CEU Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

Business Analytics Program 2022/23

Capstone Project Summary

CUSTOMER CHURN ANALYSIS

Data analysis and class prediction on GA webshop data

Gabriella Zsiros

Overview

The project was assigned by a digital consulting company with the aim to gain insight on customer churn of their clients based on webshop data. The company was able to provide the data for one of their active clients which they intend to incorporate in their collaboration with the client.

In this case, "churn" is defined as the loss of potential customers who could otherwise make a purchase. This data then can be used by the firm to target certain users for remarketing campaigns.A potent solution to this problem is to build a predictive model using the available data. However, this comes with the challenge of identifying the appropriate metrics and features.

The project kickoff took place on 28th March, and initial data was available from 2nd May. As the firm's client companies conduct business on online platforms, they utilize Google Analytics as a fundamental tool for business analytics.

Data description

The data to be used for the predictive model is provided exclusively by the client firm and is sourced from Google Analytics. GA offers several dozen views and dimensions on its user interface, but API query methods provide a customized way to extract data, with over 200 data features available. Utilizing a custom model not only offers a tailored solution to the client's requirements but also helps determine key features that can contribute to a model with potential universal applicability for the consulting firm's other clients.

Before accessing the data, I familiarized myself with Google Analytics platform and collected relevant articles from online source in the topic.

Given the time constraint of the project, especially from when the data became available my personal approach was to incorporate agile methodology and aim to deliver the first iteration of the

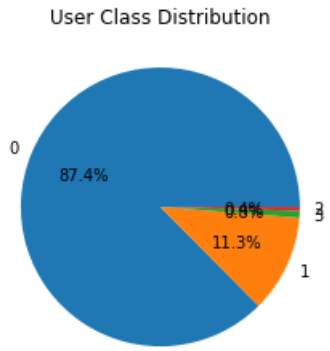
working model as soon as possible so I can gain insight on the results and fine tune the models and analysis. I designed the work in Sprint each sprint spanning the week and containing to meetings each week with the data engineer from the company to discuss the results. As the initial sprints highlighted compatibility issues in the relational databases as well as insufficient data the final ETL pipeline was set up in a later stage.

ETL

the data is extracted through Microsoft Azure from 5 predefined schemas that the data engineer of the company and I discussed the five schemas were handled as relational databases that can be joined on a user level with the variable "Client ID". To enable the database joining meticulous transformations and user level aggregations were required. Considering the data type of each feature individual aggregation methods were defined for each column. Missing values and potential errors we are considered and features not containing additional information were dropped. Eventually string variables were replaced either with labeling coding or one hot encoding. One defining feature was assigned ordinal labels.

Model preparation

Through the transformations, aggregations and designing the target variable the data frame enabled already some insights on the user classes that was not available before such as preferred browsers per user class preferred operating system per user class preferred browsing days, devices.



In the final dataframe each observation represented a user with several features most of them being categorical variables. Are used rule-based classification for labeling to create 4 user classes as in cold leads, warm leads, hot leads and users who purchased. This required some flexibility and the labels since the number of purchasing users was less than 0.001% in the whole data set.

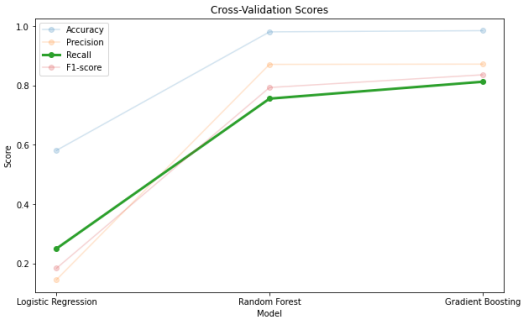
The analysis showed that the data set is heavily imbalanced, the largest category deriving from webshop actions that are not defined currently do not hold valuable information but has great potential to be informative when used properly in the future.

Classification with Machine learning

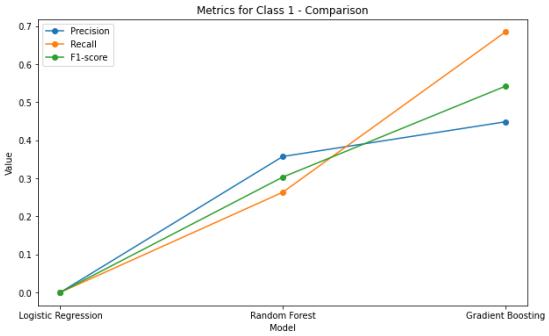
For machine learning models I used python's scikit-learn library and considered Recall as my defining metric of model performance. Recall considers the false negatives which in this context means missing marketing opportunity on users that are closer to buying than the actual negatives.

The data was split into hold out fast and training test and the five-fold cross validation was used on the training test.

Three classification models were used. To set up a baseline the first model was logistic regression while the other two were tree-based models: random forest and gradient boosting, expecting these last two to yield sensible results.



During the machine learning modeling it became apparent that the data is overfitting and biased. The latter being due to the imbalance that was carried over to the user plus targets as well. The overfitting didn't present itself in the holdout testing, that produced two good results that can be attributed to the fact that GA generates numerous variables from the same source so some variables may have perfect collinearity. After attempting to minimize these risk factors the final classification prediction considered only two classes. Results were evaluated through classification reports and confusion matrices.



Using the same three models the probability of buying was also predicted in another approach, but the imbalance of the data heavily influenced the results in this case as well. Focusing on the positive class the best model had 0.86 recall in Class 1, which can serve as a good baseline to further improve the model performance potentially with more data, spanning through several years.

Conclusion

Further exploration of the topic might include machine learning based clustering but it's prerequisite would be to produce a more balanced data set and the more conscious approach in how companies are using the potentials of Google Analytics.