User Behavior Analytics for Mortgage Lead Recognition

Public Project Summary

By Gábor Radics

Submitted to

Central European University

Department of Economics and Business

Master of Science in

Business Analytics

Program

Autumn, 2018

Business case

Hungary's market leading online real estate listing site wants to widen its activities and take advantage of its knowledge and information about real estate searchers. We know that approximately 300 mortgage loans are disbursed a day in Hungary these days. Many of the new mortgage customers were active on the site prior to taking the mortgage loan. The basic idea was to identify the potential mortgage customers among the site users and transmit them to banks for fee. So, the company wants to be a market leading online mortgage broker. They already have an online loan comparison site, but the real estate listing site and the online loan comparison site operated independently from each other. Now they want to utilize the knowledge and information they have about the online real estate market and turn real estate searchers into mortgage customers.

Starting point and desired outcome

Before starting the project, it was already possible to calculate installments for different purchase prices, Loan-to-Value (LTV), tenor etc. on the real estate site, so historical data about mortgage related activity was available. Unfortunately, the users' main purpose on this web site was not search for mortgage loans, therefore only less than 1% of the users opened the calculator (*NR Calculator* in the below table). The conversion rate i.e. the proportion of users who requested detailed information about mortgages was even lower (*NR JZ Lead*).

visit Start Time	NR users	NR Calculator	NR JZ Recalc	NR JZ Lead
2017-10	2 465 194	16 497	4 376	347
2017-11	2 236 905	14 947	4 091	237
2017-12	1 900 321	11 929	3 502	160
2018-01	2 982 129	21 353	6 324	236
2018-02	2 527 646	16 649	4 727	217

The table clearly shows that the number of users who use the calculator was very low compared to the total number of users. The task was to understand their behavior, find patterns that point to forthcoming borrowing, and increase the number of mortgage leads.

Criteria of success

The first desired outcome of the project was to come up with a prediction model that predicts mortgage taking probability based on the users' real estate search activity. The model must be robust over different samples and time periods. The business goal was to increase the number of mortgage leads based on the prediction model. The management had high expectations, they calculated with additional 120 new leads per month.

Data sources

The details of the user activity, i.e. all the user clicks on the site are stored by Google and available in Google BigQuery. There are ~4.5 million hits a day on the web site. The data in Google BigQuery is available at three levels:

1) **user** level: all computers and other devices connected to a network, such as smartphones, tablets, and fax machines, have a unique IP address. Cookies are arbitrary pieces of data, usually chosen and first sent by the web server, and stored on the client computer by the web browser. The browser then sends them back to the server with every request.
2) **session** level: a session is a period when a user is active on a site or app.
3) **hit** level: a single session can contain multiple pageviews, events, interactions, and transactions. This is the hit level. Hit level information on the web site are number of saved real estates (RE), advertisers called or messaged, type/location/size of RE viewed, details of the search etc.

Time Horizon

The observation period of the **Test** and **Train** data is Jun 01, 2017 - Jan 21, 2018. The user, session and hit level data of this period is aggregated at user level. There are two different performance periods defined: 1 week (Jan 22, 2018 - Jan 28, 2018) and 2 weeks (Jan 22, 2018 - Feb 04, 2018).

The **Out of Time** validation's (external validation) observation window is Jun 01, 2017 - Feb 04, 2018. The 1-week performance period is Feb 05, 2018 - Feb 11, 2018, the 2 weeks performance period is Feb 05, 2018 - Feb 18, 2018. The next table summarizes the different time horizons:

<table>
<tr><th></th><th>Train/Test</th><th>Out of Time</th></tr>
<tr><td>Jun 01, 2017</td><td rowspan="3">observation period (235 days)</td><td></td></tr>
<tr><td>Jun 15, 2017</td><td rowspan="5">observation period (235 days)</td></tr>
<tr><td>Jan 21, 2018</td></tr>
<tr><td>Jan 22, 2018</td><td>performance week 1</td></tr>
<tr><td>Jan 28, 2018</td><td></td></tr>
<tr><td>Jan 29, 2018</td><td>performance week 2</td></tr>
<tr><td>Feb 04, 2018</td><td></td></tr>
<tr><td>Feb 05, 2018</td><td rowspan="4"></td><td>performance week 1</td></tr>
<tr><td>Feb 11, 2018</td><td></td></tr>
<tr><td>Feb 12, 2018</td><td>performance week 2</td></tr>
<tr><td>Feb 18, 2018</td><td></td></tr>
</table>

Explanatory variables

Practically infinite number of variables can be prepared using data from Google BigQuery. We have data describing search activity and the real estates, and we created 1148 explanatory variables. The explanatory variables used in modelling are aggregated to user (i.e. a person) level, since the purpose is to identify users who are interested in taking a mortgage. Some examples of the explanatory variables:

- *Search activity*: number of pages opened in the period; devices used (desktop, mobile, tablet); time spent on site; location of the searcher (based on geo network); source of visit (typing our URL, clicking link from email, ad monitor, social media, search result etc.); pictures seen; how many times the user logged in; contacts made with sellers etc.
- *Real estate*: floor area and price (minimum, maximum, average, standard deviation); type (flat, house, summer house etc.); condition; location etc.

Target variables

The target variable is binary, i.e. if the user is interested in taking a mortgage or not. The user can calculate mortgage installments and compare banks' offers on the site, and they can express their interest giving us his/her name and availability on our website, and we call them back to discuss the details. These are called *lead*. As seen before, the number of leads is very low, so I had to use a proxy, namely the number of recalculations of bank offers with different parameters (interest rate, loan amount, tenor).

As said before, the proportion of events i.e. the number of users interested in taking mortgage is very low. In the **Train** sample there are 17709 observations, but only 185 executed repeated calculations on our web site in the 1-week performance period (Jan 22, 2018 - Jan 28, 2018), that is only 1.0%, and even the 2-week performance period definition results in low event numbers (313 / 1.8%).

Model selection

Several model types (Generalized Linear Model, Penalized Logistic Regression, Classification Tree, Random Forest, Gradient Boosting Machine) were tried combined with data transformations (continuous variables, categorized continuous variables, log of continuous variables). All these models were tested for 1 week and 2 weeks performance periods. In all, 30 different models were developed.

Basically, the task was to maximize the true positive rate, i.e. to find the ones that are predicted to be interested in a mortgage and they really are. False negative rate is not a big problem financially since the users are targeted via online campaigns (retargeting), that are usually not very expensive. Still, we don't want to address all our users, since we don't want to spam them if they are obviously not interested in a mortgage loan. So, the model selection was based on the model performance, AUC in this case. The 30 models, together with their AUC are summarized in the next table.

		Performance period					
		1 week			2 weeks		
Variables used	Model	AUC	Test	Out of Time	AUC	Test	Out of Time
Continuous	Generalized Linear Model	m00_logit_base	0.7500	0.7404	m00_2w_logit_base	0.7334	0.7293
	Penalized Logistic Regression	m02_logit_penalized	0.7499	0.7386	m02_2w_logit_penalized	0.7324	0.7313
	Classification Tree	m03_rpart	0.6629	0.6670	m03_2w_rpart	0.6827	0.6756
	Random Forest	m04_RF	0.6976	0.7177	m04_2w_RF	0.7048	0.7203
	Stochastic Gradient Boosting	m06_GBM	0.7595	0.7522	m06_2w_GBM	0.7427	0.7375
Categorized continuous	Generalized Linear Model	m10_logit_base_factor	0.7519	0.7468	m10_2w_logit_base_factor	0.7413	0.7380
	Penalized Logistic Regression	m12_logit_penalized_facor	0.7701	0.7474	m12_2w_logit_penalized_facor	0.7526	0.7484
	Classification Tree	m13_rpart_factor	0.6765	0.6646	m13_2w_rpart_factor	0.6685	0.6675
	Random Forest	m14_RF_factor	0.7232	0.7466	m14_2w_RF_factor	0.7241	0.7264
	Stochastic Gradient Boosting	m16_GBM_factor	0.7685	0.7510	m16_2w_GBM_factor	0.7525	0.7438
Log of continuous	Generalized Linear Model	m20_logit_base_log	0.7554	0.7478	m20_2w_logit_base_log	0.7516	0.7438
	Penalized Logistic Regression	m22_logit_penalized_log	0.7582	0.7660	m22_2w_logit_penalized_log	0.7544	0.7483
	Classification Tree	m23_rpart_log	0.6629	0.6670	m23_2w_rpart_log	0.6827	0.6756
	Random Forest	m24_RF_log	0.6987	0.7148	m24_2w_RF_log	0.7185	0.7232
	Stochastic Gradient Boosting	m26_GBM_log	0.7507	0.7450	m26_2w_GBM_log	0.7460	0.7426

The model with the highest AUC on the Test sample was selected for the implementation (selected models for the two performance definitions are highlighted with blue). The best model for the 1-week performance period is m12_logit_penalized_facor (categorical variables, penalized logistic regression), and m22_logit_penalized_log (log of continuous variables, penalized logistic regression) for the 2-weeks performance period. These models also have a quite high AUC on the Out of Time sample.

Implementation

The implementation of the final model took place on March 1, 2018. In this case we used *Google Ads* remarketing tool. Remarketing is a way to connect with people who previously interacted with a website or mobile app. It allows companies to strategically position ad's in front of their audiences as they browse Google or its partner websites.

The best final model for the 1-week performance period was Penalized Logistic Regression with category variables (final 1-week model), and Penalized Logistic Regression with log variables for the 2-weeks performance period (final 2-weeks model). These models were used to predict probabilities using the Production data set. The results of these models were combined, and five groups were created:

- Group 1: 8.000 random cookie id's (8000 users), that didn't use the mortgage calculator in three weeks prior to implementation.
- Group 2: 1.722 cookie id's (923 users), all of them used the mortgage calculator in three weeks prior to implementation.
- Group 3: 11.388 cookie id's (10357) with a predicted conversion rate less than 0.6% for both models.
- Group 4: 27.415 cookie id's (12823) with a predicted conversion rate between 1% and 5% for both models.
- Group 5: 6391 cookie id's (1632) with a predicted conversion rate higher than 5% for both models.

The list we submitted to Google Ads contained 54.916 cookies for 33.735 different users (i.e. persons). In Group 1 there were 8000 randomly selected cookie id's that are 8000 different users. The next table presents the five groups and the results after implementation:

	# cookies	# Users	Cookies/Users	Clicks	Impr	CTR	Clicks / cookies	Clicks / Users
Group 1	8000	8000	1.00	338	72818	0.46%	4.2%	4.2%
Group 2	1722	923	1.87	43	11649	0.37%	2.5%	4.7%
Group 3	11388	10357	1.10	303	100169	0.30%	2.7%	2.9%
Group 4	27415	12823	2.14	589	149434	0.39%	2.1%	4.6%
Group 5	6391	1632	3.92	123	38165	0.32%	1.9%	7.5%
Total	**54916**	**33735**	**1.63**	**1396**	**372235**	**0.38%**	**2.5%**	**4.1%**

Online marketers generally look at CTR (click-through rate), i.e. the ratio of cookies that click on a specific link to the number of total cookies who view a page. CTR doesn't suggest that the model has a good prediction power, since Group 1 has the highest CTR (4.2%) when I considered the *device* (desktop, mobile, tablet) as a user. So, the denominator in the calculation is the number of *devices* (this is the default calculation rule in Google Analytics).

That is why we used clicks / users instead. User identifies a *person* here who is logged in, not a device. One user can log in on many different devices. Unfortunately, I can't tell the proportion of cookies / users in Group 1 since the persons using these devices are not logged in. But I see that the higher the group number (in Group 3, 4 and 5), the higher is the proportion of cookies / users, e.g. 3.92 in Group 5. Almost four IP addresses are related to one user in average in Group 5. If we look at Clicks / Users than the rate is 7.5% for Group 5, the highest among all groups.

Conclusion

Clicks / Users of 7.5% is considered good, but the evaluation of the whole project was based on the increase of users who convert to bank leads. Unfortunately, this number didn't increase significantly. It is due to the inefficiency of the online channels. 7.5% of the users clicked on our banner (i.e. our lending page) that we placed on other sites than ours. In the moment of the click the users were not as interested in mortgage calculation as they had been when they visited our site.

However, the models were considered good since they were robust over different samples and time periods using data of the company's own sites. The next planned action step is the implementation of the personalized web site, i.e. the appearance of the site will be adjusted to the predicted user needs. It also means that the model will be implemented in an environment where the model data is coming from, that is the ideal form of implementation. We expect a much better conversion rate from this development, that verifies more the model predictions.