Building a Recommendation System for Banks

Capstone Project Public Summary

Kata Süle

Master of Science in Business Analytics

Central European University

2021

In today's business environment customer satisfaction is valued more and more by companies. Recommendation systems are one of the tools extensively used these days to improve customer satisfaction and as a result, increase sales as well as retention rates. They are mostly deployed in e-commerce settings but can be extended to any industry where the purchase histories of customers are available to the service provider. Therefore, my Client – a company in the financial sector – launched a project to build a recommendation system to assist one of their clients from the banking sector with reaching its goal of boosting sales. My Client believes that by deploying machine learning solutions, salesmen of their client will be able to provide better product recommendations to clients than the ones generated with their current practice. My tasks in this project were to provide proposals regarding the methodology, develop a data model and build a recommendation system using the chosen method which could identify products that a client would most likely purchase.

At the beginning of the project I collected methodologies which are currently used on the market. I could identify 4 main types of recommendation systems: Simple Recommender, Content-based Recommender, Collaborative Filtering and Hybrid Recommender. The most prominent differences between them were the input data requirements and the underlying models. Some only used item ratings data while others also included item metadata to generate recommendations. As for the models some used less complex techniques such as clustering, while others used more complicated ones such as singular value decomposition. The overall conclusion was that hybrid systems, that combined at least two different techniques, had the best performance. To evaluate the performances of the different approaches I listed 5 metrics. Some of these only measured whether the list of recommendations contained items that could be relevant to customers while others also put a focus on the ranks of these items within the list to validate results. Finally, I summarized a few use-cases from the banking industry. This made me spot the following: firstly, recommendation systems are not yet extensively used in the banking industry, especially not in corporate banking. Hence, companies in the industry who successfully deploy them sooner might gain significant competitive advantage. Secondly, Collaborative Filtering is less feasible in the banking industry due to the missing product ratings. Therefore, distance-based approaches such as clustering or k-nearest neighbour models were more popular to use in practice.

Based on the methodology research I decided to build a clustering-based recommendation system. To be able to do this I needed two datasets: one containing clients and the products they purchased and one containing information about the clients. I was provided with several datasets by the Client that I could use to generate these two final datasets. A challenge regarding the creation of the first dataset was that the pre-defined product categories for the three product groups that I worked with were either too vague or too precise or their descriptions were not available to me. Therefore, I had to generate new categories. These categories were defined based on domain knowledge for one of the product groups, while for the remaining two they were established based on characteristics of the data. At the end of this process I ended up with a relatively high number of categories. However, later on I decreased the number of categories to improve the quality of recommendations generated by the model.

To build the second dataset containing client information I used data exported from a subscription-based database to which the Client has access. According to the project plan I was supposed to enrich this data with information from publicly available sources. However, because I was not able to find any such sources the Client and me decided to stick with using only the data provided by them. From this data I opted for using the latest available versions of variables to make the model as up-to-date as possible. I picked features based on domain knowledge as well as some research and discussions with the Client. Furthermore, I also generated some features myself. A challenge regarding the creation of this second dataset was that the distributions of several variables were very skewed which had to be accounted for to be able to use clustering later on. It was solved by applying a commonly used data transformation technique. The whole data wrangling process was useful for two main reasons. Firstly, at the end of it I had the datasets needed to build the model. Secondly, by performing exploratory data analysis on the data the Client could have a clearer picture of the characteristics of clients as well as those of the products.

As the last step of the project, I built two models using clustering. For this I used the two datasets described above. In the first model I included all the product categories that I previously generated. The logic of the model was to first cluster clients based on their information, then generate a list of the top 5 most popular products for every cluster and use these lists as recommendations. I validated the results of the model using two different metrics which are commonly applied for evaluating the performance of recommendation systems. A further step was to compare the results of the clustering-based model to a primitive model where the recommendation for every client was the same list of the overall top 5 most popular products. I found that there was no difference between the two approaches. The reason for this was most likely that the same products of one of the three main product groups, due to their special nature, were among the most popular products in every cluster as well as in the overall top 5 list of most popular products.

Consequently, I built a second model where I excluded all the products of the above-mentioned product group. At the same time this also meant that I had to exclude those clients from the dataset who have only purchased these products before, otherwise I would not have been able to validate the results of the model. The performance of this model dropped significantly compared to the first one, however the clustering-based approach proved to be better than the primitive one. It is also important to mention that the performance of this model was much more realistic than that of the first one. At this point I experimented with a number of ways to improve the performance of the model such as adding more variables or excluding ones that seemed redundant. The one approach that worked was decreasing the number of generated product categories. The performance of this version of the model was also tested against randomly selected recommendations besides those two metrics that I used before. Since this model produced significantly better results than any other it was chosen as the final one. I believe that my results can be of use to the Client in two ways. The first one is that they can apply the same preparatory and model building steps in their own project. While the second one is that they can incorporate some of the logic of my work in their own recommendation system.

As for my learning experience, this project posed a great opportunity for me to get to know new methods which are extensively used at many companies today. Furthermore, I could also gain experience in researching topics and then implementing the solutions that I found. What is more, I learned that approaches that seem to be simpler at first might work just as well or even better than more complicated ones. It was also my first time working with raw data from the database of a company. Therefore, I could get an insight into the nature of this type of data and the challenges that come with it. Last but not least, during the numerous meetings with my Client I had the opportunity to practice my negotiation skills.