AC 297R, Spring 2022
Harvard John A. Paulson School of Engineering and Applied Sciences

Summary Report

Milestone 2

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

For context: We met with our partner last friday, and presented our Milestone 2 presentation and discussed our results 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 again also as a course deliverable for our capstone project). You may also find a technical report in attachment that records in detail all of our results. We will continue maintaining this technical report as we go to keep track of our progress.

Summary of progress so far:

1. We have developed regression models without any text in the input, i.e. a linear regression, an xgboost model and a random forest model, to predict monthly median sales volume. After some hyperparameter tuning, we discovered that our best model (xgboost) could reach an R^2-score of approximately 95%. Our understanding of this very high score comes from the fact that the most important feature used by the model is the monthly mean sales volume. This indicates that the model is using the momentum of a product to predict future sales volume which was expected.
2. We also developed two types of models predicting monthly median sales volume based only on the review texts. The first type of model is a bag-of-words model which serves as a baseline model which is highly interpretable and easy to train as well. After tuning hyperparameters and specific architectures of the model, we found that a TF-IDF model with a LASSO regression could reach an R^2-score of approximately 14%. The second type of model we developed is a transformer model using tiny BERT. This more complex model reached an R^2-score of 16%. This score is already a significant improvement compared to the baseline bag-of-words model. We have additionally several ideas to improve the score for the transformer model. All in all, we discovered that the text-based models were able to explain a non-negligible part of the future near-term performances of products. However, the non-text models are outperforming them by far.

Immediate next step:

1. We would now like to focus on predicting long-term product performances. Indeed, this task could potentially be more useful in practice as it could predict the long term success of a product right after its launch. The task will be formulated as follows: based on the review data of the first three months of a product can we predict if it will ever reach a successful rank in the year that follows. A successful rank is a relative measure which we plan on defining as being approximately the top 15% of the ranks recorded in the vitamins and supplements category. The shift to this new task will be our main focus from now on and we hope that the reviews' metadata will have a greater impact on the prediction of long-term success of a product. We also hope to find some key insights in the text of reviews to predict a product's long-term success.

We hope to make some significant advances on this next step before we meet on Friday next week. Please let us know if there is anything else we should address in the meantime.

Thanks! And looking forward to meeting next friday!

Sincerely,
Team