

Thesis_report().pdf

by IIUC4 Central Library

Submission date: 10-Feb-2024 02:03PM (UTC+0530)

Submission ID: 2291142089

File name: Thesis_report.pdf (3.68M)

Word count: 17905

Character count: 111332

An E-commerce recommendation system based on LightGBM Machine Learning Algorithm

This Dissertation is Submitted in ²² Fulfillment
of the Requirements for the Degree of

Bachelor of Science (B.Sc.)

in

Computer Science and Engineering (CSE)

by

Mohammed Ashrafujjaman Hera (C191058)

Md.Injamamul Hoque Chowdhury(C191002)

Mohammad Tohidul Alam (C191100)



TO

⁸⁴
FACULTY OF SCIENCE AND ENGINEERING
INTERNATIONAL ISLAMIC UNIVERSITY CHITTAGONG

Spring 22

DECLARATION

We hereby affirm the following statements regarding our thesis:

1. The thesis has been successfully completed as part of our undergraduate degree program at International Islamic University Chittagong.
2. The thesis work does not contain any previously published or third-party content without proper citation.
3. The thesis work ⁵⁶has not been previously submitted for any other degree or diploma at any other university or institution.
4. We have appropriately acknowledged all significant sources of contribution in the thesis.

Student's Full Name and Metric ID:

Mohammed Ashrafujjaman Hera	(C191058)
Md.Injamamul Hoque Chowdhury	(C191002)
Mohammad Tohidul Alam	(C191100)

SUPERVISOR'S DECLARATION

I formally state that I have examined this thesis and claim it to be of sufficient quality and scope to be granted ²²for the undergraduate degree of Bachelor of Science in Computer Science and Engineering.

Mohammed Mahmudur Rahman

75 Associate Professor

Department of Computer Science and Engineering

International Islamic University Chittagong

DEDICATION

This dedicated thesis report is an expression of appreciation to ourselves, our supervisor, and our family. The collaborative effort within the team was commendable, with satisfactory teamwork, and the unparalleled support from our family proved to be exceptionally remarkable. Our supervisor, characterized by dedication and unwavering hard work, has been a consistent pillar of support throughout the duration of this project. Their guidance and encouragement have played a pivotal role in our academic pursuits over these months. This document also serves as a platform to formally recognize and acknowledge the contributions made by each member of the team. It stands as a testament to the synergy within the group and the invaluable backing provided by our families. Together, this collective dedication and collaboration have culminated in the successful completion of this thesis, and we extend our gratitude to everyone involved in this journey.

ACKNOWLEDGMENT

We begin by acknowledging the tremendous assistance of the Great Allah, whose guidance allowed us to successfully navigate and complete our thesis with relative ease. Our gratitude extends to Mohammed Mahmudur Rahman, our dedicated research supervisor, whose considerate leadership and continual support played a pivotal role in our academic journey. In times of need, he accommodated our requests and provided encouragement during the most demanding phases of our research. Lastly, we owe a debt of thanks to our parents for their prayers and well wishes, without which our achievements would have been unattainable. Their steadfast support has been a constant source of strength, significantly contributing to our academic successes.

ETHICAL STATEMENT

We affirm that our thesis work was conducted without resorting to any unethical practices. The data employed for research purposes is entirely original, and we meticulously verified all citations included in this document. The authors of this work willingly assume full responsibility for any potential breaches of thesis rules. Our commitment to academic integrity and ethical standards underscores the rigor and conscientiousness applied throughout the research process. By adhering to the highest ethical standards, we ensure the credibility and reliability of our thesis, thereby upholding the principles of honesty and integrity in scholarly pursuits. This statement attests to our dedication to maintaining the highest level of ethical conduct in academic research and serves as a declaration of our accountability for the integrity of our thesis work.

ABSTRACT

⁶ In the dynamic realm of e-commerce, recommendation systems play a pivotal role in shaping user experiences and fueling business growth. This study advocates for a novel approach to online shopping recommendations, leveraging the power of the LightGBM machine learning algorithm. By focusing on item-to-item recommendations, our methodology seeks to elevate user satisfaction by swiftly and precisely offering customers highly personalized product choices.

At the core of our recommendation system lies the fusion of item-to-item association analysis and user interactions, culminating in the delivery of accurate, real-time recommendations. This research contributes significantly to the e-commerce landscape by presenting a practical and scalable method that enriches customer experiences, consequently amplifying sales and fostering customer loyalty.

Through extensive testing and evaluation, our results underscore the transformative potential of the proposed item-to-item e-commerce recommendation system. This innovative system stands poised to revolutionize digital commerce by providing users with pinpoint-accurate product recommendations. The seamless integration of machine learning, coupled with a focus on item-to-item relationships, not only expedites the decision-making process for consumers but also cultivates a deeper connection between customers and the digital marketplace. In summary, our study demonstrates the capability of our approach to usher in a new era of precision and effectiveness in digital commerce, promising a paradigm shift in the way users discover and engage with products online.

Keywords: E-commerce , recommendation system, LightGBM, machine learning, item-to-item, personalization.

Table of Contents

Abstract	vii
Table of Contents	viii
List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 Introduction	1
1.2 Research background	2
1.3 Problem Statement	3
1.4 Motivation and Scope of the Research	4
1.5 Research Flow	7
1.6 Research Question	7
1.7 Objective	8
2 Literature Review	9
2.1 Introduction	9
2.2 Content Analysis Table:	10
2.3 content analysis	19
2.4 Applied-Algorithms	21
2.4.1 Lightgbm	21
2.4.2 XGBoost	23
2.4.3 Random-Forest	25
2.4.4 Support-Vector-Machine	27
2.5 Summary	28
3 Methodology	32
3.1 Introduction	32
3.2 Models	33
3.3 Description:	35
3.3.1 First Stage	35
3.3.2 Second Stage	38
3.3.3 Final Stage	41
3.4 Mathematical Expression:	46
3.4.1 Gradient-based One Side Sampling Technique for LightGBM: . . .	46
3.4.2 Exclusive Feature Bundling Technique for LightGBM	47

53		
4	Data Analysis	48
4.1	Data-Collection	48
4.2	Data-Attributes:	49
4.3	Dataset-Properties	51
4.4	Representation-of-Scanned-Raw-Dataset:	53
4.5	Data-Pre-Processing:	54
4.6	Encoding Categorical Variables:	57
4.7	Data Standardization	57
4.8	Tools	57
5	Results and Discussions	58
5.1	Dataset-1	58
5.1.1	For 100k Data :-	58
5.1.2	For 500k Data :-	58
5.1.3	For 1 million Data:-	58
5.2	Dataset-2	60
5.2.1	For 100k Data :-	60
5.2.2	For 500k Data:-	60
5.2.3	For Data Count Of 1 million :-	60
5.3	Discussion :-	62
5.4	Feature Importance:	63
5.5	Error Rate Visualiztion:	64
5.6	Recommendation-Checking:	66
5.7	Comparison with the previous work:	70
59		
6	Conclusion	72
6.1	Contribution	72
6.2	Limitation	73
6.3	Future-Work	74
	References	75

List of Figures

1.1	Failure rate of small businesses[1]	2
1.2	Impact Of Delayed accesss[2]	3
1.3	Pandemic Accelerates Shift to Online Retail[3]	4
1.4	User activity in Amazon[4]	5
2.1	Architecture of the Lightgbm classifier[23]	21
2.2	Architecture of the Xgboost classifier[24]	23
2.3	Architecture of the Random Forest classifier[25]	25
2.4	Architecture of the SVM classifier[26]	27
57		
3.1	Overall System Architecture	34
3.2	First Stage	35
3.3	Loading dataset	36
3.4	Preprocessing Data	36
3.5	Processed Dataset	37
3.6	Sample Selection	37
3.7	Second stage	38
3.8	Transformation	39
3.9	Data split	39
3.10	Applying Lightgbm	40
3.11	Prediction Model	40

3.12 Last Stage	41
3.13 Time Calculation	42
3.14 Evaluation	42
3.15 Testing	43
3.16 Result Prediction	43
3.17 Sorting predicted data	44
3.18 Complete Architecture of the working	45
3.19 Algorithm for GOSS	46
3.20 Algorithm for Exclusive Feature Bundling Technique	47
4.1 Digitized Raw Dataset_1 Samples	53
4.2 Continuation of 4.1	53
4.3 Digitized Raw Dataset_2 Samples	54
4.4 Null Data Count For Each Column	54
4.5 Null Data Count For Each Column	56
5.1 Feature Importance	63
5.2 Multi Logloss of dataset_1	64
5.3 Multi Logloss of dataset_2	65

List of Tables

4.1	Attributes of Dataset_1	49
4.2	Attributes of Dataset_2	50
4.3	Properties of Dataset	51
4.4	Properties of Dataset	52
4.5	Preprocessed Dataset_1	55
4.6	Preprocessed Dataset_2	56
5.1	Results For 100k	58
5.2	Results For 500k	58
5.3	Results For 1 million	58
5.4	Results For 100k	60
5.5	Results For 500k	60
5.6	For 1 million	60
5.7	Input	66
5.8	Recommendation For Dataset_1	66
5.9	Input	68
5.10	Recommendation For Dataset_2	68
5.11	Previous Work	70
5.12	Our Result for 100k	70

Chapter 1

Introduction

1.1 Introduction

In the ever-changing world of electronic commerce (E-commerce), recommendation systems play a critical role in enhancing user experiences and driving business expansion. Based on the powers of the LightGBM machine learning algorithm, this paper proposes a revolutionary approach for online purchase suggestions. We aim to build LightGBM and then demonstrate a comprehensive understanding of this methodology by investigating item-to-item recommendation systems.

Lightgbm has become well-known for its efficiency, speed, and accuracy. Gradient boosting is a technique designed for large volumes of data and high-dimensional feature spaces. LightGBM constructs decision trees leaf-wise in order to minimize loss at each level and prioritize the splits that yield the most information. This approach has become LightGBM the preferred choice for many machine learning applications, including recommendation systems.

Item-to-item recommendation systems are a necessary part of E-commerce proposals. Rather of placing more emphasis on user-item interactions, these systems focus on identifying product correlations. Item-based recommendations analyze previous user behavior and product interactions to give customers with appropriate options based on their requirements and preferences. Related or complimentary items are also suggested. In this work, we examine in the context of e-commerce the synergy between LightGBM's efficiency and the accuracy of item-to-item suggestions. In order to optimize user satisfaction, engagement, and profitability, we aim to present a comprehensive understanding of how LightGBM may enhance item-to-item recommendation systems.

1.2 Research background

The coupling of LightGBM, a very powerful machine learning technique, with item-to-item recommendation algorithms in e-commerce is the subject of this study. Taking use of LightGBM's effectiveness, speed, and accuracy, the study investigates how item-based suggestions could be improved by using it. In addition, a lot of small companies exhibit declination as a result of failing to meet the expectations of their customers for improved referrals.

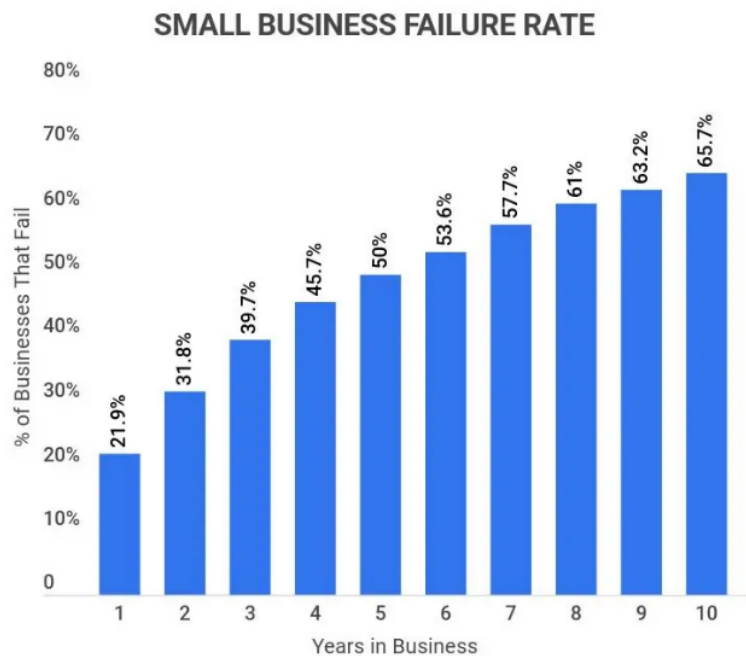


Fig. 1.1. Failure rate of small businesses[1]

Personalized suggestions derived from past user interactions are the focus of item-based systems, as opposed to user-centric approaches, which focus on product correlations. In order to maximize user engagement and financial advantage, the study aims to provide a thorough understanding of how LightGBM might enhance E-commerce recommendations. Rich experiments, approaches, and findings are presented, offering insights into a ground-breaking solution at the intersection of E-commerce recommendations and machine learning.

1.3 Problem Statement

Despite advancements in recommendation algorithms, creating personalized and entirely accurate product recommendations remains a challenging task, causing issues that have a detrimental impact on online shopping in general.



Fig. 1.2. Impact Of Delayed access^[2]

In the area of recommendation systems, the challenges of scalability, performance, time efficiency, and data sparsity are recurrent hurdles. Singular algorithmic solutions frequently grapple with the complexity of capturing the entire spectrum of user preferences. Recognizing this limitation, the exploration of novel algorithms becomes imperative for more effective recommendation systems. LightGBM, a gradient boosting framework, emerges as a beacon of promise in this landscape. Unlike its predecessors, LightGBM excels in managing high-dimensional data and deciphering intricate patterns with unparalleled efficiency. Its leaf-wise tree growth strategy and gradient-based optimization empower it to navigate the complexities of diverse datasets seamlessly. ⁸⁵ In the context of e-commerce, where the dynamic nature of user preferences demands precision and adaptability, LightGBM's performance capabilities present an intriguing opportunity. While the specific performance of LightGBM in the field of e-commerce awaits empirical validation, its intrinsic features suggest a potential paradigm shift. Its ability to handle high-dimensional data coupled with a superior pattern recognition capacity positions it as a strong contender for addressing the nuanced and evolving nature of user preferences in the e-commerce space.

1.4 Motivation and Scope of the Research

In our exploration of recommendation systems for e-commerce websites, we've delved into diverse algorithms like XGBoost, Random Forest, SVM, among others. While these algorithms exhibit competence in handling smaller datasets, a common limitation arises when scalability and real-time model building become imperative. The transition to LightGBM emerges as a strategic solution, effectively addressing these dual challenges.

As a matter of fact due to the pandemic ⁸¹ in recent years there has been an acceleration in users in online retail in other words e-commerce sites making it crucial to update the level of algorithms used in e-commerce sites all together.

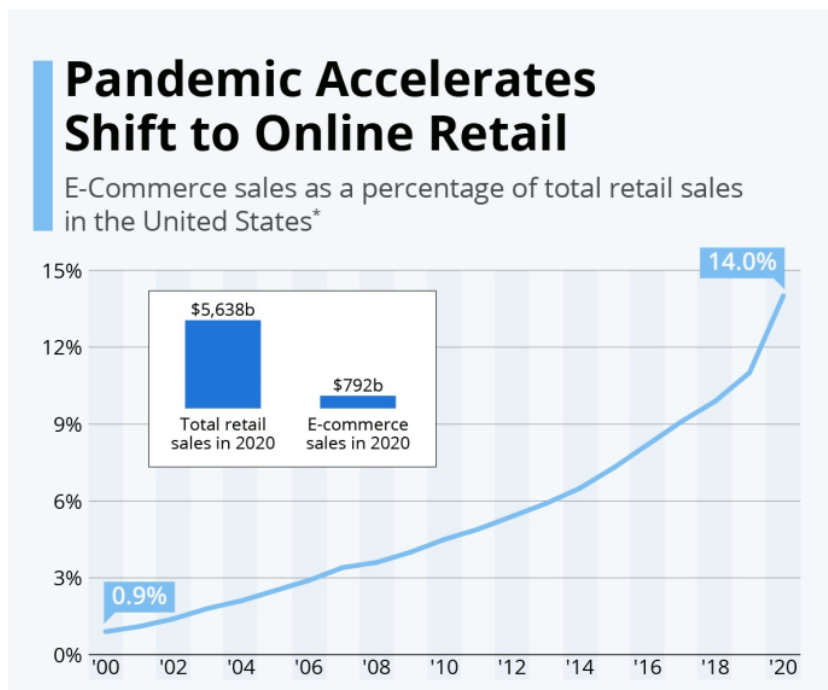


Fig. 1.3. Pandemic Accelerates Shift to Online Retail[3]

LightGBM, a standout in the realm of machine learning algorithms, particularly distinguishes itself through its exceptional performance even with modest datasets, reaching its true pinnacle when confronted with larger and more complex data in the e-commerce domain. The e-commerce sector, characterized by extensive and dynamic datasets, finds a formidable ally in LightGBM, surpassing the capabilities of traditional algorithms. As well as the fact that amazon users have seen a growth in billions due to the winter festivals during 2023-24.



Fig. 1.4. User activity in Amazon[4]

Because of its gradient-based optimization and leaf-wise tree growth technique, LightGBM is unique in that it constructs models in real time. With its ability to compute quickly without compromising accuracy, this method is the best in dynamic situations such as e-commerce. Its quick analysis of large datasets and flexibility in response to changing circumstances are essential in the quick-paced world of e-commerce, where speed is of the essence. Because of LightGBM's efficiency, organizations can quickly make data-driven decisions, enhance marketing campaigns, and tailor user experiences. LightGBM's model construction speed is a game-changer in the competitive e-commerce space, enabling businesses to remain flexible and gain a competitive advantage.

LightGBM revolutionizes recommendation systems by addressing inherent limitations and introducing unprecedented scalability and real-time responsiveness. Unlike traditional algorithms, LightGBM's implementation ensures timely and precise personalized product suggestions, critical for enhancing the user experience in e-commerce.

Its ability to handle vast and dynamic datasets is particularly advantageous in the ever-evolving e-commerce landscape. By swiftly adapting to changing trends and user behaviors, LightGBM enables recommendation systems to remain relevant and effective over time. This adaptability ensures that users receive tailored suggestions that resonate with their preferences and interests, ultimately driving engagement and customer satisfaction.

Furthermore, LightGBM's efficiency in processing data empowers e-commerce platforms to deliver recommendations in real-time, capitalizing on opportunities for upselling, cross-selling, and improving overall conversion rates. By leveraging LightGBM's capabilities, businesses can navigate the complexities of their data ecosystem with ease, uncovering valuable insights and opportunities for growth.

In essence, LightGBM transcends its role as a mere machine learning algorithm to become a strategic asset for e-commerce entities. Its ability to optimize recommendation systems not only enhances the user experience but also contributes to the bottom line by facilitating more meaningful interactions and driving revenue growth in the competitive e-commerce landscape.

1.5 Research Flow

LightGBM, a formidable supervised learning algorithm, stands out as a beacon of innovation in the realm of recommendation systems, particularly in the dynamic landscape of e-commerce. What distinguishes LightGBM from its predecessor recommendation systems is its exceptional ability to seamlessly navigate the balance between speed and processing efficiency. This algorithm excels in handling larger datasets with remarkable efficacy, accomplishing tasks in considerably less execution time than its counterparts.

In the intricate world of e-commerce, where vast and diverse datasets are the norm, LightGBM emerges as a game-changer. Its proficiency in rapidly processing extensive data sets is a testament to its ²⁶leaf-wise tree growth strategy and gradient-based optimization. This unique approach enables the algorithm to provide accurate, timely, and highly personalized recommendations, addressing the evolving needs of users in real-time.

The significance of LightGBM in recommendation systems is underscored by its capacity to surmount the challenges posed by larger datasets. By enhancing the speed and scalability of recommendation systems, LightGBM not only resolves issues of computational efficiency but also elevates the overall user experience in e-commerce. Its role in improving the effectiveness of customized product recommendations positions it as a substantial advancement in recommendation system technology, promising a future where users can enjoy more tailored and relevant suggestions, thereby fostering increased engagement and satisfaction in the e-commerce space.

1.6 Research Question

In the dynamic realm of online shopping, what are the potential benefits of enhancing item-to-item recommendation systems with LightGBM, a machine learning method that is well-known for its precision and effectiveness? To enhance user contentment, engagement, and total profitability, the research aims to comprehend how item-based suggestion variations and LightGBM's capabilities interact. The goal of the article is to investigate how this integration can disrupt e-commerce recommendation systems by using a strict methodology, extensive testing, and comprehensive conclusions.

1.7 Objective

To adopt LightGBM supervised learning algorithm in recommendation systems represents a strategic leap forward in personalized product suggestions.

- By carefully balancing speed and performance, LightGBM shines in e-commerce. It is clear how skilled LightGBM is in the fast-paced world of e-commerce, where making decisions quickly and accurately is crucial. It is a highly valued asset due to its capacity to sustain outstanding performance and ensure quick computations. With its novel leaf-wise tree growth technique and gradient-based optimization, LightGBM achieves this delicate equilibrium, making it an indispensable tool for real-time real-time dataset navigation. To put it simply, it is a lighthouse of efficiency, quickly and precisely satisfying the expectations of e-commerce.
- It's ability to handle larger datasets with ease and significantly outperform its prior recommendation systems in terms of execution time demonstrate LightGBM's supremacy. Its effective gradient-based optimization and leaf-wise tree growth method allow for quick computations without sacrificing accuracy. This ability guarantees prompt and efficient processing of enormous datasets in the context of recommendation systems, especially in e-commerce, demonstrating LightGBM's leadership in handling the difficulties presented by massive amounts of data while preserving exceptional execution speed.

77 Chapter 2

Literature Review

2.1 Introduction

The literature study examines the development of recommendation systems in e-commerce placing special emphasis on the integration of machine learning algorithms and the search for better datasets. At first, scalability was hampered by algorithmic restrictions, which prompted calls for stronger datasets. The use of dual datasets and supervised learning gained traction as machine learning gained popularity, emphasizing the need of contextual data for tailored suggestions.

Finding a balance between computing time and accuracy continued to be difficult, which led to the investigation of more effective algorithms like XGBoost and hybrid KNN combinations. These methods, however, frequently proved time-consuming and brought about disadvantages like hardware dependence. The review ends with a summary of the field's history and recommendations for future research aimed at maximizing recommendation system effectiveness.

2.2 Content Analysis Table:

Title	purpose	construct	methodology	hypothesis	findings
[5]Session-aware Information Embedding for E-commerce Product Recommendation. Chen Wu, Ming Yan	The purpose of this research is to generate recommendations for anonymous users in e-commerce websites by leveraging the user's temporal behavioral information within the current session, such as views, and purchases.	product embedding session-awareness	Paper proposes a session-aware recommendation system for e-commerce, overcoming challenges with anonymous users using a list-wise deep neural network. Emphasizes representation choices for users with limited online activities.	Session-aware e-commerce recommendation system with list-wise deep neural network excels in addressing challenges with anonymous users' limited online behaviors. List-wise deep neural network proves effective in real-world experiments.	Paper introduces session-aware e-commerce recommendation system, focusing on accurate suggestions for anonymous users with limited online behaviors. List-wise deep neural network proves effective in real-world experiments.
[6]Research on Influential Factors of E-commerce Recommendation User Behavior Intention Pang Xin-Li Economics and Business Administration Hei Long Jiang University P.R.China, Jiang Wei School of Management, Harbin Institute of Technology P.R.China,	Proposing the user adoption model of e-commerce recommendation system based on TAM model.	1.TAM Model, 2.User Adoption, 3.Behavior Intention	Empirical study using a questionnaire survey with 276 young e-commerce users, integrating TAM and IDT to propose a compatibility-focused user adoption model. Recommendations for businesses included.	Paper proposes a user adoption model merging TAM and IDT, examining factors like perceived usefulness, ease of use, and compatibility for e-commerce recommendation systems, guiding system enhancement for businesses.	Paper fills a research gap by comprehensively studying factors influencing e-commerce recommendation system adoption. It explores compatibility, perceived usefulness, and ease of use, offering valuable insights for businesses.

Title	purpose	construct	methodology	hypothesis	findings
<p>¹⁸ [7] A personalized recommendation system with combinational algorithm for online learning. Jun Xiao¹ · Minjuan Wang^{2,3} · Bingqian Jiang¹ · Junli Li²</p>	<p>Paper delves into creating and implementing a personalized recommendation system for online learning, featuring literature review, method assessment, and positive pilot testing feedback.</p>	<p>¹⁸ Knowledge and data technology · Intelligent learning systems ·</p>	<p>Paper develops a personalized online learning recommendation system, encompassing literature review, method assessment, and pilot testing with positive feedback. Integration into Shanghai Lifelong Learning Network enhances engagement and course offerings.</p>	<p>The hypothesis of this paper is that the implementation of a personalized recommendation system for online learning will improve the utilization rate of educational resources, promote learning autonomy and efficiency, and enhance the overall experience for users.</p>	<p>Paper develops a personalized online learning system, incorporating literature review, method assessment, and positive user feedback. Integration into Shanghai Lifelong Learning Network enhances engagement using combinational algorithm .</p>
<p>³⁵ [8] Product Recommendation for e-Commerce System based on Ontology Ni Made Satvika Iswari Faculty of Engineering and Informatics Universitas Multimedia Nusantara Tangerang, g. Wella Faculty of Engineering Informatics Universitas Multimedia Nusantara Tangerang, Andre Rusli Faculty of Engineering and Informatics Universitas Multimedia</p>	<p>¹² The purpose of the paper is to develop a product recommendation system for e-commerce platforms based on ontology. The study aims to provide personalized recommendations to users based on their needs and interests.</p>	<p>1.collaborative filtering, 2.slope one, 3.ontology</p>	<p>Paper creates an e-commerce recommendation system using ontology for personalized suggestions. Method involves building ontology, assigning initial ratings, and implementing in a case study, ensuring targeted sales.</p>	<p>Paper hypothesizes ontology-based e-commerce recommendation system with Slope One algorithm for personalized suggestions, anticipating user and seller benefits through targeted recommendations.</p>	<p>The research gap in the paper pertains to a lack of detailed explanation regarding the implementation stages of the proposed method in the case study. A more comprehensive overview, including specific steps, challenges, and the recommendation system's effectiveness, is needed to address this gap.</p>

Title	purpose	construct	methodology	hypothesis	findings
<p>³⁴ [9] Micro Behaviors: A New Perspective in E-commerce Recommender Systems Meizi Zhou University of Minnesota Zhuyue Ding Data Science Lab, Jiliang Tang Michigan State University Dawei Yin† Data Science Lab, JD.com yindawei@acm.org</p>	<p>To investigate e-commerce recommendations from a micro- behavior perspective</p>	<p>1.RNN 2.attention mechanism,</p>	<p>Paper presents RIB recommendation framework for e-commerce, overcoming challenges in modeling sequential information and diverse micro behaviors. Real-world evaluations demonstrate effectiveness, highlighting dwell time and micro behavior correlations.</p>	<p>43 (1) it models sequential information; (2) it captures micro behaviors (3) it models varied effects of micro behaviors.</p>	<p>The paper identifies a research gap in micro behaviors within e-commerce recommender systems. Introducing the RIB framework, it stresses capturing micro behaviors', effects, demonstrating effectiveness on real datasets, and recommends further exploration in different domains.</p>
<p>¹⁰ [10] A trust-based collaborative filtering algorithm for E-commerce recommendation system Liaoliang Jiang¹ Yuting Cheng² Li Yang Jing Lil Hongyang Yan³ Xiaoqin Wang</p>	<p>Study explores trust-based collaborative filtering in e-commerce, proposing a model combining user trust relationships to enhance accuracy and uncover opportunities.</p>	<p>1.Trusted data 2.User similarity 3.Predicting accuracy</p>	<p>The study delves into trust-based collaborative filtering for e-commerce, countering fake ratings. It proposes a model integrating user trust relationships and trust-based strength metrics, emphasizing accuracy and referencing relevant works to address e-commerce challenges.</p>	<p>19 e paper proposes a trust-based collaborative filtering algorithm for e-commerce, integrating trusted data and user similarity to enhance accuracy, outperforming traditional slope algorithms. Experiments on an Amazon dataset suggest potential improvements .</p>	<p>The paper acknowledges a limitation in the proposed trust-based collaborative filtering algorithm related to the cold-start problem, suggesting that future research could explore incorporating content information to address this challenge and enhance the algorithm's applicability.</p>

Title	purpose	construct	methodology	hypothesis	findings
<p>[38] Applying Recommender Approaches to the Real Estate e-Commerce Market</p> <p>Julian Knoll1(&), Axel Rainer Groß1 , Axel Schwanke2 , Bernhard Rinnl , and Martin Schreyer1</p>	<p>Recommender systems in real estate aim to simplify the overwhelming process of finding suitable properties by utilizing 101 approaches like deep learning, factorization machines, and collaborative filtering. The goal is to improve recommendation quality, streamline housing searches, and enhance user experience on platforms like ImmoWelt</p>	<ol style="list-style-type: none"> 1. Deep learning 2. Factorization machine 3. Collaborative filtering 4. recommendation engines 	<p>The paper assesses recommendation quality in real estate using ImmoWelt Group's data, comparing deep learning, factorization machines, and collaborative filtering algorithms in a 5-fold cross-validation setup.</p>	<p>The paper assumes that applying deep learning and factorization machines real estate e-commerce will enhance quality, reaching performance level similar to collaborative filtering. It aims to demonstrate effectiveness in handling user cold-starts and suggests future focus on real-time updates and further analysis.</p>	<p>The paper addresses the 67 research gap in applying recommender systems to the real estate e-commerce market. It explores challenges, opportunities, and the need for personalized recommendations. Additionally, it compares different approaches, addressing the gap in the evaluation of recommendation methods in real estate.</p>
<p>[20] Log-Based Session Profiling and Online Behavioral Prediction in E-Commerce Websites</p> <p>JAVIER FABRA , PEDRO ÁLVAREZ, AND JOAQUÍN EZPELETA</p> <p>Department of Computer Science and Systems Engineering</p>	<p>The paper centers on e-commerce behavioral prediction, introducing a neural network model for recommendation, customization, and profile prediction, with real-world implementation.</p>	<ol style="list-style-type: none"> 1. log analysis, 2 clustering, 3. neural networks. 4. model checking. 	<p>The paper utilizes log processing for customer profiling, business validation, and training predictive models, integrated into decision-making systems for a Magento-based e-commerce site, emphasizing practical implications</p>	<p>The use of log-based session profiling and online behavioral prediction, integrated into e-commerce websites, will lead to improved customer segmentation, personalized recommendations, and ultimately enhance the overall experience</p>	<p>The paper addresses a research gap in predicting unregistered customer behavior in e-commerce, particularly during website browsing. It emphasizes the need for fine-grained predictive models and practical applications, .</p>

Title	purpose	construct	methodology	hypothesis	findings
<p>³⁰ [13]Competitive Recommendation Algorithm for E-commerce</p> <p>Umutoni Nadine, Huiying Cao*, Jiangzhou Deng</p> <p>College of economics and management Chongqing University of Posts and Telecommunications Chongqing 400065, China</p>	<p>E-commerce recommendation systems aim to solve information overload, providing personalized suggestions. The competitive recommendation algorithm addresses limitations in single approaches by combining user-based, item-based, and Bhattacharyya methods, demonstrating enhanced efficiency and stability in electing reliable items.</p>	<p>1.collaborative filtering; 2.data mining 3.Ranking functions (RF)</p>	<p>E-commerce's competitive recommendation algorithm integrates collaborative filtering, ranking functions, and a competition mechanism for reliable, personalized suggestions, overcoming sparsity and scalability challenges with a hybrid approach.</p>	<p>The hybrid recommendation approach, combining user-based, item-based, and Bhattacharyya methods, enhances system effectiveness, proven through MovieLens dataset experiments, demonstrating increased efficiency and stability while addressing sparsity issues.</p>	<p>The research addresses gaps in single recommendation approaches and collaborative filtering sparsity. The competitive algorithm combines methods, delivering reliable and personalized recommendations, overcoming limitations.</p>
<p>⁹ [14]E-Commerce Recommendation System Using Mahout Phan Duy Hung FPT University Hanoi, Vietnam e-mail: hungpd2@fe.edu.vn</p> <p>Le Dinh Huynh FPT University Hanoi, Vietnam e-mail: huynhldmse0083@fpt.edu.vn</p>	<p>The paper opposes a process for constructing a Mahout-based recommendation system for e-commerce, demonstrated through a case study on Web log mining of a real-world online shopping site.</p>	<p>1.Recommendation system. 2.Collaborating filter. 3.Mahout</p>	<p>The paper introduces a novel e-commerce recommendation system with Mahout, detailing the algorithm using a Web log mining case study, emphasizing personalized recommendations in online retail.</p>	<p>The paper assumes that the User Log likelihood algorithm is most effective for e-commerce recommendations, anticipating insights from collaborative filtering and proposing potential improvements for future</p>	<p>The study identifies a research gap in exploring the impact of response styles on e-commerce recommendation system effectiveness, suggesting future investigation opportunities.</p>

Title	purpose	construct	methodology	hypothesis	findings
<p>[12] User Profiling Approaches for Demographic Recommender Systems</p> <p>Mohammad Yahya Al-Shamir[*]</p> <p>Department, College of Computer Science, King Khalid University, Abha, KSA.</p> <p>Department, Faculty of Engineering and Architecture, Ibb University, Ibb, YE</p>	<p>The study explores Demographic Recommender Systems, emphasizing user profiling and similarity computation methods, tackling challenges in obtaining accurate demographic data using the 100K MovieLens dataset.</p>	<ol style="list-style-type: none"> 1. Demographic data; 2. Similarity computation 3. DRS 	<p>The study suggests a single-attribute profiling approach, treating attributes as isolated profiles and merging predictions. Using the 100K MovieLens dataset, it explores user profiling and similarity computation, addressing age attribute fuzziness.</p>	<p>The paper assesses demographic recommender systems, suggesting the unified profiling approach improves recommendation accuracy, exploring collaboration hybridization, and identifying future research opportunities in the field.</p>	<p>The paper identifies research gap in exploring user profiling and similarity computation methods in Demographic Recommender Systems (DRS), suggesting future research on additional approaches and hybridization.</p>
<p>[15] Hierarchical User Profiling for E-commerce Recommender Systems</p> <p>Long Gu</p> <p>Data Science Lab, JD.com Zhuoye Ding Data Science Lab, JD.com</p> <p>Shuaiqiang Wang Data Science Lab, JD.com Dawei Yin Data Science Lab,</p>	<p>The paper presents the Hierarchical User Profiling (HUP) framework for personalized e-commerce recommendations, utilizing real-time interests at various scales and micro-behaviors, demonstrating superior performance.</p>	<ol style="list-style-type: none"> 1. Recommender systems. 2. Hierarchical user profiling. 3. Pyramid Recurrent. 4. Neural Networks. 	<p>The paper presents Hierarchical User Profiling (HUP) for personalized E-commerce recommendations, utilizing real-time interests and micro-behaviors, showing substantial performance over state-of-the-art methods in extensive experiments.</p>	<p>The paper suggests the Hierarchical User Profiling (HUP) framework to address limitations, emphasizing simultaneous modeling of real-time interests at different granularities, leveraging micro-behaviors, and capturing detailed preferences.</p>	

Title	purpose	construct	methodology	hypothesis	findings
<p>[12] [16] A Hybrid Recommendation System for E-Commerce based on Product Description and User Profile</p> <p>Tessy Badriyah, Erry Tri Wijavanto, Iwan Syarif, Prima Kristalina Informatics Department (Program Studi D4 Teknik Informatika)</p>	<p>The paper develops an e-commerce hybrid recommendation system, merging content-based and collaborative filtering, utilizing text mining for automatic tag generation, and evaluates through experiments</p>	<p>106</p> <ol style="list-style-type: none"> 1. content-based filtering in the Recommendation System. 2. cosine similarity method 	<p>The paper utilizes a hybrid e-commerce recommendation system, combining content-based and collaborative filtering with text mining for tags. Experiments assess tag analysis, recommendations, and system performance improvements.</p>	<p>The hypotheses suggest that a hybrid e-commerce recommendation system, merging content-based, collaborative filtering, and text mining with TF-IDF for tag generation, improves accuracy and overall system performance.</p>	<p>The research fills a gap by proposing a hybrid e-commerce recommendation system, merging content-based, collaborative filtering, and text mining for tag generation. The approach aims to improve recommendation accuracy and relevance.</p>
<p>[17] Online Product Feature Recommendations with Interpretable Machine Learning</p> <p>Mingming Guo The Home Depot Atlanta, GA, USA</p> <p>mingming-guo@homedepot.com</p> <p>Nian Yan The Home Depot Atlanta, GA, USA</p> <p>Xiquan Cui The Home Depot Atlanta, GA, USA</p> <p>Simon Hughes The Home Depot Chicago, IL, USA</p> <p>Khalifeh Al Jadda The Home Depot Atlanta, GA, USA</p>	<p>The paper proposes using interpretable machine learning for product feature recommendations in e-commerce. It formulates the problem as a price-driven supervised learning utilizes Shapley Values, and shows improved rates in online A/B tests over a strong baseline method.</p>	<p>2</p> <ol style="list-style-type: none"> 1. ranking features, tree-based models, linear models, shapley values, product price, left navigation algorithm, customer behaviors 	<p>The paper introduces an interpretable machine learning method for recommending product features to online customers, employing Shapley Values to interpret importance. The approach outperforms a baseline in online A/B tests.</p>	<p>The paper shows that interpretable machine learning, using Shapley Values, successfully recommends product features in a price-driven supervised learning task, outperforming a baseline in online A/B tests.</p>	<p>The research identifies a gap in the literature regarding product feature recommendations and highlights the significance of interpretable machine learning to effectively address this neglected aspect.</p>

Title	purpose	construct	methodology	hypothesis	findings
<p>[37] [18] A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. Yung-Ming Li *, Chun-Te Wu, Cheng-Yang Lai</p>	<p>This research proposes an advanced model for personalized product recommendations, prioritizing preference similarity, recommendation trust, and social relations to enhance accuracy .</p>	<p>Questionnaires interviews</p>	<p>The paper introduces a social recommender system for e-commerce, incorporating similarity, trust, and relationship analyses for improved product recommendations. It utilizes Analytic Hierarchy Process (AHP) for personalized weight determination, emphasizing accuracy in Yahoo! Shopping data.</p>	<p>The paper assumes a social recommender system for e-commerce integrating similarity, trust, and relationship analyses, aiming to surpass benchmark methodologies in recommendation accuracy. It works personalized factor importance weights applications for service improvement and product</p>	<p>Research notes limitations like data constraints and absence of social networking in e-commerce. Future works involve exploring social relations, larger experiments, and demographic factors in recommendations.</p>
<p>[10] [19] Social recommendation: A user profile clustering-based approach Sara Ouafrouh Ahmed Zellou Ali Idri</p>	<p>The paper proposes a novel social recommendation approach using a modified partitioned algorithm to cluster user profiles based on similarity, focusing on recommendations.</p>	<p>86 1. Filtering such as Bayesian Networks. 2. latent semantic, and clustering.</p>	<p>The paper proposes a novel approach to recommender systems, employing multidimensional user profiles and a unique similarity function to cluster similar preferences.</p>	<p>The hypothesis suggests that utilizing a modified partitioned algorithm for clustering user profiles improves accuracy and scalability, offering relevant product suggestions based on multidimensional structures.</p>	<p>The research gap is the absence of a technique that combines a user's personal interests with those of similar profiles as guide. The proposed approach aims to address this gap.</p>

Title	purpose	construct	methodology	hypothesis	findings
[62] [20]Collaborative filtering recommendation algorithm based on KNN and Xgboost hybrid Yingxian Lil*, Junwu 65 2, Min Yang3 College of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China	Paper introduces KNN-Xgboost hybrid collaborative filtering algorithm, addressing matrix sparseness with KNN and employing Xgboost multi-classifiers. Exhibits enhanced accuracy over traditional methods in experiments.	Evaluation metrics Sparsity of the matrix Hybrid collaborative filtering	Paper introduces KNN-Xgboost hybrid algorithm, addressing matrix sparseness with KNN and employing Xgboost multi-classifiers. Demonstrates improved accuracy optimal at tree depth 3.	The paper suggests KNN-Xgboost hybrid algorithm, enhancing recommendation accuracy by combining KNN for matrix sparseness and Xgboost for multi-classifiers. Exhibits improved accuracy over traditional methods.	Research notes limitations like data constraints and absence of social networking in e-commerce. Future opportunities involve exploring social relations, larger experiments, and demographic factors in recommendations.
[31] [21]Product Recommendation Based on Content-based Filtering Using XGBoost Classifier Zeinab Shahbazi, Yung-Cheol Byun*	Paper suggests a content-based product recommendation system with XGBoost, leveraging user click data from Jeju online shopping mall. Demonstrates system's superiority over others.	Machine Learning, Classification	Paper utilizes XGBoost in content-based filtering for product recommendations from Jeju online shopping mall. Highlights superior performance and user click data.	XGBoost-based product recommendation system, utilizing user click data, outperforms individual methods, demonstrating higher accuracy, recognition rates, and lower error rates.	The research gap is the absence of a technique that combines a user's personal interests with those of similar profiles for recommendations. The proposed approach aims to address this gap.

2.3 content analysis

A lot of the works that were made upon different outlooks in e-commerce sites for recommendation systems that gave out quite interesting ideas as well as performances, but every developing research has its future works that were not implemented at those times, such were some works that wanted to work with better datasets that can give further insights on better work [5][6][7]. But some of the works were actually unable to large dataset due to incapability of the algorithm [8][9][10]. Going further some of the later workings suggested getting involved with machine learning algorithms [8], some even insisted on supervised machine learning algorithm for a better outcome [11][12], corresponding to that authors suggested to use dual or better acknowledged datasets to verify the works with stronger impression [6][13][14], contextual data usage was also a popular suggestion [15][19]

while machine learning algorithm started to be implemented, it was not very efficient due to its lower accuracy because of the lower performing algorithms [16][22]. Even after the usage of supervised learning algorithm the efficiency was not that much feasible due to higher time taken for better performance, so the balance was off due to improper combination between accuracy and time [18][17]. Further development went towards higher performing algorithms such as xgboost using it individually and also using it in a hybrid manner with KNN which did give some better results at the expense of time both the hybrid [20] and individual [21] algorithms were really time taxing putting it off balance, as well the drawbacks were numerous while using them such as unable to work with larger datasets without higher levels of hardware or giving out incomprehensive results .

LightGBM, a gradient boosting framework developed by Microsoft, has emerged as a pivotal tool in supervised learning due to its exceptional balance of performance and efficiency, especially when handling large datasets. Its significance lies in its ability to address challenges that earlier recommendation systems faced, notably those dating back to the establishment of the first e-commerce site, the Boston Computer Exchange, in 1982.

One of the primary advantages of LightGBM is its ability to handle substantial volumes of data with remarkable execution speed. This efficiency stems from its innovative algorithms, which optimize memory utilization and leverage parallel computing capabilities. As a result, LightGBM outperforms its predecessors in recommendation systems, such as collaborative filtering and matrix factorization methods, in terms of both accuracy and time efficiency.

Furthermore, LightGBM's implementation of gradient boosting techniques ensures superior model performance. By sequentially improving weak learners, LightGBM constructs powerful ensemble models that excel at predictive tasks, including classification and regression, across various domains.

The reduced time consumption of LightGBM is particularly noteworthy, as it enables faster model training and deployment cycles. This accelerated workflow enhances the agility of data-driven processes, allowing organizations to adapt swiftly to evolving market dynamics and customer preferences.

LightGBM stands out as a significant breakthrough in machine learning, offering a compelling solution to contemporary data challenges. Its efficiency in achieving optimal performance with minimal overhead makes it a preferred choice for both researchers and practitioners across various industries. By facilitating faster model training and deployment cycles, LightGBM enhances decision-making processes and fosters innovation in domains like e-commerce, banking, healthcare, and beyond.

The framework's ability to handle large datasets with speed and precision contributes to its widespread adoption and relevance in today's data-driven landscape. As organizations strive to leverage data for competitive advantage, LightGBM emerges as a pivotal tool, enabling them to extract actionable insights and drive meaningful outcomes. Its impact extends beyond traditional boundaries, fueling advancements in predictive analytics, personalized recommendations, risk assessment, and other critical areas where data-driven decision-making is paramount.

2.4 Applied-Algorithms

2.4.1 ¹¹Lightgbm

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.

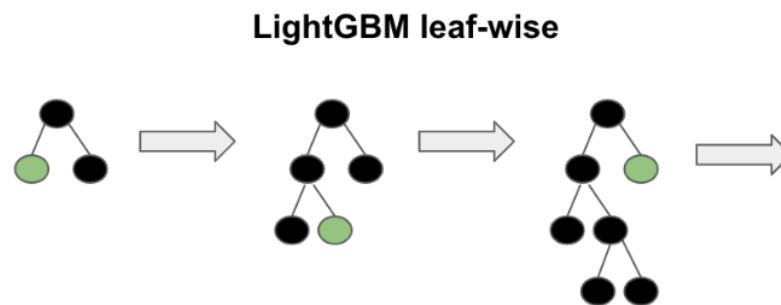


Fig. 2.1. Architecture of the Lightgbm classifier^[23]

⁴⁷LightGBM divides the tree leaf-wise, while other boosting methods expand the ⁵tree level-wise. The leaf with the greatest delta loss is the one that it grows. With respect to the ⁴⁷level-wise approach, the loss of the leaf-wise algorithm is lower because it is fixed. Leaf-wise tree development may result in overfitting in short datasets and possibly increase the complexity of the model.

LightGBM is indeed a powerful algorithm in the realm of gradient boosting frameworks. Here's a breakdown of the key features and techniques that make LightGBM stand out:

1. **Gradient Boosting Decision Trees (GBDT):** LightGBM follows the GBDT framework, which sequentially adds decision trees to the ensemble, with each tree correcting the errors of its predecessors.
2. **Gradient-based One-Side Sampling (GOSS):** GOSS is a technique used in LightGBM to efficiently handle data during the training process. It focuses on sampling the instances with large gradients while keeping the instances with small gradients. This helps in reducing the number of data instances without sacrificing the overall quality of the model.
3. **Exclusive Feature Bundling (EFB):** EFB is another technique employed by LightGBM to improve efficiency. It bundles exclusive features together during the training process, which helps reduce the number of feature combinations to consider, thereby speeding up the training process.
4. **Efficiency and Scalability:** LightGBM is known for its efficiency and scalability, which makes it suitable for large datasets. The techniques like GOSS and EFB contribute significantly to its efficiency.
5. **Handling Various Data Types:** LightGBM can handle a wide variety of data types, including categorical features, numerical features, and combinations of both.
6. **Flexibility in Hyperparameter Tuning:** LightGBM offers a diverse set of hyperparameters that can be fine-tuned to optimize model performance for different datasets and problem types.
7. **Supported Problem Types:** LightGBM supports regression, binary classification, multiclass classification, and ranking problems, making it versatile for a range of machine learning tasks.

Overall, LightGBM's combination of efficient algorithms, innovative techniques, and flexibility in hyperparameter tuning makes it a popular choice for both practitioners and researchers in the field of machine learning.

2.4.2 XGBoost

You may better comprehend your data and make decisions by using the powerful machine-learning algorithm XGBoost. A gradient-boosting decision tree implementation is called XGBoost. To improve their machine-learning models, academics and data scientists from all around the world have been using it. It functions with either department- or level-wise tree development.

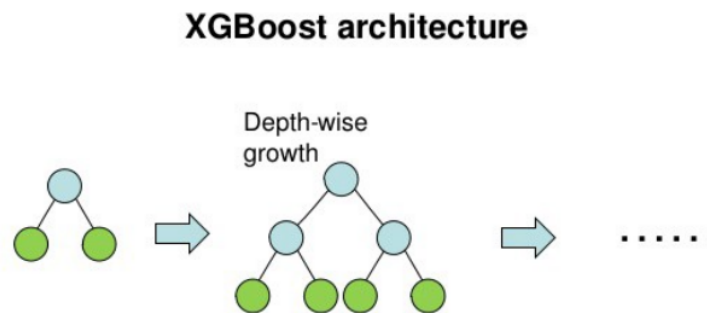


Fig. 2.2. Architecture of the Xgboost classifier[24]

Combining several weak learners into one powerful learner is how XGBoost operates. Machine learning models that perform marginally better than random guessing are referred to as weak learners. On the other hand, combining weak learners can result in a far more accurate strong learner. Training multiple decision trees is how XGBoost operates. To create the final forecast, the predictions from each tree—which is trained on a subset of the data—are added together. The GBM algorithm has been improved with XGBoost. A more regularized model is used by XGBoost, which helps to prevent overfitting. This is the primary difference.

Here's a breakdown of the features and characteristics in XGBoost:

1. **Boosting Technique:** XGBoost⁵⁸ belongs to the family of boosting algorithms, where the focus lies on iteratively improving the model by correcting the errors of previous models. XGBoost optimizes a specific loss function, which helps in improving the overall performance of the ensemble.
2. **Supervised Learning:** XGBoost is primarily used in supervised learning tasks, including regression and classification problems.
3. **Parallel Processing:** XGBoost supports parallel processing, making it efficient for training on both single and distributed systems. It can leverage frameworks like Hadoop and Spark for distributed computing.
4. **Regularization:** XGBoost⁶⁸ offers various regularization techniques to combat overfitting, including L1 and L2 regularization, which penalize the complexity of the model.
5. **Auto Tree Pruning:** XGBoost implements tree pruning strategies to control the growth of decision trees, preventing them from becoming overly complex and reducing the risk of overfitting.
6. **Handling Missing Values:** XGBoost can handle missing values in the dataset, either by learning an optimal direction during training or by placing missing values in the most appropriate side during tree construction.

Overall, XGBoost's combination of performance, scalability, and robustness has made it a popular choice for various machine learning tasks, particularly in scenarios where predictive accuracy is crucial. Its flexibility and efficiency make it suitable for both research and practical applications across different domains.

2.4.3 Random-Forest

For a range of applications, such as regression and classification, Random Forest is a powerful machine learning technique. A random forest model is composed of numerous little decision trees, or estimators, each of which generates a unique set of predictions; thus makes the method ensemble in nature. A prediction that is more accurate is generated by the random forest model by combining the predictions made by the estimators.

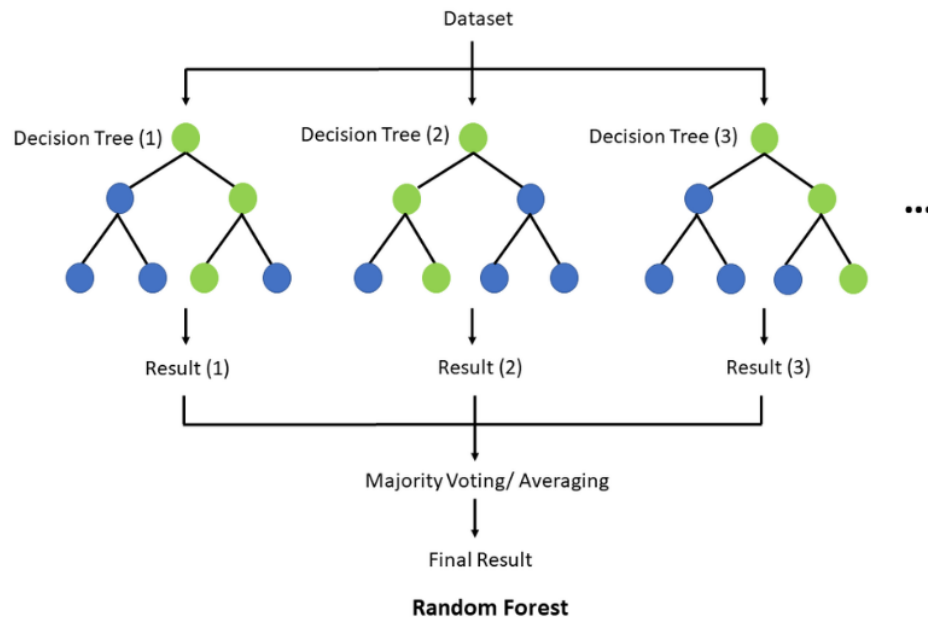


Fig. 2.3. Architecture of the Random Forest classifier[25]

Each decision tree in the ensemble of decision trees used in the random forest technique is made up of a bootstrap sample, which is a sample of data taken from a training set with replacement. One-third of that training sample—referred to as the "out-of-bag" data—is reserved as test data. Feature bagging is then used to introduce yet another randomization, increasing dataset variety and decreasing decision tree correlation. Depending on the nature of situation, the determination of the prediction will differ. The individual decision trees in a regression job will be averaged, and in a classification work, the predicted class will be determined by a majority vote, or the most common categorical variable. then cross-validation is performed using the oob sample. The features you've listed highlight the strengths and capabilities of Random Forests, a powerful ensemble learning method based on decision trees. Let's break down these features:

1. **Accuracy:** Random Forests often exhibit high accuracy compared to other machine learning algorithms, making them well-suited for a wide range of classification and regression tasks.
2. **Scalability:** Random Forests are efficient and can handle large databases with thousands of input variables without requiring variable deletion. This scalability makes them applicable to various real-world scenarios with extensive datasets.
3. **Variable Importance:** Random Forests provide estimates of the importance of variables in the classification process. This feature is valuable for understanding which features contribute most significantly to the model's predictions.
4. **Unbiased Generalization Error Estimation:** During the process of building the forest, Random Forests generate an internal unbiased estimate of the generalization error. This helps in assessing the model's performance and tuning parameters effectively.
5. **Handling Missing Data:** Random Forests have effective methods for estimating missing data, allowing them to maintain accuracy even when a large proportion of the data is missing.
6. **Balancing Error in Class Population:** Random Forests provide methods for handling class imbalance in datasets, ensuring that the model is not biased towards the majority class.

In summary, Random Forests offer a comprehensive set of features that contribute to their accuracy, robustness, and versatility across various machine learning tasks. Their ability to handle large datasets, deal with missing data, and provide insights into variable importance makes them a popular choice in both research and practical applications.

2.4.4 ¹⁵Support-Vector-Machine

A supervised learning technique called a support vector machine (SVM) is used in machine learning to perform regression and classification tasks. SVMs are especially effective at handling binary classification issues, which call for splitting a data set's elements into two groups.

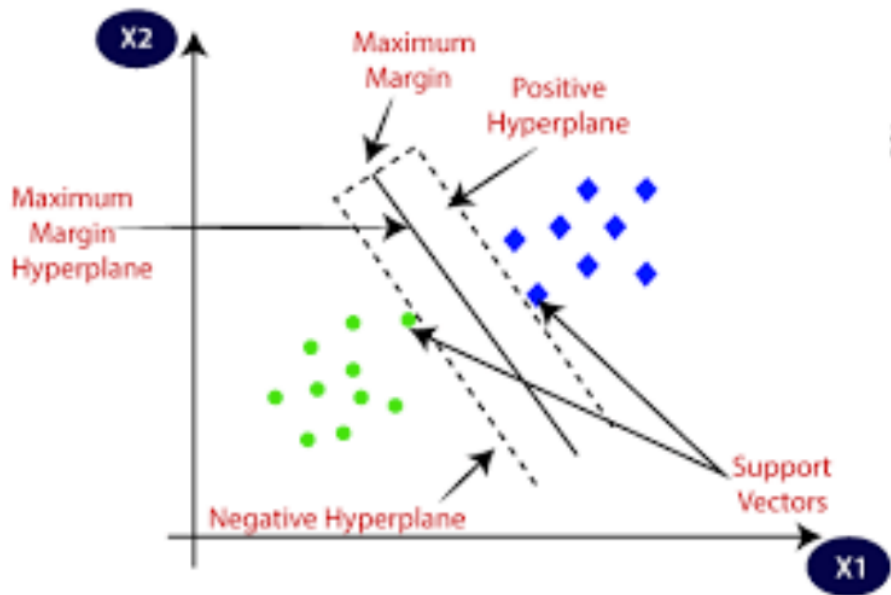


Fig. 2.4. Architecture of the SVM classifier[26]

The optimal border or region is referred to as a hyperplane and is found using the Support Vector Machine (SVM) algorithm. The SVM method locates the line segment that is closest between the two classes. Support vectors are these points. Margin is the length of time that separates the vectors from the hyperplane. And maximizing this margin is SVM's objective. An ideal hyperplane is one that has the most margin.

⁵¹ Support Vector Machines (SVMs) are powerful supervised learning models used for classification and regression tasks. Here are the key features and characteristics of SVMs:

⁵² 1. **Effective in High-Dimensional Spaces:** SVMs are effective in high-dimensional spaces, making them suitable for problems with many features, such as text classification, image recognition, and gene expression analysis.

2. **Sensitivity to Feature Scaling:** SVMs are sensitive to feature scaling, so it's essential to scale the features before fitting the model to ensure optimal performance.

⁸⁰ 3. **Kernel Trick:** SVMs can efficiently perform nonlinear classification using the kernel trick. ⁴² ⁵⁴ Kernels such as linear, polynomial, radial basis function (RBF), and sigmoid are commonly used to map data into higher-dimensional spaces where it's easier to find a separating hyperplane.

³³ 4. **Margin Maximization:** SVMs aim to find the hyperplane that maximizes the margin, i.e., the distance between the hyperplane and the closest data points (support vectors). Maximizing the margin helps improve the model's generalization ability and resistance to overfitting.

⁵⁵ 5. **Binary Classification:** SVMs are primarily used for binary classification, where they aim to find the optimal hyperplane that separates the data into two classes with the maximum margin.

²⁹ 6. **Extension to Multiclass Classification:** SVMs can be extended to handle multiclass classification problems using techniques such as one-vs-one (OVO) or one-vs-all (OVA) classification.

In summary, SVMs offer several features and characteristics that make them well-suited for various classification tasks, especially in high-dimensional spaces and when dealing with binary or multiclass classification problems. Their ability to maximize the margin, handle nonlinear data through kernel tricks, and produce sparse solutions contributes to their popularity in both academic research and practical applications.

2.5 Summary

⁶ In the dynamic realm of e-commerce, recommendation systems play a pivotal role in enhancing user experience and driving sales. Over the years, researchers and practitioners have explored various outlooks and methodologies to refine recommendation algorithms.

While numerous approaches have yielded promising insights and performances, the journey towards optimal recommendation systems is paved with challenges and unexplored avenues.

A significant aspect of advancing recommendation systems lies in leveraging datasets that offer deeper insights and better reflect user preferences and behaviors. However, many existing works have encountered limitations in handling large datasets due to algorithmic incapacities. Despite the recognition of the importance of richer datasets, some research endeavors were constrained by the scalability issues of their algorithms.

To address these challenges, recent research efforts have advocated for the integration of machine learning algorithms into recommendation systems. While machine learning holds immense potential for enhancing recommendation accuracy, early implementations faced hurdles in achieving satisfactory performance. Supervised learning algorithms emerged as a popular choice among researchers aiming for improved outcomes. The emphasis on supervised learning underscores the importance of leveraging labeled data to train models effectively and make informed recommendations.

Moreover, the significance of dual or diversified datasets has been highlighted by several authors. By incorporating diverse data sources, recommendation systems can capture a more comprehensive understanding of user preferences and behaviors. Contextual data usage has also gained traction, as it enables recommendation systems to adapt to changing user contexts and preferences, thereby enhancing user engagement and satisfaction.

Despite the advancements in machine learning algorithms, efficiency and performance remain critical concerns. Many existing algorithms suffer from lower accuracy and longer processing times, posing challenges in striking a balance between recommendation quality and computational efficiency. Supervised learning algorithms, while offering improved accuracy, often entail longer training times, thereby impeding real-time recommendation capabilities.

In response to these challenges, researchers have explored higher-performing algorithms such as XGBoost and hybrid approaches like combining XGBoost with KNN (K-Nearest Neighbors). While these approaches have demonstrated promising results in terms of recommendation accuracy, they come at the expense of increased computational complexity and processing time. The time-taxing nature of these algorithms has led to a disproportionate trade-off between recommendation quality and computational efficiency.

Furthermore, the drawbacks associated with higher-performing algorithms, such as their inability to handle larger datasets without substantial hardware resources and their tendency to produce incomprehensive results, pose significant challenges to their widespread adoption in e-commerce recommendation systems.

To overcome these challenges and chart a path towards more efficient and effective recommendation systems, future research endeavors must focus on several key areas:

1. **Algorithmic Optimization:** There is a need to explore novel algorithmic techniques and optimizations ⁸⁹ that strike a balance between recommendation accuracy and computational efficiency. This may involve refining existing algorithms or developing entirely new approaches tailored to the specific requirements of e-commerce recommendation systems.
2. **Scalability and Performance:** Addressing the scalability and performance challenges associated with handling large datasets is paramount. Researchers should explore distributed computing paradigms and parallel processing techniques to enable efficient processing of vast amounts of data in real-time.
3. **Hybrid Approaches:** Hybrid approaches that leverage the strengths of multiple algorithms while mitigating their individual drawbacks hold promise for enhancing recommendation quality while minimizing computational overhead. Exploring innovative combinations of algorithms and techniques could unlock new avenues for improving recommendation systems.
4. **User-Centric Design:** Recommendations should be tailored to the unique preferences and behaviors of individual users. Incorporating user feedback mechanisms and personalization strategies can enhance the relevance and effectiveness of recommendations, ultimately improving user satisfaction and engagement.
5. **Ethical Considerations:** As recommendation systems exert a significant influence on user behavior and decision-making, it is imperative to address ethical considerations such as transparency, fairness, and privacy. Researchers and practitioners must prioritize the development of ethically responsible recommendation systems that uphold user trust and integrity.

¹⁷ In the realm of e-commerce recommendation systems, there has been a continuous pursuit of innovation and performance enhancement. Early research efforts underscored the importance of better datasets for deeper insights, yet faced challenges with algorithmic limitations in handling large data volumes. The transition towards machine learning algorithms, particularly supervised learning, aimed to improve accuracy but encountered efficiency issues due to prolonged processing times.

LightGBM, a gradient boosting framework developed by Microsoft, has emerged as a transformative solution in supervised learning. Its exceptional performance and efficiency, especially in handling large datasets, have revolutionized recommendation systems. LightGBM's innovative algorithms optimize memory utilization and leverage parallel computing capabilities, resulting in superior execution speed and model performance compared to traditional methods like collaborative filtering and matrix factorization.

One of LightGBM's key advantages lies in its ability to construct powerful ensemble models through gradient boosting techniques, enhancing predictive accuracy across various domains. Its reduced time consumption accelerates model training and deployment cycles, fostering agility in data-driven processes and enabling organizations to adapt swiftly to evolving market dynamics.

LightGBM's efficiency has positioned it as a preferred choice for researchers and practitioners across industries, driving innovation in e-commerce, banking, healthcare, and beyond. Its capability to handle large datasets with speed and precision fuels advancements in predictive analytics, personalized recommendations, and risk assessment, empowering organizations to extract actionable insights and drive meaningful outcomes in today's data-driven landscape.

Chapter 3

Methodology

3.1 Introduction

³⁹ **LightGBM** is a gradient boosting ensemble method utilized by the Train Using AutoML tool based on decision trees. LightGBM is a decision tree-based technique that may be applied to regression as well as classification. High performance with distributed systems is the focus of LightGBM's optimization.

⁷⁸ **Predictive modeling** is a frequent statistical method for forecasting future behavior. One type of data-mining technology that helps anticipate future results is predictive modeling. It works by evaluating both historical and present data to create a model.

⁹⁶ **A Recommender System** is a program that can forecast a user's future preferences for a group of things and suggest the items that will be most popular.

A Content Based Filtering is a recommender system that utilizes machine learning algorithms to predict and suggest items to users that are new but similar to their preferences. For this method to be effective, it requires a well-defined set of features for each product and a record of the user's past choices. The recommendation process relies on assessing product characteristics and aligning them with the user's preferences.

Collaborative filtering recommender systems rely on user-item interactions to suggest new products, emphasizing collective user behavior patterns instead of individual item features. By analyzing past interactions, the system identifies similarities between users and recommends items based on the preferences of users who share similar behaviors, providing personalized suggestions without explicitly considering specific item characteristics.

3.2 Models

In the beginning we selected a dataset that met our requirements and selected our features according to their usefulness. And as we all know a machine learning algorithm works according to the selected node or data working its way towards its output and the selected features work as in the input that the output maps with accordance lightgbm creates a model, which gives a recommendation based on the learning by mapping new data to take it closer to perfection.

Note that: machine learning algorithms do not give us perfect valid results. We found our thesis result 94 percent. In case we put an input value in our model which is not available in our 75 classes for the first dataset and 70 classes for the second. We find a result but the result is not valid. but machine learning algorithm tries to give us a nearest valid result of it.

For a stepwise

- **Step 1** : we select some input feature which is used to as reference for mapping the output
- **Step 2**: we apply the Lightgbm classifier algorithm and acquire the prediction model with our given dataset feature.
- **Step 3**: Then if a user input data is used in the acquired prediction model we find an outcome.
- **Step 4**: here the generic product is predicted, which is then sorted according to the rating.
- **Step 5**: then the top 5 data is presented according to sorting for recommendation.

The workings here have been ⁵⁷ shown in figure 3.1

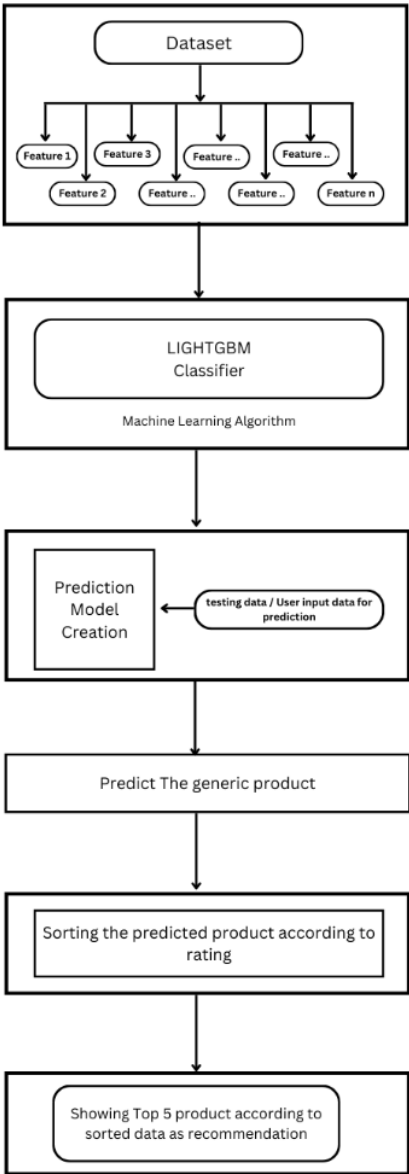


Fig. 3.1. Overall System Architecture

3.3 Description:

3.3.1 First Stage

For a broader view of the workings two large dataset are taken. One data set is from amazon(2019). Another dataset is amazon(after 2020),the reason for choosing them both are due to their features being in line with our work. As we start working on dataset for preprocessing, and start to function with our required columns for feature selection as the dataset is huge, then we start removing the null data that obstructs us from getting the valid results in stage 1. After that the dataset is ready at the end of this stage for further use, but it is still enormous (about 14 million).so we select random amount of data of 100k,500k and 1million simultaneously for going through the algorithm.

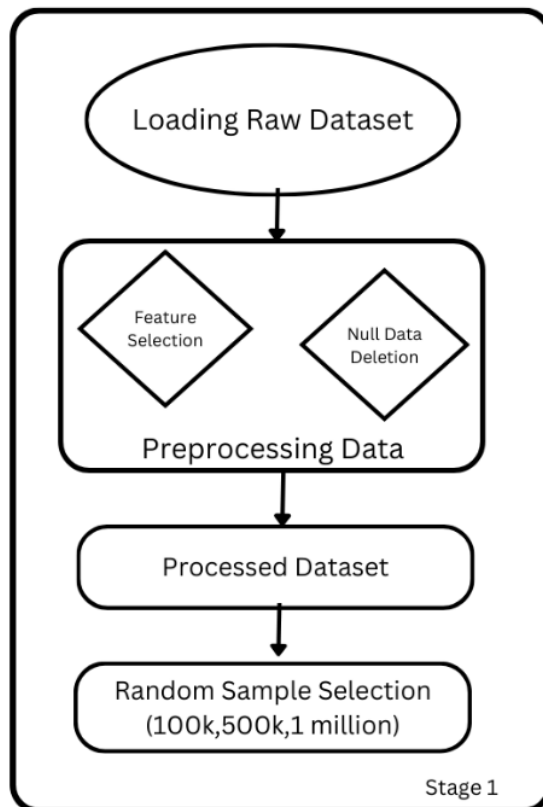


Fig. 3.2. First Stage

In order to import a raw dataset for preprocessing, data analysis for model training must be performed first. Examine the dataset in its entirety to determine its size, composition, and nature of data. Planning the preparation procedures and comprehending the data distribution are aided by this stage. In data analysis and machine learning projects, loading a raw dataset from a CSV file into memory is a typical task.



Fig. 3.3. Loading dataset

Preprocessing huge datasets entails a number of crucial procedures to guarantee the relevance and quality of the data. First, figuring out which columns are crucial for analysis facilitates a more efficient process. It is crucial to handle missing data; approaches include imputation, deletion, or sophisticated methods like predictive modeling. The results of analysis and data integrity are affected by each technique. Imputation is the process of substituting mean, median, or estimated values from the available data for missing values. When only a few variables are impacted or non-critical variables are affected, it is efficient to remove rows or columns with missing values. More sophisticated methods use correlations with other data to forecast missing numbers.

Evaluation post-processing is crucial to assess how handling strategies influence dataset integrity and analytical outcomes. Documentation of decisions and methodologies ensures transparency and reproducibility. Preprocessing extends beyond handling missing values and may include data transformation, normalization, feature engineering, and outlier detection. Systematic preprocessing ensures datasets are optimized for analysis and model training, enhancing the reliability and interpretability of results. By adhering to robust preprocessing practices, analysts can derive meaningful insights from large datasets while mitigating the risks associated with data inconsistencies and biases.

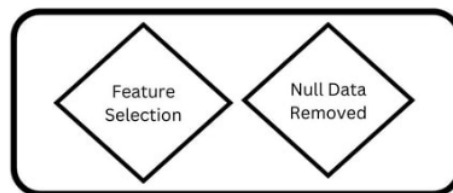


Fig. 3.4. Preprocessing Data

A processed dataset, devoid of null values and containing relevant features, is crucial for accurate analysis and model development. By addressing null values through robust pre-processing techniques, dataset integrity is preserved, minimizing biased results. Clean datasets facilitate more accurate modeling, insightful analysis, and efficient subsequent stages in the data analysis pipeline. With noise eliminated and irrelevant variables removed, the focus shifts to the most salient aspects of the data, enhancing interpretability. Furthermore, processed datasets contribute to scalability and generalizability, ensuring robust performance across diverse datasets and real-world scenarios. In essence, they lay a solid foundation for data-driven decision-making and predictive modeling.

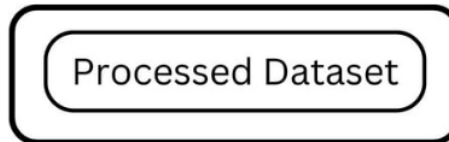


Fig. 3.5. Processed Dataset

Post-preprocessing, the dataset, though extensive at 14 million records, faces computational challenges. To address this, subsets of 100k, 500k, and 1 million records are randomly selected for algorithmic exploration. This strategy alleviates computational burdens and enhances processing efficiency while maintaining dataset representativeness through random sampling. Simultaneous evaluation of different subset sizes enables insights into algorithm performance under varying data volumes, aiding decisions on resource allocation and algorithm selection. Moreover, working with smaller subsets expedites iterative experimentation, facilitating rapid prototyping, parameter tuning, and hypothesis testing. By employing random sampling, analysts can efficiently navigate large datasets, streamline development cycles, and accelerate insights generation, ultimately enhancing the effectiveness and scalability of data analysis workflows.

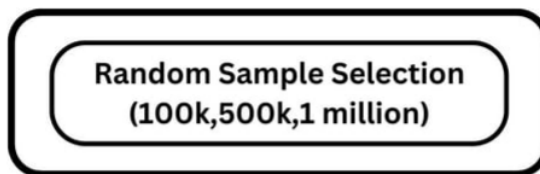


Fig. 3.6. Sample Selection

3.3.2 Second Stage

As we go forward towards stage 2 after completing processing then selection and we input our categorical data as the work that we are researching is based on e-commerce market which is already in a categorical state. In stage 2, we shall split the data into two parts which will be the test data(30 %) and the other being train data(70%) for their individual required purpose. Then Lightgbm classifier algorithm is applied upon the trained data which gives us a prediction model which is used in further conjecture.

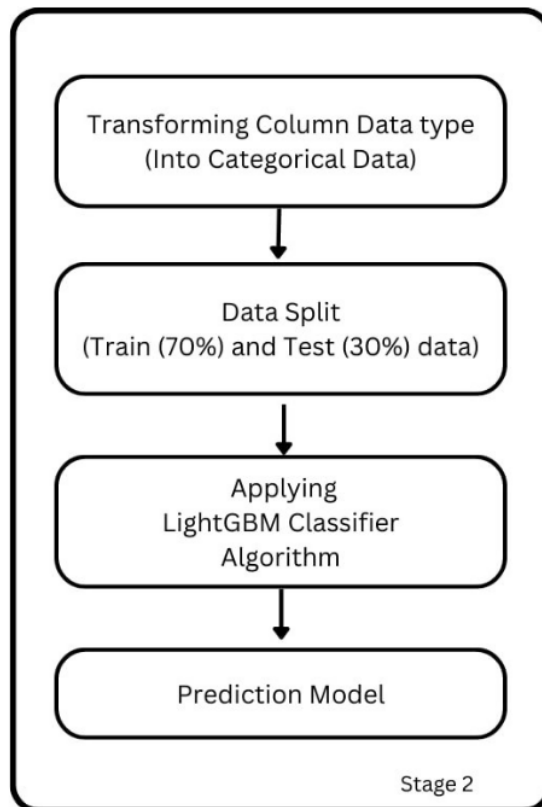



Fig. 3.7. Second stage

Transforming columns into categorical data is a crucial step before splitting datasets for analysis. This process involves converting relevant columns into categorical variables, which represent discrete, qualitative characteristics. Categorical variables can include nominal categories, such as gender or product type, or ordinal categories, like education level or customer satisfaction rating. By transforming columns into categorical data, analysts can effectively organize and classify information, facilitating clearer insights into patterns and relationships within the dataset. Categorical variables enable more intuitive interpretation of data and can enhance the performance of certain machine learning algorithms that are designed to handle categorical inputs. Splitting datasets after transforming columns into categorical data allows for stratified sampling, ensuring that each subset maintains proportional representation of the different categories. This is particularly important for maintaining the integrity and representativeness of the data across training, validation, and test sets. In summary, transforming columns into categorical data before splitting datasets enhances data organization, facilitates intuitive interpretation, and supports more effective analysis and model training processes.



Transforming Column data type
(Into Categorical Data)

Fig. 3.8. Transformation

As is standard in machine learning, 70% of the data is used for training and 30% is used for testing. Seventy percent of the data are in the training set, which is used to train the model by identifying patterns and relationships in the dataset. Thirty percent of the data is the testing set, which is used as a separate dataset to assess the performance of the model. By ensuring that the model's prediction skills are evaluated on unobserved data, this split aids in the identification of possible problems such as under- or overfitting. Reliability in model construction and precise prediction are enhanced by the 70-30 split, which achieves a balance between robust evaluation and sufficient training.



Data Split
(Train (70%) and Test (30%) data)

Fig. 3.9. Data split

After splitting the data, the LightGBM algorithm is applied to build a prediction model. LightGBM, a gradient boosting framework, is renowned for its efficiency and accuracy in handling large datasets. It employs a tree-based learning algorithm, where decision trees are constructed sequentially to minimize the loss function. LightGBM's key features include leaf-wise growth and histogram-based algorithms, which optimize training speed and memory usage while maintaining predictive performance. Its ability to handle categorical features directly without one-hot encoding further enhances efficiency and reduces computational overhead. During model training, LightGBM iteratively improves predictive accuracy by minimizing residual errors. It utilizes gradient descent techniques to update tree structures, optimizing model parameters to fit the training data. After training, the LightGBM model can efficiently predict outcomes for unseen data, leveraging the learned patterns and relationships discovered during training. Its speed, scalability, and high predictive accuracy make LightGBM a preferred choice for various machine learning tasks, particularly with large and complex datasets.

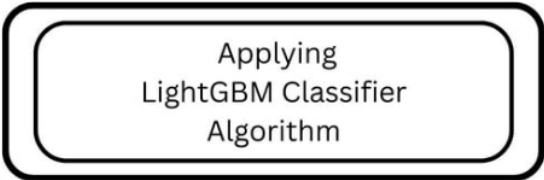


Fig. 3.10. Applying Lightgbm

After applying the LightGBM algorithm, we obtain predictions that serve as crucial insights for the final stage of our work. These predictions represent the model's estimates based on learned patterns and relationships within the data. Leveraging the efficiency and accuracy of LightGBM, we generate actionable insights for decision-making processes. These predictions inform strategic directions, optimize resource allocation, and enhance operational efficiency. By harnessing the power of advanced machine learning techniques like LightGBM, we unlock valuable insights that drive innovation, competitiveness, and success in our endeavors.

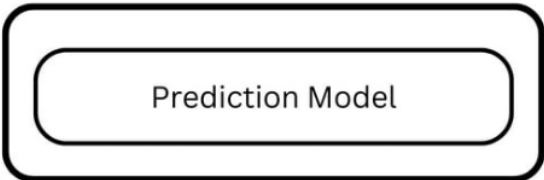


Fig. 3.11. Prediction Model

3.3.3 Final Stage

After all that in the beginning of stage 3, the time taken is calculated which is required for creating the prediction model, then using the test data in the prediction model we evaluate the accuracy percentage shown in figure 3.3 . Afterwards new data is used in the model where the algorithm performs based on what it has learned and gives off a generic product detection. Finally giving us a recommendation by sorting the predicted product data according to the top 5 ratings.

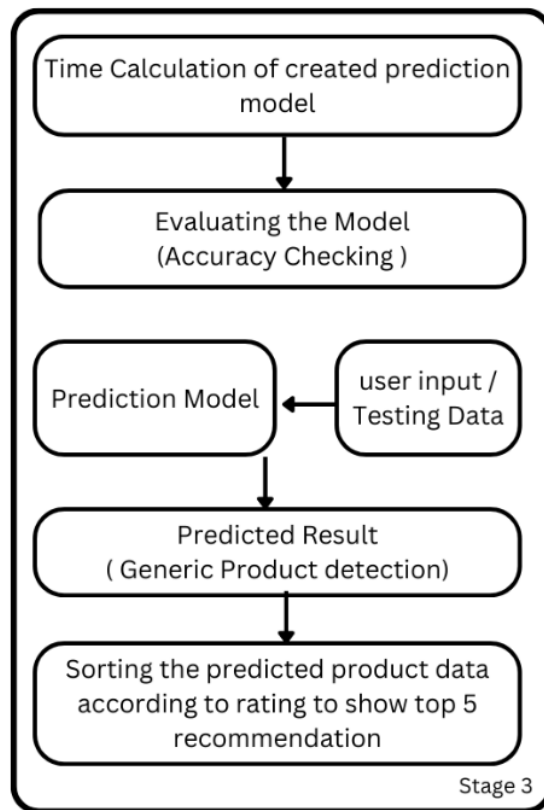


Fig. 3.12. Last Stage

Size of the dataset, model complexity, and hardware resources all affect how long it takes the LightGBM model to forecast. The efficient LightGBM algorithm usually produces quick predictions, even for big datasets. Prediction speed depends on a number of factors, including available processing power and feature count. Prediction time can be precisely estimated with the use of empirical measurement on representative data. Greater computation time may be needed for larger datasets or more complicated models, even while simpler models and smaller datasets produce predictions almost instantly. Although not perfect, LightGBM's effectiveness allows for quick prediction creation, which aids in well-informed decision-making and boosts operational effectiveness across a range of fields.

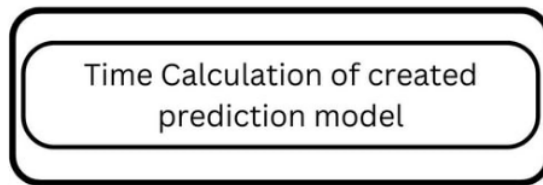


Fig. 3.13. Time Calculation

Model evaluation for accuracy assesses the predictive performance of a machine learning model by comparing its predictions to the actual outcomes in the test dataset. The accuracy metric measures the proportion of correctly predicted instances out of the total number of instances. It provides a simple and intuitive assessment of the model's overall correctness.

During evaluation, the test dataset is fed into the trained model, and predictions are generated. These predictions are then compared to the actual values in the test dataset to determine the number of correct predictions. The accuracy score is calculated as the ratio of correct predictions to the total number of predictions.

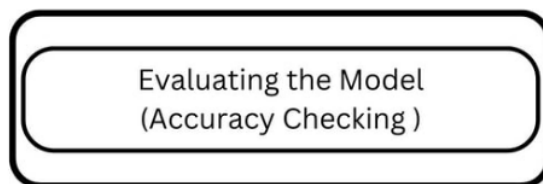


Fig. 3.14. Evaluation

After training on a dataset, a machine learning model is deployed to make predictions on new, unseen data. The model utilizes the patterns and relationships it learned during training to make predictions on this new data. During this inference or prediction phase, the algorithm applies the learned parameters and rules to the incoming data, generating predictions or classifications.

The performance of the model on new data depends on its ⁹⁷ability to generalize from the training data to unseen instances. A well-trained model should exhibit robustness and accuracy when presented with new data that shares similar characteristics to the training dataset.

⁴⁶However, it's important to monitor the model's performance over time, as its effectiveness may degrade if the underlying data distribution changes or if the model encounters data that diverges significantly from its training set. Continuous monitoring, retraining, and adaptation are essential to ¹³ensure that the model remains effective and reliable in real-world applications.

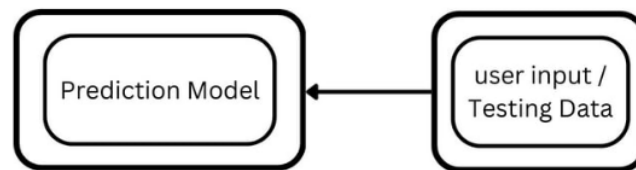


Fig. 3.15. Testing

After the model predicts outcomes based on new data, it generates results that represent generic products or predictions for various scenarios. These predictions provide insights into potential outcomes, trends, or patterns based on the input data and the model's learned parameters. The generic product represents a typical or generalized prediction derived from the model's understanding of the data. These predictions can inform decision-making processes, resource allocation, and strategic planning in diverse domains such as marketing, finance, healthcare, and more. By leveraging predictive insights, organizations can optimize operations, mitigate risks, and capitalize on opportunities in their respective industries.

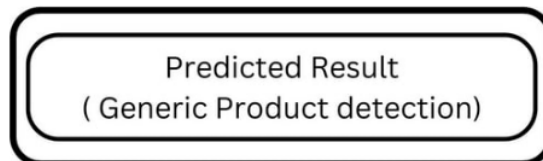


Fig. 3.16. Result Prediction

After processing new data, the model utilizes its learned patterns to detect generic product instances. These instances represent typical examples based on the model's understanding of the input data. Following this, the model proceeds to recommend products by sorting the predicted product data based on the top 5 ratings. This recommendation process leverages the model's predictive capabilities to identify products likely to perform well according to predefined criteria.

Sorting the predicted product data based on ratings ensures that the most promising products are presented first, streamlining decision-making processes for users or stakeholders. By prioritizing products with high ratings, the recommendation system guides users towards items that align closely with their preferences, needs, or business objectives.

This approach not only enhances user experience by providing tailored recommendations but also maximizes the likelihood of user satisfaction and engagement. Furthermore, it optimizes resource allocation and revenue generation by promoting products with the highest potential for success.

In essence, leveraging predictive modeling and recommendation systems enables organizations to capitalize on data-driven insights, enhance customer satisfaction, and drive business growth in competitive markets.

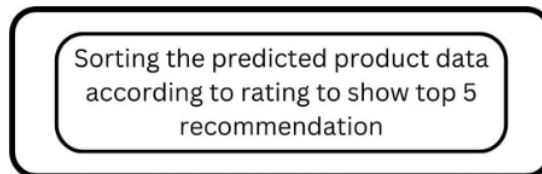


Fig. 3.17. Sorting predicted data

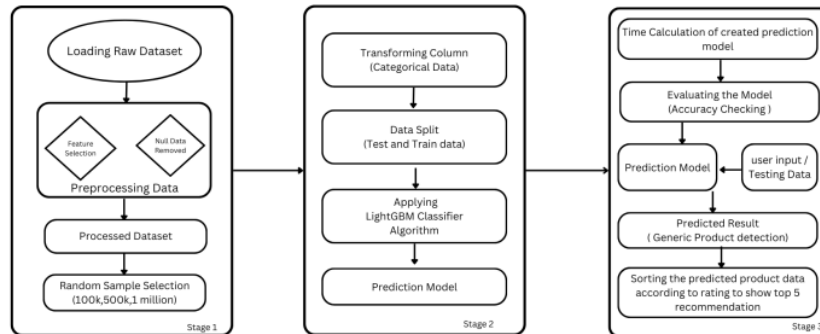


Fig. 3.18. Complete Architecture of the working

In this comprehensive study, we meticulously curated two expansive datasets from Amazon, spanning distinct periods - one from 2019 and another post-2020, chosen for their direct relevance to our research focus. The preprocessing phase was a critical step, involving meticulous feature selection and the removal of null data, culminating in a robust dataset comprising approximately 14 million entries. To facilitate effective algorithm testing, random samples were extracted, comprising 100k, 500k, and 1 million records, providing a nuanced understanding of the model's performance under varied data sizes.

The subsequent stage involved the intricate handling of categorical data, seamlessly transitioning into the division of the dataset into training and test sets. In the initial phase of the third stage, the time taken for model creation was meticulously calculated, shedding light on the algorithm's efficiency. The evaluation process was equally rigorous, focusing on testing the model's accuracy using dedicated test data. As a culmination of our efforts, the algorithm demonstrated its prowess by leveraging newly acquired data to generate recommendations based on learned patterns. This multifaceted approach not only highlights the robustness of our methodology but also underscores the algorithm's practical utility in providing accurate and insightful recommendations in real-world E-commerce scenarios.

3.4 Mathematical Expression:

3.4.1 Gradient-based One Side Sampling Technique for LightGBM:

The GOSS (Gradient-based One-Side Sampling) technique enhances gradient boosting by efficiently handling large datasets during each iteration.

Input:

```

l: training data
d: number of iterations
a: sampling ratio of large gradient data
b: sampling ratio of small gradient data
loss: loss function
L: weak learner
Models <- {} # a list of weak models
fact <- (1-a)/b
topN <- a * len(l) # number of top samples to be included
randN <- b * len(l) # number of random samples to be included

for i = 1 to d do:
  preds <- Models.predict(l)
  g <- loss(l, preds)
  w <- {1, 1, ...} # initialize sample weights
  sorted <- GetSortedIndices(abs(g))
  topSet <- sorted[1:topN]
  randSet <- RandomPick(sorted[topN:len(l)], randN)
  usedSet <- topSet + randSet # combine the top and random samples
  w[randSet] <- w[randSet] * fact # assign weight to the small gradient data
  newModel <- L(l[usedSet], g[usedSet], w[usedSet]) # train a new model on the used samples
  Models.append(newModel) # add the new model to the model list

```

Fig. 3.19. Algorithm for GOSS

In GOSS, the training set is initially sorted based on the absolute values of the negative gradients relative to the loss function. The top- $a \times 100$ percent instances with the highest gradients form subset A, while the remaining instances $(1 - a) \times 100$ percent form subset Ac. For Ac, a random subset B is sampled, sized $b \times |Ac|$.

Instances are then divided based on estimated variance gain at vector $V_j(d)$ over subset A B. This approach ensures that instances with the most significant gradients contribute more to the model's learning process, while still allowing contributions from less influential instances through random sampling. By prioritizing high-gradient instances and efficiently sampling the remaining data, GOSS optimizes the training process, improving the gradient boosting algorithm's performance on large datasets.

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i)^2}{n_l^j(d)} + \frac{(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i)^2}{n_r^j(d)} \right)$$

where the coefficient $(1-a)/b$ is used to normalize the sum of the gradients over B back to the size of A_c

$$\begin{cases} A_l = \{x_i \in A : x_{ij} \leq d\} \\ A_r = \{x_i \in A : x_{ij} > d\} \\ B_l = \{x_i \in B : x_{ij} \leq d\} \\ B_r = \{x_i \in B : x_{ij} > d\} \\ n_l^j(d) = \sum \mathbb{I}(x_i \in (A_l \cup B_l)) \\ n_r^j(d) = \sum \mathbb{I}(x_i \in (A_r \cup B_r)) \end{cases}$$

3.4.2 Exclusive Feature Bundling Technique for LightGBM

Exclusive Feature Bundles (EFBs) can be created in high-dimensional data because sparsity—the occurrence of several features infrequently occurring together—is a prevalent phenomena. Significantly lowering the complexity of the feature space, these bundles aggregate mutually incompatible features. The intricacy of creating a histogram thus changes from being proportional to the number of data points multiplied by the number of original features ($O(\text{data} \times \text{feature})$) to being proportional to the number of data points multiplied by the number of bundles ($O(\text{data} \times \text{bundle})$), where the bundle is significantly smaller than the number of original features. Without sacrificing accuracy, this reduction in complexity speeds up the training framework. The framework is an effective method for handling high-dimensional data sets because it has special characteristics that minimize processing while effectively capturing important information.

Input:

numData: the number of data points in the dataset

F: a bundle of exclusive features

Output:

newBin: a new feature vector obtained from bundling the input features in F

binRanges: a list of bin ranges used to map the original feature values to the new feature values

Algorithm:

Initialize binRanges as [0], and totalBin as 0.

For each feature f in F , add $f.\text{numBin}$ to totalBin and append the result to binRanges.

Create a new empty feature vector newBin with numData elements.

For each data point i in the dataset:

a. Initialize newBin[i] to 0.

b. For each feature j in F :

i. If $F[j].\text{bin}[i]$ is not equal to 0, add $F[j].\text{bin}[i]$ and binRanges[j] to newBin[i].

Return newBin and binRanges as the output.

Fig. 3.20. Algorithm for Exclusive Feature Bundling Technique

92

Chapter 4

Data Analysis

4.1 Data-Collection

Dataset-1

32

we use the E-Commerce behavior data from multi category store from REES46 Marketing Platform as our dataset. We define our own goal and filter the dataset accordingly.

- Data duration: October 2019 – February 2020

Dataset-1 link:

[27]<https://www.kaggle.com/dschettler8845/recsys-2020-ecommerce-dataset>

Dataset-2

14

This file contains behavior data for a one month (October 2019) from a large multi-category online store, collected by Open CDP project. Each row in the file represents an event. All events are related to products and users. There are different types of events. Reason to choose both of these datasets:

These datasets contain all the necessary data as columns we are seeking for our research. Besides this dataset is versatile with a lot of unique product and category type data which will be best for our algorithm (LightGBM). Again large dataset is also part of our goal as online shopping is gaining huge demand as days go by. These datasets provide millions of data thus getting larger datasets to work fit perfectly with our goal.

Dataset -2 link:

67

[28]<https://www.kaggle.com/code/danofer/ecommerce-store-predict-purchases-data-prep/output>

4.2 Data-Attributes:

Dataset-1:

The following features are present in our Raw dataset.

event-time, event-type, product-id, brand, price, user-id, user-session, target, cat-0, cat-1, cat-2, cat-3, timestamp, ts-hour, ts-minute, ts-weekday, ts-day, ts-month, ts-year
unique value in each column:

column	unique value
event_time	5909789
event_type	2
product_id	164453
brand	4638
price	71591
user_id	2547058
user_session	7279439
target	2
cat_0	14
cat_1	61
cat_2	91
cat_3	2
timestamp	5909789
ts_hour	24
ts_minute	60
ts_weekday	7
ts_day	31
ts_month	5
ts_year	2

TABLE 4.1: Attributes of Dataset.1

With a vast repository of over 11 million product data points spanning five months, our dataset from a multi-category online store provides a robust foundation for comprehensive testing of diverse machine learning techniques, benchmarked against the formidable LightGBM approach. The dataset's magnitude is instrumental, enabling a thorough exploration of methodologies and facilitating meaningful comparisons. Particularly crucial are specific columns like "event type," "brand," "cat 0," "cat 1," and "cat 2," essential for handling categorical data and forming the backbone of our research. The abundance of data not only empowers rigorous analysis but also allows for efficient data curation by swiftly eliminating irrelevant or blank entries. This large-scale dataset becomes a valuable asset, ensuring a holistic examination of machine learning models, with a focus on LightGBM, to derive meaningful insights and optimize recommendations in the dynamic landscape of the multi-category online store.

Dataset-2:

The following features are present in our Raw dataset.

event_type, product-id, category-id, category-code, brand, price, user-id, category-level-1, category-level-2, category-level-3 unique value in each column:

column	unique value
event_time	2621538
event_type	3
product_id	166794
category_id	624
category_code	126
brand	3444
price	65298
user_id	3022290
user_session	9244422
category_code_level1	13
user_category	7763898

TABLE 4.2: Attributes of Dataset.2

This dataset has a month of data from a multi-category online store with more than 42 million data from different products. We selected this dataset as it has all of those similar attributes to dataset 1. There are also some specific columns such as 'event_type', 'brand', and 'category_code' which are crucial for categorical data as these will play significant roles in our research. Further, we can make new attributes from the column 'category_code' which will make it similar to our dataset 1.

Moreover, a huge dataset will provide us with enough data even after data preprocessing

4.3 Dataset-Properties

Dataset-1:

Range Index: 11495242²¹ entries, 0 to 11495241

Data columns (total 19 columns):

Index	Column	Data Type
0	event_time	object
1	event_type	object
2	product_id	object
3	brand	object
4	price	object
5	user_id	object
6	user_session	object
7	target	int64
8	cat_0	object
9	cat_1	object
10	cat_2	object
11	cat_3	object
12	timestamp	datetime64[ns]
13	ts_hour	int16
14	ts_minute	int16
15	ts_weekday	int16
16	ts_day	int16
17	ts_month	int16
18	ts_year	int16

TABLE 4.3: Properties of Dataset

Dataset -2:

Range index: 42448764 entries, 0 to 42448763

Data columns (total 11 columns):

Index	Column	Data Type
0	event_time	object
1	event_type	object
2	product_id	int64
3	category_id	int64
4	category_code	object
5	brand	object
6	price	float64
7	user_id	int64
8	user_session	int64
9	category_code_level1	object
10	user_category	int64

TABLE 4.4: Properties of Dataset

4.4 Representation-of-Scanned-Raw-Dataset:

Dataset-1:

index	event_time	event_type	product_id	brand	price	user_id	user_session	target	cat_0
0	2019-11-01 00:00:14 UTC	cart	1005014	samsung	503.09	533326659	6b928be2-2bce-4640-8296-0efdf2da22a	0	electronics
1	2019-11-01 00:03:39 UTC	cart	1005115	apple	949.47	565865924	fd4bd6d4-bd14-4fdc-9aff-bd41a59482e	0	electronics
2	2019-11-01 00:05:54 UTC	cart	1002542	apple	486.8	549256216	dcdbdc6e4-cd49-4ee8-95c5-e85f3c618fa	0	electronics
3	2019-11-01 00:07:22 UTC	cart	1002542	apple	486.8	549256216	dcdbdc6e4-cd49-4ee8-95c5-e85f3c618fa	0	electronics
4	2019-11-01 00:10:45 UTC	cart	4804056	apple	160.57	522355747	0a1f37d1-71b7-4645-a8a7-ab91bc198a5	0	electronics
5	2019-11-01 00:13:40 UTC	cart	1004873	samsung	362.29	563558500	e0729b6c-eafe-4b0f-9d66-6ee777d0849	0	electronics
6	2019-11-01 00:14:09 UTC	cart	1004856	samsung	128.42	513645631	8cccf28f-136d-4941-9005-755bec0a44c	0	electronics
7	2019-11-01 00:14:23 UTC	cart	1004856	samsung	128.42	557839915	40304779-5050-440d-8325-cab689f355a	0	electronics
8	2019-11-01 00:15:23 UTC	cart	1005115	apple	949.47	561713177	821f797d-f9f4-4108-a44b-dbcba4e9625	0	electronics
9	2019-11-01 00:15:46 UTC	cart	1005116	apple	1013.86	513852314	892943dc-aacd-44d8-94ef-b2bcbcd662a	0	electronics
10	2019-11-01 00:15:56 UTC	cart	1005116	apple	1013.86	513852314	892943dc-aacd-44d8-94ef-b2bcbcd662a	0	electronics
11	2019-11-01 00:20:25 UTC	cart	1004870	samsung	275.25	563626542	e6e2cb92-9103-4e6b-a68e-e2103766e0a	0	electronics
12	2019-11-01 00:20:49 UTC	cart	1004870	samsung	275.25	563626542	e6e2cb92-9103-4e6b-a68e-e2103766e0a	0	electronics
13	2019-11-01 00:24:48 UTC	cart	4804295	xiaomi	22.8	550508533	ff30096-d345-43a4-8b01-329283095f38	0	electronics
14	2019-11-01 00:25:12 UTC	cart	4804295	xiaomi	22.8	550508533	ff30096-d345-43a4-8b01-329283095f38	0	electronics
15	2019-11-01 00:30:05 UTC	cart	1002633	apple	358.31	557994805	b3bbf6e2-f995-4482-aceb-2a5695406d9	0	electronics
16	2019-11-01 00:30:35 UTC	cart	1004856	samsung	128.42	512800858	ddb1ab7a-7667-4f79-a5dc-b5ed461e6ee	0	electronics
17	2019-11-01 00:31:35 UTC	cart	1801881	samsung	488.8	566283686	771370f8-6f8e-4584-81b9-9b5a8cd8e67	0	electronics
18	2019-11-01 00:33:37 UTC	cart	1201471	samsung	469.84	547694248	ad31f016-29a7-4d6a-9de3-d042016192a	0	electronics
19	2019-11-01 00:35:55 UTC	cart	1004767	samsung	242.63	532542999	595c35b6-269b-433d-a7cf-47158757ac5	0	electronics
20	2019-11-01 00:36:06 UTC	cart	1005031	xiaomi	181.12	550179534	831496c7-98d3-457f-8386-661a9206fb4	0	electronics
21	2019-11-01 00:36:18 UTC	cart	1005160	xiaomi	212.08	512861616	78e5556e-d5e0-4da1-8cf8-820383496f3	0	electronics
22	2019-11-01 00:37:30 UTC	cart	12711761	NA	28.83	519011828	7ccaec2a-dabe-4f9f-bca9-4e50abf2e18	0	NA
23	2019-11-01 00:38:11 UTC	cart	1004741	xiaomi	189.97	515238387	5c237d8f-c199-4091-b83c-967ef8027c7	0	electronics
24	2019-11-01 00:39:24 UTC	cart	4100151	sony	382.94	521793571	4daa12d3-bc19-4f01-8082-6e69606a639	0	NA

Fig. 4.1. Digitized Raw Dataset_1 Samples

cat_1	cat_2	cat_3	timestamp	ts_hour	ts_minute	ts_weekday	ts_day	ts_month	ts_year
smartphone	NA	NA	2019-11-01 00:00:14	0	0	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:03:39	0	3	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:05:54	0	5	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:07:22	0	7	4	1	11	2019
audio	headphone	NA	2019-11-01 00:10:45	0	10	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:13:40	0	13	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:14:09	0	14	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:14:23	0	14	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:15:23	0	15	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:15:46	0	15	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:15:56	0	15	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:20:25	0	20	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:20:49	0	20	4	1	11	2019
audio	headphone	NA	2019-11-01 00:24:48	0	24	4	1	11	2019
audio	headphone	NA	2019-11-01 00:25:12	0	25	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:30:05	0	30	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:30:35	0	30	4	1	11	2019
video	tv	NA	2019-11-01 00:31:35	0	31	4	1	11	2019
tablet	NA	NA	2019-11-01 00:33:37	0	33	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:35:55	0	35	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:36:06	0	36	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:36:18	0	36	4	1	11	2019
NA	NA	NA	2019-11-01 00:37:30	0	37	4	1	11	2019
smartphone	NA	NA	2019-11-01 00:38:11	0	38	4	1	11	2019
NA	NA	NA	2019-11-01 00:39:24	0	39	4	1	11	2019

Fig. 4.2. Continuation of 4.1

Dataset-2:

index	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session	category_code_level1	user_category
0	2019-10-01 00:00:00	view	12771	417	NaN	shiseido	35.79	8488	7406	NaN	4395863
1	2019-10-01 00:00:00	view	12158	3	appliances.environment.water_heater	aqua	33.2	11658	9465	appliances	5810260
2	2019-10-01 00:00:01	view	4977	167	furniture.living_room.sofa	NaN	543.1	4771	5560	furniture	2444320
3	2019-10-01 00:00:01	view	3010	153	computers.notebook	lenovo	251.74	10343	8014	computers	5310585
4	2019-10-01 00:00:04	view	248	84	electronics.smartphone	apple	1081.98	7662	12763	electronics	3952141
5	2019-10-01 00:00:05	view	4049	197	computers.desktop	pulser	908.62	702	835	computers	303397
6	2019-10-01 00:00:08	view	5064	39	NaN	creed	380.96	12234	5157	NaN	5893543
7	2019-10-01 00:00:08	view	10820	126	NaN	luminarc	41.16	10552	6372	NaN	5400657
8	2019-10-01 00:00:10	view	10267	287	apparel.shoes.keds	baden	102.71	5202	11085	apparel	2680854
9	2019-10-01 00:00:11	view	346	84	electronics.smartphone	huawei	566.01	7966	4133	electronics	4117011
10	2019-10-01 00:00:11	view	10376	64	appliances.kitchen.microwave	elenberg	51.46	11908	11696	appliances	5859435
11	2019-10-01 00:00:11	view	560	84	electronics.smartphone	samsung	900.64	6848	5202	electronics	3550446
12	2019-10-01 00:00:13	view	12147	3	appliances.environment.water_heater	haier	102.38	12220	9812	appliances	5893437
13	2019-10-01 00:00:15	view	12771	417	NaN	shiseido	35.79	8488	7406	NaN	4395863
14	2019-10-01 00:00:16	view	3612	105	furniture.bedroom.bed	brw	93.18	12225	8159	furniture	5893478
15	2019-10-01 00:00:17	view	7077	212	NaN	NaN	357.79	1741	1515	NaN	816362
16	2019-10-01 00:00:18	view	5680	52	electronics.video.tv	haier	193.03	7854	14629	electronics	4059841
17	2019-10-01 00:00:18	view	937	69	appliances.kitchen.mixer	bosch	58.95	4945	9233	appliances	2548963
18	2019-10-01 00:00:19	view	2921	153	computers.notebook	hp	580.89	10343	8014	computers	5310585
19	2019-10-01 00:00:19	view	646	84	electronics.smartphone	apple	1747.79	7662	12763	electronics	3952141
20	2019-10-01 00:00:20	view	56	84	electronics.smartphone	apple	588.77	12226	7126	electronics	5893496
21	2019-10-01 00:00:20	view	13311	60	electronics.audio.headphone	jbl	33.21	12216	8864	electronics	5892935
22	2019-10-01 00:00:22	view	4069	197	computers.desktop	pulser	921.49	702	835	computers	303397
23	2019-10-01 00:00:23	view	406	84	electronics.smartphone	xiaomi	197.55	4950	9802	electronics	2549307
24	2019-10-01 00:00:23	view	14500	2	appliances.environment.air_heater	midea	47.62	8064	8080	appliances	4175054

Fig. 4.3. Digitized Raw Dataset.2 Samples

4.5 Data-Pre-Processing:**Dataset-1:****step 1:**

removing unnecessary columns 'ts-weekday', 'target', 'user-session', 'event-time', 'cat-3', 'timestamp', 'ts-hour', 'ts-minute', 'ts-day', 'ts-month', 'ts-year'

step 2:

removed 'NA' from dataset according to columns 'cat-2' and 'brand'

step 3:

removing some specific product form the column "cat-2" ('sound-card', 'soldering', 'bench', 'axe', 'dictaphone', 'fryer', 'camera', 'parktronic', 'shelving', 'screw', 'steam-cooker', 'toilet', 'pillow', 'anti-freeze', 'fan') because these only occurs less the 1000 time in the whole data set as these might not trained properly.

Step 4:

removed all duplicated rows from the whole dataset.

df.isna().sum()	
event_type	0
product_id	0
brand	0
price	0
user_id	0
cat_0	0
cat_1	0
cat_2	0
dtype: int64	

Fig. 4.4. Null Data Count For Each Column

Step 5:

transforming columns which are not integer or float type into category type. Such columns are “event-type”, “brand”, “cat-0”, “cat-1”, “cat-2”

Dataset-1 after prepossessing:

Int64Index: 3988623 entries, 0 to 3988622

Data columns (total 8 columns):

#	Column	Non-Null Count	Data Type
0	event_type	3988623 non-null	category
1	product_id	3988623 non-null	int64
2	brand	3988623 non-null	category
3	price	3988623 non-null	float64
4	user_id	3988623 non-null	int64
5	cat_0	3988623 non-null	category
6	cat_1	3988623 non-null	category
7	cat_2	3988623 non-null	category

TABLE 4.5: Preprocessed Dataset.1

Dataset-2:**step 1:**

removed unnecessary columns event-time, category-id, user-session, user-category, category-code-level1

step 2:

Making 3 new columns from “category-code” which are “category-level-1”, “Category-level-2”, “Category-level-3”

step 3:

removed “category-code” columns as well

step 4:

removed some specific product form the column “category-level-3” (hammock, step-ins, anti-freeze, bench, soldering, slippers, camera, ballet-shoes, fan, padrilles, sandals, winch, bath)

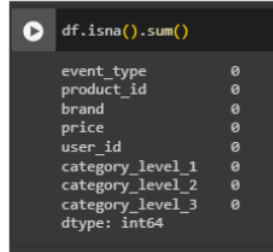
because these only occurs less the 1000 time in the whole data set as these might not trained properly.

Step 5:

removed all the duplicate data row wise.

Step 6:

removing all the Null value row wise according to columns “category-level-3” as this columns is our target for multi-class classification



```
df.isna().sum()
event_type      0
product_id      0
brand           0
price           0
user_id         0
category_level_1 0
category_level_2 0
category_level_3 0
dtype: int64
```

Fig. 4.5. Null Data Count For Each Column

Step 7:

transforming columns which are not integer or float type into category type. Such columns are “event-type”, “brand”, “category-level-1”, “category-level-2”, “category-level-3”

Dataset 2 after preprocessing:

Int64Index: 1297178 entries, 0 to 1297177

Data columns (total 8 columns):

#	Column	Non-Null Count	Data Type
0	event_type	1297178 non-null	category
1	product_id	1297178 non-null	int64
2	brand	1297178 non-null	category
3	price	1297178 non-null	float64
4	user_id	1297178 non-null	int64
5	category_level_1	1297178 non-null	category
6	category_level_2	1297178 non-null	category
7	category_level_3	1297178 non-null	category

TABLE 4.6: Preprocessed Dataset_2

4.6 Encoding Categorical Variables:

Categorical variables must be transformed into numerical representations since machine learning models typically require numerical inputs. By ensuring that the model's criteria and the data are compatible, this technique makes it possible to use categorical information effectively in the learning process.

Label Encoding:

We gave distinct integer labels to each category in ordinal categorical data, taking into account the intrinsic order of the categories. To prevent introducing erroneous associations between categories, we were careful not to use this technique to nominal categorical data. The label encoding was only done for XGBoost, Random Forest and SVM.

4.7 Data Standardization

Standardization: Alternatively, we transformed the data into a standardized form with a mean of 0 and a standard deviation of 1. This standardization method works well for algorithms that are sensitive to different feature scales, especially gradient-based algorithms. By guaranteeing a uniform scale throughout the features, it promotes better learning outcomes and convergence. This was only done for SVM due to its incapability.

4.8 Tools

1. Python 3.10
2. Google Colaboratory
3. Google sheet

Chapter 5

Results and Discussions

5.1 Dataset-1

5.1.1 For 100k Data :-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	100000	0.9487333333333333	62.23791313171387
XGBoost		0.9389333333333333	875.308699131012
Random Forest		0.8443333333333333	15.697554111480713
SVM		0.5244666666666666	186.7897274494171

TABLE 5.1: Results For 100k

5.1.2 For 500k Data :-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	500000	0.97498	333.89465069770813
XGBoost		0.97402	3178.82821559906
Random Forest		0.8212533333333333	73.31936502456665
SVM		Incomprehensible	

TABLE 5.2: Results For 500k

5.1.3 For 1 million Data:-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	1000000	0.9794766666666667	648.1630439758301
XGBoost		Incomprehensible	
Random Forest		0.8211766666666667	189.8987774848938
SVM		Incomprehensible	

TABLE 5.3: Results For 1 million

In our rigorous analysis of Dataset-1, we conducted a meticulous examination using subsets containing 100k, 500k, and 1 million data points. The purpose was to comprehensively evaluate the performance of various algorithms across varying dataset sizes, aiming to discern their adaptability and scalability in handling different volumes of data. Notably, LightGBM emerged as the clear frontrunner, consistently surpassing competitors such as XGBoost, SVM, and Random Forest.

The key observation from our analysis was the remarkable and consistent superiority of LightGBM across diverse dataset sizes. Regardless of whether the dataset contained 100k points or extended to a million data points, LightGBM demonstrated robust performance, showcasing its versatility and scalability. This demonstrated its adeptness at handling varying data volumes with ease, reaffirming its suitability for recommendation systems operating in contexts where dataset sizes can vary significantly.

The findings underscored LightGBM's adaptability to different scales of data, positioning it as a reliable and effective choice for recommendation systems in dynamic environments. Its ability to consistently deliver superior outcomes across a spectrum of dataset sizes enhances its appeal for diverse applications within the realm of recommendation systems.

Furthermore, LightGBM's versatility is a crucial attribute in real-world scenarios where datasets may evolve over time. Its consistent efficacy across varying dataset sizes ensures that it remains a dependable solution, capable of adapting to changing data dynamics and evolving requirements.

Overall, our analysis highlights LightGBM's prowess in handling datasets of different scales, reaffirming its status as a preferred choice for recommendation systems. Its consistent performance, adaptability, and scalability make it an invaluable asset in scenarios where dataset sizes can vary significantly, underscoring its significance in the ever-evolving landscape of recommendation systems.

5.2 Dataset-2

5.2.1 For 100k Data :-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	100000	0.9317333333333333	60.70483613014221
XGBoost		0.9179333333333334	802.492728471756
Random Forest		0.7442333333333333	25.996510982513428
SVM		0.5244666666666666	186.7897274494171

TABLE 5.4: Results For 100k

5.2.2 For 500k Data:-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	500000	0.9737933333333333	303.00841069221497
XGBoost		0.9712	3765.232586145401
Random Forest		0.7333266666666667	123.52909135818481
SVM		Incomprehensible	

TABLE 5.5: Results For 500k

5.2.3 For Data Count Of 1 million :-

Algorithm	Data Count	Accuracy	Time(s)
LightGBM	1000000	0.98085	600.7675633430481
XGBoost		Incomprehensible	
Random Forest		0.73717	227.8387496471405
SVM		Incomprehensible	

TABLE 5.6: For 1 million

In our concurrent evaluation of an additional dataset, we adopted a balanced approach by giving equal weight to both time and performance parameters. The aim was to comprehensively assess the effectiveness of different methods in handling various datasets. In this evaluation, LightGBM emerged as a standout performer, outperforming competitors such as XGBoost, Random Forest, and SVM.

LightGBM stands out in the realm of machine learning algorithms due to its remarkable balance between speed and efficiency. Unlike many other algorithms, LightGBM not only achieves superior performance metrics but also demonstrates agility in deriving meaningful insights from the data it processes. Its ability to swiftly analyze vast amounts of information and generate accurate predictions is a testament to its dominance in the field.

Throughout our evaluation, LightGBM consistently proved its effectiveness in handling a diverse range of datasets, showcasing its versatility and adaptability across various scenarios. Its capacity to derive actionable insights quickly and consistently deliver exceptional performance solidified its position as the top-performing algorithm in our analysis.

In recommendation systems, where responsiveness and accuracy are paramount, LightGBM truly shines. Its efficiency in processing large volumes of data and its ability to provide timely and relevant recommendations make it an invaluable asset in the domain. Whether dealing with real-time user interactions or analyzing historical data, LightGBM's capabilities remain unmatched.

Our careful analysis underscores LightGBM's exceptional performance and efficiency across a spectrum of datasets in the recommendation systems domain. Its dominance over other algorithms in terms of both speed and effectiveness reinforces its significance as the preferred choice for data-driven decision-making processes. LightGBM's reputation as a leading algorithm in the field of machine learning is further solidified by its ability to consistently outperform its counterparts and deliver actionable insights that drive business success.

In conclusion, LightGBM's remarkable balance of speed, efficiency, and effectiveness positions it as a formidable tool for data scientists and practitioners across various industries. Its ability to handle complex data scenarios and provide accurate predictions makes it a cornerstone in the development of cutting-edge recommendation systems and underscores its pivotal role in advancing the field of machine learning.

5.3 Discussion :-

In our comprehensive evaluation of various algorithms, including XGBoost, SVM, and Random Forest, LightGBM emerged as the standout performer, showcasing unparalleled advantages in terms of both time efficiency and performance. The results unequivocally demonstrated that LightGBM excelled in striking the perfect balance between these two critical factors, positioning it as the superior choice among the considered algorithms.

One key metric where LightGBM outshone its counterparts was time efficiency. The algorithm's leaf-wise tree growth strategy, coupled with its gradient-based optimization approach, allowed it to process data swiftly, significantly reducing the time required for model building. This is particularly crucial in the fast-paced landscape of e-commerce, where real-time responsiveness is paramount.

Simultaneously, LightGBM demonstrated exceptional performance, surpassing the benchmarks set by XGBoost, SVM, and Random Forest. ⁹³ Its ability to handle high-dimensional data and discern intricate patterns enabled it to generate more accurate and relevant recommendations. This not only enhances the user experience by providing timely and personalized suggestions but also contributes to the overall effectiveness of the recommendation system in driving user engagement and satisfaction.

The synergy of time efficiency and superior performance positions LightGBM as the algorithm of choice for recommendation systems. Its prowess in navigating the trade-off between speed and effectiveness not only optimizes computational resources but also ensures that users receive highly tailored recommendations in a timely manner, thereby reinforcing its status as a beacon of excellence in the landscape of recommendation algorithms.

5.4 Feature Importance:

In this subsection, the importance of different features is highlighted based on their respective amounts. Notably, event_type demonstrated the highest significance with a value of 28963. Following closely, brand maintained its importance with an amount of 10340. Subsequently, product_id held a significance of 7709, while price followed with 5293. Moving forward, category_level_2 and category_level_1 showed importance with values of 4879 and 1050, respectively.

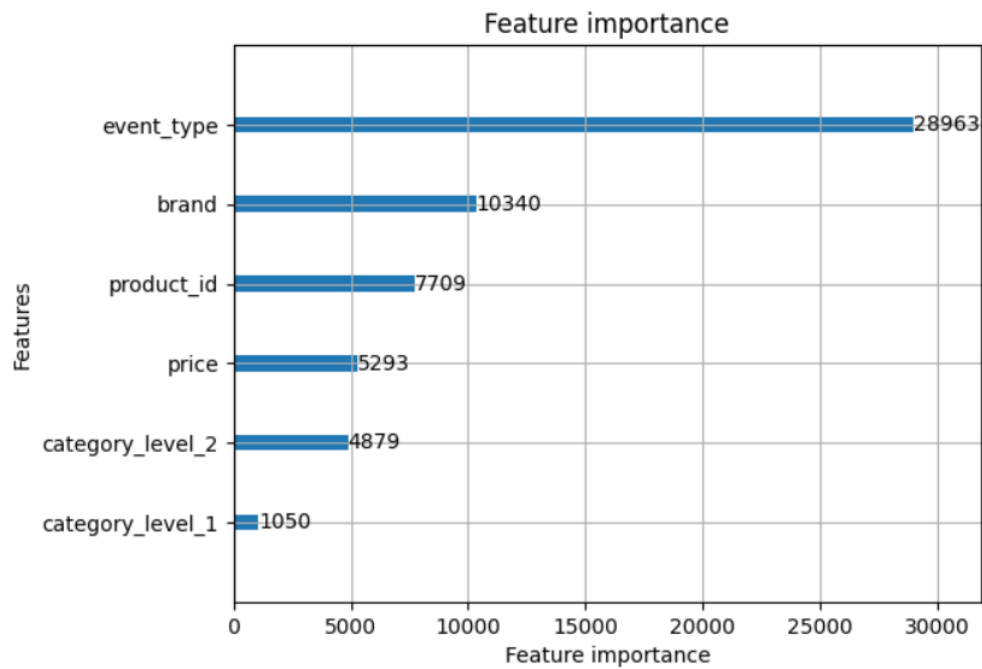


Fig. 5.1. Feature Importance

5.5 Error Rate Visualization:

The validated results in the multi_logloss showed a little discrepancy between the tested and validated results in the first dataset analysis. This convergence demonstrates the stability and dependability of the used approach, which includes LightGBM. The correctness and efficacy of the algorithm in providing dependable recommendations is highlighted by the consistency observed between the outcomes of validation and testing. A low variation of this kind indicates that the algorithms are able to perform well in all parts of the evaluation, which gives confidence about their dependability and applicability in the field of recommendation systems for a variety of datasets.

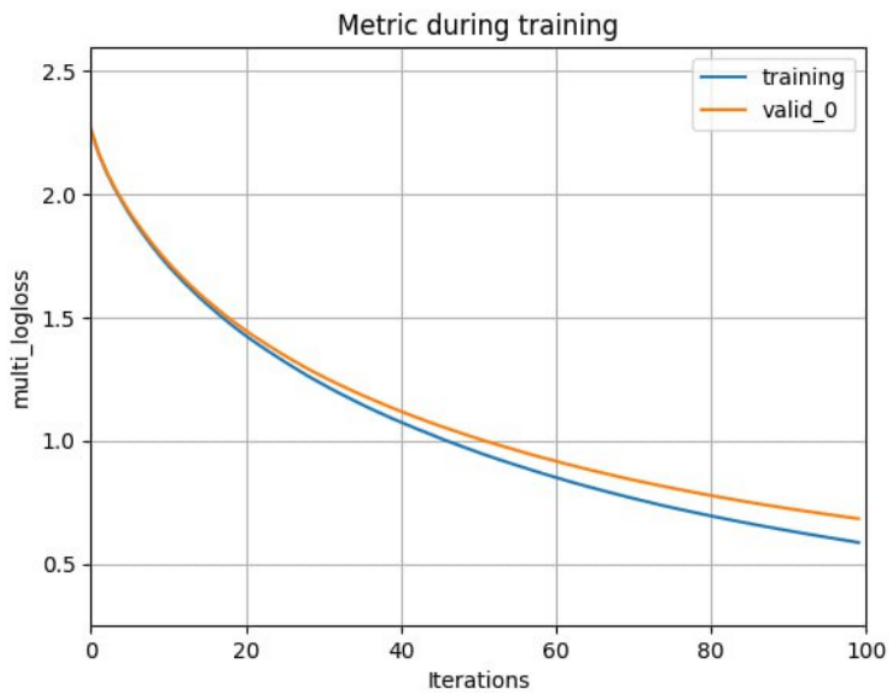


Fig. 5.2. Multi Logloss of dataset.1

Compared to the initial dataset, the subsequent dataset demonstrated a significant decrease in the margin of error between the results of tests and validations in the multi.logloss. LightGBM's outstanding performance and time efficiency are firmly reaffirmed by this significant convergence. The reduced difference highlights the algorithm's increased accuracy in making recommendations. A decrease in margin like this validates LightGBM's dependability on a variety of datasets and highlights its practical application. By demonstrating LightGBM's constant and improved performance in terms of accuracy and time efficiency for a variety of datasets, this result strengthens trust in LightGBM as a recommendation system solution.

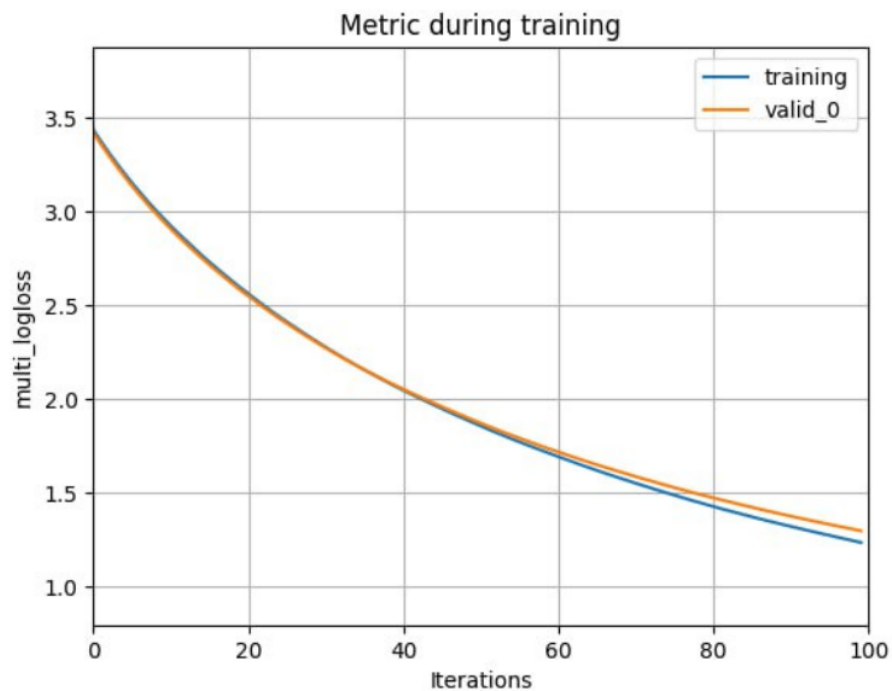


Fig. 5.3. Multi Logloss of dataset.2

5.6 Recommendation-Checking:

Dataset-1

Input:

6	event_type	product_id	brand	price	user_id	cat_0	cat_1
	cart	1801975	lg	963.01	549156314	electronics	video

TABLE 5.7: Input

Output:

Predictions: [tv]

Actual output "tv"

Recommendation :

eventtype	productid	brand	price	userid	category1	category2	rating
cart	100132465	kenwood	79.8	560729132	electronics	video	4.999396
cart	100124717	swat	30.89	515661766	electronics	video	4.999005
purchase	100124671	mydean	128.7	562840708	electronics	video	4.999003
cart	100119266	swat	69.5	568806204	electronics	video	4.998730
cart	100119255	swat	126.13	570810868	electronics	video	4.998730

TABLE 5.8: Recommendation For Dataset_1

In Dataset 1, the validation mechanism ⁵⁸ plays a pivotal role in evaluating the accuracy ¹⁶ and reliability of predictions made by the model. The input section of the dataset serves as the basis for testing the model's predictive capabilities, while the output section reveals the predicted outcome, which is compared against the expected validation result. In this particular instance, both the predicted and validated outcomes classify the input as "TV," indicating a harmonious alignment between the model's prediction and the expected classification.

This confirmation holds significant importance as it serves as a critical checkpoint ¹⁶ in assessing the model's performance and its ability to accurately categorize input data. The agreement between the predicted result and the validated outcome underscores the model's reliability and effectiveness in recognizing and classifying the specified target – in this case, the classification of the input as "TV."

The validation step not only validates the model's efficacy but also instills confidence in its predictive abilities, reinforcing its capability to yield meaningful and reliable predictions in real-world scenarios. By demonstrating consistency between the predicted and validated outcomes, the model establishes its credibility and enhances its utility in practical applications.

Moreover, the validation process provides ¹⁶insights into the model's generalization ability and its capacity to perform accurately across different datasets and scenarios. This robust validation framework ensures that the model remains dependable and resilient, even when confronted with varying data inputs and complex classification tasks.

Overall, the alignment between the predicted result and the validated outcome in Dataset 1 serves as a testament to the model's proficiency and underscores its potential to deliver valuable insights and informed decisions in real-world settings.

Dataset-2**Input:**

event_type	product_id	brand	price	user_id	category_level_1	category_level_2
view	11530	midea	99.07	758782	appliances	kitchen

TABLE 5.9: Input

Output:

Predictions: ['light']

Actual output:light

Recommendation :

eventtype	productid	brand	price	userid	categorylevel1	categorylevel2	rating
purchase	162624	comfee	108.09	2937791	appliances	kitchen	4.934849
cart	81723	samsung	373.21	1709656	appliances	kitchen	2.479896
purchase	52318	willmark	143.24	1897089	appliances	kitchen	1.587597
purchase	33474	lg	335.82	445374	appliances	kitchen	1.015773
cart	11621	beko	180.16	355335	appliances	kitchen	0.352641

TABLE 5.10: Recommendation For Dataset_2

In Dataset 2, the utilization of input data for validation purposes ensures the accuracy and reliability of predictions made by the model. The output section of the dataset clearly demonstrates a seamless alignment between the model's prediction and the validated outcome, both of which accurately identify the item as a "light." This congruence signifies the model's exceptional ability to correctly categorize the provided input, instilling a high degree of confidence in its predictive capabilities.

The validation process in Dataset 2 serves as a critical checkpoint in evaluating the model's performance and validating its consistency in delivering precise results across a spectrum of inputs. The successful correspondence between the predicted and validated classifications not only reaffirms the model's reliability but also emphasizes its effectiveness in recognizing and categorizing diverse data inputs.

This validation reinforces the model's practical applicability in scenarios where precise predictions are paramount. The model's adeptness in consistently delivering accurate results underscores its reliability and underscores its potential to inform decision-making processes in real-world settings.

Moreover, the alignment between the predicted and validated outcomes in Dataset 2 highlights the robustness of the model's predictive framework. It demonstrates the model's adaptability to different input scenarios and showcases its capacity to maintain accuracy across varied datasets.

The successful validation process in Dataset 2 not only validates the model's capabilities but also strengthens its credibility as a reliable predictive tool. It serves as a testament to the model's proficiency in delivering meaningful insights and informed decisions, reinforcing its practical utility in diverse domains.

Overall, Dataset 2's validation process underscores the model's reliability, effectiveness, and practical applicability in scenarios where precise predictions are indispensable. It accentuates the model's ability to navigate complexities and deliver valuable outcomes, positioning it as a valuable asset in data-driven decision-making processes.

5.7 Comparison with the previous work:

Z. Shahbazi and Y.-C. Byun had used XGBoost, Random Forest and Support Vector Machine for a recommendation system. Below there is a two table side by side featuring our work and the previous. [21]

31

Algorithm	Accuracy
XGBoost	89.6%
Random Forest	83.24%
SVM	78%

Algorithm	Accuracy
LightGBM	94.95%
XGBoost	93.89%
Random Forest	84.43333333333333%
SVM	52.44666666666666%

TABLE 5.11: Previous Work

TABLE 5.12: Our Result for 100k

The comparison between the results obtained by Z. Shahbazi and Y.-C. Byun and those of our study reveals intriguing differences, particularly in the selection and performance of supervised learning algorithms. While Shahbazi and Byun utilized XGBoost as their primary method and achieved an accuracy of 89.6%, our experimentation with XGBoost yielded a higher accuracy of 93.72%. However, the pivotal moment in our research occurred when LightGBM emerged as the focal point, delivering exceptional performance with an accuracy of 94.95%.

The comparison highlights the efficacy of LightGBM as a superior algorithm for our study. Both sets of results demonstrate a consistent upward trend across SVM, Random Forest, XGBoost, and LightGBM. However, it is LightGBM that stands out by showcasing a significant improvement in outcomes compared to its counterparts.

The superiority of LightGBM can be attributed to several factors. Firstly, LightGBM harnesses gradient boosting techniques, which sequentially enhance weak learners to construct powerful ensemble models. This approach enables LightGBM to effectively capture complex patterns within the data, leading to higher predictive accuracy. Additionally, LightGBM's optimization of memory utilization and parallel computing capabilities contribute to its exceptional performance, especially when handling large datasets.

Furthermore, the success of LightGBM underscores the importance of algorithm selection in machine learning applications. While XGBoost has been widely utilized and recognized for its effectiveness, our findings suggest that LightGBM surpasses it in terms of accuracy and performance. This highlights the need for researchers and practitioners to explore alternative algorithms and evaluate their suitability for specific tasks and datasets.

The steady linear increase in results across the evaluated algorithms reflects a systematic exploration of different methodologies and their impact on predictive performance. It

demonstrates the iterative nature of machine learning research, where experimentation and analysis drive continuous improvement and refinement.

The remarkable accuracy achieved by LightGBM in our study has significant implications for various domains, including e-commerce, finance, healthcare, and more. By leveraging advanced machine learning techniques, organizations can ⁸³extract valuable insights from data, make informed decisions, and drive meaningful outcomes.

In conclusion, the comparison of results between our study and previous research underscores the importance of algorithm selection and highlights the exceptional performance of LightGBM. Moving forward, further research into algorithmic advancements and their application in real-world scenarios will continue to push the boundaries of machine learning and drive innovation in diverse fields.

Chapter 6

Conclusion

6.1 Contribution

There are two main areas in which this thesis contributes significantly.

- Because LightGBM can handle large datasets (many with millions of data points), it is an important tool for modern data analysis. LightGBM shows a strong ability to handle and examine massive datasets with effectiveness by utilizing cutting-edge methods and effective algorithms. This ability makes it possible to glean deeper insights from intricate data structures, giving companies the ability to see overlooked trends, patterns, and correlations that could go unnoticed otherwise. In order to improve decision-making, predictive modeling, and actionable intelligence across a range of fields and sectors, LightGBM facilitates researchers' and practitioners' exploration of rich data landscapes by processing large amounts of data fast. LightGBM is a useful tool for pursuing innovation and knowledge discovery because of its capacity to handle the complexity of large data environments.
- Apart from its skillful manipulation of large datasets, the thesis also tackles the widespread issue of long compilation times that come with handling large amounts of data. The research delivers a notable reduction in compilation time by implementing workflow optimizations and streamlined approaches that are specifically designed for compilation operations. This achievement reduces the possibility of bottlenecks and delays in data processing pipelines by optimizing the handling of huge datasets and improving the efficiency of data analysis procedures. The thesis advances knowledge and methods for handling and drawing conclusions from massive datasets by addressing this crucial component of data analytics. In the end, these improvements highlight the thesis's influence on enhancing data handling capacities and quickening the rate of information extraction in data-driven research.

6.2 Limitation

- Hardware constraints frequently present challenges to achieving optimal results with datasets larger than one million entries. Findings from datasets with up to one million items highlight the need to modify algorithms for platforms with limited resources, which is especially clear when implementing recommendation systems. These datasets provide useful markers of algorithmic success, despite the limitations imposed by constrained hardware capabilities. They act as standards for evaluating and optimizing algorithm performance, directing the creation of more effective and scalable solutions. Researchers and practitioners can better understand the subtle difficulties of dealing with enormous datasets by realizing the significance of algorithmic adaptation to changing hardware restrictions. This insight spurs breakthroughs in algorithm design and execution, bridging the gap between hardware restrictions and practical efficiency.
- The potential benefits of utilizing a hybrid recommendation system that incorporates many algorithms may have been hidden by hardware constraints. A hybrid technique can both reduce and possibly overcome the limitations inherent in individual methods by integrating different algorithms. Additional investigation into hybrid models may reveal unrealized potential for improved performance, especially in contexts with limited resources. Examining the interrelationships between various algorithms in a hybrid framework enables the discovery of complementing advantages and the utilization of various data features. As a result, these studies aid in the creation of recommendation systems that are more resilient and flexible and that can produce better outcomes even in the face of limited computer resources. It may be possible to maximize the use of available computer resources and uncover latent efficiencies in recommendation system design by embracing hybridization techniques.

6.3 Future-Work

- Hybrid techniques, which make use of a wide range of algorithms, seem to hold great promise for recommendation systems in the future. It is possible to improve system performance by combining the special advantages of different approaches. The viability of studying hybrid models increases with the advancement of technology and computational resources. This research offers opportunities to improve recommendation systems' accuracy and efficiency while meeting changing user demands and preferences. By embracing the flexibility and adaptability of hybrid techniques, practitioners and academics may push the boundaries of recommendation system design and create more meaningful and personalized user experiences across a range of applications and domains.
- Recommendation systems in online retail can be greatly improved by utilizing high-quality hardware. Increased processing power allows algorithms to function more reliably, analyze data more quickly, and scale more effectively. Better technology gives algorithms more capacity to work as efficiently and accurately as possible to provide customized product recommendations. As a result, this improvement improves user experience overall and increases client happiness and loyalty. These developments support current business operations and open up new avenues for growth in the very competitive e-commerce space, giving retailers the advantage to remain flexible and adaptable to changing consumer demands.

References

- [1] “<https://www.zippia.com/advice/what-percentage-of-small-businesses-fail/>,”
- [2] “<https://fastercapital.com/topics/the-cost-of-delayed-access.html>,”
- [3] “<https://www.statista.com/chart/14011/e-commerce-share-of-total-retail-sales/>,”
- [4] “<https://www.tipranks.com/stocks/amzn/website-traffic>,”
- [5] C. Wu and M. Yan, “Session-aware information embedding for e-commerce product recommendation,” in *Proceedings of the 2017 ACM on conference on information and knowledge management*, pp. 2379–2382, 2017.
- [6] P. Xiu-Li and J. Wei, “Research on influential factors of e-commerce recommendation user behavior intention,” in *2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*, pp. 2485–2490, IEEE, 2017.
- [7] J. Xiao, M. Wang, B. Jiang, and J. Li, “A personalized recommendation system with combinational algorithm for online learning,” *Journal of ambient intelligence and humanized computing*, vol. 9, pp. 667–677, 2018.
- [8] N. M. S. Iswari, W. Wella, and A. Rusli, “Product recommendation for e-commerce system based on ontology,” in *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, vol. 1, pp. 105–109, IEEE, 2019.
- [9] M. Zhou, Z. Ding, J. Tang, and D. Yin, “Micro behaviors: A new perspective in e-commerce recommender systems,” in *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 727–735, 2018.
- [10] L. Jiang, Y. Cheng, L. Yang, J. Li, H. Yan, and X. Wang, “A trust-based collaborative filtering algorithm for e-commerce recommendation system,” *Journal of ambient intelligence and humanized computing*, vol. 10, pp. 3023–3034, 2019.
- [11] J. Knoll, R. Groß, A. Schwanke, B. Rinn, and M. Schreyer, “Applying recommender approaches to the real estate e-commerce market,” in *Innovations for Community Services: 18th International Conference, I4CS 2018, Žilina, Slovakia, June 18-20, 2018, Proceedings*, pp. 111–126, Springer, 2018.
- [12] J. Fabra, P. Álvarez, and J. Ezpeleta, “Log-based session profiling and online behavioral prediction in e-commerce websites,” *IEEE Access*, vol. 8, pp. 171834–171850, 2020.

- [13] U. Nadine, H. Cao, and J. Deng, "Competitive recommendation algorithm for e-commerce," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1539–1542, IEEE, 2016.
- [14] P. D. Hung and D. Le Huynh, "E-commerce recommendation system using mahout," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 86–90, IEEE, 2019.
- [15] M. Y. H. Al-Shamri, "User profiling approaches for demographic recommender systems," *Knowledge-Based Systems*, vol. 100, pp. 175–187, 2016.
- [16] T. Badriyah, E. T. Wijayanto, I. Syarif, and P. Kristalina, "A hybrid recommendation system for e-commerce based on product description and user profile," in *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, pp. 95–100, IEEE, 2017.
- [17] S. Ouaftouh, A. Zellou, and A. Idri, "Social recommendation: A user profile clustering-based approach," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 20, p. e5330, 2019.
- [18] Y.-M. Li, C.-T. Wu, and C.-Y. Lai, "A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship," *Decision support systems*, vol. 55, no. 3, pp. 740–752, 2013.
- [19] Y. Gu, Z. Ding, S. Wang, and D. Yin, "Hierarchical user profiling for e-commerce recommender systems," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 223–231, 2020.
- [20] Y. Li, J. Xu, and M. Yang, "Collaborative filtering recommendation algorithm based on knn and xgboost hybrid," in *Journal of Physics: Conference Series*, vol. 1748, p. 032041, IOP Publishing, 2021.
- [21] Z. Shahbazi and Y.-C. Byun, "Product recommendation based on content-based filtering using xgboost classifier," *Int. J. Adv. Sci. Technol*, vol. 29, pp. 6979–6988, 2019.
- [22] M. Guo, N. Yan, X. Cui, S. Hughes, and K. Al Jadda, "Online product feature recommendations with interpretable machine learning," *arXiv preprint arXiv:2105.00867*, 2021.
- [23] "<https://innovationyourself.com/lightgbm-in-machine-learning/>,"
- [24] "<https://medium.com/analytics-vidhya/introduction-to-xgboost-algorithm-d2e7fad76b04>,"

- [25] "https://en.m.wikipedia.org/wiki/file:random_forest_explain.png,"
- [26] "<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>,"
- [27] "<https://www.kaggle.com/datasets/dschettler8845/recsys-2020-ecommerce-dataset>,"
- [28] "<https://www.kaggle.com/code/danofer/ecommerce-store-predict-purchases-data-prep/output>,"

Thesis_report().pdf

ORIGINALITY REPORT

14%

SIMILARITY INDEX

10%

INTERNET SOURCES

8%

PUBLICATIONS

7%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Warwick Student Paper	1 %
2	www.researchgate.net Internet Source	<1 %
3	Submitted to University of Cape Town Student Paper	<1 %
4	Submitted to University College London Student Paper	<1 %
5	www.geeksforgeeks.org Internet Source	<1 %
6	Submitted to Liverpool John Moores University Student Paper	<1 %
7	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
8	Pang Xiu-Li, Jiang Wei. "Research on influential factors of E-commerce recommendation user behavior intention",	<1 %

2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017

Publication

9

Phan Duy Hung, Dinh Le Huynh. "E-Commerce Recommendation System Using Mahout", 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019

Publication

<1 %

10

researchr.org

Internet Source

<1 %

11

docs.auger.ai

Internet Source

<1 %

12

www.coursehero.com

Internet Source

<1 %

13

robots.net

Internet Source

<1 %

14

Submitted to UNITEC Institute of Technology

Student Paper

<1 %

15

Submitted to University of Huddersfield

Student Paper

<1 %

16

essay.utwente.nl

Internet Source

<1 %

17

www.mdpi.com

Internet Source

<1 %

18	slideheaven.com Internet Source	<1 %
19	www.sciendo.com Internet Source	<1 %
20	zaguan.unizar.es Internet Source	<1 %
21	ichatz.me Internet Source	<1 %
22	mafiadoc.com Internet Source	<1 %
23	gateway.ipfs.io Internet Source	<1 %
24	journal.access-bg.org Internet Source	<1 %
25	www.arxiv-vanity.com Internet Source	<1 %
26	M. Baritha Begum, G. Sivakannu, J. Eindhmathy, J. Sangeetha Priya, M. Mahendran, R. Ranjith Kumar. "Enhancing Agricultural Productivity with Data-Driven Crop Recommendations", 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023 Publication	<1 %

27	"Advances in Parallel Computing Algorithms, Tools and Paradigms", IOS Press, 2022 Publication	<1 %
28	Submitted to Cranfield University Student Paper	<1 %
29	Danveer Rajpal, Akhil Ranjan Garg. "Ensemble of deep learning and machine learning approach for classification of handwritten Hindi numerals", Journal of Engineering and Applied Science, 2023 Publication	<1 %
30	Umutoni Nadine, Huiying Cao, Jiangzhou Deng. "Competitive recommendation algorithm for E-commerce", 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 Publication	<1 %
31	sersc.org Internet Source	<1 %
32	kse.ua Internet Source	<1 %
33	Submitted to The Scientific & Technological Research Council of Turkey (TUBITAK) Student Paper	<1 %
34	www.wsdm-conference.org Internet Source	<1 %

35	Ni Made Satvika Iswari, Wella Wella, Andre Rusli. "Product Recommendation for e-Commerce System based on Ontology", 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), 2019 Publication	<1 %
36	ir.uitm.edu.my Internet Source	<1 %
37	scholar.archive.org Internet Source	<1 %
38	vdoc.pub Internet Source	<1 %
39	Submitted to De Montfort University Student Paper	<1 %
40	Submitted to Technological University Dublin Student Paper	<1 %
41	thesai.org Internet Source	<1 %
42	www.tandfonline.com Internet Source	<1 %
43	184pc128.csie.ntnu.edu.tw Internet Source	<1 %
44	Al-Shamri, Mohammad Yahya H.. "Power coefficient as a similarity measure for memory-based collaborative recommender	<1 %

systems", Expert Systems with Applications, 2014.

Publication

45

Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, Dawei Yin. "Hierarchical User Profiling for E-commerce Recommender Systems", Proceedings of the 13th International Conference on Web Search and Data Mining, 2020

Publication

<1 %

46

www.doria.fi

Internet Source

<1 %

47

Dharini N, Jeevaa Katiravan, Sruthi Priya D M, Sakthi Sneghaa V A. "Intrusion Detection in Novel WSN-Leach Dos Attack Dataset using Machine Learning based Boosting Algorithms", Procedia Computer Science, 2023

Publication

<1 %

48

Submitted to Mercer University

Student Paper

<1 %

49

Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, Suku Nair. "A distributed architecture for phishing detection using Bayesian Additive Regression Trees", 2008 eCrime Researchers Summit, 2008

Publication

<1 %

50	Submitted to University of Central Lancashire Student Paper	<1 %
51	Submitted to University of Sydney Student Paper	<1 %
52	Submitted to The NorthCap University, Gurugram Student Paper	<1 %
53	Submitted to University of Mauritius Student Paper	<1 %
54	Submitted to WorldQuant University Student Paper	<1 %
55	academic.oup.com Internet Source	<1 %
56	archive.org Internet Source	<1 %
57	core.ac.uk Internet Source	<1 %
58	dokumen.pub Internet Source	<1 %
59	open.uct.ac.za Internet Source	<1 %
60	doi.org Internet Source	<1 %

61	Duc Thang Nguyen, Soojin Lee. "LightGBM-based Ransomware Detection using API Call Sequences", International Journal of Advanced Computer Science and Applications, 2021 Publication	<1 %
62	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1 %
63	Ton Duc Thang University Publication	<1 %
64	Submitted to University of Sunderland Student Paper	<1 %
65	www2.mdpi.com Internet Source	<1 %
66	19aa972c-1ed1-4b46-b679-ba7325bef265.filesusr.com Internet Source	<1 %
67	Submitted to Bournemouth University Student Paper	<1 %
68	Submitted to Tilburg University Student Paper	<1 %
69	Submitted to University of Durham Student Paper	<1 %
70	Submitted to University of Essex Student Paper	

<1 %

71

Submitted to University of Portsmouth

Student Paper

<1 %

72

Submitted to University of Salford

Student Paper

<1 %

73

ds.inflibnet.ac.in

Internet Source

<1 %

74

repository.ubaya.ac.id

Internet Source

<1 %

75

www.slideshare.net

Internet Source

<1 %

76

Nabila Husna Shabrina, Siwi Indarti, Rina Maharani, Dinar Ajeng Kristiyanti, Irmawati, Niki Prastomo, Tika Adilah M. "A novel dataset of potato leaf disease in uncontrolled environment", Data in Brief, 2024

Publication

<1 %

77

eprints.usm.my

Internet Source

<1 %

78

Prasanta Baruah, Pankaj Pratap Singh. "Chapter 31 Risk Prediction in Life Insurance Industry Using Machine Learning Techniques —A Review", Springer Science and Business Media LLC, 2023

Publication

<1 %

79	opus.lib.uts.edu.au Internet Source	<1 %
80	www.techscience.com Internet Source	<1 %
81	Gina George, Anisha M. Lal. "Review of ontology-based recommender systems in e-learning", Computers & Education, 2019 Publication	<1 %
82	Mohammad Yahya H. Al-Shamri. "User profiling approaches for demographic recommender systems", Knowledge-Based Systems, 2016 Publication	<1 %
83	hashnode.com Internet Source	<1 %
84	library.iiuc.ac.bd Internet Source	<1 %
85	mail.easychair.org Internet Source	<1 %
86	onlinelibrary.wiley.com Internet Source	<1 %
87	scholarscompass.vcu.edu Internet Source	<1 %
88	"Advanced Informatics for Computing Research", Springer Science and Business	<1 %

-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|----------------|
| <div style="background-color: #008000; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">89</div> | <p>"Database Systems for Advanced Applications", Springer Science and Business Media LLC, 2019</p> <p>Publication</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------|
| <div style="background-color: #8B4513; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">90</div> | <p>"Neural Information Processing", Springer Science and Business Media LLC, 2019</p> <p>Publication</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| <div style="background-color: #8B4513; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">91</div> | <p>"Proceedings of International Conference on Emerging Technologies and Intelligent Systems", Springer Science and Business Media LLC, 2022</p> <p>Publication</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|----------------|
| <div style="background-color: #00008B; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">92</div> | <p>123doc.net</p> <p>Internet Source</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| <div style="background-color: #800080; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">93</div> | <p>Havva Elif Saroğlu, Ibraheem Shayea, Bilal Saoud, Marwan Hadri Azmi, Ayman A. El-Saleh, Sawsan Ali Saad, Mohammad Alnakhli. "Machine learning, IoT and 5G technologies for breast cancer studies: A review", Alexandria Engineering Journal, 2024</p> <p>Publication</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------|
| <div style="background-color: #008000; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">94</div> | <p>Submitted to University at Buffalo</p> <p>Student Paper</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------|
-
- | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|----------------|
| <div style="background-color: #00008B; color: white; display: inline-block; width: 40px; height: 40px; text-align: center; line-height: 40px;">95</div> | <p>ir.ahduni.edu.in</p> <p>Internet Source</p> | <p><1 %</p> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|----------------|
-

96	medium.com Internet Source	<1 %
97	pdfcoffee.com Internet Source	<1 %
98	publisher.uthm.edu.my Internet Source	<1 %
99	web.archive.org Internet Source	<1 %
100	www.ijraset.com Internet Source	<1 %
101	"Innovations for Community Services", Springer Science and Business Media LLC, 2018 Publication	<1 %
102	Javier Fabra, Pedro Alvarez, Joaquin Ezpeleta. "Log-based Session Profiling and Online Behavioral Prediction in E-commerce Websites", IEEE Access, 2020 Publication	<1 %
103	Jun Xiao, Minjuan Wang, Bingqian Jiang, Junli Li. "A personalized recommendation system with combinational algorithm for online learning", Journal of Ambient Intelligence and Humanized Computing, 2017 Publication	<1 %

104	Sara Ouaftouh, Ahmed Zellou, Ali Idri. "Social recommendation: A user profile clustering-based approach", Concurrency and Computation: Practice and Experience, 2019 Publication	<1 %
105	Shaohua Tao, Runhe Qiu, Bo Xu, Yuan Ping. "Micro-behaviour with Reinforcement Knowledge-aware Reasoning for Explainable Recommendation", Knowledge-Based Systems, 2022 Publication	<1 %
106	Tessy Badriyah, Erry Tri Wijayanto, Iwan Syarif, Prima Kristalina. "A hybrid recommendation system for E-commerce based on product description and user profile", 2017 Seventh International Conference on Innovative Computing Technology (INTECH), 2017 Publication	<1 %
107	Jinghua Jiang, Xiaolin Gui, Zhenkui Shi, Xingliang Yuan, Cong Wang. "Towards Secure and Practical Targeted Mobile Advertising", 2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN), 2015 Publication	<1 %
108	Manav Gumber, Apoorv Jain, A L Amutha. "Predicting Customer Behavior by Analyzing Clickstream Data", 2021 5th International	<1 %

Conference on Computer, Communication and Signal Processing (ICCCSP), 2021

Publication

109

Meizi Zhou, Zhuoye Ding, Jiliang Tang, Dawei Yin. "Micro Behaviors", Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18, 2018

Publication

<1 %

110

Taushif Anwar, V. Uma, Md Imran Hussain. "chapter 10 Challenges and Applications of Recommender Systems in E-Commerce", IGI Global, 2021

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On