Summary Report

Milestone 1

Team: Sehaj Chawla, Taro Spirit, Lotus Xia, Heather Liu

For context: We met with our partner last friday, and discussed our data processing procedures and baseline model architectures in detail. To avoid extensive repeated information, we phrase this email as a meeting note that summarizes the main takeaways from the meeting.

Hi Hamilton and Jacob,

Hope you had a great week so far!

We are sending this email as a summary of our discussion last Friday (and also as a course deliverable for our capstone project). You may also find the technical report in attachment that records in detail all of our data processing steps and model architectures. We will maintain this technical report as we go to keep track of some model decisions and iterations.

Summary of progress so far:

1. Due to the large noise embedded in the daily BSR data, we decided to smooth and normalize the daily rank, and work with a monthly median BSR in our modeling process (at least for now). It is our understanding that some of these noises are due to external factors that can hardly be predicted from reviews—for instance, a sudden spike of BSR may be a result of the product being out of stock for a short period of time. Therefore, we choose to focus on predicting the general trend of the BSR movement from month to month.
2. In terms of modeling, we coded up two baseline regression models (as opposed to time series models) to predict the change in monthly median BSR.
    - A model that uses only review metadata, including review rating,
    - A model that uses only review text (weighted by review votes)

    In both of these models, each unit of observation is a product-month.

Immediate next step:

1. We talked about how weighting reviews by the number of upvotes is a form of data leakage, since the upvote values are only a snapshot as of the time of scraping. To address this issue, we experiment with a version of the models that **excludes review upvotes** altogether, in order to see how much of the explanatory power of the models is due to this data leakage.

2. We are also working on a **Bag of Word** model with regularized regression to serve as a baseline NLP model. Hopefully, this model will give us some insight into whether frequent occurrences of certain words are predictive of change in monthly BSR.
3. We finally got AWS credits and can start serious training!
    a. We plan to train the baseline models on a random sample of ⅓ products (and all their associated reviews). We hope this training set will be large enough for the model to pick up some patterns (pun unintended) and avoid unnecessarily long training time.
    b. As you suggested last time, we have switched to **TinyBert** in the NLP model.
    c. We will conduct more careful **feature engineering** and **hyperparameter tuning** of the metadata based models (linear regression, random forest, and XGBoost). We also plan on trying to use the estimated sales volumes as the target variable in the set of models to see if we can gain any insights there.

We hope to accomplish most, if not all, of these immediate next steps before we meet on Friday next week (March 11). Please let me know if there is anything else we should address in the meantime.

Thanks! And looking forward to meeting next friday!

Sincerely,
Team