Towards Accurate Predictions of Customer Purchasing Patterns

Rafael Valero-Fernandez
School of Computing and Mathematics
Keele University, UK
r.valero-fernandez@keele.ac.uk

David J. Collins
School of Computing and Mathematics
Keele University, UK
d.j.collins@keele.ac.uk

K.P. Lam
School of Computing and Mathematics
Keele University, UK
k.p.lam@keele.ac.uk

Colin Rigby
Keele Management School
Keele University, UK
c.a.rigby@keele.ac.uk

James Bailey
Keele Management School
Keele University, UK
j.bailey3@keele.ac.uk

***Abstract*—A range of algorithms was used to classify online retail customers of a UK company using historical transaction data. The predictive capabilities of the classifiers were assessed using linear regression, Lasso and regression trees. Unlike most related studies, classifications were based upon specific and marketing focused customer behaviours. Prediction accuracy on untrained customers was generally better than 80%. The models implemented (and compared) for classification were: Logistic Regression, Quadratic Discriminant Analysis, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest and Multi-layer Perceptron (Neural Network). Postcode data was then used to classify solely on demographics derived from the UK Land Registry and similar public data sources. Prediction accuracy remained better than 60%.**

Keywords—classifiers, regression, segmentation, customer targeting, ecommerce, database marketing, life value cycle, churn ratio.

I. INTRODUCTION (*HEADING 1*)

Retailers are becoming increasingly aware of the value of their data and the importance of customer relationship management (CRM). Connecting customer data with wider sources allows customer identification, segmentation, product association, prediction, visualization of trends, recommenders, loyalty programs and so on. See [1,7] for more examples.

We worked with a UK '*clicks and mortar*' company (SME), our commercial partner (CP), who wanted to better understand their customer base in order to produce more effective marketing campaigns. They provided us with eight years' worth of historical transaction data. The main goal was to establish key statistics such as Churn Rate and Customer Life Cycle Value (CLV). Secondary goals were to provide predictors for the behaviour of segmented customer groups in order to better target promotional campaigns and ultimately increase profitability. This can be viewed as a straightforward classification problem that seeks to identify contribution to profits made by different groups of customers. However, a closer examination of the data reveals a highly skewed distribution within which a relatively small minority of customers are responsible for the majority of sales value.

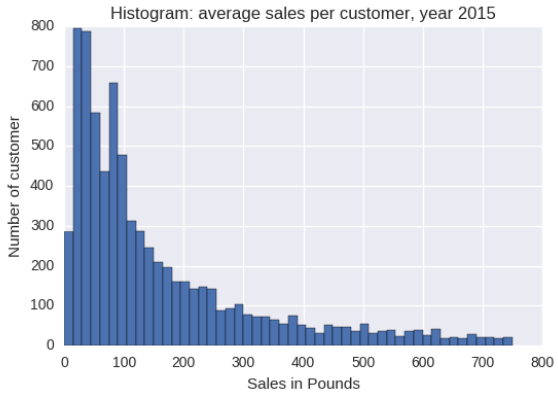

Fig. 1. Distribution of Customers. Truncated at 90%.

In Figure 1 we depict the distribution of customers by total sales per year. It is evident that the data closely approximates a traditional Pareto distribution. Similar patterns arise in *Recency*, *Frequency* and *Monetary* value (RFM) studies such as [8,11] and present considerable challenges for classification given the long tail which represent high value customers. It is essential that classification techniques applied capture all of the high value customers that the tail-end of such a distribution in order to be of any practical value. In this paper we thus examine the performance of classifiers with this important goal in mind.

II. RELATED WORK

Firstly we will explain some terms as they relate to this study, as follows.

The **Churn Rate**, or rate of attrition, is the percentage of customers who make no further purchases within a given time period. For purposes of this study, we define it as the proportion of customers who, having made a purchase, make no further purchases in the following 12 months.

978-1-5386-0958-3/17 $31.00 © 2017 IEEE
DOI 10.1109/CIT.2017.58

Customer Lifecycle Value (CLV) is an estimation of the net profit attributed to the entire future relationship with a customer. For purposes of this study we regard it as the present value of future cash flows attributed to the customer during his or her entire relationship with the company. This is primarily because we do not have cost data for the company and therefore cannot usefully speculate on net profitability.

Customer Relationship Management is an "*enterprise approach to understanding and influencing customer behaviour through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability.*" [14].

Machine learning techniques have been applied to many aspects of CRM, but relatively few to Customer Analytics [15]. Of these, most are influenced heavily by the traditional RFM approach to segmentation; for example, in [16], which requires interpretation and does not naturally allow variables to be introduced into the classification criteria.

In this study we perform classification on subsets of customers based upon typical marketing concerns – for example, which customers are most likely to buy at a discount but not at full price? We compare a range of classification techniques and prediction algorithms. We also assess the degree to which demographic data alone (without RFM data) is able to provide accurate classification and prediction with a view to identifying possible new customers in marketing campaigns in a cost-effective manner.

III. Methodology

A. Data

The data made available to us was derived from both a proprietary ecommerce system and from physical retail outlets. The data was anonymised but otherwise reasonably standard transactional data which included customer postcode [14]. We combined this data with public data sources as explained in section C below.

By looking at a subset of the dataset over time we had the opportunity to derive classifiers for selected behaviours. The behaviours that we were interested in were those that could be most allied to marketing efforts. In conjunction with the company, we selected the following customer segments for further investigation:

S(a) Customers who purchased in year *y* and purchased again in year *y+1*.

S(b) Customers who purchased at a specific time of the year (for example Easter) and purchased again at the corresponding time in the next year.

S(c) Customers who bought discounted products in one year who purchased discounted products in the subsequent year.

S(d) Customers who purchased a multiple item set in one year who purchased similarly in the next year.

S(e) Customers who bought a specific product in one year who bought the same product in the subsequent year.

The standard Churn rate is revealed by segment (a) above, but we used similar methodology to obtain churn rates for other segments (see [1, 5, 12-13]). We have observed that overall the customers of the CP have a high Churn or attrition rate, although the figures that we obtain using conventional (year on year) analysis may be slightly distorted - although churn rates are often high in SMEs [4]. There would appear to be two main reasons for this distortion - strong seasonality in sales, leading to lengthy inter-purchase delays and the false identification of repeating customers as new customers.

Instead of using the standard definition of Customer Life Cycle Value as described in section II above, we used the expected lifetime sales by segment which was preferred by the CP. Similar techniques to those used by [1] and [9] were employed.

B. Models

The analysis uses a range of models to classify consumers into segments and estimate their sales. The models implemented for classification were: Logistic Regression, Quadratic Discriminant Analysis, Linear SVM, RBF SVM, AdaBoot, Decision Trees, Random Forest, Quadratic Discriminant and Multi-layer Perceptron (Neural Network). For prediction of sales we used Linear Regressions (Reg.), LASSO (Lasso), Regression Trees (Tree Reg.) and Ridge Regression (Ridge Reg.). These classifiers were implemented using the python Scikit-Learn library [8].

The classifiers were constructed using customers who transacted during 2013-2014. For each segment, 70% of the unique customers were randomly selected for the training set. Classification parameters were then used to test prediction accuracy on the remaining 30% (Test) set (from 2013-2014 cohort) and the complete cohort sampled in 2014-2015.

C. Dependent variables and Data Selection

To enhance the CP's data, we combined it with other databases which provided additional consumer attributes at the Output Area level[1]. We used the official government Census and Land Registry databases[2]. From the Census we obtained area population, housing characteristics, household composition, age, gender, literacy, working hours, occupancy and so on. From the Land Registry Data, we obtained estimations of house prices in the consumers' neighbourhoods. In this manner we were able to significantly increase our understanding of the CP customers in order to study and understand their behaviour.

Feature selection was implemented manually, by trying to fit a linear regression to predict the sales to each customer in a period of time[3]. Non-significant variables were removed one

[1] Output Areas are the smallest geographical areas with available data in Census. They were created for statistical purposes. In 2011 the average population was 309.

[2] Land Registry data: https://www.gov.uk/government/publications/hm-land-registry-data, and Census data: https://www.ons.gov.uk/census/2011census.

[3] It is possible to generalise this by using *Recursive Feature Elimination* with cross validation.

by one until all the remaining variables were significant at 5%. The final variables were:

V(a) The logarithm of the distance of the consumer to the closest CP physical shop,
V(b) The square of V(a),
V(c) Boolean: 1 if the customer bought two years earlier,
V(d) Logarithm of the housing prices of the postcode area,
V(e) Age structure,
V(f) Rooms, bedrooms and central heating,
V(g) Occupation by sex,
V(h) Second address,
V(i) Adult life-stage,
V(j) Multiple ethnic groups,
V(k) Tenure households,
V(l) Communal establishment management and type, and
V(m) Economic activity of household reference person.

The final variables are similar to those obtained in other work on calculating Churn rate and credit card defaults as reported in [3,11,13], for example.

IV. Experiments

Using the classifiers for each segment, we sought to predict the purchasing patterns of the test groups over different time periods. Taking each segment in turn, we used:

S(a) the classifier obtained from the 2013-2014 training set on the 2014-2015 test set to identify the subset that having bought in 2014 would buy again in 2015.

S(b) the classifier obtained from the 2013-2014 training set on the 2014-2015 test set to identify the subset that having bought in a specific month in 2014 would buy again in the same month in 2015.

S(c) the classifier obtained from the 2013-2014 training set on the 2014-2015 test set to identify the subset that having bought a discounted product in 2014 would buy further discounted products in 2015.

S(d) the classifier obtained from the 2013-2014 training set on the 2014-2015 test set to identify the subset that having bought a boxed item set in 2014 would buy further boxed item sets in 2015.

S(e) the classifier obtained from the 2013-2014 training set on the 2014-2015 test set to identify the subset that having bought a specific product in 2014 would buy that same product again in 2015.

In each case, the estimations produced by the classifiers were compared with the actual purchases made and a success rate derived for each of the classification algorithms.

In addition to these estimations of purchases we also sought to estimate the total value of purchases. To accomplish this, we performed a simple linear regression of the total sale value of the training set across 2013-2014 and applied the result to estimate 2015 sales from the 2014-15 cohort. Thus we were able to use both classifiers and regressors to predict/classify in the period 2015 and compare the results with the real classification and sales of 2015.

Lasso, Regression Trees and Ridge Regression methods were also used to estimate sales value and results will be compared in the next section.

Finally, the data from the company has a total of 16762 unique customers that bought in the period 2013. These were divided into training (11732 (70%)) and test (5030 (30%)) groups. When we merged customer data with Census and Land Registry data, we reduced to 5859 (70%)+2510 (30%) customers (for which we had postcode and related information). In the same way there were a total of 18221 (100%) customers who bought in 2014, reduced to 9328 (100%) when merged with Census and Land Registry data. Table A.1 in Appendix A provides a summary.

V. Results and Discussions

A. Classifiers

Based on the segment **S(a)**, we compared the performance of different classification methods but to preserve space Table I shows only the classifiers with the best Success Ratio after the selection of the tuning parameters. The various classifiers were very similar in performance, and for subsequent analysis we adopted Logistic Regression given its simplicity and performance.

TABLE I CLASSIFIERS ACCURACY

Classifier	Success Ratio	Classifier	Success Ratio
KNN	78%	Random Forest	74%
Linear SVM	84%	Neural Network	84%
RBF SVM	75%	AdaBoost	84%
Logistic Regression	84%	Naive Bayes	84%
Decision Tree	74%	Quad. Discriminant Analysis	43%

TABLE II CUSTOMER REPETITION NEXT YEAR

Target Group	Success Ratio 2013-2014	Success Rate 2014-2015
S(a)	84%	85%
S(b)	87%	87%
S(c)	79%	79%
S(d)	84%	85%
S(e)	85%	86%

Table I shows the accuracy with which we were able to predict the purchasing pattern of consumers using the classification predictors obtained from the modelling. For four of the five examined segments it exceeds 80%. In Table II, we

show the performance of Logistic regression on the unclassified (test) 2013-2014 cohort and on the 2014-2015 cohort and the results are encouragingly similar.

B. The importance of historical data

We were particularly interested in examining the accuracy of predictors which were based solely upon the census and land registry geolocation derived variables. This would allow new customers to be more easily targeted. One of the variables used within the classifiers (V(c)) indicates whether a customer made a purchase two years previously and is the only measure of historical interaction with the company. We decided to remove this variable from the classification models and examine the results. These are presented in Table III. It can be seen that the prediction accuracy drops to circa 60%, but this should be carefully considered in view of the implicit selection bias. Furthermore, it is useful to estimate the repetition of new customers (where V(c) it is not well defined).

Table III also depicts the result of removing the Census and Land Registry data (row 2). As can be seen, the predictive capabilities remain high. The fact that the predictive capabilities of the Census and Land Registry data (row 3 in Table III) is high (59-68%) raises the possibility of far more effective marketing campaigns through (a) the selection of new postcodes, and (b) the identification of addresses from Land Registry data of people who have recently moved into highly significant areas.

C. Sales Estimators

Table IV shows the results of applying different means of estimating total sales value of the classified segments. The predictive capability is very poor, but comparable to results reported by other researchers, for example [9]. We believe that the poor performance is explained in part by the seasonal nature of sales (for this particular class of products). Without the V(c) variable, an important indicator of 'customer quality' we lose most of the predictive power.

TABLE III CUSTOMER REPETITION NEXT YEAR DIFFERENT DATA

Target Group	Success Ratio 2013-2014	Success Ratio 2014-2015
S(a)	84%	84%
S(a) only company data	83%	84%
S(a) no V(c)	59%	68%

TABLE IV CUSTOMER REPETITION NEXT YEAR DIFFERENT DATA

Target Group	Lasso	Ridge Regression	Linear Regression	Tree Regression
S(a)	0.15	0.15	0.15	0.15
S(a) only Company data	0.14	0.15	0.15	0.17
S(a) no V(c)	0.009	0.02	0.02	0.01

D. Group classification details

We can aggregate the five identified segments (**S(a)**,…,**S(e)**) as described on section III into two groups – those who are highly likely to buy in the future (**Good**) and those who are not (**Bad**). The results are shown in Table V. A few comments are in order.

i. A customer in the **Good** group produces nearly 3.6 times the sales value of a customer in the **Bad** group (355.84% against 100%).
ii. The number of **Good** customers is over-estimated whilst the number of **Bad** customers is underestimated. Despite this the split of the total sales value between the two groups are predicted accurately.
iii. Based on (ii), the classifier manages to capture the **Good** customers responsible for the majority of sales.

Tables VI and VII consider the effects of removing variables from the classifiers. In Table VI, we show the results of applying only the company data (excluding Census and Land Registry data) whilst Table VII depicts the results of applying only Census and Land Registry data.

TABLE V CLASSIFICATION DETAILS: S(A)

	Real		Predicted	
Group	Bad	Good	Bad	Good
Customers (%)	45.88%	54.12%	22.43%	77.57%
Total Sales per Customer (%)	21.94%	78.06%	21.00%	79.00%
Relative Weight (%)	100%	355.84%	100%	376.27%

Key: **Bad** group referred to those customers who will not repeat. **Good** is the opposite. Relative Weight is the measure the quality of the customers in term of sales.

TABLE VI CLASSIFICATION DETAILS: S(A) ONLY CP DATA

	Real		Predicted	
Group	Bad	Good	Bad	Good
Customers (%)	46.43%	53.57%	60.96%	39.04%
Total Sales per Customer (%)	20.31%	79.69%	20.73%	79.27%
Relative Weights (%)	100%	392.44%	100%	382.38%

TABLE VII CLASSIFICATION DETAILS: S(A) NO V(C)

	Real		Predicted	
Group	Bad	Good	Bad	Good
Customers (%)	45.88%	54.12%	22.48%	77.52%
Total Sales per Customer (%)	21.94%	78.06%	45.95%	54.05%
Relative Weights (%)	100.00%	355.84%	100.00%	117.65%

Importantly, Table VII shows that Public data alone has potential value in identifying new customers. Whilst overestimating the number of good customers and underestimating the number of bad customers it produces a valuable (17.65%) estimate of the contribution of identified good customers to total sales based solely upon geolocation factors. In marketing terms this is highly significant.

VI. Conclusions

In this paper we have shown the feasibility of identifying the probabilities of customer purchase repetition, making it possible to target important customer groups. We have further shown how public data sources can be used to augment internal data and thereby achieve improved marketing and profitability. Thus far we have only scratched the surface. In future we intended to conduct more detailed analysis by adopting unsupervised and more importantly exploratory techniques to further our understanding of the factors that influence customer behaviour in a more generic context.

Appendix A

The data from the company has a total of 16762 unique customers that bought in the period 2013. These are divided into training (11732 (70%)) and test (5030 (30%)) groups. When we merge customer data with Census and Land Registry data, we reduce to 5859 (70%) + 2510 (30%) customers (for which we have postcode and related information). In the same way there are a total of 18221 (100%) customers who bought in 2014, reduced to 9328 (100%) when merged with Census and Land Registry data.

TABLE A.1 Classification details: S(A)

Target Group	Training 2013-2014	Test 2013-2014	Total 2013-2014	Training 2014-2015	Test 2014-2015	Total 2014-2015
S(a)C	11732	5030	16762	0	18221	18221
S(a)N	5859	2512	8371	0	9328	9328

Key: S(a)C –Company Data Only, S(a)N – With Public Data (Data without Post Code locator removed).

Acknowledgement

The authors would like to thank their anonymous commercial partners for providing the data sources and other resources used in the production of this paper. We are also grateful for the financial support by the Innovate UK through the *Knowledge Transfer Partnership* (KTP) Award scheme, project 010034.

References

[1] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.

[2] Hsieh, N. C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. Expert systems with applications, 27(4), 623-633.

[3] Kaminskas, M., Bridge, D., Foping, F., & Roche, D. (2016). Product-Seeded and Basket-Seeded Recommendations for Small-Scale Retailers. Journal on Data Semantics, 1-12.

[4] Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), 101-107.

[5] Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. SAS User Group International (SUGI27) Online Proceedings, 114-27.

[6] Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830. Link to the library: http://scikit-learn.org/stable/.

[8] Verhoef, P. C., & Donkers, B. (2001). Predicting customer potential value an application in the insurance industry. Decision support systems, 32(2), 189-199.

[9] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[10] You, Zhen, et al. "A decision-making framework for precision marketing." Expert Systems with Applications 42.7 (2015): 3357-3367.

[11] Yu, X., Guo, S., Guo, J., & Huang, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Applications, 38(3), 1425-1430.

[12] Zhao, J., & Dang, X. H. (2008, October). Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer churn as the example. In Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on (pp. 1-4). IEEE.

[13] Zhu, C., Qi, J., & Wang, C. (2009, October). An experimental study on four models of customer churn prediction. In Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on (pp. 3199-3204). IEEE.

[14] Swift, RS (2001), Accelerating customer relationships: Using CRM and relationship technologies, Prentice Hall PTR, N.J.

[15] Ngai, E.W.T, Li, Z. and Chau, D.C.K,(2009) Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications ,Volume 36, Issue 2, Part 2, March 2009, Pages 2592-2602

[16] Sarvari, P.E, Ustundag, A. and Takci, H. (2016), Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, Kybernetes, Vol. 45 Issue 7.