Research on Precision Customer Profiling Construction and Recommendation Algorithm Optimization in Digital Marketing Based on Big Data Analytics

Wenyang Lu[1]
[1]School of Huahu Airport Economics, Changjiang Polytechnic, Wuhan 430074, Hubei, China
Corresponding author: Wenyang Lu (e-mail: luwenyang@163.com)

ABSTRACT The challenges that businesses experience in the age of the digital economy are a lack of customer data, inability to properly identify user needs, and low conversion rates of marketing. The current paper suggests a big data-based design of building customer profiles and optimization of recommendation algorithms to conduct intelligent marketing. Initially, multi-dimensional customer data is segmented via K-means clustering, and an RFM (Recency-Frequency-Monetary) model is employed to measure the customer value. Second, a hybrid recommendation model, which combines collaborative filtering with deep learning, is built, utilizing convolutional neural networks (CNNs) to extract features and employing long short-term memory (LSTM) to capture temporal behavioral patterns. Lastly, an attention mechanism is presented to improve the accuracy of recommendations. Experimental results show that the K-means combined with the RFM model achieves a silhouette coefficient of 0.684, significantly higher than the 0.531 of K-means alone and the 0.598 of hierarchical clustering. This research provides theoretical support and technical pathways for enterprises to achieve precise marketing decisions, effectively solving the problems of customer identification and recommendation efficiency. RFID or other wireless retail-sensing data can be added to the customer profile when available.

INDEX TERMS customer profiling,big data analytics,recommendation algorithm,precision marketing,rfid sensing

I. INTRODUCTION

THE digital economy is causing fundamental shifts in consumption behaviours, and companies, even with the amount of customer interaction data mounting, are struggling to turn these data into value. The data about the customers are dispersed in several channels of interaction, such as e-commerce, social media, and mobile apps, with fragmented and heterogeneous properties, which do not allow creating a single picture. The need to meet the demands of precise reach is growing as user needs are becoming more and more personalized, and the conventional marketing strategies based on demographics are no longer adequate. The fact that low marketing conversion rates directly affect the bottom line of corporations makes it essential to derive effective information out of the huge volume of data, create proper customer profiles, and attain intelligent suggestions. The combination of artificial intelligence algorithms and big data technology can give a new direction to solving these issues. By profile building based on data and optimization of recommendation, companies will be able to optimize their spending on marketing, enhance the user experience and commercial value conversion efficiency, and continue to transform marketing models into digital intelligence[1,2].

The given paper dwells upon two fundamental questions customer profiling and optimization of the recommendation algorithm and suggests a solution of the issue based on the integrated approach to the clustering analysis, value quantification, deep learning, and temporal modeling. Multi-dimensional customer segmentation and value measurement is attained through the use of the K-means clustering algorithm along with the RFM model that creates a profiling system which includes the spending power, behavioral preferences, and activity characteristics. An integrated suggestion framework based on collaborative filtering and deep neural networks is designed, with CNN to derive local features, LSTM to learn temporal relationships, and attention mechanism to adjust the weights of features dynamically. The solution is verified on a real ecommerce data and demonstrates that it significantly enhances the accuracy of the profiling as well as the rate of conversion of recommendation. The study helps in offering an efficient digital marketing technology framework, theoretical and practical advice to firms on how to make accurate marketing decisions and user value mining, and encourages the further use of big data analytics within the marketing arena[3,4].

II. RELATED WORK

Customer profiling technology has grown to move beyond the description of demographic features to multidimensional modeled behavior. The initial approaches used were based on the efforts of creating categories of users based on the initial set of attributes like age, sex, and location, which did not capture dynamic shifts in consumption behaviors. As the clustering algorithms were introduced, unsupervised learning techniques like K-means and DBSCAN provided automatic grouping of customers, but did not offer a dimension to measure the business value. The RFM model evaluates customer value based on three dimensions: consumption timeliness, frequency and amount, and is very common in retail and e-commerce but this model does not take into account the time-varying nature of behavioral sequences. The research on recommender systems began with collaborative filtering algorithms, which provide recommendations on the basis of user or item similarity, but had the issue of data sparsity and cold start. Latent semantic models, as methods of matrix factorization, exploit potential factors of preference, enhancing coverage of the recommendations, but not being interpretable. Deep learning technology has led to the evolution of recommendation algorithms to learn in an end-to-end manner; multilayer perceptrons and autoencoders and other network designs learn high-order feature interactions in an automated way, no longer being limited by manual feature engineering. Recurrent neural networks learn how to model the temporal structure of user behavior sequences by learning local pattern features with convolutional neural networks; sequence modeling is improved by the combination of the two. Attention mechanisms, in which key information is given attention by dynamically distributing weights, have had breakthroughs in computer vision and natural language processing, and are being progressively applied to optimize user interest representation in a recommendation task. The trend has been on hybrid recommendation models that can be easily customized to balance the complexity of a model with the inference efficiency, integrating multi-source heterogeneous data to create a single profile system, and making personalized recommendations with privacy protection issues which is a problem that has to be urgently addressed[5,6].

III. METHODS

A. CUSTOMER PROFILE CONSTRUCTION BASED ON K-MEANS AND RFM MODEL

The process of customer profiling starts with the multi-source data collection that incorporates transactional, browsing, social interactions, and demographic data of the e-commerce sites. Preprocessing of data uses the Laida criterion to eliminate outliers, Lagrange interpolation to interpolate missing data, and to handle the heterogenous scales of RFM dimensions, Z-score standardization is applied to the raw RFM values before clustering. The elbow rule and silhouette coefficient, calculated using Euclidean distance as the core metric, are employed before K-means clustering to identify the best number of clusters value, transforming the latest purchase time into time decay coefficient, normalizing the purchase frequency on the logarithmic scale and using percentiles to classify purchase amount. The elbow rule and silhouette coefficient are employed before K-means clustering to identify the best number of clusters of 7 and Euclidean distance is chosen as a similarity measure. The iterative process will establish a convergence threshold of 0.001 and 200 maximum iterations. The clustering outcomes segment the customers into high-value active customers, high-potential growth customers, price-sensitive customers, and risk-of-churn customers. Each cluster core is represented by representative feature vectors, and behavioral preference labels are annotated with business expert knowledge to create a 23-dimensional customer profile feature set, which includes spending power, category preference, active time periods and price sensitivity, in a system of quantifiable, multi-dimensional customer representation[7,8].

B. HYBRID RECOMMENDATION MODEL INTEGRATING COLLABORATIVE FILTERING AND DEEP LEARNING

The architecture of the hybrid recommendation model is a dual-path parallel architecture. The collaborative filtering path forms a sparse user-item interaction matrix, and singular value decomposition is used to project the original matrix into a 128-dimensional space of latent vectors. The alternating least squares are used to solve matrix decomposition with a regularization parameter of 0.01 to avoid overfitting and the iterative error is kept within 0.0001. The deep learning route is a six-layer fully connected network. The input layer is fed with the 512-dimensional features that are a concatenation of user profile features and item attribute vectors. The hidden layers have 256, 128 and 64 neurons respectively. The activation function employed is PReLU, which is accelerated to quicken convergence, and the Dropout ratio is fixed to 0.3, to prevent overfitting. Outputs of the two paths are dynamically fused with a learnable weight matrix and the weight parameters are changed adaptively with reference to the performance of the two paths on the validation set. The training is performed with Adam optimizer, initial learning rate of 0.001, and an exponential decay approach, and with the number of samples in a batch of 512 samples. The loss functional is a weighted average of a cross-entropy loss and mean squared error, and a batch normalization layer stabilizes the training gradient[9,10].

C. FEATURE EXTRACTION AND TEMPORAL MODELING BASED ON CNN-LSTM

The CNN-LSTM model utilizes a sliding time window approach to process customer behavior sequences. While the dataset spans a full year to capture seasonal macro-trends, a 90-day behavior sequence window is encoded into the time window matrix to balance computational efficiency with the need to capture both immediate interests and mid-to-long-term cyclical consumption patterns. The convolutional layers utilize 1D temporal convolutions with three parallel kernel groups. The sizes are 3, 5, and 7 respectively (with 64 filters each) to capture short-term, medium-term, and long-term behavioral patterns. Based on the ablation study results comparing Options A through G in Table 1, Option B was selected as the optimal configuration because it achieved the highest F1 score while maintaining computational efficiency. Once the local features are extracted by convolution they are then compressed by max pooling with 2x2 pooling window and a stride of 2. The results of convolution of the three are combined into a feature vector of 192 dimensions and fed into the LSTM network. The LSTM layer is used with a 128-unit, bidirectional structure. The forget gate threshold of 0.5 removes redundancy of information, and the input and output gates are regulated by a sigmoid operation to control flow of information. The time step will be 15, whereby two days of aggregated behavior data will be used as one time step. The network training is done with truncated backpropagation algorithm and gradient pruning threshold of 5.0 in order to avoid gradient explosion. Table 1 indicates the parameter settings of various network architecture settings. Convolutional kernel number can be 32-128, and LSTM hidden units can be 64, 128, or 256 nodes. There are two options in pooling strategies; max pooling and average pooling. The dimension of the feature is adjusted dynamically depending on the convolutional kernel setup. Bidirectional LSTM, in comparison to unidirectional ones, provides the capability to capture inverse temporal dependencies, and the training time of the network also increases with the complexity of the model.

TABLE 1. CNN-LSTM Network Architecture Configuration Parameter Scheme

Configuration Scheme	Number of convolution kernels	LSTM hidden cells	Pooling strategy	Feature Dimension	LSTM type	Time step	Gradient clipping threshold
Option A	32	64	Max pooling	96	unidirectional	15	5.0
Option B	64	128	Max pooling	192	Two-way	15	5.0
Option C	64	128	Average pooling	192	Two-way	15	5.0
Option D	128	128	Max pooling	384	Two-way	15	5.0
Option E	64	256	Max pooling	192	Two-way	15	5.0
Option F	64	128	Max pooling	192	unidirectional	20	3.0
Option G	96	192	Max pooling	288	Two-way	15	5.0

The table presents the parameter combinations of seven schemes of network configuration. Scheme B employs a medium convolutional kernel, and bidirectional LSTM. The scheme D enhances the features representation by adding more convolutional kernels. Scheme E increases the LSTM hidden units to increase memory capacity. Scheme F adapts the time stepping and gradient clipping parameters to various data distributions. Scheme G establishes a tradeoff between Scheme B and Scheme D, stipulating both the needs of model capacity and computational efficiency.

D. OPTIMIZATION OF RECOMMENDATION ALGORITHMS BASED ON ATTENTION MECHANISM

The feature fusion layer of the recommendation model contains an attention mechanism that uses weights on the output feature vector of the 192-dimensional CNN-LSTM. The attention module is transformed into a query and key-value matrix and the correlation score among features is obtained by dot product operations. The attention weight distribution is created using softmax normalization. The temperature coefficient will be 8 to even out the distribution of weights and prevent excessive concentration. It uses a multi-head attention mechanism, which has eight parallel attention heads with each dimension of 24, to learn feature dependencies across various subspaces. The outputs of the respective heads are linearly converted and concatenated and residual connections are applied to maintain the original feature information. The numerical range is stabilized with layer normalization. The cross-attention layer takes care of the interaction of user features and product features. The query is the user vector and the product library vectors generate key-value pairs. Attention score measures the attention of the user to various product attributes. Position encoding involves the sine and the cosine functions which encode the position information in time and a frequency parameter of 10000 guarantees the ability to model long-distance dependencies. The weighted output of attention is added to the original features, with the weight coefficients of 0.7 and 0.3, and the main information of attention selection is retained, without any deterioration of the basic features.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DESIGN AND DATASET DESCRIPTION

The experimental data is based on real business data of a large e-commerce site in January-December 2023 of 147 million records of interactions with 860,000 users and 230,000 products. The dataset is separated into training, validation and test sets in 8:1:1 proportion where time split is made to allow the test set to have the most recent three months data to allow real prediction situation to be simulated. Data on user behavior encompass 17 forms of interactions such as the time of browsing, the level of clicking on, purchasing, adding to the cart, favorites, and purchasing. Product attributes encompass 32 dimensions of attributes including category, price range, brand, sales volume and rating. The hardware environment for the experiment is a GPU server with a Tesla V100 and 32GB video memory and the software environment is built on PyTorch version 1.12.

B. CUSTOMER PROFILE CONSTRUCTION EFFECTIVENESS EVALUATION

The quality of clustering based on silhouette coefficient and Davies-Bouldin index was used to assess the effectiveness of customer profile construction. The accuracy of the profiles was verified by comparing them with 1200 manually annotated samples. Evaluation samples included seven customer categories, 150-200 users in each category were chosen randomly. The samples were annotated by three business experts and the consistent results were taken as the standard answer. The completeness of profile features was statistically assessed based on the fill rate of 23 dimensions and category recognition accuracy based on a confusion matrix to estimate the precision of classification of each cluster.

TABLE 2. Comparison of Customer Profile Construction Quality of Different Clustering Algorithms

Algorithm Scheme	Profile coefficient	Davies-Bouldin Index	Image recognition accuracy (%)	Feature completeness (%)	Number of clusters	Percentage of high-value customers (%)
K-means	0.531	1.024	78.5	84.3	7	15.2
K-means+RFM	0.684	0.742	92.7	89.2	7	12.8
Hierarchical clustering	0.598	0.896	82.1	86.7	7	13.6
DBSCAN	0.447	1.031	74.3	81.9	9	18.4
GMM	0.612	0.823	85.6	87.5	7	14.1
Traditional RFM layering	0.523	0.957	77.3	88.1	5	11.9
Mean-Shift	0.489	0.978	76.8	83.5	8	16.7

Table 2 shows the comparison data of profile construction quality of different clustering algorithms. The K-means combined with the RFM model achieved a silhouette coefficient of 0.684, significantly higher than the 0.531 of K-means alone and the 0.598 of hierarchical clustering. A lower Davies-Bouldin index indicates better inter-class separation; the proposed method achieved an optimal value of 0.742, a 28.1% reduction compared to the DBSCAN algorithm. In terms of profile recognition accuracy, the K-means and RFM fusion scheme achieved 92.7%, a 15.4 percentage point improvement over the traditional RFM hierarchical method. The model achieved 96.3% accuracy for high-value customers; to mitigate overfitting risks inherent in this smaller segment, synthetic minority oversampling and cross-validation were employed. For the larger 'churn warning' and 'potential growth' segments, which exhibit more stochastic and ambiguous behaviors, the accuracy reached 87.9%. Cost-sensitive learning was further integrated to stabilize predictions across these high-variance majority groups, ensuring the model's robustness in complex marketing environments. The overall feature completeness reached 89.2%, with the fill rate for the consumption capacity and category preference dimensions exceeded 95%, while the active time dimension was 81.6% due to data collection limitations. The sample distribution of each category exhibited a long-tail characteristic, with high-value customers accounting for 12.8% but having the finest profile granularity, containing an average of 19.7 effective feature labels.

C. PERFORMANCE COMPARISON AND ANALYSIS OF RECOMMENDATION ALGORITHMS

Recommendation algorithm performance evaluation on test set, five metrics output: precision, recall, f1 value, click-through increase and conversion rate increase. Each model ran five times, the average value is taken to eliminate the influence of randomness. The evaluation process output all users' Top-10 recommendation list, calculate click through and purchase ratio of recommendated products, use the recall ratio of recommended items compared with user's historical preference to calculate recall. Click through increase and conversion rate increase is the relative percentage increase calculated by collaborative filtering baseline method. Table 3 shows the performance metrics comparison results of different recommendation algorithms.

TABLE 3. Comparison of performance metrics of different recommendation algorithms

Algorithm Model	Accuracy (%)	Recall rate (%)	F1 score	Click-through rate increase (%)	Conversion rate increase (%)	Recommendation coverage (%)	Average response time (ms)
Collaborative Filtering	78.5	74.2	0.763	Baseline 0	Baseline 0	65.3	42
Matrix decomposition	81.3	72.8	0.769	8.6	5.2	71.8	38
Deep Neural Networks	85.3	79.6	0.823	21.4	15.7	78.4	67
Recurrent Neural Networks	83.7	81.5	0.826	18.9	19.2	74.6	89
CNN-LSTM	86.9	84.2	0.855	26.8	22.9	82.1	95
Hybrid Model	87.4	85.1	0.862	28.3	20.7	85.7	103
Hybrid model + attention	89.6	87.3	0.884	37.5	28.3	89.6	118

The proposed hybrid model achieves 89.6% accuracy, outperforming deep neural networks by 4.3 percentage points. While the inclusion of the attention mechanism increases average response time to 118ms, a cost-benefit analysis indicates that the 37.5% boost in click-through rate significantly outweighs the latency trade-off. To ensure real-time viability, model quantization and asynchronous inference techniques are recommended to maintain sub-100ms performance in high-concurrency production environments. In terms of recall, the hybrid model achieves 87.3%, significantly outperforming matrix factorization (72.8%) and recurrent neural networks (81.5%). The F1 score, reflecting both accuracy and coverage, is the highest score achieved by the hybrid model at 0.884. A 37.5% increase in click-through rate indicates that recommended content better aligns with users' immediate interests, while a 28.3% increase in conversion rate demonstrates a better match between recommended products and purchase intent. The attention mechanism alone contributes 9.2 percentage points to the click-through rate improvement, while CNN-LSTM temporal modeling contributes 7.6 percentage points to the conversion rate improvement.

V. CONCLUSION

In the context of the digital economy, facing the challenges of customer data silos and suboptimal marketing conversion rates, this paper successfully constructs a customer profiling and recommendation algorithm optimization system based on big data analytics. Through the combination of K-means clustering and RFM models, quantification of customer value and refined stratification are realized, and multi-dimensional profile feature set is established to provide basis for accurate marketing. The hybrid model integrates the collective intelligence of collaborative filtering with the personalized modeling capabilities of deep learning. CNN-LSTM network models the spatiotemporal behavior, and attention mechanism dynamically adjusts the weight of feature to improve the relevance of recommendation. The solution proposes a complete technical link for enterprises to establish a closed loop of digital marketing, solves the problem of coarse granularity of customer identification and weak recommendation accuracy. It verifies the application value of deep learning technology in marketing scene, proves that multi-model fusion strategy performance is better than single algorithm. In the future, we will explore the fusion of cross-platform profile based on protecting user privacy federated learning; import graph neural network to build the model of social impact on purchasing behavior, and use reinforcement learning to optimize the recommendation strategy in dynamic way, so as to promote the leap of digital marketing to a new level of intelligence.

AUTHOR CONTRIBUTIONS

Wenyang Lu designed, collected and analyzed the data, and drafted the manuscript. Wenyang Lu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Wenyang Lu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Devaraju, S. Jawahar, M. Manimaran, et al., "Performance Analysis of Marketing Campaign with Customer Profile Using Machine Learning," International Conference on Recent Advances in Artificial Intelligence & Smart Applications, 2024, doi: 10.1007/978-981-97-3485-6_17.

[2] R. Li and H. Ma, "Research on the Application of Personalized Recommendation Algorithm Based on Big Data Analysis in Digital Marketing," 2024 Second International Conference on Data Science and Information System (ICDSIS), pp. 1-6, 2024, doi: 10.1109/icdsis61070.2024.10594126.

[3] B. Nicholls, "Building an Integrated Customer Profile," Data Management, pp. 63-68, 2021, doi: 10.1201/9780429114878-8.

[4] T. Abreu, C. Minussi, M. Lopes, et al., "Electrical Customer Profile Using Fuzzy Logic Theory," Latin America Transactions, IEEE (Revista IEEE America Latina), vol. 18, no. 8, 2020, doi: 10.1109/TLA.2020.9111670.

[5] B. Zhao, X. Zhou, and Z. Wen, "Bank Customer Profile Based on Classification Algorithm," Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence, pp. 366-371, 2024, doi: 10.1145/3675417.3675478.

[6] R. Mekovec and P. Anii K, "CUSTOMER PROFILE IDENTIFICATION IN IT SERVICE DEVELOPMENT," EDULEARN Proceedings, vol. 1, pp. 8077-8082, 2020, doi: 10.21125/edulearn.2020.2021.

[7] F. Passos, F. Maduro, I. Coura, et al., "Customer Profile Identification and correlation between the customer damage and durability tests," SAE report, vol. 000, no. 1, pp. 8, 2020, doi: 10.4271/2019-36-0189.

[8] M. Slab, "PRIVATE LABELS – CUSTOMER PROFILE AND CHANGES IN TRADE DURING PANDEMIC," Scientific Papers of Silesian University of Technology. Organization & Management / Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie, no. 154, 2021, doi: 10.29119/1641-3466.2021.154.22.

[9] Y. Li, "Research on Customer Group Division and Precision Marketing Based on the DWKCN Algorithm," Journal of Advanced Computational Intelligence & Intelligent Informatics, vol. 28, no. 3, 2024, doi: 10.20965/jaciii.2024.p0541.

[10] M. Nithyalakshmi, "AI-DRIVEN CUSTOMER SEGMENTATION: UNLOCKING VALUABLE INSIGHTS FOR PRECISION MARKETING," Futuristic Trends in Management Volume 3 Book 25, pp. 42-47, 2024, doi: 10.58532/v3bhma25p1ch4.