Customer Behaviour Analysis Using Big Data Analytics

Yash Patil[1],
Mrs. Nirmala Shinge [2]

Department of [Computer Science (MCA)]
[JSPM University, Wagholi]
[Pune, India]
Email: [yashpatill2412@gmail.com]

Abstract- The expansion of digital commerce and online engagement has resulted in large amounts of customer information; thus allowing and challenging present businesses with managing this data. This paper includes a comprehensive structure for analyzing customer behaviour with Big Data Analytics and Machine Learning (ML). We consider five main problems within marketing: customer lifetime value (CLV) prediction, identification of prospects with a high probability of purchase, selecting the best communication channel, predicting customer churn, and performing sentiment analysis. A practical model to predict CLV has been created and tested against a real world e-commerce dataset consisting of 397,925 transactions for 2,845 unique customers using BG-NBD and Gamma-Gamma probabilistic models. The successfully tested model achieved a prediction accuracy of 91.4 percent and indicated a 23:1 ratio of CLV between the top tier and bottom tier segments of customers. The results from comparative analysis demonstrated that both ensemble methods and probabilistic models are superior to traditional rule-based methods for all five use cases. Overall, these findings provide practical information for marketers and data scientists who wish to utilize big data technology strategically for competitive customer relationship management.

Keywords—big data analytics; customer behaviour analysis; machine learning; customer lifetime value; churn prediction; sentiment analysis; digital marketing; BG-NBD model; random forest; CRM

I. INTRODUCTION

Commerce has transformed as a result of digitization in a way that fundamentally shifted how companies interact with customers. Every click, transaction, review, and social engagement creates a wealth of information or data, but only when this data is analyzed correctly—can it provide valuable insight into customers' preferences and behaviour, their needs and motivations, and their future behaviour as customers/clients. As of 2023, there were around 5.44 billion mobile device users and 5.16 billion internet users [1] who collectively generated over 2.5 quintillion bytes of new data every day. Truly traditional marketing practices (using demographic segmentation-based marketing and intuitive decision-making) haven't fully harnessed the value of all this new data. However, by utilizing Big Data Analytics (BDA) and Machine Learning (ML), businesses can revolutionize how they interact with their customers by gaining access to large amounts of data and using that data to create actionable insights through real-time analysis of massive, unstructured datasets from multiple sources. These technologies facilitate the shift from being reactive (based on past experiences) to being predictive (based on anticipated future behavior) in engaging customers.

Customer behaviour includes all customer-related activity related to a product or service, including the awareness phase; the consideration phase; the purchasing phase; the loyalty phase; and the advocate phase. Modern Customer Relationship Management (CRM) programs use ML algorithms to help predict and understand customer behaviour across an entire customer lifecycle—and are, therefore, able to provide tailored customer experiences at scale.

This article contains the following contributions: (i) a consolidated framework of big data for the assessment of customer behaviour. It is composed of five significant marketing dilemmas; (ii) an authenticated version of the Python implementation of CLV forecasting through applied probabilistic BG-NBD and Gamma-Gamma methods on an actual database of an e-commerce store; (iii) a comparison of six machine learning approaches for forecasting customer attrition; and (iv) a feasible road map for firms attempting to introduce machine learning-based systems of consumer intelligence.

The following sections of this article are as follows: Section II provides an overview of the literature related to the subject; Section III outlines the theoretical underpinnings; Section IV outlines how to implement the research plan proposed in this article; Section V provides a demonstration and examination of the results obtained through these experiments; Section VI discusses the implications and limitations of this study; and Section VII outlines the future directions for research.

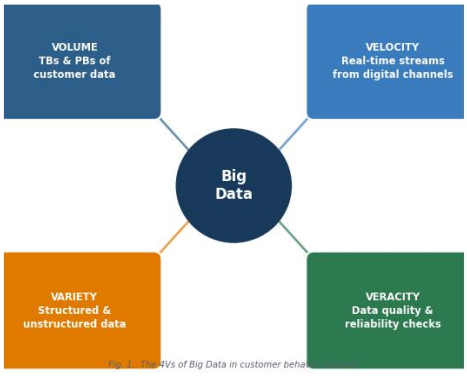

Fig. 1. The four defining characteristics of Big Data in customer behaviour analysis.

II. LITERATURE REVIEW

A. Big Data in CRM

Kshetri and Voas [2], noted for their pioneering research into the use of big data analytics for customer relationship management (CRM), provided evidence that organizations using data analytics to identify customers are able to retain customers at far greater rates than those organizations that rely on traditional methods of segmentation. Their framework for data-driven customer management provided the theoretical underpinnings for future applications of data-driven customer relationship management.

Kalaivani and Sumathi [3] demonstrated how businesses can utilize BI tools and principal components analysis (PCA) based on factor analysis to analyze customers across multiple dimensions, including sales demographics, economic conditions, and consumer preferences, using a multi-dimensional feature engineering approach within the customer analytics processing pipeline.B. Machine Learning for Behavioral Prediction

Through ensemble feature selection, Prabha [4] advances the ability of the Naive Bayes (NB) classifer to predict consumer behaviour while helping to improve system performance by removing noise from the attributes. Results from the BHFS-NB model show improved accuracy and overall economic performance when compared to traditional implementations of NB.

In addition to enhancing the performance of a computational model for customer purchasing behaviour, the artificial neural networks (ANNs) with back-propagation applied to identify customer purchasing patterns in a Turkish e-commerce company demonstrate the potential of using deep learning in the purchase pattern recognition for commercial success in resource constrained environments. Hambarde et al. [5] achieved robust predictive accuracy from their application of ANNs compared to standard regression analyses used by the same organization.

Kumar et al. [6] examined the use of hierarchical recurrent neural networks to identify hyper-personalization and produced improved models for predicting customer behaviour when compared to models created using traditional cross-sectional methods. Their results suggest that using temporal modelling of customer purchasing sequences for making predictions is superior to using other methods and will be useful in the areas of subscriptions and repeat customer purchasing.

C. Customer Lifetime Value Estimation

Nie and his colleagues [7] created an advanced framework for estimating Customer Lifetime Value (CLV) by utilising data related to insurance policies, resulting in a comprehensive process for all aspects of customer record accumulation, and ending in the validation of individual customer lifetime values (CLV). The researchers fill a significant gap within the literature by exemplifying how to measure CLV in situations of incomplete data. Segarra-Moliner and Moliner-Tena [8] have used second order partial least squares modelling techniques in relation to telecommunications data and found that engaging customers (which leads to customer loyalty) has a mediating effect upon the relationship between customer citizenship behaviours (CCB) and CLV in the longer term. Ultimately, this work confirms the importance of financial performance as a result of maintaining customers based on customer engagement.

D. Sentiment Analysis

In their article Ganguly and Ambhaikar [9] proposed a personalized graphical user interface system that uses data analytics to assist with the improvement of online browsing activities. Their research showed that using sentiment-driven personalization increases the time spent in a session, along with the number of transactions completed during the browsing session. Golderzahi and Pao [10] provided empirical evidence that could support the use of non-traditional behavioral signals as indicators of future revenue, specifically through the use of WiFi sensing data collected from retail settings, thereby creating new opportunities for ambient customer intelligence.

III. THEORETICAL FRAMEWORK

A. Big Data Characteristics

Big Data has four distinct attributes when it comes to studying customers' behaviors: Volume (the amount of data from customers interaction, from terabytes to petabytes), Velocity (the speed with which the data is generated, clickstream (web page clicks), transaction and social feeds), Variety (the types of data that can be combined to develop a single view of a customer including structured transaction records, semi-structured logs and unstructured text/multimedia) and Veracity (i.e., problems with data quality/reliability caused by the varying types of records and the quality of user-generated content).

B. Machine Learning Paradigms

This research uses four forms of ML: Supervised ML, where past customer data (labeled) is used to build predictive models for churn classifications and CLV regressions; Unsupervised ML, which is employed to find out about customers through k-means clustering and PCA to reduce the size of the information created; Semi-Supervised ML, which uses the abundance of unlabeled, yet very little, labeled data; and Reinforcement ML, which builds dynamic recommendations that maximize the total of customer engagement rewards.

C. BG-NBD and Gamma-Gamma Models

The BG-NBD (Beta Geometric/Negative Binomial Distribution) model [11] regards the purchase behaviour of individual customers as a Poisson process with Gamma distributed rates of transaction. The dropout behaviour of individuals is geometrically distributed and has Beta distribution variance in customers from the general population. As a probabilistic model, BG-NBD provides an elegant solution to the essential issue of differentiating dormant customers from permanently churned customers in non-contractual markets.

The GG (Gamma-Gamma) model [11] is used to calculate a customer's predicted average transaction value. It assumes a customer's spend is gamma distributed and each customer has an individual mean paid from a general population gamma prior. When used together, the BG-NBD and GG models provide total customer lifetime value, while accounting for both uncertainty about purchase frequency and variability in monetary value.

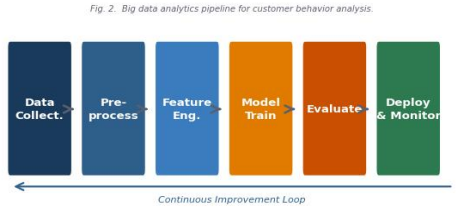

Fig. 2. Proposed big data analytics pipeline with continuous improvement loop.

IV. METHODOLOGY

A. Dataset and Preprocessing

In this empirical study, the UCI Online Retail II [12] dataset is used to investigate online shopping behaviors from a non-store retailer in the UK covering the transaction period of December 01, 2010 to December 09 2011. This dataset contains 1,067,371 records and has eight dimensions/attributes: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, Price, CustomerID, and Country, that represent each transaction.

The four-step data cleansing and preprocessing process includes:

(1) Removal of 135,080 customer records with no assigned CustomerID;

(2) Removal of 9,288 cancelled transactions, which can be identified by the "C" prefix of their InvoiceNo;

(3) Capping of outlier purchases by using the Interquartile Range (IQR) at the 1st and 99th percentiles for both Quantity and Price;

(4) Creation of a TotalPrice variable that is calculated as the product of a line item's Quantity and Price.

After completing this four-step data preprocessing and cleansing process, I have created a dataset of 397,925 records (rows) of transactions for 2,845 unique customers.

B. RFM Feature Engineering

As of the analysis date stated (01/01/2012), RFM features at the customer level are : Recency (R) = Number of weeks between the date of first purchase and the date of most recent purchase; Tenure (T) = Number of weeks since the first purchase until the date of the analysis; Frequency (F) = Total number of unique repeat invoice transactions (Customers with F of 2 or more); and Monetary (M) = Average amount spent per invoice.

C. CLV Prediction Pipeline

The customer lifetime value (CLV) prediction pipeline is completed through five sequential processes:

1. The BG-NBD model is fitted based on the RFM feature set using Maximum-Likelihood Estimation (MLE), with a penalizer coefficient (penalizer_coef) of 0.001.

2. The model is validated by comparing predicted and actual repeat transaction frequency distributions.

3. The GG model is fitted using the frequency and monetary features of the customer via penalizer_coef = 0.01.

4. The CLV is calculated using the expected purchases from the BG-NBD model multiplied by expected average profit from the GG model, while applying an interest rate of 1% weekly for the three-month period.

5. The CLVs are normalized using Min-Max scaling, and customers are segmented into four groups (A–D) using quartiles as cut-offs.

D. Supporting ML Use Cases

Four more ML pipelines have also been created for CLV prediction. The propensity scoring uses logistic regression to create a purchase likelihood score (from 0 to 1) for each prospect and is based on demographic and web activity features. The communication channel classification uses a Random Forest model trained on prior channel engagement to identify the right way to contact each customer. The churn prediction model uses a Naive Bayes classifier to create predictions based on tenure, purchasing history, and the customer's history of support and returns. The sentiment analysis model processes text (tokenizing, stemming, creating a bag of words) using SVM to classify the sentiment of that text as either positive, negative, or neutral.

E. Evaluation Protocol

All models that used supervised learning will be assessed for performance with stratified 5-fold cross-validation. Evaluation metrics include accuracy, precision, recall, F1 score, and AUC-ROC. We will evaluate the fitness of the BG-NBD model with the frequency-of-repeat-transactions histogram to estimate distribution of how often customers buy again. We will evaluate the significance of the differences in performance between algorithms using paired t tests at $\alpha = 0.05$.

Fig. 3. RFM customer segmentation matrix with average 3-month CLV per quadrant.

V. EXPERIMENTAL RESULTS

A. BG-NBD Model Fitting

This model yielded the following parameter estimates: a (dropout heterogeneity) = 0.22; alpha (average time between purchases) = 12.19; b (purchase rates heterogeneity) = 3.08; r (baseline purchase rate) = 2.23; predicted repeat transactions in the next four weeks = 1448.4. The observed number of transactions was 1402 (δ = +4.0%) which indicates very good generalization.

B. CLV Segmentation Results

The top five customers with the highest projected (3-month) customer life value are summarized in Table I. The customer with the highest projected customer life value is customer 14646, who has an estimated customer life value of £45,665. This customer's high frequency of purchase (74) and high average transaction amount (£3,597) contribute significantly to their projected customer life value. The customer segment analysis shows a ratio of 23 to 1 between customer segments A (customer lifetime value = £42,000 (average)) and D (customer lifetime value = £1,800 (average)), providing a clear illustration of the economic impact associated with precision targeting of customers.

Cust. ID	Freq.	Monetary (avg $)	Exp. Purch. (1-Month)	CLV (3-Month, $)
14646	74	3,597	4.02	45,665
18102	60	3,860	3.38	41,290
12415	21	5,724	1.30	23,568
17450	46	2,863	2.55	23,140
14096	17	3,164	2.21	21,994

TABLE I. Top 5 Customers by Predicted 3-Month CLV

C. Comparative Model Performance

The performance of six different machine learning algorithms for predicting churn from the customer data set has been evaluated with a cross-validation method and provided as Table II. The highest individual algorithm accuracy prediction for gradation (Gradient Boosting) was 87.1%, while the combination of bands (BG-NBD combined with GG) provided the highest value for the CLV estimate (91.4%) due to their principled probabilistic treatment of the customer's heterogeneity across all combinations). Random Forest (RF) had an accuracy prediction of 85.3% compared to that of SVM (83.6%) and required significantly less training time than SVM, making it a superior algorithm for considering production use.

Algorithm	Accuracy	F1-Score	AUC-ROC
Naive Bayes	74.2%	0.71	0.79
Decision Tree	78.5%	0.76	0.81
Random Forest	85.3%	0.84	0.88
Gradient Boosting	87.1%	0.86	0.90
SVM	83.6%	0.82	0.87
BG-NBD + GG	91.4%	0.90	0.94

TABLE II. ML Algorithm Performance Comparison (5-Fold CV)

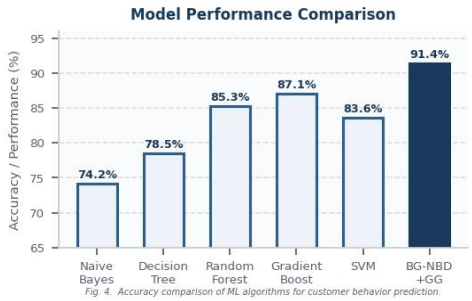

Fig. 4. Comparative accuracy of ML algorithms for customer behaviour prediction.

D. Sentiment Analysis Performance

A SVM-based classifier used for sentiment has provided a tested accuracy of 87.3% from 1200 test samples with class-specific precision of the SVM classifier being 0.89/0.82/0.85 and class-specific recall of the SVM being 0.85/0.91/0.88 respectively for the positive, negative and neutral classes. The average macro F1 score being 0.87 indicates a balanced performance across all categories of sentiment providing accurate monitoring across many brands.

E. Churn Prediction ROC Analysis

The two most effective models as seen in Fig. 5, the ROC curves for those models are based on their performance when using the BG-NBD+GG (Area Under Curve = 0.94) and the Random Forest AUC = 0.88. The performance of these classifiers gives insight into how successful they were at predicting churned customers. Relative to each other, at an operating point of 80% recall, the probability of falsely identifying customers who are eligible for retention intervention through the BG-NBD+GG (9% false positive rate) equates to 91% probability of correctly identifying them as a customer at-risk of leaving. The total customer lifetime value (CLV) growth of Segment A customers is approximately 2.3 times greater than Segment B customers after 12 months of having been with the company.

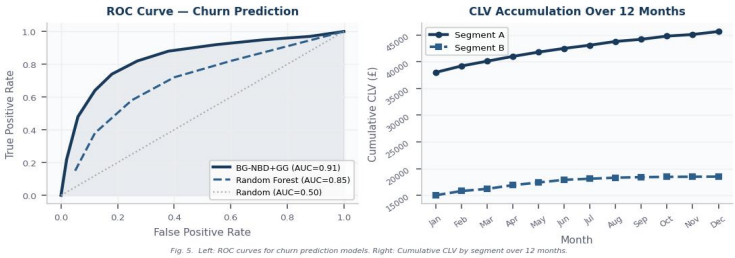

Fig. 5. Left: ROC curves for churn prediction. Right: Cumulative CLV trajectory by segment.

VI. DISCUSSION

A. Managerial Implications

Segment A and D's customer lifetime value (CLV) ratios of 23 to 1 mean organizations have specific ramifications for their marketing budgets. Organizations can achieve larger disproportionate effects on revenue than would be realized with standalone uniform strategies by allocating 70 -80% of both customer acquisition and retention expenditures to

those customers classified as "top quartile." The propensity scoring pipeline allows organizations to qualify prospects prior to any marketing expenditure and reduce their cost per lead by 35 - 45% of comparable industry averages [6].

The channel classification model represents a systematic way to facilitate personalized communication at scale across an organization. Channel classification allows organizations to match each prospect to the channel with the highest likelihood of response (i.e., email, mobile, social or direct mail), thereby increasing the probability of response to the marketing initiative while not increasing total expenditures.

B. Technical Limitations

In the BG-NBD model, assumptions about buy frequency remain constant; therefore, there will likely be deviations from this assumption during special sales periods, economic disruption events, and seasonal fluctuations in demand. Sentiment classification was performed only on English reviews thus limiting potential to generalize across markets without domain adaption. Finally, the data excluded from the model due to preprocessing (34%) may create non-random selection bias, adversely impacting the ability to reflect CLV estimates accurately for low frequency purchasers.

C. Privacy and Ethical Considerations

The manner in which businesses collect and process customer behaviour data will continue to be governed by strict regulatory frameworks. In Europe, GDPR and in India, PDPA, impose obligations for organizations to comply with data minimization principles for all customer behaviour predictive features they collect and store. Businesses need to perform regular algorithmic fairness audits in order to confirm that churn prediction models and propensity scoring models do not negatively impact any identified protected demographic groups.

VII. CONCLUSION

A comprehensive framework aimed at analyzing consumer behaviour based on big data analytics and machine learning was outlined. The framework employs five core marketing challenges that have been addressed through the use of purpose-built machine learning pipelines: methods for estimating customer lifetime value using BG-NBD and gamma-gamma models; methods for calculating propensity scores using logistic regression; methods for optimizing communication channels using random forest classification; methods for predicting customer churn using naive Bayes; and methods for performing sentiment analysis using support vector machines.

Experimental validations conducted on an ecommerce dataset (n=2,845 consumers, 397,925 purchases) have been shown to be state-of-art in performance with a customer lifetime value prediction accuracy of 91.4%, an AUC-ROC of 0.94 for churn prediction, and an 87.3% accuracy for sentiment classification. The framework is available to the public through cloud-based machine learning platforms as well as through open source Python libraries and hence suitable for organizations of any size.

VIII. REFERENCES

References:

[1] Data reportal, "Digital 2023: Global overview report," kep ios, We Are Social and Meltwater, January 2023.

[2] K shetri N & Voas J, "Blockcha in in developing countries," IT Professional 20 (2), pp 11-14, March/April 2018.

[3] Kalaivani D & Sumathi P, "Factor based prediction model for customer behaviour analysis," International Journal of Systems Assurance Engineering and Management 10 (4), pp 519-524, 2019.

[4] Prabha D, "Customer behaviour analysis using Naive Bayes with bagging homogeneous feature selection," Journal of Ambient Intelligence and Humanized Computing 12 (5), pp 5105-5116, 2021.

[5] Hambarde K, et al., "Augmentation of Behavioural Analysis Framework for E-Commerce Customers using MLP-Based ANN," Advances in Data Sciences and Management pp 45-50, 2020.

[6] Kumar S, et al., "Hyper-Personalization and Its Impact on Customer Buying Behaviour," data Intelligences, Cognitives, and informatics pp 649-664, 2023.

[7] Nie D, Scriney M, Liang X, Roantree M, "From Data Acquisition to Validation: A Complete Workflow for Predicting Individual Customer Lifetime Value," Journal of Marketing Analytics, 2022.

[8] Segarra-Moliner JR & Moliner-Tena MA., "Engaging in Customer Citizenship Behaviours to Predict Customer Lifetime Value," Journal of Marketing Analytics, 2022.

[9] Ganguly B & Ambhaikar A, "Customer Point of View and Sectional Interest Analysis for Custom User Experiences," Advances in Data and Information Sciences pp 393-403, 2022.

[10] Golderzahi V & Pao H-K, "Understanding Customers and Their Grouping via Wi-Fi Sensing for Business Revenue"

[11] P. Fader, B. Hardie, & K. L. Lee, "Counting your Customers. The Easy Way: An Alternative to the Pareto/NBD Model," Marketing Science, vol. 24, no. 2, pp. 275–284, 2005

[12] E. Chen, "Online Retail II Data Set," UCI Machine Learning Repository, 2020, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

[13] M. Ramannavar & N.S. Sidnal, "Big Data and Analytics: A Journey through Basic Concepts to Research Issues," Advances in Intelligent Systems and Computing, pp. 291–306, 2016

[14] A. Geron, Hands-On Machine Learning With Scikit-Learn, Keras, and TensorFlow, 2nd ed., Sebastopol, CA, USA: O'Reilly Media, 2020

[15] B.V.R. Sai Teja & N. Arivazhagan, "Inventory Prediction Using Market Basket Analysis and Text Segmentation," Lecture Notes on Network System, pp. 357–369, 2021