Leveraging Machine Learning to Analyze User Conversion in Mobile Pharmacy Apps Using Behavioral and Demographic Data

Sri Lestari[1,*], Kiki Setiawan[2], Raisah Fajri Aula[3]

[1,2,3]*Information Systems, School of Computer Science, Cipta Karya Informatika Institute of Technology, Jakarta, Indonesia*

(Received: June 10, 2024; Revised: July 31, 2024; Accepted: August 2, 2024; Available online: September 1, 2024)

Abstract

This study explores the use of machine learning techniques to predict user conversion in a mobile pharmacy app based on user behavior and demographic data. The analysis was conducted using two classification models: Logistic Regression and Random Forest. Key features such as time spent on the product page, adding items to the cart, and user demographics (age, gender, device type) were evaluated to determine their impact on conversion rates. Both models demonstrated strong performance, with the Logistic Regression model achieving an Area Under the Curve (AUC) of 0.88 and the Random Forest model achieving an AUC of 0.87. These results indicate that both models effectively distinguish between users who convert and those who do not, with Logistic Regression showing a slightly better overall performance. Feature importance analysis revealed that factors such as adding items to the cart and the time spent on the product page are the most significant predictors of conversion. Furthermore, demographic features like age group and device type also contributed to the model's predictive power, although they had a smaller impact compared to user engagement features. The findings suggest that machine learning models, particularly Logistic Regression, can be utilized to predict user behavior and optimize user engagement strategies in mobile apps. The study also highlights the importance of user engagement in driving conversions and the potential for targeted marketing based on demographic data. Future work should focus on hyperparameter tuning, exploring additional algorithms, and incorporating real-time data to further enhance model accuracy and adaptability.

Keywords: Machine Learning, User Conversion, Mobile Pharmacy App, Logistic Regression, Random Forest

1. Introduction

The emergence of mobile pharmacy applications signifies a transformative phase in the healthcare industry, enhancing the accessibility, efficiency, and quality of pharmaceutical services. These applications serve multiple functions, from improving medication adherence to providing drug information, ultimately contributing to patient well-being and health outcomes.

One of the primary functions of mobile pharmacy apps is to enhance medication adherence among patients. A systematic review indicated that many mobile applications deliver crucial medication reminders that foster adherence, especially in patients with chronic conditions like cardiovascular diseases [1]. This capability is particularly vital since poor medication adherence can significantly hamper treatment effectiveness. Furthermore, such apps have features that can save time for both patients and healthcare providers, thereby streamlining the treatment process [2].

Mobile pharmacy applications also facilitate access to vital drug information. Studies have illustrated that community pharmacists use mobile apps to consult medical databases, such as Lexicomp and Epocrates, to provide the most current and accurate drug information to clients [3]. The user-friendly interface and up-to-date content offered by these applications have been acknowledged as instrumental in the daily operations of pharmacy practice, enhancing pharmacists' abilities to make informed decisions [4].

The significance of mobile pharmacy apps further extends to patient engagement and satisfaction. The introduction of geolocation services, for instance, allows users to find nearby pharmacies and receive real-time updates on pharmacy hours [5]. Such features not only improve customer satisfaction but also foster loyalty by integrating services more closely into patients' daily lives. This aspect of technology in pharmacy practice also relates to the education and

***Corresponding author: Sri Lestari (srilestari@stikomcki.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v4i3.86

training of healthcare professionals, equipping them with necessary digital health competencies to effectively use such applications in their future careers [2].

It is equally important to assess the quality and usability of these mobile apps, as the user experience directly impacts their effectiveness [6]. Research conducted on various pharmacy applications in Pakistan has highlighted the need for quality evaluation frameworks, leading to more trustworthy applications that better serve the healthcare community. User perceptions regarding usability also indicate that apps like "My Medicine" effectively bridge the information gap for patients, particularly in under-resourced areas, showcasing their role in enhancing healthcare accessibility [7].

User conversion significantly impacts the long-term success of healthcare applications. Evidence suggests that mHealth apps are increasingly being adopted for chronic disease management, such as hypertension, with a usage rate exceeding 70% among hypertensive patients in certain regions, particularly noted in China [8]. This growing acceptance emphasizes not only the need for user conversion strategies but also the readiness of patients to adopt digital solutions for ongoing management of their health conditions.

Moreover, the relationship between trust and user conversion cannot be overstated. Studies indicate that a patient’s trust in their healthcare provider directly influences their willingness to engage with mobile health applications introduced by those providers [9]. Different user demographics also exhibit varying preferences for accessing health information, further complicating strategies for effective conversion and engagement. Understanding these nuances is crucial for healthcare providers aiming to foster relationships and drive usability in their mobile health solutions [10].

The challenge of increasing user conversion for a specific product in a mobile pharmacy app is critical, as a high conversion rate directly influences the product's sales and overall success. Despite the presence of the product and its visibility on the app, a significant portion of users are not making the purchase. This low conversion rate may stem from various factors, including ineffective packaging, suboptimal user experience, or the mismatch between user expectations and the product offering. Understanding why users engage with the product page but fail to convert is essential for improving the app’s sales performance. The research objective is to explore how user behavior and demographic factors influence the likelihood of conversion. Specifically, the goal is to analyze user interactions with the app, such as time spent on the page, previous purchase history, and the type of device used, in relation to their demographics like age and gender. By applying machine learning techniques, the study aims to predict conversion rates based on these factors, enabling targeted strategies to improve the user experience and optimize conversion rates.

2. Literature Review

2.1. User Conversion in E-commerce and Mobile Apps

The prediction of conversion rates has become an area of significant interest in both e-commerce and mobile applications, particularly within the healthcare sector. Existing studies have explored various dimensions of this topic, encompassing analytical methodologies, influencing factors, and sector-specific insights, all contributing to effective marketing and user engagement strategies.

In the context of e-commerce, numerous researchers have sought to understand the determinants of conversion rates through empirical studies. For instance, Fatta et al. analyze the factors affecting conversion rates on SME e-commerce websites, identifying key variables such as website quality, promotional offers, and user experience [11]. This study highlights how effective design and strategic marketing can lead to improved customer decisions and increased conversion rates. Similarly, Yu et al. explore the dynamics of traffic channeling in e-commerce, emphasizing the significance of conversion rates for developing optimal traffic management strategies, which can significantly enhance sales [12]. These findings underscore the necessity for structured approaches in understanding user pathways and influencing factors that encourage conversions.

Parallel findings emerge in the healthcare sector, where mobile applications are gaining traction. For instance, studies focused on self-management apps illustrate how user engagement can be bolstered through personalized experiences and actionable insights. Wu et al. investigate the adoption of self-management apps among COPD patients, identifying features that enhance patient interaction with the app [13]. Similarly, Wulfovich et al. highlight how design implications can influence the self-efficacy of patients managing chronic conditions through mobile apps, thereby impacting long-term user engagement [14]. Understanding how users perceive and interact with healthcare apps is crucial to predicting and enhancing conversion.

Furthermore, research on integrated mobile and web-based applications for health information dissemination indicates that these platforms can reshape user behaviors positively. The study by Mwammenywa et al. points to the role of mHealth applications in providing vital health information, which in turn fosters higher engagement rates among users [15]. This reflects the growing appreciation for how the healthcare domain can benefit from conversion rate prediction models similar to those applied in e-commerce.

2.2. Machine Learning for Conversion Prediction

The prediction of user behavior, particularly concerning conversion rates, has become increasingly sophisticated, employing a variety of statistical and machine learning methods. Prominent among these techniques are Logistic Regression, Support Vector Machines (SVM), and Neural Networks. Each of these models has distinct advantages and applications in predicting user behavior, such as conversion rates in e-commerce and healthcare domains.

Logistic Regression is one of the earliest and most widely employed models for binary classification. The method predicts the probability that a certain event occurs, making it particularly useful for conversion predictions where the outcome is often binary—whether a user converts or not. In their study, Rahman et al. utilized logistic regression among other models, illustrating its effectiveness in predicting pain volatility for users of a pain management app [16]. By employing logistic regression with techniques like ridge estimators and LASSO to mitigate overfitting, the researchers were able to extract meaningful insights from the dataset, demonstrating this model’s practical applicability in healthcare technologies. Logistic regression is favored for its interpretability and relative simplicity, facilitating a clearer understanding of how independent variables impact the likelihood of conversion. This model can effectively work with multiple predictor variables, making it suitable for applications like predicting user behavior in various healthcare and e-commerce contexts where user characteristics and behaviors can be quantified [17].

Support Vector Machines (SVM) are a more complex modeling technique that excels in high-dimensional spaces. They work by finding the optimal hyperplane that separates data points of different classes. Rahman et al. also employed SVM as one of the methods to gauge user behaviors, highlighting its robustness in handling complex data structures [16]. SVM is particularly effective in scenarios where the data is not linearly separable, as it utilizes kernel functions to project data into higher dimensions. The adaptability of SVM to map complex relationships makes it suitable for predicting user behavior in domains like social e-commerce, where numerous interacting variables may influence the likelihood of a purchase [18]. Its ability to manage class imbalance—often an issue in behavioral prediction—makes SVM a valuable tool when dealing with datasets that exhibit uneven distributions of user conversion instances.

Neural Networks, especially deep learning architectures, have gained traction due to their ability to capture intricate patterns within large volumes of data. The flexibility of neural networks allows them to model non-linear relationships, which is essential when predicting user behavior that is influenced by numerous interrelated factors. Nomura et al. discussed how they utilized neural networks to predict healthcare costs, affirming their capability to handle both discrete and continuous variables effectively [19]. Moreover, neural networks are increasingly being adapted for user conversion predictions in various applications, including e-commerce and healthcare apps. Gao et al. illustrated a context where a heterogeneous graph neural network was employed to model user behavior in social networks, emphasizing how such advanced neural architectures add value by capturing the influence of interconnected user relationships [20]. This highlights the potential for neural networks to predict complex behaviors, particularly in dynamic environments where social influence is a factor.

2.3. Behavioral and Demographic Data

Understanding how behavioral and demographic data influence conversion rates is crucial for optimizing marketing strategies and enhancing user engagement across various sectors, including e-commerce and healthcare. The interplay between these factors can dictate how individuals make purchase decisions and engage with services. This review synthesizes various studies and theories linking user behavior and demographic characteristics to conversion outcomes.

Recent research highlights the importance of user behavior in driving conversion rates. For example, Dingre's study explores how behavioral economics principles can be leveraged to enhance conversion rates in digital marketing. It emphasizes the significance of understanding human decision-making processes and suggests that incorporating behavioral elements into user profiles can inform more strategic marketing approaches [21]. The framework presented underscores the necessity to analyze user interactions, such as browsing habits and responses to promotional content, and how these behaviors correlate with actual purchase decisions.

Additionally, Utami et al. investigate the role of interactivity in enhancing customer engagement in mobile e-commerce applications. They identify key drivers of engagement, including personalized recommendations and interactive features, which directly enhance conversion likelihood [22]. Their findings align with behavioral theories suggesting that increased user interaction not only improves the shopping experience but also builds trust and motivation to complete purchases.

Demographic factors, such as age, gender, income level, and education, play a pivotal role in shaping user behavior and ultimately influencing conversion rates. Lincy and Bella's study highlights how targeted advertisements on platforms like Google and Bing yield varied conversion rates based on demographic segments [23]. By tailoring marketing campaigns to align with the specific characteristics and preferences of different target demographics, businesses can significantly enhance their conversion efficiency.

A holistic approach, integrating both behavioral and demographic data, enhances predictive modeling and marketing effectiveness. Liu et al. discuss how demographic inference can be achieved through user interactions such as ratings and reviews, thereby facilitating more personalized marketing [24]. This integration generates richer user profiles, allowing for tailored communications that significantly increase the chances of conversion. Behavioral patterns extracted from demographic data can provide actionable insights into how users interact with products and services, enhancing the strategic targeting of marketing efforts.

3. Method

3.1. Data Loading

The first step in the process is loading the dataset using the load_data function. This function accepts a file path as input and attempts to load the data into a pandas DataFrame. The function uses pd.read_csv(file_path) to read the CSV file. The function also handles potential errors such as a FileNotFoundError if the file is not found in the specified location. Once the data is loaded successfully, the shape of the dataset (i.e., the number of rows and columns) is printed to give an overview of the data size. If any issues arise during the loading process, an error message is displayed to inform the user. This loading step is critical for ensuring that the dataset is available for further processing and analysis.

3.2. Initial Data Exploration (EDA)

The initial_eda function is designed to conduct an initial exploratory data analysis to better understand the structure of the dataset. This function first displays the first five rows using df.head() to provide a quick snapshot of the data. Then, the df.info() method is used to show detailed information about the dataset, including the number of entries, column names, and data types of each feature. Descriptive statistics for numerical features are calculated with df.describe(), and categorical features are summarized using df.describe(include=['object', 'category']). This allows us to assess the distribution and range of values for each feature. The function also checks for missing values using df.isnull().sum(), ensuring that any gaps in the data can be identified and addressed during preprocessing. A contingency table is created using pd.crosstab() to examine the relationship between the group and pack_color columns. If these columns are found to be redundant (as expected in this case, where group corresponds directly to the packaging color), one of them can be safely dropped, simplifying the dataset. If any discrepancies are found, further inspection is needed.

3.3. Data Cleaning & Preprocessing

In the preprocess_data function, data cleaning and preprocessing steps are applied to prepare the dataset for modeling. The target variable, converted, is separated from the feature set X, which consists of all the other columns. First, columns that are irrelevant to the prediction task, such as user_id, pack_color (due to redundancy with group), and purchase_datetime (since it represents the outcome and not a feature), are dropped using df.drop(). For the date and time columns (visit_date and visit_time), a new feature called visit_datetime is created by combining these columns into a single datetime object using pd.to_datetime(). From this combined datetime, additional features such as visit_hour (the hour of the visit) and visit_dayofweek (the day of the week the visit occurred) are extracted using .dt.hour and .dt.dayofweek, respectively. These new features capture temporal aspects of user behavior. The original date/time columns are then dropped, leaving only the relevant features. The target variable converted is separated from the features, and the feature matrix X is prepared for modeling.

The next step is identifying numerical and categorical features. Numerical features are selected using X.select_dtypes(include=np.number).columns.tolist(), while categorical features are selected with X.select_dtypes(include=['object', 'category']).columns.tolist(). This ensures that all numerical and categorical features are identified correctly. If there are any binary features (such as group, which is encoded as 0 or 1), they are treated as numerical, but they will not be scaled, as scaling is only applied to the actual continuous numerical features. To handle missing values, preprocessing pipelines are created for both numerical and categorical features. For numerical features, a Pipeline is used, consisting of two steps: imputation using SimpleImputer(strategy='median') to fill missing values with the median and scaling using StandardScaler() to standardize the features. For categorical features, another Pipeline is created, which first imputes missing values using the most frequent value (SimpleImputer(strategy='most_frequent')) and then applies OneHotEncoder() to encode the categorical features into binary vectors. The ColumnTransformer is used to apply these pipelines to the appropriate columns, ensuring that each feature is processed according to its type.

3.4. Visual Exploratory Data Analysis (Visual EDA)

The visual_eda function performs a more detailed visual exploration of the data. This function starts by examining the distribution of the target variable, converted, using a countplot from Seaborn, which displays the count of converted and non-converted users. The palette=["#FF9999", "#66B2FF"] is used to set colors for the bars, making it visually appealing and easy to differentiate the categories. The function also annotates the bars with the percentage of each class to give additional context. A bar plot is then created to visualize the conversion rate by packaging group (group). The sns.barplot() function is used to plot the conversion rates for the two groups (red and blue packaging), and the conversion percentages are displayed on top of the bars for easy interpretation.

Next, the function explores categorical features (such as age_group, gender, device_type, and others) by plotting bar plots showing how each category affects the conversion rate. The conversion rate for each category is calculated, and the corresponding plot is displayed. The function also explores numerical features by visualizing their distribution using histplot and boxplot, which show how the features (such as time_on_page_sec and visit_hour) are distributed across users who converted and those who did not. This helps to identify trends, outliers, and patterns in the data. Finally, a correlation heatmap is generated for numerical features using sns.heatmap(), which visualizes the relationships between numerical features and can help identify potential multicollinearity or important relationships that might influence the conversion rate.

3.5. Model Training & Evaluation

The train_evaluate_model function is used to train and evaluate machine learning models on the preprocessed data. The data is split into training and testing sets using train_test_split() from Scikit-learn, with 75% of the data used for training and 25% for testing. The stratify=y argument ensures that the distribution of the target variable is similar in both the training and testing sets, preventing any potential imbalance. The function allows for different models to be used, including Logistic Regression, Random Forest, and Naive Bayes. The preprocessor is applied to the data using preprocessor.fit_transform() on the training set and preprocessor.transform() on the testing set.

The evaluation metrics used to assess model performance include accuracy, precision, recall, F1 score, and ROC AUC. These metrics are calculated using functions from Scikit-learn such as accuracy_score(), precision_score(), recall_score(), f1_score(), and roc_auc_score(). The confusion matrix is plotted to show the true positive, true negative, false positive, and false negative counts, helping to understand how well the model is distinguishing between the two classes. The ROC curve is also plotted to evaluate the trade-off between sensitivity (true positive rate) and specificity (false positive rate). Finally, feature importance is extracted from the trained Random Forest model using model.feature_importances_ and visualized with a bar plot. This helps to identify which features are most influential in predicting user conversion.

4. Results and Discussion

4.1. Data Overview Analysis

The dataset used in this analysis contains 1,000 user sessions from a mobile pharmacy app, with 15 columns in total. The dataset includes user behavior data such as session details, product page interactions, and demographic information. Upon initial inspection, the data was successfully loaded, with no missing values in most of the columns, except for the purchase_datetime column, which only has 206 non-null entries out of 1,000. This is expected as not every session resulted in a purchase. The categorical features include group (representing the packaging color, either red or blue), age_group, gender, device_type, visit_date, and visit_time, while the numerical features include time_on_page_sec, scrolled_to_reviews, added_to_cart, previous_app_user, previous_product_buyer, and the engineered features visit_hour and visit_dayofweek.

From the exploratory data analysis (EDA), it was found that the dataset includes a mix of categorical and numerical features. The target variable converted indicates whether the user made a purchase, with a mean of 0.206, suggesting that approximately 20% of the users converted (made a purchase). The group feature shows an almost equal distribution between the two packaging groups (511 users in Group A - red, and 489 users in Group B - blue). The pack_color feature was found to be redundant with the group feature and was subsequently dropped, simplifying the dataset.

4.2. Results of Descriptive Statistics

The numerical features in the dataset exhibit a range of values that help in understanding user interactions. For instance, the average time spent on the product page (time_on_page_sec) is approximately 69.7 seconds, with a standard deviation of 19.96 seconds, showing a moderate variation in how long users engage with the page. The scrolled_to_reviews feature, which indicates whether users scrolled to the reviews section, has a mean value of 0.37, suggesting that a minority of users engage with the reviews section. Similarly, the added_to_cart feature, which indicates whether users added the product to their cart, has a mean of 0.43, meaning a significant portion of users showed interest by adding the product to their cart but did not necessarily make a purchase.

Regarding the categorical features, the most common age_group is "55+" with 211 occurrences, and the most frequent gender is male (503 occurrences). The majority of users accessed the app via Android devices (503 occurrences), and the most common visit date is "2024-11-26." These descriptive statistics give an initial insight into user behavior and highlight some areas to focus on for improving conversion rates, such as targeting specific age groups or improving the Android app experience.

4.3. Visual Exploratory Data Analysis (Visual EDA)

The visual EDA provided a deeper understanding of the data. The distribution of the target variable, converted, was visualized with a count plot, showing a significant imbalance in the dataset, with only 20.6% of users converting. This imbalance is crucial for model training, as it might affect the performance of certain algorithms, especially if the models are not adjusted for class imbalance.

Figure 1 shows the conversion rate across different age groups. The highest conversion rate is observed for the 25–34 age group, at 24.24%, which is slightly higher than the 55+ group, at 24.17%. The 18–24 and 45–54 age groups show conversion rates of 19.90% and 18.88%, respectively. The 35–44 age group has the lowest conversion rate at 15.46%. This suggests that younger users (18-34) are more likely to convert, with the 25–34 group having the most significant engagement. The 55+ group also shows strong conversion behavior, potentially indicating that this age demographic responds well to the pharmacy app. These insights could help marketers and app developers target their strategies toward users in the 25–34 and 55+ age ranges for better conversion results.

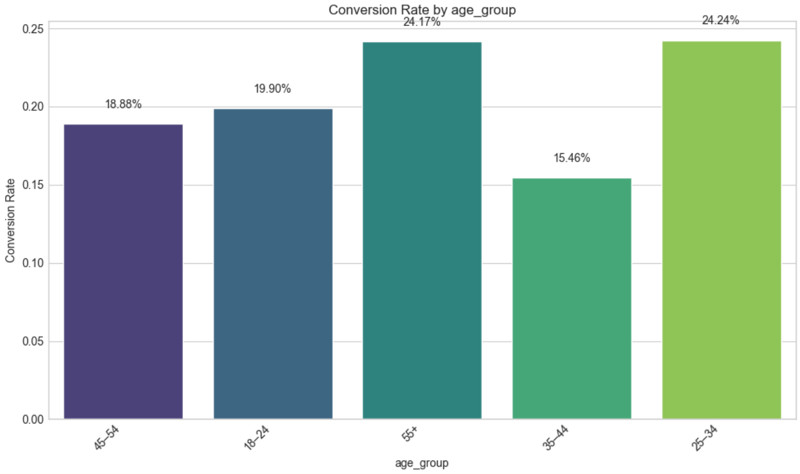

Figure 1. Conversion Rate by Age Group

Figure 2 provides a detailed view of conversion rates across different visit dates. The highest conversion rates are seen on 2024-11-07 (37.50%) and 2024-11-11 (36.67%), followed by several other high-conversion days such as 2024-11-19 (29.41%) and 2024-11-12 (27.78%). Conversion rates vary significantly across different dates, with some days showing much higher engagement than others. For example, on 2024-11-15, the conversion rate drops significantly to 7.65%, indicating that user behavior is influenced by specific factors tied to certain dates, such as promotions, app updates, or even external factors like holidays. This variation suggests that the app's conversion rate can be sensitive to external events or timing, which may be leveraged to optimize campaigns or app features for higher conversions.

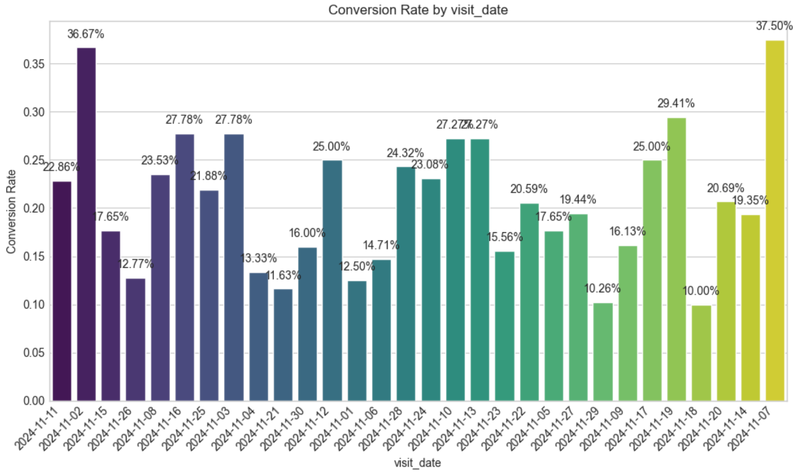

Figure 2. Conversion Rate by Visit Date

Figure 3 displays the distribution of time_on_page_sec (time spent on the product page) by the conversion status. The histogram reveals a clear distinction between users who converted (blue bars) and those who did not convert (red bars). Users who did not convert tend to have a wider spread of time spent on the page, with many users spending less than 60 seconds on the product page. On the other hand, users who converted generally spent more time on the page, as shown by the denser blue bars on the right side of the histogram. This suggests that users who engage with the page for a longer period are more likely to convert, indicating that time on the product page is a key factor in driving conversions.

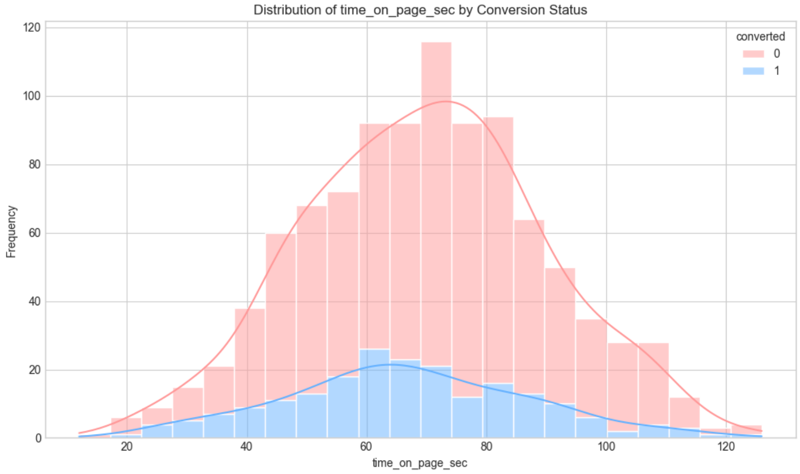

Figure 3. Distribution of Time on Page by Conversion Status

Figure 4 visualizes the distribution of visit_dayofweek (the day of the week when the user visited the product page) by conversion status. The bars represent the frequency of visits on each day of the week (0 = Monday, 6 = Sunday). While the conversion status is relatively consistent across all days of the week, there is a slight variation in the conversion rate, as shown by the smooth line for the converted users (blue) and the non-converted users (red). The conversion rate appears slightly higher on weekends (Saturday and Sunday, represented by days 5 and 6), which may suggest that users are more likely to convert on these days, potentially due to more free time for browsing and purchasing. However, the variation in conversion by day of the week is not as pronounced as it is for time_on_page_sec, indicating that the time spent on the page has a stronger influence on conversion behavior than the day of the week.

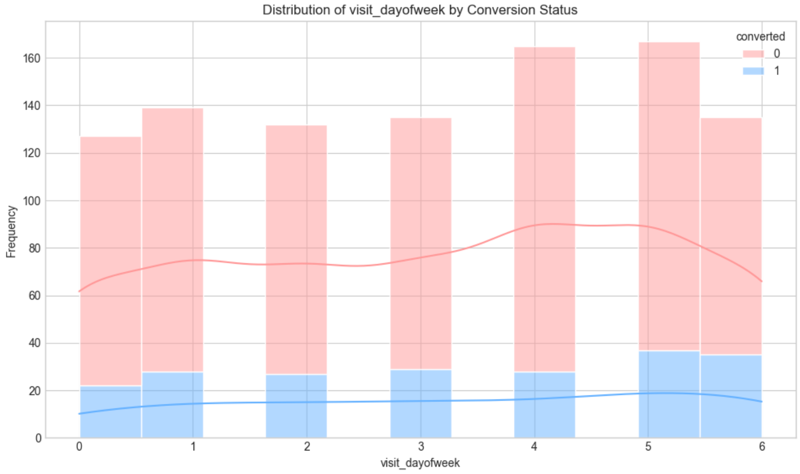

Figure 4. Distribution of Visit Day of the Week by Conversion Status

4.4. Logistic Regression Model Evaluation Results

The logistic regression model was trained on the processed dataset to predict user conversion (whether a user made a purchase or not) based on various features such as time spent on the product page, device type, and user demographics. The dataset was split into a training set (750 samples) and a testing set (250 samples), with the target variable being converted. After preprocessing, the number of features increased to 18, including both numerical and one-hot encoded categorical features. The model was trained using standard logistic regression, and the training process was completed without any issues.

The evaluation of the logistic regression model yielded notable results for both the training and test sets. On the training set, the model achieved an accuracy of 77.47%, indicating that it correctly predicted the conversion status for over three-quarters of the training data. The precision of the model on the training set was 47.84%, meaning that when the model predicted a conversion, it was correct about 48% of the time. The recall was exceptionally high at 100%, indicating that the model correctly identified all users who actually made a purchase in the training data, though this suggests a possible overfitting issue. The F1-score, a balanced metric combining precision and recall, was 64.72%, reflecting a moderate performance between the two metrics. The ROC AUC score was 0.8977, indicating that the model has good discriminatory ability, distinguishing between users who converted and those who did not.

On the test set, the model performed similarly well, achieving an accuracy of 78.80%. Precision improved slightly to 49.02%, and recall dropped marginally to 98.04%. The F1-score on the test set was 65.36%, consistent with the performance on the training set. The ROC AUC score for the test set was 0.8758, showing that the model maintained a strong ability to discriminate between the two classes, though it did show a slight decrease from the training set, which is common when transitioning from training data to unseen test data. Figure 5 shows the ROC curve for the Logistic Regression model. The area under the curve (AUC) is 0.88, which indicates that the model performs well at distinguishing between the two classes (converted vs. not converted). The curve rises steeply towards the top-left corner, meaning the model is able to achieve a high true positive rate with a low false positive rate. This suggests that Logistic Regression is effective in predicting user conversion, making it a good choice for this task.

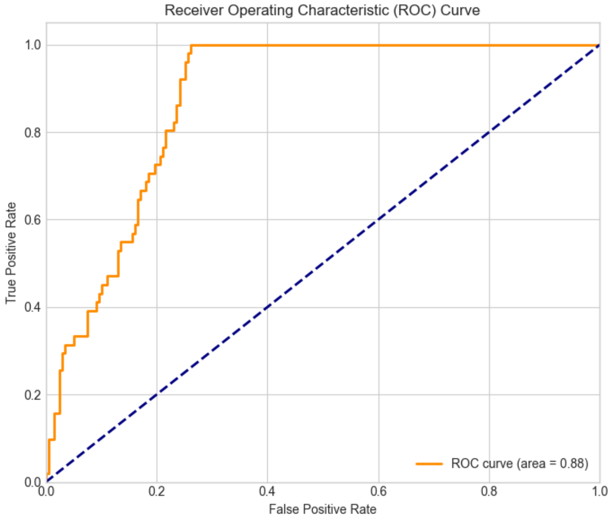

Figure 5. ROC Curve of Logistic Regression

The feature importance values from the logistic regression model provide valuable insights into which factors contribute most to predicting user conversion. The most influential feature was numerical__added_to_cart, with an importance score of 3.066434, indicating that whether a user added the product to their cart is a strong predictor of conversion. This aligns with expectations, as adding an item to the cart typically signifies high purchase intent. Following this, categorical__group_A had a negative importance score of -0.776637, indicating that being in Group A (red packaging) may reduce the likelihood of conversion compared to Group B (blue packaging), which was a key finding in the exploratory data analysis (EDA). The categorical__gender_F and categorical__device_type_Android features also had negative importance scores of -0.552340 and -0.490349, respectively, suggesting that female users and those using Android devices were less likely to convert compared to male users and iOS users.

The categorical__device_type_iOS feature had a similar negative importance score of -0.360052, reinforcing the trend that iOS users have a lower likelihood of converting when compared to Android users. Other significant features included numerical__time_on_page_sec, with a negative importance of -0.242437, suggesting that users who spend more time on the product page are less likely to convert, which could indicate that the users are engaging with the product but are uncertain or not convinced by the offering. The categorical__age_group_45–54 and categorical__age_group_55+ features, with negative importance scores of -0.271167 and -0.209214, respectively, show that older age groups tend to have a lower likelihood of conversion, whereas younger age groups such as categorical__age_group_18–24 (with an importance of -0.177605) are slightly more likely to convert.

Other features, such as numerical__visit_hour and numerical__visit_dayofweek, had relatively low importance scores, suggesting that the time of the visit did not have a strong influence on the conversion decision. The numerical__previous_app_user, numerical__scrolled_to_reviews, and numerical__previous_product_buyer features had even smaller impacts on the model, with the smallest importance score coming from numerical__previous_product_buyer, which had a score of -0.002827, suggesting that whether the user had bought this product previously had a minimal effect on conversion.

4.5. Random Forest Model Evaluation Results

The Random Forest model was trained and evaluated to predict user conversion, using the same dataset as for the logistic regression model. The dataset was split into training and test sets, with 750 samples for training and 250 samples for testing. After preprocessing, the number of features was increased to 18, including both numerical and one-hot encoded categorical features. The model was trained with the default parameters, and the evaluation metrics were assessed to understand the model's performance.

On the training set, the Random Forest model performed exceptionally well, achieving perfect scores across all metrics: accuracy (1.0000), precision (1.0000), recall (1.0000), F1-score (1.0000), and ROC AUC (1.0000). This suggests that the model has perfectly memorized the training data, which is indicative of overfitting—a common issue when models perform excessively well on training data but fail to generalize well to unseen data. When evaluated on the test set, the model’s performance dropped significantly. The accuracy on the test set was 78.80%, which is a noticeable decline compared to the training set. Precision was 45.45%, recall was 19.61%, F1-score was 27.40%, and the ROC AUC was 0.8727. The drop in recall and F1-score suggests that the model is not effectively identifying all the users who are likely to convert, and the low precision indicates that many of the positive predictions made by the model are false positives. The model’s performance on the test set highlights the potential overfitting, where the Random Forest model is too specialized to the training data and does not generalize well to new data. To improve the model's ability to generalize, hyperparameter tuning or regularization techniques such as limiting the depth of trees or increasing the number of estimators could be explored.

Figure 6 shows the ROC curve for the Random Forest model. The AUC for this model is 0.87, slightly lower than the Logistic Regression model, but still indicates a strong performance in classifying users. The curve is similar to the Logistic Regression plot, with a steady rise towards the top-left corner, though not as steep. This suggests that while the Random Forest model performs well, it is slightly less effective at distinguishing between users who convert and those who do not compared to the Logistic Regression model.

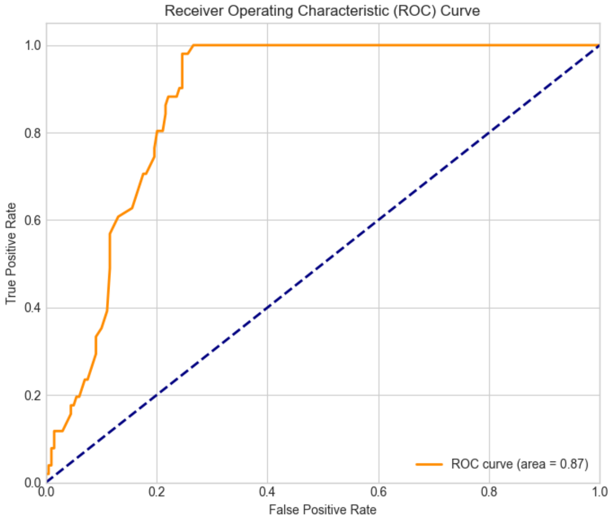

Figure 6. Random Forest

The feature importance values derived from the trained Random Forest model provide insights into which factors have the greatest influence on predicting user conversion. The most influential feature in the Random Forest model was numerical__added_to_cart, with an importance score of 0.451050, reinforcing the finding from the logistic regression model that adding a product to the cart is a strong predictor of conversion. Other significant features include numerical__time_on_page_sec (0.138838) and numerical__visit_hour (0.108521), both of which show that the amount of time spent on the product page and the time of the visit are important factors in determining conversion likelihood. Interestingly, the feature categorical__group_A (0.015106) had a relatively low importance score, suggesting that the packaging color (red) is less impactful in the Random Forest model compared to other features. This contrasts with the logistic regression model, where packaging color had a more pronounced effect. The categorical features related to user demographics, such as categorical__age_group_55+ (0.016666), categorical__device_type_iOS (0.016516), and categorical__gender_F (0.014443), had modest importance scores, suggesting that while demographic information is valuable, it is not as decisive as user engagement features like time on page or cart additions.

5. Conclusion

This study aimed to predict user conversion in a mobile pharmacy app using machine learning techniques, focusing on the impact of behavioral and demographic data. The analysis revealed that user interactions such as adding a product to the cart and the time spent on the product page were significant predictors of conversion, alongside packaging color and device type. The logistic regression model and Random Forest model both showed that engagement features, like added_to_cart and time_on_page_sec, were among the most important factors influencing conversion. However, while the models performed well on the training set, there was a notable decline in performance on the test set, indicating potential overfitting, particularly with the Random Forest model. The feature importance analysis highlighted the key factors influencing conversion, guiding future optimization strategies. This research contributes to the field by demonstrating how machine learning can be applied to predict user conversion in the mobile app industry, specifically within the context of a mobile pharmacy app. The study adds value to the growing body of work on e-commerce and mobile app analytics, showcasing the potential of machine learning to leverage user behavior and demographic data for improving conversion rates. By identifying and quantifying the impact of key features, the research provides actionable insights for app developers and marketers to enhance user engagement and optimize the user experience, ultimately driving sales and improving customer retention. While the study provides valuable insights, it also has limitations. One limitation is the potential overfitting of the models, especially the Random Forest model, which demonstrated perfect accuracy on the training set but struggled with generalization to the test set. Another limitation is the lack of real-time data, which could provide more dynamic and accurate predictions of user behavior. Future research could focus on refining the models through hyperparameter tuning or by exploring alternative algorithms such as gradient boosting or neural networks. Additionally, incorporating real-time data and considering additional user behavioral features, such as interaction frequency or in-app activities, could further enhance model performance. The findings from this research can be applied practically to improve user engagement and conversion rates in mobile pharmacy apps by optimizing the user experience, targeting specific demographic groups, and personalizing the app content based on predictive insights.

6. Declarations

6.1. Author Contributions

Conceptualization: S.L., K.S.; Methodology: S.L., R.F.A.; Software: S.L.; Validation: K.S., R.F.A.; Formal Analysis: S.L.; Investigation: S.L.; Resources: K.S., R.F.A.; Data Curation: S.L.; Writing – Original Draft Preparation: S.L.; Writing – Review and Editing: K.S., R.F.A.; Visualization: S.L.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Al-Arkee *et al.*, "Mobile Apps to Improve Medication Adherence in Cardiovascular Disease: Systematic Review and Meta-Analysis," *J. Med. Internet Res.*, 2021, doi: 10.2196/24190.

[2] J. C. Wong, L. Hekimyan, F. A. Cruz, and T. Brower, "Identifying Pertinent Digital Health Topics to Incorporate Into Self-Care Pharmacy Education," *Pharmacy*, 2024, doi: 10.3390/pharmacy12030096.

[3] B. KC *et al.*, "Positioning and Utilization of Information and Communication Technology in Community Pharmacies of Selangor, Malaysia: Cross-Sectional Study," *Jmir Med. Inform.*, 2020, doi: 10.2196/17982.

[4] G. ÜSTÜN, S. L. SÖYLEMEZ, N. UÇAR, M. Sancar, and B. Okuyan, "Assessment of the Pharmacy Students E-Health Literacy and Mobile Health Application Utilization," *J. Res. Pharm.*, 2020, doi: 10.35333/jrp.2020.125.

[5] N. Cobelli and A. Chiarini, "Improving Customer Satisfaction and Loyalty Through mHealth Service Digitalization," *TQM J.*, 2020, doi: 10.1108/tqm-10-2019-0252.

[6] A. Sattar *et al.*, "Trustworthiness of Web-Based Pharmacy Apps in Pakistan Based on the Mobile App Rating Scale: Content Analysis and Quality Evaluation," *Jmir Mhealth Uhealth*, 2025, doi: 10.2196/59884.

[7] I. Irnawati *et al.*, "Users' Perceptions of the 'My Medicine' Mobile App Usability," *J. Public Health Res.*, 2022, doi: 10.1177/22799036221115782.

[8] T. Song, J. Tang, M. Kuang, and H. Liu, "Current Status and Future Prospects of Chinese Mobile Apps for Hypertension Management," *Health Informatics J.*, 2024, doi: 10.1177/14604582241275816.

[9] R. Yin and D. M. Neyens, "Who Is Using a Mobile App and Who Is Using a Computer to Access Their Patient Portals?," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 2022, doi: 10.1177/1071181322661316.

[10] J. Cao, G. Zhang, and D. Liu, "The Impact of Using mHealth Apps on Improving Public Health Satisfaction During the COVID-19 Pandemic: A Digital Content Value Chain Perspective," *Healthcare*, 2022, doi: 10.3390/healthcare10030479.

[11] D. D. Fatta, D. Patton, and G. Viglia, "The Determinants of Conversion Rates in SME E-Commerce Websites," *J. Retail. Consum. Serv.*, 2018, doi: 10.1016/j.jretconser.2017.12.008.

[12] P. Yu, Z. J. Zhang, and Q. Li, "Traffic Channeling Under Uncertain Conversion Rates on E-commerce Platforms," *Nav. Res. Logist. Nrl*, 2022, doi: 10.1002/nav.22079.

[13] R. Wu, S. Ginsburg, T. Son, and A. S. Gershon, "Using Wearables and Self-Management Apps in Patients With COPD: A Qualitative Study," *Erj Open Res.*, 2019, doi: 10.1183/23120541.00036-2019.

[14] S. Wulfovich, M. Fiordelli, H. Rivas, W. Concepción, and K. Wac, "'I Must Try Harder': Design Implications for Mobile Apps and Wearables Contributing to Self-Efficacy of Patients With Chronic Conditions," *Front. Psychol.*, 2019, doi: 10.3389/fpsyg.2019.02388.

[15] I. Mwammenywa, M. H. Nkotagu, and S. F. Kiajage, "Integrated Mobile and Web-Based Application for Enhancing Delivery of HIV/AIDS Healthcare Information in Tanzania," *Tanzan. J. Eng. Technol.*, 2019, doi: 10.52339/tjet.v38i2.502.

[16] Q. A. Rahman, T. Janmohamed, H. Clarke, P. Ritvo, J. M. Heffernan, and J. Katz, "Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods," *Jmir Med. Inform.*, 2019, doi: 10.2196/15601.

[17] J. Hu, G. Huang, and L. Wang, "Research on User Portrait Based on Xgboost and Logistic Ensemble Learning Methods," *Highlights Sci. Eng. Technol.*, 2022, doi: 10.54097/hset.v12i.1453.

[18] J. Lv, T. Wang, H. Wang, J. Yu, and Y. Wang, "A SECPG Model for Purchase Behavior Analysis in Social E-commerce Environment," *Int. J. Commun. Syst.*, 2020, doi: 10.1002/dac.4149.

[19] Y. Nomura *et al.*, "Does Last Year's Cost Predict the Present Cost? An Application of Machine Leaning for the Japanese Area-Basis Public Health Insurance Database," *Int. J. Environ. Res. Public. Health*, 2021, doi: 10.3390/ijerph18020565.

[20] L. Gao, H. Wang, Z. Zhang, H. Zhuang, and B. Zhou, "HetInf: Social Influence Prediction With Heterogeneous Graph Neural Network," *Front. Phys.*, 2022, doi: 10.3389/fphy.2021.787185.

[21] S. S. Dingre, "An Approach to Optimize Conversion Rate Using Behavioral Economics," *J. Mark. Stud.*, 2024, doi: 10.47941/jms.1680.

[22] A. F. Utami, I. A. Ekaputra, A. Japutra, and S. V. Doorn, "The Role of Interactivity on Customer Engagement in Mobile E-Commerce Applications," *Int. J. Mark. Res.*, 2021, doi: 10.1177/14707853211027483.

[23] M. J. Kala Lincy and Dr. K. M. Jes Bella, "Effectiveness of Online Marketing Based Search Engine Advertisements: A Study on Google and Bing in Chennai," *Indian J. Inf. Sources Serv.*, 2024, doi: 10.51983/ijiss-2024.14.3.18.

[24] Y. Liu, H. Qu, W. Chen, and S. Mahmud, "An Efficient Deep Learning Model to Infer User Demographic Information From Ratings," *Ieee Access*, 2019, doi: 10.1109/access.2019.2911720.