Conversion Rate Analysis of E-commerce platform – A Big Data Approach

E-Business T-Project

M.Sc. Business Administration

Juncan Zhang (juzh22ac) (158802)

Date of Submission

03. 08.2023

Normal Pages

(15)

Characters

(4931)

Table of Contents

1 Abstract

This research work presents a comprehensive analysis of a dataset comprising 100,000 records from the Tianchi platform, focusing on user browsing behavior and essential features. Adopting the CRISP-DM methodology, the study conducts an in-depth investigation of the existing conversion flow, identifying potential obstacles in the user journey. Leveraging funnel analysis principles, the study proposes practical solutions to optimize the conversion process and improve user engagement. The data is thoroughly examined, revealing user behavior patterns and page interactions, while funnel analysis uncovers critical bottlenecks in the conversion flow. Drawing on these insights, the study devises measures to streamline the user journey and enhance the conversion rate. Additionally, machine learning techniques are employed to create a predictive model for customer conversion, enabling the identification of potential valuable customers and the customization of marketing strategies for improved customer acquisition and retention. This study showcases the value of the CRISP-DM process in generating actionable insights and offers a holistic approach to address conversion challenges and optimize user engagement.

2 Introduction

As Levitt (1986) defined customer as assets that they can only be managed after acquisicion. In today's digital era, any type of business has the need to generate and preserve customers to maintain operation even to gain profit. In other words, it is necesseary for companies to better manage a customer lifecycle, which means the process of a person choosing and interacting with the company's product or service, including awareness, acquisition, conversion, retention and loyalty (S.Lahey). Though many companies more focus on customer retention which costs less, the customer acquision, should also be given a high priority (L. Ang and F. Buttle, 2010). Obtaining new customers is not only a necessary need for companies entering a new market, it is also helpul for matural firms to supplement customer groups which exist unavoidable lost.

Customer Acquisition can be more important for E-commerce companies, a huge and potential worldwide market. According to Statista(2023), there existed 5 billion internet users among the worldretial E-commerce sales were predicted to exceed 5.7 trillion U.S dollars. To survive and even maintain competivity, it is important for companies to decide proper strategy to recogonize and develop opportunities during in marketing channel and apply them to ensure transactions happen (Gartner Glossary). Moreover, a new customer may bring more profit for a company than maintaining a same ratio of previous customers (Goodwin and Ball, 2003). However, given the importance of customer acquisition, according to the survey made by L. Ang and F. Buttle, (2010), only less than half companies partcipating the survey reportded that they have a relevant plan. Of course, due the limitation of time and sample, the actual ratio may be improved in the past decade, but we can still

conclude that developing methods for better customer acquisition will still benefit a considerable number of companies.

According to Cooper and Budd(2007), this multisatege acquiring process has been defined as “sales funnel”, collecting and filtering a large quantity of data. However, it is difficult to determine which sales leads(eg. User behavioral data) can bring orders for companies. Unproper sales forcasting may bring potential loss of bookings and cause resourse wasting. It is valuable for companies to estabish a quantitative model to predict which potential customers can be convert by analyzing their behavioral data. (Cox, 2019). Besides, in order to have a better effect of modeling, companies need to deal with data in advance, including the processes such as processing, visualization and feature engineering. These processes themselves can also provide valuable insights for business to understand user needs and market potentials. Therefore companies can make specific adjustments to optimize their business strategy and enhance profitability. (Provost & Fawcett, 2013)

In this work, we have used the JingDong’s online sales data in a given period to create business value by analyzing user behavioral data to obtain a overview of the company’s sales data. Also, we compare data historically in this project to determine the profitability of different market and target groups. Finnaly, we apply different machine-learning based model to find out the one with the best performance on predicting the possibiliy of a potential customer being converted to a real customer.

Python based data science libraries and tools including pandas, numpy, matplotlib, seaborns and scikit-learn are used in this work along with Python3 programming environments.

3 Literature Review

3.1 CRISP-DM Process

This work will follow the CRISP-DM Process, as known as the abbreviation for Cross-Industry Standard Process for Data Mining, which is popular for different data mining projects in different industries. It includes six iterative stages(Chapman et al, 2000):

1. Business understanding - In this initial phase, the primary focus lies in comprehending the project's objectives and requirements from a business standpoint. This entails transforming this knowledge into a data mining problem definition and developing a preliminary plan to effectively achieve the specified objectives.

2. Data understanding - The data understanding phase commences with an initial data collection and proceeds with a series of activities aimed at familiarizing oneself with the data. The objective is to identify potential data quality issues, gain preliminary insights into the dataset, and detect intriguing subsets that may lead to the formation of hypotheses for uncovering hidden information.

3. Data preparation - The data preparation phase encompasses all activities involved in constructing the final dataset from the initially collected raw data. This phase involves data cleaning, integration, transformation, and the creation of derived features to make the data suitable for subsequent modeling.

4. Modeling - In this phase, a variety of modeling techniques are carefully selected and applied to the prepared data. Parameters for these models are calibrated to their optimal values, ensuring that the most appropriate approach is used for analyzing the data and generating meaningful insights.

5. Evaluation - At this stage, the model(s) derived from the previous phase undergo comprehensive evaluation. Various performance metrics and validation techniques are employed to assess the models' effectiveness and reliability. Additionally, a thorough review of the steps executed during the modeling process is conducted to ensure that the models align with the intended business objectives.

6. Deployment - The completion of model creation does not mark the end of the project. Irrespective of whether the model aims to enhance data knowledge or address specific business problems, the knowledge gained must be organized and presented in a manner that facilitates its utilization by the customer. This phase involves deploying the model to the intended stakeholders, integrating it into existing systems, and providing suitable documentation and support for its implementation and usage.

Previous Researchers have done a systematic literature review about papers applying CRISP-DM Proecss Model(Chriktoph S. et,al, 2021). They found out that most authors applied this process since it is an easy, structured, reliable, commonly used and industry-independent process model and it seems that it performed well in various fields such as health, education, engineering and information technology. However, though basic steps are similar in different works, papers in different fields decribe the business understanding and data understanding stage quite differently. Hence it is important to calrify the business objective and data descriptions in specific domains (ibid.).

3.2 Sales Funnel Framework

Different from traditional consumer marketing which mainly focuses on purchaseing, customers' actions are only simply divided as response or no response, modern target marketing considers more about customer relationship (Yu & Cai, 2007). These companeis do not only consider the purchase action, they also pay high intention to consumers' decisions through whole customer journey. It is because customers who may build a closer connection with the company may bring more profitability rather than those make purchase decision in a short time. (ibid.) The potential customers' journey interacting with a company is called "Sales Funnel" (Mester, 2022). This visual representation is

because the potential clients' purchasing path is smilar to the funnel structure, where the quantity of potential buyers decreased significantly through the customer journey, untial reaching the number of real customers (Markina, A., 2021). This concept well decribes the process of customer acquisition and divides it into different phases. (Jeroen, D., et, al, 2013). However, according to Venermo et al.(2020), there is no a standard and uniform division method for "sales funnel", it should be adjusted based on the information and needs of specific business units.

As this project is aimed to analyze JingDong's sales data, which collected from users' behavioral data through online purchasing process and do not consider the post-purchase stage, it is more reasonable for us to develop a sales funnel with the basic stages based on Markina A., (2021)' paper:

1. Awareness: the stage before customer considering specific products, company usually needs to provide more attractive information for customers. In this work, we define customers who viewed home page are in this stage.
2. Interest: this is the second stage as customers show interests to buy products when company shows possible solutions for customers' need. In this work, we define customers who viewed listing page and product page are in this stage.
3. Desire: the third stage is customers' comparision stage, they usually select the one offer which has the best consideration including price and quality. In this work , we define customers viewd purchasing page are in this stage.
4. Action: the final stage is purchasing itself, company needs to give a customer a little push to show possible benefits if the customer choosing it. In this work, we define customers viewed confirmation page are in this stage.

Also, to make clearer description, we reference Jeroen, D., et, al (2013)'s work to define customers in different stages to four categories: Suspects, Prospects, Leads and Customers. They correspond to the four stages above, and the whole picture of our "Sales Funnel" framework is :

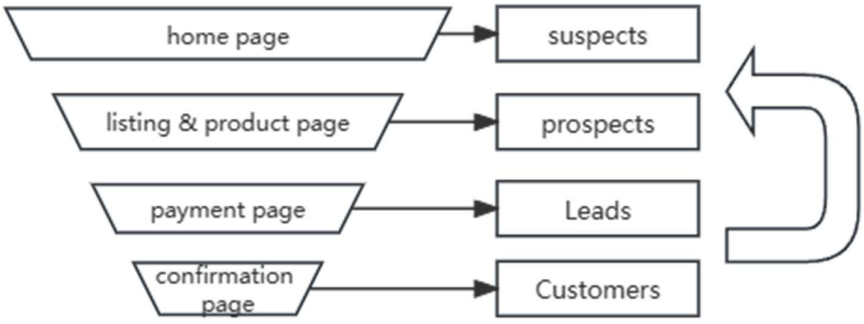


Figure 1 Sales Funnel Framework

“Sales Funnel” Concept is more used in managing the buying preparasion stage for potential buyers, the deeper they are in the funnel, the more possible they we become a real customer. (Yu & Cai, 2007) If the “Sales Funnel” runs smoothly, company will gain continuously increased profit.

3.3 Data Wrangling Techniques

Data analysis and modeling tasks often involve a significant amount of effort in data preparation, which includes tasks like loading, cleaning, transforming, and reorganizing data. Sometimes, the data may not be in the desired format for specific data processing applications. In such cases, many practitioners opt for ad hoc data processing using general-purpose programming. However, the good news is that the Python ecosystem offers powerful tools like pandas and the standard library, which provide high-level, flexible, and efficient functionalities for manipulating and transforming data into the required format with ease (McKinney, W. , 2012).

1. Cleaning: Data cleaning involves identifying and rectifying errors, inconsistencies, and inaccuracies in the dataset. This step ensures that the data is accurate and reliable for analysis. Common cleaning tasks include handling missing values, removing duplicates, and correcting data formatting issues.
2. Transforming: Data transformation involves modifying the data to suit the analysis requirements. This may include converting data types, scaling numeric values, and applying mathematical operations. Transformation can also involve creating new features or aggregating existing ones to derive meaningful insights from the data.
3. Merging: Data merging combines information from multiple datasets based on common attributes or keys. This process is particularly useful when dealing with data distributed across different sources. Merging allows analysts to consolidate relevant information into a single dataset, enabling comprehensive analysis and modeling.
4. Reshaping: Data reshaping involves reorganizing the structure of the data to make it suitable for specific analytical tasks. This may include converting data from a wide format to a long format or vice versa, pivoting data to change the way it is organized, or splitting and merging data based on certain criteria.

3.4 Machine learning methods

Machine learning is a collection of aritificial intelligence that can automatically detect data patterns, integrate information, learn from computer-based models to predict future data or find out valuable insights for other decision making (K.Murphy, 2014). This is particularly useful in today’s big data era. Machine learning basically can be divided into two types:

1. Supervised learning: Learn a mapping between the input and output in order to generate proper predictions or classifications when encountering new data.

2. Unsupervised learning: Dealing with unlabeled data, which means the input is not paired with any corresponding output labels, which is particularly useful when exploring and understanding the characteristics of a dataset without prior knowledge of the expected outcomes.

Also, we have used some classification techniques in this work (Eitle & Buxmann, 2019):

1. Random Forest: Random Forest is an ensemble learning technique for classification. It combines multiple decision trees and makes predictions by aggregating the results of individual trees. Each tree in the forest is trained on a random subset of the data and features. This randomness helps reduce overfitting and improves generalization. Random Forest is known for its high accuracy, robustness to noisy data, and ability to handle large datasets.
2. (MLP) Multi-Layer Perceptron Model: The Multi-Layer Perceptron is a type of artificial neural network known for its ability to model complex, non-linear relationships in data. It consists of multiple layers of interconnected nodes, allowing it to learn hierarchical representations from the input data.
3. Logistic Regression: Logistic Regression is a linear classification technique used to predict binary outcomes (e.g., yes/no, 0/1). Despite its name, it is primarily used for classification rather than regression tasks. It models the relationship between the features and the binary outcome by applying the logistic function to the linear combination of features. The logistic function maps the output to a probability between 0 and 1. Logistic Regression is efficient, interpretable, and works well when the decision boundary is relatively simple and linear.

4 Materials and Methods

Our research goal was to deciding proper measures to improve the platform's GMV(gross merchandise volume) by analyzing the historical user behavioral data. We will firstly find out the pipularity of products in our platform among different user groups(eg. City/region, age, gender…) to make future sales and promotion plan. Also, we will develop a proper model to predict whether a person will become our loyal customer and therefore improving the quality of customer relationship management.

In this work, we follow the CRISP-DM process, firtstly use proper methods to acquire the data, then apply exploratory data analysis(EDA) to detailedly decribe the dataset. Thirdly, we apply data engineering methods in forms of data preprocessing and data cleaning to deal with strange and missing values for a better analyzing and predicting performance. Also, we combine and deal some feature data according to its business meanings for a better explaination. Finally, we built some machine learning models with sk-learn package to predict future customer classification.

4.1 Data Acquisition

After understanding of business and its objectives, the extracted data is the stage of data acquisition. We get the dataset of 100,000 lines user behavior data from a international data provider. The dataset initially has 6 sheets, after combination, we get the dataset including following 14 features:

Column_Name	Type	Decription
user_id	categorical	Uniquely identifies each user, each user_id exists one time in one sheet
new_user	categorical	New user: 1, Old User: 0
age	numerical	User's age
sex	categorical	User's gender, range in ("Male" , "Female")
market	categorical	The market level user in, range in (1,2,3,4)
device	categorical	The device user uses, range in ("mobile" , "desktop")
operative_system	categorical	The operative system user ues, range in ("windows", "iOS","Android","Mac" ,"Linux" and "others")
source	categorical	The source user comes from, range in ""
total_page_visited	numerical	The number of pages user visited
home_page	categorical	Show "home_page" when user viewed home_page
listing_page	categorical	Show "listing_page" when user viewed listing_page
product_page	categorical	Show "product_page" when user viewed product_page
payment_page	categorical	Show "payment_page" when user viewed payment_page
confirmation_page	categorical	Show "confirmation_page" when user viewed confirmation_page

Table 1 Data Description

4.2 Exploratory Data Analysis

Here we use visualization libraries(matplotlib and seaborn) to show the characterstics of each feature to get an overview of the platform's user group:

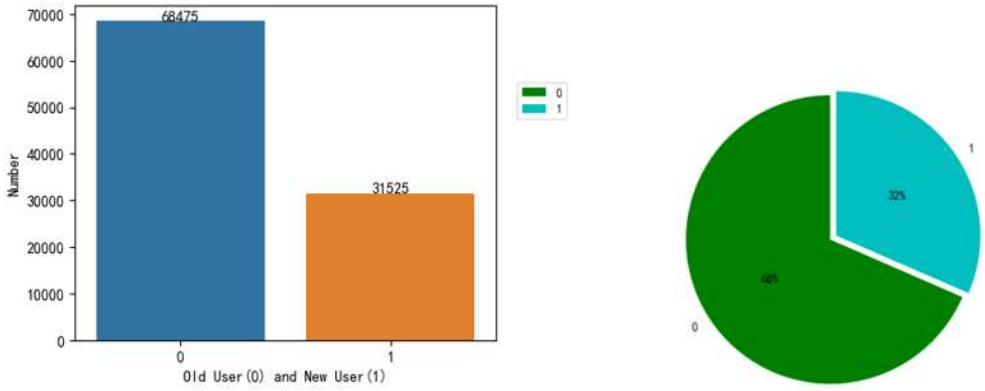


Figure 2 new_user

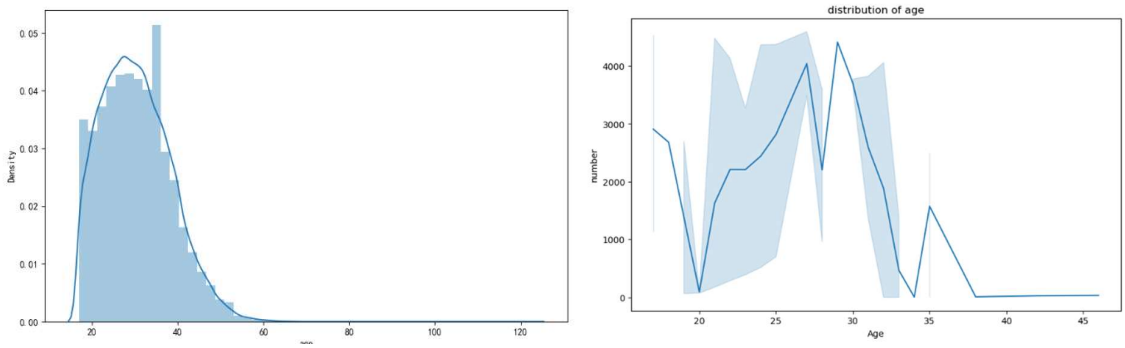


Figure 3 age

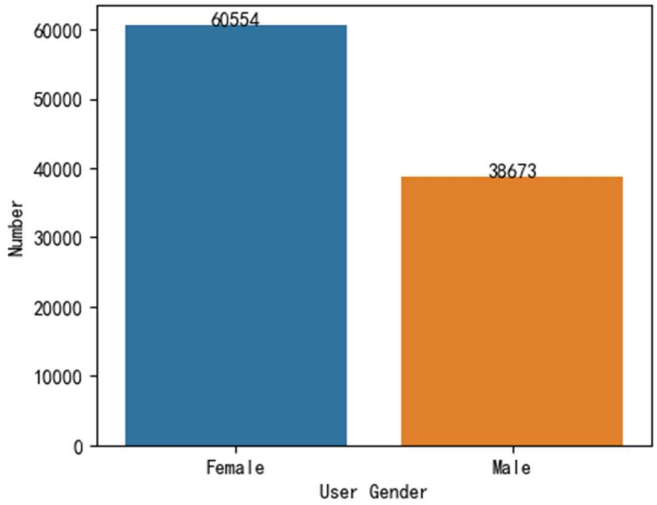


Figure 4 user gender

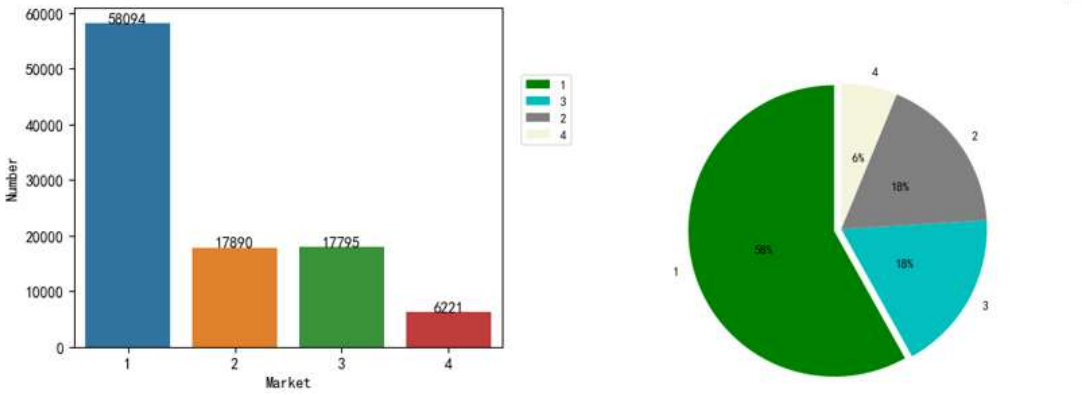


Figure 5 market

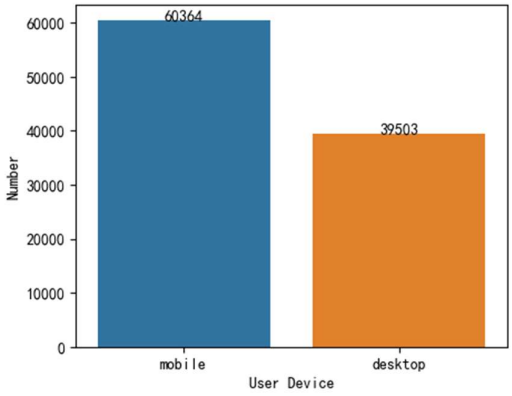


Figure 6 device

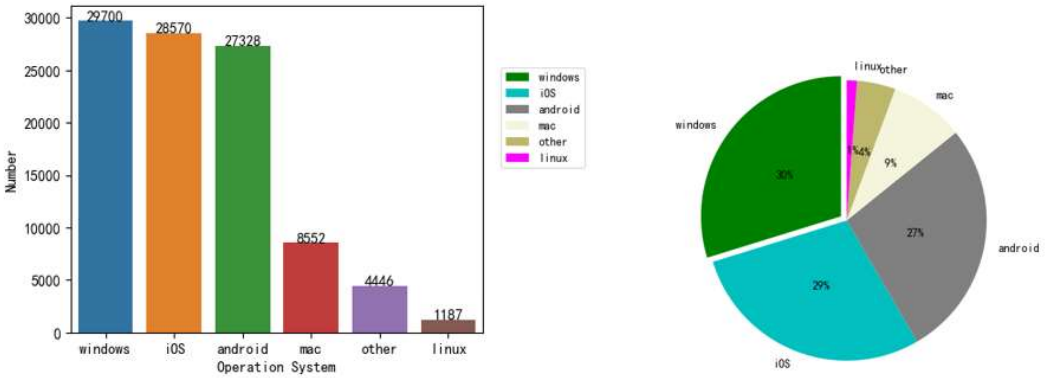


Figure 7 operative system

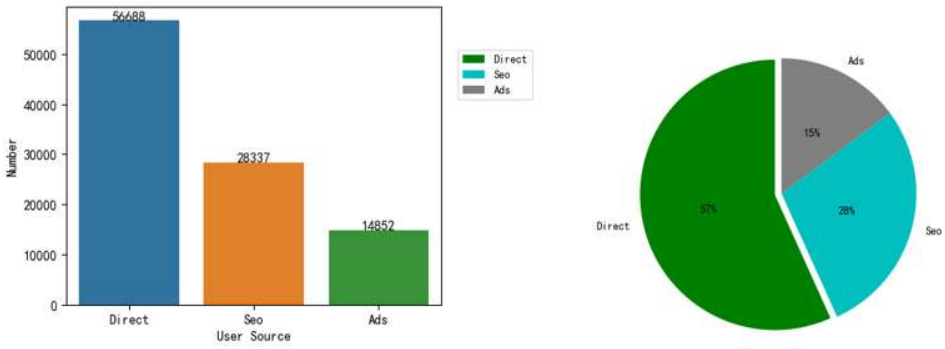


Figure 8 source

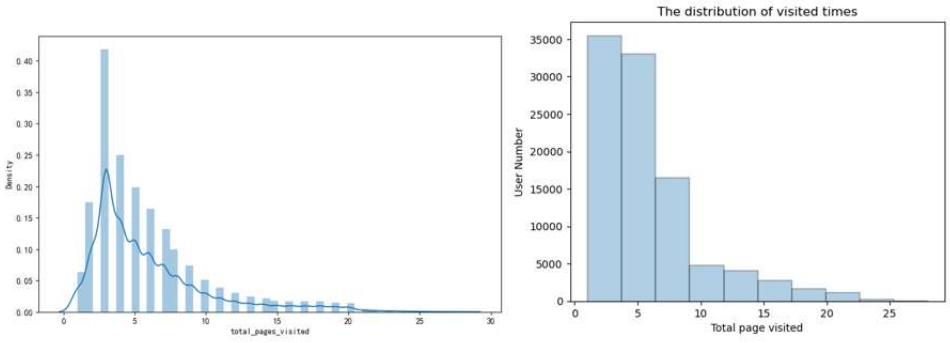


Figure 9 total page visited

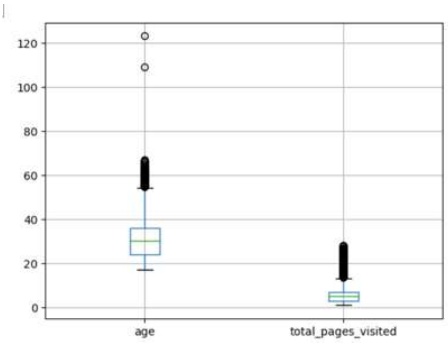
As our data visualization shows, we can observe some characteristics of our dataset:

1. Figure 2 shows the distribution of new and old users, the platform has 68% of old user and 32% new user during the observation period.
2. Figure 3 shows the adg distribution of the platform's user group, users are mainly distributed in the range between 18-35, which means currently the platform is more attractive for young adults.
3. Figure 4 shows the user gender distribution, female users take more than 60% proportion, the platform currently is more attractvie for females.
4. Figure 5 shows the user market segment distribution, the users are mainly coming from 1st level market, least users come from 4th level market.
5. Figure 6 shows the distribution of devices user used, the platform is more popular in mobile channel.
6. Figure 7 shows the operative system distribution, users are mainly from windows, iOs and android system.
7. Figure 8 shows the distribution of user source, users are mainly entering the platform directly or by Seo searching, the current advertisement does not attract considerable number of users.
8. Figure 9 shows the distribution of total visited pages. Most users visited 2-7 pages and this range has a huge difference between other ranges, which means users tent to visit less pages in this platform.

4.3 Data Engineering

We mainly apply data processing techniques in this part, which is an essential step for data science applications. Firstly we apply outlier analysis to check whether there are any data entry errors or data that does not conform to the norm. We usually follow the 3 σ principle, which means that if a data point falls outside three standard deviations from the mean, it is considered a potential outlier. In this work, we analyzed the two columns that may exist strange value: 'age' and 'total_pages_visisted'. We find out that age exist two strange value which is larger than the outlier threshold and is against the

common sense(people are usually younger than 100) . And there is no strange value exists in ‘total_pages_visited’. Then we apply missing value analysis, we find out that colum ‘market’,’device’,’operative_system’, ‘source’ exist missing value, but all in a low proportion, so we apply suitable filling methods, such as filling with mode, filling with means and deleting corresponding lines for each columns.



user_id	0
new_user	0
age	0
sex	773
market	0
device	133
operative_system	217
source	123
total_pages_visited	0
home_page	0
listing_page	0
product_page	0
payment_page	0
confirmation_page	0

dtype: int64

Figure 10 outlier analysis(left) and missing value analysis(right)

4.4 Feature Engineering

Feature engineering is the process of selecting, transforming, and creating new features (variables) from raw data to improve the performance and effectiveness of machine learning algorithms. Here we also gerneate some features based on our business need, the funnel analysis.

We firstly compute the conversion rate between consecutive pages in a user’s journey on the platform. The conversion rate is defined as the number of users who visited the next page divided by the number of users who visited the current page. The overall conversion rate of each flow is shown in Figure 11. Hence we can analyze how the conversion rate is different among different features by computing corresponding conversion rate and visualization techniques. Then we can compute the overall conversion rate (approximately 2.4%), which equals to the product of each flow’s conversion/

Also, since we plan to build models to predict whether a user will convert to a customer(entering the confirmation page), we add a categorical variable: ‘converted’, users who entered confirmation page will be assigned value ‘1’, otherwise ‘0’. Also, we combined operative_system and device since they are highly correlated (mobile users can only use system ios or android while web users can not do so) which have a bad influence to model performance. Then we delete variables 'user_id','home_page','listing_page', 'product_page','payment_page','confirmation_page' because they are unncessary for developing a predictive model.

Finally, we apply the transform-to-numeric function(OneHot Encoding, Label Encoding) to categorical features and Min-Max Scaler to 'total_pages_visited' in order to reach a better model performance.

	flow	convertion
0	home_page-listing_page	0.738945
1	listing_page-product_page	0.671484
2	product_page-payment_page	0.138297
3	payment_page-confirmation_page	0.350335

Figure 11 conversion rate overview

4.5 Predictive Modeling

The main objective of our model is to predict whether a user will convert to a real customer, so we apply some classification models to predict and choose one or more with good performance to excute this task:Logistic Regression Model, Random Forest Model and (MLP) Multi-Layer Perceptron Model

For analyzing a classification problem like predicting whether a platform user will convert to a real customer, each of these models has its advantages. Logistic regression is simple and interpretable, making it a good choice when the relationship between input features and the target variable is relatively linear. Random Forests excel in handling complex, non-linear relationships and are robust against overfitting, which is beneficial when the data has many features and potential interactions. MLPs are highly flexible and can capture intricate patterns, making them suitable for problems with large and diverse datasets. The final choice depends on the specific characteristics of the data, the interpretability requirements, and the trade-offs between model complexity and performance.

4.6 Result

There are many evaluation measurement for classification model, we use Accuracy and confusion metrics as the evaluation metrics of different models, after parameter adjusting and the testing process, we obtain the following results:

	Logistic Regression	Random Forest	MLP
Accuracy	0.9796	0.9827	0.9812
Precision	0.6176	0.7302	0.6314
Recall	0.2926	0.3544	0.3402
F1	0.3970	0.4772	0.4421

Table 2 Evaluation Metrics

5 Findings and Suggestions

Due to feature engineering, we can get the conversion rate of each stage of the funnel. Then we can analyze the conversion rate comparing with the market average data. Since this dataset does not specify what is the main product category, so we set the overall market average as benchmark.

According to Adobe's latest research, the average conversion rate in Ecommerce Market is 1.64% (Saleh, 2023). And this platform's conversion rate is 2.4%, which shows that this platform performs better than average at the given time range. We analyze gender and device features as cases (Figure 12 and Figure 13)to discuss the differences among different user groups. Other features' visualization and analysis are shown in the attached file.

1. The transition from the home page to the listing page appears to be effective, indicating users' interest in exploring products. However, there is room for improvement to maintain or increase this conversion rate and ensure more users progress through subsequent stages. The drop in conversion from the listing page to the product page suggests potential issues with product visibility, descriptions, or user experience. Optimizing the product page layout, providing clear call-to-action buttons, and enhancing product information can help retain more users through this stage. The most critical concern lies in the transition from the product page to the payment page. Various factors may contribute to this, such as a complicated checkout process, unexpected costs, or a lack of trust in payment security. Managers should conduct user testing and surveys to identify pain points and barriers that deter users from completing their purchases. Even after users reach the payment page, a significant number abandon the process, possibly due to concerns about payment security, hidden costs, or unclear shipping information. Offering reassurances, transparent pricing, and a seamless payment experience can reduce cart abandonment and increase conversions.
2. Female users generally exhibit slightly higher conversion rates at different stages of the conversion flow compared to male users. While this ratio is favorable for a platform with a substantial female user base, the platform should decide whether to focus on developing a female-oriented platform or a gender-balanced platform based on actual business needs. For example, the company can implement gender-specific marketing campaigns to cater to the unique preferences and interests of each gender.
3. According to Adobe's research, the average ecommerce conversion rates by device are 3% for desktop and 2% for mobile. However, considering that mobile users constitute a larger segment of the overall user group, continuous optimization of the mobile user experience is crucial. Ensuring that users can easily search and make purchase decisions on mobile devices will significantly benefit the platform.
4. Based on the metrics and observations, Random Forest emerges as the best-performing model among the three. It achieves high accuracy, precision, recall, and F1 score, making it suitable

as the e-commerce platform's classifier for new users. The high accuracy indicates that it can make correct predictions with a high level of confidence. Additionally, its high precision and recall indicate effective identification of positive cases (e.g., potential valuable users) while minimizing false positives and false negatives. In the future, the platform should implement related methods to facilitate the conversion process. For instance, personalized recommendations, promotional calls, incentives, and discounts can enhance user engagement and conversion rates.

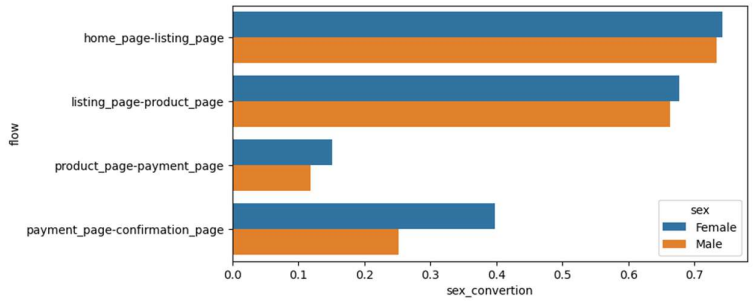


Figure 12 conversion rate by gender

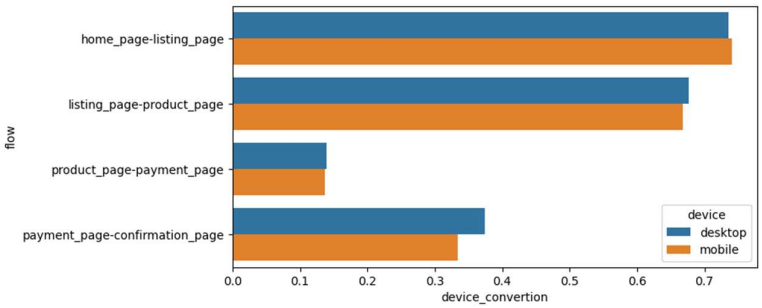


Figure 13 conversion rate by device

6 Limitations and Future works

Firstly, this analysis lacks essential business context, such as specific business goals, promotion plans, market strategies, and product types. The absence of this information hinders a comprehensive understanding of the data analysis results in a real-world business context. Additionally, insights into the target audience are limited, making it difficult to tailor marketing strategies and offerings to specific customer segments. Understanding the target audience's preferences, behaviors, and demographics is crucial for crafting personalized experiences that resonate with potential customers. Furthermore, incomplete business metrics, such as conversion rates for companies in different scales and market segments, and corresponding strategies, restrict the ability to accurately gauge the performance and effectiveness of marketing efforts. In future data acquisition processes, it is essential to incorporate more valuable features to better understand user behaviors and include more industry benchmark metrics to comprehend business data better.

Secondly, the lack of context for decision-making due to the absence of business context and metrics may result in generalized or inconclusive recommendations. Consequently, it becomes challenging to prioritize initiatives or allocate resources effectively. Although the conducted data modeling exhibited commendable performance with favorable indicators such as precision and the confusion matrix, a noteworthy limitation emerged due to the small conversion rate magnitude. Focusing primarily on managing this limited subset of users may not generate substantial business value. To address this constraint and enhance the modeling approach's efficacy, future endeavors should strive to incorporate a more comprehensive set of user behavior data and adopt more appropriate methods for analyzing user conversion rates at each flow or stage.

Moreover, future research should employ more suitable techniques for analyzing user conversion rates at different stages or flows, akin to the model-based prediction made by D'Haen and Van Den Poel (2013). Or even deeper analyzing user behaviors, like applying RFM model. Scrutinizing user actions within specific stages of the customer journey will yield a deeper comprehension of the factors influencing conversion. This finer-grained analysis will enable targeted improvements in each flow, augmenting the overall conversion rate and boosting business outcomes. Additionally, it is crucial to recognize that modern customer management is an ongoing and continuous process. Even if certain users do not immediately convert into customers, they possess the potential to contribute long-term value to the platform. Therefore, future work should adopt a holistic approach that goes beyond immediate conversions and considers the potential long-term benefits users can offer through ongoing engagement and loyalty.

7 References:

1. Managing For Successful Customer Acquisition: An Exploration https://doi.org/10.1362/026725706776861217
2. Levitt 1986 https://www.statista.com/topics/871/online-shopping/#topicOverview
3. Goodwin and Ball (2003)
4. Gartner Glossary Definition of Customer Acquisition - Gartner Sales Glossary
5. Zendesk Blog Customer lifecycle management: Definition, strategy, + 5 stages (zendesk.com)
6. Business and data analytics: New innovations in the management of e-commerce
7. ReallySimpleSystems, https://www.reallysimplesystems.com/blog/user-behaviour-data-for-sales/#:~:text=User%20behaviour%20data%20is%20a%20method%20for%20collecting%2C,do%20it%20and%20predict%20what%20they%E2%80%99ll%20do%20next.
8. S.Akter and S.F. Wamba (2016) Big data analytics in E-commerce: a systematic review and agenda for future research
9. Singh, M., Ghutla, B., Lilo, R., Mohammed, A. F. S., & Rashid, M. A. (2017). Walmart's Sales Data Analysis - A Big Data Analytics Perspective. https://doi.org/10.1109/apwconcse.2017.00028
10. Christoph, S., Felix K., Jorge Marx G. (2021) A Systematic Literature Review on Applying CRISP-DM Process Model
11. Mester, M. (2022, August 21). What is a Sales Funnel? Stages, Examples & How to create one. Email Vendor Selection. https://www.emailvendorselection.com/what-is-a-sales-funnel/
12. Yu, Y., & Cai, S. (2007). A new approach to customer targeting under conditions of information shortage. Marketing Intelligence & Planning, 25(4), 343–359. https://doi.org/10.1108/0263450071075458
13. D'Haen, J., & Van Den Poel, D. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. Industrial Marketing Management, 42(4), 544–551. https://doi.org/10.1016/j.indmarman.2013.03.006
14. Cooper, M. J., & Budd, C. S. (2007). Tying the pieces together: A normative framework for integrating sales and project operations. Industrial Marketing Management, 36(2), 173–182. https://doi.org/10.1016/j.indmarman.2006.03.005
15. Cox, D. J. (2019). The many functions of quantitative modeling. Computational Brain & Behavior, 2(3–4), 166–169. https://doi.org/10.1007/s42113-019-00048-9
16. Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51–59. https://doi.org/10.1089/big.2013.1508
17. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

18. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199
19. Cooper, M. J., & Budd, C. S. (2007b). Tying the pieces together: A normative framework for integrating sales and project operations. Industrial Marketing Management, 36(2), 173–182. https://doi.org/10.1016/j.indmarman.2006.03.005
20. Eitle, V., & Buxmann, P. (2019). Business Analytics for Sales Pipeline Management in the Software industry: A Machine Learning perspective. Proceedings of the . . . Annual Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2019.125
21. Saleh, K. (2023, July 31). The average website conversion rate by industry (Updated 2023). Invesp. https://www.invespcro.com/blog/the-average-website-conversion-rate-by-industry/

8 Appendix

8.1 Summary of the key settings in this work

Practice	Details
Programming language	Python 3
IDE	Jupyter notebook
Data collection	List, Tuple, Set, Dictionary
Control flow	If, elif, else, for loops, while loops
Functions	Built-in functions, Create own functions
numpy	ndarray
pandas	DataFrame/Series
matplotlib	pyplot
seaborn	ggplot-type graphics
sklearn	Data preprocessing, ML model training, validation and evaluation

8.2 Data Transforming Techniques

1. **One-hot encoding**: Each category in a categorical feature is converted into a binary vector. Each category becomes a separate binary feature, with a value of 1 indicating the presence of the category and 0 otherwise.
2. **Label encoding**: It is used for target/response variable in supervised learning tasks. It involves converting categorical target labels into numerical form, enabling the machine learning model to perform prediction tasks more effectively.
3. **Feature scaling/normalization**: Scaling/normalizing numerical features to a specific range ensures that features with different scales do not unduly influence the model.
4. **Feature engineering**: It means creating new features or modifying existing features to improve the performance of machine learning models. This step requires domain knowledge and creativity to derive meaningful information from the data.