PerceptiStore: An AI Driven Intelligent Shopper

by

Nazia Islam
24341083

A project submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
BRAC University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nazia Islam

24341083

Approval

The project titled "PerceptiStore: An AI Driven Intelligent Shopper " submitted by

1. Nazia Islam (24341083)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

The PerceptiStore: An AI driven Intelligent Shopper project uses advanced AI-powered personalization to improve the online shopping experience. AI-driven personalisation has emerged as a key factor influencing the direction of e-commerce in the current digital environment. Personalized shopping experiences make a difference by customizing information, recommendations, and offers to individual interests, allowing customers to navigate a wide sea of products and possibilities. Beyond merely suggesting products, AI personalisation leverages user behavior, previous exchanges, and purchase trends to anticipate customer needs before they ever become aware of them. Additionally, AI is able to generate highly relevant recommendations that increase the probability of repeat purchases by combining behavioral data with purchase history. This degree of customisation makes the buying experience efficient and pleasurable, which not only increases sales but also cultivates consumer loyalty. E-commerce personalisation fosters closer ties between companies and their clients by providing a feeling of individualized care in an otherwise congested online market. Ensuring security and privacy is essential as AI analyzes more and more personal data. Finding a balance between protecting sensitive information and providing individualized experiences is a challenge. In the end, AI personalisation signifies a profound change in the way that companies and customers communicate, making e-commerce a more dynamic, responsive, safe, and customized experience that meets each user's specific demands.

Table of Contents

Chapter 1

Introduction

1.1 Thoughts behind the AI personalization in e-commerce

The infinite topic of Artificial Intelligence (AI) in computer science focuses on creating intelligent machines that can perform tasks that normally call for human intelligence. In essence, AI systems acquire vast amounts of specific training data, then use the patterns and correlations discovered in the data to forecast future events. The mutually beneficial connection between e-commerce dynamics and consumer behavior, with an emphasis on the creative impacts, powered by artificial intelligence personalization and corresponding market trends.

In addition to quick and simple transactions, buyers demand experiences that are personalized and meaningful as e-commerce platforms become more and more integrated into everyday life. AI, with its capacity to analyze vast datasets and discern intricate patterns, emerges as a game-changer in meeting these evolving expectations [7] . The precise ways that AI-powered personalization impacts customer interactions, shapes consumer preferences, and fortifies user-platform relationships are the primary emphasis of this research. The study also explores how constantly market trends are developing which AI is helping to shape in the e-commerce industry. A new age of innovation is under way, now ML algorithms are not only used to predict consumer preferences but also to the use of predictive analytics to optimize inventory management.

The more we dive into our study, the more it becomes evident that the correlation of e-Commerce and consumer behavior is evolving into a dynamic symbiosis urged by a transformative impact of AI-powered personalization. By mediating available terrain and finding opportunities for creativity and expansion, businesses may pave the way for a digital marketplace that is more responsive and consumer-focused.

E-commerce has become an essential part of the modern age, as it is transforming the economy and consumer expectations. The main backbone of this era is AI, which basically shaping the structure on how consumers use online platforms, how business procedures are organized, and how customized recommendations are generated. One of the main forces that is taking us to the digital front line and into a future where automation not only handles transactions but also amplify the bond that ties companies to their clients is the application of AI in e-commerce.

1.2 Research Problems

AI customization is ground breaking journey which will go through various obstacles. Among them data privacy and algorithmic assumptions are two ethical issues that need to be closely examined. Striking the delicate balance between customization and intrusiveness is crucial to ensure consumer trust and satisfaction [7]. One of the main part of AI-powered personalization is the analysis of large datasets, which include user behavior, preferences, and interactions. And this large number of datasets are mostly gathered using structured quantitative method, which raise concerns about the privacy as the process might not be ample to represent the subjective opinions on customer trust, privacy concerns, and perceptions of AI-driven personalization technologies. The recommendations and content customization are enhanced by this data-driven approach, but there are significant data privacy concerns as well. Consumers are increasingly aware of the value and sensitivity of their personal information, prompting concerns about how their data is collected, stored, and utilized by e-commerce platforms [7] . To prevent coming out as intrusive, a careful balance between personalization and user privacy must be achieved.
Customers value privacy and may get uneasy if they feel that their internet usage is being improperly monitored down or misused, even while they like personalization. Being aware of how their data will be processed and providing options for changing their settings can empower users and give them a sense of control over their online experiences. Governments and regulatory bodies are increasingly recognizing the need to establish guidelines and regulations to ensure the responsible and ethical use of AI technologies [7] . Appropriate law and order need to be introduced in this era of AI to handle the privacy and security concerns in a better and effective way instructed by the government.
Data privacy concerns, algorithmic bias, the delicate balance between customization and user privacy, and the importance of regulatory frameworks and industry standards all require careful attention [7] . By taking effective measures to address these obstacles, e-commerce platforms should try to obtain customer trust, encourage the ethical growth of AI technology in the digital economy, and preach fair and transparent customer experiences. As e-commerce is advancing into the new era of customized AI ethical concerns must be the top priority.
Some questions may arise from this like:

- How can e-commerce platforms balance personalization with privacy, without crossing into intrusiveness?
- How can e-commerce platforms comply with evolving regulations and ensure ethical use of AI?
- What steps can e-commerce platforms take to ensure AI-driven personalization doesn't harm user autonomy or create dependency?

This research will try to answer these questions by studying and exploring the possible and primal way to identify to make efficient privacy policy for customers.

1.3 Research objectives

The goal of the AI-Powered Personalization project is to apply machine learning algorithms to transform the way customers interact with product recommendations. This project intends to develop an effective recommendation system that can offer highly relevant products that are customized for individual users by evaluating large datasets covering user behavior, preferences, and purchase history. For this some key features that will be included in this project are:

Preference Learning:

Machine learning models will be trained to identify and predict user preferences based on their past interactions and behavior.

Purchase History Integration:

The project will make use of past purchase data. to predict future purchasing behavior,

Real-Time Recommendations:

As users go through the platform, the system will instantly offer products based on real-time data processing, ensuring that the recommendations are relevant and up to date.

Personalized User Profiles:

A user-friendly profile that would collect information according to the interactions and preferences of the consumers.

Privacy and Security:

In order to keep the consumer data secure, privacy and security will be prioritized.

Chapter 2

Literature Review

This [5] discuss about the impact of AI is personalizing in e-commerce and how the users perceive this experience. It explores different ways that the websites creates to make shopping feel more customized to each person. These include things like suggesting products that one specific person might like, changing prices depending on that individual, and showing different information to different people. In order to understand e-commerce in depth the researchers viewed people's experience through different angles like - marketing, psychology, and how people interact with AI. And it has been found that when a website shows you things it thinks you'll like and succeed on gaining your attention, you're more likely to stick around and browse. This happens because you're not wading through a bunch of stuff you don't care about. With fewer irrelevant options, people don't get as tired from making choices. This makes the whole shopping experience feel better.

But the study also points out some problems with letting AI handle personalization. But this AI technology is not very transparent when it comes to the maintaining safety of private information. These issues are becoming bigger as companies use really advanced AI to analyze huge amounts of personal data. The paper stresses how important it is to be careful with these issues. If companies aren't careful, they could lose people's trust. Overall, this study shows how AI can make shopping more engaging and personalized. But it also warns about using this technology irresponsibly. Companies looking to improve their websites need to think carefully about the complex issues that come with using AI in e-commerce. It's about finding a balance between personalization and ethical concerns.

This [9] looks at how AI systems and machine learning are changing the way we buy things on the internet. It talks a lot about how these technologies can make ads work better and give each user a more personal experience when they're buying things online. The author of this study explains how these smart technologies are helping online stores make shopping feel special for each person who visits their website. The AI does this by looking at a huge amount of gathered datasets, trying to predict what people might want to buy, and by implementing AI personalization to make things feel more personal.

The study shows that when online stores use these combinations of AI and ML to make things feel more personal, based on what users have looked at or bought before, people often like to shop more and spend more time looking around the website. The author gives a step-by-step workflow for the combination of AI and ML in online shopping. This plan starts with collecting datasets about consumers and ends with trying to predict what they might want to buy next. According to the study it's been said that when e-commerce uses these AI to pick which ads to show people, their ads work better because they can show the right ads to the people who might actually want to buy those things.

But the study also talks about some important things to worry about, like keeping people's information safe and following the rules about how to use user's information. As more and more online stores collect information about people, these worries are getting bigger and more important. The researcher thinks that even though there are some problems that need to be fixed, like maybe making things too personal or worrying about keeping information private, the AI system should keep changing how we buy things on the internet in big ways. The study advises online stores that want to use these AI personalization systems to make shopping better for people while also being careful with the information they collect about users.

The [7] looks at how AI is shifting the way people buy things in ecommerce and how online stores work. It shows how really smart AI programs can be at gathering lots of information about what people like to buy, which helps stores to suggest people about things they might want to buy, and it helps stores answer questions from people faster with AI chatbots that can talk to users, and it even helps stores know how much stuff to keep in their warehouses. The author of this study state that AI don't just help make shopping feel more personal for each person, but they also help stores work better by guessing what people might want to buy and by making it easier to get things from the place where they're made to the store. The study also talks about some problems that can happen when stores use AI, like when the AI might not be fair to everyone or when stores might use people's information in ways that people don't like, which could make people not trust the stores anymore. To fix these problems and make people trust the stores more, the writers say it's really important for stores to use AI in ways that are good and fair, like asking people if it's okay to use their information and telling people how they're using it. The study also talks about new things that are happening because of AI in ecommerce, like chatbots that can talk to people and help them find what they want to buy, and AI programs that can guess what people might want to buy next, these new things are making stores think of new ways to tell people about what they're selling and new ways to help people when they have questions. At the end, the study gives a big picture of how AI is changing not just how online stores work but also how it feels to buy things on the internet. This study looks at why it's important to think about what's right and wrong when using AI to make shopping feel more personal, and it also looks at how using AI can change how stores work and how well they do. By looking at all these different parts, the study helps people understand a lot about how AI is changing the way we buy things on the internet and how ecommerce works. The writers think that even though there are some problems with using AI in ecommerce, it will keep changing how we shop online in big ways. They say it's really important for stores to use AI in smart ways that make shopping better for people but also keep people's information safe and follow the rules about using data. The study also says that as AI gets even smarter, it might change ecommerce even more in ways we can't imagine yet, so it's important for stores and people who make rules to keep thinking about how to use AI in ways that are good for everyone.

This [2] talks about how online stores are using AI more and more to make things better for consumers and to help the stores work better. The study says that the things that make online stores really good now are mostly because of important AI stuff like chatbots that can talk to people, product suggestion that tell people what they might want to buy, ways to make shopping feel special for each person, and ways to know how much stuff to keep in the warehouses. These AI technologies help stores tell people about things they might like based on lots of information, they help answer people's questions even on weekends when the stores might be closed, and they help make sure the stores don't have too much stuff sitting around that nobody wants to buy. The writers also talk about how AI can look at information about what people buy to find out what lots of people like, which helps stores give people better things to buy. They use Amazon as a really good example, saying that Amazon makes a lot of profit because their AI is really good at suggesting people about things they might want to buy. The article says that because stores can look at so much information about what people like, they can make shopping feel really special for each person, which makes people want to come back and buy more things. At the end, the study says that AI is really, really important for online stores if they want to keep up with other stores in the fast-changing world of ecommerce. The writers think that stores that don't use AI might have a hard time because other stores that do use AI will be able to do things better and faster. They also say that as AI gets even smarter, it might change how we buy things online in ways we can't even think of yet, so it's really important for stores to keep learning about AI and finding new ways to use it. The study also talks about how using AI can be tricky sometimes, like when it comes to keeping people's information safe or making sure the AI is fair to everyone, but they think that if stores are careful and use AI in good ways, it can make shopping online much better for everyone.

While greater transparency of behavioral data is frequently prioritized, this article [1] focuses on the role that user emotions play in e-commerce personalization. Emotions, according to the author, exhibit a big impact on consumers' judgments regarding what to buy and how they interact with personalized content. In order to provide a comprehensive concept matrix that matches different emotions with the technologies used to identify them, such as speech analysis and facial recognition, the study evaluates the literature that has already been written on emotion identification technologies and its applications in e-commerce. With the use of these technologies, companies may provide a more emotionally engaging and personalized purchasing experience by dynamically adapting material to the emotional moods of their users. The study highlights that although emotions can significantly improve personalization, there are obstacles associated with putting such technology into practice. Since the gathering of emotional data necessitates the use of intrusive monitoring methods like voice and facial recognition, privacy concerns are more urgent. The writers also go over the technological shortcomings of the systems in use these days for identifying emotions, pointing out that many of them are still in the early stages of development and might not yet be precise enough for general commercial application. The author suggests that in order to fully realize the potential of emotion detection systems to improve e-commerce customization, future research should concentrate on strengthening their accuracy and ethical deployment. The article offers a futuristic perspective on how user experiences could be transformed, becoming more intuitive and in line with user preferences, by incorporating emotional elements into personalization tactics.

The [6] discusses about the privacy concerns in the world of AI personalization of

e-commerce. By taking the focus away from the advantages of personalized online shopping experiences to the dangers of private personal data, the thesis stresses over the fact how AI-powered products impact customer trust. By following the Privacy Calculus Theory, which shares that customers should think about the usefulness and drawbacks of sharing personal information. Using a quantitative research approach and an Amazon user survey, researchers have found that customers really don't think about sharing personal information when they are offered any special offers. Among many other risks, like consent violations and data breaches, raise privacy issues and corrupt public confidence in AI systems. The survey showa an interesting finding: consumers are growing increasingly conscious of the possible risks related with revealing their personal information online, there are also scenarios where they feel their anonymity to be high. These findings are important because they indicate that companies should exercise caution when implementing AI personalization technologies in order to prevent damaging customer confidence. They should make data handling procedures more transparent and making sure that laws like the GDPR are followed should be practiced. Insights for e-commerce platforms hoping to implement AI-driven products without compromising customer trust are the results from this research, which advances our understanding of how privacy concerns directly affect consumer trust in AI technologies.
The goal of [3] is to increasing competitiveness and functional ability by improving supply chain management in e-commerce by implementing (AI) with big data. As the retail industry is growing rapidly supply chains are getting more and more complicated. In order to implement procedures like demand projections, inventory control, and route optimization the real-time data analysis, statistical modeling, and automated decision-making by AI and big data are the main factor. These capabilities lead to creative solutions. Thanks to the AI-driven models that estimate customer behavior, spot problems with the supply chain, and suggest dynamic modifications the rate of accuracy and responsiveness is increasing a lot. Especially because of machine learning (ML) businesses are thriving as it is easier to examine past data, predict future demand, and optimize logistics operations through transportation management and route optimization. Also by the presence of reliable, on-time delivery of products, these advancements helping to increase customer plea-

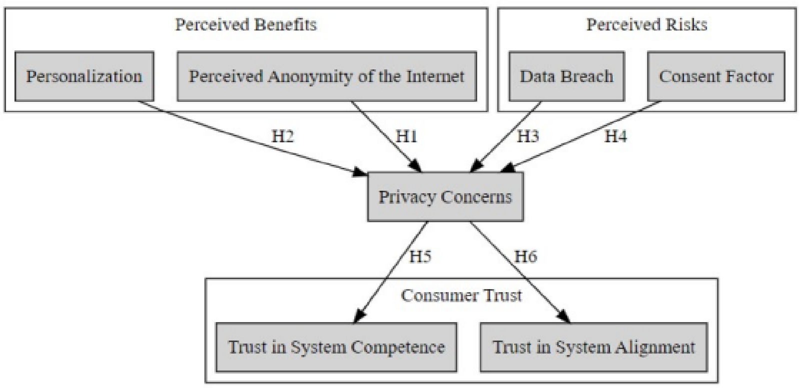

Figure 2.1: A conceptual framework on Privacy Concern

sure. The report also examines how Big Data works in combination to improve supplier management and give accurate demand predictions. The report ends with, companies that use AI and Big Data into the administration of supply chains will have a major competitive advantage because they have higher ability to make better decisions, higher customer satisfaction, and lower operating costs. E-commerce platforms can respond more quickly to changing market conditions and client requests when these technologies are used. This study highlights the role that technological innovation plays in simultaneous supply chains, highlighting the significance of artificial intelligence (AI) and big data as key components of e-commerce structure.
The [4] gives a thorough overview of AI's uses in e-commerce. To digitize e-commerce the author emphasized on improving customer experience, streamlining processes, and advancing business intelligence. Some of the key technologies of artificial intelligence (AI) that has been reviewed are Natural language processing (NLP), robotics, expert systems, artificial neural networks (ANNs), and machine learning (ML). To improve consumer interaction, among many some of the few e-commerce uses of AI Chatbots, voice assistants, recommendation engines, and personalized marketing and improve the purchasing process. Delivering personalized product recommendations, instantly responding to client inquiries, and facilitating easy transactions became much more easier for companies. By understanding and reacting to natural language commands, AI chatbots that are powered by NLP, enhance customer service. Also, the article talks about how AI can analyze large datasets in real time, improving management, enabling dynamic plans, and promoting logistics efficiency through predictive algorithms. In addition to automation business productivity and customer satisfaction is also depending on AI's ability along with facilitate data-driven decision-making. The study focuses on the application of AI in systems of suggestions, which use ML techniques to provide personalized product recommendations based on consumer behavior and past purchases. Though the need for more study on ethical issues like data protection and transparency, the paper ends with a discussion on the future of AI in e-commerce. Although AI is still in its early stages in many industries but the author believes that its uses in e-commerce are growing quickly. Economic efficiency, customer satisfaction, and market share can all see major gains for businesses integrating AI technologies. In areas like supply chains, customer behavior analysis, and ethical AI implementations, the report vouches for further research into the potential of AI.
This [10] talks about the ability of AI to improve user experiences, simplify processes, and boost revenue growth in e-commerce. The various uses of artificial intelligence, including smart search, automation, and personalization have been covered in this study. These technlogies depend on advanced algorithms to provide customized experiences. One of the key theme is concept of personalization is . To offer items that match individual interests, ML algorithms such as CF and CBF analyze user data. To process personalized recommendations and boost user engagement and revenue, these systems analyze patterns from past browsing activity and purchases. For modern e-commerce platforms (NLP) is utilized by AI-driven chatbots and virtual assistants to navigate customers' buying paths, respond to their inquiries, and enhance service effectiveness. In the article the use of AI in smart search has been talked about. Complex algorithms are used to improve the search request quality, and provide relevant product recommendations. With the help of functions like faceted navigation and auto complete, which helps the finding of products of

interest by consumers and enhance their shopping experience. Additionally, organizations may cut costs, and more efficiently use resources thanks to AI's capacity to automate human jobs like inventory management and customer care. By evaluating user behavior to detect possible leads and deliver customized advertisements or alerts that convince consumers to come back and finish their transactions AI ensures re marketing. . In conclusion, the study highlights how (AI) is changing the e-commerce by giving companies access to data-driven technologies that help them to create customized customer experiences and optimize operations. Continuing to invest in these technologies is necessary for long-term success. Long-term success and trust-building still heavily depend on ethical issues, especially those dealing with data protection.

Chapter 3

Work Plan

For the recommendation system of this AI-powered personalization project machine learning(ML) techniques will be used. A huge amount of databases on user activity, preferences and past purchases will be needed to ensure up to date relevant suggestions. In order to predict consumer preference of the collected data on user activity, preferences, and past purchase, will be analyzed using collaborative filtering and content-based filtering. Also trained algorithm models like random forest or gradient boosting will be followed to predict future preferences. This training will ensure the system is organizing each user's past purchase preferences and actions to customize user profiles. Strategies like online learning will be added to keep this profiles up to date in order to make sure the recommendation changes as the uses behavior does. In order to ensure that product recommendations are constantly up-to-date streaming algorithms such as Kafka or Spark Streaming will analyze data as users go through the platform. Security and privacy are critical. In regret of user data security the techniques like a differential privacy will be implemented so that the system can offer precise recommendation without jeopardizing user privacy. The system will go through testing and optimization in order to increase user engagement and suggestion accuracy.

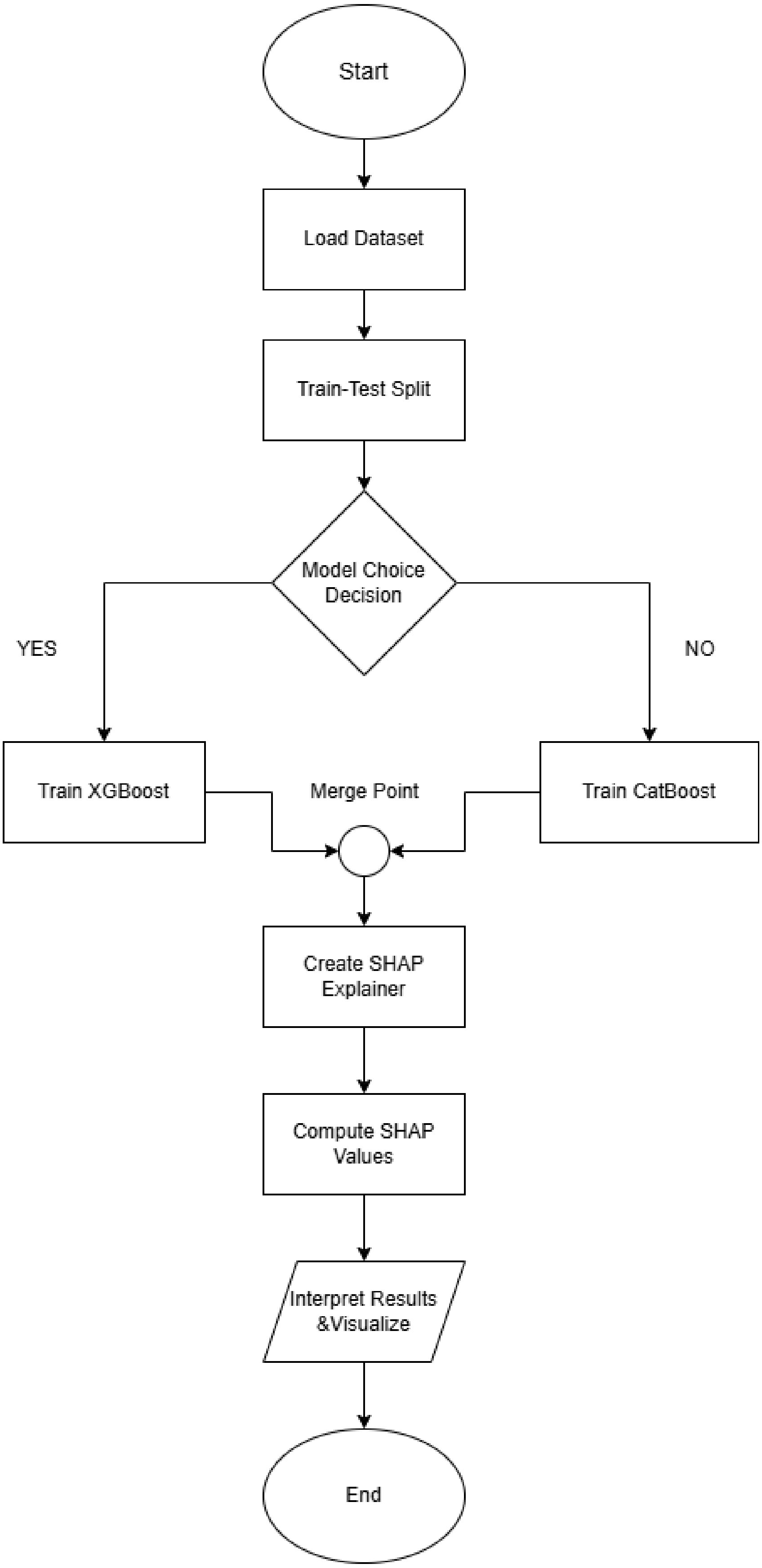

Figure 3.1: Work Flow

Chapter 4

Dataset Description

The [8] dataset that has been used for this AI powered personalization has been obtained from Kaggle titled as "Ecommerce product recommendation collaborative" which is designed to improve the research and development in the area of recommendation systems specially in the e-commerce sector. There are other recent datasets focusing on personalized e-commerce recommendation system but this dataset focus more on collaborative filtering based recommendation system which is like the core technique for modern e-commerce personalization. Unlike other datasets which primarily contain sales data without user-item interaction or focus more about general product data and customer review, this dataset is more structured for direct implementation of recommendation algorithms. Without excessive preprocessing this dataset is designed for recommendation tasks which ensures cleaner and more relevant features. This dataset contains more explicit user-item interactions which makes it the best option for collaborative filtering, matrix factorization and deep learning models. Generally e-commerce datasets are confidential for which they are not easily accessible by the general public However this dataset that includes real transactions from 2010 and 2011 was made using UCI Machine Learning Repository. From the "Online Retail" part of the headline it is indicated that the dataset of this website is up to date. These data have been obtained from a UK based registered non-store internet retailer between the duration from January 12, 2010 to September 12, 2011 which includes multinational data collection. This dataset has been provided by Dr. Daqing Chen, Director of the Public Analytics group, School of Engineering, London South Bank University, London SE1 0AA, UK by following the UCI Machine Learning Repository. The data in this dataset has been used for training and testing recommendation algorithms to predict which products the given user will be interested in based on their interaction history. The interaction history contains records of user activities like engagement with the type of products, purchases, views, ratings etc. This dataset is a collection of data related to user interactions with products in an e-commerce environment. The dataset has been organized to make the application of collaborative filtering techniques easier for both used-based and item-based collaborative filtering. By doing so it has been easier to analyze patterns of interaction to suggest items that a user might like or that are similar to items they have bought previously or interacted with. Recommendation systems requires a well-structured user behavior data which helps in building more practical personalized shopping experience. A detailed description of the attributes of this dataset and their efficiency has been discussed below:

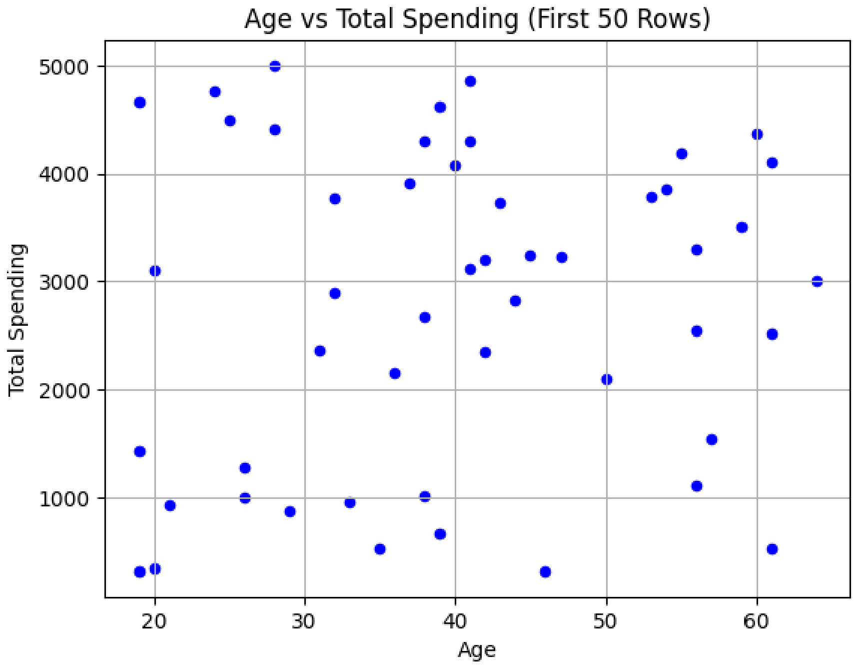

Figure 4.1: Age Vs Total Spending

User ID: To track the individual user activity and interaction within the platform every user is given an unique user id. With the help of this, analyzing each user's behavior like purchase, browsing habits and interaction with various items can be analyzed easily. And by analyzing these personalized recommendations becomes more accurate. Based on every unique user's behaviors or demographic characteristics it also helps in segmenting users for targeted marketing, promotions as well as communication.
Age: Age is one of the main factors in understanding product preferences. The attribute user age helps analyze generational preferences by giving a demographic snapshot of the people. For example , the younger the user is the more likely they are to prefer trendy fashion or tech products on the other hand older users are more likely to be interested in home decor or health supplements. By filtering these factors it becomes easier to customize marketing messages and product recommendations based on age groups.
Gender: The gender attribute helps with gender specific recommendations and promotions. Tailoring personalized advertisements or promotional emails according to gender specific interests can improve engagement and conversion rates. Gender can be influenced by product preferences like: fashion brands can recommend products based on gender for example clothing, beauty products or accessories.
Location: Geographical location of the users are very important as it helps to categorize products according to suburban, rural or urban areas. Location always influences the shopping behavior greatly. For example urban users will be more likely to purchase tech gadgets, fashion items or services while on the other hand rural users might focus on gardening tools, outdoor gears or practical household appliances. Apart from this product availability and shipping options can also be personalized based on user location, offering localized promotions, shipping rates and products suited for that area. And marketing campaigns can also be arranged

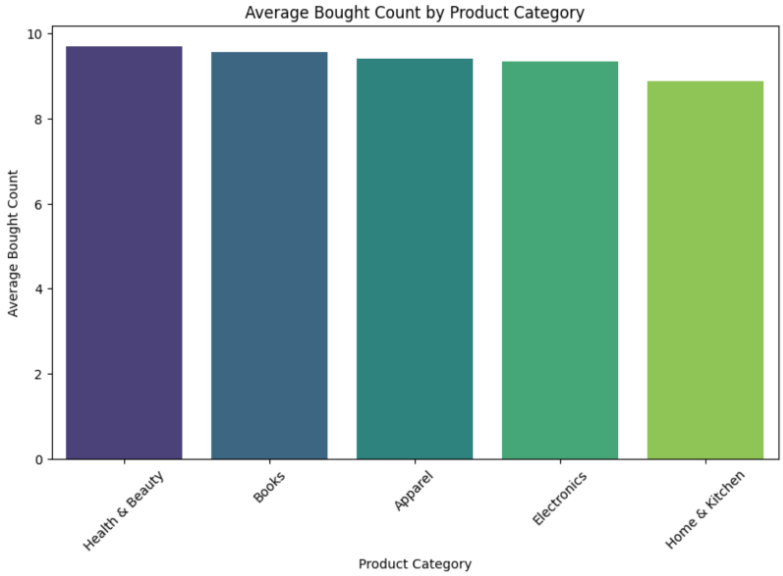

Figure 4.2: Product Category Preference

based on locations such as offering seasonal products.

Last Login Days Ago: This feature is the key metric to keep the user accessing records. With the help of last login days feature business can identify the customer engagement ratio and in according to that they can provide customized notifications for users if for example someone doesn't login for 30 days then they might receive an email with discount to reignite their interest or if any customer logs in frequently then they might get an email with exclusive discount or rewards.

Product category preference: This attribute shows the user's preference for specific product categories for example electronics, fashion etc. Users with clear preference for certain product categories can be shown new arrivals or best sellers of that specific category. For example, an user interested in electronics could recommend the latest gadgets. Also with the help of this attribute business can give category specific promotional or seasonal offers.

Total spending: By the help of this feature business owners can categorize users who have spent the most in a year and provide offers according to that to keep them interested. Businesses can allocate a fixed amount which if any user spent on purchase then they will be given the vip title and will get related offers and facilities according to that. Also with this feature business can identify who are close to that fixed amount and send them email or notification about their close milestone of becoming a vip.

Purchase Frequency: This attribute provides insight about user's loyalty and shopping habits by collecting data about how often and how much they are buying. Customers who purchase more frequently will be more likely to be interested in offers and exclusive discounts. And by focusing on those customers, businesses can recommend higher value products to increase average order value(AOV).

Pages Viewed: During the visit to the e-commerce site this attribute tracks down the pages a user views so that the business can predict their interest. Like, if a user is just browsing different category products and not buying then businesses can use that data to offer personalized discounts or promotions on the products viewed to encourage them to buy. Also a higher number of pages viewed can indicate an user

is exploring and interested in a wide variety of products. Then the recommendation system can use that data to suggest more relevant items or also they can create a personalized recommendation based on what they have visited so far.
Newsletter Subscription: With the help of this binary feature businesses can identify whether the users are subscribed to the platform's newsletter or email list. For users who are subscribed they can receive personalized product recommendations discounts and exclusive offers through email campaigns. And for those who are subscribed businesses can lure them into subscription by offering or by letting them know the benefits of the subscription. Also subscribed users will be notified about new or upcoming products, sales or promotions.
Along with minimizing preprocessing efforts it offers a solid content based, hybrid or deep learning based suggestions into practical practice with structured data designed for recommendation algorithms. For businesses that focus on developing customer engagement and increase sales rate this is an ideal dataset because of it's connection to actual e-commerce applications. It is more effective than other dataset which requires extra feature engineering or transformation required for this system because of how specific it is to recommendation tasks.

Chapter 5

Methodology

5.1 Background

In e-commerce personalized recommendation it is very important to improve the user experience and raise sales. The capability of collaborative filtering (CF) to recommend things based on user activity patterns has made it an overpowering technique among the many others. Generally there are two types of CF methods, they are: memory based and model based. Since advanced methods for machine learning have developed, model-based CF has become more effective over time. Support Vector Machine(SVM), Random Forest, Logistic Regression, XGBoost, CatBoost, and Multi-Layer Perceptron (MLP) are some of the machine learning models that can enhance collaborative filtering for effective recommendation accuracy. It helps a lot in product suggestions with historic data. SVM is a very powerful classification algorithm for learning about the user to item interaction that is used in collaborative filtering. It helps in determining the user preference from their history and also helps to filter users that will be interested in offers and discounts. On the other hand based on previous interactions, logistic regression helps in predicting whether an user will buy a product or not. In collaborative filtering, this model can help by generating binary recommendations e.g, like/dislike, buy/not buy, basically it is useful for modeling user to product relationships. Whereas random forest is best at identifying complex relations between big user to product datasets. It can effortlessly handle if there is any data missing along with analyzing different other factors such as reviews, purchase history and user behavior pattern to make everything more personalized. Then there are other advanced ML models like XGBoost, CatBoost, and Multi-Layer Perceptron (MLP) which helps to enhance CF techniques to deliver personalized recommendations. XGBoost is very effective for CF that involves large datasets and helps to rank recommendations for features like user history, item popularity etc. CatBoost which is also a gradient boosting algorithm like XGBoost specifically used to handle categorical data which helps to make an ecommerce platform very user efficient. MLP or multi-layer perceptron is a type of deep neural network which is very suited for maintaining complex, non-linear interactions between product and users. By including these models in collaborative filtering techniques e-commerce systems can improve user engagement through offering personalized recommendations, increasing sales rate and also maintaining customer satisfaction.

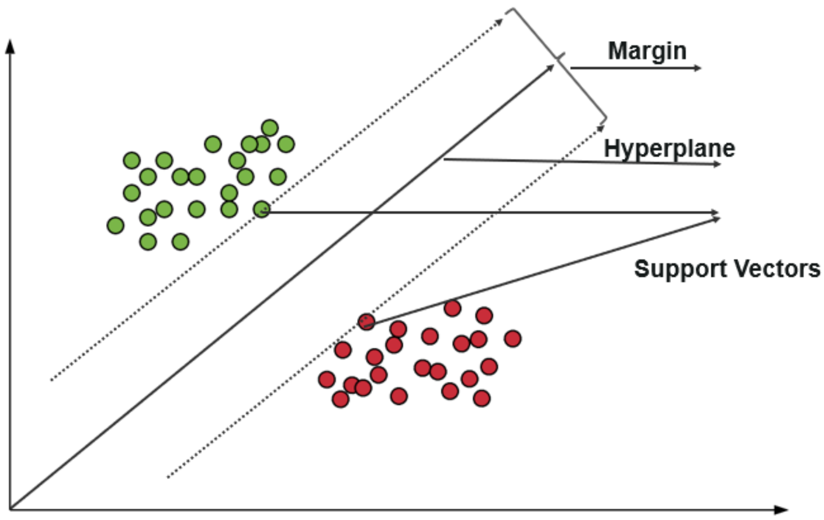

Figure 5.1: Support Vector Machine(SVM)

5.2 Support Vector Machine(SVM)

The support vector machine is a supervised machine learning algorithm that is initially used for classification and regression tasks. It is used mostly for binary classification along with different multiclass problems. To keep the data separated into different classes SVM approaches by finding a decision class which is also known as hyperplane. The key concepts of the SVM model are hyperplane, margin and support vector. The definition of hyperplane is that it is a decision boundary that separates the data points of different classes. Mostly in higher dimension it's a hyperplane whereas in 2D it's a line and in 3D it's a space. The main objective of SVM is to find a hyperplane that is best at separating the classes. Support vectors are the data points which are closest to the hyperplane. They are mostly responsible for the positions and orientation of the hyperplane. The margin is the distance between the hyperplane and support vectors. To keep the classes well separated with minimal misclassification SVM thrives to maximize the margin.
For collaborative filtering in recommendation systems, SVM are mostly used by treating problems like regression tasks like: predicting a user's rating for an item or classification task like: predicting what an user may like or dislike. From interaction feature, object features like: category or popularity or user feature like: demographic or past interaction SVM creates feature vectors that are needed for the input. This algorithm predicts values following the margin or by determining a hyperplane which divides the data into classes. Using kernels, SVM transfers data to higher dimensional spaces in order to improve separation for non-linear correlations. Training labels generally indicate user to item interactions such as: a positive (+1) refers to interaction whereas negative (-1) represents no interaction. SVM generally works by figuring out the optimal hyperplane which can separate the classes. For classification, this can be achieved by solving the following equation:

$$\min \frac{1}{2}\|w\|^2 \tag{5.1}$$

$$y_i(w \cdot x_i + b) \geq 1, \forall i \tag{5.2}$$

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{5.3}$$

w is the weight vector, b is the bias term, xi is the feature vector, and yi is the label (+1 for interaction, -1 for no interaction). SVM opts for kernel trick to map data into a higher dimensional space for non-linear detachable data using kernels like radial basis function(RBF). SVM works well with small sample sizes and usually sensitive to the high dimentional data that is why it is operationally costly as it requires precise use of features. It is commonly used for binary implicit feedback predictions or mixed recommendation systems which combines both content-based and collaborative filtering techniques.

5.3 Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary classification problems, here the main objective is to predict one or two possible outcomes e.g, spam/not spam, buy/not buy. Logistic regression predicts probabilities and design them to a binary outcome using sigmoid function whereas linear regression predicts continuous values. Logistic regression is comparatively easy to understand and implement, works efficiently for small to medium sized datasets, usually provides confidence score for predictions and it can also be extended with L1(Lasso) or L2(Ridge) regularization to prevent overfitting. However, this model struggles with non-linear separable data. It can become sensitive by a huge amount of data though works best with well-selected, organized and preprocessed features. In e-commerce logistic regression helps in predicting whether an user will purchase a product or not. Logistic regression can be used in collaborative filtering putting together a problem as a binary classification task to predict user to item interactions, here the main objective is to figure out whether an user will interact with an item(1) or not(0). Based on user to item interaction data it calculates the possibility of the interaction rather than making any specific rating predictions. All these user to item interactions are denoted as a numeric feature vector which includes interaction based features past purchases or latent embeddings from matrix factorizations as well as user variables like: demographic and past behavior also item variables like: category and popularity. The logistic regression model initially calculate a weighted sum of input features, denoted as

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b \tag{5.4}$$

Here, w represents the weights, x is the input features and b is the bias term. And after getting result from this the sigmoid function for probability prediction starts expressed as

$$P(y = 1|x) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{5.5}$$

which assists to have a probability between 0 and 1 output. If this probability cross a predefined threshold which is typically 0.5 then the model predicts as an interaction or else it predicts no interaction.

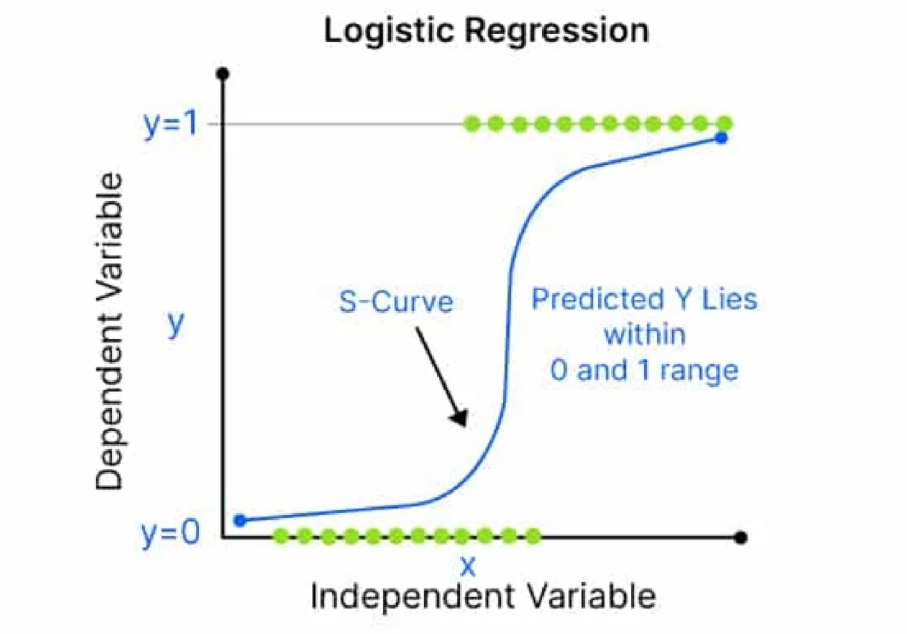

Figure 5.2: Logistic Regression

To optimize the model, logistic regression minimize the log-loss function, which is

$$J(w) = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.6}$$

using gradient descent, where weights are updated iteratively as

$$w_j := w_j - \alpha \frac{\partial J(w)}{\partial w_j} \tag{5.7}$$

Once trained, this model predicts interaction probabilities for various user to item combinations and the items that are in the highest probability are recommended to the user. Because it's easy to use, ability to be understood and effective logistic regression is very useful for dealing with big datasets, especially when working with indirect inputs like clicks and transactions. Though like all the other models it also has weaknesses like it has trouble with sparse data and is limited in its capacity to hold complex relationships as it depends a lot on a linear decision boundary. Apart from that it is commonly used in applications like online learning platforms for course enrollment prediction, streaming services for movie suggestions and e-commerce for purchase prediction. Unlike other complex models which require a huge amount of hyperparameter adjustments, logistic regression is simple to use and uses little processing power which makes it ideal for real time recommendation applications. It is also ideal for businesses as it creates an adjustable threshold for suggestions with its probabilistic feature which ensures customized experiences depending on various factors in gaining user preference confidence. Also it works really well with various machine learning techniques and is a very effective part of hybrid recommendation systems. By allocating weights to characteristics, logistic regression makes it very clear which criteria would affect recommendation, also improving system performance and gaining user confidence.

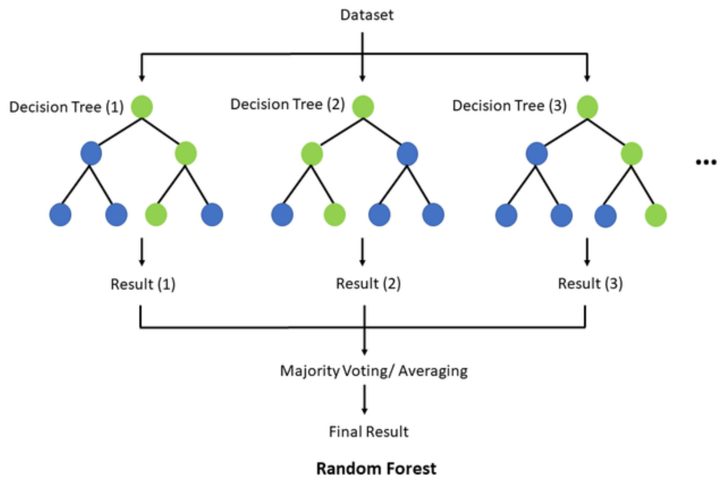

Figure 5.3: Random Forest

5.4 Random Forest

Random Forest is an assembler learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting. This algorithm is widely used for classification and regression tasks by strongly matching the results of multiple decision trees to make more strong and generalized predictions. Random forest, also known as a cooperative learning technique, generally sets up the recommendation problems as either a classification or regression problem for collaborative filtering. This model evaluates the user's rating for an item or predicts whether the user will engage with a certain product or not. For it to increase the accuracy rate and decrease the overfitting by training multiple decision trees using user to item interaction data and combining their predictions. Each and every user to item interaction pair is represented as a feature vector initially which incorporates item features like: category or popularity and user features like: past interactions. Following this bootstrap sampling is used in which a number of decision trees are trained on a random portion of the dataset. To further ensure the tree variety and to minimize the overfitting, a random subset of features are chosen at each node split in order to identify the most effective splitting rule.

For each decision tree, a random forest always creates individual predictions. For classification, each tree predicts whether the user will interact with an item; y=1 or will not interact with an item; y=0 and then the final prediction is determined by majority voting:

$$\hat{y} = \text{mode}(\{T_1(x), T_2(x), ..., T_n(x)\}) \tag{5.8}$$

For regression, each tree is assigned for a predicted rating and the final prediction is determined by averaging all the trees output:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} T_i(x) \tag{5.9}$$

Here, Ti(x) denotes the prediction from the i-th decision tree and n is the total number of the trees present in the forest. In scenarios where user preferences are being calculated from clicks, views or purchases rather than directly communicating or sharing their evaluation, Random Forest performs exceptionally well. It guides the businesses in understanding the most important variable in the recommendation system and is very strong immune to the missing data and feature importance analysis. But when it comes to the matrix factorization or deep learning techniques, operationally this model is very costly for big datasets and may not work well with extremely sparse data.

5.5 XGBoost

XGBoost which is also known as extreme gradient boosting is designed to be very efficient and fast. For Collaborative Filtering technique XGBoost can be used for learning from user-item interaction features so that it can predict how likely a user is to interact with i.e. buy, click or like an item. XGBoost is a very flexible and highly effective gradient boosting method which continuously improves the model using gradient descent to create a group of weak learners, which is mainly decision trees. The thin and high-dimensional data that is frequently found in e-commerce platforms can be handled with its effectiveness, capacity, and normalization capabilities. In CF technique the main objective is to predict how likely a user will interact with an item. In the following u - user, i - item and yui - output:

$$f(u, i) \to \hat{y}_{ui}$$

With this set of features describing both the user and the object, XGBoost represents this interaction as a function. Some of the features are activity signals, item information, browsing history, and population type data. The model is trained to reduce the gap between the expected values and actually observed interactions, turning the prediction task into a task that can be supervised. The objective function of XGBoost has two parts mathematically, they are: the training loss function and a regularization term. The general type of objective function is given below:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here l stands for differentiable loss function, omega(fk) for regularization penalty for the k-th tree in the ensemble. This regularization term is important to prevent overfitting and this is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

In the above equation T is the number of leaves in the tree, wj is the weight of each leaf node, gamma and lambda are regularization hyperparameters that control model complexity. In the training XGBoost creates new trees to minimize the objective function. At each iteration t, the model adds a new function ft which reduces the loss with respect to the current prediction. The objective function for the t-th iteration is given below:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

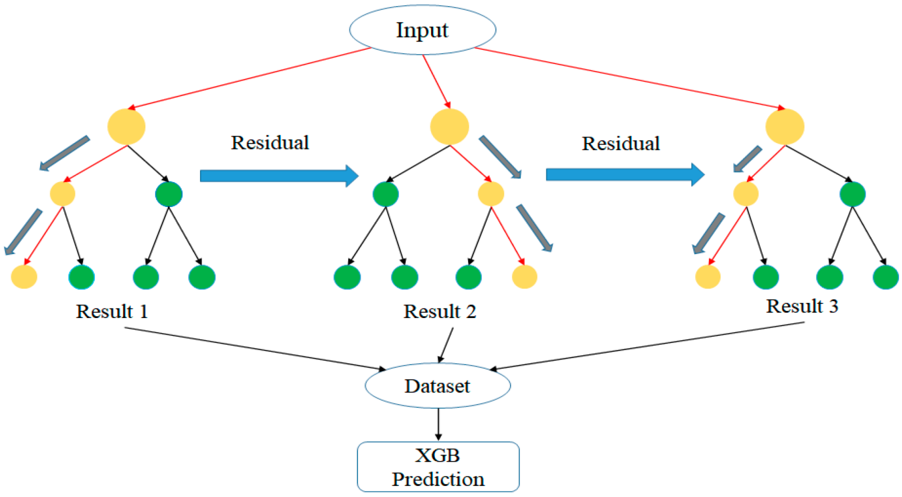

Figure 5.4: XGBoost

To optimize this efficiently XGBoost uses second order Taylor expansion of the loss function around the current prediction which is given below:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

In this equation gi and hi are denoted as the first and second derivatives of the loss function with respect to the current prediction. This second order optimization allows XGBoost to coverage quickly and handle a huge range of loss functions making it flexible for both regression and classification tasks. XGBoost performs great in collaborative filtering scenarios because of its capacity to deal with missing values, model non-linear interactions, and combine both item and user information. It performs better than conventional matrix factorization techniques when large data is present. E-commerce systems can greatly improve their capacity to offer precise, personalized product recommendations that increase user happiness and sales by including XGBoost into the recommendation pipeline.

5.6 CatBoost

CatBoost is also known as categorical boosting is a gradient boosting algorithm like XGBoost. This powerful algorithm was developed to handle common e-commerce challenges of categorical features such as demographics, product categories and other qualitative attributes. Unlike the previous gradient boosting method which required lots of preprocessing, CatBoost supports data through an innovative technique known as ordered target statistics. In collaborative filtering the main goal is to predict a user's preference or interaction with an item and in this scenario CatBoost makes this interaction like a supervised learning problem. Every pair from user-item is denoted as a feature vector which combines interaction history, item data and user features. Each decision tree in the ensemble that CatBoost creates is trained to minimize the remaining errors of the previous trees taken together. To

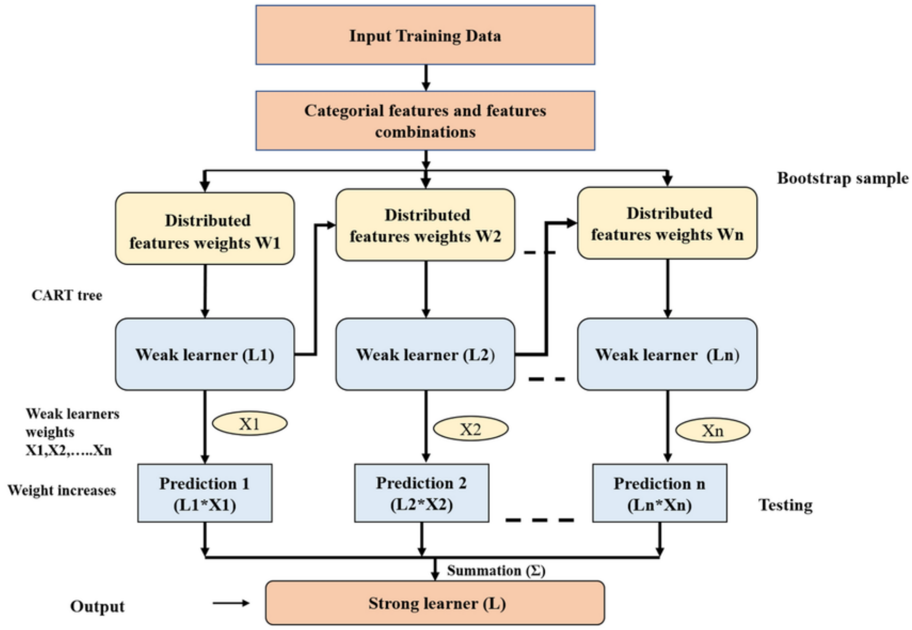

Figure 5.5: CatBoost

express this objective function mathematically the equation is given below:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

In the above equation yi is represented as a label indicating whether the user interacted with the item, yi is the predicted output for the i-th instance, l is represented as a differentiable loss function for example logistics loss for classification tasks, omega(ft) is the regularization term that penalizes the complexity of each tree ft and T is the total number of trees in the model. As the training continues at each iteration t, the model updates predictions according to the given formula:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

In the above equation n is learning rate which controls the step size for each and every update, ft is the newly trained decision tree and xi is feature vector representing user and item interaction. This incremental improvement enables the model progressively so that it can correct its mistakes in order to improve predictive performance.
There are scenarios where the recommendation problem is classified as binary classification like: predicting whether an user will buy a product or not i.e, the loss function l is the binary cross entropy loss which is defined as:

$$l(y_i, \hat{y}_i) = -\left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right]$$

Here the predicted probability pi is obtained by applying the sigmoid function to the raw model output yi:

$$p_i = \sigma(\hat{y}_i) = \frac{1}{1 + e^{-\hat{y}_i}}$$

And this probabilistic method of CatBoost helps to achieve more accurate output which can be useful to rank items for recommendation. CatBoost is perfectly suited for imitating the complex non-linear correlations found in user-item interactions formation. It allows e-commerce platforms to provide extremely accurate and personalized suggestions by identifying tiny trends in customer preferences and product characteristics that are sometimes overlooked by simpler algorithms. The result from this model is increased engagement, better client satisfaction which leads to more revenue.

5.7 Multi-Layer Perceptron

When it comes to e-commerce recommendation systems, a Multi-Layer Perceptron(MLP) is a very powerful deep learning architecture that can model complex and nonlinear user and item interactions. There are many traditional matrix factorizations that handle user-item interaction linearly whereas MLPs are very capable of learning high level abstract representations through multi-layered non-linear transformations. And because of these qualities MLPs are very suitable for neutral collaborative filtering and this is an up to date generalization which helps to extend CF methods using neural networks. To discuss the workflow of a MLP, this model consists of an input layer, one or more hidden layers with non-linear activation functions and an output layer. In this model the CF setup of each user-item is represented by a learnable embedding vector. The two vectors that are concatenated and used as the input to the MLP are:

$$\mathbf{z}_0 = [\mathbf{e}_u; \mathbf{e}_i]$$

In the above equation [.;.] this is vector concatenation. Z0 passes through multiple hidden layers. For a given layer l, the transformation can be written as follows:

$$\mathbf{z}_l = \phi(\mathbf{W}_l \mathbf{z}_{l-1} + \mathbf{b}_l)$$

Here, WL and Bl are weight matrix and bias vector of the l-th layer. Phi(.) stands for nonlinear activation function and Zl-1 is the output from the previous layer. Until the final output layer is form a scalar value representing the predicted preference score of user for item this process continues through the stack of hidden layers:

$$\hat{y}_{ui} = f(u, i) = \sigma(\mathbf{W}_L \mathbf{z}_{L-1} + b_L)$$

Here, sigma(.) is a sigmoid function used for binary classification tasks like predicting purchase/no purchase. This model is trained to minimize loss function over a set of observed user-item interactions. And for this binary recommendation tasks, the binary cross entropy loss is commonly used, which is:

$$\mathcal{L} = -\sum_{(u,i)\in\mathcal{D}} [y_{ui} \log(\hat{y}_{ui}) + (1 - y_{ui}) \log(1 - \hat{y}_{ui})]$$

Here, the ground truth label indicates whether the user interacted with the item and then predicted the probability from MLP. When it comes to collaborating filtering, the best part of Mlp is its flexibility and predictive capability. Beyond the basic dot product that is used in conventional matrix factorization, MLP is able to capture complex, high order interactions between user and item. This model can

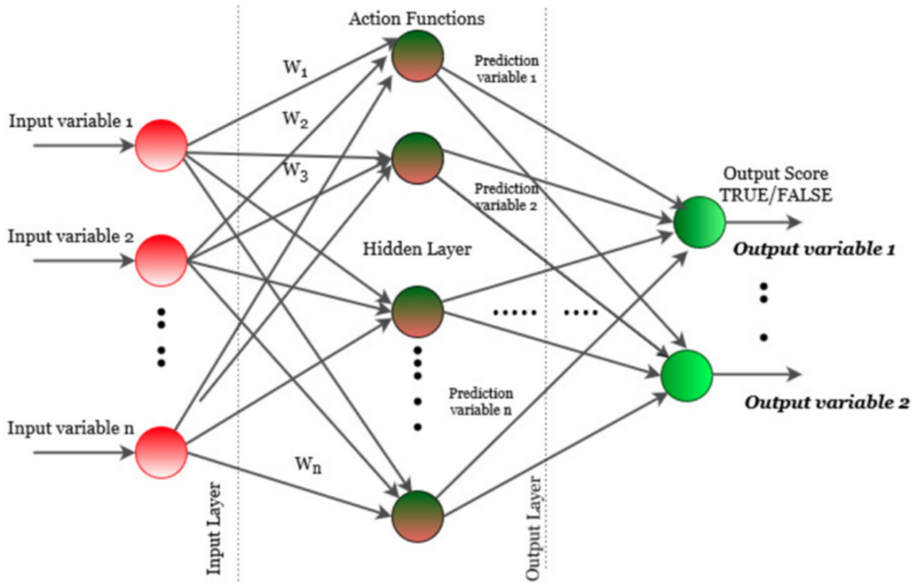

Figure 5.6: Multi-Layer Perceptron

identify patterns that are important for giving precise, personalized recommendations in e-commerce platforms where user behavior and product characteristics are affected by many contextual factors. Moreover Mlp based models are a flexible tool for creating any next generation recommendation systems as they can be easily customized to include extra information like, user characteristics, photos or any textual descriptions.

5.8 Data Privacy and Security Mechanisms

One must have feature for every e-commerce website would be strong data security measures in place, especially as e-commerce handles a huge number of clients' private data starting from financial information like credit card numbers and transaction histories to personally identifiable information(PII) like names, addresses and phone numbers. And failure in this sector would put brands reputation and customers' trust at risk which can also lead to legal penalties as in the digital world identity theft, data breach and cyber attacks are becoming more common and a conceptual framework on privacy concern is shown in Fig 2.1. A privacy focused e-commerce platform along with displaying ethical responsibilities can also gain a competitive edge by creating persistent client loyalty and guaranteeing business continuity. Because of that, ensuring data privacy only for customer trust is not it, this also helps businesses stay compliant with laws like: GDPR, CCPA and PCI-DSS. One of the first steps for this is encrypting data transit using HTTPs with TLS. By this it can be ensured that all the data exchanged between the user's browser and server is securely encrypted which will make it unreadable to attackers. And for password protection, using hashing and salting algorithms like BCRYPT can ensure that even if passwords are compromised they can't be cracked. Then in order to protect stored data in databases or servers, stronger encryption methods like AES-256 can be used. This will ensure even if the storage system is compromised, the information remains encrypted. Using secure APIs or mutual TLS in distributed or microservice based architecture, backend service and internal communication can also be encrypted. Then comes handling payment information which is very important. For that it's best to avoid storing raw credit card information and instead of that integrating with PCI compliant payment which offers tokenization to replace sensitive data with secure non-exploitable tokens would be better. Because this will reduce security burden and limit exposure to compliance liabilities. So, in conclusion, in order to turn a website from a possible threat to a trustworthy, privacy conscious space having multilayer security measures from data encryption and PCI-DSS compliant procedures to HTTPs and strong password hashing is a must. These measures will not only protect the platform from outside threats and insider abuse but also it will prepare it for future changes in technology and laws and regulations.

Chapter 6

Result and Analysis

Six algorithms- Support Vector Machine(SVM), Logistic Regression, Random Forest, XGBoost, CatBoost, and Multi-Layer Perceptron (MLP) these algorithms have different strengths and weaknesses when it comes to recommendation systems for e-commerce platforms. These algorithms are trained by collecting data from user interactions, which is used to predict purchase behaviour by using the technique of collaborative filtering.

6.1 Result

6.1.1 Accuracy

The accuracy is responsible to calculate the proportion of correct instances among all the instances. The equation for this is given in Equation (6.1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

where:

- TP = True Positives (correctly predicted purchases)
- TN = True Negatives (correctly predicted non-purchases)
- FP = False Positives (incorrectly predicted purchases)
- FN = False Negatives (incorrectly predicted non-purchases)

6.1.2 Precision

Precision is what measures actually how many predictions were correct. Here the interpretation is done in High Precision or Low Precision. High Precision means most of the predictions here correct, whereas Low Precision means most predictions were incorrect. Equation (6.2).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

where:

- TP = True Positives (correctly predicted purchases)
- FP = False Positives (incorrectly predicted purchases)

6.1.3 Recall

Recall or Sensitivity measures the actual number of positives that were predicted correctly. High Recall means model catches most of the real purchases and Low Recall means that most of the real purchases are being not predicted correctly. Equation (6.3).

$$\text{Precision} = \frac{TP}{TP + FN} \tag{6.3}$$

where:

- TP = True Positives (correctly predicted purchases)
- FN = False Negatives (incorrectly predicted non-purchases)

6.1.4 F1-Score

The harmonic mean between Precision and Recall is known as F1-Score. The purpose of F1-Score is to find the balance of trade-off between false positives and false negatives. Therefore, Higher the F1-score means higher the balance. Equation (6.4).

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.4}$$

where:

- Precision = $\frac{TP}{TP+FP}$ (Proportion of correctly predicted positive instances)
- Recall = $\frac{TP}{TP+FN}$ (Proportion of actual positive instances correctly predicted)

6.1.5 Comparison Between Algorithms

We have applied 6 algorithms i.e. SVM, Logistic Regression, Random forest, XGBoost, CatBoost and Multi Layer perceptron. Among the 6 algorithms the best performance matrix is obtained from XGBoost and CatBoost. Both of these algorithms are better for measuring imbalanced data. In XGBoost there is the *scale_pos_weight* parameter, this adjusts the weight of the positive class in imbalanced data and Catboost has its own native handling for categorical features and class imbalance, these features from these algorithms are playing a huge role in outcome for this dataset, specially with the added randomised noise, which helps us get realistic results from the algorithms. It is very evident that XGBoost and CatBoost is out performing the other algorithms.

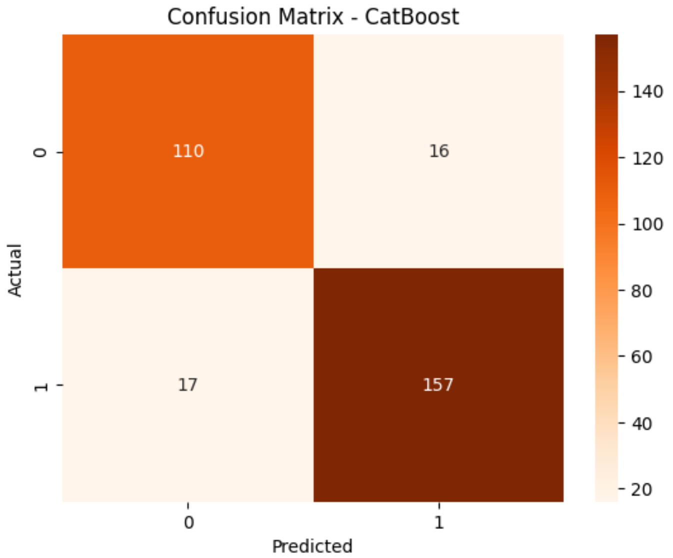

Figure 6.1: CatBoost

Table 6.1: Performance Comparison of Collaborative Filtering Algorithms

Algorithm	**Accuracy**	**F1-Score (Weighted)**	**Precision**	**Recall**
SVM	0.54	0.54	0.53	0.49
Logistic Regression	0.78	0.77	0.76	0.79
Random Forest	0.515	0.5144	0.51	0.50
XGBoost	0.88	0.87	0.87	0.87
CatBoost	0.89	0.89	0.89	0.89
Multi-Layer perceptron	0.80	0.79	0.80	0.79

6.2 Analysis

Based on collaborative filtering, six supervised machine learning models Support Vector Machine(SVM), Logistic Regression, Random Forest, XGBoost, CatBoost, and Multi-Layer Perceptron (MLP) were tested in this project. By the SMOTE oversampling technique these models were examined on a set of balanced data where Accuracy, Precision, Recall and F1-Score were utilized. The goal of this research was to understand the core advantages and disadvantages of every approach of a real world e-commerce system along with determining the best performing model. Among the six tested models, CatBoost and XGBoost have outperformed others across all key metrics. The CatBoost has shown the highest accuracy (0.89), F1-Score (0.89), precision (0.89), and recall (0.89), while XGBoost were very close with an accuracy of 0.88 and an F1-Score of 0.87.

Table 6.2: Performance of Best Performing Models

Model	**Accuracy**	**F1-Score**	**Precision**	**Recall**
CatBoost	**0.89**	**0.89**	**0.89**	**0.89**
XGBoost	0.88	0.87	0.87	0.87

These gradient boosting techniques have the capability to handle high-dimensional

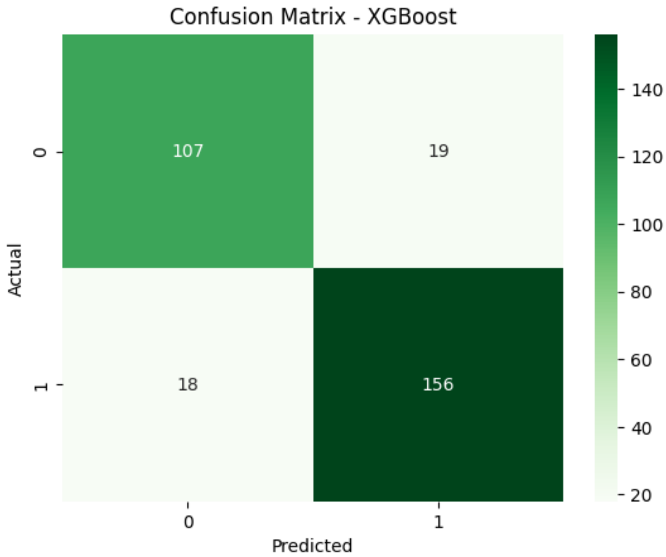

Figure 6.2: XGBoost

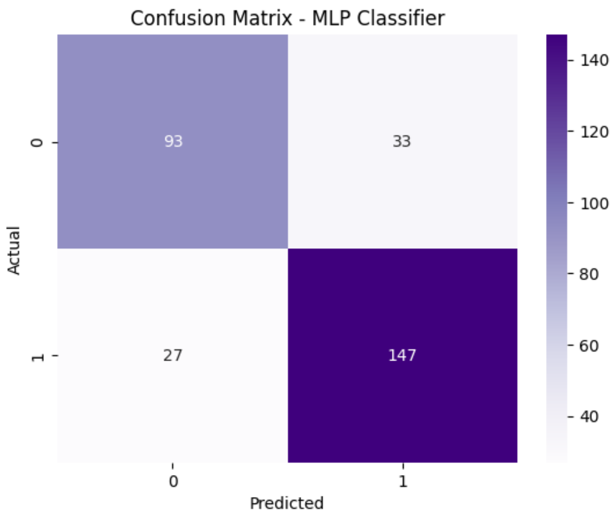

Figure 6.3: Multi-Layer Perceptron

data and model non-linear interaction which is responsible for their outstanding performance. The used dataset of this research has features like gender, location, product type which was well suited for CatBoost because it has its natural handling of categorical variables. Another reason why CatBoost performed better than others is because of its resistance to overfitting and decreased hyperparameter adjustment. On the other hand, Multi-Layer Perceptron (MLP) and Logistic Regression performed moderately. MLP showed an accuracy of 0.80 and F1-Score of 0.79, while Logistic Regression slightly fell behind with 0.78 accuracy. Even though both models have faster training and low computational cost they both fall behind when it comes to learning complex interaction in collaborative filtering.
Unexpectedly, Support Vector Machine (SVM) and Random Forest performed very poorly among these models. SVM only achieved an accuracy of 0.54 when it was very computationally demanding for large datasets. It struggled with dataset size which is very common in collaborative filtering. Similarly Random Forest had an

Table 6.3: Performance of Moderately Performing Models

Model	Accuracy	F1-Score	Precision	Recall
Multi-Layer Perceptron	0.80	0.79	0.80	0.79
Logistic Regression	0.78	0.77	0.76	0.79

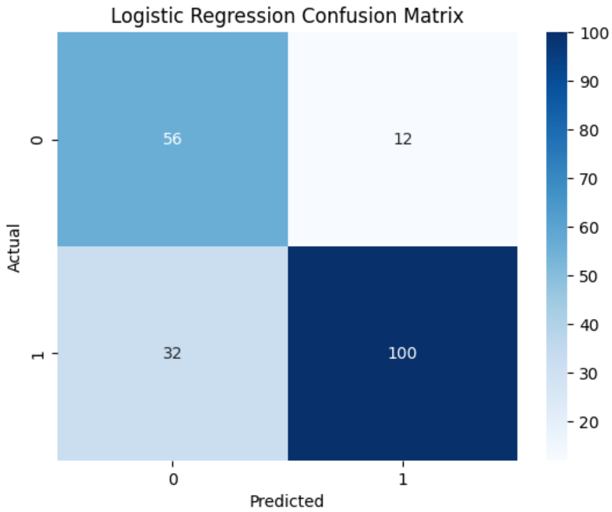

Figure 6.4: Logistic Regression

accuracy of 0.515 and the reason for that is possibly the lower efficiency with high cardinality of categorical data and it also lacks embedded management for class imbalance.

Table 6.4: Performance of Poorly Performing Models

Model	Accuracy	F1-Score	Precision	Recall
SVM	0.54	0.54	0.53	0.49
Random Forest	0.515	0.5144	0.51	0.50

For e-commerce recommendation problems with unique datasets, gradient boosting models like CatBoosting have outperformed every other model. CatBoost is the best option for real world implementation because of its resistance to overfitting and requires very low preprocessing and has native feature support. SVM and Random Forest performed poorly mostly because they were unable to adjust with the structure of user-item interaction data, even though they have historically strong performance on standard classification tasks. Random Forest failed in sensitivity of picking on small collaboration patterns whereas SVM struggled with scaling. Traditional models like Logistic Regression are very easy to understand but they are not a good choice for complex recommendation systems. So, in terms of predictive performance along with realistic implementation, the analysis concludes that CatBoost is the best choice for this project. Though XGBoost is also a good alternative, specifically in order to have flexibility over model tuning.

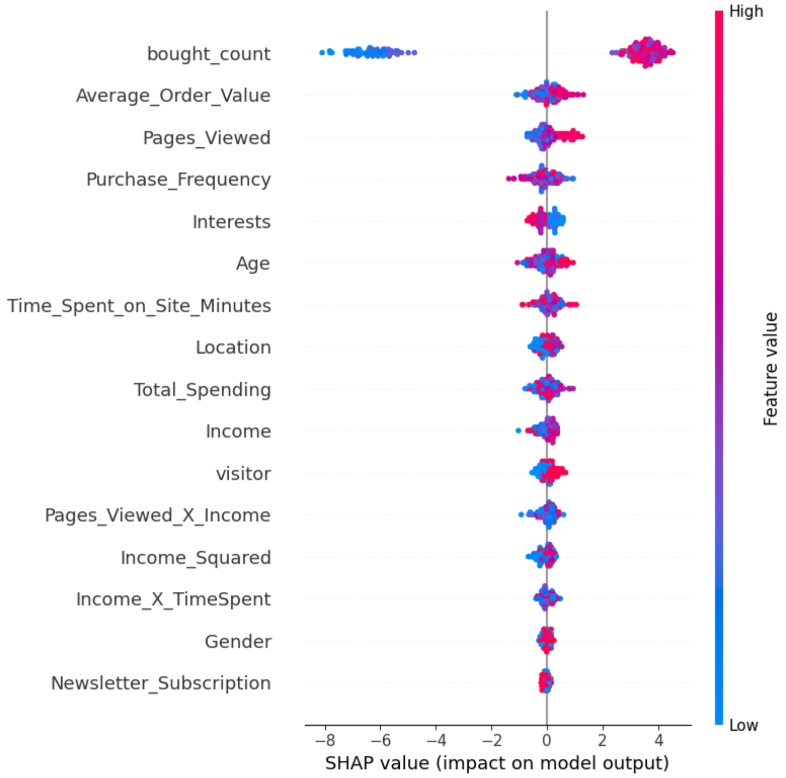

Figure 6.5: CatBoost SHAP value

6.2.1 Model Interpretation using SHAP

SHAP(SHapley Addictive Explanations) was used to analyze features for both XGBoost and CatBoost to have a better understanding of how the models have their predictions. In both models the bought count was the most prominent feature which analyzed the past purchase pattern of the user that is the most important for predicting future purchase patterns. In the XGBoost model by SHAP values the second important features were Income and Purchase Frequency. This shows the while figuring an user's purchasing habit XGBoost analyzed both financial capability along with past purchases. In addition to that the direction and spread of SHAP values shows that the predictions are constantly in the positive class by greater values of these features. On the other hand, for the CatBoost model the average order value and pages viewed were second most important features rather than Income which indicates that the deeper emphasis was on the quality of every transaction and the range of user interaction. Ad CatBoost has the automatic feature interaction management and also strong categorical feature handling that's why this was more successful to make use of behavioral signs. In case of both models the SHAP values for features like Newsletter Subscription, gender and other low rank features pooling around zero. By this it is understood that these features are found in the dataset but their predictive values are very low. As catboost was able to catch on more complex details that's why it had a better overall performance.

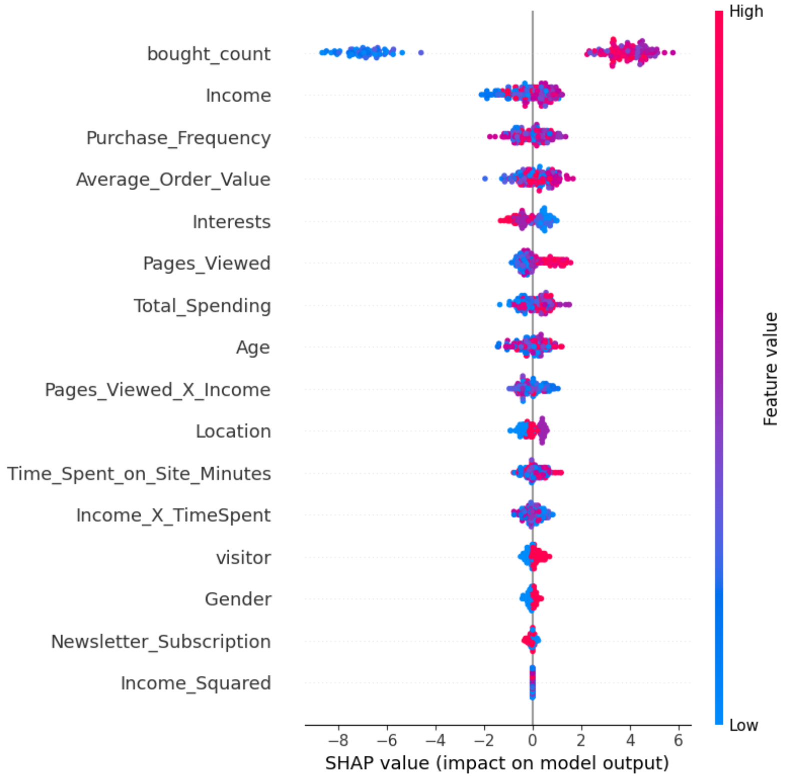

Figure 6.6: XGBoost SHAP value

6.2.2 Comparative Interpretation of SHAP Results: XGBoost vs. CatBoost

Both XGBoost and CatBoost achieved high performance but with the help of SHAP analysis the subtle variation of the model prioritizing features can be seen. Both models have bought count as the most predictive feature which shows that the models are prioritizing the purchase pattern as a reliable indicator for future purchase information. Apart from this agreement the models' feature sensitivity and prioritization varies from one another. XGBoost prioritized features like income and purchase frequency which indicates that it mostly depended on the user's financial capability more. On the other hand in CatBoost the income feature is seen to be ranked comparatively lower while they prioritized features like average order value and pages viewed which represent product engagement and transactional aka spending patterns. The reason behind this is that CatBoost is better at handling category and interaction datas which helps it to recognize the behavioral pattern which XGBoost may have missed. In addition to that both models gave less emphasis on features like gender and newsletter subscription, their SHAP values are close to zero. This agreement between the models indicates that these features don't really contribute much to the accuracy of predictions. But CatBoost handled the mid ranked features like age and time spent on site minutes in a balanced way where in XGBoost the effect of these features dropped drastically after the top five. In conclusion, even though both models are very similar in detecting behavioral patterns, CatBoost has more diffused and distinct features impact. As it takes a wider variant of important user interaction when making decisions so that could be the reason why it performs better.

Chapter 7

System Implementation and Interface Design

The e-commerce web application was created using MERN stack which includes Express.js, React.js, Node.js and MongoDB. The MERN stack technology is popular for creating efficient and maintainable full-stack online applications which includes precise separation of frontend and backend components. Along with managing a huge number of user and product data and integrating a personalized recommendation system, this system was build to provide a seamless user experience for both administrators and users. For the frontend interface of this system React.js was used providing a component based responsive design that supports the organizing states and dynamic presentation. Axios which is a promise-based HTTP client manages the API calls to the backend server and to the dedicated ML microservice which generally makes the recommendations in the frontend. Functions like product browsing, cart management, personalized product recommendations, order tracking, search and filtering are the primary user facing functions. In the homepage a feed of recommended products are displayed based on the user's past purchase and browsing history and the product pages give relevant recommendations for example; "you may also like". Also users can manage their profile, see their order history, create a wishlist along with login and register. The backend was created using Node.js and Express.js that helps to form a strong RESTful API service which manages money processing, product administration, user authentication and implementation recommendations. The main database MongoDB creates models for user, products, order

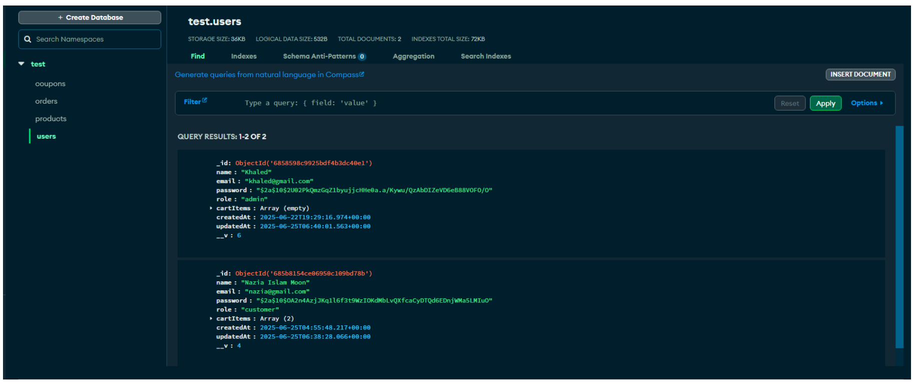

Figure 7.1: Products Collection View

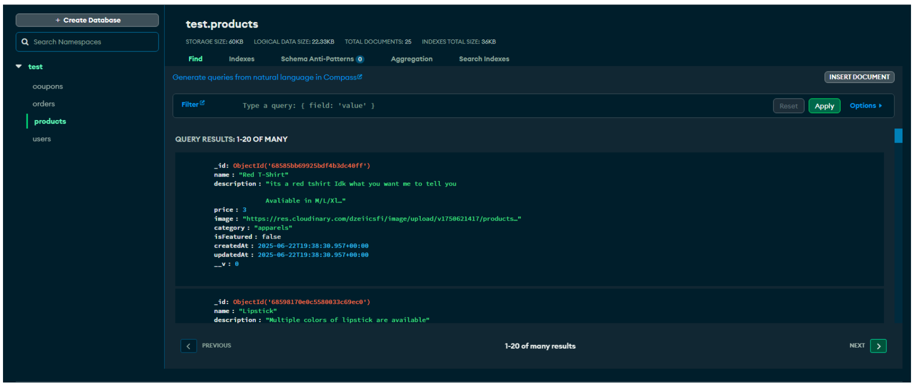

Figure 7.2: Coupons Collection View

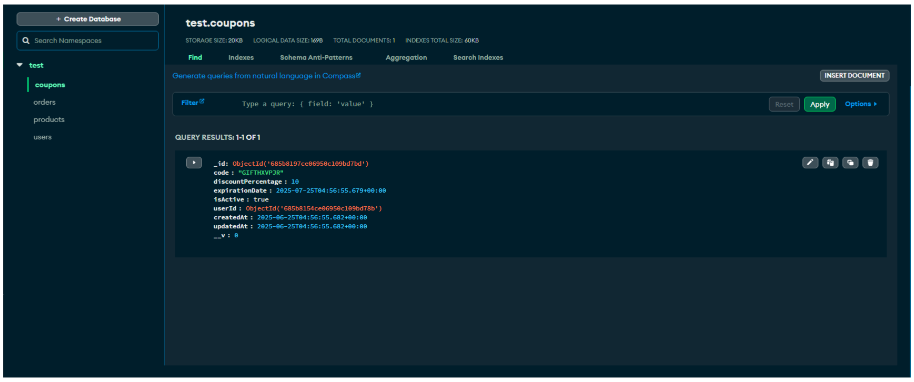

Figure 7.3: Users Collection View

review and cart items using Mongoose. To authenticate and authorize users JSON Web Tokens also known as JWTs were used in the backend. During the login or register process a token is generated and saved the client side providing a safe access to secured pathways for example: checkout, profile editing and cart actions. And every sensitive routes are secured with middleware which confirms the authenticity of the tokens and passwords safely saved in hashed form using bcrypt. There is admin panel which handles user interactions and the system's goods. Only the authorized admin user can access this panel which allows them the adding, deleting and updating of products along with the photos, prices, categories and stock level. In addition to that it allows managers to keep track of user activity, track orders and examine important performance indicators via a visible dashboard. Tables and charts were used to show product and order statics, admins can make decisions based on sales trends and user behavior. Redis has been used as a cathing layer which enhances the application speed and responsiveness specially when the user load is high. By using Redis to cache frequently requested data like product listing, user sessions and recommendations results the frequent database queries and delay. This recommendation system was created using a separate microservices Python and Flask. With models like XGBoost, CatBoost and MLP collaborative filtering method has been developed. These models were trained using the previous user interaction data like: product views, clicks, transactions and rating. Then the models provided personalized recommendations based on these behaviors which then sent to frontend through Axios-enabled API calls. The ML service can be upgraded or improved separately from the primary application thanks to its flexible structure. One of the popular

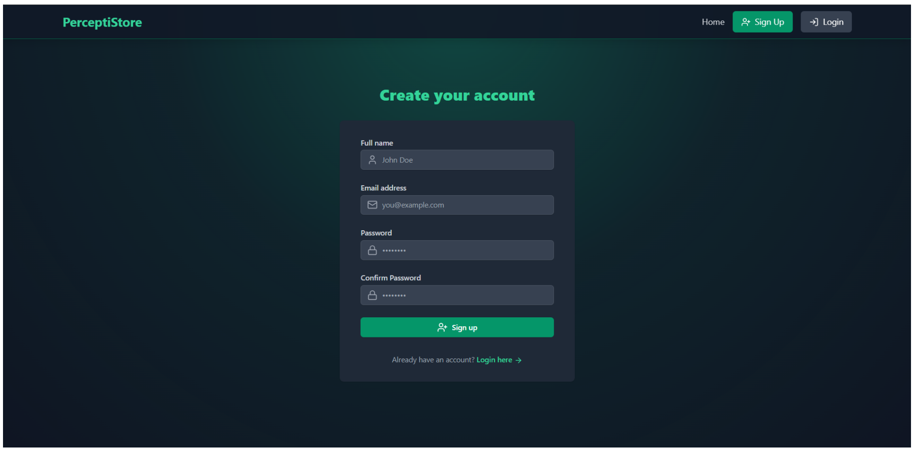

Figure 7.4: User Registration Page

payment gateway Stripe which guarantees PCI-DSS compliance has been used in the system's secure payment. During the checkout process the users are taken to a secure Stripe interface in order to enter their payment details. In this system Stripe leads order processing, updates order status and confirms the transaction by activating a link on the website for successful payment. As the system does not keep the raw payment data so danger is decreased and current security standards are upheld. For a safer communication the entire system is developed using HTTPS which guarantees encrypted data transfer between the user and server. To keep the sensitive information secured .env files were used. And to make sure that only the admins are able to carry out specific tasks like: managing products or gaining access to admin routes RBAC was implemented in the backend. This webapp is developed to be both accessible and maintainable. By separating different responsibilities between the frontend, backend and ML parts simple debugging, flexible development, and adaptable implementation were possible. A simple and intelligent e-commerce experience customized to user preferences is produced by combining efficient data access with Redis, personalized recommendations, secure user authentication, and a real-time payment system. The given screenshots are the structure and sample data of a MongoDB database named test, in which there is collections for products, coupons, and users. In the Fig 7.1: Products Collection View, there are data containing product details such as name, description, price, image URL, category, availability, and timestamps like createdAt and updatedAt. This will helps identify and manage different types of products in my system. Then if we look at Fig 7.2: Coupons Collection View which shows a single coupon document with fields like code, discountPercentage, expirationDate, and userId, that is dedicated to a specific user and can be used for discounts. The isActive field ensures only valid coupons are applied. Lastly, in the Fig 7.3: Users Collection View displays user-related data including name, email, encrypted password, role (admin or customer), and an array of cartItems. This data is essential for user authentication, authorization, and tracking shopping activity. Each document also records createdAt and updatedAt timestamps to help with data lifecycle management. This dataset collections altogether support the e-commerce system PerceptiStore.

The given screenshots are the structure and sample data of a MongoDB database named test, in which there is collections for products, coupons, and users. In the Fig 7.1: Products Collection View, there are data containing product details such

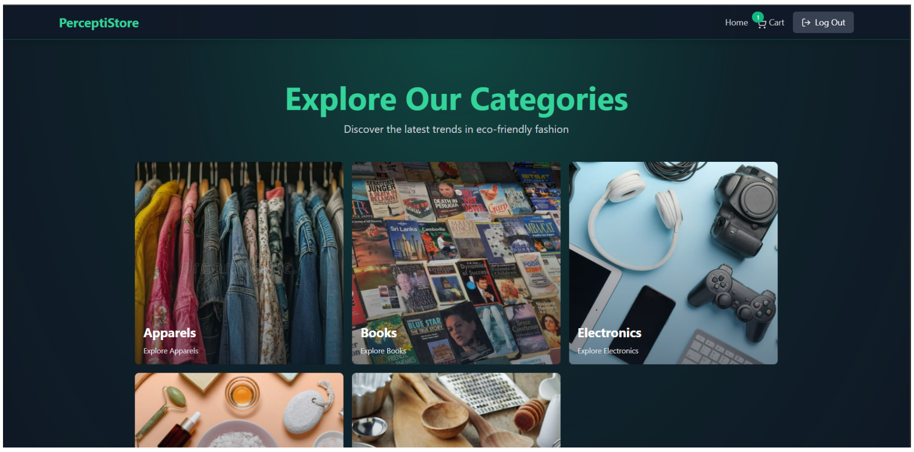

Figure 7.5: Homepage Displaying Product Categories

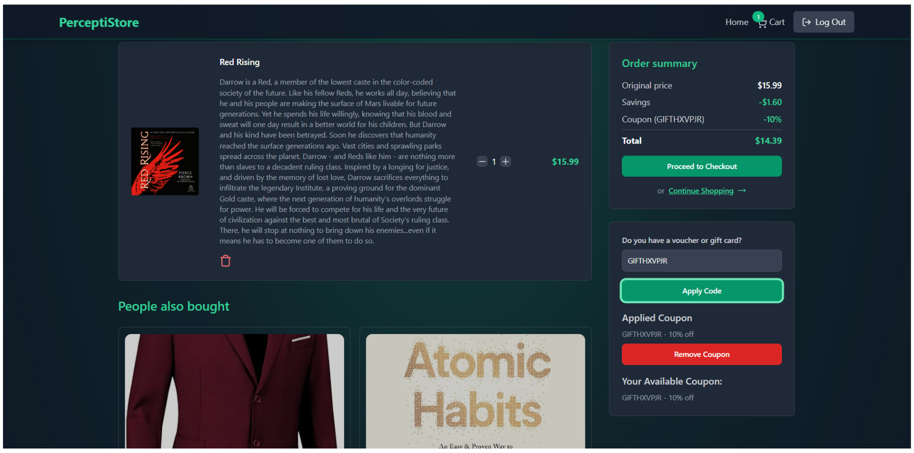

Figure 7.6: Shopping Cart Page

as name, description, price, image URL, category, availability, and timestamps like createdAt and updatedAt. This will helps identify and manage different types of products in my system. Then if we look at Fig 7.2: Coupons Collection View which shows a single coupon document with fields like code, discountPercentage, expirationDate, and userId, that is dedicated to a specific user and can be used for discounts. The isActive field ensures only valid coupons are applied. Lastly, in the Fig 7.3: Users Collection View displays user-related data including name, email, encrypted password, role (admin or customer), and an array of cartItems. This data is essential for user authentication, authorization, and tracking shopping activity. Each document also records createdAt and updatedAt timestamps to help with data lifecycle management. This dataset collections altogether support the e-commerce system PerceptiStore. The given figures represents the core features of the PerceptiStore e-commerce system, illustrating both user-facing and administrative functions. By looking at the Figure 7.4 the user registration page is displayed, where new users can create an account by entering their full name, email, and password with a confirmation field to ensure security. After signing up, users are directed to the homepage. In the Figure 7.5 the homepage is shown, here products are organized into categories for example: Apparels, Books, and Electronics, allows users to explore items easily. The navigation bar includes access to the cart, the log-in/logout page along with the admin dashboard. The Figure 7.6 shows the

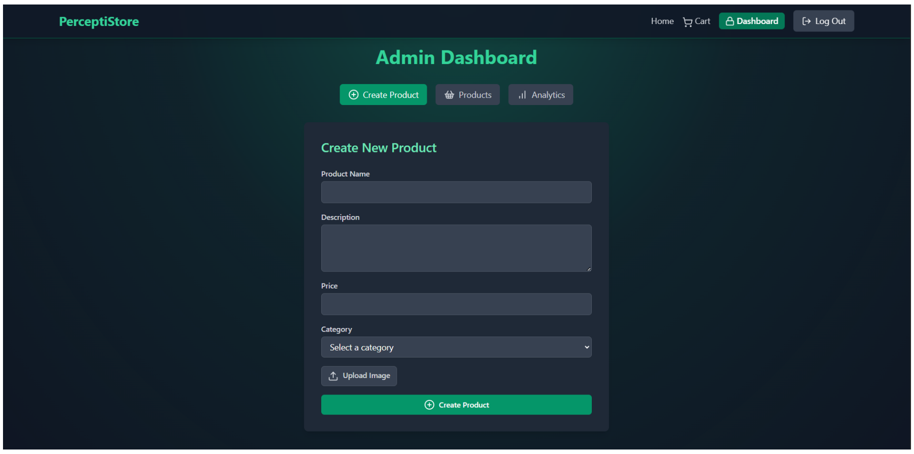

Figure 7.7: Admin Dashboard for Product Creation

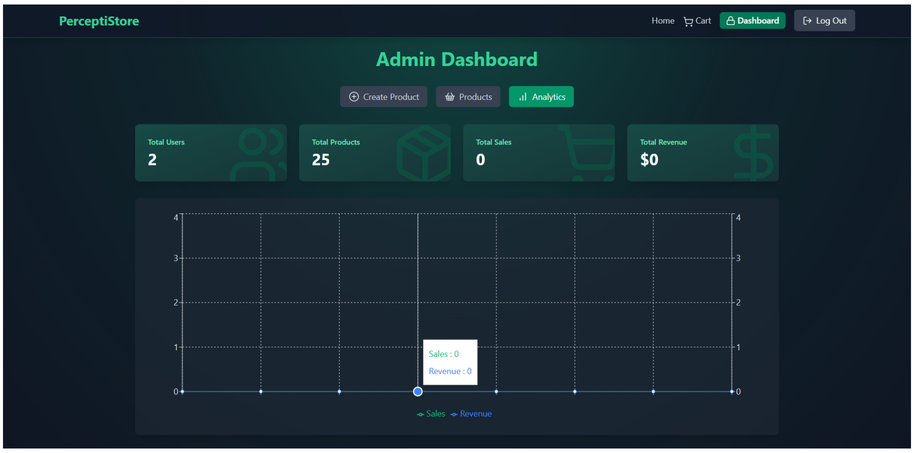

Figure 7.8: Admin Dashboard for Product and Sales Overview

shopping cart page, here a user has added the book called Red Rising, applied 10 percent discount coupon, and received a new price total of 14.39 dollar. This page also provides options to adjust quantities, apply vouchers, and proceed to checkout, there's also a recommendation section titled "People also bought". The Figure 7.7 is for the admin product creation dashboard, where adminiss can input product details, select categories, and upload images. Lastly Figure 7.8 shows the admin analytics dashboard, featuring total users, products, sales, and revenue metrics, also a performance graph and a shortcut buttons for managing products. These interfaces collectively demonstrate the usability and functionality of the PerceptiStore platform for both end users and admins.

Chapter 8

Conclusion

This study intended to provide AI-powered personalized shopping recommendations by using ML models based on collaborative filtering. The highly personalized system showed how ML and prediction can completely change online shopping. Six algorithms were tested, which are: SVM, Logistic Regression, Random Forest, XGBoost, CatBoost and MLP. The CatBoost's result showed that it performed way better than all the other algorithms in terms of accuracy, precision, recall and F1-score. CatBoost's ability to manage categorical data and avoid overfitting help to achieve such results which made it an excellent choice for used e-commerce dataset. Additionally the use of SHAP analysis helped to understand the interpretation of model choices, demonstrating the importance of features such as bought count, income and interaction history for predicting future purchases. Interestingly the demographic features like gender and newsletter didn't contribute much in prediction. Despite being organized this dataset lacks contextual signals like review or clickstream data and is only focused on purchase activity because of that it lacks diversity. Also without regular training and or online learning this system cannot react in real time as the existing system is designed using historical data. So in conclusion, despite these limitations this research prioritized protecting data privacy and using AI ethically to validate the effectiveness of AI in increasing user engagement. The PerceptiStore project shows a scalable and user-centric approach for future e-commerce development by successfully connecting the gap between AI personalization and safe, real-time recommendation delivery.

Bibliography

[1] A. Bielozorov, M. Bezbradica, and M. Helfert, "The role of user emotions for content personalization in e-commerce: Literature review," in *HCI in Business, Government and Organizations. eCommerce and Consumer Behavior: 6th International Conference, HCIBGO 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21*, Springer, 2019, pp. 177–193.

[2] A. Nimbalkar and A. Berad, "The increasing importance of ai applications in e-commerce," *Vidyabharati International Interdisciplinary Research Journal*, vol. 13, no. 1, pp. 388–391, 2021.

[3] S. Smith, "Leveraging ai and big data for smart supply chain management: Enhancing ecommerce efficiency," 2021.

[4] A. K. Kashyap, I. Sahu, and A. Kumar, "Artificial intelligence and its applications in e-commerce–a review analysis and research agenda," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 24, pp. 7347–7365, 2022.

[5] S. K. Bok, "Enhancing user experience in e-commerce through personalization algorithms," 2023.

[6] Y. Amil *et al.*, "The impact of ai-driven personalization tools on privacy concerns and consumer trust in e-commerce," 2024.

[7] M. A. Raji, H. B. Olodo, T. T. Oke, W. A. Addy, O. C. Ofodile, and A. T. Oyewole, "E-commerce and consumer behavior: A review of ai-powered personalization and market trends," *GSC Advanced Research and Reviews*, vol. 18, no. 3, pp. 066–077, 2024.

[8] K. Bartwal, "E-commerce product recommendation collaborative,"

[9] R. K. Paul and A. K. Jana, "A comprehensive framework for integrating ai and machine learning in personalization and ad targeting within e-commerce,"

[10] N. Singh, "Future of artificial intelligence in e-commerce,"