Cover Page

Capstone Project Public Project Summary

Student Information

Name: Gerold Csendes

Email: Csendes_Gerold@student.ceu.edu

Student ID: 1901724

Program: MS Business Analytics

Supervisor: Prof. Eszter Somos

Abstract

The following public report summarizes the author's Capstone Project that was implemented as a part of MSc Business Analytics at Central European University, 2020. Client is a tank station chain in the CEE region, operating hundreds of stations internationally. The decision on setting the working hours had not been backed by data before this project. This may lead to inefficiencies if operation hours are not set in an optimal way. Cutting down on them may lead to lower operation costs, broadening them may generate more revenues.

The author's task was to clean and integrate several data sources, so all necessary data is available for the analysis. Then, as a part of a data exploratory process, customer purchasing patterns with explicit focus on their time component was explored and business could better see how their financial metrics change by the hour. Finally, a simulation framework was implemented to derive the optimal set of operation hours for the tank stations in a mid-scale town in Hungary.

Tartalom

Capstone Project Discussion 3

Extensions and deployment 4

Capstone Project Discussion

In the first phase of the Capstone Project, numerous data sources like operational, financial, transactional etc.. must have been integrated and cleaned to provide business insights on how their units perform financially broken down by the hour (or even in finer scale). The data pipeline was implemented in SQL, Python and R; views, temporary and permanent tables as well as logs were generated where necessary. The goal was to provide business with a semi-automatic tool which – with proper parametrization – can return the financial data to support data-driven decisions.

On the one side of the data pipeline were the data sources, while on the other, the analysis outputs. Due to its exceptional graphics abilities, R was heavily utilized for generating figures that captured the important aspects of the data. With some manual work, business can now generate with low effort insightful data visualizations. Should a demand exist for a "self-service" tool, the framework implemented is perfectly suitable for such developments.

Monitoring how business performs is crucial but it is equally important to make actionable recommendations. However, it is not that straightforward when it comes to a complex business decision like optimizing network-level working hours. Indicative examples for the complexity of the problem may be: 1) there exists many combinations of working hours 2) competition needs to be accounted for 3) customer patterns change by segment 4) customers are moving. As an example, consider a filling station that closes one hour earlier than usual. What may the customers do? They may: 1) return at another time when the unit operates (substitute in time) 2) go to another tank station within the NETWORK (substitute in company network – stay loyal) or 3) choose a different company (churn).

One can see that the customer decision-making process is not straightforward that is influenced by its preferences. To simulate how different working hour combinations financially perform, one needs to model this decision-making process. For a successful implementation, the preferences of the customers must be modeled. This requires heavy domain knowledge and business assumptions. Without this, there is not much chance to perform such a simulation.

Luckily, Client had had such business assumptions and by providing some uncertainty to their assumptions, the real-world process of customer's choosing a tank station may be modeled efficiently. From this on, it is only the matter of coding workload to implement this simulation framework. One shall not underestimate the workload even with the easiest seeming tweak of

the simulation process. To fit into the scope of the capstone project several simplifications were used that seemed to be worth the tradeoff.

Such simplifications were used as defined in the following. (1) Only consider decreasing the working hours, so that we have data on what customer traffic to expect. When closing a unit for example one hour earlier, we know with some uncertainty how many customers they would expect in that hour. We can assume, that those customers still need gas, so they would be out looking for other options. (2) To model the movement of customers, simply distribute them within a defined distance from the tank station. (3) Distance, specifically travel times, play a key role when customers need to decide on which tank station to choose. Travel time is time-consuming to fetch, thus aerial data is considered as a substitute.

One can see that these assumptions oversimplify reality, but this is how simulations are usually designed. One cannot account for a real-life like complexity. Having this simulation framework up and running, it is only a matter of time until one concludes the optimal set of working hours. However, one should not underestimate how computationally expensive such a simulation process can be. This stems from the following factors: 1) permutations (of working hours per tank stations) increase extremely fast 2) each permutation needs to be run many times to outweigh outliers as a result of pure randomization 3) the possibly huge number of decisions to simulate.

As a result, one needs to limit the simulation scope where possible: the number of working hours permutations or the number of business units (the geographical area) to simulate. The author utilized business and legal constraints to limit the number of working hour permutations and also focused on a geographical area that fit into the Capstone Project Scope.

By implementing all this, the author was able to deliver important business insights for Client. Data visualization helped business believe that the simulation process works as expected, and the especially the assumptions are implemented as needed. Results consistently showed how the working hours of some business units may be changed to operate more efficiently.

Extensions and deployment

As it was point out in the previous sections, this simulation framework applies many simplifications, but it does not mean that complexity cannot be increased. Of course, there is a tradeoff to this, labor and computational expensiveness, especially the latter but if business considers it valuable then some simplifications may be eased. Aerial distances may be dropped

in favor of travel times that is much more life-like. This would require probably a paid navigation service implementation into the simulation framework. Runtime may be decreased by using more advanced coding methodologies, like parallelization and/or using a more powerful machine. This would allow to simulate more working hour combinations. Finally, it may be worth including the ability to track wait times at filling stations. It may be the case that the simulation advises closing one unit, thus the nearest unit's traffic would increase. This is all good up to the point that the new unit may get so crowded that wait times increase too much and customers may get mad and rather churn. By extending the simulation with these components, we may get a better picture of the true processes.

Finally, a few words on the deployment of this framework. In the author's opinion, this use case doesn't suggest for a deployment in the 'usual' sense when the code goes live and is used extensively by many. This simulation framework may rather be utilized periodically, so the code does not need to be engineered just as well as a live software but it still needs to be reliable. The author informed business that the code may be made more effective and definitely needs further development if business is to use this tool more than a couple of times in the near future.