Summary Report

Milestone 3

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

For context: We met with our partner last friday, and presented our method for long-term predictions 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 long-term predictions' results. We will continue maintaining this technical report as we go to keep track of our progress.

Summary of progress so far:

We have developed and tuned **two types of classification models** to predict long-term product performances and an **ensemble model** to combine all individual model predictions to make an overall prediction. As a reminder, the long-term task is formulated as follows: based on the review data (metadata and texts) of the first three months of a product, can we predict if it will ever reach a successful rank in the following one-year period after the first year. A successful rank is a relative measure that we defined as the top 3000 of the ranks recorded in the vitamins and supplements category.

1. The first type is classification without texts, including logistic regression, XGBoost and Random Forest. The best model (xgboost) could reach an F1-score of 0.371. We found that long-term prediction, unlike short-term predictions, depends not only on the rank features but also on the review data. Some interesting features related to reviews are, for example, the number of verified comments and the average comment rating.
2. The second type is classification with only texts, including a bag-of-words model and a transformer model using tiny BERT. For the bag-of-words model, we found unigram models overall perform the best. The highest F1-score (0.34) is achieved using a TF-IDF model with unigram and L2 penalty at a strength of 0.5. For the transformer model, the F1-score is fairly stable, while AUC does get affected significantly by the model choices. We found that a small embedding size (of 10 elements) works better and the performance of the models plateaus as we increase max sequence length beyond average length.

3. The ensemble models are fitted to the predicted probability from four individual models using logistic regression and decision tree classifier. Overall, the ensemble models outperform all individual models.

Immediate next step:

1. We will focus on the interpretation of our text-based models for the rest of this project. In particular for BERT, we are looking to gain insights into what BERT performs well on and what it doesn't do so well on. We are also looking to gain insights on how BERT's predictions change with properties of the reviews.

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

Thanks! And looking forward to meeting this friday!

Sincerely,
Team