**Improving Customer Acquisition Cost with Xgboost Classifier**

Vasudev Darshan Trivedi**1**, Akash Vasanthan**2**, Mudita Garg**3**, Fabiha Amir Lodhi**4**, Srivyshnavi Tangellapelli**5**

[vasudevtrivedi@my.unt.edu**1**](mailto:vasudevtrivedi@my.unt.edu), [akashvasanthan@my.unt.edu](mailto:akashvasanthan@my.unt.edu)**2**, [Muditagarg@my.unt.edu](mailto:Muditagarg@my.unt.edu)**3**, [FabihaAmirLodhi@my.unt.edu](mailto:FabihaAmirLodhi@my.unt.edu)**4**, [srivyshnavitangellapelli@my.unt.edu](mailto:srivyshnavitangellapelli@my.unt.edu)**5**

**Abstract**

*The aim of the report is to help maximise return of investment (ROI) for Super Server Company(SSC). The investment will be an upcoming marketing campaign to target potentially high conversion to sales customers. Given the limited budget, the machine learning model created will be vital to maximise success rate. Based on historical data, we identified potential levers that impact purchase and finalised on an ensemble learning model, XGBoost.*

*Due to data sparsity in numerous columns, the model achieved an ROC-AUC score of 72% with a good rate on true positives. The team feels that SSC should kick start the campaign focusing on medium size companies which are decision headquarters that have technical personnel which have interacted SSC’s paid media or made web inquiries about networking products.*

**Introduction**

Technology has drastically altered the product to consumer lifecycle. It is creating customer choice, and choice is altering the marketplace. Consumers are equipped with an arsenal of information at their disposal resulting in marketers having to constantly improve their approach. Since the 1990s, marketing has evolved in tangent to technology and in today’s information boom, marketing is everything and everything is marketing [1]. A successful marketing strategy depends on the marketer’s knowledge, insights, and ability to amalgamate the customer and the company.

Having an amazing marketing strategy but targeting an audience that is not interested in the product potentially results in huge losses. For simplicity, we classify the losses into two main buckets. Primary being the direct losses due to revenue invested directly in marketing and secondary being the loss due a decrease in sales of the marketed product/service. A general rule is that B2B companies spend about 5-10% of their revenue on marketing [2]. While it may not seem like a huge amount, the impact varies drastically depending on the type of business. Therefore, targeting is as crucial as the marketing campaign itself.

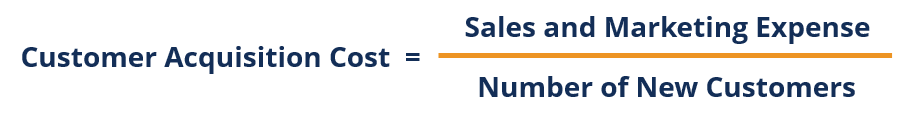


Figure 1. CAC Formula [3]

Keeping the above paragraphs in mind, the primary purpose for our team will be to help identify the right target audience that will be receptive to the SSC’c marketing campaign. We have been informed that the team is working on a tight budget therefore making it critical to have a potentially high success rate. We feel that a good success metric for both the marketing and data science team will be given by the customer acquisition cost (CAC). A lower CAC would essentially result in a good future revenue stream and higher profit margins for SSC.

**Methods**

The primary tool that was used was Python and the various machine learning packages, ranging from Pandas, Sklearn & BorutaPy

Any machine learning model is as good as the data provided to it. The aim of the model is to predict the target variable of *‘purchase\_event’*given numerous other independent features. We have been provided with two datasets, *training\_dataset* and *testing\_dataset*. The training data will be used to train the models and tune the necessary parameters while the testing data will be used to predict events to identify potential customers for the marketing team.

The first step of the data preparation was a general exploratory data analysis to gain a general understanding. The first key observation was that the target variable was extremely imbalanced with 95% of 0s and 5% of 1s(purchase). We also noticed two other columns indicating past purchases. Looking at the values, we noticed an anomaly as by definition overall purchase should be inclusive of past networking and cloud purchases. Therefore, we rectified this by ensuring that purchase\_event was correctly reflecting this and the target imbalance improved to 80% of 0s and 20% of 1s.

|  |  |  |  |
| --- | --- | --- | --- |
| Purchase | Purchase\_event | Past\_purchase\_networking | Past\_purchase\_cloud |
|  | Overall purchase from "the company" or not | Purchased a networking product from "the company" or not | Purchased a cloud product from "the company" or not |
| 0 | 0.51% | 16.2% | 0.30% |
| 1 | 94.9% | 83.8% | 99.7% |

Table 1. Discrepancy in target column

Out of the 33 independent variables, there were 5 numerical columns and remaining were Booleans. Column *‘persona\_tech*’ had missing values but in a small percentage. Furthermore, apart from a few columns, most of the Boolean columns were also skewed towards 0s. We identified this lack of information as a potential limitation when training the model.

We split the training data into two subset, training, and validation. The split was stratified on the *‘purchase\_event’* column to ensure both sets have enough instances. Deeper analysis and changes are done on this training\_data and the parameters for these changes and imputations will be used to transform the validation and testing set to prevent data leakage.

Running the Pearson Correlation matrix for the target variable shows relations with independent variables such as ‘*paid\_media\_activity’,’web\_activity\_networking’ and ‘persona\_tech*’ had the strongest correlation. Instead of just dropping the 1% of missing rows for *‘persona\_tech*’ which could result in further loss of the scare target response of 1, we imputed the missing rows the mode.

Within the numerical columns, we had three columns indicating projected budget. Checking for the variance inflation factor showed that *‘it\_spends\_cloud’* and *‘it\_spends\_others’* were highly correlated, therefore we dropped them altering summing up the two columns as ‘*it\_spends\_cloud\_others’.* Histograms for the numerical also showed a positive skewness with lots of accounts with a budget of zero and some with very high budgets.

Further investigation showed, 6% out of 20% that made a purchase with the company had zero for projected budget. Therefore, we did not drop these rows. To remedy the positive skewness, we did a log transformation for the numerical columns.

We also explored the possibility of merging different columns with skewness towards zeros based on business intuition such as product specific grouping of networking vs cloud.

**Results**

Feature importance was firstly identified using Boruta algorithm fitted with one instance of Randomforest Classification and another of XGBoost. Boruta selects features based on by creating shadow attributes of the original variables with respect to the response variable.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated



Figure 2. ROC-AUC Curve & Score & f1 score

SSC has two main branches for sales of its servers, one being for networking and other being for cloud. Majority of its past sales come from networking and therefore evidently, customers that interacted with paid media, had web activity in relation to networking to have a higher tendency to purchase the product. Overall, cloud played a minimal role in influencing the algorithms and given the existing data we feel that targeting customers interested in networking does help.

Business that are medium - size with about 10,000 to 50,000 employees, with technical personnel and decision headquarters should be targeted first.

**Conclusions**

In conclusion, the team had firstly identified networking to be a major revenue source for SSC. With a constraint on the marketing budget, we would firstly recommend narrowing the scope of the campaign theme to be around servers for networking needs. Based on that, instead of casting the net as wide as possible we would recommend to strategically target medium-size business engaging technical employees who had previous interaction with paid media communication via emails.

We feel that the limitation the approach was the lack of variation in the dataset. We had hypothesised segmenting the customers by industry type first however the industry column was skewed toward financial services with numerous overlaps with other industries.

To have a more robust model and successful campaign, we would recommend the team to collect more data especially for variables that driven by business intuition such as decision power, industry type, paid media and projected budget.

**References**

1. McKenna, R. (2014, August 1). *Marketing is everything*. Harvard Business Review. Retrieved February 20, 2022, from <https://hbr.org/1991/01/marketing-is-everything>
2. *What is an average marketing budget for a small business?* BDC.ca. (2021, August 27). Retrieved February 20, 2022, from https://www.bdc.ca/en/articles-tools/marketing-sales-export/marketing/what-average-marketing-budget-for-small-business
3. *Customer acquisition cost (CAC)*. Corporate Finance Institute. (2022, February 20). Retrieved February 20, 2022, from https://corporatefinanceinstitute.com/resources/knowledge/accounting/customer-acquisition-cost-cac/