**CMPE 256**

**Large Scale Analytics**

**Individual Project**

**Summer 2019**



**Department Recommendation System**

By:

Name: Zaid Laffta

Student ID: 009689433

**Abstract**

Recommender systems are one of the most successful and widespread application of machine learning technologies in business, education, medical field and other important aspect of our daily life. The use of the recommender system increased rapidly in the past two decades even though it was used before that in much smaller scale. You can apply recommender systems in scenarios where many users interact with many items to help them make the right and informed decision. Moreover, during the last ten years, with the rise of big social media platform like Facebook and Youtube or by ecommerce websites like Amazon many other such web services, recommender systems have taken more and more place in our lives. From e-commerce where the website show you top rated items and what best fit your needs and search to online advertisement suggest to users the right contents, matching their preferences, recommender systems are today unavoidable in our daily online journeys. Now what is recommender system? In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users based on certain models like Collaborative filtering, knowledge base filtering, and Content base filtering. Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. As a proof of the importance of recommender systems, Netflix organized a challenge called the “Netflix prize” where the goal was to produce a recommender system that performs better than its own algorithm with a prize of 1 million dollars to win. This shows the importance of Recommender system in general and how important it is in our current time. Finally, in this project we will utilize Recommender System to show top college major/department according to ratings collecting through virous ways. The goal of this project is to demonstrate some of the important aspect in designing a recommendation system and what is the steps and the process to have a successful design.

**Introduction:**  
 Choosing a college major can be hard. Some students have known what they wanted to be since they were three years old; others aren’t sure, even with college right around the corner. The major an individual choose will neither predict nor guarantee your future. Many graduates find jobs that have nothing to do with what they studied in college. According to research published lately, the average twenty years old switches jobs once every three years and the average person changes career fields two or three times in their lifetime. Choosing the right major will prepare people for a specific career path or advanced study. Maybe you already know that you want to be a nurse, a day trader, a physical therapist, or a web developer and so on. Before you declare a major, take a class or two in the relevant discipline, and check out the syllabus for an advanced seminar. Make sure you’re ready for the coursework required for the career of your dreams. If you truly have no idea what you want to study and tried all the above without any result, let’s take the opinion of professionals and try to make sense of this opinion using Recommender System. Whether you’re already in college or you’ve just started applying having a recommendation system will help to narrow down your options. Better yet, getting this recommendation from people who are in the work force will help even more to security better career path and better future.

**This project:**

This project is simply applying recommendation in education to suggest top department rated by career professional’s opinion. The idea of this project is to collect data from people in the work force about their opinion regarding what is the best department/major. I picked the top 25 college majors/departments and created a survey that ask the user to rate each major from 1 to 5. The rating takes in consideration multiple factors like; life-school balance, career trajectory, and life-work balance after graduation. The data collection process consists of two different parts; online survey and in person live survey at SJSU and different places. Once we have the ratings, I used the knowledge acquired in this course and apply it to the data set and start analyzing the data and finding out what is the best department/major. My Recommender System simply a system that seeks to predicts the rating given by users then used to recommend a department/major to the people who uses the recommendation system. The highest rating mean will be picked as the best option. I adopted Collaborative Filtering-item base model to find best department that fits the user. A cosine similarity used to find correlation between each rating. Also, I validate the correlation by using a single program and compare it to all other programs. Other calculations were made like the mean for each between departments/majors, top rated department, and used graph representation to show different aspect of the calculation. Beside all that, I did calculate other aspect and factors to fully examine the data. Moreover, the implementation of this project was in phases. I started off with the initial phase which is understanding the basic need to understand what exactly freshmen in college want and how we can help them to choose the right major/department. I can summarize the phases of the project as the following:

1. Understanding the issue: I conducted multiple interviews with student and collected all information required to come up with the right model.
2. Data understanding: where I collected and organized data
3. Modeling: where I came up with model
4. Deployment: which was the implementation of the project in terms of writing codes and debugging
5. Evaluating the implementation by take feedback for users and modify the model according to results

**Architecture and Implementation**

This project has multiple different phase. In this section I will be explaining how I worked on this project and what is the difficulties/challenges I faced. Also, I will be explaining the architecture aspect of this project in terms of coding and what type of tools I used.

**Data collection and organization:**  
 The very first step in every Recommender system project is data mining. Before getting in to data collection details we should acknowledge that data collection is the hardest part in most projects, but specifically challenging in Recommender System and ML in general. Now, before we start off this section let’s define data and how we should deal with it. Data is a broad term for large and complex information like rating, financial history, scorer an any other objects that can be mathematically analyzed. There are several challenges one can face during this integration and collection of data such as data curation, capture, sharing, search, visualization, information privacy and storage just to name a few.

The biggest challenge in this project is data collection. Once of the most important things you need in this project is contextual data. These are lines of data points that include multiple variables and came from people in the workforce which will be a useful date as those individuals know which college major/department will be more important in the job market. I used different methods;

1. Live survey at my work place and school
2. Survey Monkey to get date by using social media and group email every present in my company, friends and family

Below is the link for the paid survey I send out on Survey Monkey.

<https://www.surveymonkey.com/r/RXZJLYM>

The result where great, I got great result from the in-person survey, but not so much from the online survey. In total I got 37 people to take my online survey while I had more then 124 to take in person survey between work and school. After collecting data, I started the process of cleaning data and make sure all the data is valid and after that I merged the data into a single CSV file to have it ready to process. This phase can be summarized as the following:

1. Data acquisition
2. Data preparation
3. Data validation

Validating data and making sure it is valid, updated, and without any errors representing the end of this phase of the project.

**Building Model, Ranking and Recommendation:**

If we want to show user A a new a certain department/college major, we’ll need to show him the ranking of this department. We can extend the recommendation system by ranking the recommended college major to user A. The intuition behind this is that if many of the similar users rated a specific department/major high, it is likely that User A will find it useful as well. The system will then recommend to User A the department/majors that rank highest on the list. I have done that by using the mean function and finding similarity between department. The department with the highest mean will be the top recommendation. See figure (1) below:

A screenshot of a social media post

Description automatically generated

Fig (1) Mean rating of all departments/majors

The important part of the program is that I used Collaborative Filtering item-based model to find the best department and show correlation between different departments using cosine similarity. See figure 2.

A screenshot of a social media post

Description automatically generated

Figure (2) Department Similarity

Moreover, a visual representation was an important part of the project. I used multiple graphs to show correlation between ratings and users, top rated department/major and number of rating per department. Figure (3) shows the mean rating per department.

A close up of a map

Description automatically generated

Fig (3) Department rating

**Other Calculation and Execution:**

In the section above I talked about the main aspect of this project in terms of modeling and calculation. In this section I will be touching upon the final steps I did, and other matrices used to have a strong functional Recommender System. After compiling the data and create a program to analyze this data. I used jupyter notebook to write down my program with the help of multiple libraries. Other matrices were taking in consideration to make sure our calculations are accurate and we have a comprehensive model that covers all aspect needed by the user to make a wise informed decision. Below are a summary of the main calculations and metrics used to have a strong and accrue recommendation system:

* Loading data from CSV file
* Transferring matrix data to be DataFrame
* Data display using program ID and number of rating
* Merge data with department ID to show what ID belong to which department
* Display program and number of ratings
* Find correlation between each program and other programs
* Find cosine similarly between all user ratings
* Display top correlated department/major to the first one
* Find mean, and median for each program
* Find top rated programs and recommend it to the user
* Show top five department account to the mean rating
* Display all programs according to the rating
* Evaluation part of the program where the end user gets to rate the program as an overall.
* Display graph of all ratings entered by user to rate this recommendation system

Building this Recommendation System was the most fun part of the project. Recommender system among the most popular applications of data science in our recent time. Almost every major tech company has applied them in some form or the other: Amazon uses it to suggest products to customers, YouTube uses it to decide which video to play next on autopay, and Instagram uses it to recommend pages to like and people to follow. The Recommender System in this project offers generalized recommendations to every user based on massive amount of data collected from professionals in the workforce. The final step in the program is focus on collecting customer feedback to see how this program helped people to find the right major. This evaluation part of the project works as a simple way to show user different statistic about feedback collected from previous users. Also, it has a graph representation that shows different statistic about the feedback giving by the users.

**Findings and Conclusion**

Recommender systems have great value in recommending relevant resources to users whether the target item is a book or a college major. It can be quite useful in finding a mode that can help us in all different aspects in life. The effectiveness of recommender system relies on the algorithm it uses to find similarities. For this project I used statistical analysis to find the best department choose base on the data collected. Also, correlation between departments help users to make an informed decision regarding their college major of choice. Moreover, I used collaborative filtering which can be represented in the context of this project as a domain-independent prediction technique for content that cannot easily and adequately be described by simply reading a college manual. Collaborative filtering technique can work effectively by building a database (user-item matrix) of preferences for department/major by users. It then matches users with relevant interest and preferences by calculating cosine similarities to make recommendations. A user gets recommendations to those majors that he has not rated before but that were already positively rated by others. Recommendations that are produced by CF can be of either prediction or recommendation, in the context of this project all those top-rated department/majors are pure prediction based on another individuals’ ratings. Those predictions are represented in a numerical value related to each department. While working on the project the following downfall in recommendation system has been found:

1. Lack of data will affect the accuracy of the recommendations: Perhaps the biggest issue I faced in the project and any recommender system for that matter is that they need a big data set to effectively make recommendations. The best recommendation systems in the world typical backed up with massive amount of data. To get good recommendations, you need a lot of users, feedback, surveys so you can get a lot of data to have an accurate prediction and results.
2. Changing of job market and user preference will affect the accuracy of the system specially in the context of this project. As time goes by job market tend to change and that will affect this recommendation system, or any system for that matter. Having updated data will help creating an accurate recommendation system.
3. Cold start problem specially with CF: where some department/majors are not rated or has a low number of rating
4. Dependency on user rating: without user rating the system is useless. Also, the bigger the data set the better and more accurate the recommendation system will be.
5. Data Sparsity problem: some department/majors will not get rated by users and that effected the overall quality of the recommendation system

For the context of this project, collaborative filtering was the best chose due to the nature of the data. A hybrid model will work but it will require more work in terms of modeling and data collection.

# How to Improve:

This system can be improved by building a Memory-Based Collaborative Filtering. This is something I run across while I was researching about models in Collaborative Filtering. In this case we’d divide the data into a training set and a test set, of course we will need bigger data set in this case. We’d then use techniques such as cosine similarity to compute the similarity between the all department/majors. An alternative is to build a Model-based Collaborative Filtering system based on matrix factorization. Matrix factorization is good at dealing with scalability and sparsity than any other forms. Finally, a hybrid model would work in this case, but it will required more work. In general, multiple models can be used in this system as long as the data set is big enough to result in accurate prediction and recommendations.