Capstone Project

Public Project Summary

Business case / task description

The client company

My client company was an online start-up which is running a price comparison site for courier services. The company acts as a broker between senders and large courier service providers and as it channels large quantity of orders, it can bargain better prices with the delivery companies. It has been operating for multiple years now, but the company has reached a mature state and they observed slowdown in their expansion recently, and they were seeking areas of improvement for growing the customer base. They have not used direct marketing extensively so far, but they are planning to increase the usage of this marketing tool in the future. The pricing of the service is currently based on the freight volume and distance, and is event-based, it doesn't consider the customer lifetime value (CLV) or the purchase history of the client with the company, neither does it contain any recurring, subscription-based element.

The aim of the project was

- to be able to identify patterns in the purchase history of the clients,
- create a metric that would cut the customer base to strong and ad-hoc purchasers
- and measure customer churn that will enable the calculation of customer lifetime value for the customer base in order to be able to assess the maximum allowed expenditure for marketing activity per client.

Data description – data cleaning

The processed data came from the data warehouse of the company extracted by an SQL query and it contained transaction level data for the past 2 years going back from September of 2018.

As the starting dataset was available on transaction level, I had to aggregate it to client level. Multiple fields were referential data about the individual transactions (direction, source channel of order, etc.). When aggregating to client level I assigned the value of each categorical variable to the category where the highest number of transactions was made. Numerical variables were just simply summed up to get the client level figures.

A specific request was to segment the data to business / consumer senders and recipients (B2B / B2C / C2C / C2B deliveries). Currently the website does not collect the legal form of either the sender or the recipient, so in order to make this segmentation I had to rely on the name of the sender and the recipient. This information is entered to a free text field when an order is posted. I processed the list of recipients and sender manually in Microsoft Excel and identified 68 different keywords that were marking business organisations. Additional information could be used on the senders' side, as if there was any tax number given, I regarded the customer as a business customer.

I have also visualised the most important metrics, and checked the distributions of the number of purchases, the number of months when a client performed at least one purchase, the revenue per purchase, and the revenue per customer. As with most often with business data, these measures were found to have highly skewed distributions.

Analysis

Features extracted

For the analysis of purchase patterns, I have built the monthly purchase history of each client ID that made at least 1 purchase between September 2016 and September 2018, by flagging each month where at least 1 purchase was made by the client. I have concatenated the flags to strings for the 2-year period and also for 2017/2018, where each digit showed if at least a purchase was made. Then I have aggregated this by simply counting the number of clients that were showing the specific purchase histories. Based on this I could examine which patterns were the most typical for the largest number of clients and I could also take a look if there was any recurrence in patterns.

I have also prepared the time series of revenues for each client by months and quarters for the examined period, and I have also calculated the average number of days between purchases.

As the business is highly seasonal and lot of transactions happen before Christmas time, I have introduced a special flag to mark those clients, who have been making their first transactions around Christmas and were not part of the customer base ever before.

Models used

As one of the aims of the project was to implement a ruleset to create a metric that would define strong customers, interpretable models were preferred compared to black box models. I decided to implement decision tree models as those models perform well on datasets where association between variables might be non-linear.

After visualising the purchase patterns, it was apparent that only a smaller proportion of customers are returning on a monthly basis, significant portion of the customers are either one-off buyers, or make transactions on 2 consecutive months. I also drew a correlation plot that was visualising the correlation between subsequent months regarding purchases, and I've found that there is a slight correlation between months, that was decaying on larger time periods: if a client has made a transaction in a month, than it will have higher propensity to return the next month, but as time progresses, this propensity decreases.

Based on this I've decided to create a simple metric, that would classify those customers to be strong customers, who had made purchases at least in the previous 3 months consecutively.

To validate my decision, I analysed what predicts a purchase in a month with using decision trees. I have built regularized decision trees to predict if a purchase would be made in a month. Best predictors (cuts on the data) were the revenue on the client in the prior month and the revenue on the client in 2 months before. I've built models for different months and based on that this connection seemed to be stable. This underpinned, that only short-term purchase history of the client counts when trying to predict client return, other variables, like direction or sales channel didn't seem to have notable effect.

Conclusions

Based on the projects my proposals were the following:

- My primary proposal was that the company should focus on retaining those customers that were showing at least 3 months of continuous purchase history with the company as those are the customers, that exhibit recurring need for the service and their contribution to the revenue base is significantly larger than those customers' who don't show this pattern. Yet client churn was still found to be significant among these strong clients therefore they would be good candidates for customer retention marketing efforts.
- As there was a recurring reporting need for showing data by business / consumer senders and recipients (B2B / B2C / C2C / C2B deliveries) that was not supported by the current setup of the website of the company I suggested to add this information request as a separate field to the order form of the transactions made by clients. This would make reporting more simplified making a significant effort in data cleaning unnecessary and this would also make reports more coherent.