Improving Response Modeling with Facebook Engagement Data

Louis Lecat (SUID: Lecat)

INTRODUCTION

Predictive modeling is now a proven technique for marketing optimization and is being used more and more often by the private sector. The data sets used so far are mainly customer purchase-related data, and sometimes web and email behavior.

Facebook has totally changed the online landscape over the past few years, and companies are now trying to leverage it as a marketing resource, spending more and more resources on their "social media face". This paper intents to come up with an original way of measuring the impact of Facebook activity on a company's revenue, and how companies can leverage their "social data" to improve predicting modeling.

The potential of this project is huge for the private sector: thanks to this brand new social data, companies could significantly improve their revenue by effectively leveraging high purchase-potential customers.

DATA

In this paper, we will intent to improve the effectiveness of response modeling for one selected company. Due to privacy issues we cannot disclose the name of the company, and we will refer to it as X-Comp. However, all we need to know for this study is that X-Comp is a retailer with a strong presence on Facebook.

The data that has been used to implement the models is of four types:

- Purchase data: all data relevant to X-Comp's customers direct interaction with the company: orders, order dates, order channels, orders amount, discounts, margins, product types and categories, order frequency, platinum memberships, etc.
- Email data: data gathered through all email marketing campaigns such as emails clicked, opened, unsubscriptions, dates, etc.
- Web data: all data gathered through customers interactions with the website, such as clicks, items browsed, items carted, visit dates, etc.
- Facebook data: data gathered on X-Comp's Facebook wall, i.e. posts, comments, links and likes with the dates and attributes associated.

DATA COLLECTION

This research project was realized in collaboration with a marketing analytics company, of which X-Comp is a client. The company provided access to the purchase data, the email data and the web data in a Microsoft SQL environment. These 3 types of data together will be referred to as "standard data" in this paper.

Facebook data, on the other hand was not available and integrated with purchasing data. It was gathered using a "crawling engine" to scrape all the public information available on X-Comp's Facebook page. Note that even though most Facebook users have restricted and private profiles, all their activity on a public Facebook page such as X-Comp's page is made public. This is why we are able to access their posts, comments, links and likes on this particular page.

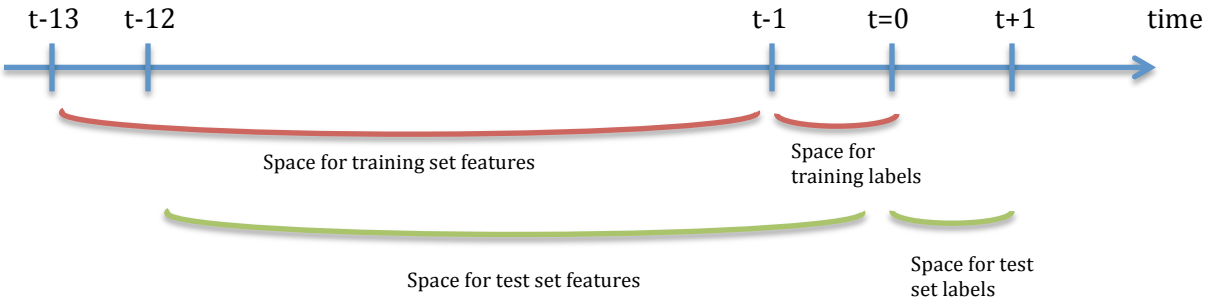
LINKING FACEBOOK DATA TO PURCHASE DATA: METHOD AND SUCCESS

All Facebook data from X-Comp's page was then imported in our SQL environment in order to be tied to the standard data.

The next challenge was to accurately link users' ID from the standard data to Facebook user IDs. Based on the data we have on hand from Facebook (first name, last name, location and gender), we computed a similarity score tying Facebook data with standard data for each customer. We chose to take into account only the customers presenting a similarity score higher than 85%, to make sure we did not mismatch any customer. As a result, we were able to match approximately 71% of Facebook users to existing customer IDs in our standard data database.

MODEL: LOGISTIC REGRESSION

Predictive modeling in our case consists of evaluating the probability of a customer to place a purchase in the upcoming month. Let's consider that we stand at t=0 (t in months).



We define our training set by looking at past data: for a given customer i, features are calculated over a 12 months period of time (between t-13 and t-1) and we associate him the label y(i) = 1 if the customer places a purchase between t-1 and t, or y(i) = 0 if he does not place a purchase during that month. We will then use the test set defined in green above (shifted 1 month forward) to predict whether customer i will place a purchase with X-Comp over the next month or not.

Our two best options to compute this likelihood are logistic regression and Naïve Bayes. On this project, we chose to work with logistic regression, since we are unsure of features correlation. We inherently need to test the model on several different features to optimize it, and logistic regression allows us to avoid the extra-

correlation induced by Naïve Bayes if we select a large number of features (e.g. let's say we work with the number of transactions realized over the past year, the average amount of these transactions, and the total amount of money spent).

Much of the implementation was realized through SQL: pre-processing of data, calculation of model features, storage of those in appropriate tables and computation of the features and labels sparse matrix. The actual logistic regressions themselves were run through Matlab using the exported sparse matrix. Once computed, coefficients were re-imported into SQL to make our predictions on the test set.

MODEL VALIDATION

A good model is a model that the company can leverage. Here is the method we used to evaluate the quality of a model: once we have computed the likelihood of purchase for customers 1..n during the upcoming month, we rank these customers in order of likelihood descending. Customers are then assigned to response buckets in that order. In the case of X-Comp, 85 response buckets were assigned each time, where customers in bucket 1 are the ones with a high likelihood of purchase and customers in bucket 85 have a low likelihood of purchase.

We then look at what actually happened over the predicted month along two metrics:

- How many **customers** actually placed a purchase in each bucket
- How much **revenue** was brought by customers in each bucket

A good model is a model in which most of the revenue and most of the customers placing a purchase are accounted for in the first buckets (the ones for which we predicted high likelihoods). We will visualize these results with the two corresponding graphs:

- Cumulative **number** of customers coming back against response buckets
- Cumulative **revenue** from customers coming back against response buckets

IMPLEMENTATION AND FEATURE SELECTION

Implementation was conducted in two steps. First, we used the standard data to implement the best response model as possible. Many iterations were conducted by refining our feature set and removing some of the features presenting a high correlation to each other. The final feature set was selected using the model validation described above: we kept the model scoring above all the others on both the plots.

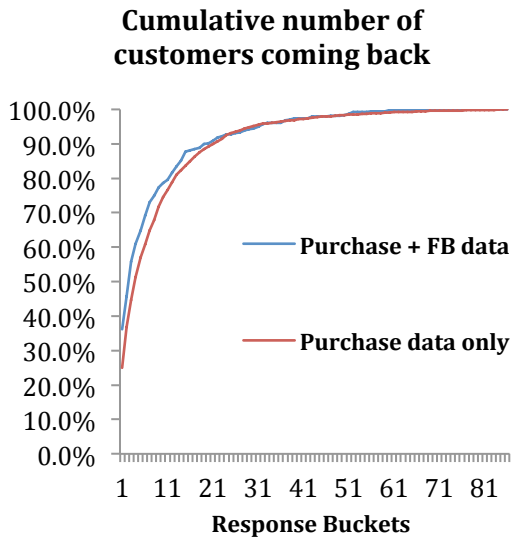
We then enlarged our feature space by including the Facebook data, namely number of posts, comments, links and likes (and filtering them out by the selected date ranges). Using the same feature selection method, and iterating as much as possible,

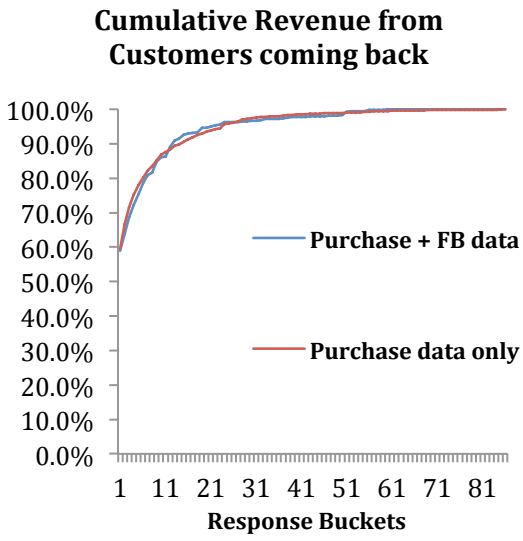
we came up with the best predictive model with Facebook data. Below is a table showing the final features selected in both cases:

Selected features	
Standard case	**With Facebook data**
log([PlatinumCustomer] +1)	log([PlatinumCustomer] +1)
log([DaysSince] +1)	log([DaysSince] +1)
log([TotalOrders] +1)	log([TotalOrders] +1)
log([TotalItems] +1)	log([TotalItems] +1)
log([AOV] +1)	log([AOV] +1)
[ReturnOrders]	[ReturnOrders]
[TimesOpened_0_30]	[TimesOpened_0_30]
[View_0_30]	[ItemsBrowsed]
[ItemsBrowsed]	[ItemsBought]
[ItemsBought]	[ItemsCarted]
[ItemsCarted]	[DaysSinceLastVisit]
[DaysSinceLastVisit]	[Posts]
	[Comments]
	[Links]
	[Likes]

RESULTS

The two graphs below show the compared results of the two models. A first comment is that we did not get any huge improvement from using the Facebook data. On the cumulative revenue side, it seems that we get no improvement at all, and the model with the standard data seems to fare slightly better on the top buckets.





However, we get a slightly significant enhancement on the measure of customers coming back: the model is notably more accurate at predicting repeat purchases on the top 25 buckets. Our hypothesis is that the Facebook features that we took into account in the model really make a difference for customers who are heavily active on X-Comp's Facebook page, and thus much more likely to be loyal customers of X-Comp and to fall in the top 25 buckets.

To double-check our predictions, we also compared movements of customers between buckets from one model to another. The table below shows these movements when we assign customers to 10 different buckets (instead of 85).

		MODEL WITH STANDARD DATA ONLY										
	Buckets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total customers**
MODEL WITH FB DATA	**1**	61%	12%	10%	9%	5%	2%	1%	0%	0%	0%	**100%**
	2	30%	31%	14%	11%	10%	3%	1%	0%	0%	0%	**100%**
	3	4%	42%	16%	8%	15%	12%	2%	0%	0%	0%	**100%**
	4	3%	10%	30%	14%	8%	15%	14%	6%	1%	0%	**100%**
	5	2%	4%	17%	22%	15%	9%	15%	13%	3%	0%	**100%**
	6	0%	1%	12%	17%	19%	13%	9%	21%	8%	0%	**100%**
	7	0%	0%	1%	15%	17%	15%	9%	10%	28%	3%	**100%**
	8	0%	0%	0%	3%	11%	23%	17%	10%	19%	17%	**100%**
	9	0%	0%	0%	0%	0%	8%	29%	22%	14%	26%	**100%**
	10	0%	0%	0%	0%	0%	0%	4%	17%	27%	52%	**100%**
	Total customers	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	

As expected, customers are mostly spread on the diagonal, and we see no outlier movements in the table (e.g. no customers were ranked in the top bucket in one model and the worst bucket in the other model), which assures us of the concordance of the two models.

IMPROVEMENTS AND CONCLUSION

Even though the use of Facebook data is crucial for companies, we believe this implementation did not show significant enough improvements for marketers to start using it on a permanent basis in response modeling. Nevertheless are the results on repeat customers predictions encouraging and we believe that predictive modeling will benefit from Facebook data in the future. Some areas for improvement of the models are:

- Constructing more features based on the Facebook data. Thus far we only looked at number of posts, comments, links and likes, but included nothing about the content and/or the quality of these. Some unsupervised learning techniques such as k-means could allow us to highlight clusters of active versus inactive users for instance.
- Including more data. We could only use public data from the company's Facebook page. Getting access to private data from users should help.