Visitor Conversion Prediction Framework

Capstone Public Project Summary – Daniel Molnar – MS in Business Analytics

Project Overview

The capstone project's goal was to create a framework for the client which based on the users' past behavior can predict future likeliness of whether they are going to become advertisers on the client's website or not.

The client for which the analysis was done is one of the biggest e-commerce companies in Hungary which helps people buy and sell certain properties on their website. The company has access to high variety of historic data on a user level which was to be used to build out a framework which can be later used for marketing purposes. The company had two major goals in mind during this process; creating visitor level predictions to identify those users that are going to post their own ads on the website while understanding what are the different characteristics that can help the company identify these people, and understand their behavior.

Project Process

As most similar analytics projects, this one also started out by gathering and understanding the data that was available at the client. Even though the company had several different databases storing different information related to their users, they restricted these so that there would be only a subset of the whole dataset to be used answering the research problem. Essentially, the reasoning behind this was that the scope of the analysis had to be restricted because one of the important questions during the analysis was whether the company can use historic data to predict such conversion or not, and this dataset had to play a key part during this exercise.

The dataset provided had to be slightly cleaned and transformed into a format that was useful for predicting the visitors' ad-posting likeliness. These cleaning, filtering and transformation steps have been thoroughly documented and the important steps have been mentioned in the final project summary handed over to the client, so that they are aware of what had to be changed.

The most important part of the project was building predictive models that can accurately identify those group of visitors that are likely to advertise on the client's site. During the prediction process several predictive models had to be tuned, built and evaluated on a test set that the model didn't use for its training.

Several different approaches were used to try and do such predictions. The problem boiled down to a binary classification, for which there are several methods one can use. These include different machine learning algorithms from easier to understand multi-linear models to slightly more complex tree based and ensemble (boosting) models.

During the building of such models one must consider the different steps through which they can get to the best performing models. One of the most important steps during building multi-linear models was the scaling and centering the explanatory variables, to account for the drawbacks of such multi-linear models that don't perform well with skewed variables. Another important step was to introduce interactions between explanatory variables.

Other algorithms such as tree based and ensemble ones that were used during the project do not require every step of the above-mentioned pre-processing of variables. However, there are different model tuning steps that the data scientist must do in order to improve model performance and understand what makes their models perform the best outside of the training sets.

The different algorithms went through a lot of iterative steps of tuning, variable selection and interaction creation until I arrived at the final point of each model. Variable selections were based on model performance and variable importance, where the not so important variables were removed, and the model was rerun with only the important ones.

Each and every one of my models were evaluated on a holdout set to ensure that my analysis and predictions are meaningful not just on the set that I ran the model on, but the one the company will run the same models on in the future as well. For binary classification problems, one of the best indicators of model performance is the AUC number which was used to compare the performance of the different models against each other.

Based on these values, the best models were picked, and were discussed with the client on how to implement them and draw takeaways from them. Another last important step was to assign yes/no values to the probabilities that the models had as outputs on a user level. This was done in accordance with the clients input about whether False Positive or False Negative errors are worse or better for them.

Lastly even though the company was mostly interested in whether it is possible to build such a predictive model, it was also important for them to understand which are the predictors that can be used in such a framework, and what is the relationship between these and advertisement likeliness. As the last step of my analysis, I looked at the most important variables in my predictive models and drew some meaningful conclusions from them. There were several variables that had the expected relationship with the target, however there were also some relations that were not so intuitive and thus provided even more meaningful input for the client.

Key Outcomes

The main output of the project consists of a detailed project analysis that explains all the steps detailed in the previous paragraph. It contains all the necessary cleaning and transformation steps that must be made on the dataset to bring it into an analyzable format. It explains the different centering and data manipulation decisions that were done before building predictive models. The document has information about the various modeling decisions and steps and underlying logic that was used to get to the end results. The company also receives ready to use models that they can use and apply on their current visitor base and implement the necessary marketing measures that they had as an end goal of the project. Lastly, the file also analysis also includes details about the interactions that were meaningful and that explain visitor behavior.

In more detail, one of the key outcomes is that it is worth exploring this dataset and use it to predict future advertisers on the site. The models that were used for prediction achieved much greater results to just random guessing. This is important, because at the start of the analysis it was a big question for the company whether they can carry out such an exercise with their available dataset or not. This was quickly proven to be true.

The best models in the analysis have a very high accuracy rate of around 90%, depending on the threshold used for predictions. More importantly, using AUC as the measure for predictions, it is possible to achieve above 0.9 here. This is important, because this makes both the False Negative and the False Positive errors to be a reasonably low number and thus the company can use it more confidently.

Regarding meaningful variables, it seems like there are a lot of variables that were important in the process of predicting future advertising visitors. Important predictors included the number of events that the user had on the website, the number of times the user visited the website, and the

frequency with which the made those visits. It was also very important how long they have been the users of the client's website. Other important predictors included values such as the device that they have been using the access the page the different events that they have had with the already advertised properties. For example, it proved to be important whether the visitor have saved an ad posted by someone else or not, or if they reached out to them. Lastly, there were some meaningful interactions between the characteristics of the ads that the users viewed in relation to whether they were about to post their own ads on the site.

Due to the nature of this project, the details of these interactions are confidential, but the client received detailed information about these and thus can use it to understand their user base and make reasonably accurate and sound predictions going forward.

Client Benefits

As it was mentioned in previous paragraphs, the client will be able to use this framework going forward for two main use cases. First, the analysis yielded a very accurate predictive model, that will be used to cluster the companies current site visitors into groups of potential advertisers or people who are just browsing, or on the site to buy.

This is especially useful for the companies marketing department, who are looking to target potential advertisers with several marketing methods. This could include personalized ads, messages and several other techniques that could get their attention and thus could convert them into actual advertisers.

Going forward it is once again also very important the client received a detailed output about the variables that the different machine learning algorithms found useful in this prediction process. These can be used to analyze further and put into great use when trying to understand customer behavior, characteristics and different clusters in their user base.

Lessons Learned

Working on such real-life analysis project has its obvious benefits at the end of one's degree program. Personally, the biggest benefit I got from this project is the fact that the project complexity was high enough to spark new ideas and problems while it was also possible to solve it using the methods and algorithms that we have learned during our studies in the Business Analytics Master's program.

I have gained valuable machine learning engineering experience that will be very useful in my career going forward. I have had the chance to oversee and manage my own development and thus have experienced how it is to lead such a development effort. It was very important that I was able to manage these efforts according to the timeline that I proposed at the start of the process, and thus the client ended up being very satisfied with the end results, and the project delivery as well, including both technical expertise and communication.

I got the opportunity to apply most of the things that we have learned during our master's program, and that is the most valuable experience that I could have gotten out of such a project.

Summary

Building such a framework where the task is to cluster customers into groups based on their past behavior is a problem that is not always straightforward. At the start, the company had doubts whether it is possible to do something like this with their existing datasets, but during the project it became apparent that it is possible, and now the client has a framework which they can use to identify this customer base, and can implement marketing measures to target them based on their goals.