National College of Ireland

Conversion Rate Optimization in E-commerce: Implementation of ML Algorithms to identify clickstream patterns

MSc Research Project
Artificial Intelligence for Business

Tooba Khan
Student ID: 23153768

School of Computing
National College of Ireland

Supervisor: Prof.Dr. Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Tooba Khan

Student ID: 23153768

Programme: MSc AI for Business **Year:** 2024

Module: Thesis

Supervisor: Prof.Dr. Muslim Jameel Syed

Submission Due Date: 12-Aug-2024

Project Title: Conversion Rate Optimization in E-commerce: Implementing ML algorithms to identify clickstream patterns.

Word Count: 7862 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Tooba Khan

Date: 12-Aug-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	

Date:	
Penalty Applied (if applicable):	

Conversion Rate Optimization in E-commerce: Implementing ML algorithms to identify clickstream patterns.

ToobaKhan
23153768

Abstract

Today's digital revolutionization in the world has immensely changed the landscape of businesses, specifically the retail segment with a 360-paradigm shift. The digital evolution in the E-commerce has essentially redesigned the commercial operations, disregarding conventional brick-and-mortar operations with digital platforms. It has facilitated continuous exchange of any form of information (transactional or informative) between buyers and sellers via online channels. Analyzing online behavior of users on websites has been a key aspect for idealizing of business performance metrics and growth. The major gap addressed in this research, is the behavioral patterns identification of website users, while using clickstream dataset from E-commerce fashion store to optimize conversion rate. Since, existing techniques from industry have not been enough to identify the factors relating link between clickstream pattern and user purchase predictions. This model intends to help e-commerce businesses to analyze and evaluate their conversion rate strategies. Relevant machine learning methodologies have been employed on clickstream data from an e-commerce (fashion store) to analyze the clickstream pattern from online sessions and their purchase intent.

Keywords: Clickstream, Intent identification, Behavioral analysis, Conversion rate, Machine Learning Algorithms, Imbalance class handling

1 Introduction

Any kind of business proceedings between customers and commercial entities via digital medium comes under the umbrella of E-commerce domain (Stair and Reynolds, 2008). In the notion of E-commerce, the prime objective of businesses is to amplify their growth and maximize the revenue. Since, the conversion rate of any website requires more systematic inputs of businesses (domain experts) to increase the rate of desired action taken by website visitors, such as add to cart, add promo or make purchase (depending on the business objective or targets). Whereas, the understanding and identification of complex user behaviors have been an intricate challenge. The extensive research and precedent analytical tools has provided opportunities to dig deeper in user interactions and drive actionable insights with the use of comprehensive datasets available.

Clickstream are the path of clicked actions (timestamped server-side traces) by users in entire web browsing sessions (Hosseinmardi etal., 2014). To target informed CRO interventions, the identification of clickstream behaviors shape the pattern to predict non-conversion users on

website (Kohavi, Tang, and Xu, 2023). While operating online it is becoming more significant for businesses to focus on user journey. To evaluate website satisfaction, user experience is a major matrix for clickstream analysis; and this has been imperatively important for professionals and strategy makers (Cai et al., 2018). To capture the complex links between online user activities and outcomes from online sessions existing methodologies had frequently failed, which lightens the issues for visualizing complex clickstream behavior (Lie et al., 2017). Whilst, the significance has been acknowledged that the effective visualization of these connections between conversion rate and clickstream has not been visualized effectively. According to Raphaeli et al. (2017), three crucial elements are required to understand online behaviors: sequence of events, durations of session and events (action taken by users). This study focuses on the identification of non-converting behavioral patterns through clickstream dataset. While focusing on transaction, product, customers and clickstream data from an e-commerce store, to link the metrics with user behavior. This research aims to amplify the user journey understanding and enhance CRO strategies.

1.1 Research Motivation:

The motivation behind this research is the significance of conversion rate optimization while understanding the user behavior on E-commerce platforms. Even the minor improvement of conversion rates can be highly competitive in increase of online retail revenue. In this study, the gap between clickstream behavior of user and website metrics are aimed to be bridged, by identifying patterns that can capture more effective conversion rate optimization (CRO) strategies. The datasets; product, transaction, customer and clickstream provide detailed view of user journey on E-commerce platform. These also contribute in identifying the key influencing factors of conversion outcomes. Findings from this analysis have potential to enhance e-commerce strategies significantly. By clickstream data, the exact behavioral actions (pinpoints) on websites that lead to track conversion funnel. In addition, (Xiao et al., 2022) discussed that temporal factors such as time of access and session durations play a crucial role in user behavior. Through which businesses can improve their action plans to increase conversion rate. This can cover optimizing website design, refining in marketing strategies, and personalizing of website on identified behavioral patterns. However, previous studies have explored several factors resulting in change of user behaviors, but there is a potential gap between its linking with website temporal and static features. This study aims to address that gap by delivering detailed analysis insights on influencing factors for conversion rate. With the implementation of advanced analytical techniques and diverse datasets.

1.2 Research Objective:

The identification of non-converting clickstream behavioral patterns is the primary objective of this analysis, to link the relation between clickstream behavior and website performance is the challenge. Considering the factors that will remain constant, following factors has been underlined in this research such as, number of views per page, new visitors, bounce rate and exit pages etc. According to Purnomo, (2023) such factors enable the user interactions website performance, which influences the conversion probabilities. Similarly, the sequences as page

views, time based factors, dates and session duration. Furthermore, the user actions pattern includes within a session; event sequence, click path and conversion funnel. Kohavi et al., (2023) discussed the mapping of behavioral pinpoint of users journey indicates the non-conversion visits and these can improve the conversion probabilities. It was a research gap, and that provides an opportunity for future researchers to explore how the machine learning methods can be utilized to illustrate further in identifying complicated connections between conversion rates and clickstream behavior. By analyzing how ML techniques can be utilized to identify and visualize the clickstream data patterns, this study aims to address these gaps that will provide insights for more effective conversion rate optimization (CRO) strategies by providing an understanding of user behavior on ecommerce platforms. Through these insights, business experts would be able to manage outcomes on relationships identification between CRO prediction of user's visit and clickstream behaviour.

Research Question:

How to optimize conversion rate using machine-learning algorithms, while utilizing clickstream patterns to evaluate overall website performance in e-commerce?

2 Related Work

2.1 Conversion Rate Optimization (CRO):

For growth and sustainability of businesses, data driven decision-making is significantly important for strategic and valuable insights. In E-commerce, the centered task is the make a purchase successful (a presumed action to be taken by the user), the success and achievement of these focused tasks is referred to as conversion (Gudigantala et al., 2016). According to Crystallize (2023) conversion rate, directly affect the revenue of business by accessing the higher percentage of converting session, through optimization of behavioral factors and website KPIs as, page load performance, navigation of click paths etc. Research concluded business revenue could be increase by adjustments in CRO strategies and less expenditures on marketing of E-commerce store. Conversion rate has always been in the center of action by business professionals and policy makers, which results in leading to the profits and sales for business. However, some of users visit websites only for browsing and price comparison, without any intend to make a purchase. This type of user visits on websites are considered as non-converting behavioral patterns (Raphaeli et al., 2017).

Conversion rate is the simple ratio of online user sessions converting in monetary transactions or planned action to be taken by user in comparison of the total number of sessions on a website; this varies in different industries and regions. In addition, Wen et al, (2023) highlighted the significance of real time clickstream data processing through predictive machine learning model, which analyzed that the engagement behavior of user on site has helped model to capture non-converting customers. Although, Kaushik et al., (2017) explored website analytics of a online retailer, focusing on user interaction which resulting converting sessions. Which shown that optimizing of few metrics as checkout processes, page performance in landing on page

time and product descriptions through continuous A/B testing and real-time analytics could improve conversion and increase the customer retention as well. For businesses, they can reconsider their strategies as segment targeting, advertising strategy and other factors as well. Although, for improvement in conversion rate clickstream data has helped industry experts in consumer experiences (UX) domain, through providing user feedback guidance for iterative modifications to websites (Narang et al., 2017). Even, (2019) used unsupervised machine learning to improve the user experience and design recommendation. Similarly, several practitioners have used machine-learning models for identification of the purchase or no purchase likelihood patterns (Szabó and Genge, 2021).

2.2 Clickstream analysis:

To obtain behavioral feedback from user professionals use clickstream behavioral patterns visualization as a tool for getting valued insights from buyer sessions data (Liu et al., 2017). It consists of all search paths in session, which is an influential source of information for online consumer behavior (Bucklin and Sismeiro, 2003). Whereas, Bigon et al., (2019) used machine-learning methods for effectively identifying the clickstream behavioral patterns to forecast conversion rate. Similarly, Sakalauskas et al., (2024) used clickstream data for identification high-valued customer for implementing and designing of personalized advertising strategies. Clickstream insight driven advertise resulted in improved conversion rates and reduced marketing costs, predominantly for high-value customers. Since, in e-commerce effective methods depend on solving the conversion prediction problem, which emphasize on whether a visitor on site will return to make a purchase later or else will make purchase in current session.

Since, clickstream data is capable of tracking online touchpoints (performed clicks and event) of consumers (Li and Kannan, 2014). While, the success rate of conversion from first site visit is only 2% of consumers execute a complete purchase, it is important to comprehend such consumer behaviour and conversion rates in order to alter marketing campaigns, which is a pivotal factor for retargeting efforts. Collaborative modeling was recommended as a strategy that incorporates both patterns; as compared to current techniques, that only focuses on customer product-level conversion patterns (Hanamanthrao & Thejaswini, 2017).Also, Montgomery et al. (2004) discussed the web design alterations could also increase the conversion rate of website performance, with the execution of sequential analysis from page-level clickstream data of online booksellers. They discoursed the path routing or direct searching on users click are more presumably to convert visits into purchase. The cumulative impact of each visit suggests that customers make more visits, increasing the probability of them making a purchase. Similarly, Chen. et al., (2023) also performed analysis on clickstream data for understanding customer intentions on e-commerce store data. To cluster clickstream patterns K-mean and decision tree clustering performed to identify customer intentions to purchase in different online sessions.

2.3 Behavioural Pattern identification:

Gudigantala et al. (2016) recognized behavioral patterns with cost adjustments of items and categories that led in a significant influence on conversion rate in both forms; economically and statistically. For further studies in research, such behavioral patterns induce an open gap to emphasize on advanced machine learning methodologies those could be utilized in analyzing and identifying of the challenging relationships between clickstream behavior and conversion pattern results. After assessing the research, Raphaeli et al. (2017) determined that click page count, events, session duration, number of session events, and average time on page correspond with potential buyer's behavior. This indicates the event form, duration and click sequences will presumably play a role in predicting conversion. Similarly, Misra et al., (2021) analyze clickstream data through application of the unsupervised clustering techniques, to understand user behavior with non-predefined labels. This helped in distinguishing behavioral patterns, whether user is having intend in converting sales based on browsing paths and engagement times. Therefore, by assessing each of the previous factors industry experts might learn more by determining the changes between converting and non-converting behavioral patterns of clickstream.

Since, clickstream data allows industry experts to be more data driven with behavioral insights of users, companies can take huge support from clickstream data for their strategic decision-making, such as, website designing, making advertising strategies and retention of consumers in terms of post-purchase management (Wedel & Kannan 2016). Also, certain factors from sequential clickstream pattern data have been significant in mapping of customer journey. Factors as, session duration and interaction sequences aided in classification of converting and non-converting sessions (Shi et al., 2019). Similarly, Tallis and Yadav (2018) studied on predicting factors of user interactions in sponsored searches, using clickstream data. Findings from this study helped advertisers to optimize click-to-conversion pathways in sponsored search environments. This methodology directly helped in informed e-commerce CRO strategies and insights.

In addition, Chen and Su (2013) discussed online user behaviors expressing their interest, containing three important measures are the category in visiting path, length of access time and browsing frequency patterns, which have been considered for refining through clickstream data. To identify the behavioral pattern of users in terms of similar sets of interests, clustering algorithm was employed. In addition, Surya and Sharma (2013) performed clickstream analysis to determine web page importance, which was aimed to enhance web crawlers' efficiency. In that, research identified a user-centric approach in comparison of traditional link-dependent metrics. The comparative analysis was performed on various clickstream methods, as clustering algorithms, sub section analysis, and user's behavior path models. Resulting insights on most effective method for non-related needs for advanced data preprocessing techniques, scenarios and integration methods for additional data sources

2.4 Machine Learning Algorithms for Conversion rate optimization:

In addition, Wang (2024) discussed the limitation of traditional data analysis techniques in E-commerce domain, especially in high dimensional data handling. Improved Bayesian algorithm

was utilized with dynamic modeling strategies and normal distributions of datasets for enhanced market trend and user behavior predictions. Similarly, advanced data collection methodologies have been integrated in predictive models of Malaysian Ecommerce stores data, to examine KPIs particularly, customer lifetime value (CLV), churn rate and conversion rate. That research concluded that Classification and Regression Trees and regression models to predict and analyze business KPIs have been effective and robust to understand the customer behavior and optimize business strategies for higher conversion rates (Teh et al., 2021). The clustering technique has been used on clickstream data of e-commerce store to highlight the issues in identifying the different users' online behavior on store (with the intention of purchase) conversion predictions (Yeo et al., 2018). This led to the guidance in strategy retargeting for business problems, specifically customer-focused behaviors aimed to optimize conversion rate with (high predictability) or not (low predictability).

Similarly, Severeyn et al., (2023) implemented artificial neural networks (ANNs) on click stream data to classify the customer behavior and other factors that have been beneficial for business to make informed decisions. And recurrent neural networks (RNNs) have been utilized on click stream data to predict online shopping behavior and preferences of user (Koehn, Lessmann and Schaal, 2020). In addition, Gumber, Jain, and Amutha (2021) studied the customer behavior using clickstream data. XGBoosst (Extreme Gradient Boosting) framework was used on data from e-commerce websites. That analysis was aimed to improve accuracy for aiding business in making data driven decisions and effective predictions on customer actions to increase customer satisfaction and conversion rate.

3 Research Methodology

This study is aimed to analyze user behavior to optimize conversion rate of E-commerce store, using the customer, product, transaction and clickstream data. This data contains information related to transaction, product, user and session related details. In this study machine, learning techniques have been utilized on integrated datasets to predict user behavioral purchase patterns that influence user conversion rate. The goal was to design and evaluate a predictive model, which could precisely predict conversion rate from website users. While delivering meaningful insights for website performance, business and marketing strategies.

4 Design Specification

To identifying clickstream patterns and conversion rate optimization (CRO) analysis using machine learning algorithms on E-commerce dataset, logistics regression has been applied as a baseline model, later Random forest , gradient boosting and CatBoost classifier have been applied according to the dataset and complex relationship in the model. These models have been selected according to the problem statement, and due to their proven effectiveness with classification problem handling. These models are robust to capture complex patterns in data and capable of dealing with imbalanced datasets. To perform further analysis, the following items have been addressed.

4.1 Dataset:

To perform this analysis dataset has been sourced from open available source kaggle with the consideration of all privacy rights protection. The dataset consists of more than 0.8 million records from Aug 2016 till July 2022 on clickstream data of an E-commerce clothing brand selling in different clothing categories in Indonesia.
Customer Data: It contains some demographic details for users such as, gender, user id, DOB and country location.
Clickstream Data: This contains all session related details served online such as event types, data and time, clicks, pages, event name, session ID, event ID, and etc
Product: This data source contains information about product and category such as, product ID, category, mastery category and season.
Transaction: Transaction data includes details are, payment method, payment status, transaction ID, payment status, shipment fee, promo code, total amount shipment details and etc.

4.2 Data preprocessing:

As a programming tool Python has been used in modeling and analysis, different libraries have been used for ML modeling as; Pandas and Numpy. For visualization, Seaborn and Matplotlib have been utilized for better understanding of driven insights. In this step of preprocessing, data cleaning has been executed to remove redundancy and missing values from the dataset. Some of irrelevant columns have dropped from data such as location, longitude and latitude, first and last name, email address and device version etc. Mean strategy has been utilized to remove 145916 records of missing values from critical numeric features such as, item price, quantity and total price. Moreover, categorical columns were imputed with mode. More than 0.6 million records were kept after cleaning of datasets.
To handle categorical variables Labelencoder from scikit-learn has been applied to encode variables. It is necessary for machine learning algorithms to keep variables in numeric form. Encoded categorical columns were payment method, payment status, traffic source, event names, gender, season and master category.

Tools and Libraries*: Python, Pandas, Numpy, Scikit-learn, CatBoost, SimpleImputer and LabelEncode*
Visualization*:Libraries like Matplotlib and Seaborn for visualizing model performance and behavioral insights.*

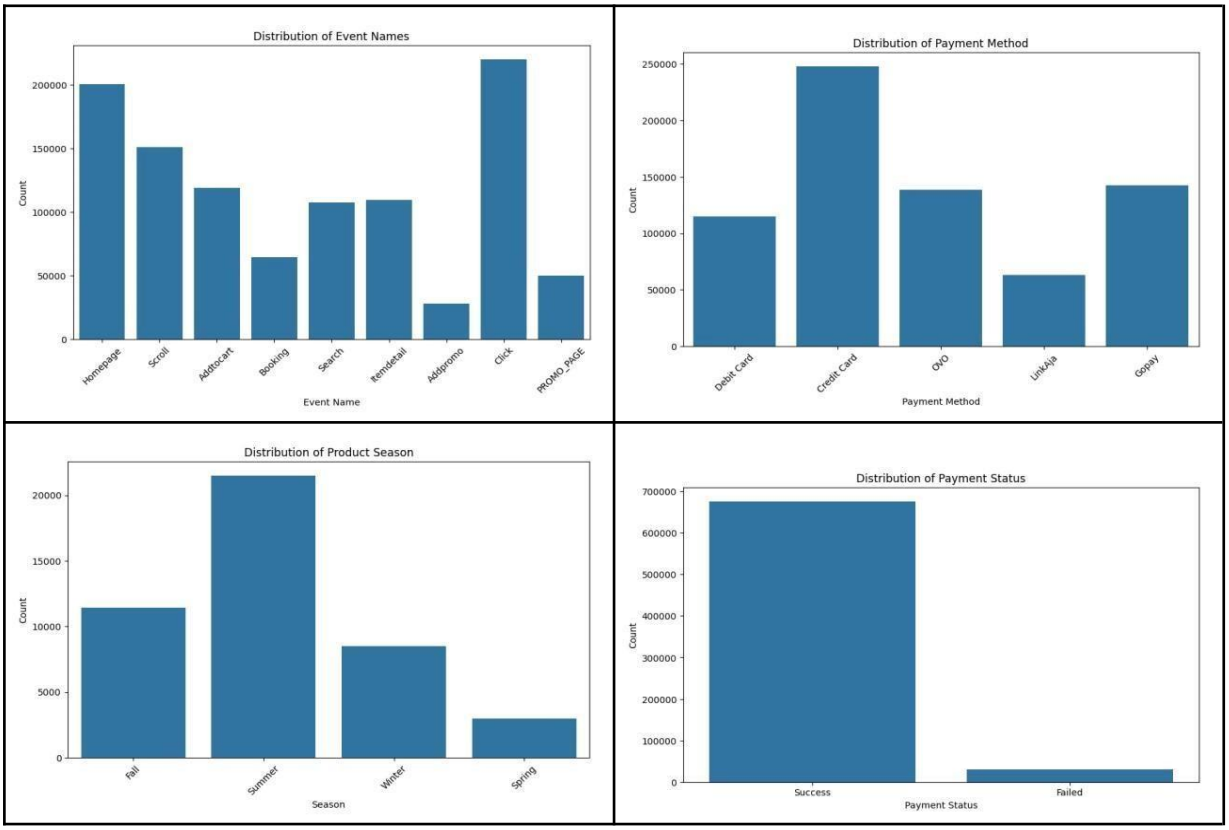

Fig:1 Distribution of variables in datasets

4.3 Feature engineering:

Feature engineering has been executed to identify and extract the useful features from multiple datasets (transaction, products, customer and clickstream data) that has represented the user behavior for purchase success predictions. Initially, session duration has been created from event start and end time differences considering session ID as unique visit on website for duration calculation. Later to improve insights bounce rate and number of page views have been created. These attributes contain a number of features about session details like, the amount of time spent on each page, navigation path, and behavior pattern as pages visited, the clicked actions named as event type and other relevant metrics. For conversion rate, the central action has been considered as purchase or not purchased. The record for purchase or no purchase has been converted to bool in analysis. For further analysis, the dataset has been divided into training and testing sets. In order to ensure that each class (0,1) has an sufficiently adequate number of samples in a balanced dataset.

4.4 Model development:

For in-depth analysis on non-converting behavioral patterns of users to enhance conversion rate machine learning models have been employed due to their robustness in managing imbalance dataset and capability of handling classification tasks efficiently. The binary classifier has been applied for conversion rate ranking with (0,1) for purchase or no purchase in the session. Appropriate binary classification algorithm has been employed in this analysis, considering the

nature of data and the desired level of interpretability from thresholds set. While maintaining some key factors in threshold consideration as, diversity and size of datasets, the complexity of the relationships between features and existing labels. These classifiers have been used to train dataset to distinguish between session behavioral patterns, which result in purchases and no-purchase by identifying output patterns and relationships in the clickstream data labels. For such predictive analysis, it is very important to select an appropriate model for optimizing the parameter for better accuracy and handling of imbalance dataset. In this section possible details are shared about model development.

Baseline model:

a. Logistic Regression: It is a simple and powerful statistical technique for binary classification problems. It is easy to interpret and implement on creating and explaining the relationship between dependent variable (DV) and independent variable (IVs). Logistics regression helps in identifying and understanding the relationships and significance of variables on each other. As a baseline model, it requires less resource to train the models and helps in establishing the performance benchmark for more advanced and complex models to be compared. Like in current analysis, it has added value in more sophisticated models used; Random Forest classifier and CatBoosting classifier for their performance improvement. Since, in this study the primary objective is the binary outcome; purchase or no-purchase (purchase = 1 or no-purchase = 0). This drives the probability of a variable to belong to a particular class, with the positive coefficient of 1 and negative coefficient with decreased value. Advance models:
b. Random Forest Classifier: In this study, clickstream data is highly dimensioned, containing various features such as event names, session duration, season, category details, payment method, promo code, etc. RF is suitable for such high-dimensional data to handle multiple variables. This model helped in interacting with complex features, with the model understanding the behavioral pattern of users. Random forest has the ability to deal with the anomalies and inconsistency in datasets. Also, it provides insights of features which helps in identi fying features that are significantly influenced in purchase or non-purchase behavior. This model is computationally efficient to deal with weak and strong trees and address the overfitting from the model, this ensures the enhanced performance of model with overall accuracy and generalizability of model on new and unseen data.
c. Gradient Boosting Classifier: This classifier is capable of delivering a strong predictive model, by utilizing weak learner outputs from data. In this study-selected dataset have several weak learners for predicting the purchase behavior from clickstream data. There are some complex non-linear relationships between target variable (purchase) and features. This technique is capable to capture and rectify error from previous tree sequentially. It also emphasizes adjusting the weights of data points on misclassified instances. This learning approach helps in dealing with imbalanced datasets, leading to enhanced model sensitivity to minority class and improved precision and recall for purchase predictions.
d. CatBoost Classifier: It’s an open sourced library of gradient boosting, very well suited for handling categorical features, this minimizes the chances of overfitting of model. Through dealing with one-hot encoding categorical variables transformed into

numerous binary features. Hence, this library has a feature of balanced bootstrapping which deals with an imbalanced dataset and ensures the bootstrapped samples are used in each iteration. In the training, this emphasizes on minority class and helps to overall performance of the model for improving accuracy and prediction of purchase.

4.5 Data balancing and splitting:

While using machine-learning algorithms on predictive analysis, particularly working with classification (binary) problems such as purchase or no purchase prediction, handling imbalanced data is a critical step. In this analysis, the dataset is highly imbalanced with higher distribution on purchase instances. To address this issue, advanced techniques have employed are:

a. SMOTE (Synthetic Minority Over-sampling Technique):

This technique is utilized to balance the class distribution through generating the synthetic data for minority classes. Instead of duplicating minority data classes, it creates new synthetic instances by interpolating between existing minority samples. To implement SMOTE, after identification of majority and minority classes (purchase or Non-purchase) the synthetic data samples synthetically generated for minority class and combined to original dataset to balance the disproportionality from data.

Before implementation of SMOTE:

```
Training target distribution:
purchase
1 (purchase)    5408660
(no purchase)     24468
```

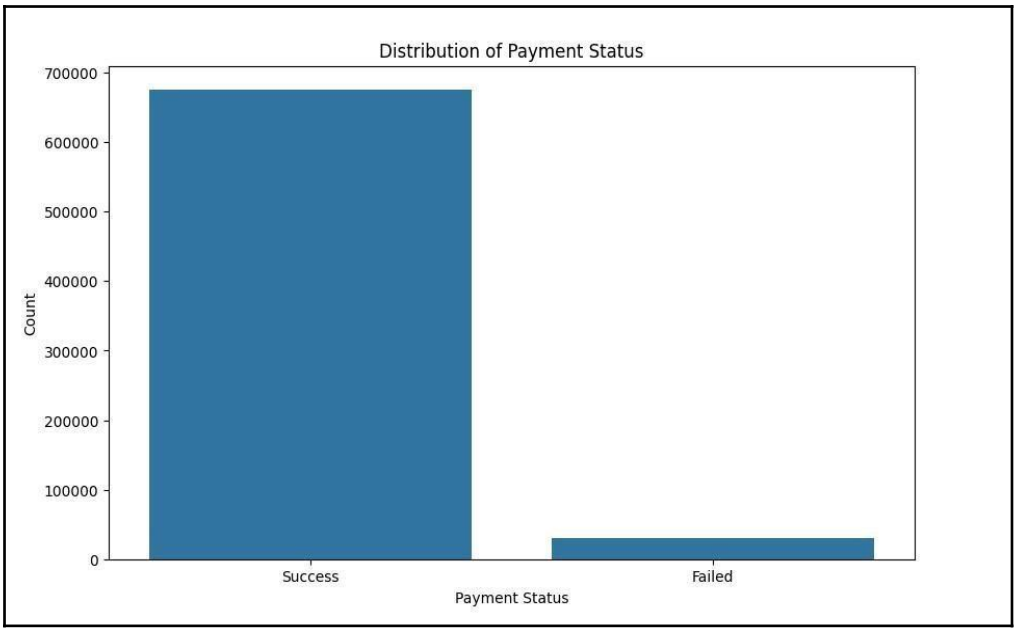

Fig2: Distribution of imbalance target variables

After employing SMOTE:

```
Training target distribution:
purchase
Class 0 (Non-purchase): 540,866 instances
Class 1 (Purchase): 540,866 instances
```

This resulted in a more representative dataset, allowing the model to perform better and predict both classes.

b. Bootstrapping:

This is also a resampling technique; here it is used to create multiple samples from the original dataset by random samples replacement. This helps in reducing overfitting of a model by estimating the performance variability of the model. To further reduce the overfitting, bootstrapping was applied with several bootstrapped samples on a balanced dataset after SMOTE implementation. For each bootstrapped sample, the model was trained separately from each instance. This process iteratively took place multiple times to get the collection of model output.

c. Downsampling:

Due to the huge difference of imbalance classes, this technique applied on the size of the majority class to be reduced, in order to be balanced with minority class. This helps to address class imbalance and ensures that the model is not biased towards the majority class. To implement this technique, after identifying both classes, the majority class 540,866 instances downsampled to 198,660 instances. In addition, the downsampled instances was combined with the majority class. So that the model was trained to be unbiased can ensured to be capable of predicting both classes accurately.

```
Balanced training target distribution:
0    198660
1    198565
Name: count, dtype: int64
Data resample complete. Shapes of train and test sets:
X_train:  (397225,  2)  y_train:  (397225,)
X_test: (99307, 2) y_test: (99307,)
```

5 Implementation

This section covers the implementation of a user behavioral predictive model to enhance the CRO of e-commerce store. The implementation of analysis includes the utilization of machine learning algorithms for analyzing clickstream data containing user, product and transaction details. The object behind analysis is to identify the patterns of user behavior that influence the success of conversion rate and develop robust and capable machine learning models for predicting desired outcomes.

For analysis execution python as a programming tool has been used. Libraries has been utilized for analysis and visualization of dataset are pandas, Seaborn, Numpy, Scikitlearn,Catboost and XGboost have been applied in google colab cloud environment.The dataset contained 1048575 records in clickstream, having details about navigation path, session id and timestamps, action taken (clicks), traffic source and page views etc. The customer data consist of 100000 entries including details of users like home location, DOB, name and customer ID. Similarly, the product dataset had 44425 records including product related details as, master category, price, IDs and season etc. The data file for the transaction had 852584 entries containing features like, payment method, promo code, shipment fees, total amount, payment status and booking and session ID.

After preprocessing of data, missing values and weak correlated variables from data have been removed. Further missing values were handled with mean imputed strategy, and categorical and normalized scalar variables were prepared for the ML algorithm. To have a comprehensive view on each customer session all dataset were merged on the bases of common shared identifiers such as customer Ids, session Ids and event Ids.

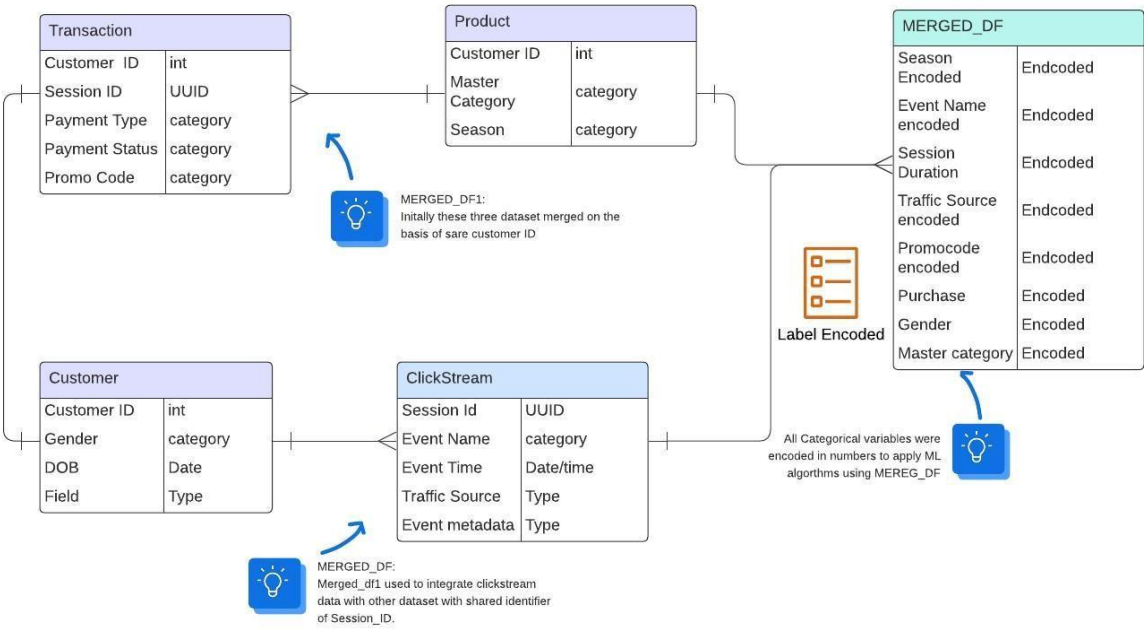

Fig:3 Flow to data integration for implementation of machine learning models.

Due to the highly imbalanced distribution of purchase and non-purchase instances in the dataset SMOTE was utilized to balance the distribution of majority and minority classes. To handle this hyper parameter tuning was applied on models including logistic regression, random forest, gradient boosting classifier to improve model performance and remove class specific outcomes from the model. That result was not satisfactory in numbers, and would be evaluated in the next section. Then downsampling was applied to manage equal distribution of each class, which resulted. Downsampling ensured that the instances in both class samples are equally applied in

training of the model, and prevented the bias for the majority class. The results from all models are discussed in the next section.

6 Evaluation

To evaluate the models performance mentioned metrics such as accuracy and precision, recall has been used, these metrics validated the effectiveness of the model in terms of CRO strategies by the proposed framework.

Evaluation Metrics:
Accuracy: It is the measure for correctly predicting the instances proportions.
Precision, Recall, and F1-Score: These scores from analysis used to evaluate the ability of the model correctly identifying the positive instance (purchase) and considers both false negative and positive.
ROC-AUC: This provides aggregately performing measures in all classification thresholds across models, useful for imbalanced datasets

6.1 Implementation of Models:

For initial implementation of ML models, logistic regression was employed to set the model accuracy benchmark, the accuracy of model on imbalanced data resulted in significant limitations in accurately classifying non-purchase records. The model revealed high accuracy of 95.65%, which showed strong performance at high level view. But this result was misleading because of the imbalance of the target class in the dataset. The number of purchase classes was significantly higher in contrast to non-purchase instances. From results, the F1 score, recall and precision of class 0 was zero. This means that the model is failing to perform correctly in identification of non-purchase classes. The findings for class 1 was having precision of 96%, recall 1.00 and f1-score of 0.98 along with macro avg of 0.49 and weight avg 0.94 which showed the skewness in model performance, more tilted towards majority class and being inefficient for minority class.

In this step the accuracy for model Logistic regression resulted is mentioned below:

```
Baseline Model - Logistic Regression before SMOTE
Accuracy: 0.9565precision    recall  f1-score   support

0       0.00      0.00      0.00      6145
1       0.96      1.00      0.98    135189

    accuracy                           0.96    141334
macro avg       0.48      0.50      0.49    141334
weighted avg       0.91      0.96      0.94    141334
```

6.2 Utilization of Imbalancing Techniques:

Results from analysis before applying smote were non-satisfactory. The accuracy to model was misleading the outcomes due to unbalanced data. Model failed in this step to identify 0 class, resulting in the all evaluating matrix - recall, precision and F1 score in zero. Nevertheless, the score for the majority class was 96% which shows poor performance for the minority class in this case. Similarly, the weighted average of the model showed the majority class skews the model. This showed the model was unable to handle imbalanced classes. To handle imbalance from results SMOTE was applied. Which adjusted the class weights and balanced the training data. After implementation of smote the accuracy of the model dropped from 95.6% to 61.9%. However, due to a big stretch in numbers for balancing the class, the model was still struggling in correctly identifying class 0. Some improvements in model number were increased for recall (0.36), f1-score (0.08) and precision for class 0 was 0.04. Despite some improvements, the model was still performing poorly. The weighted avg score of model after SMOTE achieved 0.73, which revealed the existing challenge to model in accurate classification of both models. However, the SMOTE has successfully balanced the training data set. The logistic regression performed poorly for non-purchase class with very low F1 score and precision.

Logistic Regression with SMOTE and Class Weights

Accuracy: 0.6192777392559469 precision recall

f1-score support

0 0.04 0.36 0.08 6145
1 0.96 0.63 0.76 135189

accuracy 0.62 141334
macro avg 0.50 0.49 0.42 141334
weighted avg 0.92 0.62 0.73 141334

In contrast, the random forest and gradient boosting classifiers also achieved accuracy of 100% before applying SMOTE or any other balancing techniques. The classification reports from the model are mentioned below, showing the flawed results with precision, recall, and F1-scores of 1.00 for both classes. This indicates that the models are overfitting the training sets due to biased training samples used earlier. To address this hyperparameter tuning was employed, with an objective to refine models performance for improving accuracy and generalizability through adjusting parameters. Even after that the models resulted the same with cross validation scores of 1.00 across all folds. This opens up the gap for improvement in models.

Random Forest Classifier

Accuracy: 1.0Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 6145
1 1.00 1.00 1.00 135189

```
accuracy                           1.00 141334
macro avg         1.00  1.00  1.00  141334

weighted avg      1.00  1.00  1.00  141334

Fitting 3 folds for each of 10 candidates, totalling 30 fits
Random Forest Classifier (After Hyperparameter Tuning)
Accuracy: 1.0Classification Report:
precision   recall f1-score support

0       1.00      1.00  1.00  6145
1       1.00      1.00  1.00  135189

accuracy                       1.00  141334
macro avg         1.00  1.00  1.00  141334
weighted avg      1.00  1.00  1.00  141334

Cross-Validation Scores: [1. 1. 1. 1. 1.]Mean
Cross-Validation Score: 1.0

Gradient Boosting Classifier
Accuracy: 1.0Classification Report:
precision   recall f1-score support

0       1.00      1.00  1.00  6145
1       1.00      1.00  1.00  135189

accuracy                       1.00  141334
macro avg         1.00  1.00  1.00  141334
weighted avg      1.00  1.00  1.00  141334
Fitting 3 folds for each of 18 candidates, totaling 54
fits
Gradient Boosting Classifier (After Hyperparameter Tuning)
Accuracy: 1.0Classification Report:
precision   recall f1-score support

0       1.00      1.00  1.00  6145
1       1.00      1.00  1.00  135189

accuracy                       1.00  141334
macro avg         1.00  1.00  1.00  141334
weighted avg      1.00  1.00  1.00  141334

Cross-Validation Scores: [1. 1. 1. 1. 1.]Mean
```

Cross-Validation Score: 1.0

6.3 Implementation of DownSampling:

a. Logistic Regression:

After employing downsampling, the logistic regression model resulted in some improvement of model performance. With some improved ability in predicting both classes fairly, the achieved accuracy increased to 66.5%. Revealing that now the model is able to predict two third of the instances in the test sample correctly. The score from precision and recall were 0.66 and 0.69 for class 0, and 0.67 of f1-score for class 1.Hence the model has achieved a precision of 0.68 and a recall of 0.64, with an F1-score of 0.66. Which showed the model is identifying non-purchase instances slightly better than purchase instance with moderate weightage of macro and weighted averages for precision, recall, and F1-score. These all hovered around 0.67, reflecting a consistent, but not overly strong, performance across both classes.

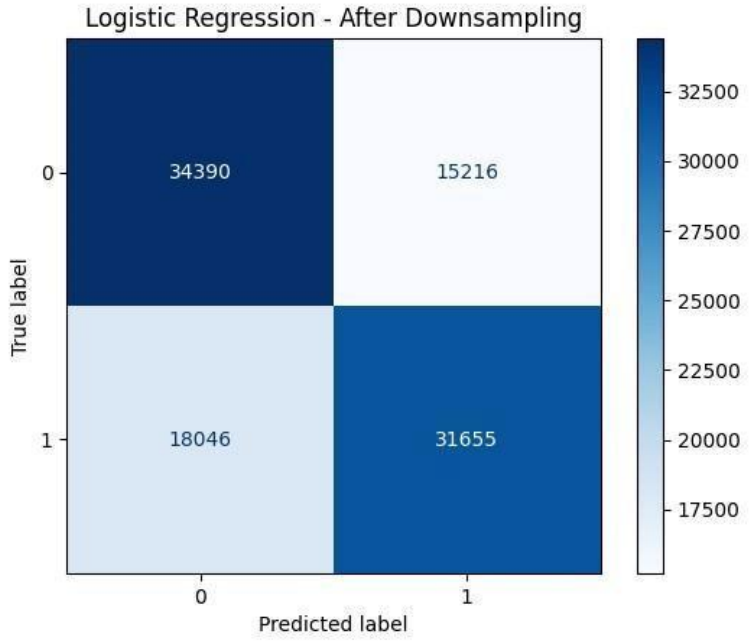

Fig:4 Confusion matrix of Logistic Regression after Downsampling

Fig:4 provides the detailed breakdown of the model after downsampling. The confusion matrix showed 31655 instances of true positive (TP), which shows that the model is correctly predicting positive instances of class 1 (purchase - 31655). And, 34390 correctly predicting instances for true negative (TN) for class 0 (non-purchase - 34390). In false positive (FP) the model is incorrectly predicting 15216 of negative instances as positive. Similarly, for false negative (FN) it is predicting 18,046 instances incorrectly as negative when they were actually positive.

Accuracy=TP+TN/TP+TN+FP+FN

Accuracy≈ 66045 ≈0.665
99292

Since, the calculated accuracy from the model after downsampling was 66.5% predicting both instances correctly (true positive and true negative).

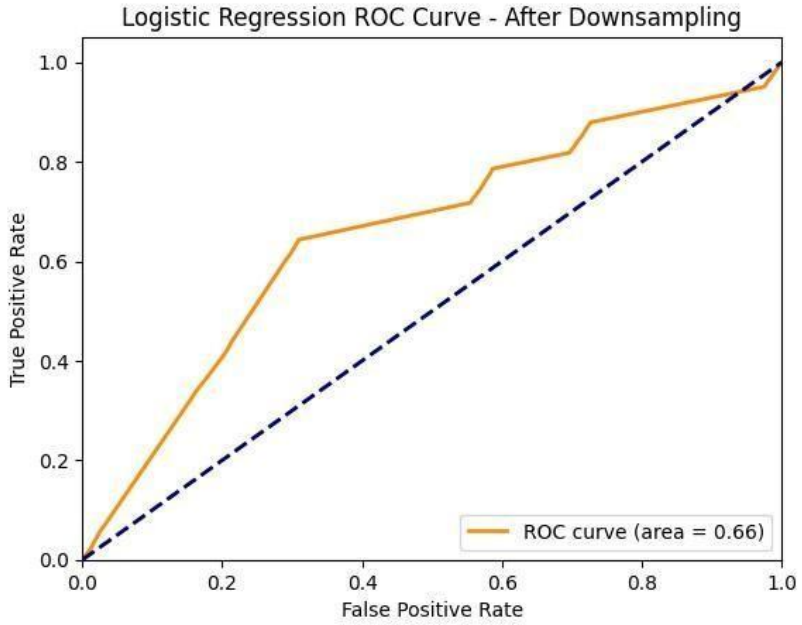

Fig: 5 ROC Curve for Logistic Regression after Downsampling

Fig.5 is demonstrating the model performance with various threshold settings. The ROC (Receiver Operating Characteristic) curve and its AUC (Area Under the Curve) value shows the curve that pass through TPR and FPR for different threshold values. The TPR (true positive rate) is the sensitivity and recall ability of a model for positive. In addition, FPR (false positive rate) is the ratio of positive observations predicted incorrectly. Since, the value of AUC 0.66 has resulted that the model is able to differentiate between positive and negative instances though these point are far from perfect line points.

The findings from fig4 and fig5 demonstrated that, after downsampling logistic regression has reasonably improved and balanced performance but still has a room for improvement to achieve higher precision and recall across both classes.

6.4 Implementation of advanced models:

The implementation of downsampling on advanced models has significantly improved the performance in handling imbalanced classes in data. All three model random forest, gradient and CatBoosting classifiers have achieved an accuracy of 71.26% representing their efficiency and effectiveness in predicting correctly after removing imbalanced effects from datasets. Random forest performed notable result, with enhanced predicting accuracy in identification of purchase instances with precision—0.77, recall—0.60, and an F1-score— 0.68 for the non-purchase class, while achieving a precision—0.67, recall—0.82, and an F1score—0.74 for the purchase class. Similarly, other classifiers (Gradient boost and CatBoosting classifier) mirrored the performance accuracy of random forest, while maintaining metrics (percision, recall and

F1-score) for both classes. The continuity in results showed that GB is well suited in dealing complex and non-linear relationships in data with its sequential tree-building approach.

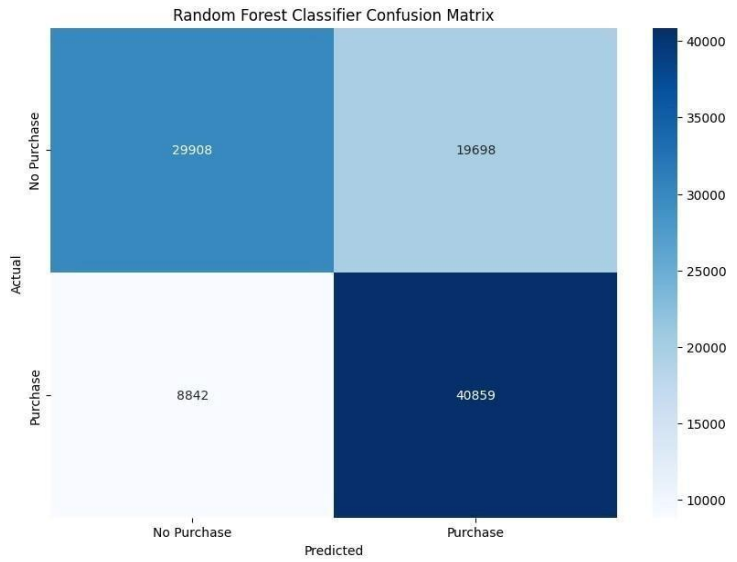

Fig.6 Random Forest Confusion Matrix after Downsampling

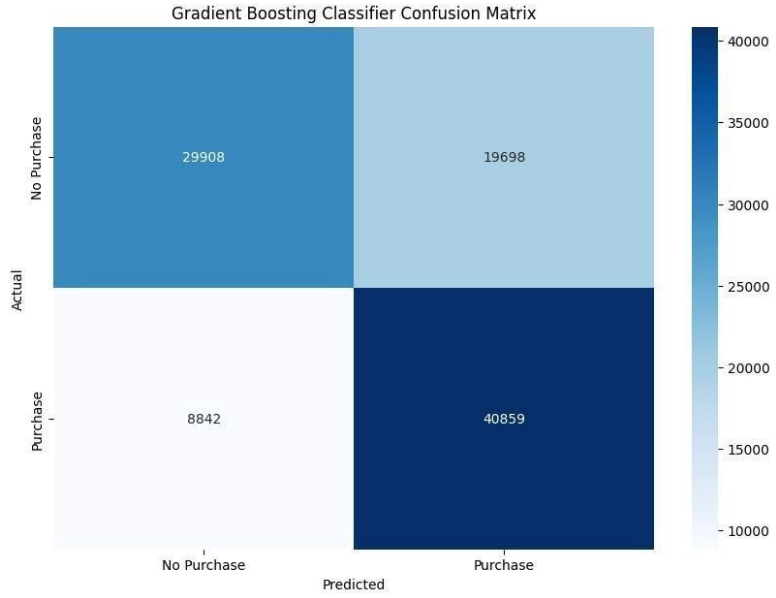

Fig.7 Gradient Boosting Classifier Confusion Martix after Downsampling

CatBoost Classifier with Downsampling
Accuracy:0.7126083760459988

Classification Report: precision recall f1-score support

0 0.77 0.60 0.68 49606

1 0.67 0.82 0.74 49701

```
    accuracy                           0.71     99307
   macro avg       0.72      0.71      0.71     99307
weighted avg       0.72      0.71      0.71     99307
```

The implementation of the CatBoost classifier also achieved improved accuracy in post downsampling. With the identical numbers from random forest and gradient boosting classifiers. CatBoost demonstrated the uniform effectiveness in handling imbalanced datasets. Catboost classifier is a most suited tool for classification tasks. It has capability of managing categorical features like those used in current analysis. In addition, uniformity in results underscores the robustness of these methods, and indicated that they are suitable for this classification problem, along with these dealt with the imbalance class issue and maintained the accuracy and generalizability.

Model	Accuracy (Before Downsampling)	Accuracy (After Downsampling)	Precision (0)	Recall (0)	F1Score (0)	Precision (1)	Recall (1)	F1Score (1)
Logistic Regression	0.90	0.66	0.04	0.36	0.08	0.96	0.63	0.76
Random Forest Classifier	1.00	0.65	0.77	0.60	0.68	0.67	0.82	0.74
Gradient Boosting Classifier	1.00	0.71	0.77	0.60	0.68	0.67	0.82	0.74
CatBoost Classifier	0.95	0.71	0.77	0.60	0.68	0.67	0.82	0.74

Table1: Comparison of models performance before and after downsampling

The comparison between the baseline model and advanced methods demonstrated the benefits of using complex models for classification. Logistic Regression contributed as an effective initiating point with its balanced performance, with a difference of 66.5% lower accuracy than ensemble models; this highlighted the limitations of linear models in capturing non-linear relationships, and complex interactions between data features as well. Advanced machine learning methods as Gradient Boosting, CatBoost and Random Forest were competitive to accomplish improved accuracy with enhanced parameter scores; these parameters (precision, recall and f1-score) aided models to be focused on misclassified points and combined multiple weak learners.

The results suggest that ensemble methods are likely to surpass simpler models like Logistic Regression, for more complex datasets with potential non-linearity and interactions. The enhanced predictive accuracy across three models and the balance performance in both classes, underlines the importance of addressing and identification of class imbalance through downsampling and other techniques. Through applied techniques, models have been ensured to perform unbiased analysis in training and testing of model. That shows model were able

handle the majority class and deliver more reliable prediction for both classes dealing them equally in training datasets.

6.5 Discussion

The performed analysis delivered valuable insights for the effectiveness and ability of machine learning models (Logistic Regression- LR, Random Forest- RF, Gradient Boosting-GB and CatBoosting classifier) while using clickstream data from the E-commerce domain for optimizing conversion rates. After downsampling, the overall performance of the models also provides valuable insights about the dataset nature and classification tasks challenges in the context of imbalanced data. Achieving similar results across utilized models suggested that the dataset was relatively straightforward for these models to learn, particularly after addressing the class imbalance; but across all models, the moderate accuracy of 71.26% indicates that there may be further need for improvement. To further enhance model performance, in the near future could explore advanced feature engineering or the use of more enlightened techniques like deep learning and neural network models. Additionally the results concluded that for balancing the dataset, the downsampling was an effective technique; but other techniques like SMOTE, bootstrapping or hybrid approaches could be checked to see if they offer additional improvements.

Also, particularly in dealing with imbalanced datasets; the results after downsampling exemplify that Gradient Boosting, Random Forest and CatBoost classifiers are well-suited for the task classification at hand. The similar performance across all these three models suggested that ensemble methods offer robust and an effective approach to handle complex classification tasks, balancing precision and recall to deliver overall strong performance. While, Logistic Regression provided a good baseline, ensemble methods clearly surpassed it, highlighting the importance of using more sophisticated models for challenging classification tasks. The results consistency also addresses the importance of class imbalance through techniques like downsampling; that can significantly improve the reliability and portability of predictive models. Additional feature engineering and further advanced technique exploration could provide even more improvements as a next step in model performance, offering valuable insights in this area for future work

6.6 Limitations and Future Recommendation

Data imbalance: The major limitation encountered in this analysis was the highly imbalanced distribution of dataset, the disproportionately of purchase and non-purchase instances was significantly imbalanced. Despite employing advanced techniques like SMOTE, bootstrapping and downsampling to address this challenge, there is still an inherent challenge in dealing with such datasets that could be found for ensuring that synthetic generated data accurately represents the real-world distribution of new data.

Feature engineering: Few of the new features have been created in this analysis, but due to time and resources constraints there is a future gap of improvement that can be addressed by applying advanced feature extraction techniques, like deep learning based feature extraction

with external data sources. Similarly, clickstream data has temporal features that were not fully exploited in this analysis. For future behavioral analysis, this will help to identify more influencing factors on purchase behavior of users, like changes in preferences of users like category, season's trends and marketing campaigns. For more valuable insights with such factors time-series analysis or recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) these methods are designed to handle sequential data, it could deliver more improved model performance in future analysis.

Overfitting and generalizability: Overfitting of the model occurs when the model learns the training dataset very well, even including outliers and noise of data. Such a model expects to perform poorly on unseen datasets. In this study, the applied models were overfitting the training dataset, despite measures like bootstrapping, hyperparameter tuning and downsampling being employed. When categorical models were utilized like CatBoost, which resulted in highest accuracy. However, it has shared the risk of overfitting due to binary outcomes of purchase or no purchase. Similarly, this model has been trained on specific dataset. The performance of this model in different domains would be uncertain to be generalized. Other factors like, website design layout, business domain and further external factors could also play a major role in change of user purchase predictions. For future analysis this model can also employee on different industry dataset, as clickstream dataset has timestamps and session details this could help in more better analysis in future, for example in education industry for online study session or courses this model could help in predicting the chances of course completion by users. Such cross-domain validation and experimentation would provide insights into the robustness of the models.

7 Conclusion and Future Work

To conclude, the research highlights the significant impact that algorithms of machine learning can have on optimizing and understanding conversion rates in e-commerce, through the clickstream data analysis. The study explained how advanced machine learning techniques such as: Random Forest, Gradient Boosting, Logistic Regression and CatBoost can be employed to enhance overall website performance and to predict user intent; by identifying users key behavioral patterns during their online sessions. This study also highlights the challenges for dealing with imbalanced datasets, presenting how techniques like downsampling and SMOTE are crucial for improving reliability and model accuracy; especially in identifying non-purchase behaviors. For digital marketers, e-commerce businesses and UX/UI designers these findings offer valuable insights; enabling them to make data-driven decisions that can lead to better customer engagement and higher conversion rates.

Moreover, broader insights of this research extends the notion of e-commerce. The methodologies driven from the study can be applied across numerous industries including: online education, financial services, media and healthcare, for improving service delivery and customer satisfaction while understanding the user behavior. By holding on clickstream data and machine learning; organizations can enhance their user experiences, refine digital strategies and can ultimately achieve better outcomes of their business. Additionally, this research not

only contributes towards academic understanding of conversion rate optimization but also provides practical tools and approaches that can implemented directly into scenarios of the real world, making it a valuable resource for both practitioner and researchers alike.

References

Bigon, L., Cassani, G., Greco, C., Lacasa, L., Pavoni, M., Polonioli, A. and Tagliabue, J., 2019. Prediction is very hard, especially about conversion. Predicting user purchases from clickstream data in fashion e-commerce. *arXiv preprint arXiv:*1907.00400.

Bucklin, R.E. and Sismeiro, C., 2003. A model of web site browsing behavior estimated on clickstream data. *Journal of marketing research*. 40(3), pp.249-267.

Cai, L., He, X., Dai, Y. and Zhu, K., 2018, September. Research on B2B2C E-commerce website design based on user experience. *In Journal of Physics: Conference Series* (Vol. 1087, No. 6, p. 062043). IOP Publishing.

Chen, L. and Su, Q., 2013, July. Discovering user's interest at E-commerce site using clickstream data. *In 2013 10th International Conference on Service Systems and Service Management* (pp. 124-129). IEEE.

Chen, S.S., Li, T.L., Wu, Y.C. and Singh, V., 2023, June. An Algorithm-based approach for Mapping customer journeys by identifying customer browsing behaviors on E-commerce Clickstream data. *In 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications* (HORA) (pp. 1-8). IEEE.

Crystallize.com. (2023). eCommerce Conversion Rate Optimization (CRO): *Boosting Your Conversion Rate Made Simple In 2023.* [online] Available at: https://crystallize.com/blog/ecommerce-conversion-rate-optimization [Accessed 16 Sep. 2024].

Even, A., 2019. Analytics: Turning data into management gold. *Applied Marketing Analytics*, 4(4), pp.330-341.

Gudigantala, N., Bicen, P. and Eom, M., 2016. An examination of antecedents of conversion rates of e-commerce retailers. *Management Research Review*, 39(1), pp.82-114.

Gumber, M., Jain, A. and Amutha, A.L., 2021, May. Predicting Customer Behavior by Analyzing Clickstream Data. *In 2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)* (pp. 1-6). IEEE.

Hanamanthrao, R. and Thejaswini, S., 2017, May. Real-time clickstream data analytics and visualization. *In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 2139-2144). IEEE.

Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Ghasemianlangroodi, A., 2014. Analyzing negative user behavior in a semi-anonymous social network. CoRR abs, 1404.

Koehn, D., Lessmann, S. and Schaal, M., 2020. Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, 150, p.113342.

Kohavi, R., Tang, D., & Xu, Y. (2023). Enhancing E-Commerce Conversion through Behavioral Analysis: Insights from Clickstream Data. *International Journal of Data Science and Analytics*, 12(2), 123-137.

Li, H. and Kannan, P.K., 2014. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of marketing research*, 51(1), pp.40-56.
Liu, Z., Wang, Y., Dontcheva, M., Hoffman, M., Walker, S. and Wilson, A., 2016. Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE transactions on visualization and computer graphics*, 23(1), pp.321-330.
Misra, G., Matteo Migliavacca and Fernando (2021). Behavioural User Identification from Clickstream Data for Business Improvement. *Lecture notes in computer science*, pp.341–354. doi:https://doi.org/10.1007/978-3-030-91100-3_27.
Montgomery, A.L., Li, S., Srinivasan, K. and Liechty, J.C., 2004. Modeling online browsing and path analysis using clickstream data. Marketing science, 23(4), pp.579-595.
Narang, B., Trivedi, P. and Dubey, M.K., 2017. Towards an Understanding of UX (User Experience) and UXD (User Experience Design), an Applicability Based Framework for Ecommerce, Intranets, Mobile & Tablet & Web usability. International Journal of Advanced Research in Computer Science, 8(5).
Pal, G., Atkinson, K. and Li, G. (2021). Real-time user clickstream behavior analysis based on apache storm streaming. Electronic Commerce Research. doi:https://doi.org/10.1007/s10660-021-09518-4.
Purnomo, Y.J., 2023. Digital marketing strategy to increase sales conversion on e-commerce platforms. Journal of Contemporary Administration and Management (ADMAN), 1(2), pp.54-62.
Raphaeli, O., Goldstein, A. and Fink, L., 2017. Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. Electronic commerce research and applications, 26, pp.1-12.
Sakalauskas, V. and Kriksciuniene, D., 2024. Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers. Algorithms, 17(1), p.27.
Severeyn, E., La Cruz, A., Matute, R. and Estrada, J., 2023, October. Neural Networks for Customer Classification Through Clickstream Analysis. In 2023 IEEE Seventh Ecuador Technical Chapters Meeting (ECTM) (pp. 1-6). IEEE.
Kaushik, U. and Grondowski, A., 2017. Conversion Rate Optimization of E-Commerce using Web Analytics and Human-computer Interaction Principles: An in-depth Quantitative Approach to Optimization of Conversion Rates.
Shi, P., Zhang, Z. and Choo, K.K.R., 2019. Detecting malicious social bots based on clickstream sequences. IEEE Access, 7, pp.28855-28862.
Stair, R. and Reynolds, G. (2008) Fundamentals of Information Systems: A Managerial Approach, 4th ed., Boston (MA): Thomson.
Surya, A. and Sharma, D.K., 2013, April. A comparative analysis of clickstream as web page importance metric. In 2013 IEEE Conference on Information & Communication Technologies (pp. 776-781). IEEE.
Szabo, P. and Genge, B., 2021. Efficient Behavior Prediction Based on User Events. Journal of Communications Software and Systems, 17(2), pp.134-142.
Tallis, M. and Yadav, P., 2018, December. Reacting to variations in product demand: An application for conversion rate (cr) prediction in sponsored search. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 1856-1864). IEEE.

Teh, Z.H., Lim, C.Y. and Chen, S.Y., 2021, September. An Exploratory Review of Malaysian E-Commerce Merchants and Their Readiness in Adopting Business Analytics Models for Assessment of Business KPIs. In 2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI) (pp. 469-473). IEEE.
Wang, L., 2024, April. Design of Intelligent Analysis Method for E-Commerce Data Based on Improved Bayesian Algorithm. In 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-4). IEEE.
Wedel, M. and Kannan, P.K., 2016. Marketing analytics for data-rich environments. Journal of marketing, 80(6), pp.97-121.
Xiao, Y., He, W., Zhu, Y. and Zhu, J., 2022. A click-through rate model of e-commerce based on user interest and temporal behavior. Expert Systems with Applications, 207, p.117896. Yeo, J., Hwang, S.W., Koh, E. and Lipka, N., 2018. Conversion prediction from clickstream: Modeling market prediction and customer predictability. IEEE Transactions on Knowledge and Data Engineering, 32(2), pp.246-259.
Ze Jun Wen, Lin, W. and Liu, H. (2023). Machine-Learning-Based Approach for Anonymous Online Customer Purchase Intentions Using Clickstream Data. Systems, 11(5), pp.255–255. doi:https://doi.org/10.3390/systems11050255.