An Engagement-Based Customer Lifetime Value System for E-commerce

Ali Vanderveld
Groupon
Chicago, IL
avanderveld@groupon.com

Addhyan Pandey
Groupon
Chicago, IL
adpandey@groupon.com

Angela Han
Groupon
Chicago, IL
angela@groupon.com

Rajesh Parekh*
Facebook
Menlo Park, CA
rgparekh@gmail.com

ABSTRACT

A comprehensive understanding of individual customer value is crucial to any successful customer relationship management strategy. It is also the key to building products for long-term value returns. Modeling customer lifetime value (CLTV) can be fraught with technical difficulties, however, due to both the noisy nature of user-level behavior and the potentially large customer base. Here we describe a new CLTV system that solves these problems. This was built at Groupon, a large global e-commerce company, where confronting the unique challenges of local commerce means quickly iterating on new products and the optimal inventory to appeal to a wide and diverse audience. Given current purchaser frequency we need a faster way to determine the health of individual customers, and given finite resources we need to know where to focus our energy.

Our CLTV system predicts future value on an individual user basis with a random forest model which includes features that account for nearly all aspects of each customer's relationship with our platform. This feature set includes those quantifying engagement via email and our mobile app, which give us the ability to predict changes in value far more quickly than models based solely on purchase behavior. We further model different customer types, such as one-time buyers and power users, separately so as to allow for different feature weights and to enhance the interpretability of our results. Additionally, we developed an economical scoring framework wherein we re-score a user when any trigger events occur and apply a decay function otherwise, to enable frequent scoring of a large customer base with a complex model. This system is deployed, predicting the value of hundreds of millions of users on a daily cadence, and is actively being used across our products and business initiatives.

*This work was done when the author was at Groupon.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08…$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939693

Keywords

Customer Lifetime Value, E-commerce, Random Forests

1. INTRODUCTION

Groupon[1] is a large global e-commerce company, operating via the web and the popular Groupon Mobile App. Currently serving more than 30 countries and 500 markets, Groupon is the place you start when you want to buy just about anything, anytime, anywhere. We offer physical merchandise through our Goods business, travel deals through our Getaways business, and we are a market leader in Local commerce. As of Q3 of 2015, we have nearly 50 million active customers worldwide, more than one million merchants connected through our suite of online tools, and over 900 million units sold. At Groupon we are trying to develop a robust marketplace, and as such we need to understand at an individual level the supply and service needed to develop a daily habit for our customers. How does featuring the local burger place down the block compare to featuring a big chain when it comes to increasing a user's future spending? Given the number of local choices a customer has, how many Groupon options do we provide to promote a daily habit? What is the added value of providing "white glove" customer service?

Answering such questions requires a comprehensive understanding of user-level value and how it is changing with time. Accordingly, in this paper we tackle the problem of modeling and monitoring customer lifetime value (CLTV) – a prediction of the net dollar value attributed to our future relationship with each individual customer. It is relatively easy to predict aggregate business measures for financial purposes. Accurately modeling the future value of individual users, however, is a far more difficult task, as users spend differing amounts of time on product research and have various engagement levels prior to a final purchase decision. Nonetheless, accurate CLTV predictions can have enormous value for an e-commerce company like Groupon. The historically most common use cases for such a system pertain to marketing and customer relationship management (CRM). With CLTV predictions it is straightforward to find the highest value customers, customers most likely to attrite, and customers most likely to make their first purchase. As

[1] http://www.groupon.com

such one can determine the ideal target audiences for promotional offers, personalized customer messaging, exclusive deals, rewards programs, and customer service treatment.

In addition to the huge marketing value, CLTV predictions are generally useful as a powerful performance metric in all areas of the business. For example, in product experimentation we can use CLTV scores to segment users and to build for longer term value returns. We can identify causal versus correlated relationships of certain behaviors with user value by encouraging and incentivising said behaviors and then measuring outcomes. We can also use these predictions in personalization, to tune results to different stages in the customer lifecycle, and for user behavior analysis. How does CLTV change based on offering (e.g. Goods versus Local) or as a result of specific actions? How does the adoption of a new platform change a user's value as a function of time? We can also use CLTV predictions for business unit goal setting; an example of such a goal might be for the CRM team to increase the value of previously one-time buyers by $X\%$ for a given quarter.

Here at Groupon we have developed a new methodology for the modeling and daily tracking of individual customer value. Our system is novel in several ways, including: 1. we engineer features that quantify the level of engagement that each customer has with us on each of our platforms, 2. we build highly accurate two-stage random forest models, 3. we use separate models for different types of purchasers, e.g. one-time buyers versus very active users, and 4. we re-score a user when any trigger events occur and apply a decay function otherwise. Using engagement-based features allows us to detect any changes in value, including attrition, far more quickly than we could otherwise. We have found random forest models to have the best performance over other machine learning frameworks, and modeling different customer types separately allows us to have different feature weights for each while simultaneously overcoming interpretability issues by giving business units a straightforward way to discover and track the most important features for each group. Employing a decay function allows us to efficiently update our scores on a daily cadence despite having a very complex model, leading to a factor of 20 speed up in overall run time. By implementing these model features we are able to achieve Spearman correlations between actual and predicted customer values of 0.53 and 0.77 for quarterly and yearly timeframes, respectively, showing moderate to high correlations at a very high statistical significance ($p < 0.0001$).

In what follows we present our CLTV system, which predicts customer value for three rolling time windows that we will denote as "short," "medium," and "long," that are on the order of a quarter to a year. We use a proxy measure for customer value that we will call "purchase value" that is based on purchasing behavior, focusing on this behavior instead of profit because the latter is subject to margins that can fluctuate and that are not related to user intent. We focus on value over these three time windows as opposed to full lifetime value; given the age of Groupon, a one-year purchase prediction is a significant portion of the overall average customer life as of today. There is not enough training data to make accurate predictions with a longer-term model but the future intent is to add longer time periods as a representative core customer base matures and the data becomes available. Hereafter and we will often refer to these values as the user's "scores." The system described herein is currently deployed in production and successfully running at scale, scoring hundreds of millions of users on a daily cadence and being used by more than half a dozen business and product use cases.

2. RELATED WORK

This work is related to a few different areas in the marketing and machine learning literature, most importantly CLTV, CRM, and random forests.

2.1 Customer Lifetime Value

Gupta *et al.* [9] provide a comprehensive review of different CLTV methodologies, relating them to what they state as the three phases of the relationship between a company and a customer: acquisition, retention, and expansion. They review six basic types of modeling approaches, including the historically popular RFM and related models that deal with the recency, frequency, and value of each customer's purchase history. They present evidence from the literature that machine learned prediction models like the one presented below are superior because they can incorporate a variety of additional variables, and they mention that random forest models had historically performed well in churn prediction competitions. See also [12, 8, 17] for further review on the history behind CLTV prediction. In all of the CLTV literature there is the argument that, rather than focusing on the short-term, CLTV modeling allows us to focus on long-term profitability [19, 2].

There are various methodologies that have been used to model CLTV in a computer science framework, including optimization [7], SVM [6], quantile regression [3], and portfolio analysis [5]. In a 2009 CLTV prediction competition [14], several different frameworks for predicting CLTV on an individual and an aggregate basis were tested against each other. We find that a new approach – two-stage random forest models for each of several user behavior segments, with a comprehensive feature set such as what we have engineered here – performs better than any of the competition entrants. This methodology allows us to incorporate features related to user engagement with our platform, which allow us to quickly detect changes in customer value.

2.2 Customer Relationship Management

The idea that we should invest resources into analyzing customer behavior and maintaining a positive relationship with our customers is nothing new. For example, Kotler and Keller [11] showed that the costs associated with obtaining new customers can be five times larger than the costs associated with maintaining a good relationship with existing customers. A thorough literature review of various subjects related to CRM, including CLTV modeling techniques, can be found in [16]. The need to fully understand our customers so that we can maintain these relationships is one of the strongest motivations for accurate customer lifetime value monitoring. This is the main impetus to build a reliable monitoring system at Groupon.

2.3 Random forests

Random forests is an ensemble learning method that works by producing many random decision trees and then bootstrap aggregating (or "bagging") the results to turn an ensemble of weak learners into a strong one that automatically

avoids overfitting. The methodology was developed by Leo Breiman and Adele Cutler [4]. We make use of the functionality in R [13] and the H2O package [1] [2]. H2O is a scalable machine learning framework that supports many of the standard algorithms, including random forests.

3. FEATURE SET

Our model uses over 40 features, which provide a comprehensive view of each user's demographics, general purchasing behavior, engagement, and overall relationship with Groupon. Here we describe each type of feature in turn. A discussion of relative feature importance is in Section 6 below.

3.1 Engagement

The most important features that we include are those that characterize user engagement with our product. These rich features allow us to detect changes in customer value far sooner than we could if we were only relying on purchase history data. In this first version of our model we include engagement scores from email and the mobile app, currently two of our most important sources of traffic. There are various key ways in which users interact with each of these platforms, for example with opens and clicks for email and with deal impressions and searches on the app. We coalesce these behaviors into one composite score for each platform by weighting a quantification of each behavior (e.g. the number of email clicks) by a numerical factor related to the historical conversion coming from that behavior. The final score is then the weighted sum. For example,

$$\text{Email engagement score} = \alpha N_{\text{open}} + \beta N_{\text{click}} \ ,$$

where N_{open} is the number of email opens, α is the corresponding historical conversion rate (average orders per email open), N_{click} is the number of email clicks, and β is the corresponding historical conversion rate (average orders per email click, given an email open). Note that most emails that are opened are not clicked on, and that a single email can be clicked on multiple times. Thus we consider each behavior (opening and clicking) separately, using the appropriate conversion rates to avoid double counting. We further track email unsubscription and mobile app downloads, and define features related to each.

3.2 User experience

There are many ways in which the Groupon user experience is correlated with future customer value. Firstly, the quantity of available nearby deals can play a large role, especially for newer and re-engaging users. We quantify this local supply by counting the number of deals within a representative radius of the user's home location. We also determine the typical purchase volume in their geographical area, which acts as a proxy for local Groupon brand recognition.

Another key part of the user experience is related to customer service. We track numbers of refunds and customer service tickets, customer service phone and email wait times, and whether the user has given positive or negative responses to post-interaction surveys. We also track average shipping times for our Goods (physical merchandise) business.

[2] `http://www.h2o.ai`

3.3 User behavior

Virtually all CLTV models use purchase history as a key variable, and ours is no exception. We have features relating to purchase value on a variety of historical timescales, along with the number of days since the most recent purchase. We also include Goods versus Local preferences and whether or not those preferences are changing as a function of time. Also redemption behavior can be important, and as such we include as features the typical time between purchase and redemption and the number of unredeemed vouchers that the user is currently holding on to. We further track how much each user is predisposed to using discount codes for their purchases or for purchasing "loss leader" deals.

3.4 Other features

In addition to the aforementioned features we also use basic demographic variables such as gender, age, and location, including city size and the distance between home and city center. We also track the circumstances regarding their original subscription and first purchase, including the cohort years for each.

4. CLTV MODEL

The features in the previous section are used by our series of CLTV models. We train models for each combination of our six user segments and our three time windows – short, medium, and long. Each model is retrained on a quarterly basis due to the relatively young age of our business and the fast pace of business changes. In this section, we describe our user segments, our two-stage random forest model framework, and our empirical seasonality and decay function models. The design of the overall system that calls upon these models is discussed in Section 5 below.

4.1 User segments

Of our 40+ features described above, some will invariably be more important predictors of future purchasing behavior than others. However, which variables are most important will be different for different types of customers. For example, for a user who purchases very frequently we may be able to accurately predict future purchasing behavior largely from previous purchasing behavior. On the contrary, for a user who has not yet purchased we will have to rely entirely on other (non-purchase history) variables. Thus we divide our users into six "purchase cohorts" based on past purchase frequency. Each of these purchase cohorts then is modeled separately from the others to allow for differing feature weights for each.

These cohorts are determined from the past five quarters of data, excluding travel and other outlier price point purchases. The total number of cohorts is six, which was determined based on a balance between model accuracy and re-training time. The time window for cohort determination (five quarters) was determined based on a balance between having enough data to detect a pattern but also allowing for the possibility that users may change their purchasing behavior over their lifetimes.

In this paper, we will focus on a subset of our purchase cohorts, defined and named as follows:

- Unactivated: never purchased
- New users: purchased but insufficient data to determine cohort (requiring at least four quarters)

- One-time buyers: only one purchase ever
- Sporadic buyers: infrequent purchasers (but more than once)
- Power users: very frequent purchasers

The specific thresholds for what constitutes a sporadic buyer versus a power user versus any other cohort will vary for different businesses that have different typical purchase frequencies. We choose thresholds that are optimized for Groupon. Note also that the cohorts developed here are optimized for the purposes of understanding user segments, modeling individual user behavior, and informing product and marketing efforts. These do not reflect the cohort strategy used to assess overall business performance. These separate concerns require very different segmentation strategies.

4.2 Random forest models

For each purchase cohort and time window, we use random forest models in R, using the H2O [1] package. In addition to previous studies finding random forests to be superior for CLTV modeling [9], we further found it to be the most accurate from a series of experiments where we tested various algorithms, including rpart [18] and SVM [15], and various normalizations and parameter tunings. We found that a lower `mtry` (the number of features randomly sampled as candidates at each split) and a higher `ntree` (the number of trees to generate) performed the best for our purposes here. Performance was judged based on a combination of RMSE and Pearson and Spearman correlations between actual and predicted values for our cross-validation data sets.

There are significant benefits and disadvantages of using random forests in this application. The advantages include allowing for missing values, accommodating many nonlinear features (e.g. the number of days since the most recent purchase), and generally better model performance including avoiding overfitting. The main disadvantage relates to interpretability because the result is an average over many decision trees. We overcome this disadvantage by modeling each of our purchase cohorts on its own to extract the most important features. Then we create cohort metrics for our end business partners (marketing, customer service, etc.) so that they can act on these metrics as they change with time. For example, we find that one of the most important features for predicting the future value of our power users is the typical number of days from purchase to redemption. Accordingly we track redemption time and CLTV as we test new redemption-related products, such as redemption reminder emails and expired voucher trade-in programs.

Our model has two stages, each using random forests:

1. Predict purchase propensity, a binary classification for whether or not the user is predicted to purchase, for the time window.
2. Predict the dollar value for users who were predicted to purchase in stage 1.

We use two stages because the data sets tend to be highly imbalanced; for example power users are far more likely to purchase than not, whereas the opposite is true for one-time buyers. Thus for stage 1 we downsample the majority class to get a 50-50 training set. The full combined training sets contain 40,000 samples for each combination of purchase cohort and time window. The classification cutoff has been optimized for each purchase cohort to minimize the bias in the number of purchasers and their overall value. We build separate two-stage models to predict future values for each of the short, medium, and long time windows for each cohort.

4.3 Seasonality and decay models

We employ two additional models which are both built empirically from historical data. Firstly, since our training data is necessarily on at least a quarterly lag, we need to make quarter-to-quarter seasonality adjustments. These adjustment factors are predicted from historical timeseries data for each purchase cohort separately and are re-calculated every quarter.

The second empirical model that we use is related to our scoring procedure. In order to get the most out of a CLTV monitoring system, we require re-scoring on a relatively fast cadence. Ideally we would like daily updates to capture sudden changes in value due to any purchases, engagement, customer service interactions, etc. However, in a large global business such as Groupon we have potentially hundreds of millions of users to score. Doing so with an extensive six-cohort two-stage random forest model would be computationally very expensive. We reduce the computation cost by leveraging our knowledge of customer behavioral patterns.

We note that the value of users who go without any interactions with Groupon, including any engagement, decays in a very predictable manner in line with Figure 1. We measure this decay function as a multiplicative factor, which is a function of the number of days since the last interaction. We calculate this function from historical time-series data and re-calculate it at the start of every quarter. We then use this function as a part of the daily scoring procedure:

1. Find users who had any "trigger events," which we define as any interactions with Groupon related to our feature set (see below).
2. For triggered users, run their updated feature set through the two-stage random forest model to calculate fresh scores.
3. For users without any trigger events, multiply their scores by the appropriate decay functions.

Trigger events include purchases, any interactions with email or the mobile app, and any customer service interactions including calls, emails, and refunds.

With this scoring procedure we reduce our daily scoring time by a factor of 20, a savings of 95%. This time quickly adds up. In this way we are able to update each user's predicted values for each of the short, medium, and long time windows on a daily cadence, with each of these time windows on a rolling basis. For instance, on September 1st, 2015, a quarterly prediction gave the expected purchase value for the current date through the current date plus 90 days, or November 30th. Similarly a one-year prediction gave the expected value until September 1st of 2016.

5. SYSTEM DESIGN

We show a high-level overview of the Groupon CLTV system in Figure 2. The computation and collection of the feature sets is done within Groupon's enterprise data warehouse. We additionally assign purchase cohorts and build

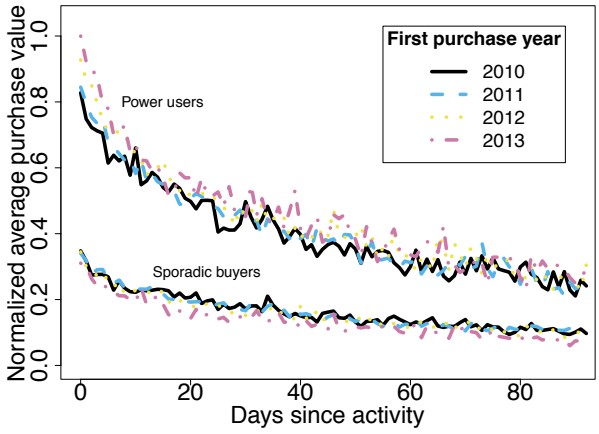

Figure 1: How average quarterly user value decays as a function of the number of days since the last interaction ("trigger event"), shown for various combinations of purchase cohort and first purchase year cohort. All customer values have been normalized by a constant factor such that the average value for 2013 power users is equal to one.

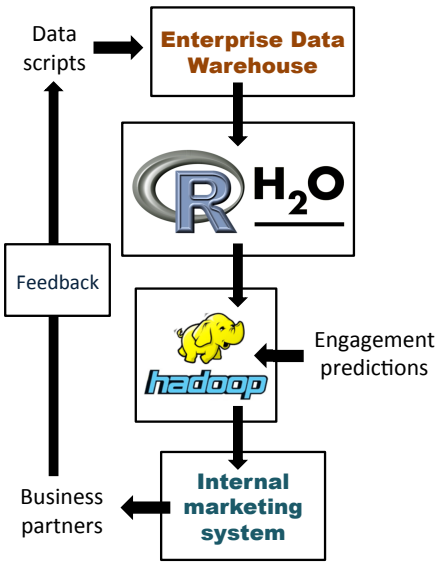

Figure 2: High-level overview of our CLTV system. Timeseries models and features are built within our enterprise data warehouse, then R and H2O [1] are used to build random forest models and calculate user values. Results are stored in our Hadoop cluster, joined with engagement score predictions, and then ported into our internal marketing system for easy access by our business partners.

our timeseries models within this database. Models are built and users are scored using R in conjunction with the H2O package [1] on a server with 44 Gb of RAM and 12 CPU cores. We then store the results in our Hadoop cluster, combine the results with engagement score predictions (from a separate, independent model), and port everything to a Groupon internal marketing system that allows our business partners to segment users for personalized marketing campaigns. Several secondary scores are also calculated, including a predicted percentage value change that compares the past quarter to the future quarter, and an associated binary 0/1 decline alert flag. Feedback from our business partners, typically in the form of feature requests and updates, is fed back into the data model.

We split the CLTV scoring process into a daily process and a quarterly process as described in the subsections below. The quarterly process runs on the first day of each quarter (January 1st, April 1st, July 1st, and October 1st) while the daily process runs on all other days of the year. Both processes have built-in accuracy monitoring and alerts at each stage of each process.

5.1 Quarterly process

On a quarterly basis we complete the following steps:

1. Assign purchase cohorts for each user based on the user's past five quarters of activity.
2. Compute seasonality functions for each purchase cohort from historical data.
3. Compute decay functions for each purchase cohort and time window from historical data.
4. Harvest new training data sets and retrain all models on new data.
5. Calculate all new CLTV scores for everyone with either a purchase in the past six months or any engagement in the past 90 days.
6. Everyone without either a purchase in the past six months or engagement in the past 90 days is predicted to have zero future value for each time window.

The requirements to receive a new score (versus an automatic score of zero value) were determined based on experimentation on the level of inactivity required to be certain of zero value according to our models.

5.2 Daily process

On a daily basis (excluding the first day of the new quarter) we complete the following steps:

1. Update the feature set for all users.
2. Find users with trigger events since the last scoring, which include any interaction with Groupon related to our feature set.
3. Re-score all triggered users for each time window.
4. Apply the decay functions to the scores for all users who were not triggered, for each time window.

As mentioned above, all of these time windows – short, medium, and long – are on a rolling basis and each set of scores is thus updated every night when the CLTV system runs. We are currently building alerts to monitor model drift.

6. SYSTEM PERFORMANCE

The system described herein has been successfully implemented at large scale, scoring hundreds of millions of users on a daily cadence. The system has been stable for several months and takes approximately four hours to run each night. In this section we review 1. the performance of our model using test data, 2. various measures from Q3 of 2015, and 3. the most important features for each stage of our model for our different user segments. In the results below, all metrics are for large random subsets of users and are purely indicative of model performance. Whenever the time window (short, medium, or long) is not specified it is to be assumed that we are defaulting to the short time period; when that is the case then it can further be assumed that the results are similar for the longer time windows.

6.1 Model evaluation

For all of our model experiments we trained with features as of the end of Q1 of 2014, thereby predicting customer values starting from the beginning of Q2 of 2014. We used out-of-time cross-validation data sets that were shifted one quarter later, such that we were predicting values starting from the beginning of Q3 of 2014. We use data from 2014 (as opposed to 2015) in order to have enough data to test the results for our long time window. For each new model experiment we drew new random subsamples of the population to create new random training and cross-validation data sets. Each of our cross-validation and testing data sets contained 10,000 samples from each purchase cohort. We evaluate the performance of Stage 1 based on accuracy, precision, recall, and false positive rate. We evaluate the performance of the overall (stage 1 and 2) model based on Pearson and Spearman correlation coefficients, RMSE, bias in the averages, and a comparison of actual versus predicted distributions.

6.1.1 *Stage 1 performance*

In the first stage of our model we predict a binary yes/no response for whether each user will make a purchase in the time window of interest. In Figure 3 we show the ROC curves – the true positive rate versus the false positive rate as we vary the cutoff parameter for the binary classification scheme – for each purchase cohort. In this figure the cutoff ranges from 0.2 to 0.8. The shallowness of these curves reflects the inherent difficulty in predicting individual user behavior. The final cutoff parameters are chosen so as to minimize the bias in predicting the total number of purchasers for each cohort.

In Table 1 we show the resulting accuracy, precision, recall, and false positive rate for each purchase cohort for this stage 1 model. The "overall" numbers are for all cohorts combined, with equal numbers of users in each cohort's sample. Note in particular the very high recall and low false positive rate for users who have never purchased, and the very high false positive rate for power users. It is easy to predict when an inactive user will make their first purchase, but it is more difficult to predict when a previously very active user will drop off. This is because we tend to have more advance warning for the former, in the form of increased engagement prior to purchase. In the latter case (when a previously very active user disengages), the shift is typically very sudden. However, we find that a majority of these dropped-off users will return in the following quarter, which means that they will be receptive to CRM efforts once we have enough engagement data to detect the drop-off. Note also that it is very difficult to predict the purchases of sporadic buyers because this is the group with the noisiest purchasing behavior.

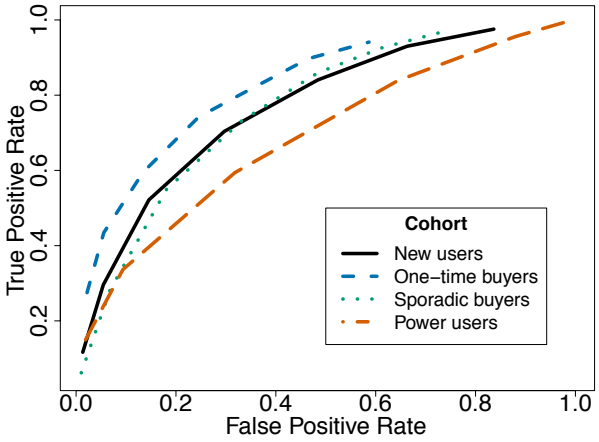

Figure 3: The ROC curves for our stage 1 model for each purchase cohort, showing the true positive rate versus the false positive rate as we vary the classification cutoff parameter from 0.2 to 0.8.

Table 1: The accuracy, precision, recall, and false positive rate (FPR) for each purchase cohort and overall for our stage 1 model. All results are percentages.

Cohort	Accuracy	Precision	Recall	FPR
Unactivated	98	43	100	2
New users	78	50	53	15
One-time buyers	95	37	32	2
Sporadic buyers	81	41	37	10
Power users	77	89	83	64
Overall	93	50	62	4

6.1.2 *Overall stage 1 and 2 performance*

After we determine a binary classification for which users are predicted to purchase, we then predict the purchase value for the positives and assign a zero value to the negatives to achieve our final results. In Table 2 we show the Pearson and Spearman correlation coefficients between actual and predicted values at the end of this process, for each combination of purchase cohort and time window. All of these show a moderate to high level of correlation [10], all with very high statistical significance ($p < 0.0001$). Typically we find that our accuracy measures tend to improve as the time period is increased. This is because purchase value per user is still a rather noisy measure for the short timeframe for this dataset.

We present the RMSE for each combination of purchase cohort and time window in Table 3. These values are given in units of the average actual purchase value for each combination in question. Once again we see that longer time windows and more active customers are easier to predict, because in these situations we have more data and less noise; the most difficult prediction by far is for the unactivated group over the short time window, whereas the easiest pre-

Table 2: Pearson (P) and Spearman (S) correlation coefficients for each cohort for each time window.

Cohort	Short P	Short S	Medium P	Medium S	Long P	Long S
Unactivated	0.38	0.65	0.55	0.95	0.56	0.85
New users	0.33	0.38	0.43	0.46	0.44	0.50
One-time buyers	0.22	0.31	0.36	0.45	0.40	0.51
Sporadic buyers	0.23	0.29	0.31	0.40	0.37	0.46
Power users	0.47	0.40	0.59	0.48	0.61	0.51
Overall	0.45	0.53	0.70	0.72	0.73	0.77

Table 3: RMSE for each cohort for each time window, presented in units of the average actual purchase value for each case.

Cohort	Short	Medium	Long
Unactivated	13.01	7.89	5.46
New users	3.30	2.55	2.44
One-time buyers	7.48	5.15	4.04
Sporadic buyers	4.24	3.07	2.59
Power users	1.20	0.95	0.90
Overall	6.08	1.82	1.67

diction is for the power users over the long time window. For reference, the typical user in our data set purchases 1-2 times over the short time window, with the unactivated users purchasing 0.008 times, the sporadic users purchasing 0.3 times, and the power users purchasing 7 times on average. Our overall results outperform the winner of the 2009 CLTV modeling competition presented in Ref [14].

Because our model was tuned to avoid a systematic bias, we find that the actual and predicted numbers of purchasers are very close, and accordingly that our systematic bias in average purchase value per user is low in each purchase cohort. Overall the bias in average value is less than 1%. This general lack of a systematic bias was verified with testing on a sample to predict Q1 of 2015 (thereby including the confounding factor of the holiday spike).

We further find a good concordance between the actual and predicted distributions of spend per user, as shown in Figure 4. The distribution of actual spend is somewhat more spread out – having both more low- and high-spend users – than the distribution of predicted spend. In particular it is impossible to predict the long tail of very rare high-value purchases. We also find that we are predicting a reduced number of very low-spend users. This is due to our two-stage model, in which these users are typically classified as non-purchasers in stage 1 and thus given an automatic zero value in stage 2. Testing the use of the probability (as opposed to a binary classification) in stage 1 to mitigate this effect is the subject of future work.

6.2 Q3 analysis

This system went live at the start of Q3 of 2015. At the end of the quarter we completed additional validation tests to determine how useful our results were for Q3. Firstly we re-computed the metrics detailed above to successfully confirm that all of our performance metrics were in line with our initial model testing. These computations are automatically checked at the end of every quarter. In particular we find that the correlation coefficients for our entire Q3 customer base are indeed consistent with what we found from

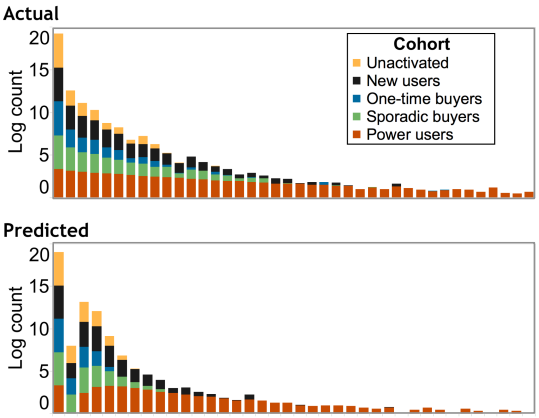

Figure 4: A comparison of the distributions of the actual (top) and the predicted (bottom) purchase value for the short time window, showing logs of counts for clarity. The horizontal and vertical scales are the same for both top and bottom panels and financial information is obfuscated.

our original out-of-time 2014 testing sets. Below we discuss a few additional metrics, all for a random subset of users from Q3 of 2015.

Comparing the top 10% to the bottom, we find that users who were predicted to be in the top 10% ended up spending on average 14 times more than users who were predicted to be in the bottom 90%, and 20 times more than those predicted to be in the bottom 10%. We further find that users who were predicted to decline in their value ended up having on average a 55% decline and users who were predicted to increase in their value ended up having on average a 62% increase.

Furthermore when we predicted that someone would attrite in Q3 of 2015, we ended up being correct 72% of the time. These attriting customers are naturally grouped according to their purchase cohort, which informs our marketing teams as they design CRM campaigns to keep these customers in Q4. We further check the feature set for each user predicted to attrite to find if there are any specific reasons for the disengagement, such as a slow shipping time or an expired voucher. Specific actions include sending promotions for general disengagement and initiating customer service treatment when a disengagement trigger has been identified. Then by tracking CLTV pre- and post-intervention, we determine the efficacy of these different strategies.

We use these predictions to analyze user behavior and to

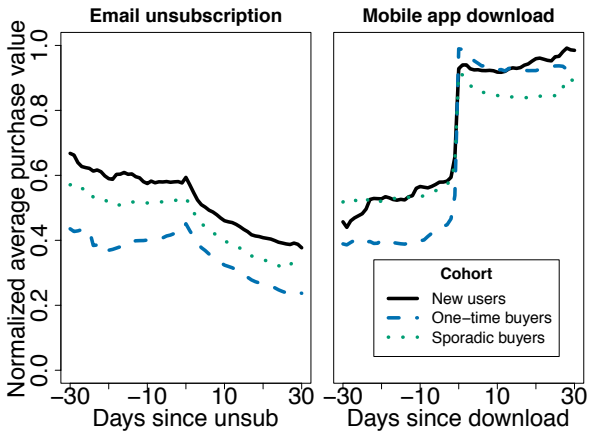

Figure 5: Average purchase value for each purchase cohort for the short time window as a function of the number of days since an unsubscription from emails (left) and a download of the mobile app (right). All customer values have been normalized by a constant factor, and the horizontal and vertical scales are the same for each panel. Power users have been left out because their value is much higher, but their behavior is qualitatively the same.

determine the value changes stemming from various lifecycle events as a function of time. We show two examples in Figure 5. Following an unsubscription from all emails we find a steady decline in predicted value due to a reduction in engagement. On the other hand, a download of our mobile app dramatically increases predicted value, due both to the download itself (a feature in the CLTV model) and due to the increase in app engagement. These behaviors are in line with what we have seen from historical data.

These CLTV predictions are so far being used at Groupon by several of our business units, including our marketing team which uses them for CRM campaigns to find the highest value customers and the customers most likely to attrite. For the sake of illustration we show some examples of the sorts of user segmenting that can be done in Figure 6 with a small sampling of data (how we employ these predictions in practice for user segmenting is outside the scope of this paper). For example, we can use these results to determine which users would benefit most from being sent promotional offers. For Q3 of 2015, users who were characterized as power users who were predicted to purchase in that quarter ended up using $4.8 million in discounts from promotional offers. That adds up to roughly $20 million per year in discounts to users who are extremely active and predicted to purchase anyway. Of course, experimentation is required to determine which users would benefit most from promotional offers. These CLTV scores can be used to select a prospective set of users and then the final decision of who gets the promotional offer would be decided based on the responses of user segments in a test campaign.

Also at Groupon we are using these results in product experimentation, both for segmenting users and as a longer-term KPI. For example, products related to redemption are geared towards increasing long-term customer satisfaction and thus cannot be adequately tested using typical shorter-term conversion-based A/B experimentation. For this we use CLTV as a performance metric and track the shifts in redemption-related features in the testing versus control groups over time.

6.3 Feature importance

As discussed above, we build different models for our different "purchase cohorts," with the intuition being that different features will be most important for customers at different stages of their lifecycle. We find that this intuition is indeed correct. For example, the top four most important features for our power users are, in no particular order:

- Purchase value in the past quarter
- Purchase value in the past year
- Lifetime purchase value
- Number of days since most recent purchase

Note that each of these features is related to purchasing behavior. On the other end of the spectrum, we cannot rely on purchase-related features to predict value for our one-time buyers. For these users the most important features include:

- Email engagement score
- Average number of orders per quarter per subscriber in the home city
- Whether or not the user has downloaded the mobile app
- Number of days since the user's last (and only) purchase

7. CONCLUSIONS AND FUTURE WORK

In this paper we have described the challenges faced, the choices made, and the lessons learned in building a customer lifetime value monitoring system for a large global e-commerce company. Our model uses engagement scores for email and our mobile app as features, and this gives us the ability to quickly and accurately detect changes in customer value. We gain additional accuracy by building separate models for differently-behaving sets of customers, such as one-time buyers and power users. We further use a triggering system and an empirical decay function to dramatically speed up the scoring process and make daily re-scoring with such a complex model technologically feasible. This system is currently running in production, scoring hundreds of millions of users on a daily cadence, and the results are being utilized by various business units including our marketing and experimentation teams.

In the future we plan to enhance this system in several ways. These include testing other machine learning algorithms (e.g. zero inflated binomial and gradient boosted decision trees) and adding more features, such as scores for other types of engagement including organic desktop web activity. We also plan to add more information regarding the use of discounts, and other features that other business units request for their specific needs. Our use of scalable technologies means that our system will continue to perform as our business grows. However we will still continue to determine ways to further speed up the system and decrease the use of resources.

Figure 6: Illustrative scatter plot of predicted future value versus past value for the short time window for a small sample, showing a few different user segments that one could define. The diagonal line denotes a one-to-one correspondence and financial information is obfuscated.

As we continue to gather more data, we will evaluate longer term trends and explore the use of longer time frames for our purchase cohorts. We will also fully quantify the business costs associated with model errors. For example, a false negative for newer users might have larger implications for building a long-tern relationship, whereas it would not be as detrimental to a frequent user. We will further continue to create cohort metrics for our end business users, for instance tracking redemption time for the redemptions team as mentioned above. We plan to similarly track cohort metrics for our other product and marketing partners and continue to iterate to develop the CLTV system and new products in tandem.

8. REFERENCES

[1] S. Aiello, T. Kraljevic, P. Maj, and with contributions from the 0xdata team. h2o: R interface for h2o. 2015. R package version 3.0.0.25.

[2] C. Bailey, P. Baines, H. Wilson, and M. Clark. Segmentation and customer insight in contemporary services marketing practive: why grouping customers is no longer enough. *Journal of Marketing Management*, 25(3/4):227–252, 2009.

[3] D. Benoit and D. V. den Poel. Improving customer retention in financial services using kinship network information. *Expert Systems with Applications*, 39(13):11435–11442, 2012.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] P. Cermak. Customer profitability analysis and customer life time value models: Portfolio analysis. *Procedia Economics and Finance*, 25:14–25, 2015.

[6] Z.-Y. Chen and Z.-P. Fan. Distributed customer behavior prediction using multiplex data: A collaborative mksvm approach. *Knowledge-Based Systems*, 35:111–119, 2012.

[7] M. Crowder, D. Hand, and W. Krzanowski. On optimal intervention for customer lifetime value. *European Journal of Operational Research*, 183(3):1550–1559, 2007.

[8] F. Dwyer. Customer lifetime valuation for marketing decision making. *Journal of Direct Marketing*, 11(4):6–13, 1997.

[9] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker, and S. Sriram. Modeling customer lifetime value. *Journal of Service Research*, 9(2):139–155, November 2006.

[10] D. Hinkle, W. Wiersma, and S. Jurs. *Applied Statistics for the Behavioral Sciences*. Houghton Mifflin, Boston, 2003.

[11] P. Kotler and K. Keller. *A framework for marketing management*. Pearson Education, England, 2012.

[12] V. Kumar and M. George. Measuring and maximizing customer equity: A critical analysis. *Journal of the Academy of Marketing Science*, 35(2):157–71, 2007.

[13] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[14] E. Malthouse. The results from the lifetime value and customer equity modeling competition. *Journal of Interactive Marketing*, 23(3):272–275, 2009.

[15] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-4.

[16] T. Mirzaei and L. Iyer. Application of predictive analytics in customer relationship management: A literature review and classification. In *Proceedings of the Southern Association for Information Systems Conference*, March 2014.

[17] S. Rosset, E. Neumann, U. Eick, and N. Vatnik. Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery*, 7(3):321–339, July 2003.

[18] T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. R package version 4.1-8.

[19] L. Valenzuela, E. Torres, P. Hidalgo, and P. Farias. Salesperson clv orientation's effect on performance. *Journal of Business Research*, 67(4):550–557, 2014.