Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A COMPARATIVE STUDY OF MACHINE LEARNING FOR AGE GROUP IMPUTATION IN LARGE-SCALE E-COMMERCE DATA

JURON PAIK
Professor. Department of Digital Information and Statistics, Pyeongtaek University, S. Korea
E-mail: jrpaik@ptu.ac.kr

ABSTRACT

Current privacy regulations and the prevalence of voluntary non-disclosure have led to significant gaps in demographic data within e-commerce platforms, severely hindering personalized marketing efforts. This study proposes a machine learning–based approach to address the problem of missing age group information in large-scale e-commerce platforms. Resolving this issue can enhance the personalization potential while respecting user privacy constraints. The dataset used in this research comprised actual e-commerce platform data, including customer behavior logs, product attributes, and temporal and regional variables. Initially, five classification models—logistic regression, decision tree, random forest, k-nearest neighbors (knn), and XGBoost—were compared. Preliminary experiments revealed that logistic regression performed relatively poorly; therefore, it was excluded from the final comparison, and hyperparameter optimization was performed on the remaining four models. Model performance was evaluated on the validation set using accuracy and F1-score as the primary metrics. Experimental results showed that the random forest (default configuration) achieved the highest performance with approximately 78\% accuracy, while XGBoost, although underperforming in the default setting, improved to a comparable level after optimization. In contrast, decision tree and knn showed limited improvement from optimization, with performance in some cases declining compared to the default setting. Feature importance analysis identified behavioral frequency, a specific event type, gender, and product attributes as key factors. This research contributes by empirically demonstrating the feasibility of constructing age group prediction models using large-scale e-commerce data, thereby offering a practical strategy for addressing missing demographic information under privacy constraints. Furthermore, the feature importance analysis provides actionable insights for target marketing and personalized recommendation system design. Conclusively, this study empirically demonstrates that behavioral logs alone are sufficient to predict demographic attributes with high accuracy. The proposed Random Forest-based framework offers a cost-effective and privacy-preserving alternative to complex deep learning models for practical deployment in real-world e-commerce systems.

Keywords: *Age Group Prediction, Machine Learning, Privacy-Preserving Data Analysis, Personalized Recommendation, E-commerce Data*

1. INTRODUCTION

In the e-commerce industry, personalized recommendation systems and targeted marketing are widely employed to enhance user experience and increase sales. The success of personalized services is heavily dependent on the demographic information contained in user profiles, with gender and age group being particularly critical attributes that significantly influence product preferences and consumption patterns. For instance, preferred product categories and responses to marketing campaigns can vary substantially depending on a user's age group. Consequently, gender and age group information are essential for improving the performance of recommendation algorithms and for effectively segmenting customers in e-commerce platforms.

However, in real-world online shopping environments, user profile information is often incomplete. Prior studies [1, 2, 3] indicate that this is largely due to privacy concerns, which lead users to avoid providing demographic attributes such as age or gender, or to enter inaccurate information. In practice, data from Google Analytics [4] also show frequent missing values for age and gender, closely linked to users' tendency to withhold personal details. This results in high rates of missing values in profile datasets, with age group data being particularly prone to omission. When new users choose not to disclose their age group, the overall reliability of the user database declines, ultimately

reducing the accuracy of recommendation systems and hindering both user experience and marketing efficiency. Accordingly, there is an increasing demand for technical methods to accurately supplement such missing profile information in e-commerce environments.

To address this challenge, prior research has explored various approaches that leverage user behavioral data to estimate missing profile attributes. As age group is among the most influential attributes for personalized recommendations and fine-grained marketing, many studies have sought to infer it. Existing research has largely focused on predicting user characteristics based on online behavioral data such as social media activity and web search logs. For example, studies on social media platforms like Twitter have used linguistic patterns and social network structures to automatically estimate user age and gender [5, 6, 7]. Other studies have demonstrated that demographic attributes can be inferred solely from web browsing and search histories [8, 9, 10, 11].

Nevertheless, empirical research that predicts user age groups and supplements missing profile information using only transactional behavior logs—such as purchase and browsing histories—directly from e-commerce platforms remains relatively scarce. For instance, Yehezkel and Resheff [12] attempted to predict certain demographic attributes based on purchase history models, yet such approaches remain limited, and studies utilizing real-world commercial platform data are rare. Similarly, while Kooti et al. [13] and Hendriksen et al. [14] analyzed purchase and clickstream data, few empirical studies have directly addressed age group prediction, highlighting a research gap in this domain.

To fill this gap, the present study proposes a method for supplementing missing profile information using real user event log data from a major domestic online shopping platform. The dataset includes over 250,000 users, approximately 5.88 million event logs, and around 280,000 products, encompassing a wide variety of user interaction events such as page views, clicks, and purchases. By leveraging such large-scale, real-world data, this study aims to ensure both the effectiveness and generalizability of the proposed method.

The proposed approach focuses on supplementing two key profile attributes: gender and age group. Gender imputation is conducted by analyzing users' product browsing and purchasing patterns, utilizing differences in preferred product categories and click paths observed between male and female user groups. Age group prediction is performed using machine learning classification models. Behavioral features are extracted from each user's event logs and used as inputs to various algorithms to predict the user's age group (e.g., 20s, 30s, etc.). This approach offers empirical evidence for a research area that has been underexplored in e-commerce and contributes to the development of profile completion techniques applicable to personalized recommendation systems and customer relationship management (CRM).

Despite advancements in data analytics, imputing demographic data remains a critical challenge due to the 'sparsity' problem in real-world tabular data and tightening privacy laws (e.g., GDPR). While recent studies explore Deep Learning or LLMs, these approaches often incur high computational costs unsuitable for real-time industrial applications. Therefore, a theoretical gap exists for a model that balances predictive performance with operational efficiency.

To address this, we posit the following hypothesis: User behavioral patterns (e.g., click frequency, purchase sequences) contain latent signals that are statistically sufficient to reconstruct missing demographic attributes (age and gender) without explicit user input. Based on this hypothesis, this study designs a two-stage imputation framework: first inferring gender via collaborative filtering-inspired methods, and subsequently predicting age groups using optimized machine learning classifiers.

The remainder of this paper is structured as follows. Section 2 reviews related work and existing approaches. Section 3 describes the dataset and details the proposed gender imputation and age group prediction methodology. Section 4 presents the experimental design, results, and performance evaluation, followed by discussion. Finally, Section 5 concludes the paper and outlines future research directions and implications.

2. RELATED WORKS

2.1 Missing Value Imputation in E-Commerce Data

In e-commerce environments, it is common for certain attributes in user profile data to be missing or incomplete. For instance, many online shopping platforms do not require users to provide demographic information such as age or gender during registration, or users may deliberately withhold such details, leading to missing values in profile datasets [15]. Various approaches have been studied to address this missing value problem. The most basic methods involve replacing missing

values with statistical estimates such as the mean, median, or mode, which offer computational efficiency for large-scale data. However, these simple imputation techniques can distort the underlying data distribution and fail to capture relationships between variables. This limitation is particularly pronounced in e-commerce datasets, which often contain a large proportion of categorical or non-numeric attributes [16,17, 18].

More advanced approaches, such as Multiple Imputation by Chained Equations (MICE) and K-Nearest Neighbors (KNN) imputation, have been applied to supplement missing user data. MICE predicts and iteratively refines missing values for each variable through sequential regression models, making it effective for structured data imputation. However, it is computationally expensive and time-consuming for large datasets [19, 20]. KNN imputation, on the other hand, leverages record similarity to infer missing values, but nearest neighbor searches become highly inefficient when applied to large-scale e-commerce data [17, 21, 22, 23, 24]. Random forest–based imputation methods have also been proposed. For example, Stekhoven and Bühlmann's missForest algorithm employs decision-tree ensembles to predict missing values from observed variables, effectively capturing complex interactions while maintaining strong performance and scalability in large datasets [25, 26].

With the advancement of deep learning, generative model–based imputation methods have emerged [27, 28, 29, 30, 31]. For example, generative adversarial network (GAN)-based approaches learn complex distributions among variables to generate plausible replacements for missing data, often outperforming traditional techniques. Yoon et al. introduced the GAIN model, which, along with later variants such as MI GAN and WSGAIN, demonstrated high accuracy and efficiency even in high-dimensional datasets with complex missing patterns. Nonetheless, GAN-based methods are hindered by high model complexity and substantial computational costs, making them difficult to implement in practical industry settings.

Recently, there has been a growing interest in using large language models (LLMs) for data imputation. For example, Lim et al. [32] proposed DrIM, an LLM-based imputation method, while Hayat and Hasan [33] introduced CRILM, which leverages natural language descriptions for missing value completion. In addition, Wang et al. [34] presented the UnIMP framework, which applies high-order message passing for mixed-type data imputation. These methods demonstrate strong generative and imputation capabilities but still face the limitation of very high computational costs.

Meanwhile, domain-specific lightweight machine learning methods tailored for e-commerce have been proposed. Antwarg Friedman et al. [35] addressed the data completion problem in product attribute datasets by introducing two practical alternatives to LLM-based methods. The first leverages unstructured text, such as product titles and descriptions, extracting keywords to populate missing structured attribute fields. The second approach exploits attribute values of similar or identical products in large-scale marketplaces, using nearest-neighbor–based voting strategies. By calculating product similarity with pre-trained language model embeddings, they supplemented missing attributes efficiently. Experiments on approximately one million product records across categories such as sporting goods, auto parts, and computers demonstrated that these lightweight methods achieved accuracy comparable to GPT-based imputation techniques while requiring significantly lower computational resources. This highlights their practicality in large-scale e-commerce environments, where traditional statistical methods are insufficient.

In summary, missing value imputation techniques have evolved from simple statistical replacements (mean, mode) to more sophisticated machine learning methods such as MICE, KNN, and random forest algorithms, and more recently, to deep learning and embedding-based approaches. Ongoing research continues to explore efficient and scalable methods tailored to the unique characteristics of e-commerce data. Building on this progression, the present study proposes an effective methodology for imputing and predicting missing demographic attributes—specifically, gender and age group—within user profiles.

2.2 User Age Group Prediction Methods

A substantial body of research has sought to predict demographic attributes (e.g., age group, gender) using user behavioral logs across various domains. Early studies in the web domain demonstrated that demographic characteristics could be inferred solely from online activity. Murray and Durrell [36] analyzed search queries and browsing histories using latent semantic analysis (LSA) vectorization combined with neural network models. Although their accuracy was limited, they highlighted the feasibility of inferring user gender and age group from behavioral data, even for anonymous users. Hu et al. [37] developed a Bayesian prediction model at Microsoft Research Asia, leveraging a user–webpage bipartite graph and

smoothing techniques based on clickstream data. Their approach significantly improved demographic inference, achieving F1-scores of approximately 30.4% for gender and 50.3% for age prediction.

Since then, demographic inference methods based on behavioral logs have evolved considerably. Beyond simple measures such as click counts or page visit frequencies, researchers have incorporated query sequences [38], high-dimensional features extracted from search logs [39], and social media content and network structures [40]. Ren et al. [41] provide a comprehensive review, noting that data sources such as web search logs, clickstreams, and social media activity offer substantial information for predicting gender and age group, thereby serving as alternative sources of demographic inference when profile data are incomplete.

User behavioral data in e-commerce contexts have also emerged as a valuable resource for demographic prediction. Purchase histories, clickstreams, and product reviews in both online and offline marketplaces capture users' interests and preferences, which are often strongly correlated with age group and other demographics. Wang et al. [42] proposed a Structured Neural Embedding (SNE) model, trained on approximately 1.2 million transaction records from a large Chinese retailer, including labeled demographic data for over 57,000 customers. The model learned user representations from shopping basket data and simultaneously predicted multiple demographic attributes (e.g., gender, age, marital status) through a multitask learning framework, achieving superior accuracy and F1-scores compared to traditional classifiers. Such purchase-based approaches provide a theoretical and empirical foundation for studies aimed at imputing missing profile attributes. Similarly, Resheff and Shahar [12] integrated multiple e-commerce transaction logs (purchases, returns, searches, and clicks) to construct user embeddings and predict attributes such as gender, age, marital status, and parental status. Their method, which employed fused embeddings across textual, numeric, and categorical variables, demonstrated robust performance even in noisy industrial settings, underscoring its practical relevance.

Beyond e-commerce, researchers have increasingly explored mobile sensor and physical activity data for demographic prediction. Zhong et al. [43] analyzed call logs, app usage, and mobility patterns from the Nokia MDC dataset, constructing machine learning–based contextual features and achieving high accuracy in predicting age and gender through multitask learning. Felbo et al. [44] applied convolutional neural networks (CNNs) to temporal patterns in smartphone call metadata, demonstrating effective age and gender prediction even under conditions of low user activity. Jiang et al. [45] utilized step-count time-series data from over 39,000 WeChat users, developing a recurrent neural network (RNN)–based model that predicted gender and classified age groups (youth, middle-aged, senior, elderly) with high weighted F1-scores, effectively addressing class imbalance.

Recent studies have extended this line of research to wearable devices. For example, Ruhan et al. [46] trained an RNN-LSTM model using accelerometer and gyroscope sensor data from walking activity, achieving a gender classification accuracy of 94% and an age prediction performance of $R^2 = 0.83$. Similarly, Roy et al. [47] employed CNN models on wearable sensor data, reaching approximately 90% gender classification accuracy.

These findings demonstrate that demographic prediction has expanded beyond web and commerce logs to include physical activity data, providing high predictive accuracy for attributes such as age and gender. Collectively, these studies underscore the potential for leveraging diverse behavioral and sensor data sources to supplement missing demographic information in user profiles.

2.3 Distinctiveness and Contributions

As reviewed in Section 2.2, prior studies have demonstrated the feasibility of predicting user profiles from behavioral logs and shown that missing demographic information can be supplemented using machine learning. Approaches that predict multiple attributes from purchase logs have provided meaningful solutions to the practical challenge of partially incomplete user data. However, most existing studies have either addressed age group prediction and profile completion separately or implicitly handled missing values within multi-attribute prediction models.

In this study, we propose an approach that simultaneously predicts user age groups using only e-commerce web log data while explicitly supplementing missing gender information—one of the input features for prediction—based on behavioral patterns. Although our method shares a conceptual foundation with previous research utilizing web search logs or offline purchase records to infer demographics, it is distinctive in that gender imputation is performed through a dedicated prediction module prior to model training. This ensures data completeness and enhances the final performance of the age prediction model.

Unlike conventional multi-attribute learning approaches, our framework treats gender as an auxiliary feature. Gender is first supplemented

through a separate prediction model and then integrated into the age group prediction stage. This design simplifies the overall model structure and improves interpretability. Furthermore, the gender correction step leverages behavioral differences observed between male and female users in their site usage patterns, offering greater data utilization efficiency compared to arbitrarily imputing or excluding users with missing gender information.

In summary, this study proposes a modeling framework tailored for predicting user age groups through e-commerce log analysis, while integrating behavior-based gender imputation to address missing values. Empirical experiments conducted on real user log data from a major domestic e-commerce platform validate the improvement in prediction performance and practical utility of the proposed approach. These contributions highlight the potential of our method as a foundational technology for more precise user targeting in personalized marketing and recommendation systems.

While prior works have focused on unstructured data (text, images) or complex deep learning architectures, they often overlook the trade-off between interpretability and resource efficiency in tabular transaction data. To bridge this gap, this study formulates the following research questions:

- RQ1:Can missing demographic data be accurately imputed solely using transactional behavioral logs?
- RQ2:Which machine learning algorithm offers the optimal balance between accuracy and computational efficiency for large-scale e-commerce data?
- RQ3: What are the key behavioral determinants (features) that distinguish user age groups?

3. METHODOLOGY

3.1 Data Description and Preprocessing

This study utilized log data collected over approximately two months, from June 3 to August 4, 2021, from an online fashion commerce platform. The dataset consisted of three main components: product data, user data, and event data. The key characteristics of each dataset are described below.

- Sampled Products: The product dataset contained information on approximately 283,326 items, including product identifier (item_no), product name (item_name), image filename (image_name), price (price), hierarchical category codes and names (category1 ~ 3_code/name), and brand identifier and name (brand_no, brand_name). Because category hierarchies were deemed critical for capturing product characteristics, category codes rather than names were used in this study. Data cleaning was required as some category names did not map one-to-one with codes. Duplication issues were observed in middle and sub-category levels (category2, category3), where identical names corresponded to multiple codes. To resolve this, the most frequent code was designated as the standard for each duplicated category name, and all records were aligned accordingly. After this refinement, the number of unique values was reduced from 114 to 105 for category2 and from 768 to 730 for category3, improving category consistency. Additionally, approximately 80 obsolete category labels (e.g., containing the marker _OLD), which represented only 0.01% of the data, were removed to eliminate noise. The final cleaned dataset contained 283,296 records. Missing values were minimal—only three cases for brand_no and 19 cases for brand_name (0.0077% of the total)—and were replaced with "unknown". Finally, duplicate, or redundant textual attributes such as product name, category name, and brand name were removed, leaving only essential features. The final product dataset therefore comprised attributes such as product identifier, price, category1/2/3 codes, and brand identifier.
- Sampled Users: The user dataset contained information on approximately 254,958 platform users, including user identifier (user_no), birth date (birth_date), and gender (gender). Missing values were identified for 40,948 gender entries (16.1%) and 23,676 birth date entries (9.3%). Although these missing rates were not excessively high, both gender and age group are critical features for personalized recommendation and behavioral analysis. Accordingly, appropriate imputation strategies were deemed necessary.
- Sampled Events: The event dataset comprised approximately 5,880,407 user interaction logs capturing various activities on the shopping platform. Attributes included session identifier (session_id), event timestamp (event_timestamp), event type (event_name), user identifier (user_no), product identifier (item_no), device type (device_type), mobile brand, and model information (mobile_brand_name, mobile_model_name, operating_system_version), user access country (country), region (region), and platform (platform). Since over 99.7% of events

originated in South Korea, the analysis was restricted to Korean users by filtering for records where country = "South Korea". Additionally, six fields considered irrelevant for analysis (e.g., session identifier and certain device details) were removed. Missing value analysis revealed that only the region field contained gaps, with 335 missing cases. Given their negligible proportion and lack of systematic patterns, these were replaced with "unknown".

Among the datasets described above, the user event logs are particularly valuable for supplementing missing gender information. In this study, for users with missing gender values, interaction patterns in the event data were analyzed to predict gender and thus complete the dataset. Since gender serves as a key input feature for the subsequent age group prediction model, it must be imputed in advance. After reviewing various alternatives, this study adopted a behavior-based gender inference method inspired by collaborative filtering. Specifically, gender was predicted from event log interaction patterns, while missing age group information (due to absent birth dates) was later estimated separately using machine learning models. The detailed procedure for gender prediction is as follows:

1. Identification of Missing Users: All user IDs with missing gender information were extracted from the user dataset.
2. Collection of User-Specific Item Lists: For each user with missing gender values, a list of product identifiers (item_no) corresponding to items they had viewed or purchased in the event logs was compiled.
3. Estimation of Product-Level Gender Distribution: The gender distribution of other users who had viewed or purchased the same items was calculated. For example, if 80% of previous purchasers of a given product were female, the product was considered to have a high female preference.
4. Gender Probability Assignment and Imputation: The gender distribution obtained in the previous step was treated as a probability, and missing gender values were imputed accordingly. For users with multiple product interactions, aggregated probabilities were computed to determine the final gender assignment. This method is conceptually like collaborative filtering, as it infers missing values by leveraging the gender information of user groups with similar product consumption tendencies.
5. Handling of Remaining Missing Users: Users for whom gender could not be imputed through this method were classified as low-activity users, since they exhibited little or no activity in the logs. As interaction-based inference was infeasible for these cases, their gender remained unassigned, and they were excluded from subsequent prediction tasks while being managed separately.

As a result of imputing missing gender values, the proportion of female users in the dataset slightly decreased, while the proportion of male users increased by a corresponding amount. This outcome indicates that many of the previously missing gender entries were inferred as male. Consequently, the gender distribution distortion was reduced, resulting in a more balanced dataset better suited for analysis. Figure 1 provides a visual comparison of the gender distribution before and after imputation.

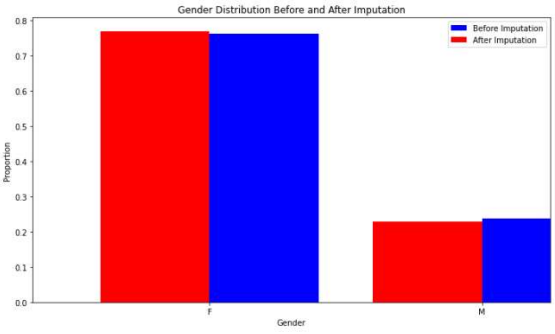


Figure 1: Gender Distribution Before and After Imputation of Missing Values

Next, missing values for age group (derived from birth date) were addressed using a machine learning–based prediction strategy. Unlike gender, age group is a numerical attribute that cannot be easily inferred through simple heuristics; therefore, training classification models that utilize diverse features was considered more effective than rule-based or statistical substitution.

To this end, an age_group feature was generated from the birth date (birth_date) attribute. First, 535 erroneous entries containing non-existent dates (e.g., "8888-01-13") were treated as missing. Then, based on August 2021 (the end of the data collection period), each user's chronological age was calculated and categorized into discrete age groups: under 10, teens, 20s, 30s, 40s, 50s, and 60 and above. This categorical variable was added as a new column (age_group) in the user dataset. For users with valid birth date entries, the corresponding age group was thus completed, while users with missing birth dates continued to have missing values for age_group. These users constituted the target group for age

prediction in this study. Rather than imputing these missing values using averages or simple neighborhood-based rules, this study applied supervised learning models. Specifically, users with complete behavioral logs and demographic information served as the training set, and the trained model was then used to predict the missing age groups.

The input features were extracted from user behavioral logs, with event data serving as the primary source of predictive variables. Building on the event dataset refined for South Korean users, additional feature engineering and exploratory visualization were performed. The event_name field in the user behavior logs consisted of four event types: click_item, like_item, add_to_cart, and purchase_success. These events represent key indicators of user interest and engagement, and both their occurrence and frequency were incorporated as predictive features. Concretely, one-hot encoding was applied to indicate whether each event type occurred, and overall frequency distributions were analyzed. The analysis revealed that click_item events accounted for most interactions, followed by purchase_success, add_to_cart, and like_item in descending order of frequency. Figure 2 visualizes the relative frequencies of these event types, illustrating that user activity was predominantly click-driven. This imbalance suggests a potential risk of biased model learning toward click events.

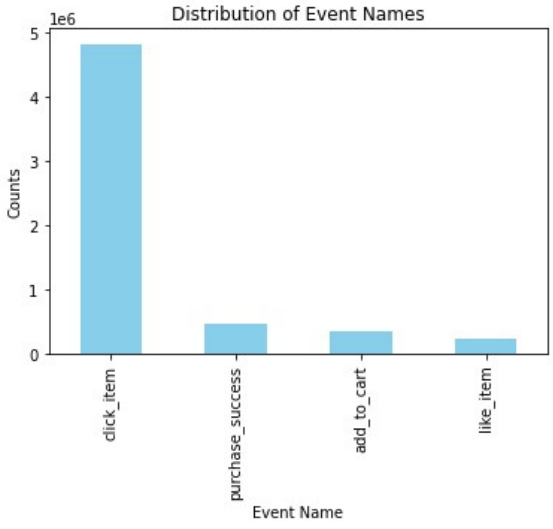


Figure 2: Raw Frequency Distribution of Event Types

To address the asymmetry observed in event distributions, this study applied a weighting strategy that assigned importance-based weights to each event type. Specifically, purchase_success events were regarded as stronger indicators of user intent than simple clicks, and weights were assigned as follows: click_item = 1, like_item = 2, add_to_cart = 3, and purchase_success = 4. Using these values, a cumulative indicator feature, weighted_event, was constructed for each user. This approach emphasized meaningful behaviors such as purchases while reducing the disproportionate influence of click events, thereby improving the model's ability to distinguish between users. Figure 3 presents the log-scale distribution of the weighted event feature, illustrating that the excessive influence of click events was adjusted while more meaningful events, such as purchase_success, gained greater relative importance in model training.

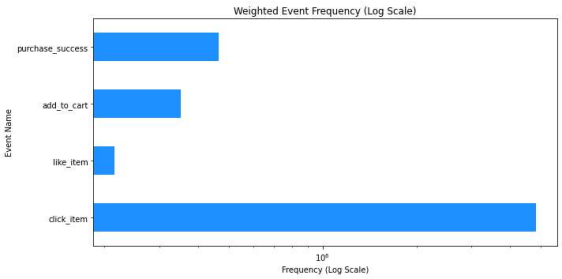


Figure 3: Log-scale Distribution of Weighted Event Types

In addition to event-type weighting, temporal aspects of user activity were also incorporated as an important dimension of feature construction. Specifically, the event_timestamp was converted into DateTime format, and event occurrences were analyzed by date and hour. Figure 4 illustrates the time-series variation in the total number of daily events, revealing substantial fluctuations in user activity. Overall activity increased after mid-July compared to early June, while sharp declines were observed on certain dates. Such analysis provides insight into the temporal distribution of training data and informs how the model should adjust its emphasis when events are concentrated at specific points in time. Moreover, examining hourly activity patterns and their variation across age groups enables the construction of more refined predictive features that embed temporal dynamics, offering valuable implications for modeling the temporal regularities of user behavior.

Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

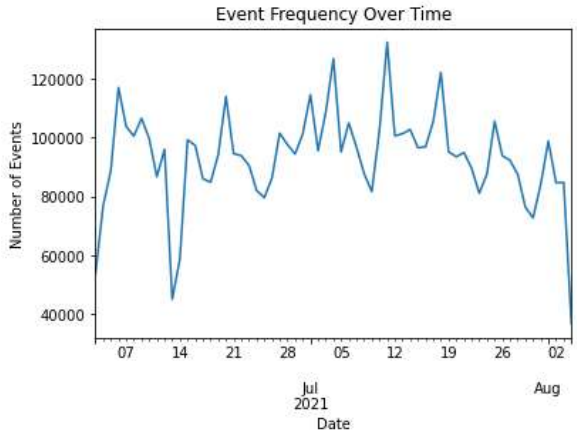


Figure 4: Time-series Variation in Daily Event Counts

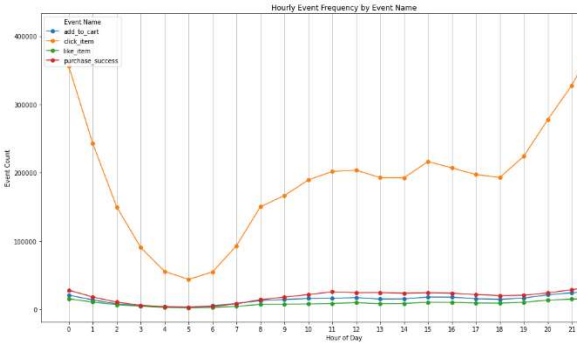


Figure 5: Hourly Distribution of Even Types

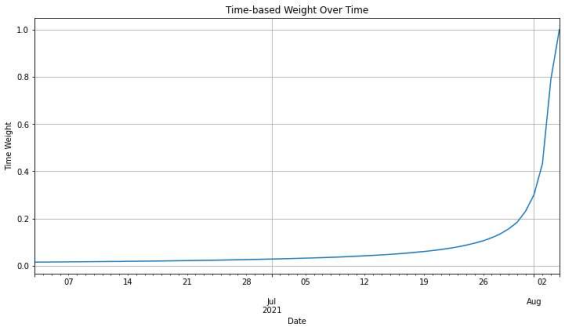


Figure 6: Pattern of Increasing Time-based Weights

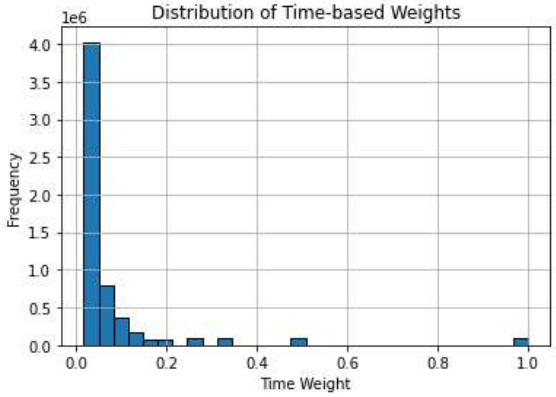


Figure 7: Histogram of Time-based Weight Distribution

Building on the temporal analysis, event distributions were further examined by hour of day. Figure 5 illustrates the hourly frequencies of each event type, showing that user activity was largely concentrated during late-night hours (22:00–01:00). The event, purchase_success, tended to occur more frequently in the afternoon and evening than during early morning hours. This finding suggests that user consumption patterns are skewed toward specific time periods, indicating that hourly activity features may serve as useful predictors for differentiating users in age group or gender classification models.

In addition, a time-based weighting scheme was designed to assign greater importance to more recent events. Figure 6 illustrates that the assigned weights gradually increase over time, with events occurring on the most recent dates receiving values close to the maximum of 1. Figure 7 presents a histogram of the time-based weight distribution, showing that while most events were assigned relatively low weights, a smaller portion of recent events were given high weights. Through this design, the model is guided to place greater emphasis on recent user behaviors as time progresses, thereby improving its ability to capture timely and context-relevant patterns.

Next, an analysis was conducted incorporating users' residential region information. While the region attribute itself does not directly explain age group, it can provide relative meaning for user activity when interpreted in relation to regional population size. Based on user logs containing regional information, event frequencies were aggregated by region. As shown in Figure 8, events were heavily concentrated in the Seoul and Gyeonggi areas. However, evaluating regional activity solely by raw frequency can lead to a metropolitan bias, since event counts are naturally higher in densely populated cities. This, in turn, may result in underestimating the activity levels of users in smaller regions. For example, if a user from a less populated area generated several events comparable to those of users from major metropolitan areas, that user could reasonably be interpreted as exhibiting much higher activity relative to the average within their own region.

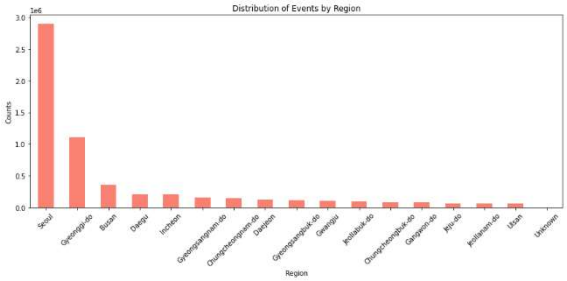


Figure 8: Regional Distribution of Event Frequencies

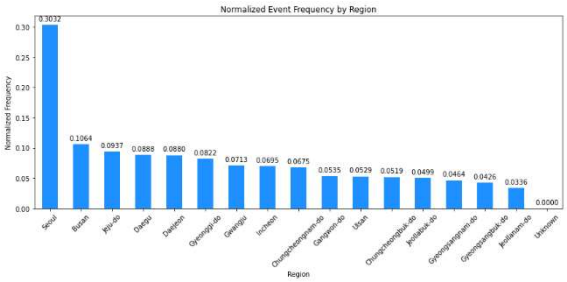


Figure 9: Population-Normalized Event Ratios by Region

To quantify this effect, monthly regional population data for July–August of the same year were obtained from the Korean Statistical Information Service (KOSIS) and combined with the event dataset. Event counts were then normalized by regional population, and the results are presented in Figure 9. After normalization, Seoul still exhibited a relatively high level of activity; however, regions with smaller populations, such as Jeju and Busan, also showed high event occurrence rates. This indicates that user activity can be evaluated more accurately by correcting for regional bias. By applying these normalized values as regional weights, the analysis ensured that active engagement in less populated areas would not be underestimated.

After preprocessing and cleaning the product, user, and event datasets, an integrated dataset for model training was constructed. To achieve this, the datasets were merged using common keys. Each event record was combined with the corresponding user and product information to form a single row, where the user ID (user_no) and product ID (item_no) served as keys to match users with the products they interacted with. As a result, each row of the event log included additional attributes such as the user's gender and age group, as well as product-related features such as price and category codes, thereby forming the final feature dataset.

In this dataset, the target variable was the user's age group (age_group), which was attached to each event record to indicate the age group of the user who generated the event. Event records associated with users who had missing age information (due to unreported birth dates) therefore lacked target labels. These cases were excluded from model training and set aside as the prediction target group. Ultimately, approximately 5.33 million event records with valid age group labels were used for training, while about 330,000 unlabeled events were reserved for testing as the final prediction targets. Details regarding dataset partitioning and model training are presented in Section 4.

3.2 Feature Engineering and Selection

Based on the integrated dataset obtained after preprocessing, a variety of derived variables were generated and meaningful features were selected to enhance predictive performance and better capture user characteristics.

Temporal features were included as described in Section 3.1. Specifically, the hour of day was extracted from the event timestamp under the assumption that user activity times may vary across age groups. In addition, a time-based weight was created from event dates to assign greater importance to more recent events. The weighting scheme was analyzed through log-scale distributions and cumulative plots, and it was designed to ensure that recent behaviors would be more strongly reflected during model training.

Event-related features were engineered by assigning importance-based weights to event types. Weights from 1 to 4 were applied to click_item, like_item, add_to_cart, and purchase_success, respectively, so that purchase-related behaviors would be considered more meaningful signals than simple clicks. In addition, one-hot encoding was applied to capture the occurrence of each event type, enabling the model to learn the contribution of individual event categories.

Regional features were derived by aggregating event frequencies based on users' access regions (region). These frequencies were then normalized by regional population figures (sourced from Statistics Korea) to produce regional weights. This adjustment allowed active engagement in less populated regions not to be underestimated. The event distributions before and after normalization were visualized to verify the effectiveness of bias correction. The derived features described above were constructed by compressing or transforming information from the original data and incorporated into the model input feature set.

For categorical variable encoding, different strategies were applied depending on the number of unique values. Gender and event type were processed using one-hot encoding, while category

and brand identifiers were retained as numerical codes but treated as nominal attributes. For continuous variables, such as price, min–max normalization was applied to ensure compatibility with distance-based models such as KNN. From a feature selection perspective, attributes that did not contribute to prediction or posed a risk of model leakage (e.g., identifiers such as IDs, duplicated textual fields such as product names) were removed.

Drawing on feature engineering methodologies from similar large-scale log analyses [13, 42, 45], which emphasized the predictive power of temporal and interaction frequency patterns, the final feature set was designed to comprehensively integrate user characteristics (gender, age group), product characteristics (price, category, brand), behavioral characteristics (event type, hour, device, platform), and weights (time-based, event-based, region-based). These features were also used in subsequent feature importance analyses of the trained models, which enabled quantitative interpretation of the most influential predictors. This facilitated the elimination of redundant variables, the identification of directions for model improvement, and the extraction of business insights.

3.3 Model Selection and Configuration

Consistent with prior empirical studies in e-commerce user profiling [12, 13], this study adopts a comparative experimental design to evaluate the efficacy of supervised learning algorithms. The age group prediction task is formulated as a multi-class classification problem. To ensure a balanced evaluation of interpretability, performance, and diversity in modeling approaches, this study employed five representative supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, KNN, and XGBoost.

- Logistic Regression was included as a classical baseline. While relatively simple, it provides probabilistic interpretation of class membership and serves as a benchmark against which more complex models can be evaluated.
- Decision Tree offers an interpretable rule-based structure and fast training speed, making it suitable as a baseline non-linear model. However, single trees are prone to overfitting, which motivated the inclusion of ensemble-based extensions.
- Random Forest extends decision trees through bagging, mitigating overfitting and improving generalization. It also provides robust performance across age groups in preliminary experiments and allows straightforward extraction of feature importance for interpretability.
- KNN was selected as a non-parametric, distance-based method that does not assume specific data distributions. While computationally intensive on large datasets and sensitive to feature scaling, it provides modeling diversity by relying on similarity-based classification rather than explicit model learning.
- XGBoost represents the boosting paradigm, capable of achieving high accuracy and computational efficiency even with large-scale data. Its regularization mechanisms prevent overfitting, and GPU acceleration enables fast training. Although its initial performance was slightly lower than Random Forest, tuning showed potential for improvement, making it a strong candidate model.

These five models were chosen to cover a broad spectrum of learning paradigms, from regression-based baselines to ensemble methods and instance-based approaches. Such diversity enables a comprehensive comparison of performance and provides objective evidence for selecting the most suitable model for this study. To address class imbalance in the dataset, the class_weight='balanced' parameter was applied to Logistic Regression, Decision Tree, and Random Forest, ensuring adequate weight was assigned to underrepresented age groups.

4. MODEL TRAINING AND PERFORMANCE EVAUATION

4.1 Data Splitting and Evaluation Metrics

For model training and the assessment of generalization performance, the dataset was divided into training, validation, and test sets. Approximately 330,000 event records without age group labels were separated as the prediction target (test set), while the remaining 5.33 million labeled events were used for model training. The training data was further split as follows:

- X_train_basic (≈ 70%): the main training set used for parameter learning.
- X_val (≈ 15%): the validation set used to periodically evaluate model performance during training and to prevent overfitting.
- X_hyper (≈ 15%): the set dedicated to hyperparameter search and optimization.

This data splitting strategy follows commonly adopted ratios in machine learning (training 60–80%, validation 10–20%, test 10–20%), ensuring both sufficient data for model fitting and reliable evaluation. The explicit separation of X_val

and X_hyper allowed for objective assessment by preventing overfitting while enabling robust hyperparameter optimization. After training was completed, the final predictions were performed on the 330,000 unlabeled test samples (X_test), which were strictly excluded from the training and tuning process. This ensured that the final evaluation reflected the true generalization performance of the model.

The predictive performance of the models was primarily evaluated using Accuracy and the F1-score. Accuracy represents the proportion of correctly classified samples out of all predictions and provides an intuitive measure of classification performance. However, in multi-class classification with imbalanced class distributions, Accuracy alone may overlook biased predictions toward majority classes. To address this limitation, additional metrics such as Precision, Recall, and their harmonic mean, the F1-score, were measured. Since this study deals with a multi-class classification problem, the Macro-F1 score, which averages the F1-scores of all classes regardless of their frequencies, was adopted as a supplementary evaluation metric. Macro-F1 provides a fairer assessment of model performance under class imbalance. In addition, the training time of each model was compared to evaluate efficiency from a practical deployment perspective.

4.2 Model Performance Comparison and Analysis

Based on the integrated log dataset after preprocessing and feature engineering, five machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, KNN, and XGBoost—were applied to train the age group prediction models. Each model was trained under the same data splitting strategy and feature configuration to ensure fairness in comparison. Performance evaluation was conducted using the metrics defined earlier, namely Accuracy, Precision, Recall, and F1-score.

Table 1: Performance Comparison of Machine Learning Models for Age Group Prediction.

Model	Accuracy	Precision	Recall	F1-score	Time(sec)
Logistic Regression	0.2024	0.3749	0.2024	0.2538	1273.12
Decision Tree	0.7464	0.7468	0.7464	0.7466	33.18
Random Forest	0.7832	0.7918	0.7832	0.7813	532.82
KNN	0.5859	0.5845	0.5859	0.5795	23.40
XGBoost	0.3415	0.4807	0.3415	0.3854	10.07

Table 1 summarizes the results, presenting each model's predictive performance in terms of accuracy, precision, recall, and F1-score, as well as the corresponding training time (in seconds). This comprehensive comparison not only highlights the relative predictive performance across models but also provides insights into computational efficiency, which is critical for practical deployment.

The experimental results indicate that the Random Forest model achieved the best overall predictive performance, with an accuracy of 0.7832 and an F1-score of 0.7813. Notably, it maintained relatively balanced precision and recall even under class imbalance—a common challenge in multi-class classification problems—demonstrating its practical applicability. Furthermore, the ability to interpret feature importance inherent in its tree-based structure provides a valuable advantage for real-world analysis. The Decision Tree model, despite its relatively simple structure, delivered solid results with an accuracy of 0.7464 and an F1-score of 0.7466. With a short training time of approximately 33 seconds, it is suitable for exploratory analysis and rule-based system design. However, the vulnerability of single trees to overfitting requires caution in practical deployment. The KNN model achieved an accuracy of 0.5859 and an F1-score of 0.5795, reflecting moderate performance. In this study's high-dimensional, multi-class setting, the limitations of distance-based classifiers became apparent, and the computational burden for large-scale datasets must also be considered. Nevertheless, its relatively short training time of about 23 seconds highlights a practical advantage of KNN. The XGBoost model recorded the shortest training time of 10 seconds, but its accuracy (0.3415) and F1 score (0.3854) were comparatively low. This is attributed to insufficient hyperparameter optimization in the current experiment, suggesting that XGBoost remains a promising candidate with substantial potential for improvement through further tuning. Finally, Logistic Regression exhibited the weakest performance, with an accuracy of 0.2024 and an F1-score of 0.2538, while requiring the longest training time (1,273 seconds). This outcome can be explained

by the inability of linear models to capture the non-linear relationships inherent in multi-class classification tasks involving user behaviors and product attributes. Taken together, these findings suggest that Random Forest is the most suitable model for the age-group prediction task based on the integrated log dataset. Decision Tree also offers meaningful value given its interpretability and computational efficiency. KNN and XGBoost serve as useful comparative baselines due to their distinct algorithmic approaches, while Logistic Regression is deemed unsuitable for this type of problem.

To facilitate a more intuitive understanding of the comparative analysis, the visualization results of model performance are presented in Figures 10–12. Figure 10 illustrates both the validation accuracy and training time of each model. The left panel compares the validation accuracy across models, while the right panel shows the corresponding training time (in seconds). The Random Forest model achieved the highest validation accuracy, followed by Decision Tree and KNN. In contrast, Logistic Regression demonstrated no meaningful performance, and XGBoost also exhibited relatively low accuracy compared to other models. In terms of training time, XGBoost completed training the fastest, with KNN and Decision Tree also requiring only minimal computation time. Conversely, Logistic Regression consumed the longest training duration, while Random Forest, despite its superior accuracy, required comparatively more training time due to the computational overhead of its ensemble structure.

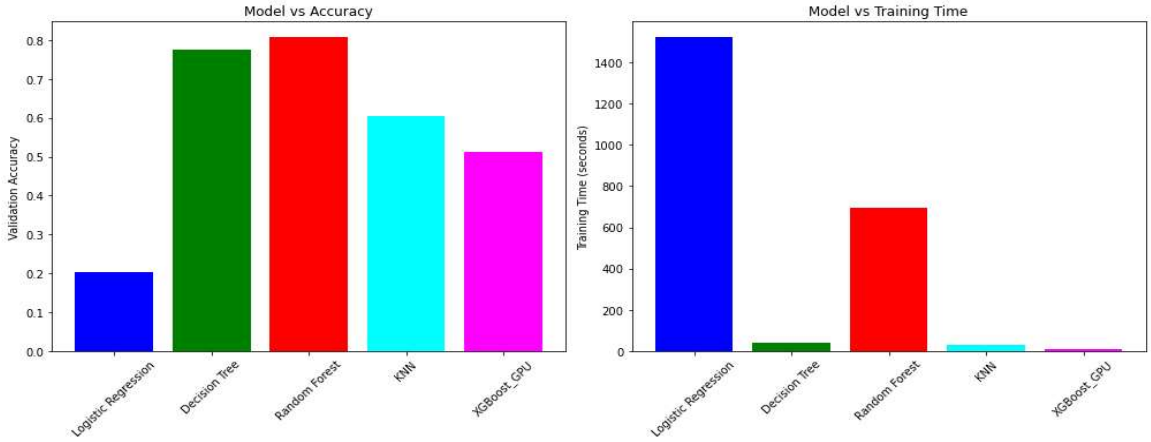


Figure 10: Model Performance Comparison

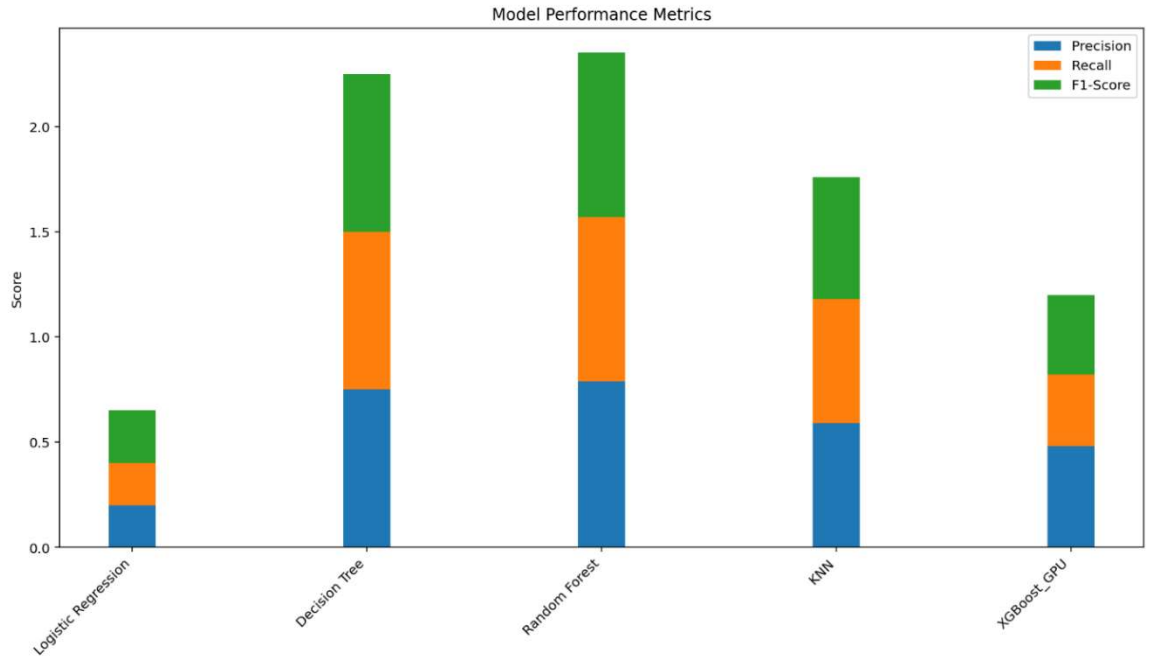


Figure 11: Stacked Visualization of Model Performance Metrics

Figure 11 presents the stacked bar visualization of Precision, Recall, and F1-score for each model. Overall, Random Forest and Decision Tree demonstrated balanced performance across all three-evaluation metrics, indicating both precision and coverage in prediction. In particular, Random Forest achieved the highest values in all three metrics, proving its robustness and predictive power even in multi-class and imbalanced data environments. Decision Tree, despite its simple structure, also showed consistently high scores in Precision, Recall, and F1-score, highlighting its potential as a practical and interpretable alternative model. KNN recorded mid-range values for all metrics, with relatively stable performance in Recall. However, due to the nature of distance-based algorithms, scalability and computational complexity issues may arise as the dataset grows larger. XGBoost showed low performance, with Precision and Recall around 0.34, and a slightly higher F1-score of 0.38. This suggests that in its default configuration, the model struggled with the complexity of the multi-class classification task; however, further hyperparameter tuning is expected to improve performance. Finally, Logistic Regression performed the worst across all metrics, with both Precision and Recall remaining at 0.20, indicating performance close to random guessing. This result reflects the structural limitation of linear models in capturing nonlinear behavioral patterns inherent in user interaction data.

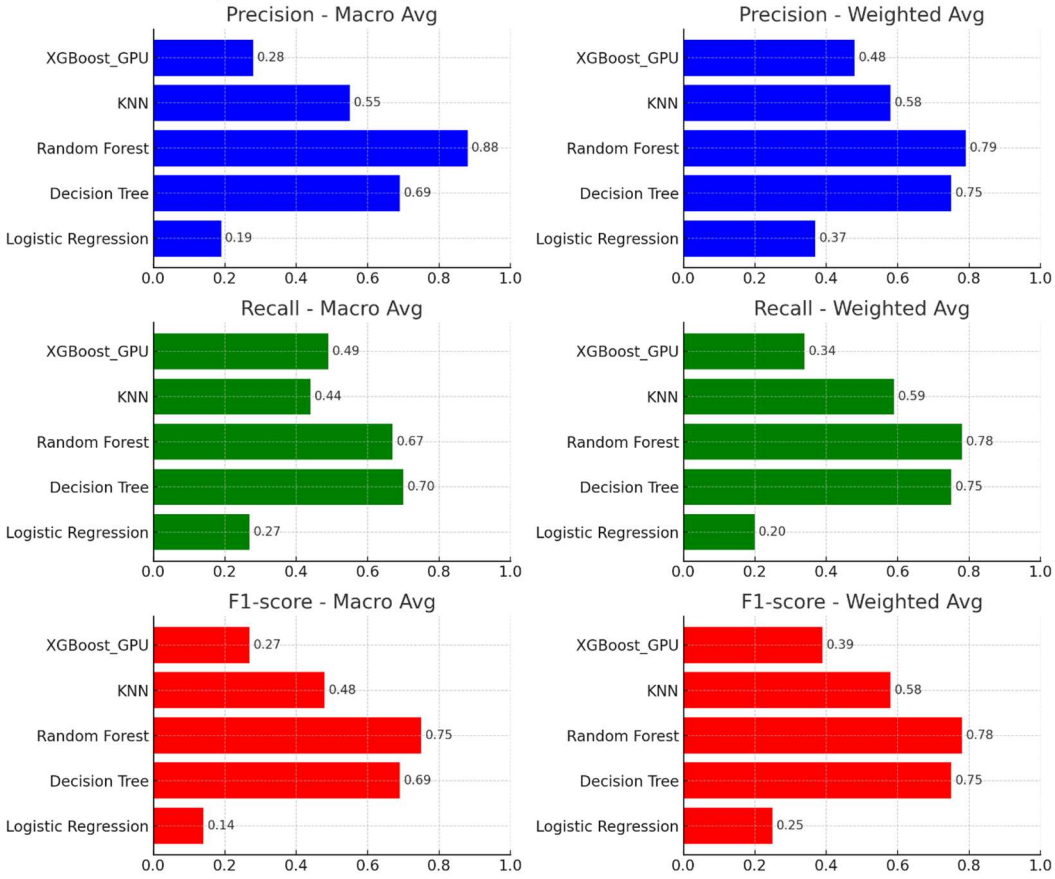


Figure 12: Comparison of Macro and Weighted Averages across Models

Figure 12 compares the macro and weighted averages of Precision, Recall, and F1-scores for each model. The macro average calculates the simple mean of class-specific metrics, minimizing the influence of class imbalance by treating all classes with equal importance. In contrast, the weighted average incorporates the sample size of each class as weights, thereby reflecting the differences in class distribution. Across all three metrics, Random Forest achieved the highest performance in both macro and weighted averages. Its F1-scores reached 0.75 (macro) and 0.78

(weighted), confirming its ability to deliver stable predictive power even under imbalanced class distributions. Decision Tree recorded the second-best performance across all metrics, with its weighted averages showing results close to Random Forest. On the other hand, XGBoost showed relatively poor performance, with both Precision and Recall at low levels and an F1-score of only 0.27. This suggests that the model struggled to effectively address the multi-class classification problem, likely due to overfitting on certain classes or structural vulnerability to class imbalance. KNN maintained overall moderate performance; however, under the macro average, both Precision and Recall remained at a mid-level, indicating potential degradation in highly imbalanced environments. Logistic Regression remained at the lowest performance level, with both macro and weighted averages generally between 0.14 and 0.37. This result reflects the model's inability to sufficiently capture the diversity and complexity of patterns across classes. These average-based comparison results are useful for understanding how consistently each model performs across the entire dataset and can serve as an important reference metric when seeking to prevent predictions biased toward specific classes in practical applications. Taken together, the findings confirm that the most suitable model for the characteristics of the e-commerce data and the classification objectives in this study is the Random Forest. This model not only provides high predictive performance but also demonstrates robustness against real-world data issues such as class imbalance. Therefore, it represents a highly practical alternative, particularly for applications such as user attribute prediction or personalized recommendations. However, its relatively long training time should be taken into consideration in environments requiring real-time analysis, where a trade-off between performance and speed may be necessary. Meanwhile, beyond the overall model performance comparison, analyzing how predictive power varies across different age groups is essential for suggesting more refined applications. Accordingly, the next section provides an in-depth analysis of performance differences by age group.

4.3 Comparative Analysis of Predictive Performance by Age Group

Figure 13 presents a comparative analysis of precision, recall, and F1-score across six age groups, ranging from under 10 to over 60, for the five models. Random Forest achieved the highest F1-score across all age groups, which can be attributed to the ensemble method's ability to combine multiple decision trees and thereby enhance predictive stability and accuracy. Decision Tree also demonstrated relatively strong performance, although consistently lower than Random Forest. Logistic Regression showed low values in precision, recall, and F1-score for all age groups, highlighting its limitations in capturing complex patterns. KNN maintained moderate performance overall, with scores comparable to Decision Tree in certain age groups. XGBoost exhibited relatively higher precision in specific age groups (e.g., individuals in their 20s and 30s), but very low recall, reflecting a conservative prediction tendency that identifies positives correctly but misses many actual positive samples.

Figure 14 visualizes the age group specific F1-scores of the five models using a heatmap. Random Forest and Decision Tree consistently maintained high values across all age groups, while Logistic Regression recorded the lowest scores in every age group. For example, in the 20s, Random Forest (0.82) and Decision Tree (0.79) outperformed Logistic Regression (0.30) by approximately 0.5 points. Similarly, in the 60+ age group, Random Forest and Decision Tree achieved strong performances of 0.79 and 0.73, respectively, whereas Logistic Regression yielded only 0.04. The F1-scores of KNN and XGBoost generally fell between those of Random Forest/Decision Tree and Logistic Regression, but showed greater fluctuations across age groups, indicating that their performance was relatively strong only within certain cohorts. Specifically, KNN peaked at 0.64 in the 20s but dropped to 0.42 in the 50s, while XGBoost reached its highest score of 0.58 in the 20s, then sharply declined to 0.24–0.27 in the 40s–50s before partially recovering to 0.45 in the 60+ group. Based on this analysis, the next subsection explores the hyperparameter optimization process and results conducted to further enhance the predictive performance of each model.

Journal of Theoretical and Applied Information Technology
28th February 2026. Vol.104. No.4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

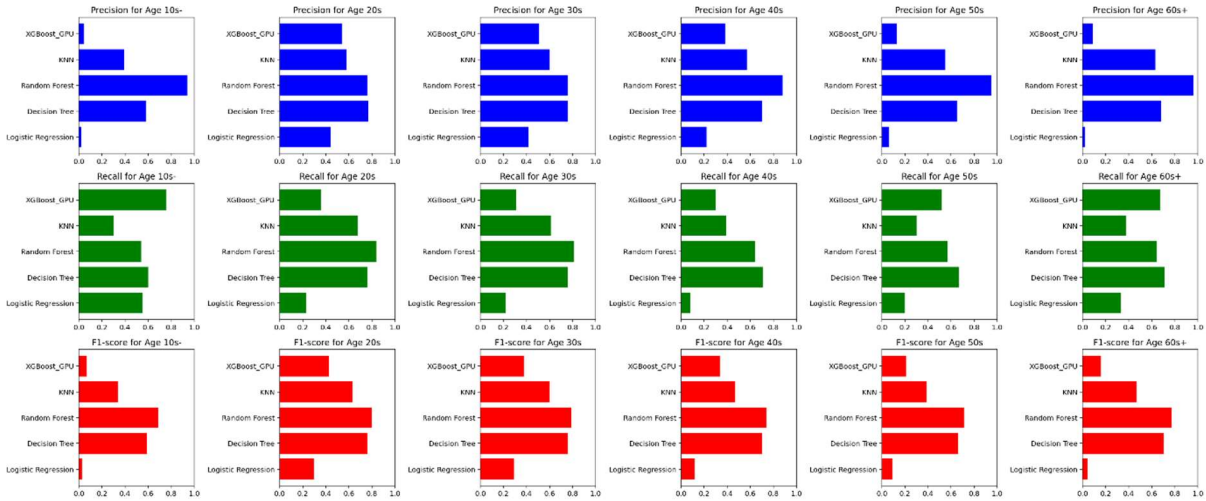


Figure 13: Comparative Visualization of Precision, Recall, and F1-score across Six Age Groups for Five Models

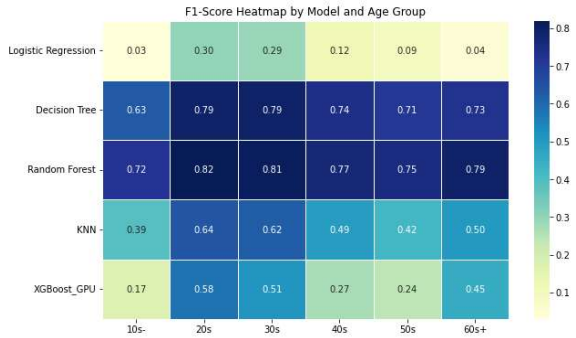


Figure 14: Heatmap of Age Group-Specific F1-score

4.4 Hyperparameter Optimization

Based on the performance comparison of the five models presented in Section 4.4, this study selected four models—Random Forest, Decision Tree, XGBoost, and KNN—as the primary candidates for hyperparameter optimization. Random Forest and Decision Tree consistently achieved high F1-scores across all age groups. In particular, Random Forest was deemed highly applicable in practice due to its feature importance interpretability and strong generalization capability. Decision Tree was also included as an optimization target because of its relatively short training time and high interpretability. Although KNN exhibited relatively lower overall performance, it was selected to ensure model diversity, given its non-parametric nature and distinctive distance-based prediction mechanism that requires no distributional assumptions. XGBoost, despite showing very low F1-scores in certain age groups, was also considered due to its fast-training speed leveraging GPU acceleration and the scalability of boosting techniques. In contrast, Logistic Regression was excluded, as it consistently demonstrated poor performance across all age groups. To improve the generalization performance of the selected four models, key hyperparameter ranges were defined for each model, and both GridSearchCV and RandomizedSearchCV were employed to identify optimal configurations. The primary hyperparameters and search ranges were as follows:

- Decision Tree: max_depth (3, 5, 7, 10, 15), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), criterion (gini, entropy), splitter (best, random)
- Random Forest: n_estimators (50, 100, 150, 200), max_features ('auto', 'sqrt'), max_depth (3, 5, 7, 10, None), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), criterion (gini, entropy), bootstrap (True, False)
- XGBoost: learning_rate (0.01, 0.05, 0.1, 0.3), n_estimators (50, 100, 150, 200, ...), max_depth (3, 5, 7, 10, 15), subsample (0.6, 0.7, 0.8, 0.9, 1.0), colsample_bytree (0.6 … 1.0), gamma, min_child_weight, objective
- KNN: n_neighbors (3, 5, 7, 9, 11, ...), weights ('uniform', 'distance'), metric ('euclidean', 'manhattan', 'minkowski')

The search process was primarily conducted using GridSearchCV; however, when the parameter space was large, RandomizedSearchCV was adopted, sometimes in combination with data sampling, to ensure computational efficiency. RandomizedSearchCV is particularly effective for large-scale searches as it evaluates random parameter combinations within a limited number of iterations. The optimal parameters derived through

this process, along with the resulting performance improvements, are presented in detail in the subsequent subsections.

4.4.1 Decision Tree Model Tuning Results

The hyperparameter search for the Decision Tree model was restricted to a maximum depth of 15 (excluding None) and was performed sequentially using GridSearchCV and RandomizedSearchCV. The GridSearchCV (60 parameter combinations, 5-fold CV) identified the optimal parameters as criterion = 'entropy', max_depth = 15, and splitter = 'best', yielding a cross-validation accuracy of approximately 0.49. This was significantly lower than the baseline model accuracy (approximately 0.746), suggesting that an entropy-based tree with limited depth suffered from reduced generalization due to overfitting. Subsequently, RandomizedSearchCV (100 iterations) selected splitter = 'random', min_samples_split = 7, min_samples_leaf = 2, max_depth = None, and criterion = 'entropy' as the optimal configuration. This combination produced a slightly higher cross-validation accuracy of 0.57, but it still did not surpass the baseline model. Notably, the unrestricted tree growth (max_depth = None) combined with random splits likely led to a model overly specialized to the training data. When retrained on the full training set and evaluated on the validation set, the tuned model did not outperform the baseline and, in fact, showed marginally reduced performance. Ultimately, the default configuration (e.g., criterion = 'gini', max_depth = None) proved to be more stable than the tuned alternatives.

4.4.2 Random Forest Model Tuning Results

Due to the large parameter space of Random Forest, a full grid search led to memory errors. Therefore, RandomizedSearchCV (100 iterations) was first conducted, followed by partial grid searches and data subsampling for refinement. RandomizedSearchCV identified the optimal configuration as n_estimators = 90, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, criterion = 'entropy', and bootstrap = False, achieving a cross-validation accuracy of approximately 0.64. This indicates that a deep forest using the entire dataset without bootstrapping, combined with entropy-based splitting, was most effective for this dataset. Subsequent grid searches with reduced parameter ranges (e.g., max_depth = {10, 20}, criterion = 'gini', bootstrap = True) yielded an alternative optimal configuration: max_depth = 20, n_estimators = 90, min_samples_split = 2, min_samples_leaf = 1, criterion = 'gini', bootstrap = True, with an accuracy of 0.55. However, this differed from the RandomizedSearchCV outcome in terms of split criterion and bootstrap usage. To further examine data sensitivity, grid searches were performed on subsampled training sets (30%, 50%, 70%, 80%). The results generally aligned with those from RandomizedSearchCV. Accuracy improved as the sample size increased, with values of 0.54 (30%), 0.58 (50%), 0.61 (70%), 0.62 (80%), and 0.64 (100%). In summary, the optimal configuration was a deep, entropy-based forest without bootstrapping, yielding a validation accuracy of about 64%. This represented a modest improvement over the baseline model.

4.4.3 XGBoost Model Tuning Results

Given the extensive hyperparameter space of XGBoost, a complete grid search was impractical. Instead, RandomizedSearchCV (50 iterations) was employed. The optimal parameters were learning_rate = 0.1, max_depth = 15, n_estimators = 350, subsample = 0.6, colsample_bytree = 1.0, gamma = 0.2, min_child_weight = 3, and objective = 'binary:logistic'. This configuration yielded a cross-validation accuracy of approximately 0.65, the highest among all tuned models. Interpretation of the optimal configuration suggests that while a relatively deep and large number of trees were employed (max_depth = 15, n_estimators = 350), overfitting was mitigated through subsampling (subsample = 0.6) and regularization (gamma = 0.2, min_child_weight = 3). The learning rate of 0.1 further stabilized the training process. An additional grid search using 30% of the data was attempted, but the process was terminated after more than two days due to the excessive number of parameter combinations. Consequently, narrowing the parameter space based on RandomizedSearchCV results was deemed a more efficient strategy. The final tuned XGBoost model achieved a validation accuracy of approximately 65%, representing a slight improvement over the baseline.

4.4.4 KNN Model Tuning Results

The KNN model involved relatively few hyperparameters, allowing both GridSearchCV and RandomizedSearchCV to be applied effectively. Both methods produced the same optimal configuration: n_neighbors = 36, weights = 'distance', and metric = 'manhattan'. This combination achieved the highest cross-validation accuracy of approximately 0.57, an improvement over the baseline KNN model (n_neighbors = 5, weights = 'uniform'). The performance gain suggests that referencing a larger number of neighbors while weighting closer neighbors more heavily provided better generalization than relying on fewer neighbors

with uniform weights. Furthermore, the adoption of the Manhattan distance metric indicates that the L1 distance was more effective than the L2 (Euclidean) distance in this feature space. Although the absolute accuracy of KNN remained lower than that of other models (with a maximum of about 57%), hyperparameter tuning successfully improved its performance relative to the baseline.

4.5 Comparison of Model Performance on the Validation Dataset

In this subsection, the classification performance of the four models was compared on the validation dataset. Each model was evaluated under two conditions: with default hyperparameter settings and with optimized settings. The evaluation focused primarily on accuracy and F1-score. As shown in Table 2, Random Forest (default configuration) consistently achieved the highest performance overall, while the effectiveness of hyperparameter optimization varied considerably across models.

The Decision Tree model achieved 74.53% accuracy and 74.56% F1-score under default settings, which were substantially higher than those under optimized settings (58.37% accuracy, 58.03% F1-score). This indicates that hyperparameter tuning degraded performance, likely because excessive restrictions on model complexity prevented the tree from adequately capturing data patterns. While the default model maintained a good balance between recall and precision, the optimized configuration introduced constraints that reduced learning efficiency. The Random Forest model outperformed all others, achieving the highest accuracy (78.54%) and F1-score (78.35%) with default settings. However, its optimized configuration resulted in lower performance (66.79% accuracy, 66.29% F1-score).

Table 2: Performance Comparison on the Validation Dataset with Default and Optimized Settings

Model	Accuracy		Precision		Recall		F1-score	
	Def.	Opt.	Def.	Opt.	Def.	Opt.	Def.	Opt.
Decision Tree	0.745	0.584	0.750	0.580	0.750	0.580	0.746	0.580
Random Forest	0.785	0.668	0.790	0.680	0.790	0.670	0.784	0.663
KNN	0.586	0.583	0.580	0.590	0.590	0.580	0.579	0.574
XGBoost	0.510	0.665	0.540	0.680	0.510	0.660	0.486	0.660

Like the Decision Tree, the tuned model's reduced complexity or overfitting during the optimization process may have led to diminished generalization ability. The KNN model showed very little difference between default (58.59% accuracy, 57.95% F1-score) and optimized settings (58.26% accuracy, 57.44% F1-score). Overall, its performance remained relatively low, likely due to KNN's memory-based mechanism being less effective in capturing complex patterns within the dataset. The XGBoost model demonstrated the most dramatic improvement through hyperparameter optimization. In its default configuration, performance was the lowest of all models (50.98% accuracy, 48.61% F1-score). However, after optimization, its performance improved markedly to 66.50% accuracy and 65.97% F1-score. This suggests that the appropriate choice of learning rate, tree depth, and regularization parameters substantially enhanced its ability to model data patterns. Notably, the optimized XGBoost achieved significant improvements in detecting minority classes. For example, recall for the 10s age group increased from approximately 9% in the default model to 32% after optimization, which contributed to the overall improvement in the F1-score.

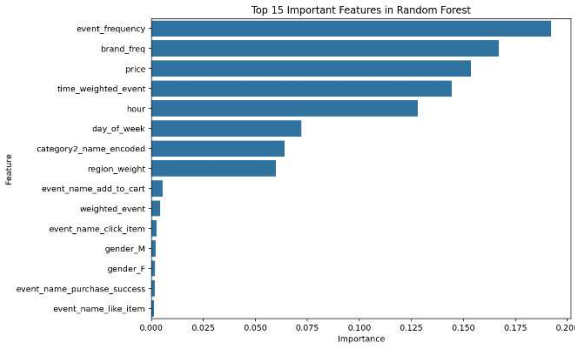
In summary, Random Forest with default settings remained the best-performing model overall, while XGBoost demonstrated the highest potential for improvement through tuning. By contrast, both Decision Tree and KNN yielded relatively low performance, with Decision Tree notably performing better in its default configuration than in its optimized version, highlighting the need for careful consideration when selecting hyperparameters. Furthermore, across all models, recall for minority classes remained comparatively low, underscoring the importance of incorporating class imbalance mitigation strategies such as resampling or class-weight adjustments in future work.

Critiquing our outcomes against the current state-of-the-art, although recent studies [45, 46, 47] utilizing Deep Learning architectures (e.g., RNN, CNN) have reported high accuracy, they often require substantial computational resources and large-scale labeled data. In contrast, our results demonstrate that the Random Forest model achieves a competitive accuracy of approximately 78% while

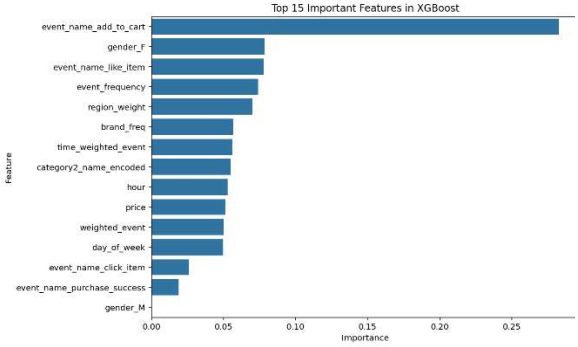
maintaining significantly lower training times (approx. 532 seconds) compared to typical deep learning baselines. This confirms that for tabular e-commerce data, ensemble-based machine learning offers a superior trade-off between performance and operational efficiency.

4.6 Feature Importance Analysis and Model Interpretation

In this subsection, we analyze feature importance for the tree-based models Random Forest and XGBoost to interpret the relative influence of different variables on age group prediction. Using the feature_importances_ attribute, the top 15 most influential features were extracted from the full set of input variables.



(a) Random Forest Model



(b) XGBoost Model

Figure 15: Top 15 Important Features

As shown in Figure 15(a), the Random Forest results highlight that variables related to behavioral frequency and product attributes—such as event_frequency, brand_freq, price, and time_weighted_event—exhibited high importance. This suggests that repeated behavioral patterns and product preferences play a critical role in predicting user age groups. In addition, time-related variables such as hour and day_of_week also demonstrated notable contributions, implying that preferred shopping times vary across age groups. In contrast, the XGBoost results in Figure 15(b) placed features such as event_name_add_ to_cart, gender_F, and event_name_like_item at the top. Notably, "add to cart" behavior emerged as the most important feature in this model, underscoring that purchase patterns and preferences differ significantly across age groups depending on user interactions. The high importance of the gender_F variable further indicates that gender remains an influential factor in predicting age categories. Although both models shared several key features (e.g., event_frequency, brand_freq, region_weight), the relative ranking and interpretation of these features differed. Random Forest, which aggregates multiple decision trees in parallel, evaluates overall feature contributions, while XGBoost assigns greater weight to variables that contribute most to reducing residual error through its sequential boosting process. Furthermore, Random Forest computes feature importance based on gini impurity reduction, whereas XGBoost relies on gradient-based changes in the loss function, leading to methodological differences in the importance calculation. These distinctions explain why the two models, even when applied to the same dataset, emphasize different variables. Importantly, such differences provide complementary perspectives for feature interpretation in age group classification, enriching the analytical insights that can be derived from model comparison. Taken together, these results provide a foundation for the conclusions and future research directions presented in the next chapter. In particular, the key factors identified through feature importance analysis serve as actionable evidence for targeted marketing, personalized recommendation, and customer segmentation, while the comparative performance results offer important guidance for model selection and operational strategies.

5. CONCLUSIONS

This study addressed the problem of missing age group information in large-scale e-commerce data by comparing and analyzing various machine learning models. Leveraging multidimensional variables—including customer behavioral logs, product attributes, and temporal and regional information—five models, including Logistic Regression, were initially evaluated. As Logistic Regression exhibited relatively low performance, it was excluded from further analysis, and four models—Decision Tree, Random Forest, KNN, and XGBoost—were subjected to in-depth hyperparameter optimization and performance evaluation.

The analysis demonstrated that the Random Forest model (default configuration) achieved the highest predictive accuracy at approximately 78%, while the XGBoost model, though initially weaker, improved to a comparable level through hyperparameter tuning. The primary scientific contribution of this work is twofold. First, it establishes empirical evidence that 'shallow' machine learning models, specifically Random Forest, can achieve competitive performance comparable to complex deep learning approaches for tabular e-commerce data, but with significantly lower computational overhead. Second, unlike previous studies that treat missing values as mere noise, this study presents a validated methodological framework that integrates gender imputation as a precursor to age prediction, thereby enhancing overall data integrity. Feature importance analysis further revealed that behavioral frequency, specific event types, gender, and product attributes were among the most influential factors. These results highlight the potential to enhance the precision of marketing targeting and personalized recommendation systems, as well as to establish early personalization strategies for customers with restricted or missing personal information.

However, this study is subject to certain limitations. The analysis was conducted on data from a single e-commerce platform, which raises the need for further validation to ensure generalizability across different platforms and domains. Additionally, the reliance on accumulated behavioral logs introduces a 'cold-start' problem, rendering the model less effective for new subscribers with sparse interaction history. In addition, the study did not directly implement or evaluate the integration of age group prediction results into an operational recommendation system.

Future research will focus on incorporating predicted age groups into recommendation systems to validate performance improvements and on conducting comparative analyses using time-series models and deep learning techniques. Furthermore, expanding the dataset to include multimodal information such as product images and review texts, as well as applying class imbalance mitigation strategies through data augmentation and weighting techniques, will be explored. From a managerial perspective, these findings provide actionable insights for e-commerce practitioners. In particular, the identification of key behavioral and product-related features offers concrete guidance for developing more accurate customer segmentation, optimizing targeted marketing campaigns, and enhancing personalization strategies. By enabling more precise profiling even for users with incomplete demographic data, firms can improve customer engagement, increase conversion rates, and ultimately strengthen their competitive advantage in highly dynamic online marketplaces.

ACKNOWLEDGEMENT

This paper was supported by the Research Fund, 2024, Pyeongtaek University in Korea.

REFERENCES:

[1] Lee, H., Wong, S.F., and Chang, Y. "Confirming the effect of demographic characteristics on information privacy concerns", *Proceedings of the 20th Pacific Asia Conference on Information Systems*, Chiayi, Taiwan, 27 June - 1 July, 2016.

[2] Benamati, J.H., Ozdemir, Z.D., and Smith, H.J. "An empirical test of an antecedents – privacy concerns – outcomes model", *Journal of Information Science*, Vol. 43, 2017, pp.583-600, doi:10.1177/0165551516653590.

[3] Bartol, J., Vehovar, V., Bosnjak, M., and Petrovčič, A. "Privacy concerns and self-efficacy in e-commerce: Testing an extended APCO model in a prototypical EU country", *Electronic Commerce Research and Applications*, Vol. 60(C), 2023, doi:10.1016/j.elerap.2023.1012.

[4] Analytify, "Why is your Google analytics demographic data missing?", *Analytify*, 16 May, 2023, Available: https://analytify.io/google-analytics-demographic-data-missing/ (accessed on 26 July 2025).

[5] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. "How old do you think I am? A study of language and age in twitter", *Proceedings of the 7th International AAAI Conference on Web and Social Media*, Boston, MA, USA, 8-11 July, Vol. 7, No. 1, 2013, pp.439-448, doi:10.1609/icwsm.v7i1.14381.

[6] Chamberlain, B.P., Humby, C., and Deisenroth, Marc. "Probabilistic inference of twitter users' age based on what they follow", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Skopje, Macedonia, 18-22 September, LNAI 10536, 2017, pp. 191-203. doi:10.48550/arXiv.1601.04621.

[7] O'Connor, K., Golder, S., Weissenbacher, D., Klein, A.Z., Magge, A., and Gonzalez-Hernandez, G. "Methods and annotated data sets used to predict the gender and age of twitter

users: scoping review", *Journal of Medical Internet Research*, 26:e47923, 2024, doi:10.2196/47923.

[8] Jian, H., Hua-Jun, Z., Hua, L., Cheng, N., and Zheng, C. "Demographic prediction based on user's browsing behavior", *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 8-12 May, 2007, pp. 151-160, doi:10.1145/1242572.1242594.

[9] Phuong, D.V. and Phuong, T.M. "Gender prediction using browsing history", *Proceedings of the Fifth International Conference on Knowledge and Systems Engineering*, Hanoi, Vietnam, 17-19 October, Vol. 244, 2013, pp. 271-283, doi:10.1007/978-3-319-02741-8_24.

[10] Ren, Y., Tomko, M., Salim, F.D., Chan, J., and Sanderson, M. "Understanding the predictability of user demographics from cyber-physical-social behaviors in indoor retail spaces", *EPJ Data Science,* Vol. 7, No. 1, 2018, doi:10.1140/epjds/s13688-017-0128-2.

[11] Hinds, J. and Joinson, A.N. "What demographic attributes do our digital footprints reveal? A systematic review", *PLoS ONE*, Vol. 13, No. 11: e0207112, 2018, doi:10.1371/journal.pone.0207112.

[12] Yehezkel, S. and Resheff, M.S. "Fusing multifaceted transaction data for user modeling and demographic prediction", arXiv:1712.07230, 2017, doi:10.48550/arXiv.1712.07230.

[13] Kooti, F., Lerman, K., Aiello, L., Grbovic, M., Djuric, N., and Radosavljevic, V. "Portrait of an online shopper: understanding and predicting consumer behavior", *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, San Francisco, California, USA, 22 - 25 February. 2016, pp. 205-214, doi:10.1145/2835776.2835831.

[14] Hendriksen, M., Kuiper, E., Nauts, P., Schelter, S., and de Rijke, M. "Analyzing and predicting purchase intent in e-commerce: anonymous vs. identified customers", *ACM SIGIR 2020 Workshop on eCommerce*, Virtual Event, China, 30 July, 2020, doi:10.48550/arXiv.2012.08777.

[15] Stojanovic, J., Gligorijevic, D., and Obradovic, Z. "Modeling customer engagement from partial observations", *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, Indianapolis, Indiana, USA, 24 - 28 October, 2016, pp. 1403-1412, doi:10.1145/2983323.29838.

[16] Jadhav, A., Pramod, D., and Ramanathan, K. "Comparison of performance of data imputation methods for numeric dataset", *Applied Artificial Intelligence}*, Vol. 33, No. 10, 2019, pp. 913-933, doi:10.1080/08839514.2019.1637138.

[17] Alwateer, M., Atlam, E., El-Raouf, M., Ghoneim, O., and Gad, I. "Missing data imputation: a comprehensive review", *Journal of Computer and Communications,* Vol. 12, 2024, pp. 53-75, doi:10.4236/jcc.2024.1211004.

[18] Aracri, F., Bianco, M.G., Quattrone, A., and Sarica, A. "Bridging the gap: missing data imputation methods and their effect on dementia classification performance", *Brain Science*, Vol. 15, No. 639, 2025, doi:/brainsci15060639.

[19] Azur, M.J., Stuart, E.A., Frangakis, C., and Leaf, P.J. "Multiple imputation by chained equations: what is it and how does it work?", *International Journal of Methods in Psychiatric Research*, Vol. 20, No. 1, 2011, pp. 40-49. doi: 10.1002/mpr.329.

[20] Austin, P.C., White, I.R., Lee, D.S., and Buuren, S. "Data in clinical research: a tutorial on multiple imputation", *Canadian Journal of Cardiology*, Vol. 37, No. 9, 2021, pp. 1322-1331, doi:10.1016/j.cjca.2020.11.010.

[21] Poulos, J. and Valle, R. "Missing data imputation for supervised learning", *Applied Artificial Intelligence*, Vol. 32, No. 2, 2018, pp. 186-196, doi:10.1080/08839514.2018.1448143.

[22] Faisal, s. and Tutz, G. "Nearest neighbor imputation for categorical data by weighting of attributes", *Information Sciences}*, Vol. 592(C), 2022, pp. 306-319, doi:10.1016/j.ins.2022.01.05.

[23] Memon, Shaheen M.Z., Wamala, R., and Kabano, I. H. "A comparison of imputation methods for categorical data", *Informatics in Medicine Unlocked*, Vol. 42, 101382, 2023, doi:10.1016/j.imu.2023.101382.

[24] Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., and Guo, H. "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets", *BMC Medical Research Methodology*, Vol. 24, No. 1:41, 2024, doi:10.1186/s12874-024-02173-x.

[25] Stekhoven, D.J. and Bühlmann, P. "MissForest--non-parametric missing value imputation for mixed-type data", *Bioinformatics*, Vol. 28, No. 1, 2012, pp. 112-118, doi:10.1093/bioinformatics/btr597.

[26] Tripet, A., Eustache, E., and Tillé, Y. "Improving donor imputation using the prediction power of random forests: a combination of SwissCheese and missForest", *Journal of Survey Statistics and Methodology*.

Vol. 12, No. 5, 2024, pp. 1389-1404, doi:10.1093/jssam/smad040.

[27] Yoon, J., Jordon, J., and Schaar, M. "GAIN: missing data imputation using generative adversarial nets", *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 10 - 15 July, Vol. 80, 2018, pp. 5689-5698.

[28] Dai, A., Bu, Z., and Long, Q. "Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems", *20th IEEE International Conference on Machine Learning and Applications*, Pasadena, California, USA, 13-16 December 2021, pp. 791-798, doi:10.1109/ICMLA52953.2021.00131.

[29] Dong, W., Fong, D.Y.T., Yoon, J., Wan, E.Y.F., Bedford, L.E., Tang, E.H.M., and Lam, C.L.K. "Generative adversarial networks for imputing missing data for big data clinical research", *BMC Medical Research Methodology*, Vol. 78, 2021, doi:10.1186/s12874-021-01272-3.

[30] Neves, D.T., Naik, M.G., and Proença, "A. SGAIN, WSGAIN-CP and WSGAIN-GP: novel GAN methods for missing data imputation", *Proceedings of the 21st International Conference on Computational Science*, Krakow, Poland, 16–18 June, 2021, Part I, pp. 98-113, doi:10.1007/978-3-030-77961-0_10.

[31] Shahbazian, R. and Trubitsyna, I. "DEGAIN: generative-adversarial-network-based missing data imputation", *Information*, Vol. 13, No. 12, 575, 2022, doi:10.3390/info13120575.

[32] Lim, J., An, S., Woo, G., Kim, C., and Jeon, J.J. "DrIM: context-driven missing data imputation via large language models", *submitted to International Conference on Learning Representations*, 2025. Available: https://openreview.net/forum?id=b2oLgk5XRE.

[33] Hayat, A. and Hasan, M.R. "A context-aware approach for enhancing data imputation with pre-trained language models", *Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, UAE, 19-24 January, 2025, pp. 5668-5685. Available: https://aclanthology.org/2025.coling-main.380.

[34] Wang, J., Wang, K., Zhang, Y., Zhang, W., Xu, X., and Lin, X. "On LLM-enhanced mixed-type data imputation with high-order message passing", *arXiv preprint*, arXiv:2501.02191, Jan. 2025, Available: https://arxiv.org/abs/2501.02191.

[35] Friedman, L. A., Lavee, G., Shapira, B., and Shmaryahu, D. "Data completion in e-commerce. Advances in Database Technology", *28th International Conference on Extending Database Technology*, Barcelona, Spain, 25-28 March, Vol. 28, No. 3, 2025, pp. 1048-1056, doi:10.48786/edbt.2025.88.

[36] Murray, D. and Durrell, K. "Inferring demographic attributes of anonymous internet users", *International Workshop on Web Usage Analysis and User Profiling*, San Diego, California, USA, 15 August, 1999, LNAI 1836, pp. 7-20, doi:10.1007/3-540-44934-5_1

[37] Hu, J., Zeng, H.J., Li, H., Niu, C., and Chen, Z. "Demographic prediction based on user's browsing behavior", *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 8-12 May, 2007, pp. 151-160, doi:10.1145/1242572.1242594.

[38] Jones, R., Kumar, R., Pang, B., and Tomkins, A. "I know what you did last summer: query logs and user privacy", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, Portugal, 6-10 November, 2007, pp. 909-914, doi:10.1145/1321440.1321573.

[39] Bi, B., Kosinski, M., Shokouhi, M., and Graepel, T. "Inferring the demographics of search users social data meets search queries", *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 13-17 May, 2013, pp. 131-140, doi:10.1145/2488388.2488401.

[40] Kosinski, M., Stillwell, D., and Graepel, T. "Private traits and attributes are predictable from digital records of human behavior", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 110, No. 15, 2013, pp. 5802-5805, doi:10.1073/pnas.1218772110.

[41] Ren, Y., Tomko, M., Salim, F.D., and Chan, J. "Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces", *EPJ Data Science*, Vol. 7, No. 1, 2018, doi:10.1140/epjds/s13688-017-0128-2.

[42] Wang, P., Guo, J., Lan, Y., Xu, J., and Cheng, X. "Your cart tells you: inferring demographic attributes from purchase data", *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, San Francisco, California, USA, 22-25 February, 2016, pp. 173-182, doi:10.1145/2835776.2835783.

[43] Zhong, E., Tan, B., Mo, K., and Yang, Q. "User demographics prediction based on mobile data", *Pervasive and Mobile Computing*, Vol. 9, No. 6,

2013, pp. 823-837, doi:10.1016/j.pmcj.2013.07.009.

[44] Felbo, B., Sundsøy, P., Pentland, A., Lehmann, S., and Montjoye, YA. "Modeling the temporal nature of human behavior for demographics prediction", *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases*, Skopje, Macedonia, 18-22 September, 2017, LNAI 10536, pp. 140-152, doi:10.1007/978-3-319-71273-4_12.

[45] Jiang, Y., Tang, W., Gao, N., Xiang, J., Zha, D., and Li, X. "Your pedometer tells you: attribute inference via daily walking step count", *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, Leicester, UK, 19-23 August, 2019, pp. 834-842. doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00173.

[46] Ruhan, R.J., Wahid, T., Rahman, A., Leshob, A., and Rab, R. "Using wearable sensors for sex classification and age estimation from walking patterns", *Sensors*, Vol. 25, 3509, 2025, doi:10.3390/s25113509.

[47] Roy, S., Nakisa, B., Pathirana, P.N., and Dazeley, R. "A wearable multi-sensor fusion approach for gender recognition based on deep learning", *Proceedings of the 10th International Conference on Bioinformatics Research and Applications*, Barcelona, Spain, 22-24 September, 2023, pp. 114-119, doi:10.1145/3632047.3632065.