Analyzing the Effect of Customer Cannibalization by the Opening of New Sites: Causal Analysis

Capstone Project Summary for CEU MSc Business Analytics 2022

Mahrukh Khan

Contents

Description of the project

Companies face uncertainty when making the decision to construct a brand-new store in an area where one of their stores is already situated. My company finds themselves in the same position. Even though it belongs from the retail industry, selling items of necessities that will always have a demand, it could face lower performance in one of its store due to the changing behavior of customers who are attracted to shopping at its newly opened stores. But at the same time, store openings are important for the company to capture a larger share of the market. The impact the new stores could have on already existing stores is called the effect of cannibalization. Cannibalization will be studied by taking the difference of the change in number of customers in two type of store categories: stores that had new stores opening in nearby distance in a defined time-period versus those that did not have any new store opening throughout that time-period. This project aims to explore whether the company faced the effects of cannibalization over the years by gauging the trend in monthly average number of customers through causal analysis. It will use event studies method to study the effect of opening new stores, our causal variable, on the subjects that are the previously existing stores.

Summary of the work done

Databricks, a cloud-based platform, was used to explore and analyze the data. Many data frames had to be merged for getting the final datasets that could be used for the analysis. Firstly, the data

was loaded as a sparks data frame and had to be converted to pandas for thorough data cleaning. European datasets can have decimals being represented by a comma which had to be converted to a point, some columns required their type to be correctly specified, and whilst calculating distance between stores, the distance with the store itself – a zero- had to be dropped otherwise it would falsely show that the store gets cannibalized by itself. This was followed by data exploratory analysis, in which data was subset for the first and second half of the project based on relevant years to be studied. Distance was used as a measurement for a store being impacted by cannibalization. The terms specifying a store being impacted were selected using a certain criterion for both distance and time period. Using both these measurements, the unimpacted stores were also defined.

The analysis to study cannibalization had been divided into two parts. In the first part, a detailed exploratory analysis was conducted to make the decision of finding a suitable six-month period, taking into account all the years, to best capture the optimum number of stores impacted by building new stores. It also balanced the selection by considering an optimal number of stores that do not have new stores opening nearby in that time period for precise comparison. The performance metric was based on the change in the monthly average number of customers visiting the stores, as well as on a three-month average. The comparison was made whilst looking at the performance of impacted and unimpacted stores six months before and after the intervention of new stores. The results were supported by visuals as well as a fixed-effects regression.

A second part of the project was introduced due to the implementation of the Hungarian Law: resulted in a significant reduction in the number of stores opening in recent years. This automatically caused a bias in not selecting the recent years as they did not provide enough data points for a six-month coverage period defined in the first half of the project. Hence, in the second part, to study cannibalization, the intervention or period in which a new store is established near another store was manually selected. Three sets of two years were selected. The steps taken were the same as mentioned in the first half of the project. Only the intervention period of new stores varied, but the performance for both categories was similarly compared six months before and after with visuals and a fixed-effects regression.

Key outcomes

The findings from the report bend towards the possibility of considering cannibalization as an important factor that contributes negatively to the affected stores' performance. The first part of the project captured a strong differential between stores' performance in both categories after the intervention. In comparison to impacted stores, the unimpacted stores had approximately a doubled percentage point increase in the average number of customers six months after the intervention period compared to their performance six months before the intervention. The interaction terms in the regression also showed how belonging to the impacted group had a negative impact on the three-month average number of customers studied in the two intervals after the intervention of new stores. Similar results were found for all three sets of years analyzed in the second part of the project. The unimpacted stores always had a better performance, that is the change in the average number of customers visiting the store was higher than the impacted stores after the intervention period. All three fixed effects regressions had a negative sign associated with the interaction term of belonging from an impacted group and the average number of customers. The repeated findings in various years tend to give validity to the possibility of cannibalization and its impact on the stores.

Benefit to the client

This report is the steppingstone in addressing the company's concern, as this is the very first time, they have tried testing the impact of this phenomenon. The department making decisions in this field can find valuable insights when selecting the location for a new store. This finding can also be integrated with their key performance metrics, to find solutions to counter the result of cannibalization. Moreover, the deep dive in the exploratory data analysis phase will not only enable the company to view this problem from multiple perspectives but also develop this analysis further. The criteria set for considering stores cannibalized is provided in such detail that it is easily understandable and adjustable if the company wishes to change the range for distance or time period. The entire code for the project is deftly accessible for the company as it's uploaded on an IPython notebook on Databricks. The code is made reproducible by using multiple functions, hence making it easy to tweak for future analysis. The method for event time studies is very clearly laid out and the fixed effects regression is straightforward and compliments the visuals. This makes it uncomplicated for all managers, especially outside of the data team, to understand this project.

Learning and experience

This project challenged every aspect of the lessons I had learnt during my master's. It was a unique topic from the list of projects we had to select. I chose it simply because it seemed like an unknown territory that would push me out of my comfort zone. It strengthened my capabilities to do extensive decision-making using both domain knowledge and data outputs as it heavily required data exploration before indulging in the analysis. Since this was the first time the company was testing out this phenomenon there was no prior work I could refer to. This was a delightful challenge as it gave me more authority over the work I was conducting. The company started us off with two field visits to their warehouse and stores, familiarizing us with every detail of how business was conducted. Multiple sessions were kept providing us with information on the business processes and functioning of the organization. This allowed me to learn about the retail industry not only in discussion but also through fieldwork. As this was done before the projects were introduced, it allowed me to make an informed decision with the selection of my project. On the technical aspect, this project deepened my skills in using the cloud-based platform Databricks. This was the first time I was conducting such a lengthy, detailed project on it. This project sharpened my python coding skills to a large extent. From data aggregation, cleaning, exploration, visualizations, and statistical analysis – all was done in python programming. It refreshed my statistical knowledge and gave me an open end to choose what I saw fit. Setting the various criteria for defining cannibalization was a very challenging part of the project and demanded intensive data exploration. This taught me how significant it is to familiarize yourself completely with the data before making any decisions. Through mistakes and errors, I learnt that the more I explored the data the more my issues became solvable. Furthermore, this project disciplined me on how to conduct myself with such a renowned company through multiple meetings and a presentation.