2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Customer Lifetime Value Prediction with K-means Clustering and XGBoost

Marius Myburg
Computer Science Department
University of Cape Town
Cape Town, South Africa
mybmar003@myuct.ac.za

Sonia Berman
Computer Science Department
University of Cape Town
Cape Town, South Africa
sonia@cs.uct.ac.za

***Abstract*—Customer lifetime value (CLV) is the revenue expected from a customer over a given time period. CLV customer segmentation is used in marketing, resource management and business strategy. Practically, it is customer segmentation rather than revenue, and a specific timeframe rather than entire lifetimes, that is of interest. A long-standing method of CLV segmentation involves using a variant of the RFM model – an approach based on *R*ecency, *F*requency and *M*onetary value of past purchases. RFM is popular due to its simplicity and understandability, but it is not without its pitfalls. In this work, XGBoost and K-means clustering were used to address problems with the RFM approach: determining relative weightings of the three variables, choice of CLV segmentation method, and ability to predict future CLV segments based on current data. The system was able to predict CLV, loyalty and marketability segments with 77-78% accuracy for the immediate future, and 74-75% accuracy for the longer term. Experimentation also showed that using RFM alone is sufficient, as augmenting the features with additional purchase data did not improve results.**

Keywords— XGBoost, K-means clustering, customer lifetime value, recency, frequency, monetary value.

I. Introduction

Machine learning is increasingly being used in Customer Relationship Management (CRM). A key component of CRM is Customer Lifetime Value, or CLV, which is a measure of a customer's expected future value to a company. To construct models that can predict a customer's future CLV, the input features and target variable must first be determined. A popular approach to customer segmentation is the RFM method [1] which is based on three features: Recency (of their last purchase), Frequency (how many purchase visits over a fixed time period) and Monetary value (total revenue over all their purchases during the fixed time period). This work therefore chose to use the three RFM features as the basis for CLV prediction. A variety of target variables have been explored, most of which aim to segment customers with similar CLV, rather than predict individual revenue amounts. Five different segmentation approaches were investigated as target variables, three of these segmenting customers on future value to the company, one on loyalty, and the remaining one on marketability.

Traditionally, determining current CLV based on current RFM amounts involves the following: R, F and M are converted to quantile values and their sum is used as CLV. The possible sums are then divided into different ranges representing CLV segments. For example, suppose R, F and M are converted to quintiles, and 4 CLV segments are targeted. Then the sums, that range from 3 to 15, are divided into 4 segments in a manner appropriate for that business, such as: Poor (sums of 3, 4 and 5), Low (sums of 6 or 7), Fair (sums of 8, 9, 10 or 11) and Good (sums from 12 up to 15). Variations of this approach exist, the most popular being AHP (analytical hierarchy process) methods, in which R, F and M are scaled based on the views of industry experts [2,3,4].

As regards predicting future CLV, stochastic models and linear regression are typically applied [5,6,7]. This study explored the use of XGBoost as an alternative method for predicting CLV from Recency, Frequency and Monetary value inputs. It further examined the robustness of this approach by comparing prediction accuracy for the immediate future against accuracy in predicting a more distant time period; and by comparing models given only RFM values with models given additional purchase information to supplement the RFM values.

II. Background

Marketing and costing can be more effective when finely tuned to requirements of specific customer segments. Attributes that have been used to segment customers include demographic, geographic and buying behaviour information. Segmentation is expensive and challenging; data is often unavailable and methods capable of addressing the needs of individual customers are difficult to develop. Customer Lifetime Value (CLV) is defined as "the present value of the future profit stream expected over a given time horizon of transacting with the customer" [8]. In addition to its importance as a business metric, CLV is also used as a customer segmentation tool [9]. One of the most popular methods of determining CLV is the Recency Frequency and Monetary value (RFM) model [1]. This proposes using a quantile to represent each variable, and taking the sum of a customer's three quantile values as representative of CLV. Several variations of this method exist; the use of clustering algorithms - in place of quantile sums - to obtain the final segments, is a common alternative also based on R,F,M data. Several studies have clustered customers according to their RFM values, with K-means clustering the most widely used technique [4]. In other studies, some researchers have excluded one of the three variables, while others have augmented R,F,M with additional attributes such as average item value, profit, etc. A common variation is to use AHP or Fuzzy AHP. Rather than weighting R, F and M values equally, (F)AHP uses a weighted-RFM method instead. These weights are intended to capture the relative importance of the three variables in a specific industry, based on the opinion of experts in that industry [2,3,4].

Limitations of the RFM approach to determining CLV addressed by this work are:

- predicting customer CLV in the future can provide greater business benefit than mere segmentation of current behaviour

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 ©2022 IEEE

- effective prediction in the longer term is also investigated; the lack of such research having been noted e.g. in [10]
- difference in importance of the three variables is difficult to ascertain [11]
- the best quantile choice for representing RFM values is arbitrary and can compromise results due to insufficient granularity [12].

XGBoost is one of the most popular classification algorithms due to its scalability, speed and ease of use, among other benefits [13]. In a 2020 review of 119 papers on machine learning in CRM [14], only 3 studies used XGBoost, none involving CLV prediction. This study therefore explored the use of K-means clustering for customer segmentation, and XGBoost for predicting customer segments in both the immediate future, and a later period.

III. Segmentation

Five segmentation methods were investigated, both as a multi-class classification and as a binary classification problem. The former aimed at predicting which segment each customer would fall into in the future, and the latter at determining whether or not a customer would drop to an inferior segment in the future. The five target segmentations respectively aimed at predicting: revenue cluster, loyalty cluster, marketability cluster, revenue quantile and weighted-RFM quantile. Revenue clusters were generated through K-means clustering of the single variable M, monetary value of purchases. K-means was also used to obtain Loyalty clusters and Marketability clusters. Loyalty was based on Recency and Frequency, as loyalty is a function of visits, not expenditure. Loyal customers can be enticed to spend more through promotions such as "buy 2 save X". Marketability clusters were based on Recency and Monetary values, indicating how attractive more frequent visits by the customer would be. Limited-time special offers, for example, is indicated for those with high marketability. The final two approaches segmented customers based on the quantile into which their CLV fell. The revenue quantile used monetary value as CLV. The weighted-RFM quantiles used weighted sums of R, F and M; but in place of expert judgement, these weightings were set equal to the information gain of each variable as determined by the XGBoost model trained to predict monetary cluster.

IV. Method

Fast Moving Consumer Goods sales data was used in the project. This was selected as an extreme example, based on the view expressed in [10] that effective use of CLV models in that specific industry was particularly doubtful. Three timeframes of six months' duration were extracted from the data; the first, T1, represented the training data, the second (T2) the six months following T1, and the third (T3) the six-months following T2. Predicting CLV segments in T2 would be more useful for the business than a mere segmentation of customers according to their current standing. T3 prediction was also investigated, in order to assess the efficacy of the method for longer-term prediction. Predicting more than a year ahead, on the other hand, would be inappropriate for planning purposes, due to being too far in the future. The CRISP-DM methodology was followed. Data preparation included normalisation of RFM values and removal of outliers based on Z-score. An upper limit of z=4 was used, to avoid overly-aggressive removal of outliers with high values, since these are potentially valuable to the business. Silhouette score determined the number of clusters for CLV, Loyalty and Marketability clustering, as this gave clearer indication of the optimum number of clusters/segments than did the Elbow method.

Once the data exploration and preparation had been completed, K-means clustering of customers during each of the three timeframes T1, T2 and T3 was performed to obtain Revenue clusters, Loyalty clusters and Marketability clusters. As mentioned, Loyalty and Marketability were both 2-dimensional clusterings, on Recency and Frequency, and on Frequency and Monetary value, respectively. XGBoost was then used to create models able to predict future Revenue, Loyalty and Marketability segments accordingly. At the same time, the information gain of these three variable, as returned by XGBoost in these experiments, were recorded and averaged to obtain the three weightings to use in the weighted-RFM experiments. For the dataset in this experiment, these weightings were 1 for Frequency, 2 for Recency and 8.5 for Monetary value. The division of T1, T2 and T3 customers into weighted-RFM terciles followed accordingly; the division into revenue terciles accounted for the fifth and final target segmentation.

One set of XGBoost models was constructed for predicting each type of customer segment, and another set of XGBoost models created for predicting which customers would drop to inferior segments of each type during the later timeframes. All models were trained on T1 data, and tested on the immediately following period T2 and on the subsequent period T3. In addition to training models given only the normalised R, F and M values, additional models were constructed from data that augmented R, F and M with a variety of other purchase data. This additional data comprised features such as average basket (transasction) value, number of transactions, number of national stores visited and coupon usage, each calculated for every customer over the entire period as well as on a per-weekday, per-product-department and per-time-of-day basis. Table I gives an overview of the models constructed, and the abbreviated names associated with each. After applying PCA to the augmented data, the number of features was reduced to 31. Class imbalance was corrected with SMOTE prior to training XGBoost, and hyperparameter optimisation was applied to optimise model configuration. RandomizedSearchCV was used in preference to GridsearchCV as it is able to evaluate a wider range of possible parameters in reasonable time; the number of tuning iterations was set to 1500 throughout. For the binary classification models, predicting if customers would drop to a lower segment or not, the use of SMOTE gave the same results as when the XGBoost scale-pos-weight parameter was set to the recommended (sum of negative instances)/(sum of positive instances), and hence SMOTE was only used for constructing the multiclass classification models.

V. Results

Three methods of CLV segmentation were investigated: K-means clustering on revenue (Monetary value), revenue tercile, and a weighted-RFM method in which information gain replaced expert estimates of relative R, F, M weights. Table II gives the results of the experiments predicting the segments into which each customer would fall in the immediately following (T2) and subsequent (T3) six-month periods. All methods grouped customers into 3 segments; Low, Medium or High. Of these three types of target segment,

TABLE I. EXPERIMENTS PREDICTED 7 SEGMENTATIONS, USING RFM ALONE AND WITH AUGMENTED DATA, FOR 2 TIME PERIODS

Target variable	Experiment abbreviation: Input features and target timeframe			
	RFM data only (R)		RFM with Added features (A)	
	Timeframe 2	Timeframe 3	Timeframe 2	Timeframe 3
C - revenue Cluster – multiclass classification	R-2C	R-3C	A-2C	A-3C
T - revenue Tercile – multiclass classification	R-2T	R-3T	A-2T	A-3T
W – Weighted RFM segment – multiclass classification	R-2S	R-3S	A-2S	A-3S
L – Loyalty cluster – multiclass classification	R-2L	R-3L	A-2L	A-3L
M – Marketability cluster – multiclass classification	R-2B	R-3B	A-2B	A-3B
TD - revenue Tercile Drop – binary classification	R-2TD	R-3TD	A-2TD	A-3TD
WD – Weighted RFM segment Drop – binary classification	R-2SD	R-3SD	A-2SD	A-3SD
LD - Loyalty Drop – binary classification	R-2LD	R-3LD	A-2LD	A-3LD
MD - Marketability Drop – binary classification	R-2BD	R-3BD	A-2BD	A-3BD

TABLE II. XGBOOST SEGMENT PREDICTION RESULTS FOR RFM DATA ALONE (R) AND FOR RFM WITH PURCHASE DATA (A)

	Cohen's Kappa	Training Accuracy	Holdout Accuracy	Weighted Precision	Weighted Recall	Weighted f1-score	Micro AUC
R-2C (R-3C)	0.59 (0.53)	79.79 (75.26)	78.21 (75.57)	0.78 (0.75)	0.78 (0.76)	0.78 (0.75)	0.92 (0.89)
A-2C (A-3C)	0.55 (0.46)	78.19 (63.64)	76.89 (63.77)	0.77 (0.63)	0.77 (0.64)	0.76 (0.63)	0.91 (0.81)
R-2T (R-3T)	0.54 (0.48)	71.10 (65.34)	69.62 (65.38)	0.69 (0.65)	0.70 (0.65)	0.69 (0.65)	0.86 (0.82)
A-2T (A-3T)	0.52 (0.46)	68.93 (63.64)	68.02 (63.77)	0.69 (0.63)	0.68 (0.64)	0.68 (0.63)	0.85 (0.81)
R-2W (R-3W)	0.48 (0.43)	68.56 (60.62)	65.57 (62.45)	0.65 (0.61)	0.66 (0.62)	0.65 (0.61)	0.84 (0.79)
A-2W (A-3W)	0.44 (0.43)	68.08 (58.92)	63.53 (61.98)	0.64 (0.61)	0.64 (0.62)	0.64 (0.61)	0.81 (0.79)
R-2L (R-3L)	0.42 (0.36)	78.38 (73.56)	76.70 (74.25)	0.68 (0.66)	0.77 (0.74)	0.72 (0.69)	0.91 (0.88)
A-2L (A-3L)	0.40 (0.34)	83.57 (76.02)	75.85 (72.64)	0.70 (0.66)	0.76 (0.73)	0.71 (0.68)	0.89 (0.86)
R-2M (R-3M)	0.48 (0.35)	78.75 (76.86)	78.58 (74.06)	0.70 (0.71)	0.79 (0.71)	0.74 (0.69)	0.91 (0.88)
A-2M (A-3M)	0.48 (0.35)	84.99 (78.56)	77.92 (74.34)	0.74 (0.70)	0.78 (0.74)	0.74 (0.69)	0.90 (0.88)

TABLE III. XGBOOST PREDICTION OF CUSTOMERS DROPPING TO INFERIOR SEGMENT

	Training Accuracy	Holdout Accuracy	Drop class Precision	Drop class Recall	Drop class f1-score
R-2TD (R-3TD)	59.66 (66.85)	58.17 (58.31)	0.33 (0.34)	0.81 (0.59)	0.47 (0.43)
A-2TD (A-3TD)	72.78 (67.00)	64.37 (53.94)	0.34 (0.32)	0.60 (0.66)	0.43 (0.43)
R-2WD (R-3WD)	64.56 (61.04)	61.88 (54.15)	0.30 (0.34)	0.72 (0.76)	0.42 (0.47)
A-2WD (A-3WD)	73.57 (55.70)	66.67 (48.48)	0.33 (0.31)	0.71 (0.76)	0.45 (0.44)
R-2LD (R-3LD)	69.91 (63.28)	67.27 (58.72)	0.25 (0.23)	0.76 (0.81)	0.38 (0.36)
A-2LD (A-3LD)	85.76 (83.51)	76.26 (76.90)	0.28 (0.31)	0.52 (0.49)	0.36 (0.36)
R-2MD (R-3MD)	75.07 (74.48)	76.92 (74.26)	0.35 (0.32)	0.93 (0.93)	0.51 (0.48)
A-2MD (A-3MD)	82.91 (81.20)	73.08 (72.65)	0.26 (0.31)	0.57 (0.66)	0.36 (0.42)

XGBoost was best able to predict future revenue cluster, as shown in Table II. Revenue tercile was better predicted than weighted-RFM tercile. Metrics dropped very slightly when the more-distant time period was predicted, as compared with the period directly following the training data. When predicting which customers would drop to a lower revenue tercile in the next timeframe, recall was good (0.81) but precision poor (0.33). Recall (0.93) and accuracy (76.92%) were particularly good in predicting whether customers would drop to a lower Marketability cluster – i.e. would spend less and/or become inactive. In each set of experiments, models given the augmented data did not achieve better results than their counterparts trained only on RFM values, confirming the robustness of the RFM . In predicting which customers would drop to a lower segment, XGBoost achieved good recall scores but precision was poor, as shown in Table III.

VI. CONCLUSION

A popular method of determining Customer Lifetime Value (CLV) is based on Recency, Frequency and Monetary (RFM) values of past purchases. The relative weightings of these three variables is generally based on the opinion of industry experts [2,3,4] rather than machine learning. K-means clustering and XGBoost were explored as a means of CLV segmentation, and of predicting future CLV based on RFM, respectively. Monetary cluster prediction gave better results than both revenue quantile and weighted-RFM quantile prediction. Quantile methods are the norm for RFM-based CLV prediction; this result indicates that K-means clustering should be a focus of future work in RFM-prediction of CLV. K-means is an easily understood algorithm in general, and particularly so when data can easily be visualised because there are no more than 3 variables.

XGBoost was able to predict customers' monetary cluster with 78% accuracy, an f1-score of 0.78, and a micro AUC of 0.92. K-means clustering of Loyalty and of Marketability was also undertaken. XGBoost prediction of Loyalty and Marketability cluster achieved accuracy of 77% and 78% respectively. Recall is more important in marketing, as reaching the right customers is paramount, even at a cost of targeting others who would not have dropped. While RFM for CLV prediction is generally applied only to the immediate future, the XGBoost models showed a negligible decline in all metrics when predicting was the more distant period, and achieved accuracy of 74-75%. In addition, augmenting RFM input with additional purchase data features did not improve results. We conclude that K-means clusters are good representations of Customer Lifetime Value, Loyalty and Marketability segments, and that XGBoost can effectively predict CLV, Loyalty and Marketability from RFM.

The current system is constructed as a pipeline through which data flows along from data ingestion and preparation, oversampling, separation of training, testing and holdout data sets, model construction and evaluation to production of model success metrics and a variety of data plots showing confusion matrices and 2-D clusterings. This pipeline can be enhanced with a user-friendly interface to serve as an interactive tool for semi-automated prediction of future CLV, customer loyalty and marketability. Future work is planned to compare the effectiveness of other machine learning techniques with that of XGBoost, and of other clustering algorithms with that of K-means clustering. The prediction of the five types of target segment in this way should also be extended to other industries and data sets, over and above that of the Fast Moving Consumer Good industry in-store purchases used here. Finally, whereas augmentation of RFM values with purchase product and temporal data was shown to give no better results than using RFM alone, the inclusion of demographic data was not explored due to lack of availability, and should be considered.

REFERENCES

[1] A. Hughes, Strategic database marketing. McGraw-Hill, 2005.

[2] P. Amin, T. MohammadJafar and A. Hossein, "Combining data mining and group decision making in retailer segmentation based on LRFMP variables", Int. Jnl. Industrial Engineering and Production Research, vol. 25, 2014, pp. 197-206.

[3] M. Ray and B. Mangaraj, "AHP based data mining for customer segmentation based on customer lifetime value", Integrated Intelligent Research, vol. 5, 2016, pp.28-34.

[4] P. Pramono, I. Surjandari, and E.Laoh, “Estimating customer segmentation based on customer lifetime value using two-stage clustering method”, 16th Int. Conf. on Service Systems and Service Management (ICSSSM): IEEE, 2019, pp. 1-5.

[5] J. R. Bernat, Modelling customer lifteime value in a continuous, non-contractual time setting, Masters Thesis, Econometrics and Mnagement Science, Erasmus University Rotterdam, 2018.

[6] P. Jasek, L. Vrana, L. Sperkova, Z. Smutny nd M. Kobulsky, “Predictive performance of customer lifeime value models in e-commerce and the use of non-financial data”, Prague Economic Papers, vol. 28, 2019, pp. 648-669.

[7] P. E. Pfeifer, and R. L. Carraway, “Modeling customer relationshps as Markov chains”, Journal of Interactive Marketing, vol. 14, 2000, pp. 43-55.

[8] P. Kotler, “Marketing during periods of shortage”, Journal of Marketing, vol. 38, 1974, pp. 20-29.

[9] S. Gupta et al, “Modeling customer lifetime value”, Journal of Service Research, vol. 9, 2006, pp. 139-155.

[10] D. Bell, J. Deighton, W. Reinartz, R. Rust and G. Swartz, “Seven barriers to customer equity management”, Journal of Service Research, vol. 5, 2002, pp. 77-85.

[11] A. Dursun and M. Caber, “Using data mining techniques for profiling profitable hotel customers: an application of RFM analysis”, Tourism Management Perspectives, vol. 18, 2016, pp. 153-160.

[12] R. Blattberg, B. Kim and S. Neslin, “RFM analysis”, in Database Marketing: Springer, 2008, pp. 323-337.

[13] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system”, 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining: ACM, 2016, pp. 785-794.

[14] Chagas et al, “A literature review of the current applications of machine learning and their practical implications”, Web Intelligence, vol. 18: IOS Press, 2020, pp. 69-83.