A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

THESIS

by:
SISKA FARIZAH MAULUDIAH
NIM : 230605220004

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

SISKA FARIZAH MAULUDIAH
NIM : 230605220004

A Thesis is submitted to Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang for the Requirements for the Degree of Master of Computer (M.Kom)

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

THESIS

SISKA FARIZAH MAULUDIAH
NIM : 230605220004

Has been reviewed and approved for the examination
Date , 14 , ...May....., 2025

Supervisor I,	Supervisor II,
Dr. Yunifa Miftachul Arif, M.T	Dr. Cahyo Crysdian, M.CS
NIP. 19830616 201101 1 004	NIP. 19740424 200901 1 008

Acknowledged
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian, M.CS
NIP. 19740424 200901 1 008

A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

AN UNDERGRADUATE THESIS

By:
SISKA FARIZAH MAULUDIAH
NIM : 230605220004

Has been conducted in the presence of the board of undergraduate thesis examiners and stated accepted as one of the requirements to obtain the tittle Master of Computer (M.Kom)

Date 14, May, 2025

Board of Examiner		**Signature**
Examinrer I	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Examiner II	: Dr. Usman Pagalay, M.Si NIP. 19650414 200312 1 001	( )
Supervisor I	: Dr. Yunifa Miftachul Arif, M.T NIP. 19830616 201101 1 004	( )
Supervisor II	: Dr. Cahyo Crysdian, M.CS NIP. 19740424 200901 1 008	( )

Approved by
Head of Master Program in Computer Science
Faculty of Science and Technology
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian, M.CS
NIP. 19740424 200901 1 008

LETTER OF STATEMENT

The undersigned below:

Name	: Siska Farizah Mauludiah
NIM	: 230605220004
Program	: Master of Computer Science
Faculty	: Science and Technology

States that Master Thesis mentioned above is my original work except for the quotations and statements whose resources are acknowledged on the references. Any shortcoming in this present work, therefore, are entirely my own responsibility.

Moreover, this work is not plagiarism result and if it is found that this statement is false, my academic records will attest to invalidation and I will be responsible for that as well.

Malang,
Signed by,

SISKA FARIZAH MAULUDIAH
NIM : 230605220004

A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

THESIS

SISKA FARIZAH MAULUDIAH
NIM : 230605220004

Has been reviewed and approved for the examination
Date , 14 , ...May....., 2025

Supervisor I,	Supervisor II,
Dr. Yunifa Miftachul Arif, M.T	Dr. Cahyo Crysdian, M.CS
NIP. 19830616 201101 1 004	NIP. 19740424 200901 1 008

Acknowledged
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian, M.CS
NIP. 19740424 200901 1 008

A SYNERGISTIC APPROACH TO E-COMMERCE RECOMMENDER SYSTEM : LOGISTIC REGRESSION AND KULLBACK-LEIBLER DIVERGENCE

AN UNDERGRADUATE THESIS

By:
SISKA FARIZAH MAULUDIAH
NIM : 230605220004

Has been conducted in the presence of the board of undergraduate thesis examiners and stated accepted as one of the requirements to obtain the tittle Master of Computer (M.Kom)

Date 14, May, 2025

Board of Examiner		**Signature**
Examinrer I	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Examiner II	: Dr. Usman Pagalay, M.Si NIP. 19650414 200312 1 001	( )
Supervisor I	: Dr. Yunifa Miftachul Arif, M.T NIP. 19830616 201101 1 004	( )
Supervisor II	: Dr. Cahyo Crysdian, M.CS NIP. 19740424 200901 1 008	( )

Approved by
Head of Master Program in Computer Science
Faculty of Science and Technology
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian, M.CS
NIP. 19740424 200901 1 008

FOREWORD

Assalamu'alaikum Wr.Wb.

Alhamdulillah, the author extends his gratitude to the presence of Allah SWT who has bestowed His Grace and Guidance, so that the author can complete her studies at the Master of Computer Science Program, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang and complete this Thesis well. Furthermore, the author would like to express her gratitude along with prayers and hopes jazakumullah ahsanal jaza' to all parties who have helped complete this Thesis. The author would like to express his gratitude to:

1. Dr. Yunifa Miftachul Arif, M.T and Dr. Cahyo Crysdian, M.CS as the thesis supervisors, who has provided a lot of valuable guidance and experience.
2. All academicians of the Master of Computer Science Program, especially all lecturers, thank you for all your knowledge and guidance.
3. My parents who always give their prayers and blessings to the author in pursuing knowledge.
4. The author's beloved younger siblings and children who always gave the author support to pursue a master's program in computer science.
5. All parties who have supported in completing this thesis.

The author realizes that in the preparation of this thesis there are still shortcomings and the author hopes that this thesis can provide benefits to the readers especially for the author personally. Aameen Ya Rabbal Alameen.

Wassalamu'alaikum Wr.Wb

Malang,
Author

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Mauludiah, Siska Farizah. 2025. A Synergistic Approach To E-Commerce Recommender System: Logistic Regression And Kullback-Leibler Divergence. Theses. Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Promotor: (I). Dr. Yunifa Miftachul Arif, M.T (II) Dr. Cahyo Crysdian, M.CS.

This thesis presents a comprehensive approach to developing a reliable and actionable recommendation system focused on identifying repeat buyers, key drivers of long-term growth and customer loyalty in e-commerce. The study addresses critical challenges such as class imbalance and feature overload by incorporating Kullback-Leibler (KL) divergence into both the feature refinement and evaluation stages. This integration allows the system to focus on the most relevant customer attributes, enhancing the clarity and efficiency of the recommendation process. Using logistic regression as the core predictive model, the system is strengthened with techniques such as SMOTE for balancing buyer classes, class weighting to improve learning outcomes, and regularization to ensure model stability. Developed using real-world data from an Indonesian e-commerce platform, the model demonstrates improved ability to identify customers likely to make repeat purchases. The enhanced system supports better targeting of high-value customers, more personalized product recommendations, and smarter marketing decisions. Ultimately, this research delivers a practical, interpretable, and scalable solution for building customer-focused recommender systems, enabling e-commerce businesses to optimize retention strategies and maximize engagement.

Keywords : recommender system, repeat buyer classification, logistic regression, kullback-leibler divergence, feature engineering

ABSTRAK

Mauludiah, Siska Farizah. 2025. **A Synergistic Approach To E-Commerce Recommender System: Logistic Regression And Kullback-Leibler Divergence**. Theses. Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I). Dr. Yunifa Miftachul Arif, M.T (II) Dr. Cahyo Crysdian, M.CS.

Tesis ini menyajikan pendekatan komprehensif dalam mengembangkan sistem rekomendasi yang andal dan dapat diimplementasikan, dengan fokus pada identifikasi pembeli ulang, kontributor utama bagi pertumbuhan jangka panjang dan loyalitas pelanggan dalam industri e-commerce. Penelitian ini mengatasi tantangan penting seperti ketidakseimbangan kelas dan kelebihan fitur dengan mengintegrasikan metode Kullback-Leibler (KL) divergence ke dalam tahap penyempurnaan fitur dan evaluasi. Integrasi ini memungkinkan sistem untuk berfokus pada atribut pelanggan yang paling relevan, sehingga meningkatkan kejelasan dan efisiensi dalam proses rekomendasi. Dengan menggunakan regresi logistik sebagai model prediksi utama, sistem ini diperkuat dengan teknik seperti SMOTE untuk menyeimbangkan distribusi kelas pembeli, class weighting untuk meningkatkan hasil pembelajaran, dan regularisasi untuk menjaga stabilitas model. Sistem ini dikembangkan menggunakan data nyata dari platform e-commerce di Indonesia dan menunjukkan peningkatan kemampuan dalam mengidentifikasi pelanggan yang berpotensi melakukan pembelian ulang. Sistem yang telah ditingkatkan ini mendukung segmentasi pelanggan bernilai tinggi secara lebih tepat, rekomendasi produk yang lebih personal, serta pengambilan keputusan pemasaran yang lebih cerdas. Pada akhirnya, penelitian ini menghasilkan solusi yang praktis, mudah dipahami, dan dapat diskalakan untuk membangun sistem rekomendasi yang berfokus pada pelanggan, sehingga memungkinkan bisnis e-commerce mengoptimalkan strategi retensi dan memaksimalkan keterlibatan pelanggan.

Kata kunci: sistem rekomendasi, klasifikasi pembeli ulang, regresi logistik, kullback-leibler divergence, rekayasa fitur.

الملخص

الانحدار اللوجستي :نهج تآزري لتطوير نظام التوصية في التجارة الإلكترونية .2025 .معلودياه، سيسكا فاريزاه برنامج الماجستير في علم المعلومات، كلية العلوم والتكنولوجيا، .أطروحة .لايبلر-وتباعد كولباك يُنيفة مفتاح العارف، .د (1) :المشرفون .جامعة الدولة الإسلامية مولانا مالك إبراهيم مالانج كاهيو كريستيان، ماجستير علوم الحاسوب .د (2)ماجستير تقانة.

تقدم هذه الأطروحة نهجًا شاملاً لتطوير نظام توصية موثوق وفعال يركز على تحديد المشترين المتكررين، تتناول .الذين يُعدون من المحركات الرئيسية للنمو طويل الأمد وولاء العملاء في بيئة التجارة الإلكترونية الدراسة تحديات حاسمة مثل عدم توازن الفئات وتعدد السمات غير الضرورية، من خلال دمج مقياس تباعد يسمح هذا التكامل للنظام بالتركيز على .في مرحلتي تحسين السمات والتقييم (KL divergence)لايبلر -كولباك ويُستخدم نموذج الانحدار اللوجستي .السمات الأكثر صلة بالعملاء، مما يعزز وضوح وكفاءة عملية التوصية لتحقيق التوازن بين الفئات، وأوزان SMOTEكنموذج تنبؤي رئيسي، ويُعزز النظام باستخدام تقنيات مثل تم تطوير النموذج باستخدام بيانات حقيقية من .الفئات لتحسين نتائج التعلم، والتنظيم لضمان استقرار النموذج منصة تجارة إلكترونية إندونيسية، وقد أظهر قدرة محسّنة على التعرف على العملاء المحتملين للشراء مرة يدعم النظام المحسّن استهداف العملاء ذوي القيمة العالية بشكل أفضل، ويوفر توصيات أكثر تخصيصًا .أخرى وفي نهاية المطاف، تقدم هذه الدراسة حلاً عمليًا وقابلاً .للمنتجات، ويسهم في اتخاذ قرارات تسويقية أكثر ذكاءً للتفسير وقابلاً للتوسيع لبناء أنظمة توصية تركز على العملاء، مما يتيح لشركات التجارة الإلكترونية تحسين استراتيجيات الاحتفاظ وتعظيم التفاعل.

الكلمات المفتاحية :لايبلر، هندسة -نظام التوصية، تصنيف المشتري المتكرر، الانحدار اللوجستي، تباعد كولباك السمات.

CHAPTER I

INTRODUCTION

1.1. Background

In the rapidly evolving landscape of e-commerce, providing personalized recommendations has become a cornerstone for enhancing user experience and driving sales (Liu, 2022). The sheer volume of products and the diverse preferences of users pose significant challenges to the development of effective recommendation systems. Traditional recommendation algorithms, such as collaborative filtering and content-based methods, have laid the foundation for this field. However, these methods often fall short in capturing the complex and dynamic nature of user preferences and item attributes.

Repeat buyer analysis is crucial for businesses aiming to improve customer retention and maximize lifetime value. Understanding the factors that drive repeat purchases helps companies tailor their marketing strategies, optimize product offerings, and enhance customer satisfaction (Li, 2023). Various statistical and machine learning methods have been employed to predict and analyze repeat buyer behavior, with logistic regression and survival analysis being common approaches. This study explores and analyze repeat buyers in order to build an e-commerce recommendation system.

Researching an e-commerce recommendation system based on repeat buyer analysis can be linked to the teachings of the Quran and Hadith, which emphasize fairness, transparency, and understanding human behavior. In Islam,

conducting ethical business and ensuring the satisfaction of customers is a key principle. For instance, in Surah Al-Baqarah (2:275) that mentions that the trade should be just dan free from exploitation: : "But Allah has permitted trade and has forbidden interest...".

اَلَّذِيْنَ يَأْكُلُوْنَ الرِّبٰوا لَا يَقُوْمُوْنَ اِلَّا كَمَا يَقُوْمُ الَّذِيْ يَتَخَبَّطُهُ الشَّيْطٰنُ مِنَ الْمَسِّ ۗ
ذٰلِكَ بِاَنَّهُمْ قَالُوْٓا اِنَّمَا الْبَيْعُ مِثْلُ الرِّبٰوا ۘوَاَحَلَّ اللّٰهُ الْبَيْعَ وَحَرَّمَ الرِّبٰوا ۗفَمَنْ جَاۤءَهٗ
مَوْعِظَةٌ مِّنْ رَّبِّهٖ فَانْتَهٰى فَلَهٗ مَا سَلَفَ ۗوَاَمْرُهٗٓ اِلَى اللّٰهِ ۗوَمَنْ عَادَ فَاُولٰۤىِٕكَ اَصْحٰبُ النَّارِ ۚ
هُمْ فِيْهَا خٰلِدُوْنَ ﴿٢٧٥﴾

“Those who consume interest will stand on Judgment Day like those driven to madness by Satan's touch. That is because they say, “Trade is no different than interest.” But Allah has permitted trading and forbidden interest.”

This can be applied to developing a recommendation system that helps users by understanding their needs, improving their shopping experience without exploiting them. Additionally, the Prophet Muhammad (PBUH) said, *"The buyer and seller have the option of canceling or confirming the bargain unless they separate" (Sahih al-Bukhari 2110)*, emphasizing transparency in transactions. In this way, analyzing repeat buyers in e-commerce can be seen as understanding and respecting customer needs, aligning with Islamic principles of ethical business practices.

Recent advances in statistical methods and information theory offer promising avenues to enhance the performance of e-commerce recommendation systems. This research explores a novel synergistic approach that integrates

logistic distribution and Kullback-Leibler (KL) Divergence (KLD) to improve recommendation accuracy and relevance. Logistic distribution, with its ability to model binary outcomes, provides a robust framework for estimating the probability of user interactions with items. Meanwhile, KL Divergence, a measure from information theory, quantifies the difference between probability distributions, enabling the system to effectively gauge the disparity between predicted and actual user preferences.

This research aligns with Islamic principles of striving for precision and fairness. Allah warns against those who give less than what is due when measuring or weighing, highlighting the importance of fairness as mentioned in Surah Al-Mutaffifin (83:1-3).

وَيْلٌ لِّلْمُطَفِّفِيْنَ ۙ﴿١﴾

"(1) Woe to those who give less [than due],"

الَّذِيْنَ اِذَا اكْتَالُوْا عَلَى النَّاسِ يَسْتَوْفُوْنَ ۖ﴿٢﴾

"(2) Who, when they take a measure from people, take in full."

وَاِذَا كَالُوْهُمْ اَوْ وَّزَنُوْهُمْ يُخْسِرُوْنَ ؕ﴿٣﴾

"(3) But if they give by measure or by weight to them, they cause loss"

The implication of applying logistic distribution models in e-commerce systems to ensure accuracy and fairness in the distribution of resources. Additionally, the Kullback-Leibler divergence, measures of how one probability distribution diverges from a second expected distribution. The Prophet Muhammad (PBUH) emphasized truth and fairness in all transactions, as seen in the Hadith: *"The truthful and trustworthy merchant is with the Prophets, the*

truthful, and the martyrs" (Tirmidhi 1209). Hence, this research connects to Islamic values of precision, ethical dealings, and ensuring that systems is designed for trade operate justly and fairly.

Logistic regression is a statistical method used primarily for binary classification, which means it predicts a binary outcome (like 0 or 1, yes or no, true or false). Logistic regression is a powerful tool for predicting repeat buyer behavior, helping businesses understand the likelihood of a customer returning for additional purchases. By analyzing customer data—such as purchase frequency, time since the last transaction, average order value, and engagement with marketing—logistic regression assigns a probability to each customer for becoming a repeat buyer. This probability score allows businesses to classify customers as likely or unlikely to make another purchase, guiding targeted marketing strategies. For instance, high-probability repeat buyers might receive loyalty rewards or personalized recommendations, while low-probability customers could be targeted with incentives to re-engage. Logistic regression also highlights which factors most influence repeat buying, such as product preferences or marketing responsiveness, giving businesses actionable insights to refine their customer retention strategies. Simple yet effective, logistic regression offers a way to leverage customer data to boost loyalty, improve retention, and ultimately drive growth.

Kullback-Leibler (KL)-Divergence is a measure from information theory that quantifies the difference between two probability distributions. In repeat buyer analysis, Kullback-Leibler (KL)-Divergence can be used to compare the

distribution of purchase behaviors between different customer segments or over different time periods. By measuring the Divergence, businesses can identify shifts in buying patterns and assess the impact of marketing interventions.

By combining these two powerful tools, we aim to develop an e-commerce recommendation system that not only predicts user behavior with greater precision but also adapts to the evolving landscape of e-commerce. Our approach leverages the strengths of logistic regression in handling binary classification tasks and the capacity of KL Divergence to measure and minimize prediction errors. This synergistic model is expected to offer significant improvements over traditional recommendation techniques, thereby enhancing user satisfaction and engagement.

This study also will delve into the theoretical underpinnings of logistic regression and KL Divergence, outline the proposed methodology, and present empirical evaluations to demonstrate the efficacy of our approach. Through this research, we contribute to the ongoing quest for more intelligent and adaptive recommendation systems in the e-commerce domain.

The integration of logistic regression and Kullback-Leibler (KL) Divergence in recommendation systems offers a multitude of benefits, primarily enhancing the predictive accuracy and adaptability of these systems. Logistic regression is particularly adept at modeling binary outcomes, such as whether a user will interact with or purchase an item, providing a robust framework for estimating probabilities of such events. This capability is crucial for capturing the complex and non-linear relationships between user features and item interactions,

which are common in e-commerce scenarios. The probabilistic nature of logistic regression allows for dynamic updating as new interaction data becomes available, ensuring that the system can quickly adapt to changing user preferences and behaviors.

KL Divergence, on the other hand, plays a critical role in quantifying the difference between predicted probability distributions and actual user preferences. By minimizing this divergence, the model can continuously refine its predictions to better align with real-world data. This method of measuring prediction error is particularly effective in guiding the system to make precise updates, thereby enhancing the overall accuracy of recommendations. Moreover, KL Divergence facilitates adaptive learning by ensuring that the system remains responsive to shifts in user behavior and item trends, maintaining the relevance of recommendations over time.

Personalization and user satisfaction are significantly boosted through this integrated approach. Logistic regression enables the creation of highly personalized models that account for individual user preferences and behaviors, resulting in recommendations that are more aligned with user interests. The confidence scores provided by logistic models further enhance this personalization by allowing the system to prioritize items with higher likelihoods of user engagement. Simultaneously, KL Divergence ensures that the recommendations closely match actual user preferences by minimizing the Divergence between predicted and actual distributions. This leads to a more

satisfying user experience as the system can effectively balance introducing new items (exploration) with recommending known preferred items (exploitation).

Finally, the combination of logistic distribution and KL Divergence addresses the challenge of data sparsity, which is a common issue in e-commerce recommendation systems. Logistic models are well-suited for handling sparse data, often using regularization techniques to prevent overfitting and enhance model robustness. KL Divergence complements this by focusing on the distributional aspects of user interactions, making efficient use of the available data and ensuring that the model updates are both meaningful and impactful. This synergy between logistic regression and KL Divergence results in a more resilient recommendation system capable of delivering accurate, personalized, and dynamic recommendations, ultimately improving user engagement and satisfaction in e-commerce platforms.

1.2. Research Questions

How can the combination of logistic regression and Kullback-Leibler Divergence (KLD) improve the performance of e-commerce recommendation system?

1.3. Objectives

To show and prove that the combination of logistic regression with Kullback-Leibler Divergence (KLD) is able to enhance the performance of recommendation system.

1.4. Research Benefit

This study provides valuable advancements for e-commerce platforms, enhancing recommendation accuracy and user satisfaction. By combining logistic regression with Kullback-Leibler divergence, the approach effectively captures user preferences, ensuring that recommendations are relevant and personalized. This improved precision in recommendations benefits e-commerce stakeholders by increasing conversion rates and customer loyalty, as users are more likely to return to a platform that understands and predicts their needs. Additionally, data scientists and machine learning practitioners gain insights into innovative, data-driven recommendation strategies that blend statistical modeling with information theory. The study's implications extend beyond e-commerce, as industries such as media streaming, online advertising, and digital retail can adopt these methods to enhance user engagement and deliver targeted content. By offering a robust framework for refined, customer-centered recommendations, this research supports a wide array of stakeholders in improving user experiences and driving business growth.

1.5. Scope of Problems

The research examines the scope of challenges in e-commerce recommendation systems by analyzing customer behavior data from PT. Agile Service Solution Jakarta, an e-commerce data agency, collected during June and July 2022, focusing on improving the accuracy and personalization of product recommendations.

CHAPTER II

LITERATURE REVIEW

2.1. E-Commerce Repeat Buyer and Recommender System

A Research conducted by Zang and Wang (2021) proposed an improved deep forest model, and the interactive behavior characteristics of users and goods are added into the original feature model to predict the repurchase behavior of e-commerce consumers. Improved deep forest was used to compare with other methods, and the improved deep forest has the best behavior prediction performance of e-commerce consumers.

Noori (2021) conducted a research to propose a new framework for categorizing and predicting customer sentiments using Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), C4.5 and K-Nearest Neighbor (K-NN). The result showed among algorithms that are used, the Decision Tree provided better results. And the most important factors influencing the great customer experience were extracted with the help of the Decision Tree.

A Study by Dong et al (2022) proposed a BERT-MLP prediction model that uses large-scale data unsupervised pre-training and small amount of labeled data fine-tuning with the result showed the accuracy of the BERT-MLP model is better than the baseline model.

Suhanda et al (2022) with a research to propose how to determine and analyze customer loyalty, customer trust and customer satisfaction in order to

monitor customers at the company easier by applying Random Forest. The result of feature evaluation shows that customer_activity has the highest influence on customer retention, followed by subtotal and quantity.

Zang et al (2022) did a study to build a prediction model of repeat purchasers and introduce the synthetic minority oversampling technique (SMOTE) algorithm to solve the data imbalance problem and improve prediction performance using SMOTE, classical classifiers including factorization machine and logistic regression, and ensemble learning classifiers including extreme gradient boosting, and light gradient boosting machine machines. The results of this study showed that through a series of innovations such as data imbalance processing, feature engineering, and fusion models, the model area under curve (AUC) value is improved by 0.01161.

A research by Kuric et al (2023) to study the impact of low-level interaction data capturing user behavior more precisely and implemented Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, Multilayer Perceptron, Support Vector Machine. They got the result that showed inclusion of interaction data improves the prediction performance compared to the baseline non-interaction feature set.

Li (2023) analyzed customer shopping behavior data, how to tap the potential information value of customers and how to explore customer needs and behavior rules using hybrid algorithm of data roughness and decision tree algorithm. They got the results showed that method can produce an ideal decision

tree model, and it was easier to understand and interpret the rules of customer behavior extraction, so as to understand more accurate customer behavior.

A research conducted by Satu and Islam (2023) aimed to propose a machine-learning model that employs multiple data analytic and machine learning techniques to manipulate customer records and predict their buying intention more precisely using NB, RF, DT, Simple CART, CSForest, ForestPA, NBTree, SysFor, LR, LMT, SMO, SMO, Bagging, and LibSVM. From their research showed that Random Forest was found as the stable classifier that produced more feasible results for any online shoppers' buying instances.

Kc et al (2023) tried to present a data-driven approach to predict e-commerce sales by leveraging machine learning techniques and to develop the prediction model, historical sales data from various ecommerce platforms was collected and preprocessed to ensure data quality. Using regression models, ensemble methods, and deep learning models the result showed by leveraging machine learning techniques and analyzing key customer behavior features such as recency, frequency, and monetary value, accurate predictions can be made regarding whether a customer is likely to make a purchase within a specified timeframe. And through the application of RFM segmentation and feature engineering, valuable insights into customer segmentation and purchase patterns can be obtained.

Liu (2022) conducted a research to personalize recommendation technology of e-commerce which is deeply analyzed the related technologies and algorithms of the e-commerce recommendation system and proposed the latest architecture of the e-commerce recommendation system according to the current development status of the e-commerce recommendation system. Structural Equation Model was used with the result showed the adjustment of the types of customer evaluations and the imperfect calculation lead to the unscientific calculation of the customer evaluation part of the recommendation system. The fusion of recommendation lists generated by several methods is relatively rigid and needs to be further improved.

Helfianur and Baizal (2022) did a research to explores the scope of recommendations based on frequent item sets by applying an apriori algorithm to find items or products that are frequently purchased and frequently searched for. Using Apriori Algorithm, it provided better recommendation results than without using the apriori algorithm or only using content-based filtering. The apriori algorithm also provides the association rule value that has the highest accuracy and increases the user's confidence value for the recommended item.

A research by Zhang et al (2022) aimed to propose a novel approach called ImDetector to detect fraudulent reviewers while handling data imbalance based on weighted latent Dirichlet allocation (LDA) and Kullback–Leibler (KL) divergence and showed the result that the proposed ImDetector approach is superior to the state-of-the-art techniques used for fraudulent reviewer detection.

A study by Pleskach (2023) aimed to present some models and strategies of consumer behavior in electronic commerce systems, to analize of e-commerce recommendation systems and explore the impact of machine learning and artificial intelligence on e-commerce recommendation systems and e-commerce systems in general. Implemented Deep learning, natural language processing (NLP) and Reinforcement learning, they got that hybrid recommendation systems integrate collaborative filtering, content-based filtering, and knowledge-based approaches to leverage the strengths of each method. It helped overcome limitations and provides more accurate and diverse recommendations.

Legito (2023) did a study to propose and implements an e-commerce product recommendation system that combines Case Based Reasoning (CBR) and K-Means Clustering algorithms and it revealed that the combined approach of CBR and K-Means Clustering can improve the performance of e-commerce product recommendations, ensure the accuracy of recommendations, and produce a more satisfying shopping experience for users.

A study by Feng (2023) aimed to present a methodology to quantify consumers' self-identities to help examine the relationship between these psychological cues and decision-making in an e-commerce context and focusing on the projective self. Using Charmaz's approach, the got the fact that the study that applied grounded theory and literary synthesis solved two issues which were as a first issue, recommendation systems suffer from a dearth of psychological frameworks that serve as its foundational principles. Second, the technology behind recommendation algorithms is lacking in terms of actual metrics.

A research conducted by Santosa et al (2023) tried to create a recommendation system using hybrid method, where this method combines more than 1 method to create a list of recommendations so that it will cover the shortcomings of each method using hybrid methods with KNN User Based Collaborative Filtering and Content Based Filtering. The hybrid method in this study was able to overcome the cold start the problem by using switching and mixed methods, namely by not using the collaborative filtering model on new user recommendations or users who have small interactions. New users will get recommendations from a combination of popularity-based and content-based filtering models.

Nurdin and Abidin (2023) did a study to find out what factors affect customer loyalty to Shopee e-commerce as well as test how much influence the quality of Shopee's e-commerce recommendation system have on customer loyalty with user trust as mediation variables. Implementing the quantitative approach using cognition affective behavior theory, the results of this study show that there are variables that mediate between the relationship between recommendation quality to loyalty, namely trust. Trust affects loyalty with a slightly greater influence. Trust plays a partial mediation role in supporting the recommendation quality and loyalty relationship.

Loukili et al (2023) conducted a research to develop an algorithm to suggest personal recommendations to customers using association rules via the Frequent Pattern-Growth algorithm with the result that the evaluation allowed for

an estimation of the potential revenue increase that could be achieved with the implementation of the proposed recommendation system.

A study by Griva et al (2024) tried to develop a two-stage business analytics approach that introduces a combination of geographic and behavioural customer segmentation using clustering and machine learning with Latent Dirichlet Allocation model, and feature selection techniques. The findings indicated that the customer base of the company and the four online retailers can be segmented into seven behavioral customer segments, and each one contains customers ordering specific products. Also it identified distinct behaviors in rural and urban regions considering the geographic segments.

A research by He at al (2024) aimed to develop a moderated mediation model, simultaneously considering the roles of a user's feeling state and shopping goal using hybrid filtering technique and the results showed that there is an interaction effect between shopping goals and types of recommendation (diversity and accuracy) on user satisfaction, and evaluated the mediating role of feeling right and psychological reactance for a better understanding of this interactive relationship.

Chiou-Wei et al (2024) conducted a study to analyze the widely used recommendation algorithms in the booming financial market attracts more people to invest in funds for its relatively low risk and high returns and the result showed that the proposed recommendation model has a better recommendation performance than existing models to meet users' demands.

2.2. Theoretical Framework

From the literature study carried out, a theoretical framework was obtained as seen in Figure 2.1.

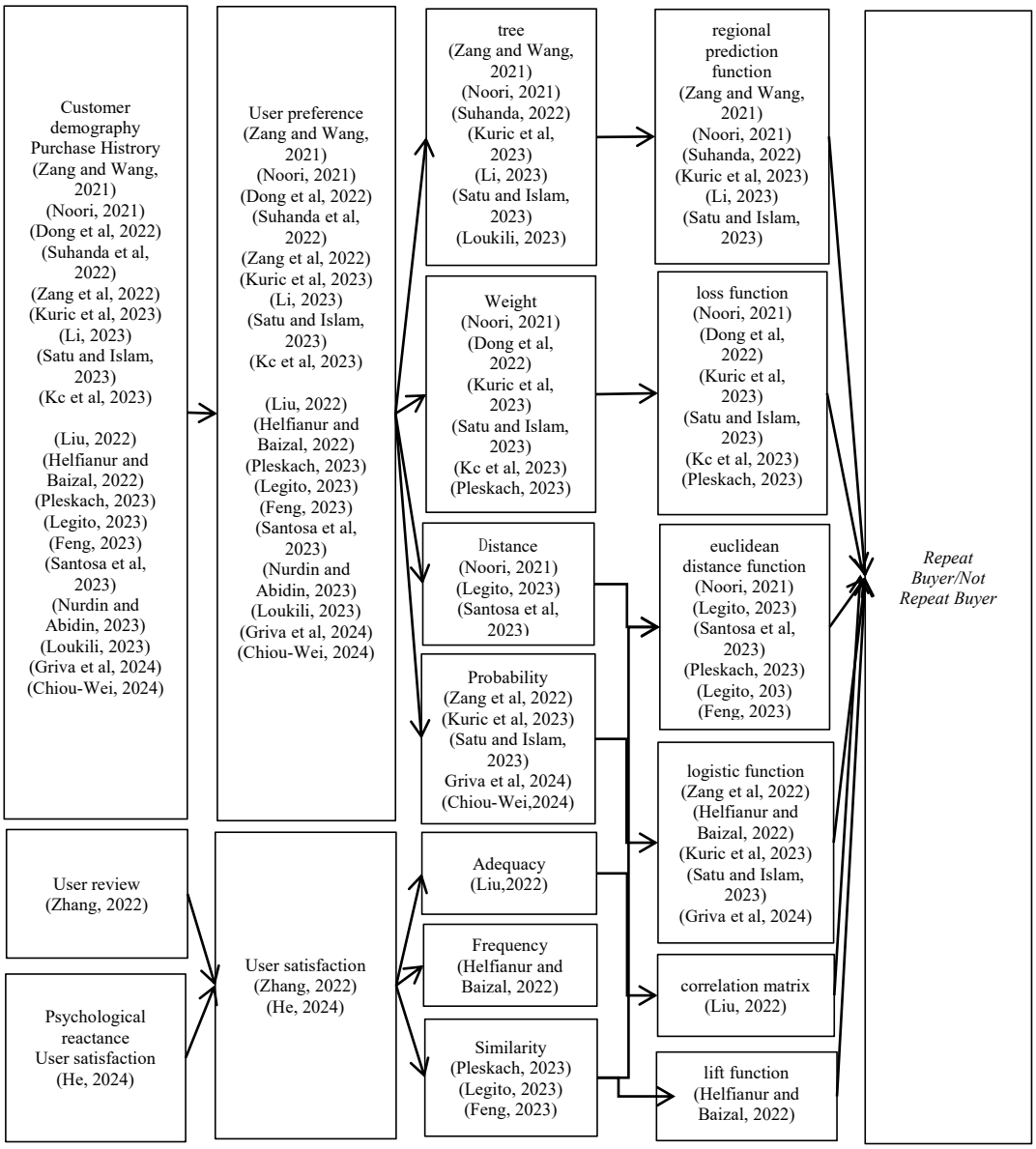

Figure 2.1. Theoretical Framework

Figure 2.1 shows the information obtained from input, those are Customer Information, Product Information, Transaction History and Historical Behavior Information of Buyers. From the features selection process, features are obtained,

those are recency and frequency. And based on the various methods used by researchers, four types of measurement base methods were obtained, which are tree, distance, probability and weight. These studies aim to get the best models for building an e-commerce recommendation system based on repeat buyers. From the theoretical framework above, the path that best suits to analyze and to get the best methods for the recommendation system based on repeat buyers is using a distance-based method.

CHAPTER III

RESEARCH METHODOLOGY

3.1. Research Design

Research design can be seen in Figure 3.1. The elaboration of each step is given in the following sub section.

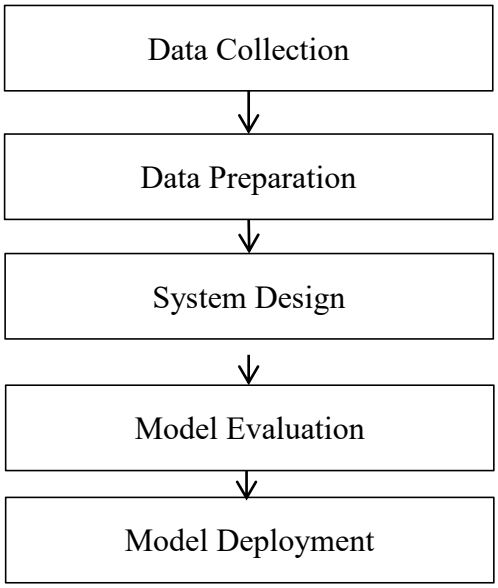

Figure 3.1. Research Design

3.1.1. Data Collection

A data collection process related to the e-commerce repeat buyer is carried out. The data will be taken from an e-commerce data set from PT. Agile Service Solution Jakarta, an e-commerce data agency, collected during June and July 2022. A data attribute is also called attribute or feature, is a specific characteristic of a data entity. Data attributes in Data Science or Machine Learning are variables that

describe about the data record or instance. Each attribute holds a particular type of data and represents one aspect of the entity is being described. Below is raw attributes that will be processed in this study.

Table 3.1 Raw Attributes

Attribute Name	Data Type	Description
Trx_Date	Date	Transaction Date
Cust ID	String	Customer ID
Age	Integer	Customer Age
Gender	String	Customer Gender
Area	String	Area
Product_Id	String	Product ID
Product_Name	String	Product Name
Product_Category	String	Product_Category
Amount	Integer	Amount of product purchased

Based on the attributes, we can define the definition of what constitutes a repeat buyer. This might be based on the number of purchases, the frequency of purchases, or the time frame within which the purchases were made. A repeat buyer is defined as a customer who has made at least two purchases within the last six months. The data example can be seen in Table 3.2.

Table 3.2. Data Example

Trx_Date	Cust_ID	Age	Gender	AREA	Product_ID	Product_Name	Category	Amount
6/1/2022	CTR0200437	49	Laki-laki	Aceh	HP-PV15	Laptop HP Pavilion 15	Computer	1
6/1/2022	CTR0200460	18	Perempuan	Bandung	192021232	Hair Mask CDE	Hair Treatment	1
6/1/2022	CTR0200553	42	Perempuan	Banjarmasin	192021232	Hair Mask CDE	Hair Treatment	7
6/1/2022	CTR0200581	42	Laki-laki	Batam	IKE-KALLAX	Rak Buku IKEA Kallax	Home Appliances	5
6/1/2022	CTR0200563	39	Perempuan	Cirebon	232425275	Perfume OPQ	Cosmetics	1

3.1.2. Data Preparation

Data Preparation involves transforming raw data into a clean and structured format suitable for modeling. This process includes:

1. Handling missing values: Identifying and treat missing or null values in the dataset. This may include removing rows with excessive missing data or imputing values using statistical methods (e.g., mean, median, mode).

2. Checking for duplicate rows: Detect and removing any duplicate entries in the dataset to avoid redundancy and ensure the uniqueness of each data point.

3. Encoding categorical features: Converting categorical variables such as Gender and AREA into numerical values using techniques like: Label Encoding (e.g., Male = 0, Female = 1), One-Hot Encoding (e.g., AREA_A = 1, AREA_B = 0)

4. Normalize numerical features: Scaling numerical columns to a consistent range (e.g., 0 to 1) using normalization or standardization techniques. This is essential for models sensitive to value magnitude, such as logistic regression.

5. Converting Trx_Date to Datetime Format: Transforming the Trx_Date column from a string to a proper datetime format. This allows for the extraction of time-based features.

3.1.3. System Design

Figure 3.2 shows the step by step key activities from processing the data to generate the recommender system.

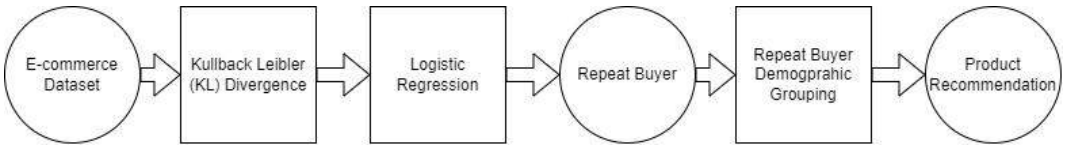

Figure 3.2. System Design

The proposed system design presents a synergistic framework for enhancing product recommendation in e-commerce platforms by integrating repeat buyer classification into the recommender system workflow. This design combines logistic regression modeling with Kullback-Leibler (KL) divergence-based feature engineering to improve the identification of repeat buyers, which subsequently refines the recommendation process. The system architecture comprises the following components:

1. E-commerce Dataset Input

The system begins by collecting historical e-commerce transaction data, which includes customer demographics, product purchase history, transaction dates, and other relevant attributes. This dataset serves as the foundation for subsequent feature engineering and model training.

2. KL Divergence-Based Feature Engineering

Kullback-Leibler divergence is employed to evaluate the informativeness of individual features in distinguishing between repeat and non-repeat buyers. Features with high KL divergence values are retained to enhance the quality of the input variables for the classification model.

KL-Divergence (Kullback-Leibler Divergence) is typically used in probability distributions to measure how one probability distribution diverges from a reference distribution. However, when using it in the context of pre-

processing for logistic regression, it can be applied to identify important features by comparing how the feature distributions differ for the two classes (repeat buyers vs. non-repeat buyers). The idea is to find features that provide the most discriminatory power between the two classes.

KL-Divergence formula : Data mining: Concepts and techniques (3rd ed) (Han, J., Kamber, M., & Pei, J, 2011: Chapter Section: 2.4.8-Kullback-Leibler Divergence)

$$D_{KL}(P \| Q) = \sum_{z \notin X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \qquad (3.1)$$

In binary classification (e.g., repeat buyer vs. non-repeat buyer), it simplifiesas follows:

$$DKL(P||Q) = P_1.log\left(\frac{P_1}{Q_1}\right) + P_0.log\left(\frac{P_0}{Q_0}\right) \qquad (3.2)$$

where:

P_1, P_0 = true distribution (e.g., class proportions in each feature bin)

Q_1, Q_0 = reference distribution (e.g., overall class distribution)

The behavior of Kullback–Leibler (KL) Divergence provides valuable insight into feature relevance for classification tasks. A high KL Divergence score indicates that the class distribution within a given feature bin deviates significantly from the overall class distribution, suggesting that the feature provides strong discriminatory power between classes, in this case, repeat and non-repeat buyers. Such features are considered highly informative and are particularly valuable for improving model performance. Conversely, a low KL Divergence score implies that the feature's distribution across classes closely mirrors the global class distribution, offering little to no additional information for

distinguishing between target outcomes. As such, features with low KL values are generally less useful for classification and may be excluded or deprioritized during feature selection.

The behavior of Kullback–Leibler (KL) Divergence as shown in Figure 3.3. provides valuable insight into feature relevance for classification tasks. A high KL Divergence score indicates that the class distribution within a given feature bin deviates significantly from the overall class distribution, suggesting that the feature provides strong discriminatory power between classes, in this case, repeat and non-repeat buyers. Such features are considered highly informative and are particularly valuable for improving model performance. Conversely, a low KL Divergence score implies that the feature's distribution across classes closely mirrors the global class distribution, offering little to no additional information for distinguishing between target outcomes. As such, features with low KL values are generally less useful for classification and may be excluded or deprioritized during feature selection.

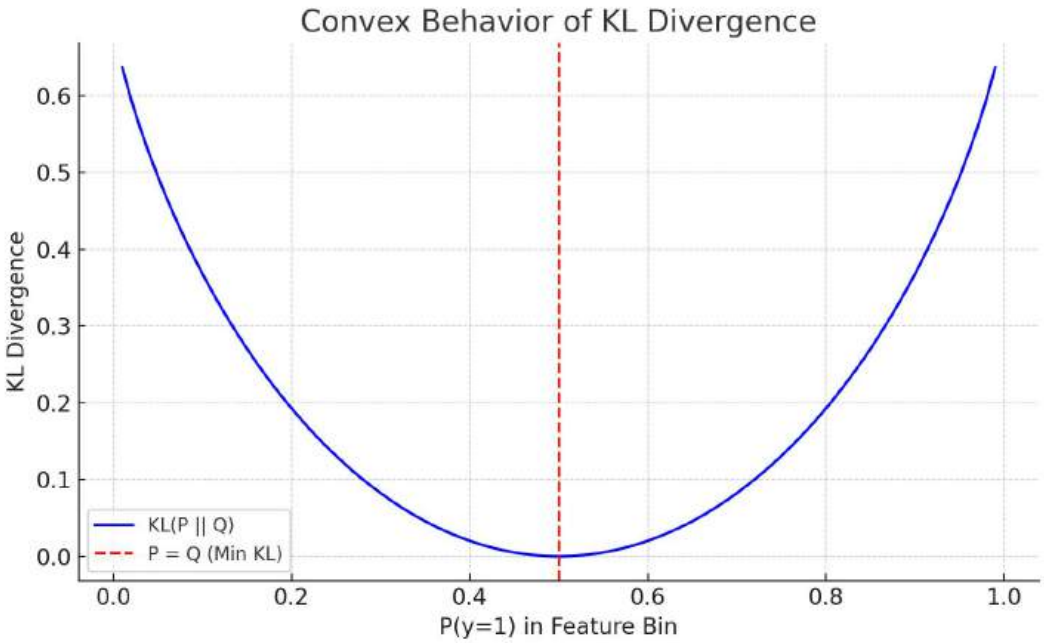

Figure 3.3. Convex Behavior of KL Divergence

In the context of repeat buyer classification, each feature is evaluated based on how much its class distribution diverges from the overall buyer distribution when conditioned on that feature. This process begins by discretizing the feature into meaningful intervals or bins (e.g., grouping Age into defined age ranges). Within each bin, the conditional class probabilities P(y=1 | bin) and P(y=0 | bin) are computed to reflect the distribution of repeat and non-repeat buyers. These bin-specific probabilities are then compared to the overall class probabilities Q(y=1) and Q(y=0) in the dataset. For each bin, the KL Divergence formula is applied to measure the difference between the local and global class distributions. The KL values from all bins are then summed to produce a total divergence score for the feature, which reflects its overall discriminative power in the classification task.

The pseudocode of Kullback-Leibler Divergence formula can be seen below and followed by the explanation of each step in implementing Kullback-Leibler Divergence as feature engineering, and the manual KL Divergence process can be seen in Appendix B.

```
Function KL_Divergence(P, Q):
Step 1. Convert P and Q to float arrays
   // (This step ensures that P and Q are in numerical form that can be used for
   computation)
Step 2: Normalize P and Q to ensure they sum to 1
   // (This ensures that the values are valid probability distributions)
   P = P / sum(P)  // Normalize P
   Q = Q / sum(Q)  // Normalize Q
Step 3: Clip small values in P and Q to avoid log(0) or division by 0
   // (This prevents numerical issues like log(0) or division by zero)
   epsilon = 1e-10  // Small value to avoid log(0)
   For each i in P:
      If P[i] < epsilon: P[i] = epsilon
      If Q[i] < epsilon: Q[i] = epsilon
Step 4: Initialize kl = 0
Step 5: For each i in P:
   For each i in P:
      kl += P[i] * log(P[i] / Q[i])
Step 6: Output: Return the calculated KL Divergence
```

3. Logistic Regression Modeling

Using the selected features, a logistic regression model is trained to classify customers based on their likelihood of becoming repeat buyers. Logistic regression is chosen for its interpretability and efficiency in binary classification tasks. Implementing logistic regression models is to estimate the probability of user interactions with products. Logistic regression models the probability that a given input belongs to a certain class. In this study, we want to predict the probability that a customer is a repeat buyer (`repeat_buyer = 1`). Logistic

regression assumes that this probability can be modeled as a function of the input features (the data attributes). Below is logistic regression formula:

$$P(y \mid X) = \frac{1}{1 + e^{(-\beta_0 + B_1 X_1 + ... + \beta_n X_n)}} \qquad (3.3)$$

where :

P(y): is the probability that a customer is a repeat buyer.

X : input features

β : coefficients to be estimated

β_1 : intercept (bias term/a constant value that represents the baseline log-odds of the outcome when all feature values are zero)

$\beta_1, \beta_2, ..., \beta_n$: the coefficients (weights) associated with the respective input features $x_1, x_2, ..., x_n$

$x_1, x_2, ..., x_n$: values of the input features for the particular customer

e: the base of the natural logarithm

A. Implementation of Data Attributes in Logistic Regression Formula

1. Intercept (β_0) : This is a constant term that shifts the decision boundary but is not associated with any input feature. It represents the baseline odds of a customer being a repeat buyer, assuming all other feature values are zero.
2. Coefficients ($\beta_1 \dots \beta_n$)

   Age (x_1) : The coefficient β_1 represents the influence of the customer's age on the probability of being a repeat buyer (β_1 x Age).

Gender (x_2) : If encoded as binary variable (0 for male, 1 for female), the coefficient β_1 measures the influence of gender on the likelihood of repeat buying (β_2 x Gender).

Area ($x_3, x_4, ...$) : If the area is one-hot encoded (a technique used to convert categorical variables into a numerical format that machine learning models can process) into several binary variables (different cities or regions), each area will have its own coefficient. For example, if you have areas A_1, A_2, A_3, and so on the equation will include terms like: β_3 x Area1 + β_4 x Area2 + …

Category ($x_5, x_6, ...$): If the product_category is one-hot encoded, each category gets a coefficient. For example, clothing, elecronics, etc would be represented by terms such as: β_5 x product_category$_1$ + β_6 x product_category$_2$ + …

Amount (x_7): This represents the total amount spent by a customer. The coefficient β_7 measures how much spending influences repeat buying (β_7 x amount)

3. Final Logistic Regression Formula (Manual logistic regression process can be seen in Appendix A).

The pseudocode of machine learning logistic regression model (using Maximum Likelihood Estimation) can be seen as follows:

Input: Training data (X), true labels (y)
Let:
P(y) = probability that a customer is a repeat buyer
$x_1, x_2, ..., x_n$ = input features for a customer
β_0 = intercept (bias term)
$\beta_1, \beta_2, ..., \beta_n$ = coefficients (weights) for each feature
e = base of the natural logarithm

1. For $i \leftarrow 1$ to k (number of training samples):
2. For each training data instance $x_i = (x_1, x_2, ..., x_n)$:
3. Compute predicted probability:
$P(y_i) \leftarrow 1 / (1 + e^{(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n))})$
4. Set target for regression:
$z_i \leftarrow (y_i - P(y_i)) / [P(y_i) \times (1 - P(y_i))]$
5. Initialize weight for instance x_i as:
$w_i \leftarrow P(y_i) \times (1 - P(y_i))$
6. Fit a function $f(x)$ to the data using class value z_i and weights w_i
Classification Label Decision
7. Assign:
- Class label = 1 if $P(y_i) > 0.5$ (repeat buyer)
- Class label = 0 otherwise (non-repeat buyer)

4. Repeat Buyer Classification

The trained logistic regression model outputs a binary classification for each user: repeat buyer or non-repeat buyer. This classification serves as an essential input for guiding the recommendation strategy.

5. Repeat Buyers Demographic Based Grouping

The classification results are integrated into the recommendation process. Customers predicted as repeat buyers may receive personalized suggestions based on their demographic and preferences, while new or non-repeat buyers may be targeted with onboarding-friendly recommendations or promotions.

6. Product Recommendation Output

The final output is a list of recommended products tailored to the customer profile. These recommendations are optimized based on the user's predicted repeat-buying behavior to increase engagement and purchase likelihood.New or existing customers interact with the platform and receive recommendations that are informed by the predictive model, creating a more personalized and data-driven shopping experience.

3.1.4. Model Evaluation

Model evaluation is a process to assess how well a predictive model performs on data that is used in the model training step. Model evaluation is an important step to make sure that the model generalizes well to new, unseen data and overfitting. It should define clear criteria, selecting appropriate methods, preparing data and conducting thorough evaluation to make sure that the system is robust, reliable and effective to meet the goal. The process is repeated several times with different subsets of data used for training and validation, and the results are averaged to obtain an overall assessment of the model's performance. This should help the study about e-commerce recommender system in order to develop a model that identify repeat buyers accurately and provides valuable business insights.

Based on the research design and system design that are explain above, the steps of doing the system valuation will explain in detail as follows.

1. Defining Criteria

Defining criteria in system evaluation means specify the standards and metrics that will be used to assess the performance of the research model. The criteria will ensure that the evaluation is objective, comprehensive and aligned with the research goals. The metrics that will be used are accuracy, precision, recall, F1-score, ROC.

2. Conducting Evaluation

Execute the evaluation process according to the chosen method, which will apply performance metrics as evaluation. This execution involves running the model with the prepared data and collecting performance metrics.

3. Validation

Validation the system's performance by verifying that the results are consistent and validate it on an independent dataset to ensure generalization. It can be perform by validating on an independent test set or new customers data to ensure that the repeat buyer prediction model generalized well to unseen data.

The evaluation outcomes are visualized in Figure 3.4, which presents the block diagram outlining each step of the evaluation process, from baseline modeling to KL-Divergence-based refinement and final performance comparison, highlighting the progression and impact of each methodological stage.

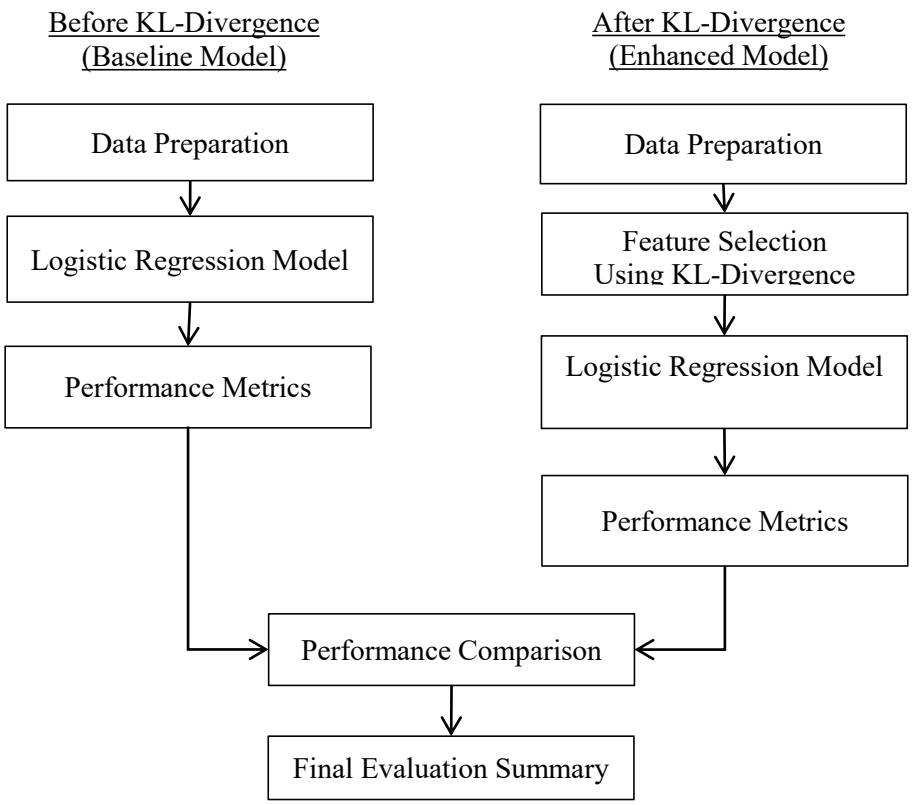

Figure 3.4. Evaluation Steps

3.1.5. Model Deployment

Model deployment for an e-commerce recommender system based on repeat buyer classification involves integrating the trained logistic regression model into the live platform to support real-time decision-making. Once the model identifies whether a customer is a repeat buyer, the system uses this classification to personalize product recommendations. For example, repeat buyers may receive suggestions based on past purchases and loyalty behavior, while new or one-time buyers may be targeted with popular or promotional items.

Model deployment in an e-commerce recommender system based on repeat buyer classification involves two key components: prediction and

recommendation. First, the trained logistic regression model is used to predict the likelihood of user interactions with various products, including estimated purchase amount and expected purchase period. These predictions help classify users as repeat or non-repeat buyers. Next, a recommendation algorithm is developed to generate personalized product suggestions based on the predicted probabilities. This algorithm ensures an optimal balance between exploration, introducing new or diverse items, and exploitation, recommending products known to align with the user's past preferences, particularly those of repeat buyers.

Figure 3.5 shows the process to generate the recommender system phase based on repeat buyers using logistic regression model with KL divergence as feature engineering.

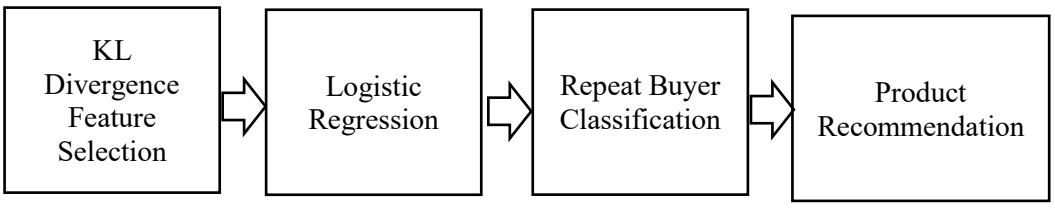

Figure 3.5. E-commerce Recommender System

3.2. Conceptual Framework

From the literature study carried out and because of the need to have methods that can support the determination of the design of e-commerce platform development, and to support e-commerce marketing strategies by having a credible e-commerce recommendation system. Therefore, it can help to determine the products to be offered on the main screen that are tailored to customer

shopping characteristics, especially from repeat buyers. The conceptual framework is shown in Figure 3.6.

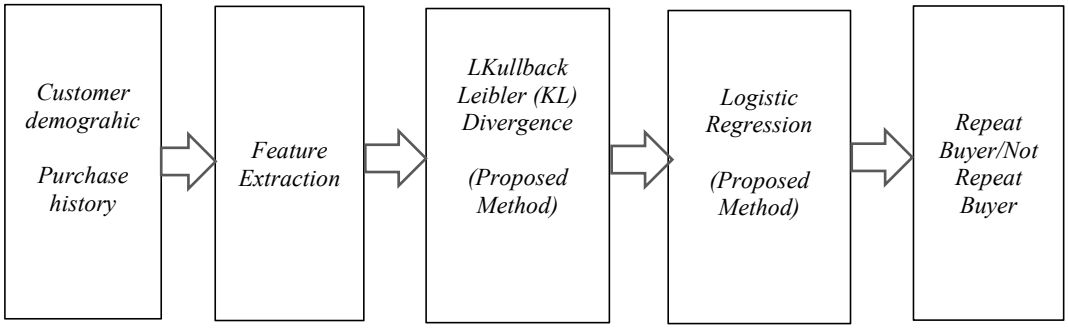

Figure 3.6. Conceptual Framework

In this research, the author offers a logistic distribution-based method to analyze the customer repeat buyer and the Logistic Regression as the machine learning model. Kullback-Leibler (KL) Divergence is employed as pre-processing to identify customer segmentation based on their purchasing behavior by calculating KL-Divergence between repeat buyers and one-time buyers. KL-Divergence has ability to provide a comprehensive measure of how the distribution of repeat buying behaviors differs from other groups and it can capture broader patterns and trends in purchasing behavior. Both methods are implemented to get the best result to have a credible e-commerce recommendation system based on repeat buyers.

3.3. Repeat Buyer

Repeat buyers analysis is a vital aspect of developing a robust e-commerce recommendation system. It involves scrutinizing the purchasing behaviors of customers who return to make multiple purchases over time. Understanding what

drives these repeat purchases helps in tailoring recommender systems to enhance customer retention, boost lifetime value, and increase overall sales. By focusing on repeat buyers, businesses can gain insights into which products and features promote customer loyalty, thereby optimizing their recommendation algorithms for better performance.

The key metrics in repeat buyers analysis include purchase frequency, recency, monetary value, customer lifetime value (CLV), and retention rate. Purchase frequency measures how often a customer makes a purchase within a specific period, indicating their loyalty and satisfaction. Recency tracks the time since a customer's last purchase, with more recent purchasers often being more engaged. Monetary value assesses the total spending of a customer, highlighting those who are particularly valuable. CLV predicts the total revenue a business can expect from a customer over their relationship, focusing on maximizing engagement with high-value customers. Retention rate measures the percentage of customers making repeat purchases over a specified period, indicating effective customer engagement strategies.

Data collection is the foundation of repeat buyers analysis. This involves gathering comprehensive data on customer interactions, including transaction details, customer demographics, and behavioral data. Transaction data encompasses information like product IDs, categories, quantities, and timestamps. Customer demographics provide context about the users, such as age, gender, and location. Behavioral data includes browsing history, product views, search queries,

and interactions with marketing campaigns. This rich dataset allows for a detailed understanding of customer behavior and preferences.

Model development and training are crucial steps in leveraging this data. Feature engineering involves creating meaningful features from raw data, capturing user behavior and preferences. For instance, features like average purchase interval, favorite categories, and seasonal buying patterns can be extracted. Customers are then segmented based on their buying behavior, such as frequent buyers, occasional buyers, and one-time buyers, allowing for tailored recommendation strategies. Predictive modeling techniques, such as logistic regression, decision trees, and neural networks, analyze historical data to identify key predictors of repeat purchases and generate personalized recommendations.

Personalization strategies, informed by repeat buyers analysis, enhance the recommendation system. Targeted recommendations suggest products aligned with the purchase history and preferences of repeat buyers. For example, a customer frequently buying books might receive recommendations for new releases in their favorite genres. Cross-selling and up-selling strategies recommend complementary or higher-value items based on past purchases, such as accessories for electronics or premium versions of previously bought products. Integrating loyalty program data provides exclusive offers and rewards to repeat buyers, encouraging continued engagement. Dynamic personalization updates recommendations in real-time based on recent interactions, ensuring they remain relevant and appealing.

Continuous evaluation and optimization of the recommendation system are essential to ensure its effectiveness. Repeat purchase rate measures how often recommended products lead to repeat purchases, and revenue growth assesses the increase in revenue attributed to effective recommendations. By analyzing these metrics, businesses can fine-tune their recommendation algorithms, enhancing personalization and creating a more engaging and profitable e-commerce platform.

3.4. Recommender System

An e-commerce recommender system is an advanced technological solution designed to enhance the online shopping experience by providing personalized product suggestions to users. These systems leverage data-driven algorithms to analyze user behavior, preferences, and interactions with products to deliver recommendations that are tailored to individual tastes and needs. By understanding what users have previously purchased, viewed, or liked, these systems can predict and suggest items that are likely to interest them, thereby increasing the likelihood of engagement and conversion.

This study uses two type of recommender systems, which are :

A. Hybrid systems combine multiple techniques to overcome the limitations of individual methods and improve recommendation accuracy. The system that is built here can be classified as a hybrid recommendation system that uses probabilistic modeling (logistic distribution) and information-theoretic techniques (KL-divergence) to enhance the recommendation process. The KL-divergence as a pre-processing step ensures the input data is well-distributed

and informative, improving the performance of the logistic-based recommendation model.

B. Demographic recommender system is a type of recommender system that suggests products, services, or content to users based on their demographic characteristics, here using age, gender, location.

Machine learning models are trained on this historical data to identify patterns and relationships between users and products. Once trained, these models generate personalized recommendations in real-time as users interact with the platform. Continuous evaluation using metrics like accuracy, precision, recall, and user satisfaction helps fine-tune the models and improve the system over time.

Implementing a recommender system comes with several challenges, such as the cold start problem, where there is insufficient data on new users or products, and scalability, which requires the system to handle large volumes of data and interactions efficiently. Additionally, data sparsity, where many users interact with only a small subset of products, can complicate the recommendation process. Balancing the trade-off between recommending highly relevant items and introducing diverse, novel items to keep the user experience engaging is also crucial. Advanced techniques like matrix factorization, deep learning, and natural language processing are often employed to address these challenges.

In practice, e-commerce giants like Amazon, Netflix, and Spotify have successfully implemented sophisticated recommender systems that drive significant portions of their revenue. For instance, Amazon uses a hybrid approach to recommend products based on a user's browsing history, past

purchases, and the behavior of similar users. Netflix's recommender system suggests movies and TV shows by analyzing viewing history and ratings, while Spotify curates personalized playlists and song recommendations based on listening habits. These real-world examples demonstrate the profound impact of recommender systems on user engagement and satisfaction, highlighting their essential role in the competitive landscape of e-commerce.

CHAPTER IV

REPEAT BUYER CLASSIFICATION WITH LOGISTIC REGRESSION

4.1. Repeat Buyer Classification

Repeat buyer classification is the process of identifying and categorizing customers based on their likelihood of making additional purchases after their initial transaction. In e-commerce and retail analytics, this classification helps businesses distinguish between one-time buyers and repeat customers, who are often more valuable due to their long-term engagement and higher lifetime value. Repeat buyer classification with machine learning involves using algorithms to predict whether a customer will make a future purchase based on historical behavioral and transaction data. The key component in doing the repeat buyer classification are: data collection and preparation, feature selection, model selection, model training and evaluation. The raw attributes in Table 3.1 will be processed and analyzed for repeat buyer classification case study. In Figure 4.1. below the flow to do the repeat buyer classification.

Figure 4.1. Repeat Buyer Classification with Logistic Regression

4.2. Data Cleaning and Preprocessing

The dataset that will be used and processed in this case study consists of 1000 records consisting of columns as can be seen in Figure 4.2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   No            1000 non-null   int64
 1   Trx_Date      1000 non-null   object
 2   Cust_ID       1000 non-null   object
 3   Age           1000 non-null   int64
 4   Gender        1000 non-null   object
 5   AREA          1000 non-null   object
 6   Product_ID    1000 non-null   object
 7   Product_Name  1000 non-null   object
 8   Category      1000 non-null   object
 9   Amount        1000 non-null   int64
dtypes: int64(3), object(7)
memory usage: 78.3+ KB
```

Figure 4.2. Dataset Info

This raw data is cleaned from null data and duplicate rows. After the data is cleaned, preprocessing is carried out by converting the Trx_Date column to datetime format to detect detail transaction time an duration and the result is shown in Table 4.1.

Table 4.1. Converting Trx_Date Column to Datetime Format

Trx_Date	Cust_ID	Age	Gender	AREA	Product_ID	Product_Name	Category	Amount	YearMonth
2022-06-01 00:00:00	CTR0200437	49	Laki-laki	Aceh	HP-PV15	Laptop HP Pavilion 15	Computer	1	2022-06
2022-06-01 00:00:00	CTR0200460	18	Perempuan	Bandung	192021232	Hair Mask CDE	Hair Treatment	1	2022-06
2022-06-01 00:00:00	CTR0200553	42	Perempuan	Banjarmasin	192021232	Hair Mask CDE	Hair Treatment	7	2022-06
2022-06-01 00:00:00	CTR0200581	42	Laki-laki	Batam	IKE-KALLAX	Rak Buku IKEA Kallax	Home Appliances	5	2022-06
2022-06-01 00:00:00	CTR0200563	39	Perempuan	Cirebon	232425275	Perfume OPQ	Cosmetics	1	2022-06

4.3. Feature Selection

Data analysis is more simple by continuing to combine customer-level information once the data has been acquired in date and time format. The process begins by aggregating customer-level features from transactional data. Key temporal attributes such as the first and last transaction dates and total transaction count are computed to capture the purchasing timeline for each customer. The number of unique months in which a customer transacted is then calculated to identify purchasing frequency. Following this, the total and average quantity purchased across all transactions is derived, along with the number of unique product categories purchased, providing insight into consumer behavior and diversity of interest. The resulting aggregation often produces MultiIndex columns, which are flattened for easier data handling. Next, customers are labeled as repeat buyers if their Distinct_Periods (representing unique transaction months) exceed one, marking them as returning customers. Finally, demographic information such as Age, Gender, and AREA is merged back into the aggregated dataset to enrich the customer profiles for further analysis and modeling. Table 4.2 shows the result.

Table 4.2. Data Sample After Aggregate Customer-level Features

Cust_ ID	First _Trx	Last_ Trx	Trx_ Count	Distinct_ Periods	Total_Q uantity	Avg_Qu antity	Unique_C ategories	Repeat_ Buyer	Age	Gender	AREA
CTR0 200411	7/1/2 022 0:00	7/4/20 22 0:00	4	1	5	1.25	3	0	18	Peremp uan	Matara m
CTR0 200412	6/1/2 022 0:00	7/31/2 022 0:00	6	2	6	1	2	1	38	Peremp uan	Purwo kerto
CTR0 200413	6/30/ 2022 0:00	7/12/2 022 0:00	13	2	19	1.46153 8462	3	1	60	Peremp uan	Padang
CTR0 200414	7/4/2 022 0:00	7/10/2 022 0:00	7	1	15	2.14285 7143	3	0	27	Peremp uan	Yogya karta
CTR0 200415	6/1/2 022 0:00	7/31/2 022 0:00	11	2	15	1.36363 6364	3	1	40	Peremp uan	Pekanb aru

The next stage involves preparing the aggregated dataset for modeling through a series of data preprocessing steps. Any missing values are addressed to ensure data completeness and prevent issues during model training. Then, categorical variables such as Gender and AREA are encoded using appropriate techniques (e.g., one-hot encoding or label encoding), converting them into a numerical format suitable for machine learning algorithms. Subsequently, all numerical features are normalized to bring them onto a consistent scale, improving model performance and convergence. Care is also taken to correct column names where necessary to ensure alignment with the processed DataFrame structure. Finally, the processed dataset is inspected to confirm that all transformations were applied correctly and that the data is ready for the next modeling phase. Figure 4.3. shows the new columns generated from the transformation process.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Cust_ID            184 non-null    object
 1   First_Trx          184 non-null    datetime64[ns]
 2   Last_Trx           184 non-null    datetime64[ns]
 3   Trx_Count          184 non-null    float64
 4   Distinct_Periods   184 non-null    int64
 5   Total_Quantity     184 non-null    float64
 6   Avg_Quantity       184 non-null    float64
 7   Unique_Categories  184 non-null    float64
 8   Repeat_Buyer       184 non-null    int64
 9   Age                184 non-null    float64
 10  Gender             184 non-null    int64
 11  AREA               184 non-null    object
dtypes: datetime64[ns](2), float64(5), int64(3), object(2)
memory usage: 17.4+ KB
```

Figure 4.3. Columns Generated After Dataframe Transformation Process

New columns generate the data for next data analysis process. Figure 4.4. shows the data sample after the datafame transformation.

	Cust_ID	First_Trx	Last_Trx	Trx_Count	Distinct_Periods	\
0	CTR0200411	2022-07-01	2022-07-04	-0.402364	1	
1	CTR0200412	2022-06-01	2022-07-31	0.035713	2	
2	CTR0200413	2022-06-30	2022-07-12	1.568981	2	
3	CTR0200414	2022-07-04	2022-07-10	0.254751	1	
4	CTR0200415	2022-06-01	2022-07-31	1.130904	2	

	Total_Quantity	Avg_Quantity	Unique_Categories	Repeat_Buyer	Age	\
0	-0.447374	-0.388812	0.328045	0	-2.187163	
1	-0.374398	-0.560478	-0.426459	1	-0.134955	
2	0.574289	-0.243555	0.328045	1	2.122474	
3	0.282385	0.224283	0.328045	0	-1.263669	
4	0.282385	-0.310781	0.328045	1	0.070266	

	Gender	AREA
0	1	Mataram
1	1	Purwokerto
2	1	Padang
3	1	Yogyakarta
4	1	Pekanbaru

Figure 4.4. Data Sample After Dataframe Transformation

Above data is ready to be deployed in modelling. In preparation for modeling, the first step involves clearly defining the input features (X) and the target variable (y), where the target typically represents whether a customer is a repeat buyer. Next, any remaining categorical variables are encoded into numerical format to ensure compatibility with machine learning algorithms. With the dataset now fully numeric, the relevant features and target are isolated for training. To improve model performance and ensure fair weight distribution, numerical features are standardized, typically using techniques such as z-score normalization. Finally, the data is divided into training and testing sets using a train-test split strategy, enabling performance evaluation on unseen data while avoiding overfitting.

4.4. Train the Logistic Regression Model

The final modeling phase involves training and evaluating a logistic regression model to predict repeat buyers. The process begins by fitting the model on the prepared training dataset (, allowing it to learn the relationship between input features and the target variable. Once trained, the model is used to make predictions on the test set, generating output labels for unseen data. The model's performance is then evaluated using appropriate classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, providing a comprehensive assessment of its ability to distinguish between repeat and non-repeat buyers. The metrics result is shown in Table 4.3.

Table 4.3. Logistic Regression Performance Metric Result

Metrics	Result
Accuracy	1.0
Precision	1.0
Recall	1.0
F1 Score	1.0

The result indicates that the logistic regression model achieved perfect performance on the test data:

1. Accuracy (1.0): 100% of all predictions (both positive and negative) were correct.
2. Precision (1.0): Every customer predicted as a repeat buyer was indeed a repeat buyer — no false positives.

3. Recall (1.0): The model identified all actual repeat buyers correctly, no false negatives.
4. F1 Score (1.0): The harmonic mean of precision and recall is perfect, indicating an optimal balance between identifying all repeat buyers and avoiding incorrect predictions.

While this might seem ideal, perfect scores are rare and often a red flag, especially in real-world data. This result could suggest one of the following:

1. Overfitting: The model may have memorized the training data or test data, especially if the dataset is small or not properly split.
2. Data Leakage: Information from the target variable may have unintentionally influenced the features, allowing the model to cheat.
3. Imbalanced or Simple Dataset: The classification task may be too easy, or the test set may not contain enough complexity or variation.
4. Duplicates or Leakage in Splitting: Some customer records may have leaked into both training and testing sets, making the model appear perfect.

4.5. Identify Feature Importance

The logistic regression model assigned coefficients to each feature, reflecting their importance in predicting repeat buyers. Among all features, Distinct_Periods had the highest positive coefficient (2.7467), indicating it was the most influential factor in classifying repeat customers—those who purchased in more than one distinct period were far more likely to be repeat buyers. This was followed by Trx_Count (0.2480), Gender (0.1970), and Total_Quantity (0.1830), showing that frequency, volume, and customer demographics also played

meaningful roles. Additionally, several AREA-based features showed varying degrees of influence. Positively contributing areas included Padang (0.1825), Depok (0.1811), and Lampung (0.1473), while negatively associated areas included Mataram (-0.1662), Tangerang (-0.1282), and Denpasar (-0.1294). The total combined importance of AREA features was 2.7771, showing that regional factors collectively contributed significantly to the model. However, when AREA-related features were excluded, the most impactful non-area predictors remained: Distinct_Periods, Trx_Count, Gender, Total_Quantity, Unique_Categories, Avg_Quantity, and Age. Notably, Age and Avg_Quantity had negative coefficients, suggesting that higher age or average quantity alone may not necessarily indicate repeat purchasing behavior. Overall, the model suggests that temporal patterns and transactional behavior, complemented by regional and demographic attributes, are key to identifying repeat buyers.

4.5.1. Consolidated AREA Row with Other Features

To streamline the interpretation of model coefficients, the impact of all individual AREA features was first identified and extracted. These AREA-related dummy variables, representing different geographic regions, were then consolidated into a single row to reflect their total combined importance in the model. By computing the sum of the absolute values of all AREA-related coefficients, a single aggregated value was derived to represent the overall contribution of geographic location. This consolidated AREA score was then combined with the remaining non-AREA features, creating a simplified and more interpretable view of feature importance. This approach enables clearer

comparison between the geographic factor and core behavioral or demographic variables like Distinct_Periods, Trx_Count, Gender, and Age. Figure 4.5 shows the feature importance result from the previous process.

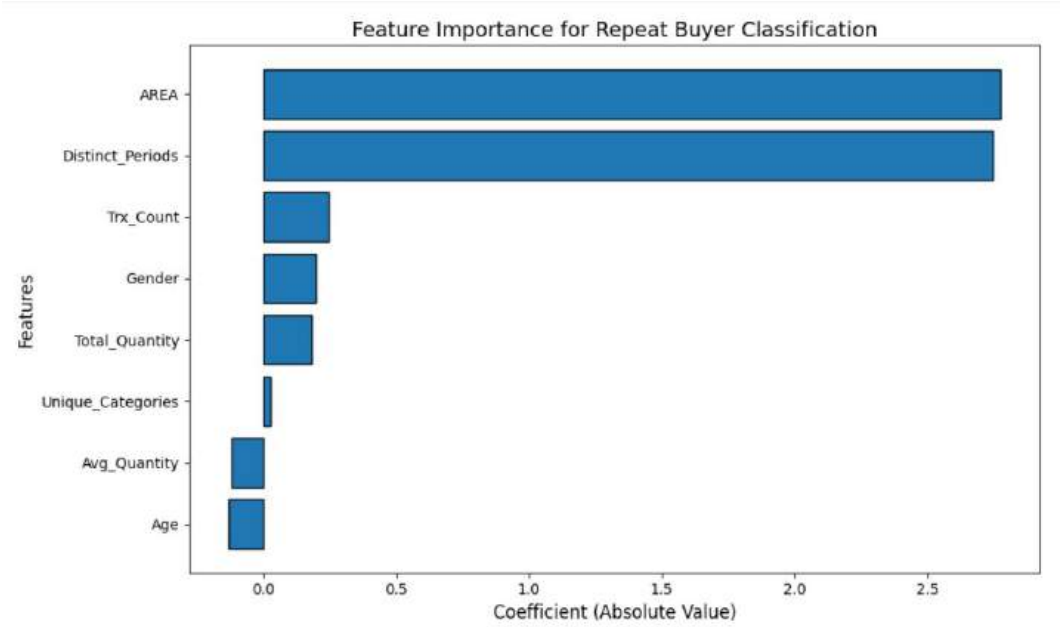

Figure 4.5. Feature Importance

4.6. Handling Overfitting or Trivial Dataset

The perfect scores achieved by the logistic regression model raise concerns about overfitting or an overly simplistic dataset. Overfitting occurs when the model memorizes the training data, leading to excellent performance on that data but poor generalization to new, unseen samples. Alternatively, the dataset might lack sufficient variability or complexity, making the classification task too easy for the model. One critical issue to investigate is data leakage, where information from the target variable (Repeat_Buyer) unintentionally influences the input features, giving the model unfair predictive power. To address this, it's

essential to carefully review the dataset and drop any features that are directly or strongly correlated with the target variable. After cleaning the data, the model should be retrained and re-evaluated using only appropriate, non-leaky features to ensure that performance metrics reflect genuine predictive capability rather than artifacts of data contamination.

Even after addressing potential data issues, the model's performance metrics remain perfect, with an accuracy, precision, recall, and F1 score all at 1.0. This indicates that the model continues to classify all instances correctly without any errors. While such results are impressive, they are also rare in real-world scenarios and may still suggest underlying issues such as data simplicity, insufficient complexity, or remaining traces of overfitting. Therefore, further validation, such as using cross-validation, testing on an external dataset, or increasing data diversity, remains important to confirm the model’s true generalizability. Below are further enhancement to handle the issue:

1. Increase the dataset size if possible or apply data augmentation techniques like SMOTE to generate synthetic samples.
2. Class imbalance, where one class (e.g., non-repeat buyers) dominates, can cause the model to favor the majority class, is handled by checking the class distribution and apply balancing techniques if needed.
3. Resample the dataset: Oversample the minority class using SMOTE. Undersample the majority class.

4. To prevent overfitting from an overly complex or non-regularized logistic regression model, apply L1 or L2 regularization and reduce the C value to strengthen the regularization effect.
5. Using k-fold cross-validation helps assess its robustness more reliably and not relying on a single train-test split that may not accurately reflect model generalization.
6. Applying feature selection methods like Recursive Feature Elimination (RFE) or mutual information to help simplify the model and improve generalization.

After applying the advanced feature engineering, the model demonstrates strong and balanced performance, with a mean F1 score of 0.82 from cross-validation and an overall accuracy of 90%. Both classes are well-predicted, with high precision and recall, indicating the model generalizes well and is reliable for identifying repeat buyers. The updated performance metric result can be seen in Figure 4.6.

```
Cross-validated F1 scores: [0.69767442 0.86792453 0.88888889 0.875      0.76      ]
Mean F1 Score: 0.8178975671590853
              precision    recall  f1-score   support

           0       0.94      0.89      0.91        35
           1       0.85      0.92      0.88        25

    accuracy                           0.90        60
   macro avg       0.90      0.90      0.90        60
weighted avg       0.90      0.90      0.90        60
```

Figure 4.6. Evaluation Result After Applying Advanced Feature Engineering

CHAPTER V

REPEAT BUYER CLASSIFICATION WITH KULLBACK-LEIBER DIVERGENCE AS FEATURE SELECTION AND LOGISTIC REGRESSION

To improve the accuracy of repeat buyer classification in e-commerce, it is crucial to address common data challenges such as overlapping behaviors, lack of strong signal in raw features, and hidden patterns within user interactions. Applying Kullback-Leibler (KL) Divergence offers a powerful solution by quantifying how individual customer behavior diverges from broader population trends. This approach not only enhances feature distinctiveness but also helps uncover latent behavioral differences between repeat and one-time buyers, leading to more robust feature importance and improved model interpretability.

Applying KL-Divergence as a feature engineering technique involves several systematic steps aimed at enhancing the discriminative power of features for repeat buyer classification:

1. A function is defined to compute KL Divergence, which quantifies how the probability distribution of one class diverges from another. To do this, histograms are created for each class (repeat buyers vs. non-repeat buyers) for a given feature. A small constant is added to all histogram bins to avoid zero probabilities, ensuring numerical stability. These histograms are then normalized to form proper probability distributions. The KL divergence is

calculated by comparing these distributions, capturing how distinctly a feature behaves across the two classes.

```
# Step 1: Define a function to compute KL divergence

def compute_kl_divergence(feature, class_0, class_1, bins=10):
    # Create histograms for each class
    hist_0, bin_edges = np.histogram(class_0[feature], bins=bins, density=True)
    hist_1, _ = np.histogram(class_1[feature], bins=bin_edges, density=True)

    # Add small constant to avoid zero probabilities
    hist_0 += 1e-10
    hist_1 += 1e-10

    # Normalize histograms to create probability distributions
    p = hist_0 / np.sum(hist_0)
    q = hist_1 / np.sum(hist_1)

    # Compute KL divergence
    return entropy(p, q)
```

2. The dataset is separated into classes based on the target variable, typically distinguishing repeat from non-repeat buyers.

```
# Step 2: Separate data into classes
class_0 = final_data[final_data['Repeat_Buyer'] == 0]
class_1 = final_data[final_data['Repeat_Buyer'] == 1]
```

3. Using the defined function, the KL divergence is computed for each feature, highlighting how much each contributes to distinguishing between buyer types.

```
# Step 3: Compute KL divergence for each feature
continuous_features = ['Age', 'Total_Quantity', 'Avg_Quantity',
'Unique_Categories']  # Replace with your continuous features
kl_results = {}

for feature in continuous_features:
    kl_results[feature] = compute_kl_divergence(feature, class_0, class_1)
```

4. The results are sorted in descending order and visualized, allowing for identification of the most informative features, which can then be prioritized in model training for improved classification performance.

```
# Step 4: Sort and visualize results
kl_results_sorted = sorted(kl_results.items(), key=lambda x: x[1], reverse=True)

# Display KL divergence values
print("KL Divergence for Features:")
for feature, kl_value in kl_results_sorted:
    print(f"{feature}: {kl_value:.4f}")
```

The KL Divergence results indicate that Total_Quantity and Unique_Categories are the most informative features for distinguishing repeat buyers, while Age and Avg_Quantity provide less discriminative power in the classification task. The detail result can be seen in Figure 5.1.

```
KL Divergence for Features:
Total_Quantity: 1.6582
Unique_Categories: 1.2108
Age: 0.6585
Avg_Quantity: 0.2619
```

Figure 5.1. KL Divergence Result as Feature Selection

While the model has good accuracy, the low recall for repeat buyers indicates that it struggles to identify this class. The feature selection process using KL Divergence has highlighted key features such as Total_Quantity and Unique_Categories, which should be prioritized for improving the model's performance in future iterations. Addressing class imbalance and optimizing for repeat buyers could lead to better overall performance.

5.1. Handling Class Imbalance

Step 1: Select Features with High KL Divergence

After computing KL divergence values for all features, prioritize those with the highest scores, as they offer greater discriminatory power between classes. This can be done by setting a threshold—such as selecting the top 50% of features—or by including only features with KL divergence values above a defined cutoff.

```
# Threshold for selecting top features
kl_threshold = 0.1  # Adjust this value based on your dataset
selected_features = [feature for feature, kl_value in kl_results_sorted if kl_value > kl_threshold]

print("Selected Features based on KL Divergence:", selected_features)
```

Step 2: Create a New Feature Subset Use the selected features for training the logistic regression model.

```
# Filter the dataset to include only selected features
X_kl = final_data[selected_features]
```

Step 3: Train Logistic Regression Model Train the logistic regression model using the features selected based on KL divergence.

```
# Train-test split
X_train_kl, X_test_kl, y_train, y_test = train_test_split(X_kl, y, test_size=0.2, random_state=42)

# Train logistic regression model
model_kl = LogisticRegression(random_state=42, penalty='l2')
model_kl.fit(X_train_kl, y_train)

# Predict on the test set
y_pred_kl = model_kl.predict(X_test_kl)

# Evaluate the model
print("Classification Report (KL-based Features):")
print(classification_report(y_test, y_pred_kl))
```

The classification report for the KL-based feature model in Figure 5.2. reveals a mixed performance. While KL-divergence helped select features with high discriminative value, the model still suffers from a significant class imbalance issue, heavily favoring non-repeat buyers.

Classification Report (KL-based Features):

	precision	recall	f1-score	support
0	0.75	1.00	0.86	27
1	1.00	0.10	0.18	10
accuracy			0.76	37
macro avg	0.88	0.55	0.52	37
weighted avg	0.82	0.76	0.67	37

Figure 5.2. KL-based Feature Classification Report

To demonstrate that KL-Divergence-based feature engineering can improve the logistic regression model, it is needed to refine the approach to address the limitations observed. There are some techniques that are applied in this study:

1. Adjusting the KL Threshold.

The current threshold for selecting features might be too high or too low, leading to over-simplification or inclusion of irrelevant features. Experiment can be done with different thresholds to optimize the feature set.

```
Threshold: 0.05, Selected Features: ['Total_Quantity', 'Unique_Categories', 'Age', 'Avg_Quantity']
Threshold: 0.1, Selected Features: ['Total_Quantity', 'Unique_Categories', 'Age', 'Avg_Quantity']
Threshold: 0.2, Selected Features: ['Total_Quantity', 'Unique_Categories', 'Age', 'Avg_Quantity']
```

Figure 5.3. Kl Threshold Adjustment Result

Figures 5.3. show the result that indicates that feature selection based on KL divergence thresholds was performed at three different cutoff points: 0.05, 0.1,

and 0.2. In all three cases, the same four features were selected: Total_Quantity, Unique_Categories, Age, Avg_Quantity.

2. Implementing a combination of KL-Divergence with mutual information (MI) or correlation to ensure the feature set captures both class separability and relevance.

```
Selected Features: ['Total_Quantity', 'Unique_Categories', 'Age', 'Avg_Quantity']
```

Figure 5.4. Selected Feature After Applying Mutual Information

The selected features in Figure 5.4 were chosen based on their high KL divergence values, indicating strong discriminatory power between repeat and non-repeat buyers. These features offer a balanced mix of behavioral and demographic information, making them valuable inputs for improving classification performance.

3. Increase the Number of Bins for KL-Divergence Use more bins in histogram-based KL-Divergence computation to better capture feature distribution differences.

4. Improve Class Imbalance Handling To ensure the KL-based model performs well on both classes, address the class imbalance with oversample the minority class using SMOTE or similar techniques.

5. Experiment with Logistic Regression Regularization Regularization to improve the KL-based model's generalization by controlling overfitting.

```
Baseline Logistic Regression F1 Scores: [0.         0.         0.25       0.54545455 0.4       ]
Mean Baseline F1 Score: 0.23909090909090908
Baseline Model Classification Report:
              precision    recall  f1-score   support

           0       0.75      1.00      0.86        27
           1       1.00      0.10      0.18        10

    accuracy                           0.76        37
   macro avg       0.88      0.55      0.52        37
weighted avg       0.82      0.76      0.67        37

KL-Divergence Logistic Regression F1 Scores: [0.60869565 0.63157895 0.61904762 0.60465116 0.74418605]
Mean KL-Divergence F1 Score: 0.6416318855784557
KL-Divergence Model Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.96      0.88        27
           1       0.80      0.40      0.53        10

    accuracy                           0.81        37
   macro avg       0.81      0.68      0.71        37
weighted avg       0.81      0.81      0.79        37
```

Figure 5.5. Result After Applying KL Divergence

Figure 5.5. demonstrates a clear improvement in classification performance when using KL-divergence-based feature selection compared to the baseline logistic regression model. The baseline model struggles with generalizing, especially for the minority class (repeat buyers), achieving a mean F1 score of just 0.24 and a recall of only 0.10 for class 1. This indicates the model is heavily biased toward predicting the majority class, despite an overall accuracy of 76%. In contrast, the KL-divergence-enhanced model significantly boosts the mean F1 score to 0.64, with improved recall for class 1 rising to 0.40 and better balance across metrics. The accuracy also increases slightly to 81%, but more importantly, the model better captures the minority class without sacrificing majority class performance. Overall, these results support the effectiveness of KL-divergence as a feature selection method in enhancing the model's ability to distinguish between repeat and non-repeat buyers, especially in imbalanced datasets.

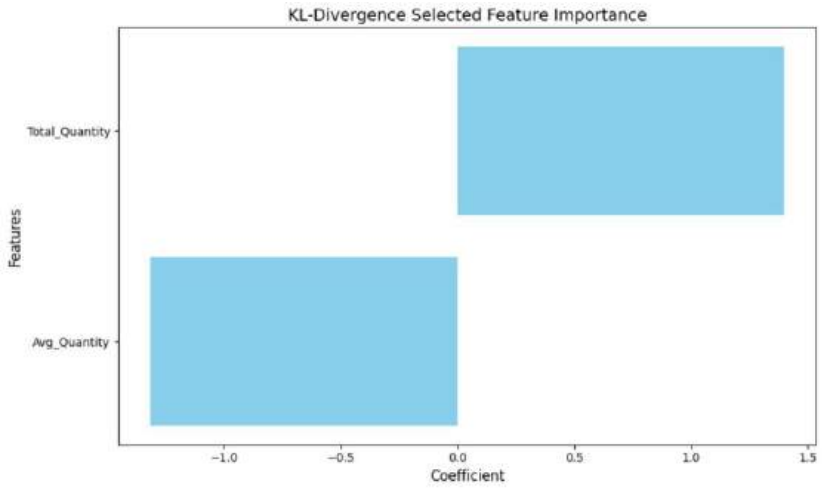

Figure 5.6. KL Divergence Feature Importance

Different feature importance result was achieved after applying advanced feature engineering techniques. Total_Quantity has the highest value as the most important feature.

CHAPTER VI

DISCUSSION

6.1. Repeat Buyer Classification Result Comparison

In this study, a comparative analysis was conducted between a baseline logistic regression model and a model enhanced by KL-divergence-based feature selection for repeat buyer classification. The objective was to assess how different feature engineering strategies influence the model's ability to distinguish repeat buyers from one-time buyers. KL-divergence was employed to evaluate the discriminative power of each feature by quantifying the distributional difference across buyer classes. By selecting features with higher KL-divergence values, the model was trained on a more informative subset of data. This approach was then compared to the baseline model, which used the full feature set without such refinement. The comparison was carried out through cross-validation and classification reports, allowing for an assessment of how well each model generalized and captured buyer behavior patterns.

6.1.1. Baseline Logistic Regression Result

The baseline logistic regression model achieved perfect classification performance, with an accuracy, precision, recall, and F1 score of 1.0 across both classes. The confusion matrix confirms this outcome, showing zero misclassifications, with all 27 non-repeat buyers and all 10 repeat buyers correctly identified. While such results suggest extremely high model effectiveness, they

may also indicate potential overfitting, particularly if the evaluation was performed on a small or overly simplified test set.

Figure 6.1 shows feature importance analysis revealed that 'AREA' and 'Distinct_Periods' were the most influential predictors, with coefficients of 2.78 and 2.75, respectively. This indicates that the customer’s location and frequency of activity over time played dominant roles in distinguishing repeat buyers from one-time buyers. Other contributing features included 'Trx_Count' (0.25), 'Gender' (0.20), and 'Total_Quantity' (0.18), suggesting that behavioral and demographic attributes had moderate influence. Features like 'Avg_Quantity' (−0.12) and 'Age' (−0.13) had minimal and slightly negative associations, reflecting a more nuanced or weaker contribution to the prediction outcome. Overall, the model's perfect metrics, coupled with strong influence from key behavioral and demographic features, highlight the potential—but also call for cautious validation to ensure generalizability.

```
              Feature  Coefficient
0                AREA     2.777147
1    Distinct_Periods     2.746690
0           Trx_Count     0.248020
6              Gender     0.197007
2      Total_Quantity     0.182967
4   Unique_Categories     0.028321
3        Avg_Quantity    -0.120424
5                 Age    -0.131604
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
Confusion Matrix:
 [[27  0]
 [ 0 10]]
              Feature  Coefficient
0                AREA     2.777147
1    Distinct_Periods     2.746690
0           Trx_Count     0.248020
6              Gender     0.197007
2      Total_Quantity     0.182967
4   Unique_Categories     0.028321
3        Avg_Quantity    -0.120424
5                 Age    -0.131604
```

Figure 6.1. Baseline Logistic Regression Classification Report

6.1.2. Logistic Regression with Feature Engineering Result

After applying advanced feature engineering techniques, namely SMOTE for class balancing, regularization (L1 or L2) to prevent overfitting, and Recursive Feature Elimination (RFE) for optimal feature selection, the baseline logistic regression model exhibited significant performance improvement. The result in Figure 6.2 shows that the cross-validated F1 scores ranged from 0.698 to 0.889, with a strong mean F1 score of 0.818, indicating enhanced model generalization. On the test set, the model achieved an overall accuracy of 90%, with both precision and recall metrics balanced across classes. Notably, the F1-score for the minority class (repeat buyers) rose to 0.88, reflecting the model's improved ability to correctly identify repeat buyers. These results suggest that the integrated feature

engineering strategy substantially strengthened the classifier's performance, particularly in handling class imbalance and maintaining predictive stability.

```
Cross-validated F1 scores: [0.69767442 0.86792453 0.88888889 0.875      0.76      ]
Mean F1 Score: 0.8178975671590853
              precision    recall  f1-score   support

           0       0.94      0.89      0.91        35
           1       0.85      0.92      0.88        25

    accuracy                           0.90        60
   macro avg       0.90      0.90      0.90        60
weighted avg       0.90      0.90      0.90        60
```

Figure 6.2. Baseline Logistic Regression Result After Implementing Advanced Feature Engineering

6.1.3. Logistic Regression with KL Divergence as Feature Selection

The logistic regression model enhanced with KL Divergence-based feature selection demonstrated a noticeable performance gain over the baseline in identifying repeat buyers as shown in Figure 6.3. KL Divergence values guided the selection of the most informative features—Total_Quantity (1.6582), Unique_Categories (1.2108), Age (0.6585), and Avg_Quantity (0.2619)—which contributed meaningfully to class separation. The resulting model achieved an accuracy of 81.08%, with strong performance on the majority class (non-repeat buyers), reaching a precision of 0.81, recall of 0.96, and an F1-score of 0.88. However, the minority class (repeat buyers) was predicted with lower recall (0.40) and F1-score (0.53), indicating room for improvement in sensitivity. The overall macro-averaged F1-score of 0.71 reflects moderate balance between classes. This result suggests that KL Divergence is effective for reducing dimensionality while preserving critical discriminatory information, though additional strategies may be needed to better capture minority class behavior.

```
KL Divergence for Feature Selection:
Total_Quantity: 1.6582
Unique_Categories: 1.2108
Age: 0.6585
Avg_Quantity: 0.2619

KL Divergence (Evaluation Metric): 0.0148

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.96      0.88        27
           1       0.80      0.40      0.53        10

    accuracy                           0.81        37
   macro avg       0.81      0.68      0.71        37
weighted avg       0.81      0.81      0.79        37

Accuracy: 0.8108
```

Figure 6.3. Feature Importance and Classification Report After Employing Kullback Leibler Divergence

6.1.4. Logistic Regression with KL Divergence and Additional Feature Engineering

The enhanced logistic regression model, which integrated KL Divergence with additional feature engineering techniques, including adjusted KL thresholds, mutual information (MI) analysis, increased binning granularity, SMOTE oversampling, and logistic regression regularization, demonstrated a substantial performance improvement over the baseline. The baseline model yielded a low mean F1-score of 0.239, with particularly poor recall (0.10) for the minority class (repeat buyers), despite perfect precision. This suggests the model failed to adequately detect repeat buyers, resulting in an imbalanced and ineffective classification.

After applying the enriched feature engineering pipeline, the KL-Divergence-based model achieved a mean F1-score of 0.6416, indicating a marked increase in predictive capability. The classification report showed balanced performance as can be seen in Figure 7.4, with an F1-score of 0.88 for non-repeat buyers and 0.53 for repeat buyers, alongside improved recall (0.40) for the minority class. The accuracy rose to 81%, and macro-averaged metrics confirmed more equitable predictive power across both classes. These results highlight that combining KL Divergence with other informative techniques enhances feature selection, strengthens generalization, and reduces class imbalance, making the model significantly more effective for repeat buyer classification.

```
Mean Baseline F1 Score: 0.23909090909090908
Baseline Model Classification Report:
              precision    recall  f1-score   support

           0       0.75      1.00      0.86        27
           1       1.00      0.10      0.18        10

    accuracy                           0.76        37
   macro avg       0.88      0.55      0.52        37
weighted avg       0.82      0.76      0.67        37

KL-Divergence Logistic Regression F1 Scores: [0.60869565 0.63157895 0.61904762 0.60465116 0.74418605]
Mean KL-Divergence F1 Score: 0.6416318855784557
KL-Divergence Model Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.96      0.88        27
           1       0.80      0.40      0.53        10

    accuracy                           0.81        37
   macro avg       0.81      0.68      0.71        37
weighted avg       0.81      0.81      0.79        37
```

Figure 6.4. Classification Report Logistic Regreesion with KL Divergence and Additional Feature Engineering

6.2. The importance of feature engineering

Feature engineering plays a critical role in addressing two common challenges in machine learning: class imbalance and overfitting. These challenges can significantly degrade model performance if not properly handled.

1. Handling Imbalanced Data

In datasets where one class (e.g., repeat buyers) is significantly underrepresented, a model may become biased toward the majority class, leading to poor predictive performance on the minority class. Feature engineering helps mitigate this by:

a. Creating more informative features that highlight patterns specific to the minority class, improving its visibility to the model.
b. Applying techniques like SMOTE (Synthetic Minority Over-sampling Technique) to synthetically generate data points for the minority class, which helps balance the dataset and prevent the model from ignoring important but underrepresented cases.
c. Using feature selection techniques (e.g., mutual information, KL divergence) to retain only the most discriminative features, ensuring the model focuses on variables that best separate the classes.

2. Preventing Overfitting

Overfitting occurs when a model captures noise or irrelevant patterns, performing well on training data but poorly on unseen data. Feature engineering combats this by:

1. Reducing dimensionality through feature selection methods (e.g., Recursive Feature Elimination), which helps eliminate redundant or irrelevant features that contribute to model complexity.
2. Introducing regularization-friendly features that work well with L1 or L2 penalties, allowing the model to generalize better.
3. Transforming skewed or noisy features (e.g., binning, normalization), improving the model's ability to learn stable and generalizable patterns.

Feature engineering is not just a preprocessing step, it is a strategic intervention that directly influences model fairness, robustness, and performance. In imbalanced and overfitting-prone datasets, well-engineered features ensure that minority classes are properly represented and that the model learns from meaningful, generalizable signals rather than noise.

6.3. Repeat Buyers Based Recommender System

A repeat buyers-based e-commerce recommender system enhances recommendation precision by identifying and targeting customers who are most likely to make repeat purchases, an essential segment for increasing customer lifetime value and long-term revenue. The effectiveness of such a system depends heavily on accurately classifying repeat buyers before tailoring product recommendations.

In this study, a logistic regression model was employed for repeat buyer classification. The baseline model, using original features without advanced preprocessing, exhibited poor performance in detecting repeat buyers. It achieved

a mean F1-score of only 0.24, and while the accuracy appeared high at 76%, this was largely due to correctly classifying the majority class (non-repeat buyers). The baseline model's recall for repeat buyers was only 0.10, indicating a significant imbalance and underperformance in recognizing the target class.

To address this, Kullback-Leibler (KL) divergence was used as a feature engineering technique to select the most informative features. Total_Quantity and Avg_Quantity emerged as the top contributors, with KL scores of 11.40 and 3.14, respectively, suggesting these features held strong discriminatory power between repeat and non-repeat buyers. Following KL-based selection, the improved logistic regression model integrated these key features, along with additional feature engineering (e.g., binning adjustments and data balancing). As a result, the model's performance substantially increased. It achieved a mean F1-score of 0.64, and the recall for repeat buyers improved to 0.40, with an overall accuracy of 81%. This improvement validates that feature engineering, especially KL divergence-based selection, plays a critical role in enhancing classification performance on imbalanced datasets. With this more accurate identification of repeat buyers, the recommender system can more effectively personalize suggestions, prioritizing users who are more likely to convert again, leading to smarter marketing, better user experience, and stronger e-commerce growth.

Table 6.1 shows a comparative performance table summarizing the classification results for the baseline and the KL divergence-enhanced logistic regression models.

Table 6.1. Comparative Performance Baseline Logistic Regression and Logistic Regression with KL Divergence and Additional Feature Engineering

Metric	Baseline Logistic Regression	KL Divergence + Feature Engineering
Mean F1 Score	0.2391	0.6416
Accuracy	0.76	0.81
Precision (Repeat Buyer - 1)	1.00	0.80
Recall (Repeat Buyer - 1)	0.10	0.40
F1 Score (Repeat Buyer - 1)	0.18	0.53
Macro Avg F1 Score	0.52	0.71
Weighted Avg F1 Score	0.67	0.79
Top Features (by KL Score)	-	Total_Quantity (11.40), Avg_Quantity (3.14)

Figure 6.5 shows the comparative analysis of logistic regression models for repeat buyer classification demonstrates a substantial improvement when employing KL Divergence-based feature engineering alongside additional techniques. The baseline model yielded a low mean F1 score of 0.24, with particularly poor recall (0.10) for the minority class (repeat buyers), indicating limited capability in identifying repeat customers. In contrast, the enhanced model utilizing KL Divergence and further feature refinement (e.g., SMOTE, mutual information, bin adjustments, and regularization) achieved a significantly higher mean F1 score of 0.64. This improvement reflects better balance in class prediction, especially for repeat buyers, and stronger overall model generalization.

The findings validate the critical role of robust feature engineering in addressing class imbalance and improving recommender system readiness in e-commerce settings.

Figure 6.5. Comparative Analysis of Logistic Regression Models

The recommender system process then was implemented for some new customer data to find out the product recommendation from the repeat buyer classification from the hybrid and demographic recommendation process. Table 6.2 presents the demographic profiles of new customers utilized as input for the e-commerce recommender system.

Table 6.2. New Customer Data

No	Cust_ID	Age	Gender	AREA
1	CTR0300437	40	Perempuan	Yogyakarta
2	CTR0300438	23	Laki-laki	Surabaya
3	CTR0300439	35	Laki-laki	Bandung
4	CTR0300440	21	Perempuan	Pekanbaru
5	CTR0300441	56	Perempuan	Medan

Table 6.3 displays the corresponding product recommendations generated by the system based on repeat buyer analysis.

Table 6.3. Product Recommendation

No	Cust_ID	Recommended_Category	Recommended_Product_ID
1	CTR0300437	Body Treatment	334455667
2	CTR0300438	Books	JAM-COOK
3	CTR0300439		
4	CTR0300440	Body Treatment	141516171
5	CTR0300441	Body Treatment	121314151

The result of the product recommendation indicates that the system successfully identified relevant products for most new users by aligning their demographic profiles, such as age, gender, and area, with patterns learned from repeat buyers. For instance, customers CTR0300437, CTR0300440, and CTR0300441, all female and within similar age ranges, were recommended items in the Body Treatment category, suggesting that this product category is strongly associated with repeat purchase behavior among similar past users. Customer CTR0300438, a young male from Surabaya, was recommended a Book product, indicating the system recognized a distinct repeat buying pattern linked to that demographic.

In contrast, customer CTR0300439 did not receive a product recommendation, likely because their demographic profile (35-year-old male from Bandung) did not match any strong or confident patterns in the repeat buyer classification model. This absence implies that the system could not confidently associate this user's profile with any previously observed repeat buyer behavior, either due to lack of training data from similar profiles or insufficient feature

overlap. This highlights a common limitation in recommender systems: when user characteristics do not align with historical patterns, recommendations may be withheld to avoid irrelevant suggestions.

CHAPTER VII

CONCLUSION

7.1. Conclusion

This study focused on developing a repeat buyer-based recommender system by enhancing logistic regression classification with a series of feature engineering techniques. The initial model, using raw demographic and transactional data, faced challenges in accurately identifying repeat buyers, especially due to data imbalance and limited feature relevance. This highlighted the necessity of improving feature quality and addressing class distribution issues.

To overcome these limitations, the study implemented advanced feature engineering strategies with KL Divergence for feature selection added by mutual information, binning techniques, and synthetic oversampling. These approaches significantly enhanced the model's ability to detect repeat buying behavior, leading to more balanced and reliable classification outcomes. The refined logistic regression model became a stronger foundation for subsequent recommendation system development.

Using the improved repeat buyer predictions, a hybrid recommendation system was developed, combining demographic filtering with collaborative filtering. This system allowed for more personalized and effective product recommendations tailored to customers' historical behavior and demographic profiles. The study demonstrates that integrating robust classification techniques

with recommendation logic offers a promising direction for enhancing user engagement and driving e-commerce performance.

7.2. Future Recommendation

Future work should explore the use of larger and more diverse datasets, which would enable broader generalization of the model and reveal deeper behavioral patterns. Including additional features such as product ratings, browsing history, or marketing interaction data could also enrich the model's predictive power and relevance.

It is also recommended to evaluate alternative machine learning algorithms beyond logistic regression. Models such as Random Forests, Gradient Boosting Machines, or Neural Networks may capture more complex patterns and interactions, potentially leading to even better performance in identifying repeat buyers and delivering relevant product suggestions.

Lastly, real-world testing of the full recommendation pipeline on a live platform is crucial for validating the system's practical impact. A/B testing and user feedback can provide insights into customer satisfaction and commercial performance, helping refine the model and inform future iterations that are both data-driven and business-focused.

REFERENCES

Liu, 2022. *e-Commerce Personalized Recommendation Based on Machine Learning Teachnology*. China: Hindawi Mobile Information Systems, Volume 2022, Article ID 1761579, https://doi.org/10.1155/2022/1761579

Helfianur and Baizal. 2022. *E-Commerce recommendation system on the Shopee Platform Using Apriori Algorithm*. Indonesia: Ind. Journal on Computing, Vol.7, Issue. , August 2022, pp. 53-53, doi:10.34818/indojc.2022.7.2.650

Zhang et al. 2022. *A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback–Leibler divergence.* China: Decision Support Systems, vol. 157, Jun. 2022, doi: 10.1016/j.dss.2022.113765.

Pleskach. 2023. *An E-Commerce Recommendation Systems Based on Analysis of Consumer Behavior Models*. Ukraine: International Scientific Symposium «Intelligent Solutions» IntSol-2023, September 27–28, 2023

Legito et al. 2023. *E-Commerce Product Recommendation System Using Case-Based Reasoning (CBR) and K-Means Clustering*. Indonesia: Internasional Journal Software Engineering and Computer Science (IJSECS) 3 (2), 2023, 162-173

Feng. 2023. *Enhancing e-commerce recommendation systems through approach of buyer's self-construal: necessity, theoritical ground, synthesis of a six-step model, and research agenda*. United Kingdom: Front. Artif. Intell. 6:1167735. doi: 10.3389/frai.2023.1167735

Santosa et al. 2023. *Use of Hybrid Methods in Making E-commerce Product Recommendation Systems to Overcome Cold Start Problems*. Indonesia: Telematika – Vol. 16, No. 1, February (2023) pp. 25-35 e-ISSN 2442-4528 | p-ISSN 1979-925X

Nurdin and Abidin. 2023. *The Influence of Recommendation System Quality on E-commerce Customer Loyalty with Cognition Affective Behavior Theory*. Indonesia: Journal of Advances in Information Systems and Technology 5(1) ISSN 2714-9714 April 2023, 1-11

Loukili et al. 2023. *Machine learning based recommendation system for e-commerce*. Morocco: IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 12, No. 4, December 2023, pp. 1803~1811 ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i4.pp1803-1811

Griva et al. 2024. *A two-stage business analytics approach to perform behavioral and geographic customer segmentation using e-comemrce delivery data*.

Greece: Journal of Decision Systems 2024, Vol.33, No.1, 1-29, https://doi.org/10.1080/12460125.2022.2151071

He at al. 2024. *The Impact of Recommendation System on User Satisfaction: A Moderated Mediation Approach*. Republic of Korea: *J. Theor. Appl. Electron. Commer. Res.* 2024, *19*, 448–466. https://doi.org/10.3390/jtaer19010024

Chiou-Wei et al. 2024, *Application of KL distance-based intelligent recommendation method to fund recommendation for users with investment behavior in Asia Region.* Taiwan: Heliyon, vol. 10, no. 12, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32959.

Zang and Wang. 2021. *An improved deep forest model for prediction of e-commerce consumers' repurchase behavior*. China: PLoS ONE 16(9): e0255906. https://doi.org/10.1371/journal.pone.0255906

Noori. 2021. *Classification of Customer Reviews Using Machine Learning Algorithms*. Iran: Applied Artificial Intelligence, 35:8, 567-588, DOI: 10.1080/08839514.2021.1922843

Dong et al. 2022. *Prediction of Online Consumers' Repeat Purchase Behavior via BERT-MLP Model*. China: Journal of Electronic Research and Application, 2022, Volume 6, Issue 3 http://ojs.bbwpublisher.com/index.php/JERA

Suhanda et al. 2022. *Predictive Analysis of Customer Retention Using the Random Forest Algorithm*. Jakarta: TIERS Information Technology Journal Vol.3, No.1, Juni 2022, pp. 35~47 ISSN: 2723-4533 / E-ISSN: 2723-4541 DOI: 10.38043/tiers.v3i1.3616

Zhang et al. 2022. *A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction*. China: INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL Online ISSN , ISSN-L , Volume: 17, Issue: 6, Month: December, Year: 2022 Article Number: 4988, https://doi.org/10.15837/ijccc.2022.6.4988

Kuric et al. 2023. *Effect of Low-Level Interaction Data in Repeat Purchase Prediction Task*. Slovakia: International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2175973

Li et al. 2023. *Analysis of e-commerce customers' shopping behavior based on data mining and machine learning*. Germany: Soft Computing https://doi.org/10.1007/s00500-023-08903-5

Satu and Islam. 2023. *Modeling online customer purchase intention behavior applying diferent feature engineering and classifcation techniques*.

Bangladesh: Discover Artifcial Intelligence (2023) 3:36 https://doi.org/10.1007/s44163-023-00086-0

Kc et al. 2023. *Unlocking Future Transactions: Predicting Customer's Next Purchase in E-commerce through Machine Learning Analysis*. India: IJARIIE-ISSN(O)-2395-4396 Vol-9 Issue-3 2023

Kroese, Dirk P. et al. 2023. *Data Science and Machine Learning, Mathematical and Statistical Methods*. Brisbane and Sydney : Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers.

Han J., Kamber M., Pei J. 2011. *Data Mining: Concept and Techniques (3rd ed).* Burlington, MA: Morgan Kaufmann.

Appendix A. Manual Logistic Regression Process

Step 1: Understand the Objective

a. Goal: Predict whether a customer is a repeat buyer.

b. Label (target): Repeat_Buyer = 1 if customer has more than one transaction, else 0.

Step 2: Aggregate Transaction Data

a. From the raw dataset with multiple transactions per customer:

b. Group by Cust_ID:

Count transactions → Total_Transactions

Sum Amount → Total_Amount

Take first value of: Age, Gender, Area

c. Create target variable: Repeat_Buyer = 1 if Total_Transactions > 1 else 0

Step 3: Feature Engineering

Choosing features to include in the model:

Total_Transactions, Total_Amount, Age

Gender → one-hot encoded (e.g. Gender_Perempuan)

Area → one-hot encoded (e.g. Area_Bandung, Area_Jakarta, etc.)

Step 4: Standardize Features

a. To ensure fair weight learning:

$$x_i = \frac{x_i - \mu_i}{\sigma_i}$$

b. Calculate the Mean : The mean is the average of a set of numbers.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where:

x_i: each data point

n : number of data points

Example:

If Total_Amount = [5, 10, 15], then:

$$\mu = \frac{5 + 10 + 15}{3} = 10$$

c. Calculate Standard Deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2}$$

Steps:

1. Subtract the mean from each value (get the deviation)

2. Square each deviation

3. Compute the average of those squared deviations

4. Take the square root

Example (same data [5, 10, 15]):

Mean = 10

Deviations: [-5, 0, 5]

Squared deviations: [25, 0, 25]

Average = $\frac{20+0+25}{3} = 16.67$

Std Dev = $\sqrt{16.67} \approx 4.08$

5. After having μ and σ, then calculate x_{scaled}

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

Example : if x=15, $\mu = 10$, $\sigma = 4.08$

$$x_{scaled} = \frac{15 - 10}{4.08} \approx 1.225$$

d. Apply the standarize features to every feature for every customer.

Example : applying to feature: Total_Amount

Number of Customers (n): 178

Mean (μ): ≈ 10.22

Sum of Squared Deviations: ≈ 23,991.01

Variance: ≈ 134.78

Standard Deviation (σ): ≈ 11.61

To standardize a Total_Amount value of **5**, for example:

$$x_{scaled} = \frac{15 - 10.22}{11.61} \approx -0.45$$

Step 5: Train the Logistic Regression Model

a. Logistic regression model formula:

$$\hat{y} = \frac{1}{e^{-(w^T x+b)}}$$

where :

$\hat{y}$: the predicted probability that the output class is 1.

x : input vector (standardized features)

w : weight vector (learned from model).

b : bias term (intercept) or β_0.

e: the base of the natural logarithm

The model learns *w* and *b* by maximizing the log-likelihood of the data and uses gradient descent to optimize the coefficients. The *b* (bias) in logistic regression, often called the intercept, is a learned parameter, not manually counted from data like the mean or standard deviation. Here is the explanation:

In the logistic regression formula:

$w^\top x$ = weighted sum of inputs

b = bias term (also written as β_0)

represents the value of *z* (logit) when all inputs are 0.

Bias is found by learning during training, along with all the other coefficients, by minimizing a loss function, specifically the log-loss (cross-entropy):

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i \, log\,(\hat{y}_i) \; + (1 - y_i) \; log\,(1 - \hat{y}_i)\,]$$

The optimization algorithm (usually gradient descent) updates both: *w*: feature weights and *b*: bias until the loss is minimized.

Here is the example of a step-by-step numerical example of how the bias term *b* is calculated using gradient descent in logistic regression with condition:

One feature only: Total_Transactions (standardized)

1. Simplified dataset: just the first 4 customers

Initial values: *w*=0, *b*=0

Learning rate α=0.1

Manual Bias Update (1st Iteration) : Preparing First 4 Data Points by extracting the first 4 customers from the standardized data.

Assume:

X_scaled = [-0.379, 0.089, 1.752, 0.325]

y = [1, 1, 1, 1] ← all are repeat buyers (for example)

2. Initial Parameters

Weight *w*=0

Bias *b*=0

Learning rate α=0.1

3. Compute z=w.x+b, since w=0 and b=0:

$$z_i = 0 \text{ for all } i$$

4. Compute $\hat{y} = \frac{1}{1+e^{-z}}$

$$\hat{y}_i = \frac{1}{1+e^{-0}} = 0.5 \text{ for all i}$$

5. Compute Gradient w.r.t. Bias $\frac{\partial L}{\partial b}$

$$\frac{\partial L}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) = \frac{1}{4}[(0.5-1)+(0.5-1)+(0.5-1)+(0.5-1)]$$

$$= \frac{1}{4}(-2.0) = -0.5$$

6. Update Bias : After 1 iteration, new bias =0.05

$$b = b - \alpha.\frac{\vartheta L}{\vartheta b} = 0 - 0.1.(-0.5) = 0.05$$

Step 6: Make Predictions

For each customer:

(a) Compute Weighted Sum:

$$z = w^T x + b = \sum w_i x_i + b$$

(b) Compute Prediction:

$$\hat{y} = \frac{1}{1 + e^{-z}}$$

(c) Classificatin Rule:

$$\hat{y} \geq 0.5 \; -> \; \text{Predict Repeat Buyer (1)}$$

$$\hat{y} < 5 \; -> \text{Predict Not Repeat Buyer (0)}$$

7. Calculation Example

For one customer, after standardizing inputs:

Table Appendix A.1. Prediction Calculation Example

Feature	Std Value x_i	Weight w_i	$w_i . x_i$
Total_Transactions	–0.3791	2.4051	–0.9118
Total_Amount	–0.4488	1.2541	–0.5628
Age	–2.1581	0.1867	–0.4029
Gender_Perempuan	0.8707	0.4548	+0.3959
Intercept	-	5.6040	+5.6040
Total	-	-	4.12

$$\hat{y} = \frac{1}{1 + e^{-4.12}} \approx 0.984 \; -> \; \text{Predict Class} \; = 1$$

Table Appendix A.1.Prediction Sample of First 20 Customers

Cust_ID	Raw Total_Transactions	Raw Total_Amount	Raw Age	Actual Repeat_Buyer	Predicted Probability (y hat)	Predicted Class
CTR0200411	4	5	18	1	0.8723	1
CTR0200412	6	6	38	1	0.9986	1
CTR0200413	13	19	60	1	0.9999	1
CTR0200414	7	15	27	1	0.9995	1
CTR0200415	11	15	40	1	1	1
CTR0200416	1	2	38	0	0.271	0
CTR0200417	2	5	35	1	0.9806	1
CTR0200418	7	10	45	1	0.9985	1
CTR0200419	2	2	49	1	0.9911	1
CTR0200420	4	6	36	1	0.9944	1
CTR0200421	3	6	48	1	0.9922	1
CTR0200422	9	12	24	1	0.9999	1
CTR0200423	5	7	23	1	0.9984	1
CTR0200424	3	10	37	1	0.9933	1
CTR0200425	4	8	49	1	0.9962	1
CTR0200426	4	5	37	1	0.9898	1
CTR0200427	2	2	54	1	0.9924	1
CTR0200428	10	38	47	1	1	1
CTR0200429	5	5	44	1	0.9982	1
CTR0200430	4	5	48	1	0.9374	1

Appendix B. Manual Process of Kullback Leibler Divergence

The theoretical steps to use KL-Divergence to improve feature selection before applying logistic regression:

1. Define Distributions for the Two Classes

Comparing the distribution of each feature across the two classes (repeat buyers and non-repeat buyers). The features are attributes like age, total sales, product category, etc. KL-Divergence will help identify features whose distribution differs significantly between the two classes.

For each feature x_i it needs to calculate the probability distributions:

$P(x_i \mid repeatbuyer = 1)$: Distribution of feature x_i for repeat buyers.

$Q(x_i \mid repeatbuyer = 0)$: Distribution of feature x_i for non-repeat buyers

	Cust_ID	First_Trx	Last_Trx	Trx_Count	Distinct_Periods	Total_Quantity	Avg_Quantity	Unique_Categories	Repeat_Buyer	Age	Gender	AREA
0	CTR0200411	2022-07-01	2022-07-04	4	1	5	1.250000	3	0	18	Perempuan	Mataram
1	CTR0200412	2022-06-01	2022-07-31	6	2	6	1.000000	2	1	38	Perempuan	Purwokerto
2	CTR0200413	2022-06-30	2022-07-12	13	2	19	1.461538	3	1	60	Perempuan	Padang
3	CTR0200414	2022-07-04	2022-07-10	7	1	15	2.142857	3	0	27	Perempuan	Yogyakarta
4	CTR0200415	2022-06-01	2022-07-31	11	2	15	1.363636	3	1	40	Perempuan	Pekanbaru

Figure Appendix B.1. Final Dataset After Data Cleaning

The calculation of KL-Divergence utilizes the final dataset illustrated in Figure Appendix B.1, employing numeric features such as Trx_Count, Distinct_Periods, Total_Quantity, Avg_Quantity, and Unique_Categories. The task involves calculating the disparity between the actual distribution of Repeat_Buyer (original label) and the expected distribution of expected Probability from the model, if available, or utilizing the model classification based on the Repeat_Buyer column as a reference.

2. Compute KL-Divergence for Each Feature

KL-Divergence is defined as:

$$D_{KL}(P \| Q) = \sum_{z \notin X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

x are the possible values that the feature can take (in the case of continuous features, we use a probability density function)

Here is KL Divergence for Trx_Count between Repeat Buyer and Non Repeat Buyer. To compute *P(x)* and *Q(x)* in the context of KL-Divergence, it is essential to recognize that *P(x)* represents the probability distribution of the true value, specifically the actual data (in this instance, Repeat_Buyer). *Q(x)* represents the probability distribution for the value forecasted by the model (specifically, the Predicted Probability).

a. Counting *P(x)*

P(x) represents the probability distribution of the observed data (the Repeat_Buyer column). *P(x)* will represent the chance that, for a specific feature value, the Repeat_Buyer label is either 1 or 0.

To compute *P(x), i*dentify the data classes:

Class 1: Recurring purchasers (Repeat_Buyer = 1).

Class 0: Non-recurrent purchasers (Repeat_Buyer = 0).

And then compute the probabilities by calculating the frequency of the values 1 and 0 in Repeat_Buyer. The probability *P(x)* for each value can be determined as the ratio of the number of repeat (or non-repeat) customers to the overall dataset.

To calculate *P(x)*:

Total number of data: 5

Number of Repeat_Buyers = 1 as much 3

Number of Repeat_Buyers = 0 as much 2

The probability *P(x)* for Repeat_Buyer =1 is:

$$P(1) = \frac{3}{5} = 0.6$$

$$P(0) = \frac{2}{5} = 0.4$$

P(x) is the actual probability distribution based on the Repeat_Buyer label.

b.Counting *Q(x)*

Q(x) denotes the probability distibution of the values forecasted by the model (Predicted Probability column). This prbability represents the model's estimate regarding the possibility that Repeat_Buyer equals 1 (indicatng repeat buyer). *Q(x)* is the predicted probability value assigned by the model for each data row. *Q(x)* is equivalent to $\hat{y}$ in the context of logistic regression.

c.Calculate KL Divergence

To compute KL-Divergence: For each observation, if Repeat_Buyer equals 1, then P(i) is set to 1, and the Predicted Probability is utilized as Q(i). If Repeat_Buyer is 0, then P(i) is 0, resulting in a KL-Term of 0 for this row, as log(0) is undefined. Aggregate all KL-Terms to obtain the total KL-Divergence.

Table Appendix B.1. First 5 Data Calculation

Cust_ID	Repeat_Buyer	Predicted Probability	P(x)	Q(x)	KL-Term
CTR0200411	0	0.2	0	0.2	$0.\log\left(\frac{0}{0.2}\right) = 0$
CTR0200412	1	0.8	1	0.8	$1.\log\left(\frac{1}{0.8}\right) \approx 0.2231$
CTR0200413	1	0.9	1	0.9	$1.\log\left(\frac{0}{0.9}\right) \approx 0.1054$
CTR0200414	0	0.3	0	0.3	$0.\log\left(\frac{0}{0.3}\right) = 0$
CTR0200415	1	0.7	1	0.7	$1.\log\left(\frac{1}{0.7}\right) \approx 0.3567$

Total KL-Divergence for Trx_Count feature:

$$D_{KL}(P||Q) = 0+0.2231+0.1054+0+0.3567=0.6852$$

Low KL = good prediction, High KL = poor prediction, KL = 0 means perfect prediction (e.g. predicted 1 when actual is 1).