Product Category Purchase Prediction

CEU Business Analytics MSc

Ádám József Kovács

Faculty Supervisor: Eszter Windhager-Pokol

2022 June

Contents

1 Client introduction

The client of my capstone project was Hiflylabs, a Hungarian data consultancy company. The different teams of the firm have participated in multiple Business Intelligence projects in various industries for over 10 years now. The clients of these projects cover a wide range of sectors such as the financial, telco, energy, manufacturing or retail. The company delivers a wide portfolio of innovative business solutions including Data Warehouses, Business Intelligence, Data Analytics and Modeling.

2 Project description

It is of utmost importance for Hiflylabs to provide the best possible service and solutions for all clients from these various sets of indutries. To this end, they not only work on the projects coming in from the clients, but also run inner innovation projects on some of their clients data. The goal of such projects is to be prepared and develop some basic insights, look for challenges so that when future clients come with similar data and similar problems, they will be prepared to solve it faster and more effectively. I worked on such a project and used the data of a retailer with several stores in Hungary.

The key problem to be solved in my capstone project was building a predictive model that can tell on what days or weeks customers of this retail chain are going to buy from different kinds of product categories. The relevance of this project from a business perspective for future clients lies in allowing for more personalized targeting of promotions, product recommendations, or even churn prevention. Besides making correct predictions, identifying customers that behave in more or less predictable ways is also useful to learn more about the customer base's behavior. As for my clients perspective, they were rather keen on the exploration of the appropriate time aggregation for which useful models can be built, the kind of features that work best, and the overall feasibility of the project.

3 Project summary

This capstone project touched upon many of the topics covered during my studies at the Business Analytics program at CEU from data collection, cleaning and transformation to model building and evaluation.

3.1 Data and tools

For my project, two main sources of data were provided to me. The first is two and half years of historical transaction data from the retail stores at the level of transactions and product items as well. The second data source is a product hierarchy table with three levels above the product. In terms of external data collection, since the observed period of the transaction data (beginning of 2019 to end of May 2021) overlaps with the biggest waves of the coronavirus pandemic in Hungary, I collected daily new cases and daily new deaths due the pandemic from the website of a Hungarian news outlet (source). I also used an API to get all the public and national holidays in Hungary that I presumed could also affect consumer behavior (source).

The two main tools used in the project are Microsoft SQL, since the data from my client needed to be queried from a SQL database, and python. I used python and the jupyter environment for all my analysis, especially the pandas and scikit-learn packages were used most heavily in my project.

I would also like to mention that before the modeling part, several data cleaning (included only in the Technical report) and data transformation steps needed to be made to get to the analysis. The transaction data and the product hierarchy data needed to be merged. In the resulting dataframe, each record is a product that is part of a product category and a transaction by a customer. This raw data is grouped by customer and transaction date and the target of the models is a dummy of whether the customer purchased on these days (consumption records are filled up with 0s for days when the customers did not buy). Each customers consumption record is started from their first purchase recorded.

3.2 Findings

During the modeling part, first, my client told me to start with a simplified problem, predicting purchase of any product by the customers. Two time aggregation levels were tested in this framework, daily prediction, which is the original resolution of the data, and weekly prediction, which was achieved through aggregation. Then, I moved on to the problem defined in the title and built models predicting product category purchase.

In terms of feature engineering, several features were created from the sheer date information. From the serial number of the day and week in their respective years, dummies of whether the particular day is a weekend and for the seasons as well. As mentioned before, holidays, long-weekends and the covid situtation were also accounted for. Other important features were drafted from the consumption patterns of the customers, these are the days and weeks elapsed since the last purchase and the rolling window of the number of times people purchased in the past week (or month for weekly aggregation).

As for the design of the modeling again two kinds were tested. In the first one an 80-20 train-test split was completed on each individuals' consumption record and the prediction was either on the last 20% of his consumption record, or using 4-fold cross validation, other 20% segments of his record as well. In the second design, 80% of the customers was chosen randomly and their full record was used for training different models, and the rest of the customers were used for evaluation. Regarding evaluation metrics, the two main ones I considered were the area under the receiving operator characteristic curve (auc) and the accuracy. Further measures specificity, recall and the F1-score were also presented.

Turning on the results, the daily predictive models turned out to perform very poorly, not being able to surpass the baseline accuracy and improving only mildly on the random choice 0.5 auc value, thus I decided to concentrate on weekly predictions. The best models using weekly aggregated data on the simple purchase prediction in the first framework managed to make significant improvement to the baseline model,

especially the random forest and the classification tree with cross-validation proved to be useful. As for the most important features based on the impurity reduction, the best model showed that the date and calendar related features were not too important, the covid-related measures on the other hand contributed over 5%. The simple trend in the time series was also very important and so was the previous periods' total sales. But the most important feature (above 20%) was by far the weeks elapsed since the last purchase made. This shows that frequency of consumption is a very important predictor. Also, the most predictable customers were given additional inspection and it turned out that they either purchased a lot or very few times over the period considered, while the most unpredictable ones generally wait a long time between any two purchases.

Next, these models were also applied to one product category (personal hygiene). It turned out that the improvement in accuracy is somewhat lower this way, but is still there. As for the second framework, it was tested for the top 5 most purchased product category and different kinds of machine learning models showed the best performance for different categories. Personal hygiene and baby care consumption turned out to be the most predictable.

4 Next steps

There are several areas where this project has potential for improvement. One of the most obvious ones is enlarging the sample. I could only work with 1% of the available customers data due to computational limitations, and the higher sample size is very important for higher efficiency of models, especially since machine learning algorithms were tested mainly.

Another part of the analysis that could be improved, though I tried to put deep emphasis on it, is feature engineering. One of the additional features I would have created if I had the time include the calculating not just the quantity bought in the period before the one for which the prediction is being made, but also the sum money paid.

A further plan of mine was to extend my modelling from the classification problem of whether or not a customer will make a purchase from the given product category to a regression problem of how much they are going to buy from it each week.

Finally, I would also consider going further down on the product hierarchy and fitting models on these more specific ones like sun cream. This would provide higher business value for the clients, but also is a much harder task as consumption of it is rarer.

5 Summary

In the retail sector, it is of very high value and importance to predict customer behavior, especially in the case of fast-moving consumer goods. In my capstone project, I used real-life data from this sector to uncover patterns and build models that my client can use to provide better service later on. It was a great technical challenge and I learned a lot while going through the whole data science process. I also learned a lot about the retail sector and its dynamics and I will try to complete some of the ideas in the Next Steps section if my client shows further interest.