Product Recommendation on Scraped Usage Data

Atherv Gole, Cristian Razo, Eric Cha, Natasha Leodjaja, Qimin Tao

Sponsors: Rob Fox, Lauren Wong, HG Insights

UC SANTA BARBARA

HG Insights

Project Description and Goal

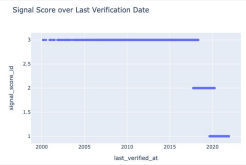
- The primary goal of this project is to analyze patterns in product adoption to effectively engage with existing and/or potential customers. Our primary dataset consists of two large datasets regarding many companies' qualities across the United States over a span of 20 years ranging from 2001 to the present.
- We're trying to predict future technology purchases based on products adopted by companies. Such that companies get to ask question like 'Show Me Automotive Companies in Germany who HG Insights predicts is going to purchase SAP HANA'.
- Two of the main features we used are 'signal score' which is a 1-3 score of how recently the product was observed, and 'intensity' which is a measure of how much a product is observed over time

Data

- Data was provided to us by HG Insights. We were given two datasets: Install and time series which were scraped from companies' job listings, product pages, etc.
- The observations in the *'install'* dataset were of a single product usage at a single company, as well as auxiliary company information like the industry, employee number range, etc.
- *Time Series Dataset:* product_hit_id, date, weighted_intensity
- There were some missing values (~5% of the entire data) in our dataset. Due to the small number of missing observations, we dropped the NA values to have a fully intact dataset.
- Since all the data is scraped, we expect to have some variance in actual vs. recorded usage of a product. Some products may be occasionally mentioned on a company's web presence, but only to describe a task or to attract candidates.

Preparing Data for Collaborative Filtering

Recency Penalty = log(current year - most recent verification year)



Aggregate Score = *weighted intensity ÷ recency penalty*

- In order to implement collaborative filtering, we wanted our data to resemble a user-ratings matrix. To do this, we created a rating-score that incorporates how recently used a product is and its weighted intensity value.
- We then scaled the values and binned the scores 1-5 company-wise (to avoid penalizing companies with low overall product usage)

Methodology

- Our best performing model to date is a collaborative filtering model with a Baseline or SVD estimator. With these methods, we were able to predict a company's rating for a product with an average RMSE of ~0.88
- To train and test these models, we used sklearn and scikit-surprise - a recommendation system library with prebuilt collaborative filtering methods.
- While the results are promising, it's important to remember that the predicted scores are based on our assumptions as to what metrics indicate a company's effective product usage, those being the 'intensity' value, and the recency of a product's usage. Collaborative filtering handles the rest by leveraging company similarity to predict product ratings.
- To include more features in our predictions, we're working on a few boosting methods using company information like number of employees and revenue to layer onto our recommendations

Company	Product	Rating
A	Quickbooks	4
A	jQuery	2

Collaborative Filtering Model

Company	Product	Rating (predicted)
A	Microsoft Excel	3.89
A	Java	1.87

Final Results

- We found that we're able to predict a company's rating for a product with an RMSE of around ~0.88 and an MAE of around ~0.57 using Singular Value Decomposition and Baseline estimation methods.
- As seen in the figure below, RMSE approaches a minimum at around 1.3 million included rows of data.

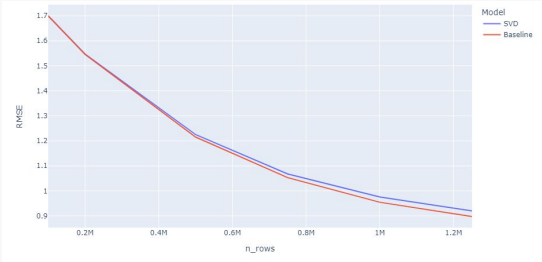


Figure. RMSE as more rows are included in the trainset.

Conclusions

- User-item recommendation is difficult- conceptually and computationally
  - Measuring user preferences can be difficult, and in our case, there are implicit business related reasons as to why a company selects a product.
  - Runtime costs for training a model at the scale of which HG collects data can be very high.
- The general goal of the project is to understand whether we can provide relevant recommendations to companies based on the information that they make publicly available
- While user-item recommendations have been done many times before (most famously in the video streaming and e-commerce industries), the opportunity to explore those same strategies with very different data presented unique challenges.