DOI: 10.15439/2022F167
ISSN 2300-5963 ACSIS, Vol. 30

An End-to-end Machine Learning System for Mitigating Checkout Abandonment in E-Commerce

Md Rifatul Islam Rifat[1], Md Nur Amin[2], Mahmud Hasan Munna[3], and Abdullah Al Imran[4]

[1,3]Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh
Email: irifat.ruet@gmail.com, munna.ete15@gmail.com

[2]University Jean Monnet, Saint Etinne, France
Email: nuramin.aiub@gmail.com

[4]American International University-Bangladesh, Dhaka, Bangladesh
Email: abdalimran@gmail.com

***Abstract*—Electronic Commerce (E-Commerce) has become one of the most significant consumer-facing tech industries in recent years. This industry has considerably enhanced people's lives by allowing them to shop online from the comfort of their own homes. Despite the fact that many people are accustomed to online shopping, e-commerce merchants are facing a significant problem, a high percentage of checkout abandonment. In this study, we have proposed an end-to-end Machine Learning (ML) system that will assist the merchant to minimize the rate of checkout abandonment with proper decision making and strategy. As a part of the system, we developed a robust ML model that predicts if someone will checkout the products added to the cart based on the customer's activity. Our system also provides the merchants with the opportunity to explore the underlying reasons for each single prediction output. This will indisputably help the online merchants in business growth and effective stock management.**

***Index Terms*—E-Commerce, Checkout Prediction, Checkout Abandonment, Decision Support, Explainable AI (XAI), LIME**

I. Introduction

As we are living in an era where digitalization and technology are evolving day by day, our dependency on the internet has noticeably increased. E-commerce has made shopping easy and safe for all internet users all over the world. People nowadays prefer exploring websites to find their daily needs rather than walking around shopping malls, supermarkets, and shops. They do not have to take the hassle of finding a product and waiting for a long billing queue, which makes purchasing simple and quick. On the contrary, e-commerce also makes it easier for companies to reach out to new customers all over the world. A report from Statista [1] shows that global sales have jumped from 1,336 billion to 5,542 billion USD in the last 6 years in the e-commerce industry. In the near future, undoubtedly the dependency on online shopping will increase significantly.

Recently, the online retailers are encountering numerous business challenges such as the lack of trust, customer churn, product return and so on. With the rapid technological advancement in the data science domain, researchers have already started to solve these type of problems by utilizing data science approaches. Some of the existing research works are related to the product review classification such as the authors in [2] proposed DNN networks to train a classifier for identifying the product quality from product reviews. One of the major problems in e-commerce industry is the return of the product. In [3] and [4], the researchers tried to address this issue by using different predictive modeling techniques.

Apart from these, one of the most common business challenges in the e-commerce industry is high checkout abandonment rate. According to the study of Baymard Institute, a research institute in Denmark, the average checkout abandonment rate is 69.82% [5]. Also during the COVID-19 pandemic, online shopping behaviors have been significantly changed. When individuals browse an online store for a particular product, as a natural consequence, they often add many additional items to their cart. Among the added items in the cart, some are in need and the others may be their favorite but are not in great demand, and most of the time the majority of them are never checked out which results in high checkout abandonment rate. However, very few researchers have contributed to solve the issues related to online shopping carts. Jian et al [6] proposed a framework to predict buyers' repurchases intention from the cart information. In another study [7], the author built a recommendation system using the shopping cart information.

In this research, we have tried to address the aforementioned business problem of the e-commerce industry and proposed an end-to-end ML system that will automatically perform all the steps such as data collection, transformation, preprocessing, statistical analysis, and predictive analytics. In the case of predictive analytics, we have conducted an extensive experiment and found CatBoost as the outperforming approach that predicts the checkout possibility of users with the highest accuracy (=0.76) and precision (=0.694). Moreover, we have applied a model agnostic local explanation approach for the explainability that will help the e-commerce merchants to analyze how every single customer gets influenced by different factors.

Our contribution to this study can be considered from two perspectives: one from a research standpoint, and the other from a commercial standpoint. The research perspective is that no previous study attempted to solve this particular problem and proposed any end-to-end ML based solution. On the other hand, if business people conduct targeted marketing or apply other business strategies to consumers who are most likely to purchase the products in the cart, then the sales will be increased. Apart from that, the prediction will assist the merchant in maintaining effective stock management as well. Indisputably, the combination of statistical insights, predictive output, and local explanation aid the seller in developing proper strategies for business growth.

II. Proposed Solution

As a solution to the checkout abandonment issue in e-commerce business, in this phase, we have proposed an end-to-end system shown in Fig. 1.

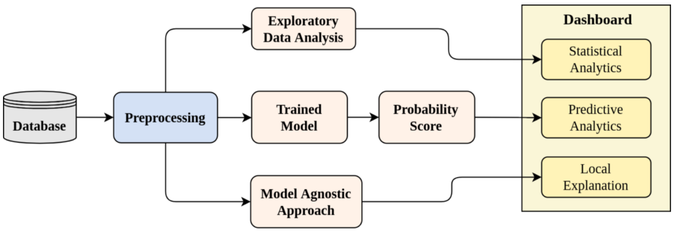

Fig. 1. System diagram

From Fig. 1 it can be observed that our proposed system takes raw data from the database and outputs analytical and predictive insights to a dashboard. Development of this system is composed of several steps: data understanding, preprocessing, exploratory data analysis, data modeling, and a model-agnostic approach.

To conduct the experiment, a large real dataset has been collected from a prominent SaaS platform that integrates with online stores to track behavior in real-time. Our dataset contains 27 features in total, where 21 features are numerical and the other 6 features are categorical. Table I and Table II shows the description of all the numeric and categorical features respectively. Among all the instances(=28410), 55% instances belong to class 0 (‘not checked out’) and the other 45% belongs to class 1 (‘checked out’) which indicates that the dataset is almost balanced.

After collecting the data, in the preprocessing phase, we have applied the James-Stein encoder to convert the categorical features into informative numerical representations. The mathematical expression of the James-Stein encoder is as follows:

$$JS_i = (1 - B) \times mean(y_i) + B \times mean(y) \tag{1}$$

Where

$$B = \frac{var(y_i)}{var(y_i) + var(y)} \tag{2}$$

The idea of the James-Stein encoder is to shrink the category’s mean target towards a more median average.

As most of the features do not fall into a Gaussian distribution, in this experiment, we have applied min-max scaling to scale the features from 0 to 1.

TABLE I
Data Description of Numerical Features

Feature Name	Description
visited_cart	How many times did the user visit the “Cart” page so far in the current visit.
total_add_cart	How many products did the user add to the shopping cart.
total_clicked_products	How many times did the user click on products or visit the “Product Page” of products.
session_length_steps	Length of the visit in steps/events.
session_length_sec	Length of the visit in seconds.
mean_viewed_price	Mean price of clicked products.
max_viewed_price	Max price of clicked products.
min_viewed_price	Min price of clicked products.
sum_viewed_price	Sum price of clicked products.
total_orders	Total orders the user checked out so far.
total_purchased_sum	Total sum of orders the user checked out so far.
total_visits	Total previous visits of the users.
returning_visitor	Is it a new or returning user.
customer_since_days	Total minutes since user’s first visit.
cart_sum	The sum of the shopping cart the user has/sees it right now.
cart_total_prd	Total products the user has in the shopping cart.
cart_total_sale_prd	Total discounted products the user has in the shopping cart.
cart_total_prd_in_sale	The ratio of discounted products vs non discounted products in the cart.
cart_total_saved	The amount of money the user is saving due to the presence of sale products in the cart.
cart_sum_without _discount	Total sum of the cart on original price of the products.
cart_total_saved_%	Total percentage of saved money in the cart.

TABLE II
Data Description of Categorical Variables

Feature Name	Description
landing_page	The page where the user started the visit.
week_day	Day of the week
origin	From which origin did the user land on the store.
utm_source	From which source did the user land on the store.
utm_medium	From which medium did the user land on the store.
device	The user’s device type.

Then, during the data modeling phase, we have followed a proper and systematic workflow that has been illustrated in Fig. 2. This workflow has been started just the completion of the data preprocessing steps. At first, we have segregated the entire processed dataset into training (=80%) and validation (=20%) data to eliminate biases during the model evaluation phase. Then, in the second stage, we have chosen the ML algorithms based on the objectives as well as the characteristics of the data. In this study, we have applied 5 SOTA algorithms such as XGBoost [9], LightGBM [10], CatBoost [11], mGBDTs [13], and TabNet [12] not only targeting the best prediction output but also with a special focus on interpreting the models. To create a baseline performance, we have also included a DNN model.

Then in the evaluation phase, we have applied 5 evaluation metrics namely Accuracy, Precision, Recall, $f_{0.5}$-score, and ROC-AUC. The fourth and most important step is to tune the hyperparameters of the model very attentively to obtain the best prediction output without over-fitting or under-fitting. For

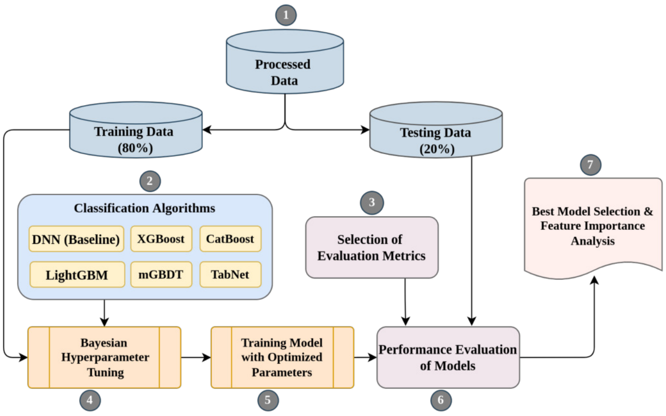
1
Processed Data
Training Data (80%)
Testing Data (20%)
2
Classification Algorithms
DNN (Baseline)
XGBoost
CatBoost
LightGBM
mGBDT
TabNet
3
Selection of Evaluation Metrics
7
Best Model Selection & Feature Importance Analysis
Bayesian Hyperparameter Tuning
4
Training Model with Optimized Parameters
5
Performance Evaluation of Models
6

Fig. 2. Flow-diagram of Data Modeling

TABLE III
MODEL PERFORMANCE ON TRAINING AND TESTING DATA.

		ROC -AUC	Precision	Recall	$f_{0.5}$ Score	Accu-racy
DNN	Train	0.565	0.586	0.972	0.858	0.606
	Test	0.565	0.585	0.971	0.636	0.606
XGBoost	Train	0.798	0.742	0.778	0.771	0.802
	Test	0.753	0.689	0.729	0.720	0.758
CatBoost	Train	0.804	0.755	0.779	0.774	0.809
	Test	0.755	**0.694**	0.725	**0.719**	**0.760**
LightGBM	Train	0.782	0.724	0.759	0.752	0.786
	Test	0.746	0.680	0.723	0.714	0.751
TabNet	Train	0.757	0.688	0.746	0.733	0.739
	Test	0.730	0.655	0.722	0.702	0.723
mGBDT	Train	0.793	0.743	0.764	0.760	0.798
	Test	0.733	0.673	0.694	0.690	0.741

tuning the hyperparameters, the Bayesian approach has been chosen in this experiment since it takes less time compared to others to find the best set of parameters and improves generalization performance on the test data. After getting the optimized hyperparameters for all of the models, in the fifth stage, we have trained our models using the training dataset. With the completion of the training phase, we went on to the sixth stage, in which we have utilized the trained models to make predictions on the unseen validation dataset and track their performance against each evaluation metric. Then, in the following step, we have compared the performance among the models to figure out the best model for this particular business problem.

III. Best Model and Features Selection

In this phase, we have divided our analysis into two parts. We have started the analysis by explaining the performance of each model and selecting the best model from their comparison. Finally, we have explored the top features of our best model to find insight that can help to make important business decisions.

Table III shows both of the training and testing results for each of the models of our experiment. From the Table III, it can be observed that our obtained test results are very close to the training results, which indicates that the model has learned the underlying patterns well from the data without over-fitting. By considering the business problem we were trying to solve, we have mainly focused on Precision, $f_{0.5}$-score, and Accuracy. Significantly, it has been appeared that all the models have been performed better than our baseline DNN model in terms of Precision, $f_{0.5}$-score, and Accuracy. Although TabNet pretrained model performs well for tabular data, in our case it fails to outperform the tree-based models. Similarly, the mGBDT fails to outperform the other non-differentiable boosting algorithms. All of the tree-based boosting algorithms: XGBoost, Catboost, and LightGBM have yielded the results close to each other. Comparing the results of each models, it can be seen that CatBoost has consistently outperformed all other models with accuracy (=0.76) and precision (= 0.694). As the conclusion of all comparisons, we chose CatBoost as the best performing model for checkout prediction.

Furthermore, we have extracted the most effective 5 features from the CatBoost classifier. Table IV shows the best features in descending order based on the feature importance.

TABLE IV
FEATURE IMPORTANCE (TOP 5) OF CATBOOST CLASSIFIER

Rank	Features	Importance(%)
1	total visits	14.09
2	customer since days	13.56
3	total purchased sum	6.46
4	max viewed price	5.90
5	min viewed price	5.75

TABLE V
EXAMPLES FROM VALIDATION DATA FOR LOCAL EXPLANATIONS.

Features	Instance 1	Instance 2
visited_cart	1	1
total_add_cart	1	2
total_clicked_products	6	2
session_length_steps	38	5
session_length_sec	1564.05	111.54
mean_viewed_price	213.62	48.9
max_viewed_price	239	48.9
min_viewed_price	169.9	48.9
sum_viewed_price	1281.7	97.8
total_orders	0	1
total_purchased_sum	0	185
total_visits	10	1
returning_visitor	1	1
customer_since_days	0	1704.9
cart_sum	239	97.8
cart_total_prd	1	2
cart_total_sale_prd	0	2
cart_total_prd_in_sale	0	100
cart_total_saved	0	42
cart_sum_without_discount	239	139.8
cart_total_saved_%	0	30.04
landing_page	home page	home page
week_day	wednesday	thursday
origin	google	google
utm_source	unknown	google
utm_medium	unknown	cpc
device	mobile	desktop
checkout_status	not-checkout	checkout

IV. Model Explanation and Decision Support

To make our model's decision more transparent, in this study, we have built a support system using the model agnostic local explanation technique, LIME [14]. This method will assist us in comprehending the factors that influence a complex black-box model around a single instance of interest.

In the following Table V, we have taken two instances from the unseen validation data for local explanations. Instance 1 has a checkout-abandonment status, and Instance 2 has a checkout status. The prediction and explanations for these examples can be found in Fig. 3 and 4.

Fig 3 illustrates the decision rules and feature significance based on which the CatBoost model made the decision for instance 1, which was actually abandoning the checkout after adding items to the cart. We can observe that the model predicts with a 95% probability that this person will abandon the checkout. Also, the aforementioned 3 most important features: "total visits", "customer since days", and "total purchased sum" significantly influenced the model to decide in favor of checkout abandonment.

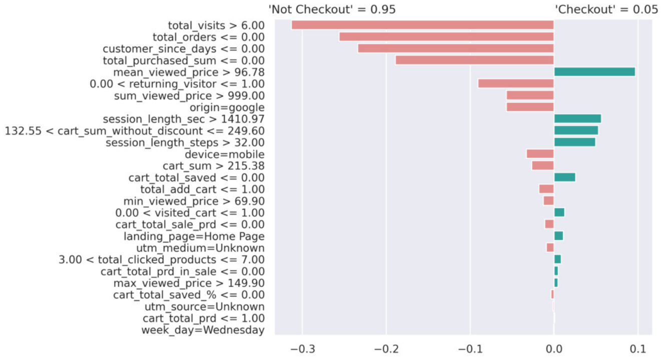

Fig. 3. Explanation of instance 1.

In Fig 4, we can observe that the model predicts with 99% probability that instance 2 will checkout the added item which is actually correct. From the value of "customer since days" feature, it is obvious that being the old customer gave the model confidence that instance 2 will checkout the products.

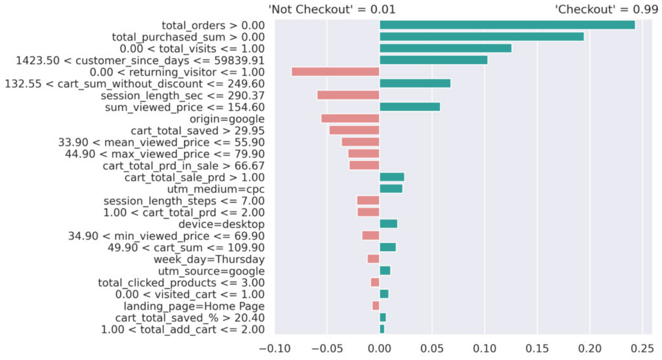

Fig. 4. Explanation of instance 2.

The aforementioned illustrations of local explanations of predictions for individual examples can immensely assist business analysts or decision-makers in making meaningful and unbiased decisions. Especially, this explanation technique can especially assist in making decisions in confusing situations where it is difficult to decide whether a person will checkout the items from the cart or not.

V. Conclusion

In this study, we have aimed to minimize the checkout abandonment rate, which is a key concern in modern e-commerce business, by proposing an end-to-end system. One of the most important components of the system is the ML model that predicts the probability of checkout abandonment for each of the customer. Also, it provides the explanation of the decision taken by the model for further business support. In case of predictive analytics, the CatBoost was found as the best performer with 0.694 (Precision), 0.719 ($f_{0.5}$-score), and 0.76 (Accuracy). For reliable decision making with additional support, we have integrated the LIME, that interprets the model's output as well as extract the decision rules.

References

[1] Staista,https://www.statista.com/statistics/379046/ worldwide-retail-e-commerce-sales. Last accessed 8 Apr 2022

[2] M. H. Munna, M. R. I. Rifat and A. S. M. Badrudduza, "Sentiment Analysis and Product Review Classification in E-commerce Platform," *2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020*, pp. 1-6, DOI: 10.1109/ICCIT51783.2020.9392710.

[3] Al Imran, A. and Amin, M.N., 2020. Predicting the return of orders in the e-commerce industry accompanying with model interpretation. *Procedia Computer Science, 176*, pp.1170-1179. DOI: 10.1016/j.procs.2020.09.113

[4] Urbanke, P., Kranz, J. and Kolbe, L., 2015. Predicting product returns in e-commerce: the contribution of mahalanobis feature extraction.

[5] Baymard Institute, https://baymard.com/lists/cart-abandonment-rate. Last accessed 8 Apr 2022

[6] Mou, J., Cohen, J., Dou, Y. and Zhang, B., 2017. Predicting Buyers'repurchase Intentions in Cross-Border E-Commerce: A Valence Framework Perspective.

[7] Budnikas, G., 2015. Computerised recommendations on e-transaction finalisation by means of machine learning. Statistics in Transition. *New Series*, 16(2), pp.309-322.

[8] Cox, N.J., 2010. Speaking Stata: The limits of sample skewness and kurtosis. *The Stata Journal*, 10(3), pp.482-495. DOI: 10.1177/1536867X1001000311

[9] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).DOI: 10.1145/2939672.2939785

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

[11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

[12] Arik, S.Ö. and Pfister, T., 2021, May. Tabnet: Attentive interpretable tabular learning. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 8, pp. 6679-6687).

[13] Feng, Ji, Yang Yu, and Zhi-Hua Zhou. "Multi-layered gradient boosting decision trees." *Advances in neural information processing systems 31 (2018).*

[14] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (2016) DOI: 10.1145/2939672.2939778