Zalora's Data Science Technical test.

Introduction & Objective:

Zalora is a online fashion retailer. We have hundreds of thousands of products accross all of our ventures - Singapore, Indonesia, Malaysia, Hong Kong, Thailand, Philippines, Vietnam, Brunei, Australia and New Zealand.
One of our challenging problems, is to determine which products are better than others, essentially a venture- wide ranking for all of its products. The ranking can be different for each different product category / subcategory.
The result of this ranking serves a lot of purposes:

- Displaying best selling products on our homepage.
- Sorting our catalogs / search results whenever users surf our homepage.
- Sending better, more profitable email campaigns.

Dataset:

(Attached as a single products.csv file) Luckily, our Data Engineering team has done a lot of data cleaning and aggregation to produce everything in a clearly formatted table. The given data set contains 4000 Indonesia products and their metadatas that are currently available on our store, sorted by a random order.
[Notice: a large part of this was randomized since it contains confidential information]

Questions:

Responder: Sundar Lakshmanan

1) Give us your suggestions on how we could make our data set better / more useful.

The dataset is very extensive, although a few improvements can be made as suggested below.
i) Some of the parameters and predictors should be normalised for accurate comparison.
e.g.
a) sales to impressions, canceled sales to sales and other suitable ratios should be added as additional features.
b) for all numeric features like impressions, sales, etc. median values for that category can be an additional parameter for normalisation.
c) to know the current hotness of products, a ratio of last 7 days to all time values can be added as a parameter.
ii) User-related features should be added. This is very important since product ranking is highly personal. Since the data has details like gender, it will be useful to get some info about users like their gender, past buying history etc. can be added
iii) since there is a seasonal element, day of year, week number, month, quarter etc. can be other features.
iv) If possible, impressions from search and navigation should be separately counted. This is because, a featured article will naturally get more impressions regardless of its product rank. Search, on the other hand, reflects buyers' intent.

2) With the given dataset, can you come up with a scientific approach and model for our ranking?

We can build a regression model and derive weights for the various predictors. Some parameters may have low prediction, which can be dropped to simplify the model. More complex models can be used if absolutely justified in the tests described below.

3) How would you test, train, and evaluate your model?

a) Identify key features by talking to the various teams like marketing and products.
b) Write scripts to generate features in a scalable and repeatable manner (if required, using Hadoop Pig)
c) Choose a simple regression model to start with. That will be much more interpretable than complex models.
d) Train it with 90% of the data.
e) Test mean squared error (MSE) on training data as well as the remaining data which was not used for training.
f) Calculate Bias-Variance trade-off for the model to make a decision on whether a different model should be used. Iterate.
g) Try the candidate model on x% of live traffic for A|B testing. Iterate.

I can expand upon each of the above as required.