ANALYZING USER ACTIVITY BASED ON RFM MODELS COMPLEMENTED WITH WEBSITE VISITS AND SOCIAL NETWORK INTERACTIONS

Pavel Jašek

Department of Information Technologies
Faculty of Informatics and Statistics
University of Economics, Prague
pavel.jasek@vse.cz

Keywords

RFM Analysis, Customer Segmentation, Social Media

Abstract

Analyzing customer's behavior based on Recency-Frequency-Monetary value (RFM) modelling serves well both for customer segmentation purposes and for its predictive ability to hit customers with high probability of repurchase. This paper presents an innovative approach to extend data sources for RFM modelling apart from purchase data only to using website visits and social network interactions as factors of valuable customer activities that help to leverage the predictive power of a model. A comparison of transactional and enhanced model with visit-level data was conducted in order to demonstrate useful additional information. The paper outlines and discusses an incorporation of data sources about individual customer interactions on Facebook and Twitter into RFM analysis.

1. Introduction

(Birant, 2011) followed on work by (Bult & Wansbeek, 1995) and described RFM Analysis as a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions.

(Miglautsch, 2000) points out that the purpose of RFM is to provide a simple framework for quantifying customer behavior and that RFM is a superior method for selecting customers. The traditional approach to RFM Analysis as described by (Bult & Wansbeek, 1995) and with practical implementations in R software by (Ohri, 2012) and (Han, 2013) consists of dividing data into intervals of five breaks for every RFM component. This simplicity of quintiles is useful for visualization and practical application. Miglautsch also offers a form of weighting of R, F and M scores together.

Very little literature has explored different sources of customer interactions that can be valuable for RFM analysis. (Li, Lin, & Lai, 2010) combine opinion-mining techniques and adaptive RFM models to develop a framework to evaluate the influential capability of online reviewers and recommend appropriate ones to support word-of-mouth marketing. This paper aims to use similar approach as (Aggelis & Christodoulakis, 2005) that used RFM scoring for active e-banking users and also continues the previous work of (Novotny & Jasek, 2013) where indirect links to future behavioral measures of Customer Lifetime Value were proposed, with great focus on website interactions. Linking and integrating social media to company website is an important issue. (Smutny, Reznicek, Kalina, & Galba, 2013) state that only by using a proactive approach to identify user interaction, company will be able to reallocate its resources adequately and effectively manage marketing not only in the internet environment. This gives great motivation of integrating data from all possible sources of customer interactions together.

This paper compares two quantitative models of RFM analysis using transactional and visit-level data from the website. The author expects that visit data can update the information whether customer is "alive" in terms of engagement with company. As it proves that enhancement of transactional data is useful and adds value to the RFM analysis, several practical steps towards individual customer data from social networks Facebook and Twitter are described and analyzed in more detail.

1.1. Methodology

The paper used anonymized quantitative purchase and website visits data from a Czech online retailer obtained with permission. More information about the dataset is described in part 1.2. The dataset was divided into two parts of transactional data only and enhanced data with website visits tied to an individual customer. On each of this subsets a quasibinomial logistic regression model was constructed with the goal of estimating probability of purchasing given RFM characteristics of a customer. The model family was selected accordingly to (Han, 2013). In order to build the models, these subsets were divided into training and validation parts of same length. Proposed cut-off date for this 50/50 model validation is December 15, 2012.

As RFM repurchase probability presents an obstacle in hard to visualize 4-dimensional dataset, this paper followed an approach by (Han, 2013) that plots relationships between Repurchase Rate and each RFM component individually. Three submodels of same family (quasibinomial logit) were used to support such relationship visualization of Buy ~ Recency, Buy ~ Frequency, and Buy ~ Monetary value. Apart from that, two quasibinomial logit models of Buy ~ Recency + Frequency + Monetary value were built and compared.

1.2. Sample dataset description

For the purpose of this paper real-world data from a Czech online retailer was used. The business sells fashion primarily for mid-aged women and regularly twice a year changes a large portion of product catalogue in order to match summer and winter season. According to the customer base classification done by (Fader & Hardie, 2009, p. 63), this dataset has non-contractual relationship with customers and continuous opportunities for transactions.

This historical log contained 77 289 logged-in visits to the e-commerce website and 33 613 online purchases made by 29 589 different customers from the time period of September 1, 2011 to March 31, 2014 (134 weeks in total). The data source was Google Analytics. The data was anonymized and none of attributes could be used to link pro personally identifiable information.

Table 1 exposes an example of this dataset for one specific customer. Customers are often visiting the website without intention to purchase and it can be seen that purchases vary with quantity and amount of goods sold.

Date	Client ID	Visits	Transactions	Amount (CZK)	Quantity
2011-11-02	22862	1	0	0	0
2011-11-17	22862	1	2	1 540	10
2011-11-26	22862	1	1	434	4
2011-11-30	22862	1	0	0	0
2012-05-05	22862	1	1	1 120	5

Table 1. Data for online visits and purchases. Source: Sample dataset, filtered for a specific customer.

Unfortunately, the sample dataset does not include data from social networks. There were two reasons of not obtaining such data: 1) the company itself doesn't operate any Facebook page or Twitter profile, 2) due to the demographical characteristics of the target group only a tiny fraction of customers are supposed to actively use social networks.

2. Comparing RFM models

This chapter describes two quasibinomial logit models based on RFM metrics. Formal model comparison is done later in part 2.3.

2.1. RFM with transactional data

The work of (Han, 2013) includes implementation of RFM analysis in R software with all necessary calculations to calculate recency from last purchase date until now and to sum up customer frequency of interactions and her profit or revenue in past. Although scatter plot seems to be a good way to shed some light on the distribution of the data, for more explanative reasons also individual models were constructed.

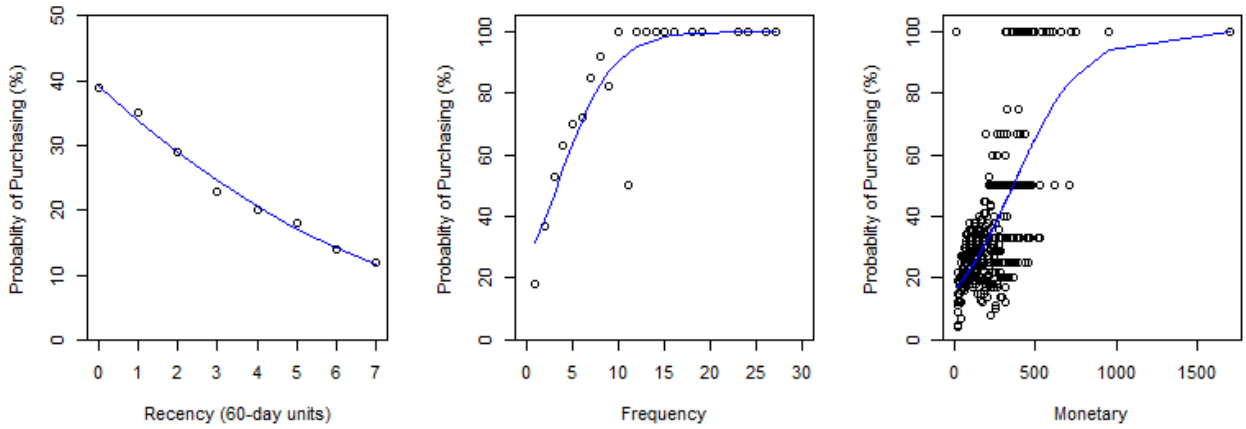

Figure 1. Visual interpretation of RFM components for transactional data. Each figure plots actual data and quasibinomial logit model for Buy ~ Recency, Frequency, Monetary. Source: Author, based on sample dataset, computed with R software.

Figure 1 shows a scatter plot of RFM components and its probability of repurchasing and quasibinomial logit models for every RFM component. Details of coefficients and their significance are listed in Table 2. Figure 1 clearly shows strong relationship of Recency to Probability of repurchasing with natural interpretation of customers recently active within last three months to have twice as high probability of purchasing as customers with last order a year ago. This negative relationship would also be expected in a model with visit-level data described later in part 2.2. The role of Frequency is evident: customers with high number of transactions (with frequency higher than 10 transactions) seem to be very loyal as their probability of repurchasing is greater than 80 %. Yet the company should be worried about new acquired customers with only 18 % probability of repurchasing.

	Recency model		Frequency model		Monetary value model	
Coefficients	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	-0.4439	0.0303 ***	-1.1235	0.5180 __*	-1.6716	0.0851 ***
Recency	-0.2263	0.0083 ***	X	X	X	X
Frequency	X	X	0.3368	0.0772 ***	X	X
Monetary	X	X	X	X	0.0046	0.0003 ***

Table 2. Estimated coefficients of quasibinomial logit models for Buy ~ Recency, Frequency, Monetary. Coefficient estimates are significant at the 0.1% level (marked as *) and 5% level (marked as __*). Source: Author, based on sample dataset with transactional data, computed with R software.**

2.2. RFM Model enhanced with visit-level data

As an additional step, individual customer-level website visit data can be added, as seen within Table 1. Customer identification is achievable by current logged-in session information or by setting up long-term browser cookie that identifies future visits of a customer.

One can expect that visit data can update the information whether customer is “alive” in terms of engagement with company. The model parameters are shown in Table 3. Changes between such data (shown on Figure 1 and Figure 2) are better explained in part 2.3 of this paper.

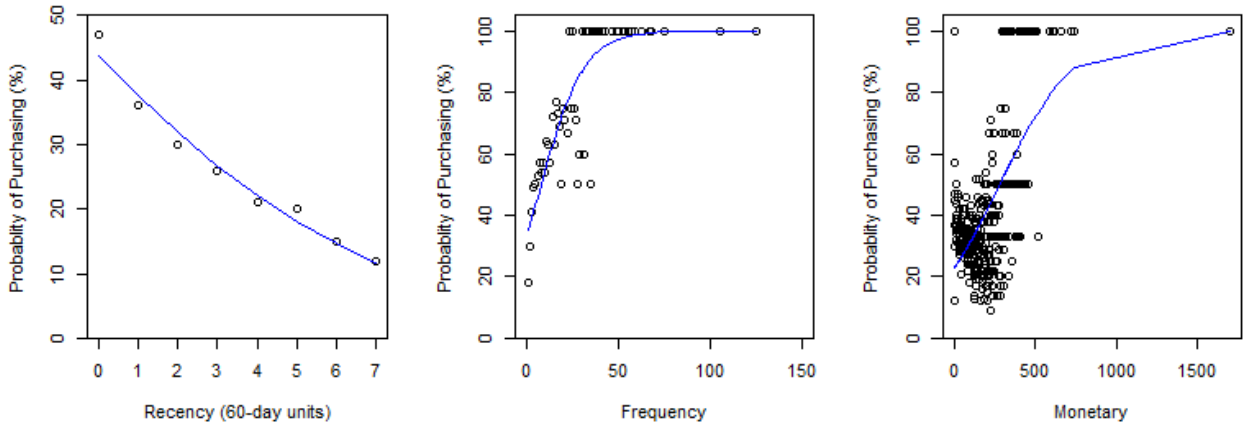

Figure 2. Visual interpretation of RFM components for visit-level data. Each figure plots actual data and quasibinomial logit model for Buy ~ Recency, Frequency, Monetary. Source: Author, based on sample dataset, computed with R software.

	Recency model		Frequency model		Monetary value model	
Coefficients	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept	-0.2510	0.0588 _**	-0.7026	0.2392 _**	-1.2228	0.0777 ***
Recency	-0.2514	0.0162 ***	X	X	X	X
Frequency	X	X	0.0861	0.0115 ***	X	X
Monetary	X	X	X	X	0.0043	0.0003 ***

Table 3. Estimated coefficients of quasibinomial logit models for Buy ~ Recency, Frequency, Monetary. Coefficient estimates are significant at the 0.1% level (marked as *) and 1% level (marked as _**). Source: Author, based on sample dataset with visit-level data, computed with R software.**

2.3. Model Comparison

Coefficients and other statistics of studied models have to be discussed. Table 4 shows important metrics for each model and Figure 3 clearly indicates the difference between models.

	Transactional model		**Visit-level model**	
Coefficient	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Intercept	-1.6768	0.0647 ***	-0.7066	0.0537 ***
Recency	-0.1500	0.0107 ***	-0.2148	0.0100 ***
Frequency	0.5915	0.0247 ***	0.1719	0.0091 ***
Monetary	0.0009	0.0003 ***	0.0011	0.0003 ***

Table 4. Comparison of coefficients and other statistics for two studied models. Quasibinomal logit model for Buy ~ Recency + Frequency + Monetary. All coefficient estimates are significant at the 0.1% level (marked as *). Source: Author, based on sample dataset, computed with R software.**

Negative values for Recency were as expected in part 2.1. Very interesting difference can be seen in terms of Frequency: visit-level model weakens the marginal importance of this coefficient by 70%.

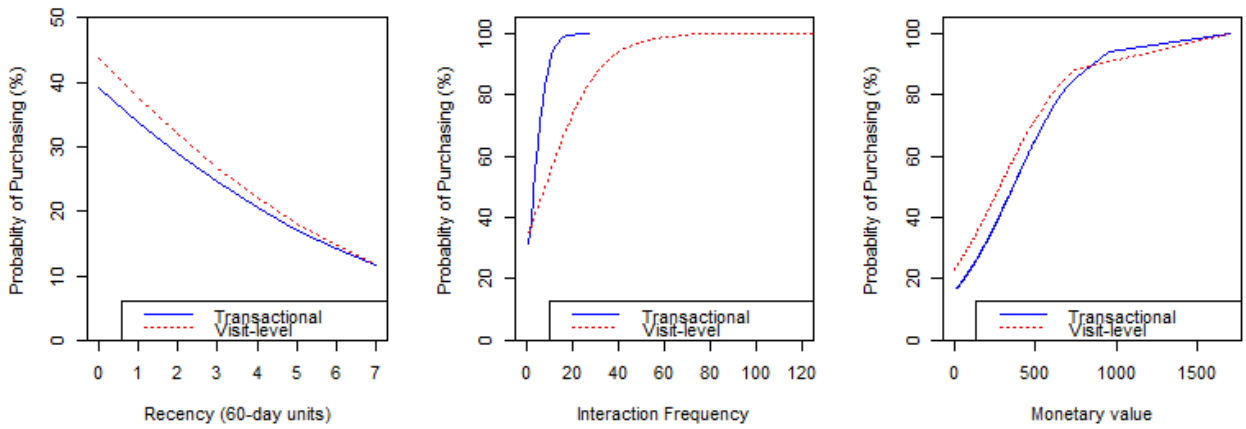

Figure 3. Visual comparison of RFM probabilities when using transactional and visit-level models based on three quasibinomial logit models for Buy ~ Recency, Frequency, Monetary. Source: Author, based on sample dataset, computed with R software.

The role of Monetary value is very doubtful in overall result: the estimate 0.0011 would mean that high value purchase of 1700 would add just 1.87% into resulting probability of repurchase.

3. Towards Enhancing RFM Models with Social Network Data

Similarly to the shift from transactional data to incorporating "internal" data from company's owned channels, the focus will now be placed into adding "external" channels data as well. These include customer activities on social networks which may or may not be managed by a company. Also, demographic, interest and personal data about specific customers could be retrieved as well. According to (Pavlicek & Pechar, 2012), such personal data is easily available by careless social media users.

In the scope of this paper the author outlines possible ways to enhance RFM analysis with interaction data from two largely known social networks Facebook and Twitter. Such data could be used in a similar way as data from website visits.

3.1. Facebook

In case of Facebook the interest is on customer interactions for specific content a brand publishes. Graph API described in (Facebook, Inc., 2014) offers the */likes* edge for post nodes and */comments* edge for comment replies to a post. Also, */posts* edge for a specific Facebook page returns published posts, so all data retrieval can be done programmatically.

An output array represents each of the people who liked the object, specifying user with application-specific ID of this person's user account and full name. An example of HTTP/1.1 GET request on host *graph.facebook.com* can be */661704773864624/likes*. The output contains following data field:

{ "data": [{ "id": "10201427875893354", "name": "Alice Placeholder" },

{ "id": "10202251797591562", "name": "Bob Placeholder" }] }

This list of customer identifiers can be tied with internal CRM records of a customer. A possible implementation is outlined in Table 3. The relationship between Internal Customer ID and Facebook User ID is supposed to be 1:1, but one user can like or comment multiple posts.

Internal Customer ID	Facebook User ID	Post ID	Liked?
123	10201427875893354	661704773864624	1

Table 5. Structure of data retrieved and processed from Facebook API. Source: Author

3.2. Twitter

In case of Twitter the focus is on customer's interactions with a specific content a brand publishes or when a customer mentions a brand. For the former one, Twitter's API (Twitter Inc., 2014) doesn't allow direct requests as Facebook does, so a developer has to work with the list of 20 most recent Tweets favorite by a specific user. An example of an authenticated HTTP/1.1 GET request on host *api.twitter.com* can be */1.1/favorites/list.json?user_id=14979466*, where *user_id* value is a specific Twitter user identifier. The output contains following data fields (some other were omitted for simplicity):

[{ "created_at": "Fri May 02 06:51:28 +0000 2014", "id": 462122180470640640,

"retweet_count": 2, "favorite_count": 5,

"favorited": true, "retweeted": false, }]

Although outlined implementation would require additional development, the final output shown in Table 4 demonstrates the desired data structure. The relationship between Internal Customer ID and Twitter User ID is supposed to be 1:1 – yet one should consider possible multiple Twitter accounts for a customer in specific situations, e.g. director of a company who tweets both by her personal account and by her company's account. Favorites and Retweets would be linked to Twitter User ID with an obvious 1:N relationship.

Internal Customer ID	Twitter User ID	Tweet ID	Favorited?	Retweeted?
123	14979466	661704773864624	1	0

Table 6. Structure of data retrieved and processed from Twitter API. Source: Author

3.3. User Profile Matching

Integration of customer-level data from social networks relies on one important assumption that customer can be accurately identified by 1:1 relations between internal and external user identifiers. This topic was largely studied by different authors, such as (Malhotra, Totti, Wagner, Kumaraguru, & Almeida, 2012), (Paridhi, Ponnurangam, & Anupam, 2013) or (Raad, Chbeir, & Dipanda, 2010).

The work of Paridhi et al. found 39% Facebook identities to a studied group of Twitter users. An algorithm by Anshu et at. matched profiles with 64% accuracy, gaining impressive 98% of accuracy, 99% of precision and 96% of recall using the most promising set of features.

Methodologies for user profile matching are outside the scope of this paper, but it is clearly evident that automated processing of user profiles is highly reliable and with regards to the privacy and legal requirements can be used to tie internal and external customer data.

4. Conclusion

The paper outlined possible ways of enhancing traditional RFM Analysis with additional data that demonstrate customer's activity and engagement with the company's marketing channels. An anonymized data sample from one Czech online retailer was analyzed to support this theory.

Studied comparison of transactional and visit-level data on data sample showed that Recency calculated from visit-level data shows higher probability of repurchasing, meanwhile Frequency calculated from both interactions of transactions and visits doesn't always show that a customer would have higher probability of repurchasing in case of her frequent visits to a website.

Because such enhancement of RFM Analysis proved to be valuable, in part 3 additional data sources as Facebook and Twitter were discussed. Information retrieval of customer interactions can be done automatically using API. Such data integration raised important issue of user profile matching. The data could be treated in the same manner as with website visits: enhancing RFM model with new interactions to better understand long-term purchase behavior of a customer. A combined log of customer purchases, visits to website, comments, likes favorites and retweets would thus serve in a model for probability to repurchase. Unfortunately, this enhancement could not be done with the sample dataset.

As a future work, further exploration of social network data and proposal of new links between customer multi-channel engagement and Customer Lifetime Value are needed. Also, the social network data described in part 3 should be used in a practical case study of RFM analysis.

5. Acknowledgment

This paper was written with the financial support of IGA grant F4/18/2014.

6. References

Aggelis, V., & Christodoulakis, D. (2005). RFM analysis for decision support in e-banking area. *WSEAS Transactions on Computers 4.8* (pp. 943-950). The World Scientific and Engineering Academy and Society.

Birant, D. (2011). Data Mining Using RFM Analysis. *Knowledge-Oriented Applications in Data Mining.* InTech.

Bult, J. R., & Wansbeek, T. (1995). Optimal Selection for Direct Mail. *Marketing Science* (pp. 378-394). Catonsville: Informs.

Facebook, Inc. (2014). *The Graph API.* Retrieved from Facebook for Developers: https://developers.facebook.com/docs/graph-api

Fader, P. S., & Hardie, B. G. (2009). Probability Models for Customer-Base Analysis. *Journal of Interactive Marketing*, pp. 61-69.

Han, J. (2013). *Calculating Customer Lifetime Value with Recency, Frequency, and Monetary (RFM).* Retrieved from Data Apple: http://www.dataapple.net/?p=133

Han, J. (2013). *RFM Customer Analysis with R Language.* Retrieved from Data Apple: http://www.dataapple.net/?p=84

Li, Y.-M., Lin, C.-H., & Lai, C.-Y. (2010). Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications* (pp. 294-304). Elsevier.

Malhotra, A., Totti, L., Wagner, M. J., Kumaraguru, P., & Almeida, V. (2012). Studying User Footprints in Different Online Social Networks. *International Conference on Advances in Social Networks Analysis and Mining* (pp. 1065-1070). Istanbul: IEEE Computer Society.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing* (pp. 67-72). Nature Publishing Group.

Novotny, O., & Jasek, P. (2013). Enhancing Customer Lifetime Value with Perceptual Measures Contained in Enterprise Information Systems. *CONFENIS - 2013* (pp. 211-220). Linz: Trauner Verlag.

Ohri, A. (2012). *Doing RFM Analysis in R.* Retrieved from Decision Stats: http://decisionstats.com/2012/03/27/doing-rfm-analysis-in-r/

Paridhi, J., Ponnurangam, K., & Anupam, J. (2013). @i seek 'fb.me': identifying users across multiple online social networks. *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 1259-1268). Geneva: International World Wide Web Conferences Steering Committee.

Pavlicek, A., & Pechar, Z. (2012). Availability of User's Personal Data on Facebook. *IDIMT-2012 – ICT Support for Complex Systems* (pp. 377–380). Linz: Trauner Verlag universitat.

Raad, E., Chbeir, R., & Dipanda, A. (2010). User Profile Matching in Social Networks. *NBIS '10 Proceedings of the 2010 13th International Conference on Network-Based Information Systems* (pp. 297-304). Washington, DC: IEEE Computer Society.

Smutny, Z., Reznicek, V., Kalina, J., & Galba, A. (2013). Interaction of Social Media and Its Use in Marketing Management. *IDIMT-2013 Information Technology Human Values, Innovation and Economy* (pp. 167–174). Linz: Trauner Verlag.

Twitter Inc. (2014). *REST API v1.1 Resources.* Retrieved from Twitter Developers: https://dev.twitter.com/docs/api/1.1