necessarily mean a good book.
- The other 8 hashing vectors has little interpretation value so we will ignore them. Their values vary wildly.

Random Forest

Applying random forest using GridsearchCV to see how much improvement we can expect from a more non linear method. The RMSE we get from random forest is 119.53 which is about 29 units better than linear methods. Again, this is little tuning of variable.

In this case, linear methods give us good results in a very fast training time comparing to random forest.

Feature importance is as followed:

- genre 0.046709
- sold_by 0.015321
- publisher_type 0.008728
- average_rating 0.048635
- sale_price 0.062769
- total_reviews 0.528626
- 0 0.020690
- 1 0.046067
- 2 0.015812
- 3 0.026295
- 4 0.120756
- 5 0.031050
- 6 0.004612
- 7 0.004647
- 8 0.008477
- 9 0.010806

This confirms our analysis above using Ridge Regression about the impact of each feature. However, it also gives us an interesting insight: hashing feature 4 has a good impact on the daily sales value. This will require a much more careful look to see why this is the case.

Limitation

The analysis is performed on ebook sales data only and thus might not apply to traditional brick and mortar book store.

Moreover, the algorithm is fitted with quick tuning of hyperparameter and untouched features. To further improve the model, we can engineer more features using transformation (square, root, …)