Machine Learning Final Report

Kerui Lu, Anne Chambers, Pradyoth Hegde
Keruilu2020, annechambers

Our Task

The primary goal of this project is to use past purchase and browsing history in order to predict which coupons a customer would buy in a given period of time. Our secondary goal is to perform an exploratory data analysis on the transactional data for 22,873 users on Ponpare, a leading Japanese coupon site. We are building a Recommender System, which is also known as a Collaborative Filtering System. Recommender Systems are very common today, especially for online shopping and video streaming services. Websites with the best recommendations will have the most customer satisfaction and retention.

Exploratory Analysis

We first wanted to obtain a solid understanding of our dataset. Therefore, we investigated our dataset and these were some of the key findings.

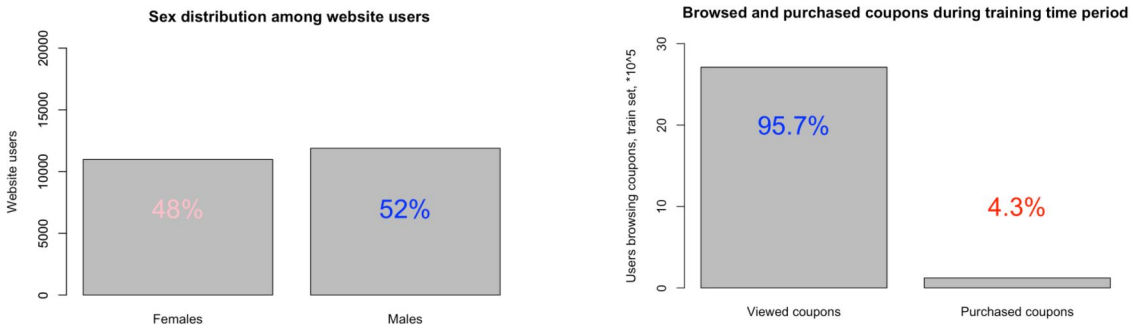

Fig1. Shows the sex distribution among website users

Fig 2. Shows the Percentage of Browsed and purchased coupons during Train period

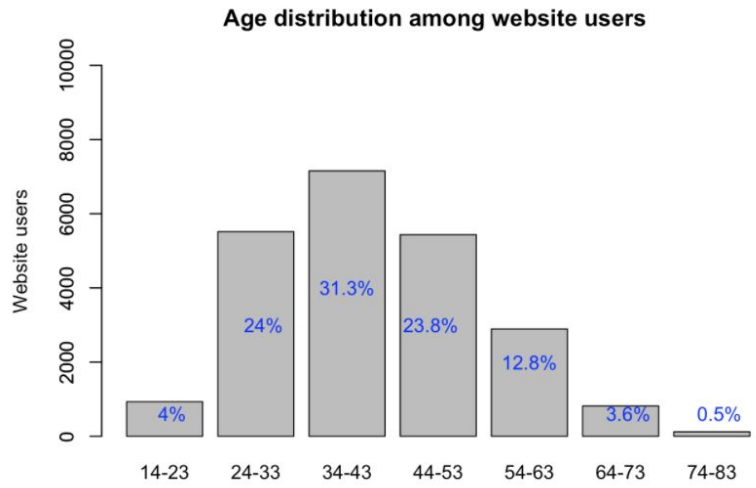

Fig 3. Shows the Age Distribution among website users

Fig 4. Shows the Coupon discount rate in the training set

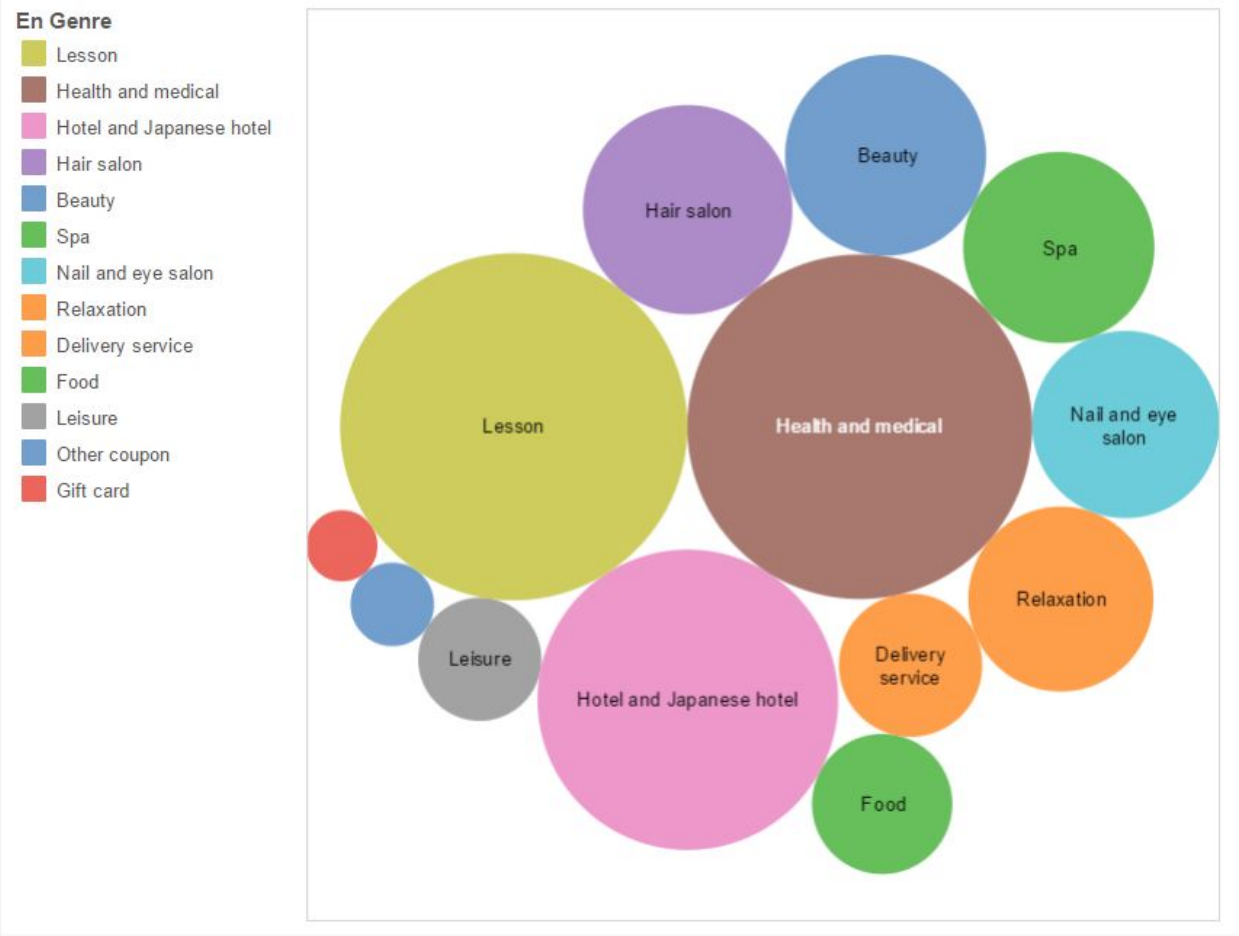

Fig 5. Shows the bubble chart of the different coupon categories

Approaches

Very little customer demographic information was provided by our data, so we decided to focus on the information provided by the available coupons.

- Predicting the most popular coupon: In order to predict which coupon would be bought by a customer, we simply predicted the most popular coupon in a given area, and checked for which area a customer was from and then made the final prediction based on those attributes.

  In order to predict the most popular coupon, we used different classifiers:

  - Nearest Neighbors classifier: One approach we tried was predicting the popularity of coupons based on their attributes such as **dates available**, **genre**, **price**, and **discount**. We could then recommend popular coupons across all users, so that even ones with very little past information could have something to fill out their list.

    - Decision Tree: A decision tree is a support tool that uses a tree-like graph or a model of decisions and their possible consequences. The attributes used for the Decision Tree approach include **dates available**, **genre**, **price**, **discount**, etc.
    - Naive Bayes: A Naive Bayes classifier is a classification technique which is based on Baye's Theorem with an assumption of independence amongst the classifiers.
- Cosine Similarity methods: Similar to Problem Set 4, where we used Cosine Similarity to predict the Caption of an image, we used a Cosine Similarity approach to predict whether a user purchased a particular coupon or not. The input vector has a list of features which correspond to the Coupon attributes like **Usable Date**, **Usable Period**, **Location**, **Genre**, etc.

For users with more past history, we recommended coupons based on their location and most frequently purchased or viewed genre of coupon. Within these categories, we would recommend the coupons we predicted to be popular.

Key Results

In our baseline approach, we only went on the predict the most popular coupons in a given area. We also used different methods to predict the most popular coupons. They were:

- Nearest Neighbor:
    - **3-NN: 37.91%**
    - **5-NN: 34.08%**
    - **10-NN: 29.44%**
- **Decision Tree: 29.48%**
- **Naive Bayes: 31.07%**

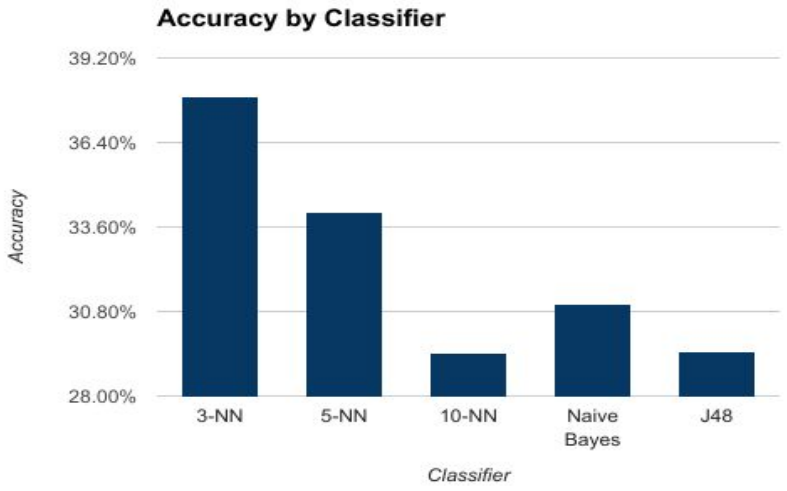

Figure 6. Coupon popularity accuracy on various classifiers

Using our best prediction, we went on to predict which coupons would a customer buy in the given period and we got a validation accuracy of **0.602%**

Since this was a Kaggle competition, we can compare our results to the leaderboard to get an idea of how we are doing. This was not as high as we might have liked, but is well within the range of results achieved by others on Kaggle.

The second approach tried was a cosine similarity between users and coupons.

- The accuracy was **0.78%** for this dataset.
- We tried the same method but including the coupon view log in our calculation. It cost us more time but helped us get a higher prediction accuracy of **0.89%**
- Since the combined matrix we get is a sparse matrix, we used Pearson Correlation Between Columns (Sparse Matrices) to evaluate the similarity between coupon and users. This gives us a highest accuracy of **1.095%.**

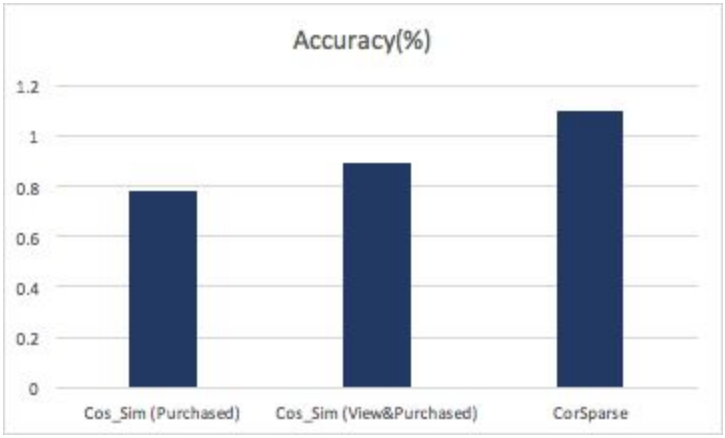

Figure 7. Coupon prediction accuracy on various similarity methods

Key Findings

- We found that 3-Nearest Neighbors works the best when it comes to predicting the most popular coupon because it has a higher accuracy when compared to the additional noise brought by the additional neighbors.
- The Coupon Area plays a major role in predicting whether a user would purchase them or not, this is in comparison to other attributes such as Time Period, or on which day it is usable.
- Using the View-log to check whether a user has looked at a coupon before significantly increases the accuracy of the prediction, but increases the training time.

Future Work

If we continued to work on this project, we would like to try some feature engineering from the available data. Unfortunately, since the dataset is so large, if we had added many more attributes to it the algorithms would have taken much longer to run. Most of the top scorers on Kaggle

reported run times between 24 hours and days, all while running on very high end computers. That is why we had a lesser flexibility to try out different approaches.

Anne: baseline method using popular coupons
Kerui and Pradyoth: Cosine Similarity Implementation