Arvind Kanesan Rathna - ak4728
Ling Sun - ls3759
Mohit Medini Singh - mms2308
Zeyu Jin - zj2325
Hongtao Jiang – hj2554

Proposal 1 - H&M Personalized Fashion Recommendations

Background/Context:

H&M Group is a Swedish multinational clothing company known for its fast-fashion apparel and accessories. It is the second-largest global clothing retailer, with nearly 5,000 stores globally. It is present in 53 online markets, where it utilizes product recommendations to help its customers navigate through its extensive catalog. Our task is to apply machine learning techniques on purchase history of customers across time and its supporting metadata to predict what articles of clothing a customer would purchase in the 7-day period following the end of the training data. This task has been defined explicitly under the Kaggle competition through which the datasets have been provided.

Identification/Description of Data Set:

The dataset contains information about articles of clothing sold by the retailer H&M, customer information, and past history of customer transactions. The clothing data is mostly categorical, with information like product name, classifications into larger clothing groups, color groupings, and some English descriptions. The numeric data for articles of clothing is mostly just encoded categories. On each customer, we have their unique identifier, age, as well as participation in H&M's rewards club and receptiveness to marketing emails. Transaction data includes the date of purchase, what article of clothing it was and who bought it, as well as a price and through what sales channel the purchase was made. There are upwards of a million transactions and unique customers included in the dataset. There are around 45 thousand articles of clothing in the dataset. There are also images for many of the garments in the dataset.

Proposed ML Techniques:

For our task, we need to generate 12 candidate clothing articles to recommend to each customer. One approach is content based systems. Here, we recommended articles that are similar to what the same customer has purchased in the past. The article image data can be used to find clothing items that are similar in appearance as well. Another approach is collaborative filtering. Here, we make recommendations based on other customers with similar tastes in clothing. A baseline can be modelled by recommending the same articles as their previous purchase. We would also like to experiment with deep learning-based recommendation systems if time permits.

Proposal 2 - Restaurant Recommendation Challenge

Background/Context:

The restaurant recommendation project targets one of the trending recommendation needs in recent years, the vendor (restaurant delivery) recommendations. Due to the COVID pandemic in recent years, customers are shifting their preferences from in-door dining to online ordering, which leads to the rapid growth of the delivery businesses, including Yelp, Doordash, etc. All the delivery businesses aim to provide accurate vendor recommendations to different groups of customers empowered by well-designed recommender systems. In this project, our task is to apply data analytics and machine learning techniques to build a relatively simple recommender system to provide vendor recommendations to customers. (This task has been defined explicitly under the Kaggle competition through which the datasets have been provided.)

Identification/Description of Data Set:

The dataset consists of customer, vendor and order data. The customers' geographical dataset includes location number, type, longitude, and latitude, and the demographic dataset includes personal information such as gender, birth year, language, etc. This information would help us build user portraits to understand customer backgrounds and preferences. The vendor dataset contains food types, restaurant ratings, locations, and promotions. The order dataset contains the record of all orders made by customers from the customer dataset. The granularity of the vendor order dataset is set by order id. The features include general information of the respective order, such as item counts and total cost. It also includes payment-related features (payment mode, promo code, etc) and vendor-related features (if the customer, etc rate the vendor). There are ~35,000 customers for training and ~10,000 customers for testing. Among them, 61% have made at least one order from 100 vendors.

Proposed ML Techniques:

While the problem is a classification problem, an intuitive solution is to generate classifiers to classify a customer interested in a specific vendor. However, there are potential advanced solutions. The first technique we would apply is collaborative filtering. While the dataset includes geographical and demographical customer data, we have enough information to build models to study how users responded to these same vendors while certain conditions are satisfied. In this project, content-based filtering is not applicable because we do not have detailed information on vendors. The second technique we could potentially apply is applying unsupervised learning on the customer base first and then training supervised classifiers. This approach would reduce the risks of having a complex model while also increasing the interpretability of the model.