Customer Life Time Value

Tomer Kalimi, Jacob Zahavi and Ronen Meiri

Contents

Introduction

You are considering launching a campaign to acquire new customers through the web. How much are you willing to invest in acquiring a customer?

You suspect that a given customer is about to defect and switch to a competitor. Would you try to keep this customer or let her go?

You want to assess the value of your business. How would you evaluate the worth of your customers?

Well, all of these questions boil down to assessing the life time value of your customers, often abbreviated as LTV. Certainly one would be willing to acquire a new customer only if the cost of acquiring the customer is lower than the life time value of the customer. Not only companies are happy to get rid of unprofitable customers but in many cases companies may take active actions to make these customers defect to their competitors. And, finally, in the new world, and also in the more "conventional" industries, the value of a company is given by the aggregation of the life time value of their customers.

So what is the LTV?

LTV is a prediction of the net profit attributed or the expected income to the entire future relationship with a customer. The prediction model can have varying level of sophistication and accuracy, ranging from a crude heuristic to the use of complex predictive analytics techniques. In predictive modeling one estimates the LTV by means of a variety of attributes, mainly the history of activities (e.g., purchases in the retail industry, donation amounts in the charity industry, deposits in the gaming industry,…) the customer has conducted with the company and, to a lesser extent, demographics characteristics. The duration of the activity history varies depending on the domain. In the automotive industry the history period may consist of several years, perhaps up to 10 years or more; in the retail and charity industries one or two years; in the gaming industry perhaps a few months.

The gaming industry is especially noted for the short-time duration as many customers stick around for several days or weeks and then walk away. Consequently, in this industry it is mandatory to assess the LTV based on the activities short after the first deposit, sometime as little as 24 or 48 hours since their first deposit. The objective here is to identify the high-value customers, often referred to as the VIPs or whales , up front, in order to gain more profits from them by, say, giving them more incentives to increase their deposits and extend their tenure with the company. Indeed, even a small increase in deposits of these customers worth a lot to the gaming outfit.

In this white paper, we exhibit an approach to calculate the LTV in the gaming industry based on the customer's activity in the first 24 hours since his first deposit.

LTV in the Gaming Industry

To calculate the LTV in the gaming industry we need to gather three key information components in the customer life cycle:

1. Registration information
2. The deposit period
3. The LTV period

Figure 1 describes the flow stream for LTV modeling schematically.

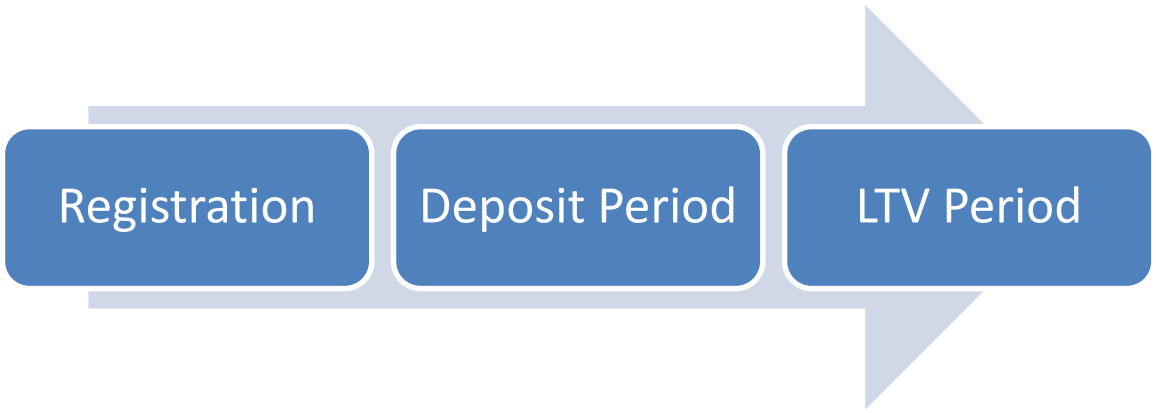

Figure 1 LTV life cycle components

The registration information is concerned with the basic customer's demographics at the time of registration, primarily gender, referral, email, location, credit card type and others.

The Deposit period is the duration of time we want to base our LTV calculations on which, because of the short-lived activity periods in the gaming industry, is set, in this study, to 24 hours after the first deposit. Based on the activities in these 24 hours we want to predict the LTV.

And finally, the LTV period is concerned with the time duration, following the deposit period used for predicting the LTV, e.g., 30 days, 90 days or any other duration. By and large, the LTV measure of interest is defined in money-related terms, e.g., the sum of deposits over the period of the LTV period. But money-related measures are not necessarily the only relevant metric. Another measure could be the number of clicks which could be a good proxy for profitability -

usually the larger the number of clicks, the larger is the likelihood that the customer will deposit more money. Alternatively, one can partition the customers into several segments, based on the available history following the LTV period, e.g. 30 days, 60 days, 90 days,… and build a separate model for each.

The Modeling Process

The modeling process in data mining consists of several steps which are described schematically in Figure 2

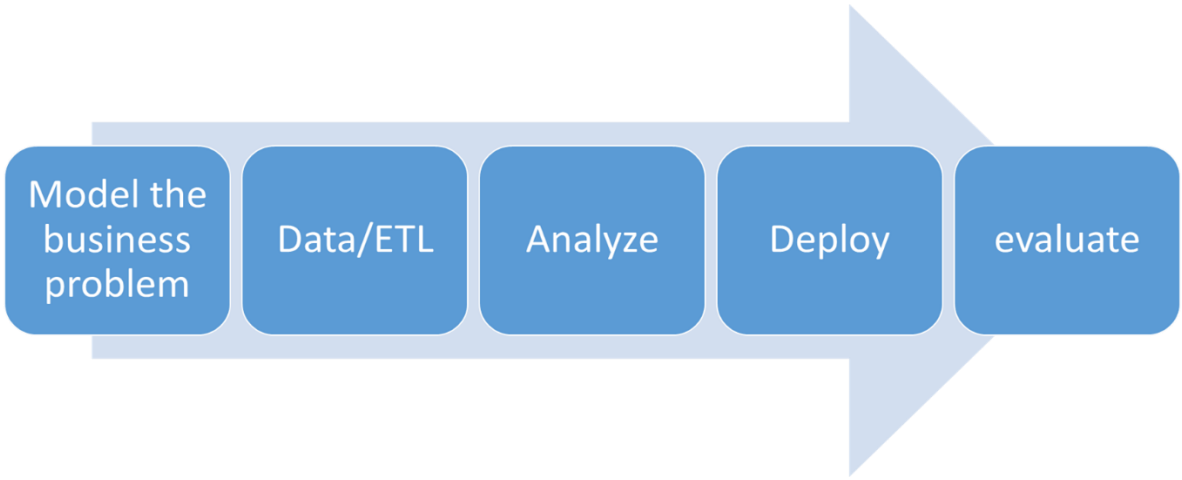

Figure 2 – LTV Modeling process

The process starts with defining the business problem at hand. As alluded to earlier, in our case we are interested in estimating the LTV of a customer who registers to the site and makes at least one deposit. The first deposit marks the start of the deposit period. Our objective is to estimate the LTV over a period of 90 days based on the customer's activities in the first 24 hours since the first deposit.

We note that the first deposit time may, or may not, coincide with the registration time, as a customer can sign up for the site and makes a deposit only later. If the customer registers to the site but does not make any deposit, she does not participate in the modeling process.

Data Modeling

Oftentimes, the information for modeling comes from different sources - the registration information from one source, the deposit information from another source, etc. One needs to merge the information from the various sources, based on a user ID, to create a flat file (the analysis dataset) which contains all the attributes for each observation.

Figure 3 describes the data flow. The User initiates the process by making one or more deposit and opens up sessions to play games which result in gains/losses to the User. The deposits and the game results affect the customer balance which, in turn, creates transactions that are stored in the company's database.

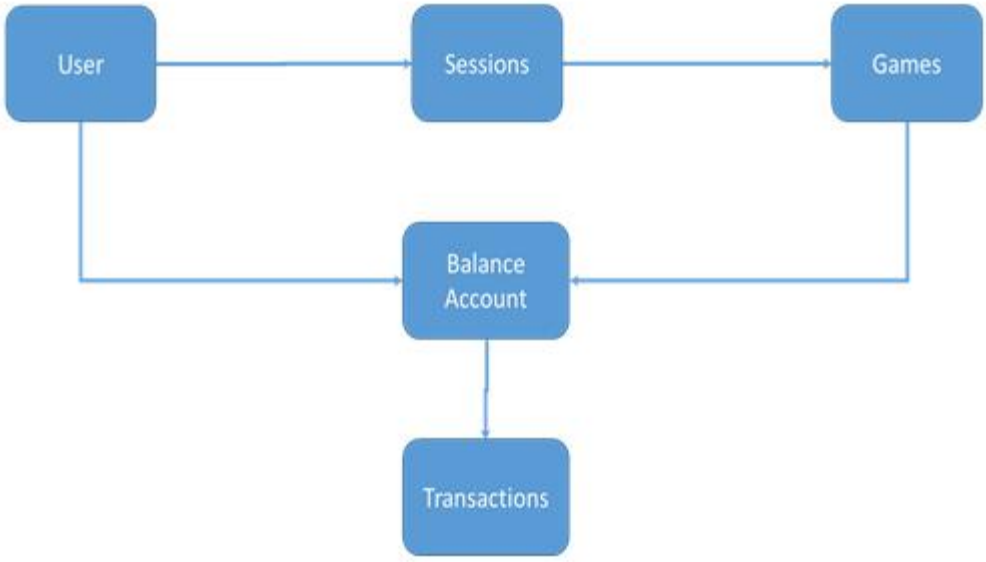

Figure 3 – LTV Data flow

The transactional database is often too detailed for data mining, requiring that the information be aggregated according to some classifications, as appropriate for the problem domain. Some transformations of the original variables may be necessary to convert these variables into "predictors". For example, summing up all deposits in the deposit period to create the dependent variable; defining "flags" for categorical variables in the LTV period, and others.

Another major concern is timing issue. Since each customer makes the first deposit at a different time, then, to bring all data points to the same denominator, we set the reference date as the date of the first deposit and normalize all time-related variables relative to this reference.

The following lists some of the data manipulations that we took in our case:

Use all registration data, as is:
Gender
Country
Language
Affiliation

Normalizing all relevant dates as distances from the reference date:
Birth date – customer age at the reference date
Registration date – number of days to reference date

For each numerical variable in the Deposit period – deposit, sessions, durations,…, calculate:

- First value in period
- Last value in period
- Average in period
- Frequency in period

For each categorical variable in the Deposit period – games, deposit method,…, calculate:

- First value in period
- Last value in period
- Most frequent in period
- Frequency in period

For each relevant date – deposit date, registration date,…, find the corresponding:

- Day of week
- Whether Weekend
- Time of day
- Time period (night, morning, noon, evening)

Finally, for the LTV period we created the target variable by summing up all deposit amounts during the period.

This process ended up with 207 variables, 1 key variable (customer ID number) and one target variable.

The LTV Model

In predictive modeling we predict the value of the target (also, dependent) variable based on a host of attributes. In our case, the target variable is the sum of the customer's deposit values over the LTV period, the independent, or explanatory, variables are derived, as above, based on the first 24 hours after the first deposit.

Since the target variable is continuous, we use a linear regression model to predict the LTV. To avoid over fitting and make sure the model is stable enough, the audience is split into two mutually exclusive and exhaustive datasets – a training dataset used for building the model and a validation dataset used for validating the model.

When using linear regression, best results are often obtained when the distribution do not have a lot of extreme values. Plotting the frequency distribution of the LTV variable yields a distribution which is far from normal (left side of Figure 4). To fix this phenomenon, we took the log of the LTV values to obtain a much more normal-like distribution with the extreme values of the long tail falling within the reasonable range (right side of Figure 4). We then run the linear regression model using the two target variables – the original numeric LTV values and log(LTV value +1). The (+1) was added to account for zero LTV values.

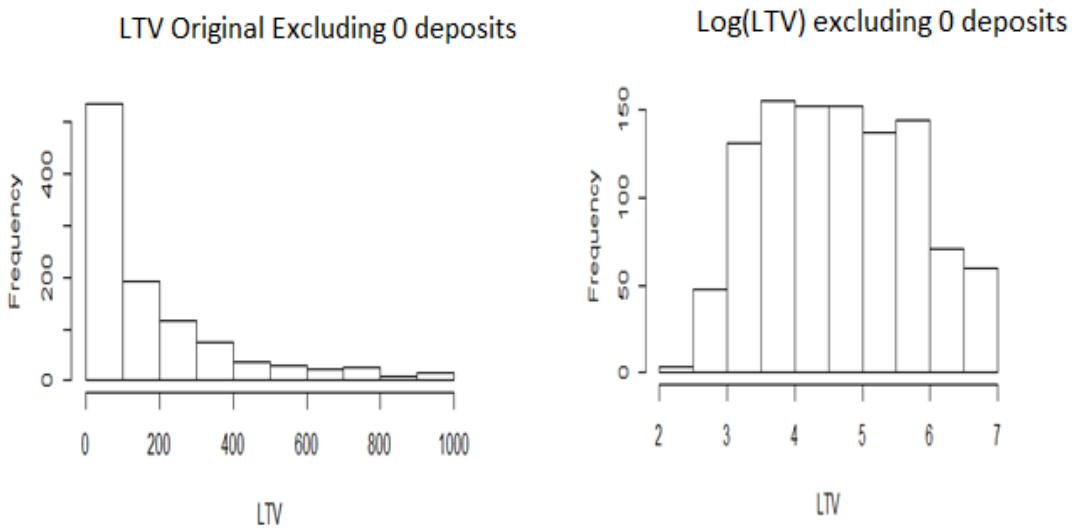

Figure 4 – frequency distribution of the original LTV values and the logarithm of the LTV values

The Modeling Results

The Modeling Results are presented by means of Gains Charts in Figure 5. Figure 5a presents the gains charts for the training dataset, Figure 5b the chart for the validation dataset. The gains charts are created off the corresponding gains tables which are exhibited in tables 1 and 2.

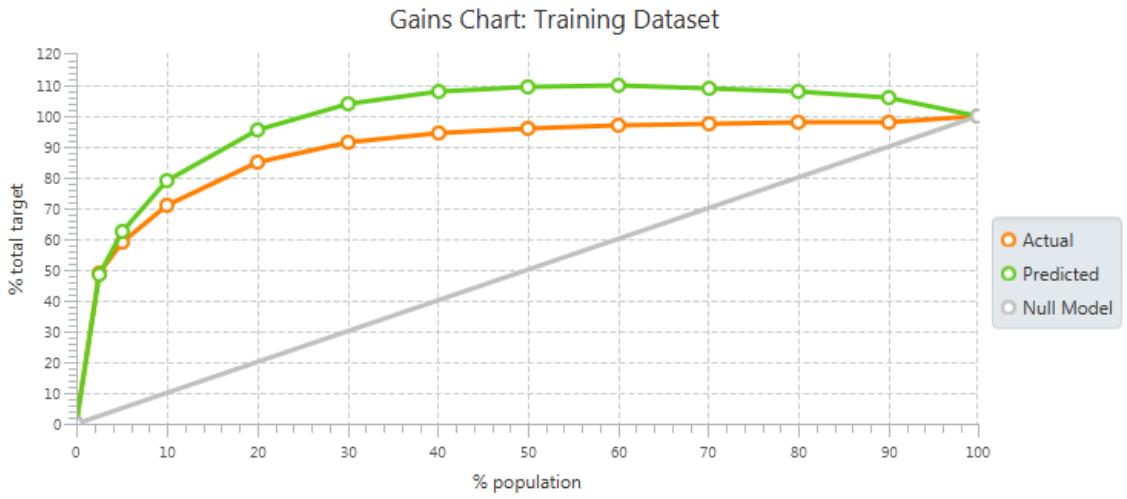

Figure 5a – Gain charts for LTV Values in training dataset

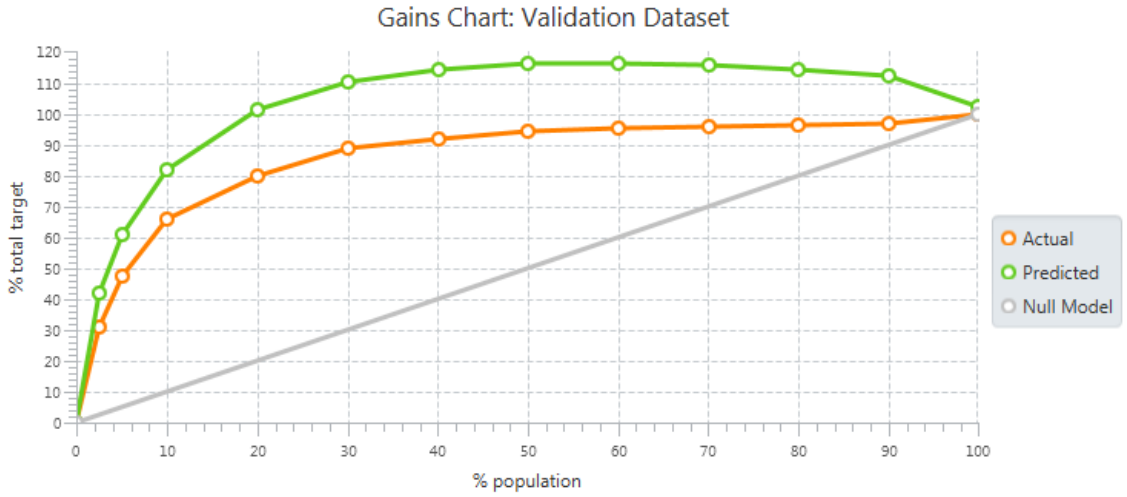

Figure 5b – Gain charts for LTV Values in validation dataset

Table 1 – Gains table for training dataset

% Population	# Prospects	Actual Response	Pred. Response	Actual Response	Pred. Response	% Total Response	% Total Pred. Response
2.53%	96	331425.5	326280.3	3452.35	3398.75	49.14%	48.38%
5.03%	95	66686.55	94077.19	701.96	990.29	59.03%	62.32%
10.03%	190	80322.97	111678.3	422.75	587.78	70.93%	78.88%
20.03%	380	93116.27	112547.4	245.04	296.18	84.74%	95.57%
30.03%	380	46348.05	55269.27	121.97	145.45	91.61%	103.76%
40.04%	380	18881.85	25648.22	49.69	67.5	94.41%	107.57%
50.04%	380	10667.01	11293.84	28.07	29.72	95.99%	109.24%
60.04%	380	5974.432	2366.54	15.72	6.23	96.88%	109.59%
70.04%	380	2936.092	-3759.89	7.73	-9.89	97.31%	109.03%
80.05%	380	1996.869	-8311.71	5.25	-21.87	97.61%	107.80%
90.05%	380	1269.148	-12176.1	3.34	-32.04	97.80%	106.00%
100.00%	378	14851.71	-40437	39.29	-106.98	100.00%	100.00%

Table 2 - Gains table for validation dataset

% Population	# Prospects	Actual Response	Pred. Response	Actual Response	Pred. Response	% Total Response	% Total Pred. Response
2.59%	46	85210.04	114797.4	1852.39	2495.6	31.23%	42.07%
5.07%	44	44727.14	51117.6	1016.53	1161.76	47.62%	60.81%
10.08%	89	49693.93	57257.52	558.36	643.34	65.83%	81.79%
20.06%	177	38546.1	53075.03	217.77	299.86	79.96%	101.24%
30.08%	178	24465.88	24383.61	137.45	136.99	88.93%	110.18%
40.06%	177	8421.891	11592.09	47.58	65.49	92.01%	114.43%
50.08%	178	5988.447	4815.63	33.64	27.05	94.21%	116.19%
60.06%	177	2627.517	699.62	14.84	3.95	95.17%	116.45%
70.08%	178	2258.895	-1881.53	12.69	-10.57	96.00%	115.76%
80.06%	177	1372.204	-3915.24	7.75	-22.12	96.50%	114.32%
90.08%	178	846.8274	-5748.02	4.76	-32.29	96.81%	112.22%
100.00%	176	8698.431	-26877.9	49.42	-152.72	100.00%	102.37%

Gains charts are very convenient and self-explanatory way to visualize data mining model results. To create the gains charts, we sort out the model predicted values in decreasing order and then summarize the results at the decile level, from the top decile to the bottom decile. The x-axis represents percentage of the population, the y-axis the percentage of the cumulative predicted value. The orange curve exhibit the actual values, the green curve – the predicted values.

For example, 20% of the audience in the training dataset (Figure 5a) capture almost 85% of the total actual LTV, a very hefty lift. 50% of the audience creates captures almost 100% of the total actual the LTV. The predicted values (the orange curve), on the other hand, are somewhat off with a relatively large difference between the actual and predicted values.

The curves for the validation dataset (Figure 5b) are somewhat different than the corresponding curves for the training dataset but the difference between the actual and predicted values are still pretty large.

Both these results may indicate over fitting, most likely caused by the skewed shape of the LTV values which are far from being normal. As a results, we ran the linear regression model also on the log of the LTV values. The corresponding gains charts are presented in Figure 6. We note that the y axis in the gains chart are expressed in terms of % of the total of the log LTV. The corresponding gains tables for the training and validation datasets are exhibited in tables 3 and 4, respectively. In the gains tables we also exhibit the results in terms of the log LTV values.

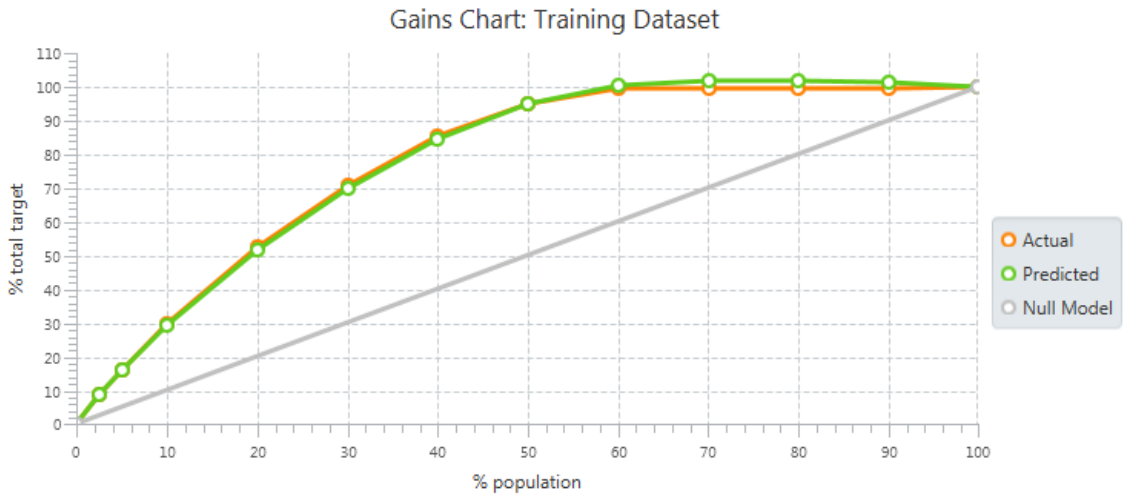

Figure 6a – Gains charts for the Log(LTV) model for the training dataset

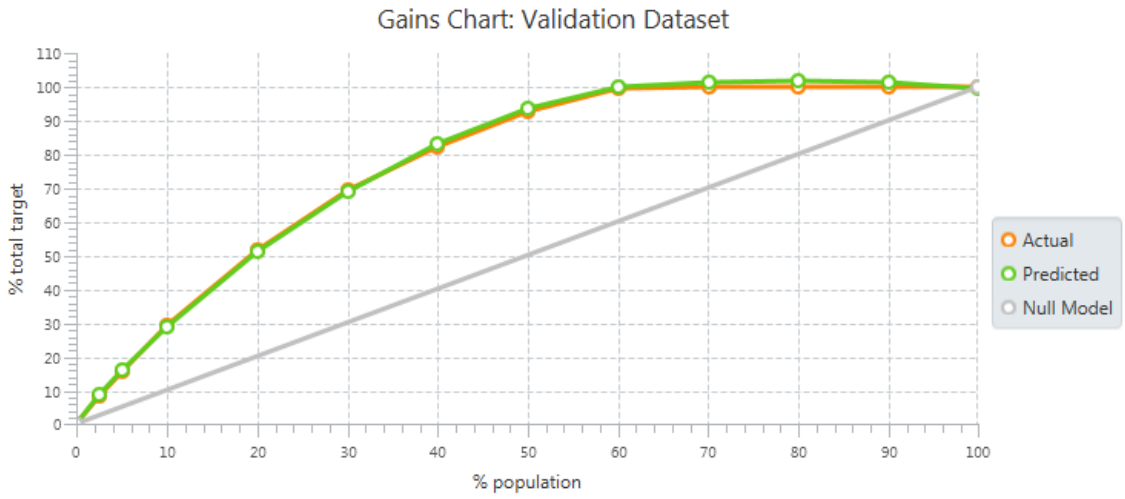

Figure 6b – Gains charts for the Log(LTV) model for the validation dataset

Table 3 Gains table for training dataset

% Population	# Prospects	Actual Response	Pred. Response	Actual Response	Pred. Response	% Total Response	% Total Pred. Response
2.55%	97	699.1664	701.68	7.21	7.23	8.93%	8.96%
5.05%	95	583.6348	571.43	6.14	6.02	16.39%	16.26%
10.04%	190	1041.423	1024.97	5.48	5.39	29.69%	29.36%
20.03%	380	1804.352	1767.4	4.75	4.65	52.74%	51.93%
30.04%	381	1429.905	1436.67	3.75	3.77	71.01%	70.29%
40.03%	380	1122.997	1127.48	2.96	2.97	85.35%	84.69%
50.04%	381	770.534	819.98	2.02	2.15	95.20%	95.17%
60.03%	380	349.3357	430.98	0.92	1.13	99.66%	100.67%
70.04%	381	13.86153	106.97	0.04	0.28	99.84%	102.04%
80.03%	380	5.663735	7.59	0.01	0.02	99.91%	102.13%
90.04%	381	0	-31.44	0	-0.08	99.91%	101.73%
100.00%	379	7.189922	-135.66	0.02	-0.36	100.00%	100.00%

Table 4 – Gains table for validation dataset

% Population	# Prospects	Actual Response	Pred. Response	Actual Response	Pred. Response	% Total Response	% Total Pred. Response
2.60%	46	322.2367	334.21	7.01	7.27	8.64%	8.96%
5.09%	44	263.4748	269.12	5.99	6.12	15.70%	16.18%
10.06%	88	509.5279	484.4	5.79	5.5	29.36%	29.16%
20.07%	177	839.3277	829.9	4.74	4.69	51.87%	51.41%
30.07%	177	658.0326	665.68	3.72	3.76	69.51%	69.26%
40.08%	177	473.2268	528	2.67	2.98	82.20%	83.42%
50.08%	177	396.2765	390.61	2.24	2.21	92.82%	93.89%
60.09%	177	259.232	230.72	1.46	1.3	99.77%	100.07%
70.10%	177	8.520911	57.63	0.05	0.33	100.00%	101.62%
80.10%	177	0	7.44	0	0.04	100.00%	101.82%
90.11%	177	0	-13.1	0	-0.07	100.00%	101.47%
100.00%	175	0	-62.06	0	-0.35	100.00%	99.80%

Table 5 presents selected variables that made it to the final model. For example the country of registration was grouped into several country categories to render them more prediction power. But only variable X005, containing the countries Bolivia, Macadonia, Romania, Moroco,…made it to the model; all the other country categories were eliminated from the model by the DMWay Analytics algorithm.

The variable X013 was represented by a piecewise linear representation, containing several segments, two of which have made it to the model (1570.22<=X013<7123.75) and (7123<=X013<infinity). And likewise for the other predictors.

Table 5 – Model Summary

Variable	Transformation	Predictor	Description	Estimated Coefficient	Standard Error	t-value	P.value
	(Intercept)	(Intercept)	(Intercept)	-0.01843	0.063624	-0.2897	0.77206
X002	piecewise	X002_piecewise_0004	X002 in(32.76,51.966]	0.023705	0.005575	4.2523	2.17E-05
X005	groupCat	X005_groupCat_0005	X005 in{Bolivia,Macedonia,Romania,Morocco,Serbia}	-0.58008	0.17757	-3.2668	0.001098
X012	piecewise	X012_piecewise_0003	X012 in(33.7988,78.8293]	0.019345	0.003777	5.1219	3.18E-07
X013	piecewise	X013_piecewise_0004	X013 in(1570.22,7123.75]	5.48E-05	1.93E-05	2.8432	0.004492
X013	piecewise	X013_piecewise_0005	X013 in(7123.75,Inf)	5.48E-06	9.82E-07	5.5762	2.64E-08
X015	piecewise	X015_piecewise_0003	X015 in(35.0362,88.9216]	-0.01398	0.003223	-4.3371	1.48E-05
X022	groupNum	X022_groupNum_0003	X022 in(60.8251,135.077]	1.2119	0.25482	4.7559	2.05E-06
X022	groupNum	X022_groupNum_0004	X022 in(135.077,392.753]	1.919	0.30578	6.2758	3.89E-10
X022	groupNum	X022_groupNum_0005	X022 in(392.753,942.357]	2.4601	0.37269	6.6008	4.68E-11
X022	groupNum	X022_groupNum_0006	X022 in(942.357,5048.75]	3.0359	0.44567	6.8119	1.12E-11
X022	groupNum	X022_groupNum_0007	X022 in(5048.75,Inf)	4.1125	0.55695	7.3839	1.89E-13

Conclusions

In this white paper we presented an approach to calculate the LTV of customers in the gaming industry based on her activities in the of activities in the first 24 hours since the customer made her first deposit. The target variable consists of the total deposit amount in the succeeding 90- day period. Because the distribution of the original LTV values where ill-distributed with many extreme values, we defined a new target variable which involve the log of the original LTV values. The modeling results were evaluated for stability and over fitting by splitting the audience into a training and a validation datasets, building the model on the training dataset and validating it on the validation dataset.

The results for the log transformation yielded results which are stable and exhibiting no over fitting.

DMWay Analytcs

The modeling task of the LTV model was conducted by DMWay's state-of-the-art predictive analytics software, DMWay Analytics. DMWay Analytics uses an expert system approach to build large scale predictive analytics models that mimics the way that an experienced data scientist goes about building models. The process starts by creating transformations of the original numerical variables to account for potential nonlinear relations between the target and the explanatory variables. Categorical variables are grouped together based on their relation with the target to render them more prediction power. The feature selection process of DMWay Analytics then kicks in, applying a multistep procedure, involving statistical tests, and business and domain rules, to identify the most influential predictors affecting the target variable. This process reduces the set of potential predictors, which may contain hundreds, if not more, of predictors to the handful of the most influential predictors affecting the target variable. Those influential predictors then take part in the modeling process. The resulting data mining model, whether linear regression or logistic regression, or other, is then translated into a scoring code that allows to deploy the model results for scoring new observations either online or offline.

Contact us

dmway.com
info@dmway.com
Tomer Kalimi: tomer@dmway.com
Ronen Meiri: ronen@dmway.com
Jacob Zahavi: jacob@dmway.com