Personalized Fashion recommendation system

Team: Sai Krupa Jangala(sj3140), David Heagy(dh2868), Karunakar Gadireddy(kg2911), Jugal Shah(js5950), Jiayuan Cui(jc5670)

Overview:

Many online shopping sites face the issue of customer's scrolling through their page but not making a purchase. We decided to build a product recommendation system based on H&M's data from previous transactions, as well as the customer's and product meta data. The data includes garment type, customer's data, product description, and image data of the product. The recommendation system will help the customer decide which product to buy easily from the plethora of options. A well developed recommendation system will help businesses improve their shopper's experience on websites and result in better customer acquisition and retention. An efficient recommendation system reduces the risk of return and thereby reducing the cost the firm pays for the logistics on the return policy.

Data Description:

This dataset gives us the purchase history of the customer at different times. Our goal is to predict what each customer will buy in 7 days based on the customer purchase history after training the data. We have three csv files with different information along with the image of each article. File articles.csv contains article_id, which shows available articles that can be purchased. Folder images/ contains images corresponding to each article_id. File customers.csv contains customer_id, which has the detailed information of customers. File transactions_train.csv contains the training data, consisting of the purchases of each customer for each date.
Link to the dataset.

Proposed ML techniques:

We will perform basic EDA and data visualization on the data to get a sense of what features are most important and whether transformations are necessary. Since the data is primarily text, we will likely employ categorical preprocessing techniques, such as One-Hot and target encoding.Based on briefly looking at other recommendation systems, we propose use of K Nearest Neighbors and Neural networks. KNN is commonly used in tasks like collaborative filtering because it is non-parametric (does not make assumptions about underlying distribution) and uses distance (cosine similarity) to rank items. If we don't have categories associated with the data, we plan to try out topic modeling approaches as well. We may use NLP techniques, such as creating and feeding embeddings into Neural Networks for latent factorization. We plan to look at image data as well. We will try to bring up a system that clusters similar images together. We would then like to compare our predictions with the NLP model that we plan to build and evaluate both the approaches and then pick those predictions which are common from both the models(if they exist). If the results are very different from each other, we would consider hyperparameter tuning and then select the model for which the accuracy is high on test data.