# Identifying Visitors of a Real Estate Portal with a Higher Probability of Posting a Paid Ad

Eszter Balogh

CEU - MS Business Analytics - Capstone Project Summary

**Abstract**

The aim of the project is building a predictive model which is capable of identifying visitors to one of the Hungarian real estate portals with a higher probability of posting a paid advertisement. The data used for the analysis is a Google Analytics clickstream data containing information on more than 100,000 visitors' activity. The predictive model is built with the aim to identify the sellers with a higher probability of switching to a paid plan. A key outcome of the project is that the clickstream data largely available for the company proved to be useful for identifying the target group. The machine learning model built on the data performed well on the test set achieving over 0.91 AUROC value and the variance importances extracted from it are in line with the results of the bivariate analysis. It can be concluded that visitors belonging to the target group are in general more active and also visit the website more frequently. Furthermore, they prefer using desktop and tend to save the interesting ads more often than contacting its seller and they, on average, view properties on sale with a higher price.

## Background

In this project, I analyse the characteristics of the visitors to a Hungarian real estate portal with the aim of identifying those with a higher probability of posting a paid property advertisement. Anyone intending to sell a property through this website has to first register and choose a plan. It is possible to go with a free plan which offers only basic features, e.g. no photos of the real estate can be included.

Still, not everyone decides to pay for the wide range of services on the website - many choose the free version with limited options. This leads to the conclusion that there is a large number of 'passive' sellers who do not put a lot of effort into the sales process probably because they are not aware of the added value of the paid services. The goal of the project is to build a predictive model which assigns a probability of upgrading to any of the paid plans to the sellers, however, as the company has not tried to use this data for this particular problem, the capstone also serves as a pilot project.

CEU eTD Collection

# Data

The data used for the analysis is a Google Analytics clickstream data with a detailed log of how visitors navigate through the website. The actual data contains information on more than 100,000 visitors' activity in over 2.5 million sessions amounting to nearly 15 million observations in total, with time reference between 2018 July 1 and 2019 March 31. Information is available on visitors' every session from this period with variables indicating e.g. the device used for opening the website, property type of the viewed ads, a list of actions related to property ads (saving to the favourites, contacting the seller or simple just viewing the ad) and some features of the real estate like price, floor area and county.

I used all the information included in the clickstream data to create visitor-level features and after that, I transformed the dataset to a one-visitor-one-observation structure. While some property types are significantly underrepresented (like public education institutions or warehouses) and also, there are obviously some errors in the data due to data recording by the seller (e.g. the posted price is a thousand/million times larger than the intended one), I decided not to drop these to make the predictions as general to the population (in this case, all the visitors to the website) as possible. It is also worth mentioning that the real data is highly unbalanced (the ratio of visitors posting a paid ad is low). To tackle this, filtering was applied on the group with no target event. In the end, six percent of the visitors belonged to the target group.

# The analysis workflow

As a first step, I performed a bivariate analysis when I inspected all features created from the original dataset with respect to the target variable. After that, for benchmarking the final boosting model, I fitted a simple logistic regression. This model does not include as many variables as the final one and also, the variables derived from categorical features saved in a 'counting' format (e.g. if a visitor saved 16 ads to the favourites, the corresponding variable takes the value of 16) were converted to binary due to converging issues. Another difference is that the logit model was not fitted using cross validation, however, the training data is the same 75% of the whole dataset. It is important to note that I excluded from the list of features the number of times a visitor posts a free ad as these people are easily identifiable by the company, including this feature in the predictive model would not add business value.

Contrary to the limited list of explanatory variables, the logit model performed well in the prediction task achieving 0.733 AUROC value. Fitting an interpretable statistical model before proceeding to the more sophisticated machine learning models is also useful due to the fact that not only the magnitude of effect but also its sign can be extracted. Based on the average marginal effects, the most important features increasing the probability are the number of visits and the mean price of the viewed properties on sale. On the other hand, the use of a mobile or tablet instead of a desktop for opening the website or the viewed ads' prices' larger interquartile range has a negative effect on the probability of posting a paid ad.

CEU eTD Collection

To improve the predictions, I decided to implement an extreme gradient boosting (xgboost) model. I started with tuning, focusing on five parameters: the maximum number of iterations (which, in a classification problem like this, is similar to the number of trees grown), the maximum depth of trees (larger the depth, more complex the model), eta (controls the learning rate), subsample ratio of features for constructing trees and finally, the ratio of samples supplied to a tree. The final model has 1,000 number of iterations, trees' maximum depth is 10, eta equals to 0.1 while both the features' and observations' ratio used for growing the trees is 0.8. The final model achieved 0.9129 AUROC value on the test dataset which is slightly higher than the cross-validated train AUROC (0.9063). When choosing a threshold for classification, a typical approach is to use the ratio of the target group in the whole dataset which is 6% in this case. However, I found the threshold based on Youden's J statistic (one optimisation method performed on the ROC curve) more suitable for the problem. From a business perspective, identifying the largest possible ratio of the target group is desired, False Positives are less of a problem. Using Youden's threshold allowed me to increase the number of True Positives (correctly classified visitors with the target event).

# Key outcomes

The first and most important outcome of the project is that the clickstream data largely available for the company proved to be useful for identifying the target group. The machine learning model built on the data performed well on the test set achieving over 0.91 AUROC value and the variance importances extracted from it are in line with the results of the bivariate analysis. Visitors belonging to the target group in general spend more time on the website and also visit it more frequently. Furthermore, they prefer using desktop and tend to save the interesting ads mor often than contacting its seller. Also, they view more expensive properties on sale than the average.

# Summary and conclusion

The goal of the project was building a predictive model that is able to identify visitors with a higher probability of posting a paid advertisement. The data provided for the problem was a Google Analytics clicktream data, containing over 100,000 visitors' activities from 2018 July 1 to 2019 March 31 and was structured in a way to tackle the unbalanced data problem. In the feature engineering phase, I transformed the data to a one-visitor-per-row structure. The types of models presented in this technical discussion are a logistic regression and an xgboost model. While they are not directly comparable due to the differences in the list of explanatory variables, the xgboost model performed undoubtly better in this problem, the variance importances were in line with what could have been concluded based on the bivariate analysis.

CEU eTD Collection