Use of Machine Learning for Classification and Regression Models to Forecast Customers' Purchase Intentions

[1] Masoom Jabri, [2] J.A. Paulson,

[1,] PG Student, Department of CSE, Varaprasad Reddy Institute of Technology, Sattenapalli, Kantepudi (Village)

[2] Associate Professor, Department of CSE, Varaprasad Reddy Institute of Technology, Sattenapalli, Kantepudi (Village)

Abstract

In this research, we utilize JD.com as a case study to examine, after considering all the available information, whether people would really place an order for a certain item on an e-commerce platform. It predicts purchases by analyzing a broad variety of relevant characteristics using machine learning techniques. In order to train predictive models, the dataset is preprocessed and designed with features. Classification problems are handled by using important algorithms such as Decision Tree, Random Forest, and Logistic regression.

To achieve maximum performance, hyperparameters are fine-tuned. A high Area Under the Curve (AUC) of 0.9998, indicating a high chance of exact predictions, and an accuracy score of 0.999817, the Random Forest model exhibits outstanding performance in the research. Notably, the coupon discount level and quantity discount level of the product are the factors that contribute most significantly to prediction, making up more than half of the predictive power. This study adds to the literature on e-commerce analytics by providing a data-driven method for studying and forecasting customer behavior; this method may find use in marketing and tailored suggestions.

Keywords-

machine learning, JD, feature importance analysis, and purchase behavior prediction

I. INTRODUCTION

Consumers now have access to an almost infinite range of items and services because to the fast rise of e-commerce, which has transformed the retail business [1]. With their convenient mix of competitive price, tailored suggestions, and user-friendly interfaces, platforms like JD.com have become indispensable to contemporary shopping experiences (Ibid). Researchers and practitioners alike continue to face a formidable obstacle in trying to grasp what motivates customers to buy a particular product, thanks to the dizzying array of alternatives and the intricate nature of the decision-making process. Product quality (e.g., price, qualities, and kind), promotional activities (e.g., discounts and coupons), and user-specific preferences (e.g., browsing history and purchase habits) are a few of the many aspects that impact consumer buying behavior on e-commerce platforms.

There has never been a better chance to apply machine learning to study and forecast consumer purchases than with the abundance of comprehensive data from user interactions now at our fingertips [3]. With the use of these models, platforms may fine-tune their strategies—including customized marketing, inventory management, and pricing policies—to boost user happiness and revenue. Lacking a comprehensive analysis of all relevant variables, most recent research on consumer buying behavior on ecommerce platforms uses rudimentary statistical methods to focus on the effect of individual or uniform elements on purchase choices.

This research overcomes that shortcoming by combining several machine learning models—Decision Tree, Random Forest, and Logistic regression—into a thorough analysis. By filling a large gap in the current literature, we improved the model's performance via sophisticated feature engineering and parameter adjustment. This research

challenges conventional assumptions and offers data-driven insights for refining customized recommendation systems and marketing tactics. It demonstrates the major influence of product discounts on customer purchase choices.

II. DATA

Section A. Dataset The study relies on data acquired from the Institute for Operations study and the Management Sciences (INFORMS), which in turn received it from JD.com as part of the 2020 MSOM Data Driven Research Challenge initiative [4]. All the data pertaining to the transactions that took place in March 2018 is available in this dataset at the transaction level. List of stock keeping units (SKUs), users, clicks, orders, delivery, inventory, and network are the seven tables that make up this database (Ibid). Each one is composed of a unique set of characteristics that may be studied.

B. Selected Original Data Finding the right variables in the skus, users, clicks, and orders tables is the main goal of this study. Table 1 shows the distribution of numerical and categorical variables used to examine the prediction of individual consumption habits using product and user information.

TABLE I. DISTRIBUTIONS OF VARIABLE TYPES

Data Type	Quantity	Percentage
Categorical	6	31.6%
Numerical	13	68.4%

Table 2 lists the specific information of the original variables
selected for the study.

TABLE II. DISTRIBUTIONS OF VARIABLE TYPES

Table	Key Variable	Description
Skus—Describing the characteristics of all	sku_ID	The nique identifier for a product.
31,868 SKUs that belong to a single product category receiving at least one click during March 2018	type	Whether the sku was sold by JD itself or a third-party seller.
	attribute1	One attribute that tells the performance of each sku, taking integer values between 1 and 4.
	attribute2	Another feature that taking integer values between 30 and 100.
Users—Describing the characteristics of all 457,298 users who purchased at least one of the SKUs in the given category during March of 2018	user_ID	A user's unique identification code.
	user_level	The degree of total purchase value of the user in the past.
	plus	Whether the user was a PLUS membership on February 28, 2018.
	gender	User's gender.
	age	The degree of the user's age.
	education	User's education level.
	city_level	The degree of the user's living city.
	purchase_power	The degree of the user's purchase power.
	marital_status	Marriage situation of the user.
Clicks—Establishing the linkage between users and SKUs through their 20 million browsing histories	channel	The platform that a user chose and clicked in to check the sku.
	request_time	The point of time recorded while user made a click.
Orders—Containing 486,928 unique customer orders associated with the focused product category that were placed during the month of March 2018	original_unit_price	Original unit price of each sku.
	direct_discount_per_unit	The direct discount of each sku by the time the user checked it.
	quantity_discount_per_unit	The quantity discount of each sku by the time the user checked it.
	bundle_discount_per_unit	The bundle discount of each sku by the time the user checked it.
	coupon_discount_per_unit	The coupon discount of each sku by the time the user checked it.

C. Investigating Data

Preliminary data research has shown that in March 2018, JD's sku inventory distribution was ordered by its two qualities, as shown in Figure 1.

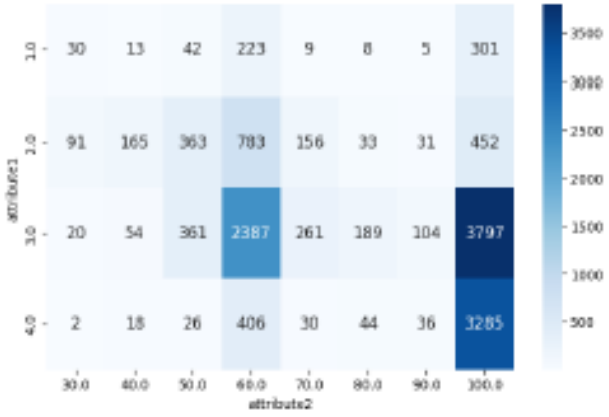

Figure 1. The distribution of skus considering attributes combination.

One interesting thing about this distribution is that it shows a few skus with certain combinations of characteristics being quite popular, including (3.0,100.0), (4.0,100.0), and (3.0,60.0). In line with the specified distribution of attribute values, this suggests that the majority of items are of higher quality. This makes perfect sense, considering the size and popularity of the JD e-commerce platform, and it guarantees that customers will have the best possible shopping experience. Also, it seems that people really like high-quality items, according to this observation. items with the largest inventory also have the highest transaction rate, as seen in Figure 2, which displays the transaction ratios for items with different attribute combinations.

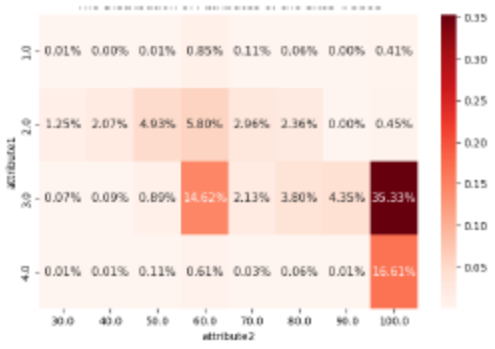

Figure 2. The distribution of skus' trading rates considering attributes
combination.

D. New Variables Created

To enhance the model's construction, this study derived and introduced several new variables. The research aims to reduce the complexity of subsequent models and mitigate the risk of overfitting by categorizing five consecutive numerical variables in the order table into discrete intervals. This approach alleviates the necessity for the model to accommodate minor fluctuations or noise present in the data. Moreover, interval discretization can attenuate the influence of outliers on the model. Direct utilization of raw values can lead to extreme prices or discounts disproportionately affecting model training, whereas interval partitioning can diminish this impact. Intervals also endow the data with a more intuitive interpretation; for instance, recognizing a product's price level as '3' may be more readily comprehensible than perceiving the exact price (e.g., 135), particularly when making strategic, high-level decisions.price_level: In terms of the variable "original_unit_price",on the premise that all original prices have been processed and known, the intervals were appropriately designed based on their distribution. Table 3 shows the distribution of original price intervals

TABLE III. DISTRIBUTIONS OF THE ORIGINAL PRICE

Interval	Frequency
[0,50]	17672
(50,100]	2054523
(100,150]	4085579
(150,500]	1733562
(500,max]	708

Hence, numbers from 1 to 5, indicating low and high, are allocated to the five intervals. The newly created variable "price_level" uses the intervals between the original sku prices to give values to them. The levels of direct discount, quantity discount, bundle discount, and coupon discount. Every one of the four discount kinds went through the same procedure. To begin, we split the initial price by four different ratios; next, we established each level from 1 to 10 based on 0.1, left open, and right closed. For this product, a specific '0' level was assigned to indicate that there is no guaranteed discount. So, "direct_discount_level," "quantity_discount_level," "bundle_discount_level,"and"coupon_discount_level " were the names given to the four additional variables. "weekday_or_weekend" is a field in the "request_time" variable that should be included in the inputs as a judgment on the date's status as a weekday or not. A value of "1" signifies a weekday and a value of "0" implies a weekend. "purchased" is a newly-created variable that serves as both the dependent variable for the final model prediction and a representation of the user's ultimate buying choice (a binary outcome with values of 0 or 1) about the creation of the transaction. The study states that for the same "sku_ID" and "user_ID," any click history within 30 minutes before an order is placed will be combined with the actual order record and marked as a "purchased" status. On the other hand, if there are only click records and no orders appear within 30 minutes, these clicks will be treated as a separate group and all records within this group will be marked as "not purchased." We shall arrange the click

data among two or more orders made within thirty minutes according to the following order records. Rows in groups that include order records will be treated as "1" in this manner, while rows in other groups will be treated as "0". Therefore, a user's buying habits would be evaluated at regular intervals. Processing and Cleaning of Data (E)

Handling missing values: Some variables may have missing values when the original dataset is merged using sku_ID or user_ID. For example, because details like price and discount are only saved when a transaction is initiated, they are not kept when a user chooses not to buy a product. There are two separate cases of missing data that this research deals with. When the percentage of missing values for a certain variable is small, the rows that include such values are removed from the dataset. On the other side, if there is a large amount of missing data, the remaining data from complete instances is used by a Random Forest regression model to impute missing values. The variables "channel", "attributes (1&2)", "type", and "weekday_or_weekend" encapsulate the sku's intrinsic characteristics and the circumstances of its discovery, which this investigation takes to mean that these factors impact the original price and different types of discounts linked with the sku. The research uses these five predictor variables in a Random Forest regression model to fill in the blanks when missing values are known. Then, using the insights derived from the current data, it changes the discount levels and prices that are absent. The user's data is subject to the same procedures. 2) One-hot encoding: This feature engineering approach is often employed to transform categorical variables into a numerical format that is conducive to machine learning models [5]. The main concept is to make a binary vector of length kk for every category, where kk is the total number of categories. Ibid. states that ordinal associations in categorical characteristics may be efficiently eliminated using the one-hot encoding approach. To make sure this non-numerical variable could be easily included into the model, the one-hot encoding approach was used to split "channel" into five dummy variables, as the categorical variable had five categories. Data simplification (number three): All entries in a group would have almost identical information because of the pre-processing, which might lead to duplication. In order to prevent model prediction errors caused by data redundancy and repetition, only the most recent row will be kept. If the group is labeled as "1," then the order row will be retained; if it's labeled as "0," then the click row will be kept. The last dataframe had 7892044 rows when the phase was over.

III. MODELS AND RESEARCH QUESTION

1. Introducing the Model One common supervised learning technique for classification and regression is the decision tree [6]. In order to make predictions, Decision Tree models use a tree-like structure that is created by recursively splitting the data depending on feature values. In order to make predictions, Decision Trees rely on analyzing feature values as they go through the tree structure to the leaf nodes.

2) Random Forest: The Random Forest model is one of the ensemble learning approaches utilized in this study. It builds several decision trees to increase prediction accuracy [7]. The Random Forest model constructs a network of trees by randomly choosing feature and sample subsets, and then averages or votes on the aggregated predictions. The use of an ensemble method improves model generalizability and decreases overfitting (Ibid). Third, there's logistic regression, a linear model often used for binary classification jobs [8]. It works by feeding a logistic function, which converts continuous values to probabilities, with a linear combination of features. To predict the likelihood of an outcome, logistic regression uses a linear combination to weight information and then passes the result via the logistic function.Method for Tuning Parameters (B) The parameters are tuned in this study using the Random Search method: Determining an appropriate range and step size for each important model parameter, and then utilizing its hyperparameter optimization technique to determine which values may provide the best accuracy score. By focusing on random combinations, which may land in highperforming portions of the search space, this tuning approach is efficient in high-dimensional spaces and frequently discovers near-optimal solutions quicker than grid search [9]. It avoids analyzing every combination of parameters. C. Question for Research Predicting, given all the information provided,

whether a user will make an order for a certain item they are browsing is the main emphasis of this research. To categorize purchases, machine learning methods like logistic regression, Decision Tree, and Random Forest are used.

The whole dataframe was divided into a training set and a testing set using a seed value of 21, at a ratio of 0.6 to 0.4. Following the split, there are 4,735,226 data points in the training set and 3,156,818 data points in the testing set. The testing set includes 22 independent characteristics and a dependent variable("purchased"). The random seed is the same for all three models that use this train-test set.

IV. RESULTS AND DISCUSSION

Results of the Models' Predictions (A) Using these three models with their default settings doesn't provide very nice results, so we had to adjust our tuning process. This worked like a charm, and now we have much more accurate models and outputs. The work addressed overfitting in the models by using suitable regularization approaches. The penalty term was adjusted in the context of logistic regression to limit the complexity of the models. Overfitting owing to too complicated model architectures was avoided for the tree-based model by meticulously monitoring the tree depth and the number of leaf nodes. To further guarantee the model's resilience and generalizability on new data, model validation was carried out utilizing a separate test set that had not been used during training or previous validation rounds. The detailed outcomes of the three categorization methods are shown in Table 4. The assessment results show that the Random Forest model achieves near-perfect accuracy and AUC values, making it the best model out of the three.

TABLE IV. DISTRIBUTIONS OF THE ORIGINAL PRICE

Model	Accuracy Score	Precision	Recall	AUC
Decision Tree Classifier	Train: 0.999914 Test: 0.999863	0.999259	0.996749	0.9991
Random Forest Classifier	Train: 0.999965 Test: 0.999871	0.999306	0.996943	0.9998
Logistic Regression	Train: 0.988523 Test: 0.988259	0.959085	0.687056	0.9296

The Importance of Features (B) Decision Tree and Random Forest models were analyzed using feature significance statistics in the research. Table 5 shows the six most important factors that affected the two models' prediction abilities.

TABLE V. FEATURE IMPORTANCE RANKING IN THE RANDOM FOREST AND DECISION TREE (TOP 6)

Random Forest		Decision Tree	
Variable	feature_importance	Variable	feature_importance
coupon_discount_level	0.391027	coupon_discount_level	0.714901
quantity_discount_level	0.339364	quantity_discount_level	0.143351
direct_discount level	0.094437	attribute2	0.041032
attribute2	0.035272	attribute1	0.039060
bundle_discount_level	0.025207	direct_discount level	0.026816
attribute1	0.024960	channel_1	0.007692

The chart shows that the two most significant characteristics in both the Random Forest and Decision Tree models, accounting for more than half of the overall significance, are coupon discounts (0.3910 / 0.3394) and quantity discounts (0.7149 / 0.1434). The power of discounts to influence consumer spending is one possible explanation for this phenomena.

One way coupons might make a purchase seem more appealing is by reducing the amount actually spent on products. There is a correlation between steep discounts and increased purchasing. Also, people may choose to purchase in bulk when there are discounts for bigger amounts so they may save money overall.

The availability of discounts could cause customers to change their shopping decisions, even if they hadn't planned to buy more things to begin with. Section C: Analyzing Regression Coefficients In addition, the study obtained additional parameters from Logistic Regression using statsmodel. The results of the regression analysis are shown in Table 6, which also includes the positive and negative coefficients for the important factors that affected buying behavior.

TABLE VI. PARTIAL OLS REGRESSION RESULT OF THE KEY VARIABLES

Variables	Parameters				
	coef	std err	P>|z|	[0.025	0.975]
coupon_discount level	-3.9269	0.007	0.000	-3.940	-3.914
quantity_discount level	-0.4053	0.003	0.000	-3.424	-3.383
attribute1	0.3990	0.005	0.000	0.390	0.408
attribute2	0.0117	0.000	0.000	0.011	0.012

The "coupon_discount_level" variable has a confidence range of [-3.940, -3.914] and a coefficient of -3.9269. This indicates a strong and substantial negative impact ($p < 0.001$). With a coefficient of -0.4053 ($p < 0.001$) and a confidence range of [-3.424, -3.383], "quantity_discount_level" reveals a less significant negative effect. Results show that discounts had the opposite impact in the regression model as predicted.More specifically, there is a correlation between a rise in discount coupons and a decline in the desire to buy from consumers. Reasons for this phenomena might include customers becoming hesitant to buy after seeing steep reductions. the product's worth or notice a decline in its quality. In comparison to coupon discounts, the negative impact of "quantity_discount_level" is not as severe. This shows that whereas quantity discounts might encourage more purchases, they can also make people think that purchasing too much is wasteful. With a coefficient of 0.3990, "attribute1" has a strong positive effect on consumers' propensity to buy the goods. This trait probably stands for a practical benefit of the product, which increases the likelihood of consumers buying it. Similarly, "attribute2" influences consumers' propensity to buy, but to a less degree (coefficient = 0.0117). This impact might be related to an underappreciated secondary product feature or extra value.

V. CONCLUSION

Utilizing rigorous feature engineering approaches and multidimensional user and product variable data, this work constructs a complete dataset for forecasting consumer purchase behavior on the JD e-commerce platform. For the purpose of predictive analysis, the study employs three traditional categorization models: Logistic regression, Decision Tree, and Random Forest. The Random Forest model attained the maximum prediction accuracy of 0.999871 after thorough parameter adjustment and testing, highlighting its usefulness in solving comparable challenges. Product discounts, contrary to popular belief, play a pivotal part in the research since they greatly ease client purchase choices, according to the statistical analysis of the variables impacting forecasts. Having said that, mindless sales do not always cause people to buy on the spur of the moment. It should be noted that there are still issues with this research that need to be explored further. To begin, instead of starting at 1 and increasing in a linear fashion, the value of characteristic 2 begins at 30 and continues up to 100, with an addition of 10 for each number. Because of this, accuracy might be compromised due to higher fluctuations and uneven data dimensions. In order to maintain the data's natural scale-related properties, especially between attributes 1 and 2, this article did not normalize some variables in the original dataset that could have unknown specific assignment implications. In addition, the first data analysis showed that over half of the inventory is comprised of only two distinct combinations of sku features. Overfitting may occur if the model incorrectly interprets certain characteristics because of this. Future research must prioritize addressing and improving this issue. Future research in this area will also try to add users' locations or logistics warehouse sites into the prediction model in order to make it more comprehensive. Applying the model to real-time, large-scale datasets may present difficulties related to processing resources and reaction times, despite its excellent performance in controlled experimental conditions. Therefore, in order to enable decision-making in real-time, future studies should focus on improving and implementing these models inside efficient platforms, including cloud computing environments. The goal is to maintain a high level of predicted accuracy while adding additional data variables to better understand and cater to particular customer preferences. Improving user happiness and platform income are the goals of this method, which seeks to guarantee the accuracy of suggestions and adverts. The work's ultimate goal is to help improve e-commerce platforms' recommendation algorithms.

REFERENCES

[1] M. Cao, "Competitive Advantages and Challenges of E-Commerce," Adv.Econ. Manag. Polit. Sci., vol. 47, pp. 114-119, December 2023.

[2] S. C. S. C. Yong, R. T. Huan, W. S. Poh, M. Osman, and D. C. W. Ng, "Assessing the Factors Influencing Consumer Behaviour in E-Commerce Platforms," Int. J. Multidiscip. Appl. Bus. Educ. Res., vol. 4, pp. 3725-3735, October 2023.

[3] Y. Zhang, "Utilizing machine learning algorithms for consumer behaviour analysis," Appl. Comput. Eng., vol. 49, pp. 213-219, March 2024.

[4] M. Shen, C. S. Tang, D. Wu, R. Yuan, and W. Zhou, "JD.com: Transaction-Level Data for the 2020 MSOM Data Driven Research

Challenge," Manuf. Serv. Oper. Manag., vol. 26, pp. 2-10, February 2024. [5] A. Biswas, "Feature Engineering: Categorize your data using One-Hot Encoding," Medium, June 2024.

[6] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," Front. Artif. Intell., vol. 6, July 2023.

[7] N. Donges, "Random Forest: A Complete Guide for Machine Learning," BuiltIn, March 2024.

[8] J. Brownlee, "Logistic Regression for Machine Learning," Machine Learning Mastery, December 2023.

[9] M. H. Hassan, "Tuning Model Hyperparameters With Random Search," Medium, November 2023.