Mall Recommendation System : Project Report

Rishabh Jain
Department of Computer Science
Courant Institute of Mathematical Sciences
New York University
rj955@nyu.edu

1 Introduction

Building a mall is huge investment and hence its necessary for the owner to well plan what kind of stores and companies one should rent the space within a mall. It is considered as a business decision primarily to find the right mix of shops which could increase the footfall and spending in the mall. But in this project, we try to solve it using data based approach, where after analyzing the existing data we try to recommend stores that can be added to a mall.

2 Previous Work

There has been a significant work done on deciding a suitable shopping mall location, how the stores within a mall be oriented to and how to decide a rent for a store. But a limited work has been done on finding a right mix of stores for a mall.

Charles Carter and Marcus Allen, did explain some principles useful for finding a right mix of tenants in a mall in their work on Deciding Optimal Tenant Mix in Shopping Centers[1]. The main point made in their work was that the shops should be decided based on what they sell, such that there is sufficient competition for motivating shoppers to come in but it should not increase so much that it reduces individual shop's profits significantly.

A previous work for this particular project was done using Crab based recommendation system[2] to come up with categories of shops, given the list of current list of shops in a mall and some shops that are already present in the current mall.Crab uses collaborative filtering technique to make recommendations based on past data. The data being used was a list of all US based malls with all the stores in them.

3 Dataset

The main data being used is a list of all US malls with their addresses, geographic coordinates and and list of stores present. This is the major data to use as training set for understanding an optimal combination of stores in any mall.

Since, the studies suggest that the store mix should be based on the kind of products they sell, data for each stores categories is also being used. This data was retrieved using Google places Api, which can return the data about a store's category when looked up by its name and Geographic location.[3]

We also used national survey data for regional incomes. This is an attempt to try recommending the stores based on income assuming it to be a representative of spending in the region

where mall is located. The data was retrieved from US census website, which shares the data for national level census being conducted. We used American Community Survey(ACS) for getting income data. Apart from the income, data to map geographic coordinates to ACS's division of areas for surveys called called Public Use Microdata Area(PUMA) was obtained to map each mall to PUMA for analysis. [4]

4 Methodology

Since we couldn't get the revenue and rent based data from malls and stores, we cannot used profit maximization objective to come up with the best store combination. So the problem was tackled with a clustering approach [5]. One of the idea was to utilise Crab's collaborative filtering, but the library doesn't have any active support, so it could not be used.

We used Association Analysis Technique to find associations and dependency between the current store combinations. It utilizes Apriori Algorithm to find combination of categories and stores which are most frequently occurring together. This technique finds most common associations between the elements. It takes in the list of occurrences of categories or stores across malls and then depending on how many times which stores occur together it generates a list of commonly occurring combinations and their scores. Also, it is used to find a suggested category from a combination of categories by determining a causal behaviour in them. Then these results were used to recommend the new set of stores based on missing categories.[6]

The census income data was utilized to categorize the locality of malls based on the average income of the region they are present in. The regions being utilized are smaller areas within a state divided based on population for the purpose of survey collection. These areas are called Public Use Microdata Area(PUMA). Using this income data, scores were given to all regions by dividing each PUMA's average Income by the Maximum average Income across PUMAs.
Using these scores for the malls present, each shop was scored by averaging out the scores of each malls they were present in. So the recommendations were generated by obtaining the income score of a new mall and matching it to list of scored stores.

5 Experimentation

5.1 Clustering

We used Association Analysis approach for clustering similar categories and stores using existing malls' store combination as training data. Here, we generated a list of categories for all the stores we had in all the malls. Using these categories and Apriori algorithm we could find scored combinations of stores occurring together and most frequently.
Based on these scored combination of categories, our recommender then generates a list of combination of shops in those categories along with the categories. This list of stores is generated based on most highly rated combination of stores in each category. A threshold of 0.5 was used to decide minimum number of times we want a certain combinations of categories occurring together. This means that any category combination needs to be present in at least 50 percent of the malls to come up as a probable recommendation. Also the top occurring store combination threshold was set to be around 0.04.
Using this setup we tried to do validation of the approach. 90 per cent of the malls data was used as training and remaining 10 percent as validation data.

5.2 Income based Analysis

Apart from the category based classification, an income based analysis was also performed. Here, we tried to obtain the income data for each census and map it to the malls using the Geographic coordinates. Based on income each PUMA was scored and eventually an average score for each

shop was also generated. These scores indicated the income patterns of people visiting the malls and stores. The data was again classified into 90 percent training data to come up with scores, and 10 percent validation. For testing, the recommender was given a list of existing stores and it was tested if it would be able to come up with the missing stores using the scores. For this validation, it was simply tested whether the scores for missing stores lied within the range of the malls income score (plus some margin).

6 Results

The Association analysis resulted in accuracy ranging from 65 percent. This accuracy depended on how common was the store to be predicted, across US malls. The algorithm was able to come up with more commonly occurring stores accurately most of the times, but if a store was only present in few malls, it usually doesn't show up in the recommendations.

The results for income based analysis were discouraging. If we took a margin of around 0.25, i.e. the recommendations be in at least within 0.25 of the mall's current score , we got only around 45 per cent accuracy. This means using income as a primary criteria to score may not be very helpful.

7 Website Guide

Current Server: http://linserv1.cims.nyu.edu:47152/
Current code structure: Flask based web application.
A folder FlaskApp with app.py which is a controller for the app. it renders the HTML files located in templates folder. It utilizes data of existing malls and categories from data folder.

Homepage: http://linserv1.cims.nyu.edu:47152/home It contains several check boxes of stores already known. It also contains check boxes for categories to be selected for the shops not in the list above. Both the lists are alphabetically sorted.

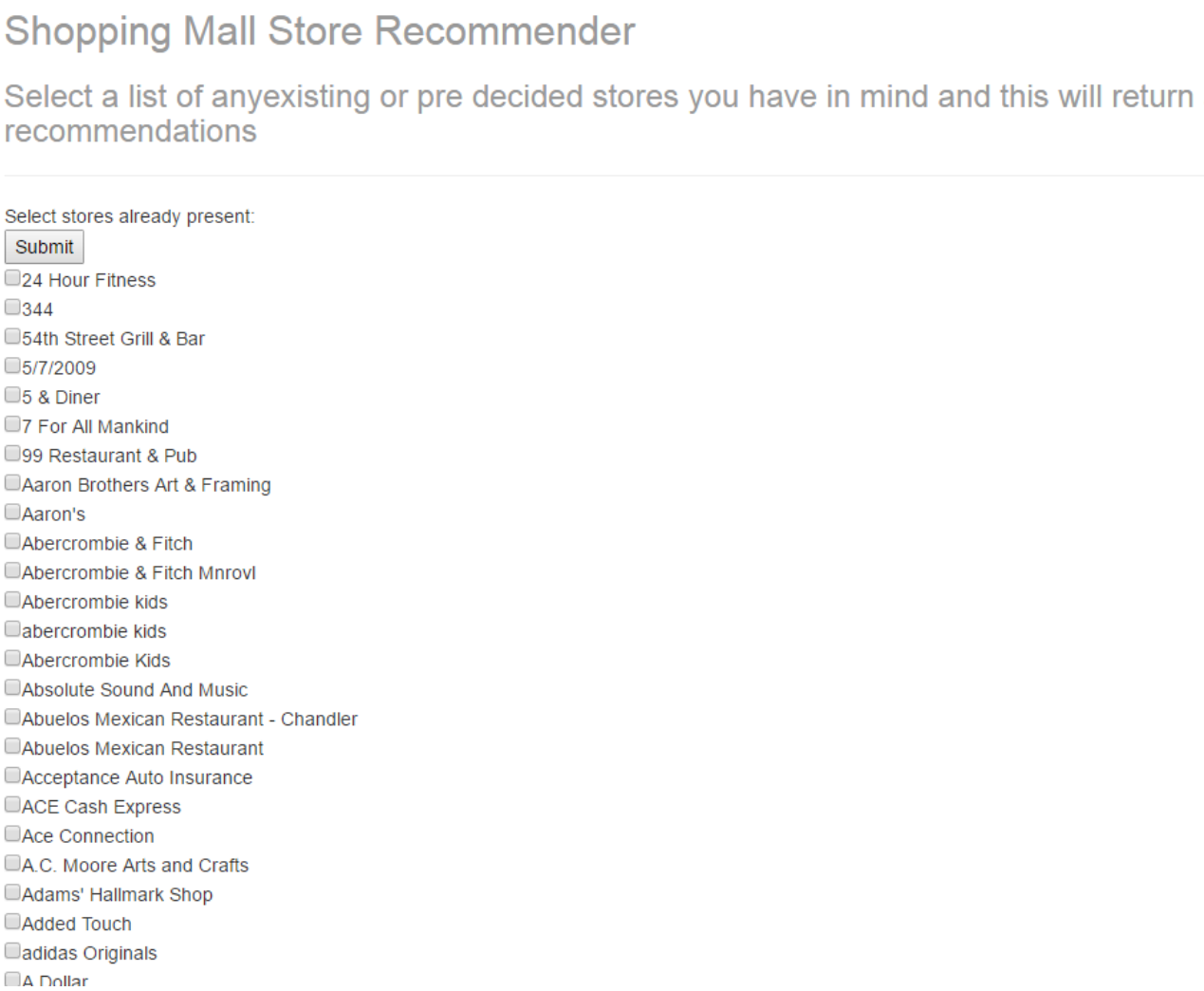

Figure 1: Home Page Top.

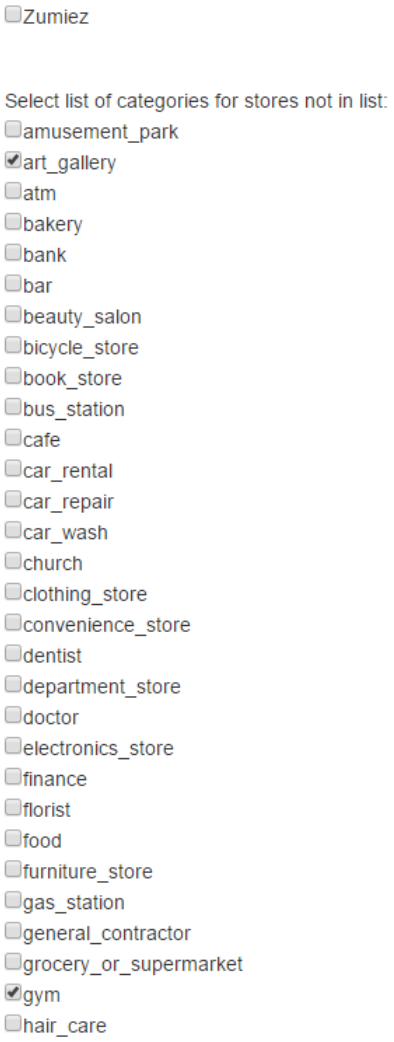

Figure 2: Home Page Bottom.

Results Page:
It displays categories and recommended combination of stores within each category

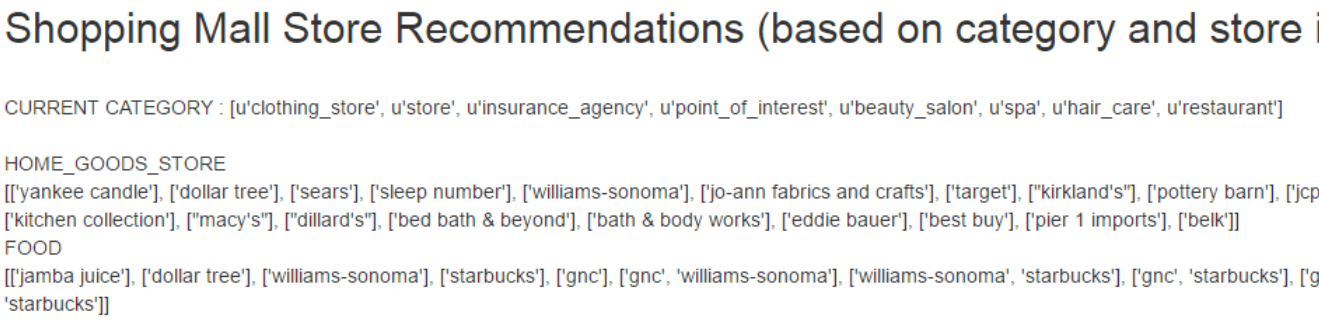

Figure 3: Results Page.

8 Challenges

One of the major challenge in the problem was data gathering. It was difficult to map mall addresses to PUMA codes since there is no direct mapping between zip codes and PUMA codes. Had to then match them using geographic coordinates. Also, absence of revenue or rent data makes it impossible to create an objective function to map the success of existing malls store combinations

Another Challenge in current system is that many stores might be present by multiple names, for example Apple Stores present at various locations includes the name of the place in their name. Hence, its frequency of occurrence might be slightly distorted.
Existing system is unable to take into account regional stores, which do not have many outlets across different malls, for recommendation. This is because it relies on frequently occurring patterns

9 Conclusion and Future Work

This project tries to tackle a conventional business problem from the technical approach of data based learning and finding patterns. Using the various data sources, we were able to build a website which can recommend probable store combinations based on existing mix in any Mall.

It could further be extended to try to arrange revenue data and use it to predict success based on optimal store mix. Additional analysis components would include purchase patterns based on geography to recommend region specific favourite store categories and to be able to apply collaborative filtering using some success metric for individual stores like ratings for the stores.

10 Acknowledgement

I would like to take this opportunity to thank Professor Dennis Shasha for the great project problem and his support during the entire project.

11 References

1. Carter Charles C., Allen Marcus (2012) ”A method for determining optimal tenant mix in shopping centers” Cornell Real Estate Review

2. Caraciolo Marcel, Melo Bruno, Caspirro Ricardo (2011) ”Crab: A recommendation engine framework for Python”.

3. Google Places API Web Service : https://developers.google.com/places/web-service/

4. American Community Survey (ACS) Data : https://www.census.gov/programs-surveys/acs/data/pums.html

5.Aggarwal Charu C., Zhai, ChengXiang ”A survey of Text Clustering algorithms”

6.Machine Learning and Data Mining: Association Analaysis with Python. http://aimotion.blogspot.fr/2013/01/machine-learning-and-data-mining.html