Casptone Project Summary

Ágnes Kollaine Stark – MS in Business Analytics

Executive Summary

The aim of this project is to support an e-commerce company in lighting business with marketing and business decisions, by conducting a thorough analysis of their sales and transaction data. The client is a fast growing Hungarian company founded in 2012, operating one of the top webshops in the region. It supplies a broad range of lighting products for homes, cars, outdoor etc. The venture is continuously improving its operations for keeping its leading position in the highly competitive market. The goal of the project is to generate insights and recommendations for sales and marketing through the analysis of accumulated customer and transaction data.

The Data

I was very lucky to have received great sources from my client, in the form of a Google Analytics page which provided plenty of information, accessible data and option to customize the data to my needs.

With the help of the Query Explorer and the api, I was able to pull all the necessary data to R, using the GoogleAnalyticsR package.

For the second and third part of my analysis, I used the order transaction dataset of 2020 which was also provided by the client. In both cases, naturally some data cleaning and feature engineering steps were required, however it was nearly not as complex as it is usually with real life case studies. I was very lucky in this aspect, because I was able to focus more on the analysis.

Summary of Parts 1,2,3

My analysis started out with analysing the device usage during the Black Friday Campaign – which is one of the most important parts of the year for my client- whether there is a change in user behavior compared to previous years and compared to yearly average.

I continued with the analysis of sold quantities and total revenue coming from the Black Friday Campaign compared to a different weekend during the year. I chose a comparison weekend based on the information I received from the client about seasonality of sales. I summarized the results of both weekends, both in terms of quantity sales and revenue, and made some conclusions about the type of products that bring higher quantities than usual, and higher revenues than usual, during the promotional weekend. I also identified price ranges, and applied that to both time periods to check the differences in prices, both nominal and in percent. This allowed for conclusions on the effects of price reductions in the various price ranges.

I continued my analysis with a Market Basket Analysis which uses transactions data to identify products that are purchased together. I was very lucky as the data I had contained observations within orders, in each row. This was a perfect base for a transactions dataset, which is crucial to have if one is to use the apriori algorithm. Although I spent very much time on converting the dataset to the correct format, the algorithm was only able to find 4 rules, which actually pointed to only 4 products, purchased in a different order. Having given it some thought, I realized that a monthly breakdown would have helped to identify pairs of purchases better, as the monthly data would have allowed for seasonality differences in product purchases to show up. After the market basket analysis, based on my client's request I identified products that were put into the basket first and products which were put into the basket as last. My client was interested to see what are the main characteristics of these products. I checked the price distributions, the brands, the product categories, and the average discount per month for the product types in mention, and managed to find some conclusions.

The third part of my analysis was creating linear and multivariate models to identify relationships of variables in the data that have an effect on the gross price of a transaction. For this, I first cleaned and transformed the orders dataset, and picked the variables of interest. I conducted descriptive statistics steps, created histograms and scatterplots for the continuous variables and boxplots for the categorical ones. I also checked a correlation heatmap, to see the correlations between the variables. I continued with the actual modelling, creating a linear model first, then adding in more variables, for several multvariate models. The R-squared value of the model increased as more variables were added to the regression. The p- values in most cases were significant, below 0.01, however in some of the results, I recevied not significant estimates as well. Lastly I added two models with interactions. These interactions allowed for some improvement in the models, boht models with interactions ended up with an R-squared value if 9 %. I was able to highlight some of the relationships of variables in the data with the help of the models, and the interpretation of the estimates.

Conclusions

I listed a number of summarizing conclusions at the end of my technical discussion, and some recommendations to the client. I am also aware that there are many areas in which an improvement can be made to my analysis. In terms of the data, it would be very useful to conduct the market basket analysis on a monthly basis, for which the scope of the project did not allow, as I would have exceeded 50 pages in that case. The data could be sorted in a way where one observation is one product, and the analysis of product sales could then be conducted. I refrained from doing that, as it would have taken me weeks to convert the dataset in such a format, and I would have lost a lot of order-related information that was crucial for the second part of the analysis. The product categories were missing in a number of cases, which was rather sad, because it did not allow for major conclusions about the product categories. The modelling part should definitely be

extended with predictions, by setting up training and test data. Some more accurate models could be built using more sophisticated models such as deep learning models. I chose not to do so for two reasons: my experience with machine learning tools is rather limited, and the interpretation of black box models is not as self- explanatory as with regression. In my analysis my aim was more to identify various relationships in the data, rather than to make precise predictions.

Thank you note

Finally, I would like to thank my client for providing the outstanding quality data, for the recommendations and comments, for the positive approach and the opportunity to get to know the company and its products better.

I would also like to thank Professor Miklós Koren for supporting me to apply to the program, and Professor György Bőgel for his supportive approach during the past years.

Most importantly I would like to thank Eszter Fuchs for her amazing support during my years of study, and for listening to all the recommendations I made about the program.