Customer Wallet Share Estimation for Manufacturers based on Transaction Data

Xiang Li[1,2], Ali Shemshadi[2], Łukasz P. Olech[2,3], and Zbigniew Michalewicz[1,2,4,5]

[1] School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia
http://www.adelaide.edu.au
[2] Complexica Pty Ltd.
155 Brebner Drive, West Lakes, SA 5021, Australia
{xl,as,lo,zm}@complexica.com
http://www.complexica.com
[3] Wroclaw University of Science and Technology
Wybrzeze Stanislawa Wyspianskiego 27, 50-370 Wroclaw, Poland
https://pwr.edu.pl/en/
[4] Institute of Computer Science, Polish Academy of Sciences,
Ordona 21, 01-237 Warsaw, Poland
https://ipipan.waw.pl/en/
[5] Polish-Japanese Academy of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
https://www.pja.edu.pl/en/

Abstract. The value of customers for any business cannot be overemphasised, and it is crucial for companies to develop a good understanding of their customer base. One of the most important pieces of information is to estimate *the share of wallet* for each individual customer. In the literature a related concept is often referred to as *customer equity* that provides aggregated measures such as the business market share. The current trend in personalising marketing campaigns have led to more granular estimation of wallet share, than the entire customer base or aggregated segments of customers. The current trend in personalising marketing and business strategies have lead to more granular estimation of wallet shares than the entire customer base or aggregated segments of customers. Existing research in this area requires access to additional information about customers, often collected via various surveys. However, in many real-world scenarios, there are circumstances where survey data are unavailable or unreliable. In this paper, we present a new customer wallet share estimation approach. In the proposed approach, a predictive model based on decision trees facilitates an accurate estimation of wallet shares for customers relying only on transaction data. We have evaluated our approach using real-world datasets from two businesses from different industries.

Keywords: Wallet Share Estimation · Customer Equity · Random Forest · Real-World Case Study

1 Introduction

Wallet share can be defined as the ratio of money that a customer spends with a brand compared to all of his/her expenditure on similar brands that can be considered as competitors. It can help businesses to understand and evaluate their relationship with their customers [7]. Thus it is crucial for businesses to have the ability to measure the share of wallet of their customers. Different businesses, including manufacturers and distributors, usually record a significant amount of data on their customers.

One of the main approaches to estimating share of wallet is to use Voice of Customer (VoC) data [13]. VoC data can be collected by sales representatives or call centres either by surveying customers (pull) or by monitoring the messages sent from customers when they initiate. Many industries such as retailers [3] or banks [2] have access to an extensive amount of VoC data. However, this is not the case manufacturers, which usually do not possess a well-sized sample of VoC data compared to other industries, even when they also operate as distributors.

Research in this area has identified different requirements in real-world scenarios of wallet share estimation. Thus different trends can be observed in the literature including "the analysis of customer wallet share and its impacts in different environments" and "the development of new approaches to estimate the share of wallet for customers based on the availability of the data". In this paper, our focus is more on the latter while working on a novel application area (manufacturers) from the former trend's perspective.

Given that a manufacturer has access to transaction data only through retailers and has a limited amount of VoC data through the sale process, we investigate a novel approach for measuring the share of wallet for manufacturers. The main research questions, which we aim to answer in this paper, are as follows:

1. How is it possible to accurately measure the share of wallet for individual customers for a given manufacturer based only on transaction data?
2. What are the most important features of transaction data to build an effective predictive model of wallet share?
3. Can decision tree-based modelling be deployed to facilitate a real-life predictive model for wallet share estimation?

Fig. 1 illustrates our scenario. In this paper, we focus on manufacturers that distribute products to their customers via a chain of retail shops. Dashed lines denote the flow of material, and solid lines denote the flow of information.

Occasionally, a manufacturer may contact its customers for a survey, however, the information collected usually does not represent the general population due to its limited sample size. In order for the manufacturer to optimise its offers and promotions in different areas, it needs to estimate its wallet share with each customer. In our scenario, a decision tree-based model is developed to estimate the wallet share solely through transaction data that have been collected through retail shops.

The rest of this paper is organised as follows. First, an overview of the related research is presented in Section 2. Section 3 describes the proposed approach to

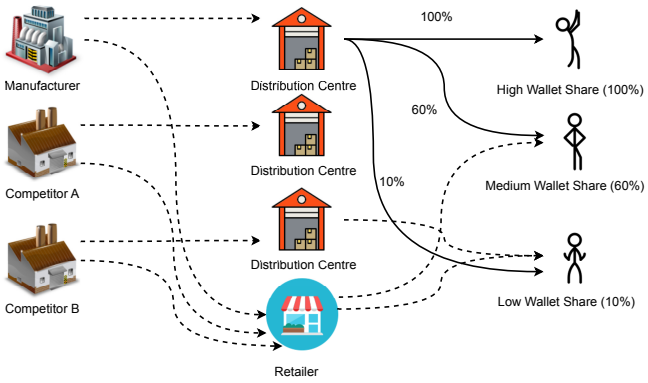

Fig. 1. Conceptual illustration of wallet share scenario

estimate wallet share estimations based on transaction data for manufactures. Details of the evaluation results and experiments on two real-world datasets are presented in Section 4. Section 5 concludes the paper and provides some direction for future research.

2 Related Work

In this section, we review the research related to our work. Wallet share estimation is a fundamental problem that has been investigated by many researchers for more than a decade [6]. Different methods have been proposed by different researchers to tackle the challenges present in this area.

In the literature, two main research trends are identified. The first trend focuses on the analysis of customer wallet share and its impacts in different environments. At the highest level, the impact of wallet share has been assessed as crucial in both Business to Business (B2B) as well as Business to Customer (B2C) [6] environments. At the lower level, research narrows it down to particular business domains (e.g., retail, wholesale, manufacture).

In the domain of retailing, research has demonstrated a relationship between customer satisfaction and the share of wallet [10]. While this relationship has been described as positive, yet it is being considered relatively weak. However, research in some particular areas of retailing (e.g., retail banking, massmerchant retail, and Internet service providers) shows a more sophisticated correlation between the share of wallet, customer satisfaction and other business goals [5, 1].

The second trend (which is more relevant to the research presented in this paper) focuses on new approaches for estimating the share of wallet for customers when provided with a different level of available data. Distinguishing criteria in this area are the prediction (estimation) approach, as well as the input data.

The research prior to 2005 considered two main approaches for wallet share estimation: top-down and bottom-up. In the former one, the market share is dis-aggregated, whereas in the latter the share for individuals are directly estimated and then aggregated [12]. In the following years, a white-box modelling based on regressive analysis has been adopted to develop predictive models [11]. The advantage of this approach is its simplicity, although its usefulness would be limited, particularly in more complex, real-world environments. Another approach uses estimates of customer potential by assuming optimistic conditions, which is referred to as customer opportunity [14]. Based on such an approach, a number of criteria are picked and used to compare the similarity of each customer to the set of customers whose predicted opportunities and actual sales are a close match. However, one limitation for such an approach is that in many cases, the customer potential can not be verified exactly and thus, the validity of all estimates are subject to assumptions.

Finally, addressing requirements based on input data specifications is a trending field of research in oCthis area — and this is directly related to this paper. Many papers have addressed a number of different requirements. In particular, the use of V data is one of the main approaches to estimate the share of wallet [13]. VoC data can be collected using survey data or unstructured data gathered. However, in many circumstances, VoC data are not available. In this case, existing approaches have tackled data availability issues by utilising transaction data for credit cards and focusing on inter-purchase times [2]. However, unlike manufacturers, credit card companies usually hold many records on their customers' transactions. Other research, which addresses the same issue, do not rely on real-world data and need to be further extended to include only the transaction data [4].

3 Methodology

This section starts with a brief introduction of the available raw data (Section 3.1). Then Section 3.2 presents the set of features extracted from the raw data. Estimation algorithm details are reported in Section 3.3.

3.1 Datasets

In this research, two datasets from different Australian manufacturers are used. The first one came from the paint industry and the second one — from a major producer of air conditioning products. In both cases, the companies behave both as a manufacturer and a distributor. As distributors, they have performed surveys on selected customers. But the coverage of these surveys was limited and skewed towards positive feedback. In our research, we have used the existing survey results as the training input to build the estimation models.

For each dataset, transaction data up to 3 years and around 250 customers are available. Customers are usually small or medium-sized enterprises, e.g., builders, handymen, electricians or painters. The data are extracted directly from

the manufacturer's sales database. The transaction data contained: *Order Date*, *Customer ID*, *Product ID*, *Product Quantity*, and *Product Price*. The response variable is an integer *walletshare* ranging from 0 to 100. 0 means the customer doesn't buy anything, and 100 means the customer buys everything needed from this manufacturer. The value was provided by sales representatives who had high confidence in the reported score. This confidence arises from a number of reasons, e.g., long-lasting cooperation, built trust. No data from the survey is used except the response variable *walletshare*.

3.2 Factorisation

The raw data is factorised into a row-based matrix, where each row represents one customer. The columns of the matrix are the features extracted from the raw data. The list has been created after extended discussions with domain experts. The main intuition behind those features is that high wallet share customers should buy products in a more consistent way than low wallet share customers. Also, historical peak sales could be helpful to identify the customer's business size.

The extracted features are:

1. *average_12_months*: customer average monthly spent in the last 12 months
2. *average_36_months*: customer average monthly spent in the last 36 months
3. *average_12_36_Ratio*: feature 1 / feature 2
4. *top_1_month_spent*: the highest monthly spent in the last 36 months
5. *top_3_months_avg_spent*: the average of top 3 highest monthly spent in the last 36 months
6. *top_6_months_avg_spent*: the average of top 6 highest monthly spent in the last 36 months
7. *avg12_top_1_month_ratio*: feature 1 / feature 4
8. *avg12_top_3_months_ratio*: feature 1 / feature 5
9. *avg12_top_6_month_ratio*: feature 1 / feature 6
10. *std_12_months*: the standard deviation of monthly spent
11. *spring_average*: the average monthly spent in Mar, Apr, May
12. *summer_average*: the average monthly spent in Jun, Jul, Aug
13. *autumn_average*: the average monthly spent in Sep, Oct, Nov
14. *winter_average*: the average monthly spent in Dec, Jan, Feb
15. *month_with_purchase_in_12_months*: no. of months with at least one purchase in the last 12 months
16. *month_with_purchase_in_36_months*: no. of months with at least one purchase in the last 36 months

The final column of the matrix was *wallet_share* representing the response value.

The first fourteen features (numbered from 1 to 14) come in two flavours: one that provides dollar values (and we use *dollar_* to precede the name of the feature), and the second one that provides the number of different product purchased (and we use *products_* to precede the name of the feature). For

example, the original feature *average_12_months* is replaced by two features: *dollar_average_12_months* and *products_average_12_months*. So in total, we have 30 features columns (28 features generated from the first fourteen plus features 15 and 16) and *wallet_share* is the response column.

3.3 Estimation Model and Synthetic Data

The model is build using the Random Forest (RF) algorithm [9], with the number of trees set to 50. Training is done by 75% – 25% random split and 10-fold cross validation. RMSE is used as the evaluation metric.

In addition, as the raw data is skewed, there are only a few cases for training sample of very low wallet share. This is very common in all real-world cases, since customers may not want to put low numbers on the survey. Thus, we have created synthetic data by introducing new empty entries and then set every column to a very low value. In addition to that, we also duplicated low wallet share samples and applied a small random variance to the feature values to create slightly different ones. Furthermore, there are many unlabeled customers in the transaction data that we could confidently assign a very low wallet share score without much analysis. For example, it is possible to assign a wallet share score of 0% to all customers that spent less than $100 during the last 12 months.

4 Results

In this section, the results of experiments are presented. First, Section 4.1 shows the importance of features for predicting the share of wallet. In Section 4.2, the accuracy of the developed method is compared with other existing methods. The impact of different training datasets and the additional synthetic data are investigated in Section 4.3 and Section 4.4, respectively.

4.1 Feature Selection

We have deployed the Boruta algorithm to select the most important features [8]. By default, Boruta runs Random Forest internally, testing each original feature against randomly generated features to check whether the original features can improve the prediction accuracy. Among the randomly generated features, the best performing one has been named ShadowMax and the mean performing random feature is called ShadowMean. Then, each original feature is compared with the randomly generated features. A feature that contributes positively should perform better than the best random feature (ShadowMax).

Results are shown in Fig. 2 for the paint dataset. Similarly, Fig. 3 shows the result for the air conditioner dataset. The randomly generated Shadow features are coloured as blue. We have validated the results with domain experts to assure the soundness of our approach.

Based on the comparison with the Shadow features, the features shown in green are confirmed, the features in yellow are questionable, whereas the algorithm suggests rejecting the features in red.

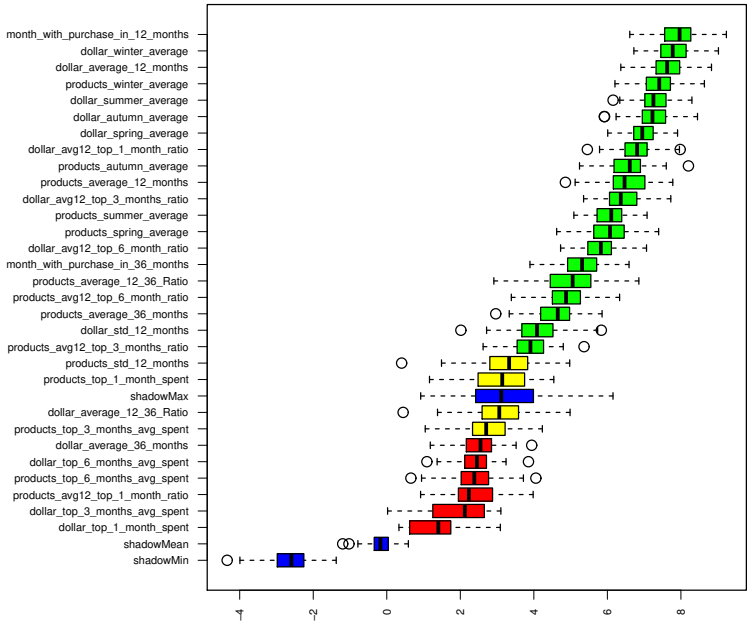

Fig. 2. Feature Importance for the Paint Dataset. Shadow features have corresponding blue bars

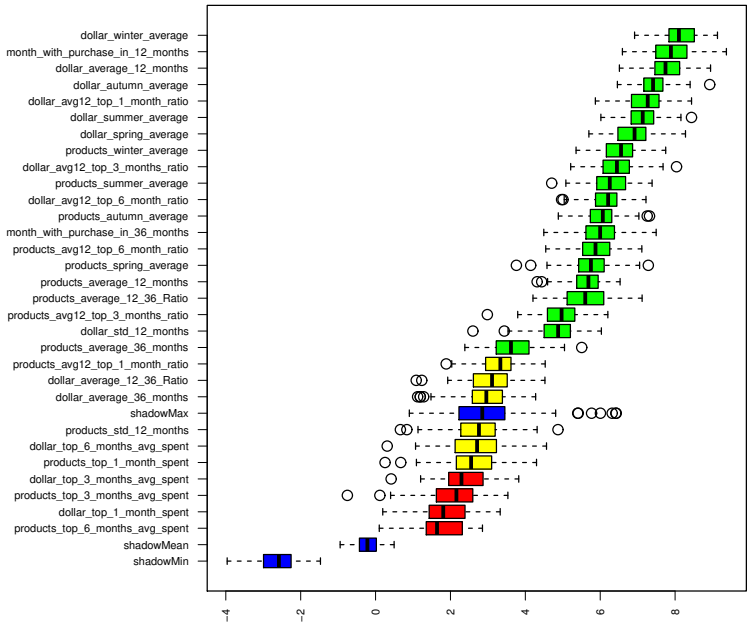

Fig. 3. Feature Importance for Air Conditioner Dataset. Shadow features have corresponding blue bars

For the paint dataset, the *month_with_purchase_in12_months* is the best-performing feature, as expected by domain experts. The feature *dollar_std_12_months* is not performing as expected: it outperforms ShadowMax only by a small margin. After some discussions with domain experts, we concluded that this might indicate that the customer purchases do have a natural variance between months, due to public holidays, seasonality, and most importantly account budget. Each customer has an account budget and could buy (to the budget limit) and pay later. When the customers are running out of budget, they simply go to another vendor (this is not captured in our model).

Around one-third of the features have questionable or near-random performance but all performing better than the ShadowMean feature. We believe they may still contribute to the predictive model, especially if applied to a different industry. Thus we did not reject any of them.

The results for the air conditioner dataset are very similar to those of the paint dataset. Here, seasonality related features show slight more importance than the cases in the paint dataset.

4.2 Accuracy Comparison

Our approach (Random Forest (RF)) is compared with two other algorithms: SVM and Linear Regression (LR). These algorithms are quite popular and have been tested in a variety of applications. Tables 1 and 2 present the accuracy of the results in the paint and air conditioning datasets, respectively.

Table 1. Accuracy (RMSE) on Paint Dataset

Paint	**RF**	**SVM**	**LR**
0 <15%	5.1	6.2	2.5
15<50%	16.3	16.8	18.4
50<80%	17.2	17.4	18.1
80<100%	21.6	22.8	22.3
Overall	16.4	17.2	18.5

Table 2. Accuracy (RMSE) on Air Conditioner Dataset

Air Cond.	**RF**	**SVM**	**LR**
0 <15%	3.2	5.1	4.4
15<50%	17.9	19.3	18.9
50<80%	17.8	20.1	21.1
80<100%	20.3	19.9	20.5
Overall	18.7	19.2	20.5

As shown in both tables, the results have been grouped into 5 segments, each row representing a different customer segment. The segments are based on customers' wallet share value, and marketing activities usually target those segments separately. For example, the 0 to 15% wallet share group consists mainly of occasional customers, and the 80 to 100% group should be all loyal customers. Due to data gathering imperfections, the lower bound of RMSE is estimated to be 10%. It stems from the fact that some customers were sampled multiple times during the data gathering process and some sales representatives assigned different wallet share values when surveyed at different times. This has been con-

firmed by the data provider, as there is no perfect data, and the system should not over-fit the training set.

Random Forest shows the best overall performance in both test sets, and in nearly all the wallet share segments. However, it still scores an unimpressive 5.1 and 3.2 RMSE in the lowest wallet share group and 21.6 and 20.3 RMSE in the highest group. A limited number of training samples in those two groups are the main reason for such results.

SVM holds the middle position. In all but the 80% to 100% group in the air conditioner group, SVM has a lesser accuracy than Random Forest, but it seems to have less deviation between other groups.

Linear Regression shows the worst overall result, but it performs well in the lowest group (0 to 15%). This is probably a result of the use of synthetic data in the lowest group. But in all other groups, the performance of Linear Regression is inferior to the other two algorithms.

Additionally, the two businesses which provided the data have tested the model in real life. In both cases, this testing (evaluation) has been conducted for more than a year now, and the feedback has constantly been very positive. The estimation accuracy is in-line with the results shown in the tables.

Meanwhile, the data providers also keep collecting customers surveys and self-stated wallet shares are one of the focusing points. We found that the wallet share estimation system can also be used to identify inaccurate records in the survey results if the self-stated wallet share in the survey differs too much with the estimation. One of the examples of incorrect survey results could be: reporting 100% wallet share by a customer who has spent only 10 dollars with the business.

However, the users of our model have occasionally detected cases where the system appears to assign a low wallet share score to some of the known loyal customers who should have rather a high value of the wallet share. A detailed analysis of those individual cases often revealed that the customer in question has spent all his/her budget and stopped buying temporally. Due to the ad-hoc nature of such cases, we can consider them as outliers.

4.3 Training Size vs. Performance

Many businesses have overlooked the potential to apply data science to their operations. They may have heard the term big data and are afraid that they have not accumulated enough. Furthermore, it is unknown whether the inclusion of all historical data or all available data improves the accuracy of the results. Thus, we measured the accuracy of the results based on the varying size of the training data.

As shown in Fig. 4, the performance gradually reaches a plateau as the size of the training approaches 80. As we mentioned earlier, the model has been tested in real life. This indicates that it is possible to perform wallet share prediction with good accuracy even with a relatively small dataset.

Similar to the first dataset, the second dataset shows a gradually reducing error rate when the training size approaches 80. Fig. 5 summarises this test.

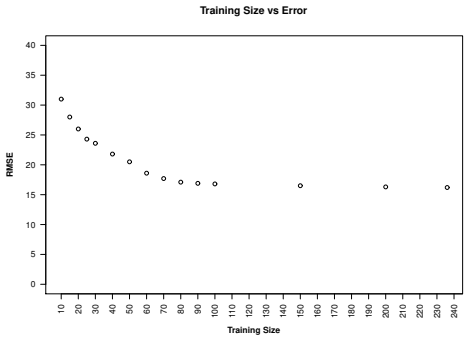

Fig. 4. Paint Training Size vs Error

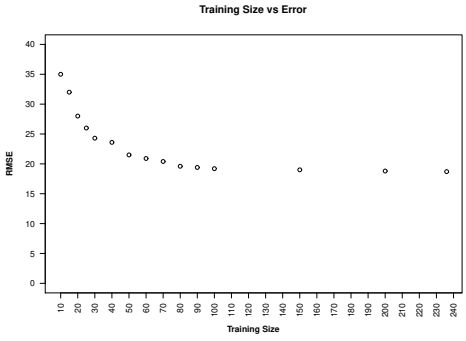

Fig. 5. Air Conditioner Training Size vs Error

4.4 Effect of Synthetic Data

As mentioned in section 3.3, skewed data represent a common issue in real-world problems. In our case, we had very limited samples from the lowest group (wallet share between 0 and 15%), Table 3 presents the estimation accuracy using only the original data (paint dataset). Table 4 presents the estimation accuracy result in the air conditioner dataset using only the original data.

Table 3. Accuracy (RMSE) on Paint Dataset without Synthetic Data

Paint	**RF**	**SVM**	**LR**
0 <15%	9.5	8.7	6.4
15<50%	16.5	16.6	17.4
50<80%	16.9	19.2	18.3
80<100%	21.7	21.5	19.3
Overall	17.1	18.2	18.1

Table 4. Accuracy(RMSE) on Air Conditioner Dataset without Synthetic Data

Air Cond.	**RF**	**SVM**	**LR**
0 <15%	7.4	6.4	6.2
15<50%	17.7	18.6	18.4
50<80%	17.8	19.5	20.4
80<100%	20.1	20.7	20.2
Overall	18.8	19.1	19.9

Table 5 presents a comparison between cases with and without the synthetic dataset on both original datasets (paint and air conditioner). Clearly, the performance improvement is significant, especially on the lowest wallet share group.

5 Conclusions

This paper presents a novel approach for the analysis and estimation of customer wallet share for manufacturers. Two major manufacturers which collect

Table 5. Comparison: with/without Synthetic Data

	Paint		**Air Conditioner**	
	With	**Without**	**With**	**Without**
0 <15%	**5.1**	**9.5**	**3.2**	**7.4**
Overall	**16.4**	**17.1**	**18.7**	**18.8**

transaction data related only to their own products are investigated. The approach accurately predicts wallet shares scores based on transaction data only and does not rely on any additional survey data, as it is normally used in other approaches.

The proposed approach is evaluated using two real-world datasets. The first dataset consisted of the transactions data from one of the largest manufacturers of paint products in Australia. The other dataset came from a major Australian manufacturer and distributor of air conditioning products. Furthermore, the analysis of the most important features for wallet share estimation is provided. These findings can be helpful for similar problems as well. Additionally, it is shown that the proposed model can work with a limited training input and a data augmentation approach is presented to address the data skew issue. To the best of our knowledge, no existing research has investigated a similar problem with the same data limitations.

The contributions of this paper can be summarised as follows:

1. To the best of our knowledge, there is no similar research to estimate customer wallet share for manufacturers based on transaction data only, as described in the scenario. We developed a Random Forest predictive model and extended the existing features of transactions by creating new features.
2. We analysed and selected the most important features and provided a simplified and scalable model, which then was used to analyse a large customer base. This assisted in accurately estimating the share of wallet for customers for a manufacturer's product.
3. We showed that it was possible to build an estimation model with a small training set. Furthermore, we demonstrated the application of synthetic data to address the data skew problem.

This paper aims at encouraging companies to apply modern data science techniques in approaching their business problems and to start collecting more surveys to begin the process. We are currently experimenting with many other businesses to check the model's generalisation capability. In future research, we plan to augment transaction data with survey results for cross-referencing. The additional benefit of this step would be a possible identification of inaccurate survey entries. Moreover, using this data, a business could aim to outperform the original survey in wallet share estimation accuracy.

References

1. Baumann, C., Burton, S., Elliott, G.: Determinants of customer loyalty and share of wallet in retail banking. Journal of Financial Services Marketing **9**(3), 231–248 (2005)
2. Chen, Y., Steckel, J.H.: Modeling credit card share of wallet: Solving the incomplete information problem. Journal of Marketing Research **49**(5), 655–669 (2012)
3. Çifci, S., Ekinci, Y., Whyatt, G., Japutra, A., Molinillo, S., Siala, H.: A cross validation of consumer-based brand equity models: Driving customer equity in retail brands. Journal of Business Research **69**(9), 3740–3747 (2016)
4. Glady, N., Croux, C.: Predicting customer wallet without survey data. Journal of Service Research **11**(3), 219–231 (2009)
5. Keiningham, T.L., Cooil, B., Aksoy, L., Andreassen, T.W., Weiner, J.: The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet. Managing service quality: An international Journal **17**(4), 361–384 (2007)
6. Keiningham, T.L., Perkins-Munn, T., Evans, H.: The impact of customer satisfaction on share-of-wallet in a business-to-business environment. Journal of Service Research **6**(1), 37–50 (2003)
7. Keiningham, T.L., Cooil, B., Malthouse, E.C., Lariviere, B., Buoye, A., Aksoy, L., De Keyser, A.: Perceptions are relative: an examination of the relationship between relative satisfaction metrics and share of wallet. Journal of Service Management **26**(1), 2–43 (2015)
8. Kursa, M.B., Rudnicki, W.R., et al.: Feature selection with the boruta package. J Stat Softw **36**(11), 1–13 (2010)
9. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R news **2**(3), 18–22 (2002)
10. Mägi, A.W.: Share of wallet in retailing: the effects of customer satisfaction, loyalty cards and shopper characteristics. Journal of retailing **79**(2), 97–106 (2003)
11. Merugu, S., Rosset, S., Perlich, C.: A new multi-view regression approach with an application to customer wallet estimation. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 656–661. ACM (2006)
12. Rosset, S., Perlich, C., Zadrozny, B., Merugu, S., Weiss, S., Lawrence, R.: Wallet estimation models. In: International Workshop on Customer Relationship Management: Data Mining Meets Marketing (2005)
13. Subramaniam, L.V., Faruquie, T.A., Ikbal, S., Godbole, S., Mohania, M.K.: Business intelligence from voice of customer. In: 2009 IEEE 25th International Conference on Data Engineering. pp. 1391–1402. IEEE (2009)
14. Weiss, S.M., Indurkhya, N.: Estimating sales opportunity using similarity-based methods. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 582–596. Springer (2008)