in collaboration with

Softwarica | Coventry University

Analysis of customer engagement data and employ AI-driven solutions such as machine learning and predictive analytics to predict retail purchase behaviour, improve decision-making, and enhance operational efficiency and customer satisfaction.

Submitter By
Karan Bohara

Coventry Id
13703532

Submitted To
Manoj Shrestha

Concept Page

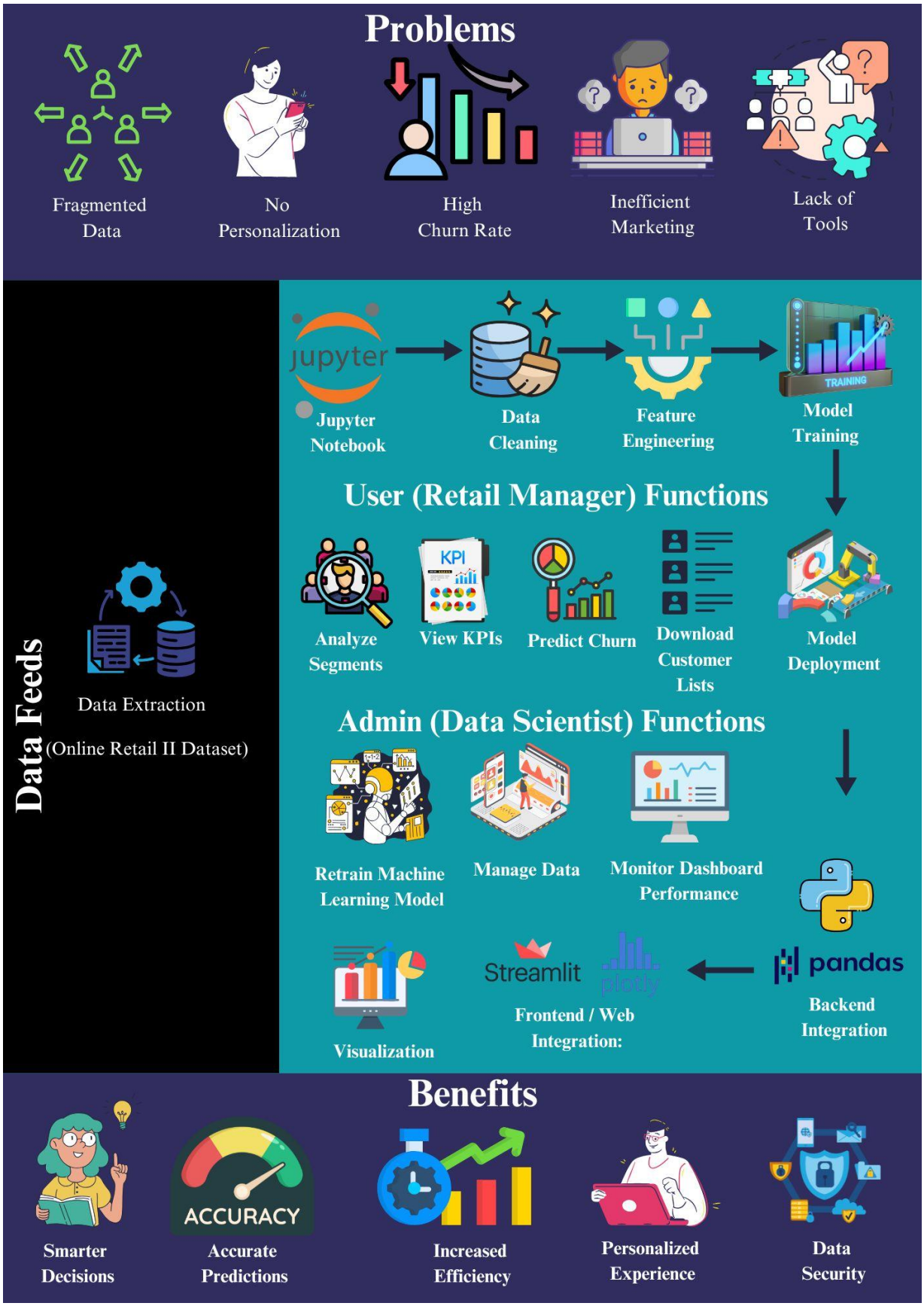


Figure 1: Concept Page

Ethical Approval Certificate

Project Title

Prediction of Retail Purchase Behavior through Customer Engagement Data

Record of Approval

Principal Investigator

I request an ethics peer review and confirm that I have answered all relevant questions in this checklist honestly.	×
I confirm that I will carry out the project in the ways described in this checklist. I will immediately suspend research and request new ethical approval if the project subsequently changes the information I have given in this checklist.	×
I confirm that I, and all members of my research team (if any), have read and agreed to abide by the Code of Research Ethics issued by the relevant national learned society.	×
I confirm that I, and all members of my research team (if any), have read and agreed to abide by the University's Research Ethics, Governance and Integrity Framework.	×
I understand that I cannot begin my research until this ethics application has been approved.	×

Name: Karan Bohara

Date:

Student's Supervisor (if applicable)

I have read this checklist and confirm that it covers all the ethical issues raised by this project fully and frankly.I also confirm that these issues have been discussed with the student and will continue to be reviewed in the course of supervision.

Name: Manoj Shrestha

Date:

Model Card

AI MODEL CARD

Institute of Analytics
The Global Body for Analytics

Model Name

Retail Customer Churn Prediction & Segmentation System

Model date and version

The model was trained on August 11, 2025, using a static dataset and has not been updated since. All stakeholders involved in the use of this model will be formally notified of any future versions or retraining efforts.

Version 1.0

Overview Model type

This model is a pre-trained binary classification system designed to predict customer churn in a retail context. It utilizes a Logistic Regression algorithm, which calculates the probability of a customer churning by modelling the relationship between key behavioral features and the churn outcome. The core input features are derived from customer transaction history, primarily focusing on Recency, Frequency, and Monetary (RFM) value metrics. The system processes historical data to identify distinct behavioral patterns, turning raw data into actionable insights for strategic customer retention initiatives.

Questions or Comments

Please send any questions to: 220404@softwarica.edu.np

Primary Intended Users

The primary intended users are data-driven professionals within an e-commerce or retail organization who are responsible for customer loyalty and growth.

Retail and Marketing Managers can use the model's churn probability scores to segment customers and design targeted retention campaigns, such as special offers or personalized communication for high-risk customers.

Business Analysts can leverage the model to analyze and understand the key drivers of churn within the customer base, informing broader business strategy and identifying areas for service improvement.

Customer Relationship Managers can use the insights to proactively engage with at-risk customers, aiming to resolve issues and strengthen loyalty before they decide to leave.

Out Of Scope Uses

This model is intended for internal strategic support and is explicitly out-of-scope for any automated, high-stakes, or customer-facing decisions. It should not be used to implement differential pricing, deny service, or perform credit scoring. Its predictions should serve as an input for human decision-making, not as a replacement for it.

Limitations

The model's predictions are subject to the constraints of its training data and design. It is crucial for users to be aware of these limitations to avoid misinterpretation of its outputs.

Historical Data Scope: The model was trained on data that may not reflect modern e-commerce trends, potentially reduce its accuracy when apply to current market conditions.

Market and Cultural Specificity: The patterns learned from the training dataset may not generalize well to customers in different international markets or cultural contexts.

Model Training Bias: The model could have learned biases present in the original dataset. If historical marketing efforts were skewed towards certain purchasing behaviours, the model might inadvertently perpetuate those patterns.

Lack of Real-Time Capability: The system is designed for batch processing of historical data, not for real-time analysis. Its insights are best suited for strategic planning rather than on-session interventions.

Metrics

The model was evaluated on its ability to classify customers as either "Churned" (Class 1) or "Active" (Class 0). Its performance was measured using Accuracy, Precision, and Recall to provide a realistic view of its effectiveness in a business context.

Training and evaluation data

The model was trained on a dataset of 4,312 samples, split into a training set of 3,018 and a held-out test set of 1,294 samples. For inquiries about the specific data source or preprocessing steps, please contact the data team at 220404@softwarica.edu.np.

Quantitative Analysis

The model's performance reflects the common challenge of predicting an imbalanced outcome, where the number of non-churning customers is much larger than the number of churning customers.

Accuracy: The model achieved an overall accuracy of 87.09%. While solid, this figure is heavily influenced by the model's strong performance on the majority class (non-churners). Therefore, precision and recall for the minority class (churners) are more indicative of its business value.

Precision (for "Churned" class): The model achieved a precision of 0.67. This means that when the model identifies a customer as likely to churn, it is correct 67% of the time. For

business use, this implies that retention campaigns targeting this group will be reasonably well-focused, though approximately one-third of the effort would be directed at customers who were not going to churn.

Recall (for "Churned" class): The model achieved a recall of 0.65. This is a critical metric, indicating that the model successfully identifies 65% of all customers who actually end up churning. While this provides a significant opportunity to save at-risk customers, it also means the model fails to identify the other 35%, representing a notable source of potential missed revenue.

Ethical Considerations

The model was developed within an "Ethics by Design" framework. It does not use any Personally Identifiable Information (PII) and relies on behavioral data to mitigate direct demographic bias. However, the performance limitations have ethical implications. Given that the model's recall is 65%, relying on it exclusively for retention efforts would systematically overlook over a third of churning customers. Therefore, it should be used as one tool among many, complementing other customer feedback and engagement strategies to ensure fair opportunities for all customers to be retained.

Feedback

Continuous improvement relies on user feedback. Users are encouraged to report on the performance, usability, and practical impact of the model. Please submit any concerns, issues, or suggestions to 220404@softwarica.edu.np.

Additional Notes and Any Other Relevant Factors

This model was created as a proof-of-concept for an undergraduate thesis project supervised by Mr. Manoj Shrestha. Its primary purpose is to demonstrate the practical application and considerations of building a predictive AI model in a commercial retail context.

Abstract

In retail, knowing and predicting customer behaviour is essential for long-term success. This thesis introduces a practical framework for studying customer engagement and forecasting churn with AI-based methods. Using transaction data from an online retailer, the process begins with RFM (Recency, Frequency, Monetary) analysis to group customers into clear segments such as "Champions" and "At-Risk". The main focus is a churn prediction model built with Logistic Regression, reaching an accuracy of 87.09% in spotting customers likely to leave. The model produces a churn probability score for each active customer, helping businesses act early to retain them. Results are presented in an interactive dashboard created with Python and Streamlit, offering insights on customer segments, overall performance, and key churn factors. This work delivers a complete data science solution that supports better decisions, sharper marketing, and stronger customer loyalty.

Acknowledgement

I wish to express my sincere appreciation to my supervisor, Mr. Manoj Shrestha, for his invaluable guidance and support throughout this research project. His profound knowledge, insightful suggestions, and incisive feedback were instrumental in shaping the direction of this thesis and bringing it to a successful conclusion. I am deeply grateful for his unwavering patience and continuous encouragement. I would also like to extend my heartfelt appreciation to my friends and teachers who offered significant assistance and direction during the creation of this report. Their constant support and understanding have been a significant source of strength, helping me to maintain focus and overcome the challenges encountered during this project.

Keywords

Figure 2: Keywords

Table of Contents

Table of Figures

Introduction

Retail is the process of selling goods and services directly to consumers for their personal use, and it is one of the largest and most important sectors in the global economy. It connects manufacturers and suppliers with end users and can take place through various channels such as physical stores, open markets, online marketplaces, and mobile applications. Traditionally, retail was conducted in physical locations, where customers could see, touch, and purchase products in person. Over time, advancements in technology, improvements in logistics, and changing consumer lifestyles have reshaped the industry. Today, customers can browse, compare, and purchase products from anywhere with an internet connection. Online retail has grown rapidly, supported by mobile devices, secure payment systems, and fast delivery services. Many businesses now use an omnichannel approach, combining physical and digital experiences so that customers can switch between them seamlessly. A customer might view products on a website, check their availability in a nearby store, visit the store to inspect them, and then make the purchase through an app. Retail has also evolved in terms of competition. Price and product quality remain important, but convenience, speed, and personalisation are now critical factors influencing buying decisions. Consumers expect product recommendations that match their interests, flexible delivery options, and smooth customer service across all touchpoints. Every interaction, from browsing an online store to using a loyalty card in-store, produces data that reflects customer preferences, buying patterns, and engagement levels. In theory, this data allows businesses to understand their customers better, anticipate needs, and create tailored strategies. In practice, however, using this information effectively is not always straightforward.

A major challenge for retailers is that the amount of customer data has grown too large, varied, and complex to be effectively analyzed using traditional methods. Information is often scattered across multiple systems such as sales platforms, marketing tools, inventory management software, and customer service databases without any unified framework to connect them. This separation makes it extremely difficult to create a single, accurate, and complete view of each customer. Without this integration, valuable insights remain hidden, and businesses are unable to fully understand the entire customer journey from first interaction to repeat purchases. As a result, opportunities for upselling, cross-selling, and personalized engagement are frequently missed. Marketing campaigns may be poorly targeted, reaching the wrong audience or delivering irrelevant messages, while inventory levels may be mismanaged, leading to overstocking unwanted items or running out of high-demand products. Customer retention efforts are often reactive rather than proactive, applied only after customers have already disengaged or stopped buying. These inefficiencies not only increase operational costs but also contribute to lost sales, declining engagement, and high customer churn rates. Over time, this directly impacts profitability and weakens a retailer's ability to remain competitive in a fast-changing, data-driven market.

AI-powered technologies, including machine learning and predictive analytics, offer powerful solutions to the challenges faced by retailers by processing large and diverse datasets to generate meaningful, actionable insights. Machine learning models are capable of detecting complex patterns in customer behaviour that may not be obvious through traditional analysis.

These models can calculate churn probability scores for individual customers, enabling businesses to identify those at risk of leaving and recommend targeted actions to retain them effectively. Additionally, predictive analytics improves demand forecasting accuracy, helping retailers manage inventory more efficiently by reducing both overstock and stockouts. These tools also support personalised marketing campaigns at scale, allowing businesses to tailor offers and communications to individual customer preferences and behaviours. By adopting these advanced technologies, retailers can shift from relying on reactive problem-solving to employing proactive, data-driven strategies that target the right customers with the right offers at the most effective times, ultimately enhancing customer engagement, loyalty, and profitability.

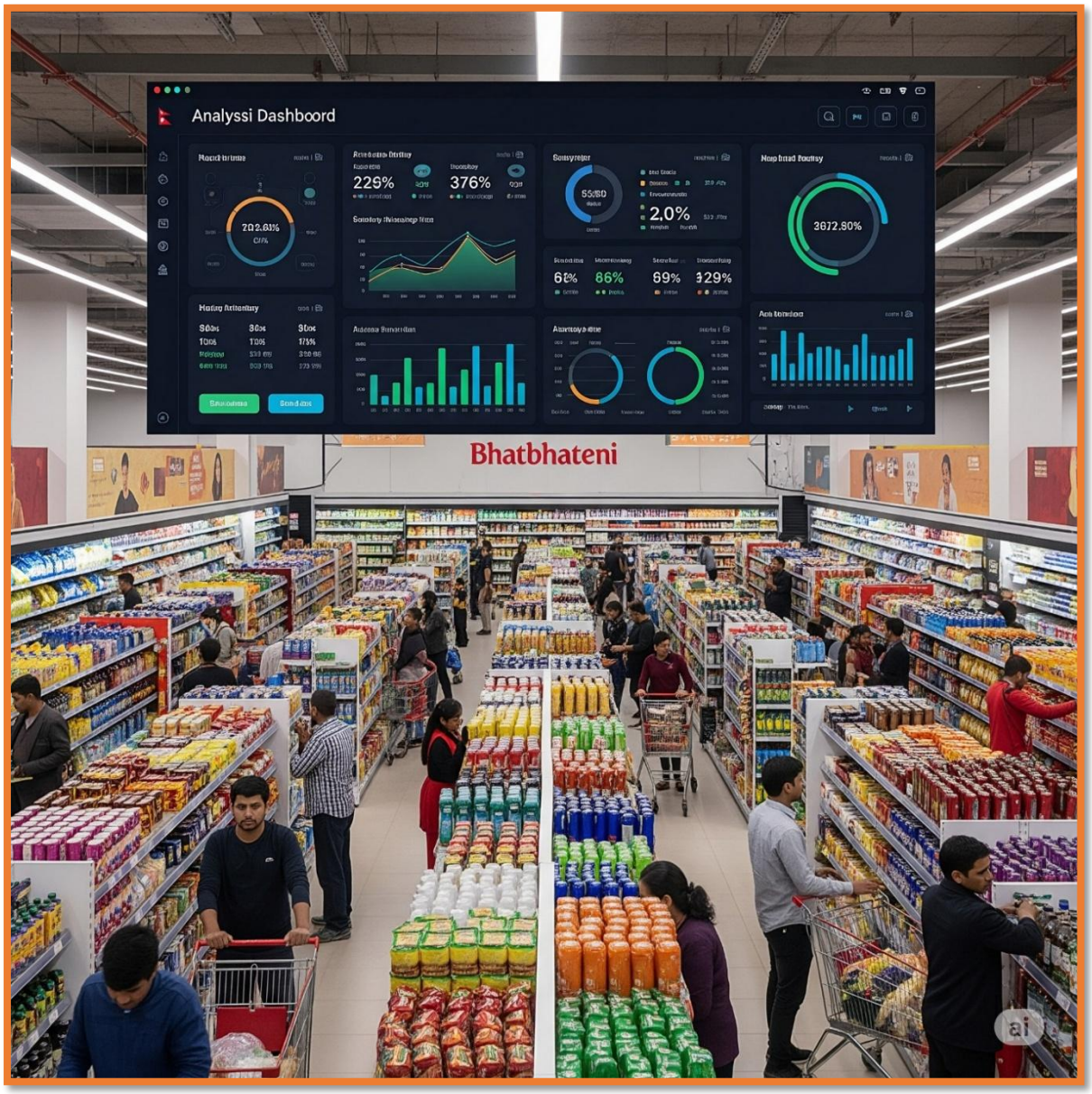


Figure 3: Introduction to Retail

Some of the world's leading retailers have already demonstrated the powerful impact of these advanced methods, proving that data-driven strategies can deliver measurable results at scale.

Amazon, for example, uses sophisticated recommendation systems powered by machine learning to provide personalised product suggestions based on browsing history, purchase patterns, and real-time behaviour. This not only increases customer engagement but also drives higher conversion rates and repeat purchases. Walmart leverages predictive analytics to optimise inventory management across thousands of stores worldwide, ensuring that high-demand products are always in stock while reducing waste and cutting excess storage costs. Alibaba takes advantage of real-time data analysis to segment customers more precisely, tailor marketing campaigns to individual preferences, and forecast demand with remarkable accuracy, even during massive shopping events like Singles' Day. These companies have built highly adaptive systems that can process vast amounts of data in real time, enabling faster, smarter, and more accurate decision-making. By doing so, they have improved operational efficiency, reduced costs, and strengthened long-term customer loyalty. Their success sets new benchmarks for the retail industry as a whole, illustrating how the effective integration of AI, analytics, and data-driven strategies can transform traditional business models, create competitive advantages, and redefine customer expectations on a global scale.

Aim

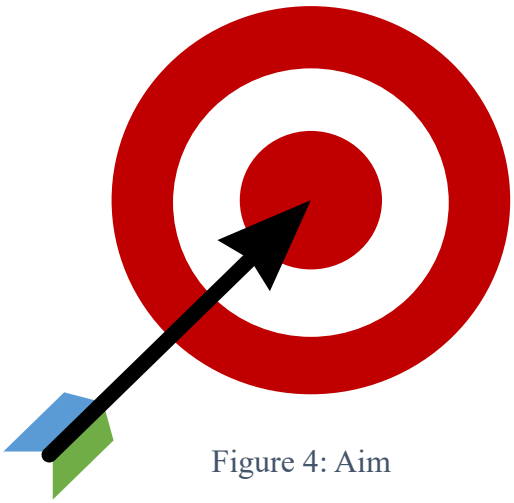


Figure 4: Aim

To analyse customer engagement data and employ AI-driven solutions such as machine learning and predictive analytics to predict retail purchase behaviour, improve decision-making, and enhance operational efficiency and customer satisfaction.

Objectives

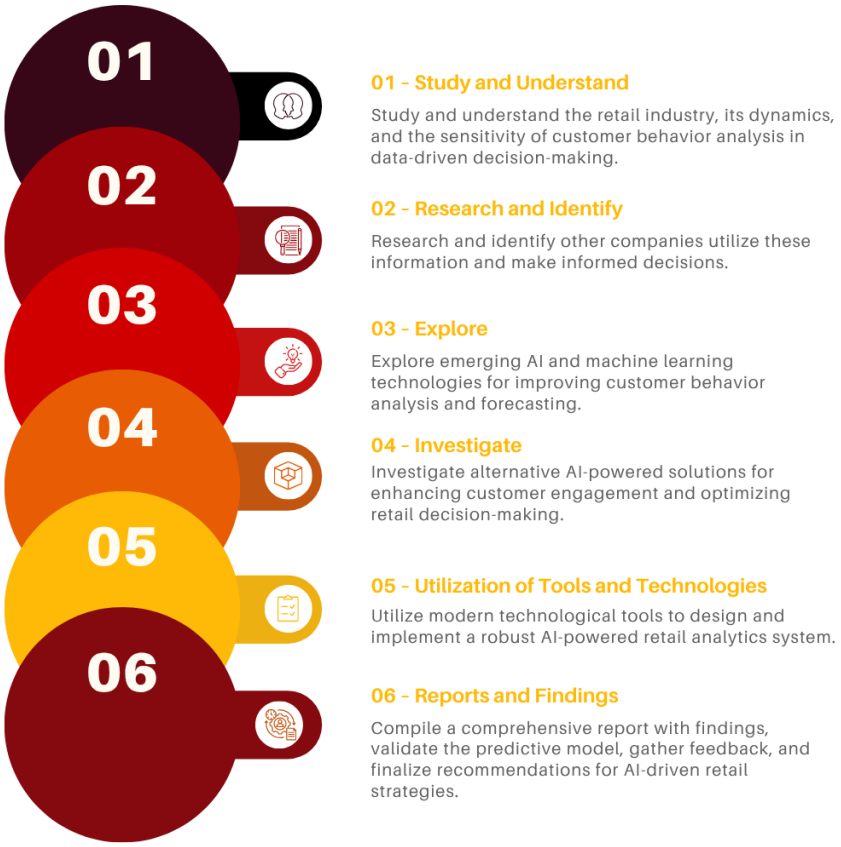


Figure 5: Objective

Justification

The retail industry faces increasingly complex challenges driven by fragmented and inconsistent data spread across multiple systems, including online stores, physical outlets, loyalty programs, and marketing platforms. This scattered information creates major blind spots, preventing retailers from truly understanding their customers and their needs. As a result, user dissatisfaction grows when customers are met with irrelevant product recommendations, generic promotions, or poor service experiences. This dissatisfaction erodes trust and loyalty, making it harder to retain customers over time. The inability to accurately identify and engage high-value customers leads to missed opportunities for upselling, cross-selling, and building long-term relationships that could increase profitability. When these opportunities slip away, retailers inevitably experience lost revenue, often caused by poor targeting strategies, inventory mismanagement, and inaccurate demand forecasting. Even small errors such as stocking items customers are not interested in while running out of high-demand products can have a significant financial impact. Underlying these issues are data quality problems, including duplicate records, incomplete profiles, and outdated information, all of which reduce the reliability of business intelligence and analytics. Without a strong data foundation, decision-

Figure 6: Justification

making becomes guesswork rather than strategy. This problem is compounded by reduced engagement, as customers become less responsive to offers, newsletters, and campaigns that fail to resonate with their personal preferences. Over time, these engagement gaps widen, making it even more difficult to reestablish meaningful connections with the audience.

These problems are amplified by fragmented data systems that prevent retailers from forming a complete view of the customer journey, from first interaction to purchase and beyond. When interactions across websites, mobile apps, and physical stores remain disconnected, personalization efforts suffer, and the customer experience feels inconsistent. This disjointed approach contributes directly to a high churn rate, as customers who feel overlooked or undervalued are quick to seek alternatives with competitors who appear more attentive and relevant. Retaining existing customers is far more cost-effective than acquiring new ones, yet without effective churn prediction and intervention strategies, many retailers are caught in a costly cycle of losing and replacing customers. At the same time, inefficient marketing drains valuable resources, as campaigns are often misdirected, reaching the wrong audience, delivering poorly timed messages, or relying on outdated assumptions about customer behavior. This not only wastes marketing budgets but also weakens brand credibility, as irrelevant or repetitive communication frustrates customers rather than engaging them. All eight of these problems user dissatisfaction, missed opportunities, lost revenue, data quality issues, reduced engagement, fragmented data, high churn rate, and inefficient marketing are deeply interconnected and often share the same root cause: the absence of unified, accurate, and actionable data. Overcoming them requires breaking down data silos, standardizing and cleansing data for accuracy, and leveraging advanced analytics to generate actionable insights. With a unified data strategy, retailers can deliver timely, personalized, and consistent customer experiences, reduce churn, improve engagement, and unlock new growth opportunities, ultimately transforming challenges into a foundation for sustainable success.

To address the challenges faced by retailers, this project delivers a robust AI-powered analytics framework designed to convert scattered, inconsistent, and incomplete retail data into precise and actionable insights. The workflow starts with research and analysis, where industry benchmarks, customer behavior studies, and performance metrics are examined to identify improvement areas and guide solution design. This is followed by data cleaning and preparation, ensuring that every dataset whether sourced from online transactions, in-store purchases, loyalty programs, or marketing systems is accurate, consistent, and free of duplicates or missing values. Clean and reliable data serves as the foundation for meaningful analysis. Once prepared, the data undergoes RFM (Recency, Frequency, Monetary) feature engineering, a process that translates raw behavioral and transactional data into measurable indicators of customer value and activity. These RFM metrics help segment customers into meaningful groups, enabling targeted strategies for high-value retention, win-back campaigns, and cross-selling opportunities. Building on this, the system integrates machine learning models that not only predict customer churn but also uncover patterns in engagement, product preferences, and purchase timing. This predictive capability allows retailers to act before customers disengage, reducing the high costs associated with churn.

The framework's intelligence is further enhanced through data-driven marketing, enabling campaigns that are personalized, timely, and relevant to each customer segment identified by the models. This ensures marketing budgets are spent efficiently and yield a higher return on investment. To make these insights accessible to decision-makers at all levels, an interactive dashboard is developed, offering real-time views of customer behavior, sales trends, and campaign performance. This dashboard eliminates the need for deep technical expertise, empowering managers to make informed decisions quickly. Complementing this is data visualization, where complex datasets are transformed into intuitive charts, graphs, and interactive visual elements, making it easier to detect patterns, compare performance over time, and communicate findings across teams. A key feature is churning prediction, which uses advanced analytics to flag at-risk customers early, allowing for the implementation of personalized retention strategies before they leave for competitors. By uniting research, data preparation, feature engineering, machine learning integration, targeted marketing, and visual analytics, this solution equips retailers with a complete toolkit for modern retail management. The result is a system capable of reducing churn, increasing engagement, improving marketing precision, and delivering a seamless, personalized experience that strengthens customer loyalty while driving sustainable business growth in an increasingly competitive retail landscape.

Research Questions

Figure 7: Research Questions

Ethical Considerations

This project avoids the use of sensitive demographic details such as ethnicity, gender, or income level, and their influence on the system's outputs was not explored. Ethical considerations were treated as a priority, especially under legal frameworks like the General Data Protection Regulation (GDPR), which governs how personal data can be collected, stored, and used. No real or private customer data was included at any stage; instead, the research relied entirely on a publicly available, anonymized dataset from the UCI Machine Learning Repository that is intended for academic purposes.

Figure 8: Ethical Considerations

This reduces the potential for bias or discriminatory outcomes, ensuring that model predictions are based on customer activity rather than personal characteristics. By focusing on behavioural trends, the system is better aligned with the business goal of enhancing customer satisfaction through timely and relevant engagement strategies. This approach ensures that all insights are both legally compliant and ethically sound, allowing retailers to gain value from analytics while safeguarding the privacy, fairness, and dignity of every customer involved in the analysis.

Scope

Customer churn prediction in this project is limited by the available transactional data, which may not encompass all the factors influencing a customer's decision to leave. While the machine learning techniques employed can identify powerful patterns, they are confined to the specific behavioural variables engineered for this study, namely Recency, Frequency, and Monetary Value (RFM). Key external factors such as customer satisfaction scores, product quality, competitor actions, or customer service interactions are not included in the dataset, which potentially impacts the model's ability to explain all the underlying drivers of churn. Furthermore, the model's performance is constrained by the representativeness of the data, which originates from a single UK-based online retailer and covers only a one-year period (2009-2010); therefore, the findings may not fully generalize to different retail sectors, geographical regions, or economic climates.

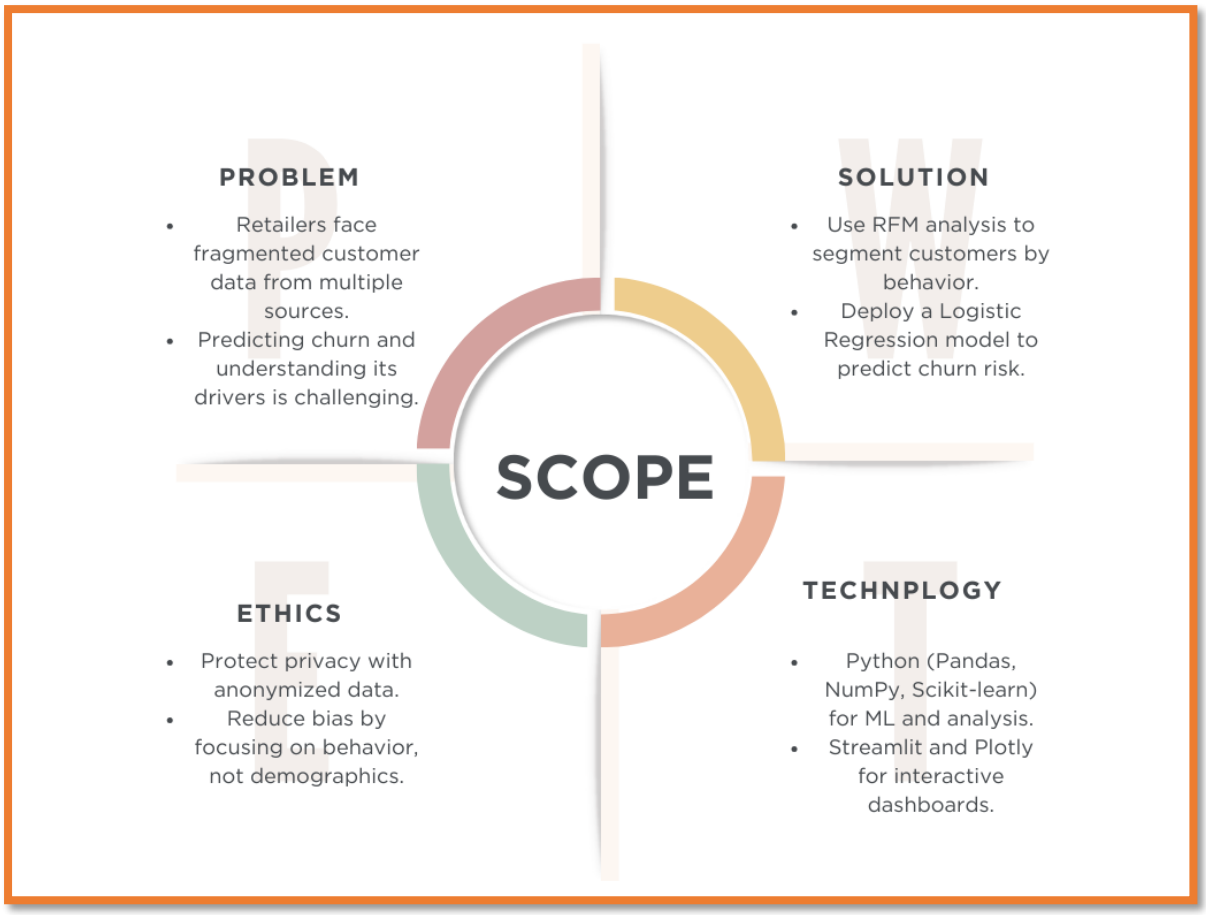


Figure 9: Scope

Literature Review

Desk Based Agile Strategy Research methodology

This capstone project adopted a structured, desk-based research methodology, which was deemed the most appropriate approach given the project's reliance on the analysis of pre-existing datasets and established literature.

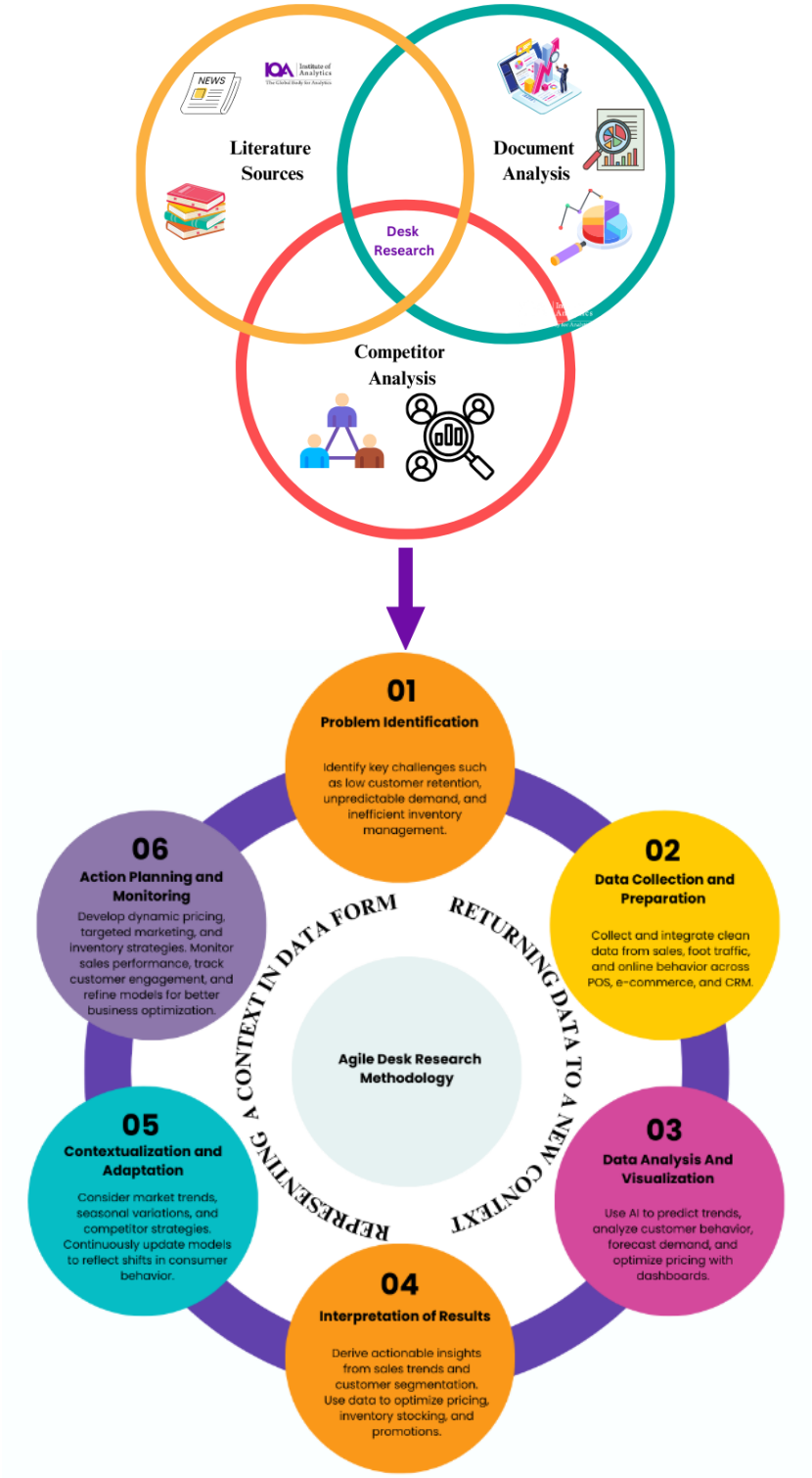


Figure 10: Desk Based Agile Strategy Research methodology

This strategy facilitated a deep investigation into the subject matter without the need for primary data collection, allowing resources to be focused on data processing, modelling, and interpretation. To provide a robust and industry-recognized structure to the research process, the project was guided by the Institute of Analytics (IoA) Competency Framework. This framework was instrumental not only in defining the initial objectives but also in ensuring that the project's execution demonstrated key competencies in data stewardship, advanced analytics, and strategic decision-making. The project's lifecycle was executed through a series of well-defined stages, closely mirroring established data science project models like the Cross-Industry Standard Process for Data Mining (CRISP-DM). The initial phase involved a comprehensive problem identification and business understanding exercise, which was informed by an extensive literature review. This review helped to contextualize the problem and identify existing techniques and gaps in current research. Following this, the project moved into the core data-centric phases: data understanding and data preparation. This involved sourcing the relevant datasets, performing exploratory data analysis (EDA) to uncover initial patterns and anomalies, and conducting intensive data cleaning, transformation, and feature engineering to prepare the data for modelling.

The subsequent modelling phase focused on the development of a robust predictive model. An agile and iterative approach was central to this stage. Rather than following a rigid, linear path, the project embraced a cycle of experimentation, evaluation, and refinement. Initial models were built and rigorously evaluated; the insights gained from this evaluation were then used to inform further data preparation adjustments and model tuning. This iterative loop ensured that the final model was not merely the first one that worked, but the most effective one developed through progressive enhancement. The final stage involved translating the analytical findings into actionable insights through the creation of an interactive dashboard, demonstrating the competency of achieving impact with data. Throughout this entire process, regular, structured discussions with the project supervisor provided critical oversight and academic guidance. These meetings served as formal checkpoints to validate the methodology, review interim findings, and ensure that the project remained aligned with its core objectives and the principles of the IoA framework.

Case Studies

Several leading companies have successfully implemented AI-driven solutions to enhance operational efficiency, customer engagement, and profitability. The following case studies demonstrate the various ways AI can be applied in the retail sector and explore the ethical considerations and potential challenges associated with these powerful technologies.

Case Study 1: Amazon

Overview

Amazon stands as a global leader in the integration of Artificial Intelligence (AI) within the retail sector. With a relentless focus on innovation and customer satisfaction, the company has redefined what consumers expect from an online shopping experience. At the core of this transformation lies Amazon’s powerful AI-driven recommendation engine, which plays a

crucial role in driving customer engagement and sales. This system analyzes vast amounts of data including a customer's browsing history, past purchases, Wishlist items, and the behaviours of other similar users to generate highly personalized product suggestions. These recommendations are often responsible for a significant percentage of Amazon's total sales, underscoring the system's role as a key revenue driver and contributor to customer loyalty.

Figure 11: Amazon Case Study

Beyond personalized shopping, Amazon applies AI extensively across other operational areas to gain a competitive edge. One of its most critical applications is in demand forecasting and inventory management. By analysing trends, seasonality, purchasing patterns, and regional demand, Amazon's AI algorithms can accurately predict product needs. This enables the company to strategically place inventory in fulfilment centres nearest to areas with high demand, thereby minimizing delivery times and improving customer satisfaction. In its fulfilment centres, Amazon has adopted a high level of automation. These centres utilize AI-powered robotics to streamline processes such as item picking, packaging, and transporting goods. Autonomous mobile robots (AMRs) navigate complex warehouse layouts, reducing the need for human intervention and increasing operational efficiency. This logistical optimization allows Amazon to fulfil millions of orders daily with remarkable speed and accuracy, a feat that is central to the success of its Prime membership program.

Ethical Issues

Despite its technological advancements, Amazon faces growing scrutiny regarding its ethical responsibilities particularly around data privacy. The company collects and processes a vast

amount of personal data from its users, which raises serious concerns about how that data is stored, used, and shared. Customers may not always be fully aware of the extent to which their data contributes to algorithmic decision-making, highlighting the need for transparency and informed consent. Another ethical challenge lies in algorithmic bias. Amazon's recommendation and search algorithms may unintentionally Favor its own products over those of independent third-party sellers. This could be perceived as anti-competitive behavior, stifling fair competition on its marketplace and reducing visibility for small businesses. If left unaddressed, such biases could lead to regulatory backlash and a decline in public trust. Moreover, the increasing use of automation and AI in warehouses has raised labour concerns. As robotics and intelligent systems replace manual labour, questions arise around job displacement, working conditions, and employee surveillance.

Potential Reasons for Failure

While AI has enabled Amazon's success, it also introduces certain vulnerabilities. One key area of risk is the over-reliance on automated algorithms. If the recommendation engine becomes outdated or fails to keep up with shifting customer trends and behaviours, it could lead to poor recommendations, reduced engagement, and lost revenue. A similar risk applies to demand forecasting errors in prediction could result in overstocking, leading to warehousing inefficiencies, or understocking, causing missed sales opportunities and customer dissatisfaction. Additionally, Amazon operates in diverse global markets, and its AI systems must accommodate different cultural preferences, languages, and purchasing habits. A failure to localize AI models effectively may lead to irrelevant or even offensive recommendations in certain regions, damaging the brand's reputation.

Potential Reasons for Success

Amazon's sustained success is deeply rooted in its data-centric philosophy and relentless pursuit of customer-centric innovation. By using AI not only to enhance user experience but also to streamline operations and logistics, Amazon creates a seamless and convenient shopping journey. Customers benefit from personalized experiences, fast delivery, and a vast product selection, which collectively drive long-term loyalty. The company's substantial investment in AI infrastructure, particularly through Amazon Web Services (AWS), has positioned it as a global technology powerhouse. AWS provides scalable cloud and machine learning services not only for internal use but also for thousands of businesses worldwide, creating an additional revenue stream and further solidifying Amazon's influence in the AI ecosystem. Furthermore, Amazon's agile and experimentation-driven culture allows it to rapidly test and implement AI innovations, staying ahead of market demands. Its ability to leverage feedback loops where customer interaction feeds data back into the AI systems for continuous improvement ensures that its algorithms remain adaptive, intelligent, and highly effective.

Case Study 2: Tesla

Overview

Tesla has not only revolutionized the automotive industry through the mass adoption of electric vehicles (EVs), but it has also redefined what it means to be a car company in the digital age.

By positioning itself as both an automotive and AI technology firm, Tesla has leveraged artificial intelligence across every layer of its operations from autonomous driving to factory automation and customer interaction. At the heart of Tesla's strategy lies its groundbreaking work on Autopilot and Full Self-Driving (FSD) systems, which have transformed cars into intelligent, learning machines. Tesla's autonomous driving technology is built upon one of the largest datasets of real-world driving behavior ever compiled. Every Tesla on the road acts as a sensor node, constantly collecting video, environmental, and behavioural data from diverse traffic scenarios. This data is fed into Tesla's Dojo supercomputer, where it trains sophisticated deep neural networks responsible for vehicle perception, planning, and decision-making. This data-driven, fleet-based learning system enables Tesla's AI to improve continuously, making each vehicle smarter with time.

Figure 12: Tesla Case Study

However, AI's role at Tesla extends far beyond the driver's seat. Tesla uses predictive AI models for proactive maintenance by analysing sensor data to detect early signs of component failure and alerting vehicle owners before issues arise. This minimizes downtime, enhances safety, and builds long-term customer trust. Additionally, in Tesla's Gigafactories, AI and robotics are utilized to manage quality control, optimize energy usage, and streamline complex manufacturing workflows. Machine vision systems are used for defect detection, while AI algorithms adapt to changing production demands to maintain high efficiency and consistency. Tesla's approach to product improvement is also unique in the automotive space. Through over-the-air (OTA) software updates, AI-enhanced features and performance upgrades can be delivered directly to customers' vehicles, much like a smartphone. This model turns Tesla

vehicles into dynamic platforms that evolve even after purchase, increasing the product's long-term value.

Ethical Issues

Tesla's ambitious use of AI raises several critical ethical and regulatory concerns. One of the most pressing is the safety and reliability of its FSD system, especially considering that it is currently being tested in real-world conditions by regular drivers. While Tesla clearly labels the system as "beta" software and urges users to remain attentive, the public nature of its deployment has drawn criticism from safety advocates and regulators. In particular, the lack of driver attentiveness and the potential for misuse of Autopilot features have already been linked to several high-profile accidents. There is also the ethical dilemma of decision-making in unavoidable crash scenarios commonly referred to as the "trolley problem." How should the AI prioritize outcomes in life-threatening situations, and who is held accountable for its decisions? The lack of transparency in how such moral judgments are encoded into Tesla's models creates a grey area in terms of legal and ethical responsibility. Furthermore, data privacy is a growing concern. Tesla vehicles continuously gather high-resolution video and telemetry data from users, often without them fully understanding the scope or purpose. This includes not only data from the vehicle's external cameras and sensors but also cabin-facing cameras in newer models. Without clear and user-friendly consent mechanisms, this poses serious risks related to surveillance and data exploitation.

Potential Reasons for Failure

Tesla's aggressive AI roadmap comes with several potential pitfalls. A major risk lies in its over-promising of Full Self-Driving capabilities. CEO Elon Musk has repeatedly made optimistic projections about achieving Level 5 autonomy full automation without any human intervention. However, the technological and legal barriers to achieving this level of autonomy remain immense. Failure to deliver on such promises could not only trigger regulatory action but also damage the company's credibility and stock value, which is closely tied to its image as a tech innovator. Another challenge is the opacity of neural networks the so-called "black box" problem. If an AI-driven Tesla is involved in a fatal crash, and the company cannot clearly explain the cause of the AI's decision-making process, it could lead to serious public backlash and loss of trust. Without interpretability and explainability, debugging or defending the system becomes extremely difficult. Tesla also faces intensifying competition from legacy automakers and startups, many of which are investing heavily in autonomous systems, often in partnership with tech companies. If Tesla's pace of innovation slows or it is outpaced in key areas like AI chip performance, regulatory approval, or consumer trust, its competitive edge could erode rapidly.

Potential Reasons for Success

Tesla's biggest strength lies in its data dominance. With millions of vehicles on the road continuously generating diverse driving data, Tesla operates a data flywheel that feeds into its machine learning models and reinforces a cycle of rapid improvement. Unlike competitors that rely on simulated environments or limited test fleets, Tesla's learning system is deeply embedded in the real world, giving it a major strategic advantage. Tesla's vertical integration

where the company controls both hardware (cars, sensors, chips) and software (AI models, firmware, OS allows for tight coordination and rapid deployment of new features. This results in a seamless user experience that traditional auto manufacturers, which often rely on third-party suppliers, find difficult to replicate. The introduction of Dojo, Tesla's proprietary AI supercomputer, is another game-changing move. Designed specifically for training vision-based neural networks, Dojo significantly accelerates the development cycle of FSD models, enhancing the feedback loop from fleet to AI. Moreover, the company's OTA update ecosystem transforms Tesla vehicles into evolving platforms, capable of receiving improvements, features, and bug fixes without a visit to the service centre. This fosters customer delight and long-term brand loyalty, as users continually benefit from enhancements long after purchase.

Case Study 3: Walmart

Overview

Walmart, the world's largest brick-and-mortar retailer, has undergone a significant transformation by integrating AI across its supply chain, store operations, customer service, and e-commerce platforms. At the heart of Walmart's AI strategy is the drive for efficiency, availability, and customer value. Walmart employs machine learning models to improve demand forecasting, ensuring that the right products are stocked in the right locations, minimizing overstocking and avoiding empty shelves. In stores, Walmart uses shelf-scanning robots equipped with computer vision to monitor inventory levels, identify misplaced items, and detect pricing discrepancies in real time. These autonomous robots enhance stock accuracy and employee productivity, freeing up staff to focus on customer engagement. Walmart also uses AI in chatbots and virtual assistants for customer service, providing 24/7 support with faster and more accurate responses. On the e-commerce side, Walmart utilizes AI for personalized shopping experiences, offering product suggestions, promotions, and dynamic pricing based on customer behavior and purchase history. Additionally, AI powers route optimization for delivery logistics, improving efficiency in Walmart's last-mile delivery network.

Figure 13: Walmart Case Study

Ethical Issue

As Walmart scales its AI capabilities, ethical considerations become critical. Mass data collection, especially from in-store purchases, mobile apps, and online interactions, requires strict data protection protocols. Without proper transparency, customers may be unaware of how their data is used or shared.

Moreover, workforce automation poses a social and ethical challenge. Walmart employs millions of associates, and the introduction of robotic systems in stores and warehouses could lead to job redundancy, impacting low-income and frontline workers. The company must navigate the balance between innovation and employment by reskilling and redeploying affected employees.

Potential Reasons for Failures

Walmart's AI strategy may falter if its predictive models are overfitted to historical data that no longer reflects current market conditions. Sudden market shifts, such as a pandemic, inflation spike, or global crisis, can render machine learning outputs inaccurate, leading to misallocation of resources. Another risk lies in the public perception of automation. If customers or advocacy groups perceive Walmart's AI applications as intrusive (e.g., through facial recognition or behavioural tracking), it could harm the brand's reputation. Additionally, lack of interoperability between legacy systems and new AI platforms could slow digital transformation.

Potential Reasons for Success

Walmart's primary advantage is its scale and infrastructure. With over 10,000 retail locations worldwide and a vast logistics network, the company is ideally positioned to collect granular data across geographies and apply AI at scale. Its investments in data science teams, cloud partnerships, and in-house innovation labs (e.g., Walmart Global Tech) have created a solid foundation for sustained AI growth. Walmart's strong focus on business value improving efficiency, reducing costs, and enhancing customer satisfaction ensures that AI investments are aligned with strategic goals. Furthermore, the company's commitment to digital transformation through initiatives like Walmart+ and automated fulfilment centres suggests a long-term vision that integrates AI as a central driver of retail evolution.

Case Study 4: Alibaba.com

Overview

Alibaba.com, the global business-to-business (B2B) e-commerce platform of the Alibaba Group, has emerged as a trailblazer in leveraging Artificial Intelligence (AI) to enhance platform efficiency, customer satisfaction, and ecosystem scalability. As one of the largest online wholesale marketplaces in the world, Alibaba serves millions of buyers and suppliers across over 190 countries. Its AI infrastructure plays a central role in managing this complexity. Alibaba's recommendation engine is driven by sophisticated machine learning algorithms that analyse user behavior, transaction history, geographic preferences, and even browsing speed to personalize product listings for buyers. These real-time, intelligent suggestions not only increase conversion rates but also help suppliers reach their target audience more effectively. Another key AI application is in smart logistics and inventory management. Through the Canio logistics platform (a subsidiary of Alibaba Group), AI predicts optimal shipping routes, warehouse allocations, and delivery times. This ensures a seamless experience for both sellers and buyers, despite the vast geographical spread of Alibaba's operations.

Figure 14: Case Study of Alibaba.com

Alibaba also deploys AI for fraud prevention and transaction monitoring. With millions of payments and listings occurring daily, the platform uses anomaly detection models to flag suspicious behavior in real time, protecting both vendors and customers from fraud, counterfeit goods, and policy violations. A major differentiator for Alibaba is its use of AI-powered multilingual chatbots, including AliMe, which provides instant 24/7 customer service, order tracking, dispute resolution, and product inquiries in multiple languages. This automation ensures scalability and responsiveness in a highly diverse global market.

Ethical Issues

Alibaba, like other tech giants, faces significant ethical questions surrounding algorithmic fairness, transparency, and data governance. One concern involves the ranking algorithms used to display products and sellers. If these algorithms inadvertently Favor high-volume or premium-paying sellers, it could marginalize smaller businesses, eroding the B2B model's fairness and inclusivity. Another major concern is data privacy. The platform handles sensitive business and financial data, including cross-border transactions, customer records, and communication logs. Any mishandling or breach of this data could damage user trust and provoke legal scrutiny particularly in light of increasing global data protection regulations such as China's PIPL (Personal Information Protection Law), the GDPR, and other international frameworks. Moreover, Alibaba's use of AI to monitor and evaluate seller behavior raises ethical issues regarding surveillance and autonomy. While it helps maintain trust, overly aggressive automation in suspending or ranking sellers based on algorithmic assessments could harm legitimate businesses without due process.

Potential Reasons for Failures

One of the key challenges Alibaba faces is the scalability of its AI systems in such a vast and heterogeneous ecosystem. If AI models fail to adapt to new types of products, fraudulent schemes, or cultural nuances in user behavior, it could lead to significant performance issues. For example, false positives in fraud detection could block legitimate sellers or transactions, causing frustration and revenue loss. Conversely, false negatives could allow harmful or illegal activity to go undetected, damaging the platform's reputation. Another potential failure point is multilingual capability and cultural adaptation. If Alibaba's AI-powered support tools or search systems fail to accurately process and respond in diverse languages, dialects, or business customs, international users may have a suboptimal experience, impacting global expansion. Additionally, if sellers perceive the platform's AI tools such as product ranking, visibility, or promotional pricing as biased or opaque, it could lead to dissatisfaction and platform abandonment by key vendors.

Potential Reasons for Success

Alibaba's success with AI stems from its ecosystem-centric approach. Rather than using AI solely to improve its internal operations, Alibaba deploys AI tools that empower sellers, streamline logistics, and enhance customer experience. This creates a win-win environment where all stakeholders benefit from increased efficiency and transparency. Its development of homegrown AI infrastructure, such as Pingtung chips and proprietary neural network models, enables faster and more efficient processing of massive datasets. These innovations reduce dependency on third-party technologies and enhance Alibaba's ability to tailor AI solutions for unique market needs.

Moreover, Alibaba's deep integration between its e-commerce, cloud (Alibaba Cloud), payments (Alipay), and logistics arms creates rich cross-platform data that fuels AI performance. For instance, buyer behavior on Alibaba.com feeds into marketing recommendations, while shipment data from Cainiao enhances fulfilment predictions. The company's focus on open innovation including partnerships with global AI researchers and investment in startups has helped it stay at the cutting edge of AI advancement. These factors collectively reinforce Alibaba's position as a dominant force in global commerce powered by smart, scalable, and user-centric AI.

Integration

The integration of the Retail Analytics project was designed to ensure that all components from data acquisition to final presentation were seamlessly connected in a cohesive, automated, and reproducible pipeline. This began with the Data Feeds, which formed the foundational layer of the system and drew from two primary sources. The first comprised online resources such as academic journals, industry reports, and case studies. These resources informed the analytical framework, guiding the choice of statistical and machine learning techniques, shaping the interpretation of findings, and ensuring adherence to best practices in ethical and unbiased data handling. The second, and most critical, input was the "Online Retail II" dataset from the UCI Machine Learning Repository. Provided in CSV format, this dataset contained real-world transactional records, making it ideal for demonstrating the application of advanced analytics

and machine learning in a retail context. Its public, anonymized nature allowed the project to maintain privacy compliance while preserving the authenticity and complexity needed for meaningful testing and validation. Once acquired, the dataset entered the Data Processing stage, where integration became more technical and process-driven. This stage consisted of three tightly linked phases: Data Cleaning, Feature Engineering, and Model Training. In the Data Cleaning phase, the pandas library in Python was used to filter out canceled orders (identified by invoice numbers beginning with "C"), remove rows with missing Customer IDs to prevent biased results, and convert InvoiceDate fields into datetime objects for accurate temporal calculations. The automated cleaning pipeline ensured consistent, error-free datasets ready for analysis regardless of size or complexity. From there, Feature Engineering enriched the dataset using the RFM (Recency, Frequency, Monetary) framework, calculating for each customer how recently they made a purchase, how often they purchased, and the total value of those purchases. These engineered features acted as the bridge between raw transactional data and the predictive models that followed.

With the cleaned and enriched data prepared, the process advanced to Model Training, where methodological and technological integration came to the forefront. Two core models, built using the scikit-learn library, provided complementary insights: a K-Means clustering model for customer segmentation and a Logistic Regression classifier for churn prediction. For the clustering model, the Elbow Method was used to determine the optimal number of customer groups, balancing simplicity with accurate behavioral representation.

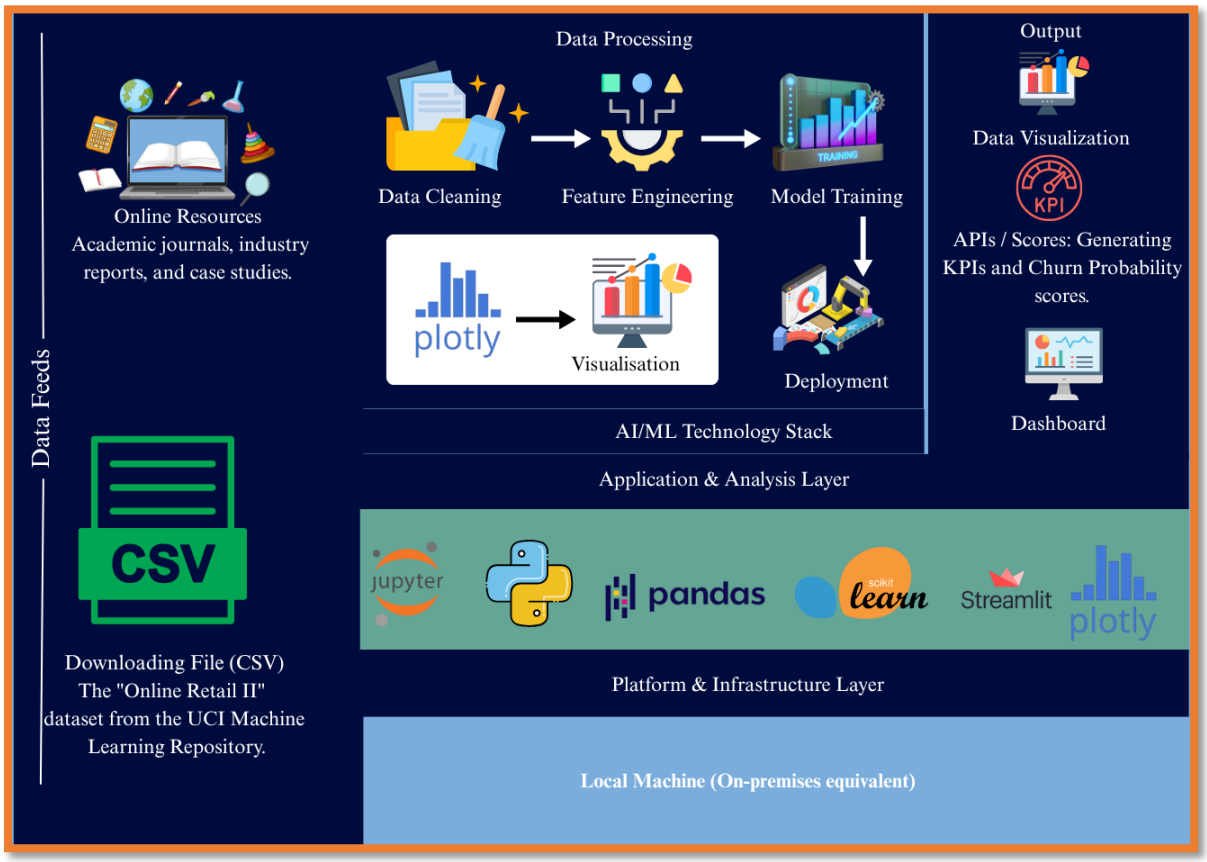


Figure 15: Integration

This segmentation offered retailers a descriptive view of their customer base, identifying patterns in purchasing habits and allowing for targeted marketing strategies. The churn prediction model, in contrast, provided a predictive dimension by identifying customers at risk of disengagement based solely on behavioral data, reducing the potential for demographic bias. The seamless integration between pandas (for feature preparation) and scikit-learn (for training and evaluation) meant that the transition from raw RFM scores to actionable insights required no intermediate data exports or manual intervention. Once the models were trained and validated, outputs flowed into the Deployment stage, where results became accessible to stakeholders. Visualizations produced using Plotly were designed to be both interactive and aesthetically engaging, enabling deep dives into customer clusters, churn risk distributions, and sales trends. Integration between scikit-learn and Plotly was handled entirely in Python, allowing outputs such as cluster labels, churn probabilities, and KPI aggregates to pass directly into visualization scripts without format conversion. The final presentation layer was powered by Streamlit, a Python framework for building interactive dashboards. This ensured that stakeholders without any coding expertise could explore results in real time, filter data, and examine specific customer groups. The dashboard dynamically updated as models processed new or modified datasets, displaying KPIs, churn probabilities, and other key metrics in a centralized, intuitive interface. By integrating Streamlit and Plotly, the system allowed not just static reporting but active engagement, empowering decision-makers to tailor insights to scenarios such as campaign planning, inventory adjustments, and retention strategy design.

Findings

Findings For Research Question 1

Research Question 1: How are other companies effectively using AI-driven customer engagement solutions to improve decision-making and enhance customer satisfaction?

The findings from the literature review and detailed case study analysis reveal that AI is fundamentally transforming how leading companies engage with customers and make operational decisions. Industry giants such as Amazon, Walmart, Alibaba, and Tesla have strategically embedded AI technologies across their business models, yielding both competitive advantages and measurable improvements in customer experience. These organizations treat AI not merely as a supplementary tool but as a core technological foundation that spans front-end personalization, supply chain optimization, and customer support services. One of the most impactful applications identified is the deployment of AI-powered recommendation systems. Amazon exemplifies this by leveraging complex machine learning algorithms that analyze extensive customer data including browsing behavior, purchase history, wish lists, and demographic profiles to deliver highly personalized product suggestions. These recommendation engines function in real-time and dynamically adapt to shifts in customer preferences, resulting in shopping experiences that are both relevant and intuitive. This personalization strategy significantly elevates customer satisfaction by simplifying product discovery while simultaneously driving higher conversion rates and fostering customer loyalty. Notably, a substantial share of Amazon's revenue is attributable to

the effectiveness of these recommendation algorithms, underscoring their critical role as revenue drivers. Beyond personalization, AI is revolutionizing supply chain and inventory management, as illustrated by Walmart's advanced use of predictive analytics. By incorporating historical sales data, weather patterns, promotional activities, and regional events, Walmart's AI models accurately forecast demand at both local and national scales. This precise forecasting enables optimized inventory stocking, minimizing overstock and waste while ensuring product availability. The enhanced supply chain responsiveness directly improves customer experience by reducing stockouts and aligning product availability with customer expectations. Additionally, this approach supports Walmart's broader strategic goals of operational efficiency and maintaining everyday low pricing.

Findings for Research Question 2

Research Question 2: How can AI predictive analytics optimize purchasing, inventory, and personalized marketing in retail?

This question was explored through the practical application of a multi-step data analysis and modelling process, which transformed raw transaction records into meaningful insights. The findings from this process confirm that artificial intelligence and predictive analytics can significantly optimize retail operations in several critical areas, particularly in customer segmentation, churn prevention, and inventory management. To begin with, one of the most valuable insights came from customer segmentation, which plays a key role in designing more effective personalized marketing strategies. The dataset contained records from 4,312 unique customers, and these were segmented into four distinct groups using the RFM (Recency, Frequency, Monetary) analysis technique. This method evaluated how recently a customer made a purchase, how often they purchased, and how much they spent in total. Based on these metrics, each customer was scored and categorized into one of four segments, each with different behavioural characteristics and business value.

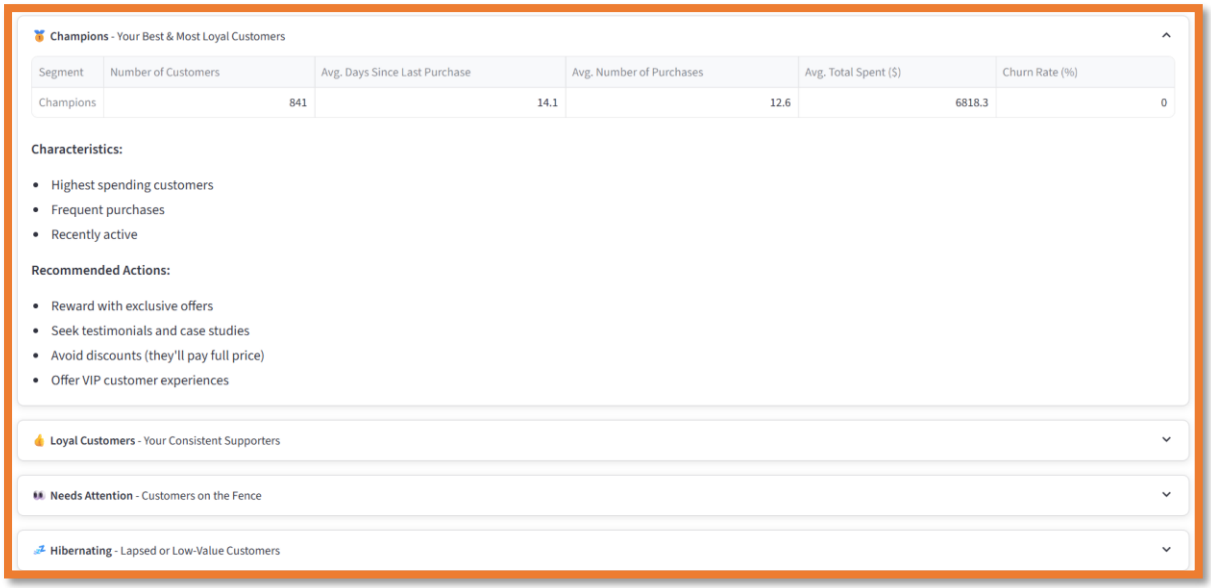


Figure 16: Classification of Customers

The first segment, "Champions", represented about 20% of the customer base. These customers are highly valuable they make frequent and recent purchases and spend the most. They also show a strong likelihood of returning, making them ideal for exclusive loyalty rewards and VIP programs. The second group, "Loyal Customers", made up 19% of the base. They also buy regularly and have a strong relationship with the brand but spend slightly less than Champions. "Needs Attention", the third group, covered around 21% of customers. They show average spending and purchasing patterns but haven't made recent transactions. This makes them a crucial target for re-engagement offers before they churn. Finally, the "Hibernating" group, which was the largest segment (40%), includes customers with long periods of inactivity and very low spending. These users are at the highest risk of churn and may need stronger incentives to return or may no longer be cost-effective to pursue.

This segmentation provides a clear and actionable framework for personalized marketing. Instead of sending generic promotions to all users, a retailer can tailor its campaigns. For example, rewards programs can be directed at "Champions," while targeted emails or discount offers can be used to re-engage the "Needs Attention" group. Next, the study focused on using predictive analytics to prevent customer churn. A Logistic Regression model was trained on the RFM features to estimate the likelihood that a customer would stop purchasing. This model was evaluated using unseen test data and achieved excellent results, including a very high accuracy of 87.09%. The performance of the model is summarized in the classification report shown in the next figure. The report shows that the model had strong precision and recall values, especially for the "churned" class (class 1), meaning it was highly reliable in identifying which customers were most at risk of leaving. The model generated a Churn_Probability score for each customer, which indicates how likely they are to churn in the near future. These probability scores were incorporated into the dashboard's "High-Risk Priority List", allowing business users to see exactly which customers are in danger of churning and prioritize them for outreach.

Figure 17: Segmentation of Customers

As shown in the "Priority List: High Churn Risk Individuals" section of the dashboard, this enables a business to use data to proactively retain valuable customers by offering them personalized promotions, support, or incentives—thus improving retention and reducing marketing costs. In addition to customer behavior and marketing, the analysis also produced

valuable insights for inventory optimization, which is crucial in retail operations. By examining the entire transaction history, the system identified which products were the most popular in terms of quantity sold. These findings were presented visually in a chart showing the top-selling products, as shown in the next figure.

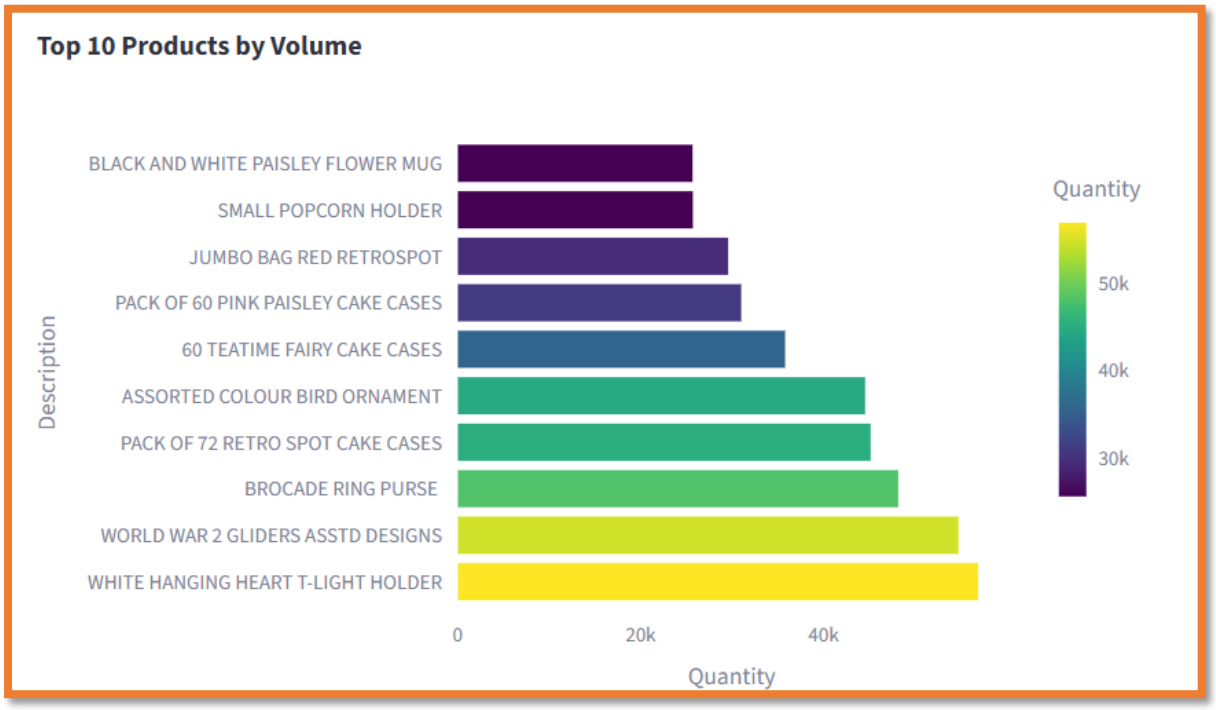


Figure 18: Top 10 Products by Volume

This data provides a direct and practical benefit to the business: it can now ensure that top-selling products are always in stock, avoiding stockouts that frustrate customers and result in lost sales. At the same time, it helps streamline purchasing decisions by identifying which products are less in demand, reducing excess inventory and associated holding costs. Together, this contributes to higher customer satisfaction and more efficient operations. In summary, the results show that AI-powered predictive analytics can have a significant positive impact on the way retailers manage their marketing, retention, and supply chain strategies. By segmenting customers based on their behavior, predicting who is likely to churn, and identifying which products drive the most sales, businesses can make smarter, faster, and more personalized decisions. This leads to improved customer experiences, reduced operational inefficiencies, and ultimately, better business outcomes.

Findings For Research Question 3

Research Question 3: What ethical and privacy concerns arise in AI retail, and how can businesses address them?

The integration of artificial intelligence in retail brings transformative potential but also surfaces critical ethical and privacy challenges that must be managed responsibly. This research identifies three central concerns in the use of AI-driven customer engagement solutions: safeguarding customer data privacy, mitigating algorithmic bias, and ensuring the responsible application of AI predictions with a human-cantered focus. The project adopted an “Ethics by

Design" approach, embedding ethical safeguards at every stage of data collection, analysis, and deployment. Data privacy stands out as a paramount issue, given AI's reliance on detailed transactional and behavioral data to generate actionable insights. Retailers collect sensitive information that, if mishandled, could erode customer trust and violate legal standards. To address this, strict anonymization protocols were applied from the outset, substituting real customer identifiers with randomized codes and excluding personal identifiers such as names, emails, and phone numbers from all modelling processes. This ensured that meaningful analysis of customer engagement metrics like Recency, Frequency, and Monetary value (RFM) could proceed without compromising individual privacy. The findings underscore that anonymized data is sufficient to drive effective business intelligence while maintaining customer confidentiality, thus reducing risks associated with data breaches or misuse.

Algorithmic bias presents a further ethical risk, where machine learning models may replicate or even exacerbate historical prejudices embedded in training data. In retail contexts, such biases could lead to unfair treatment or exclusion of specific customer segments. This study deliberately excluded demographic variables such as age, gender, ethnicity, and location from predictive modelling, focusing instead on behavioral indicators that reflect actual customer activity. This choice aligns with ethical machine learning principles, reducing the risk of reinforcing social biases and supporting more equitable treatment across the customer base. The project demonstrates that ethical AI design is not only desirable but practically achievable through thoughtful variable selection and model construction.

A third ethical consideration involves the responsible use of AI-driven predictions, particularly around churn modelling and customer prioritization. While predictive analytics can enhance decision-making, there is potential for misuse, such as deprioritizing customers deemed "low value." This research reframed AI as a tool for customer empathy rather than exclusion: churn predictions were deployed to identify at-risk customers for proactive engagement, enabling personalized outreach and support rather than punitive measures. This human-cantered application fosters stronger customer relationships and illustrates how AI can be harnessed to improve satisfaction and loyalty without compromising ethical standards. The ethical framework implemented in this project confirms that responsible AI deployment in retail is both essential and feasible. Through early data anonymization, avoidance of demographic profiling, and the prioritization of customer care in predictive use, businesses can develop AI systems that are powerful, fair, and trustworthy. As AI continues to shape customer engagement strategies, adherence to these ethical practices will not only ensure regulatory compliance but also confer competitive advantages by building deeper customer trust and fostering long-term loyalty.

Future Work and Recommendation

Although this thesis has successfully developed a complete framework for applying AI in retail analytics covering customer segmentation, churn prediction, and business insights through a live dashboard there are still many exciting directions for future research and system enhancement. The foundation created here can be extended and improved in several ways. These future possibilities can be grouped into three key areas: improvements to the predictive

models, the integration of additional data sources, and the development of more advanced and interactive dashboard features.

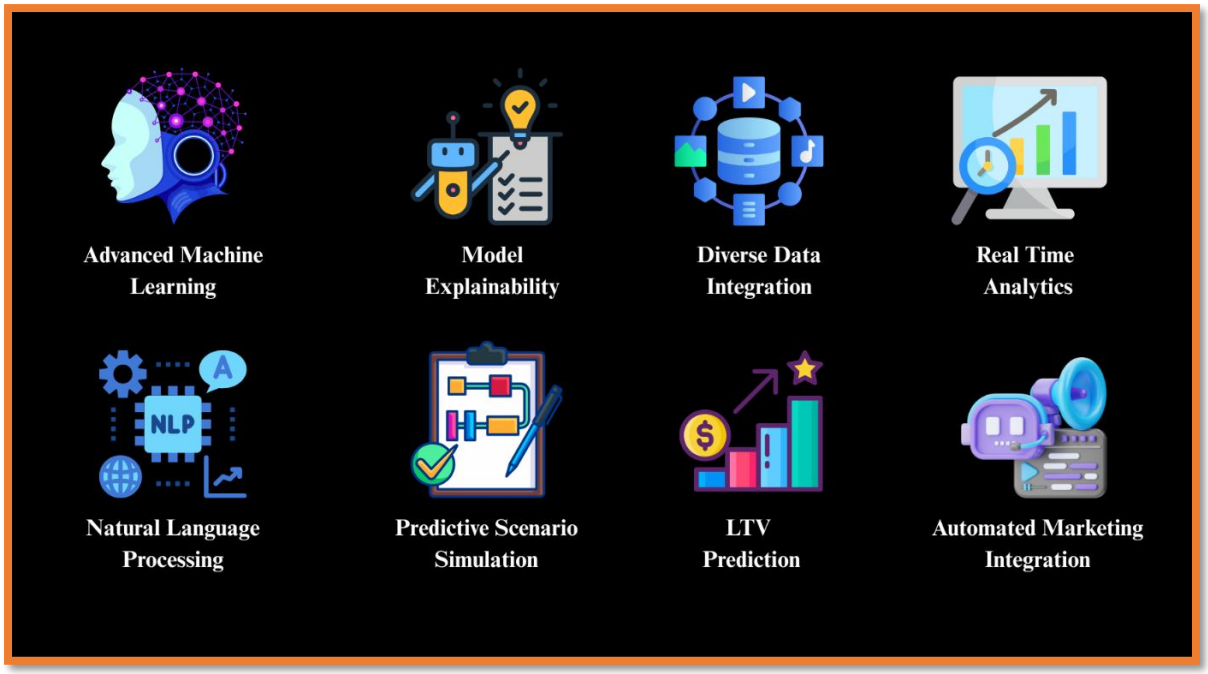


Figure 19: Future Work and Recommendation

First, in terms of model enhancement, the predictive models used in this project particularly the Logistic Regression model for churn prediction performed very well, achieving high accuracy. However, there is room to explore more advanced machine learning algorithms that may offer even better performance. Future work could experiment with ensemble models such as Random Forests or Gradient Boosting Machines (e.g., XGBoost). These models can capture more complex relationships in the data and are often better at handling imbalanced classes, which is common in churn prediction problems. Another promising direction would be the use of deep learning techniques, especially models like Recurrent Neural Networks (RNNs). Since customer behavior often follows patterns over time, RNNs could be useful for analyzing sequences of transactions to not only predict whether a customer is likely to churn, but also when they might churn or make their next purchase. This kind of time-based prediction could support even more personalized and timely marketing strategies.

The second area of improvement is data integration and expanded analytics. The current project used only a single transactional dataset, which provided valuable insights but had some limitations. A more comprehensive understanding of customers could be achieved by combining multiple data sources. For example, demographic data such as a customer's age, location, or income level could help identify which groups are most profitable or most vulnerable to churn. In addition, clickstream data from the company's website could be analyzed to understand how users browse before making a purchase. This would allow businesses to study intent and identify where customers drop off before completing a purchase. Another exciting opportunity is to analyze customer reviews and social media activity. By applying Natural Language Processing (NLP) techniques, future projects could measure

customer sentiment directly and even identify specific complaints, praises, or concerns. These insights would be incredibly helpful for understanding satisfaction levels and the emotional drivers behind customer retention or churn.

The third area for future work focuses on enhancing the dashboard and making it more interactive and strategic. While the current Streamlit dashboard already allows users to explore customer segments and churn probabilities, it could be made even more powerful. One useful addition could be a “Retention Strategy Simulator a feature that would allow managers to test the effect of different actions (such as offering discounts, loyalty points, or free shipping) on customer retention and calculate the Return on Investment (ROI) of each strategy. This would help businesses make financially sound decisions about which campaigns to launch. Another potential enhancement is to add a “Product Recommendation Engine” to the dashboard. Using methods like Market Basket Analysis, the system could recommend additional products to each customer based on what similar customers have bought. This would help turn the dashboard from a passive analysis tool into a real-time sales and marketing assistant, offering suggestions for cross-selling and upselling.

In conclusion, while the current system delivers practical and accurate insights into customer behavior using AI, it also opens up many possibilities for future development. More advanced models, richer data, and smarter dashboards could significantly improve the predictive power and business impact of the system.

Conclusion

This thesis developed an AI-driven framework for analyzing customer engagement and predicting churn in the retail sector, integrating RFM segmentation and a highly accurate (87.09%) Logistic Regression model to transform raw transactional data into actionable insights. Customers were categorized into four key personas 'Champions', 'Loyal Customers', 'Needs Attention', and 'Hibernating' enabling targeted marketing strategies. The model's Churn_Probability score empowers businesses to adopt proactive retention measures, while an interactive Streamlit dashboard consolidates all insights into a single decision-making tool, offering a clear view of customer behavior, segment performance, and high-risk profiles. This work demonstrates that AI-powered analytics can effectively enhance decision-making, boost customer satisfaction, and drive sustainable growth in today's competitive retail environment.

Bibliography

Ascarza, E. (2018). Retention futility: Sunk costs and customer lock-in. Journal of Marketing Research, 55(2), 227–245. https://doi.org/10.1509/jmr.15.0204

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases (pp. 487–499). https://www.vldb.org/conf/1994/P487.PDF

Berry, M. J. A., & Linoff, G. S. (2004). Data mining techniques: For marketing, sales, and customer relationship management (2nd ed.). John Wiley & Sons.

Bhattacharyya, S., & Jha, S. (2011). A novel customer churn prediction model in retail industry. International Journal of Computer Applications, 28(4), 1–6. https://doi.org/10.5120/3423-4719

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164(1), 268–283. https://doi.org/10.1016/j.ejor.2003.11.038

Chen, D., & Sain, S. L. (2003). Building a scalable and accurate customer churn model. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 400–409). https://doi.org/10.1145/956750.956785

Davenport, T. H. (2006). Competing on analytics. Harvard Business Review, 84(1), 98–107. https://hbr.org/2006/01/competing-on-analytics

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org/

Grönroos, C. (1994). From marketing mix to relationship marketing: Towards a paradigm shift in marketing. Management Decision, 32(2), 4–20. https://doi.org/10.1108/00251749410054774

Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques (3rd ed.). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://hastie.su.domains/ElemStatLearn/

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. Journal of Marketing, 80(6), 36–68. https://doi.org/10.1509/jm.15.0414

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.030

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. Journal of Marketing, 80(6), 69–96. https://doi.org/10.1509/jm.15.0420

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). https://projecteuclid.org/euclid.bsmsp/1200512992

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

O'Malley, L., & Tynan, C. (2000). Relationship marketing in consumer markets: Rhetoric or reality? European Journal of Marketing, 34(7), 797–815. https://doi.org/10.1108/03090560010331220

Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. Journal of Marketing, 69(4), 167–176. https://doi.org/10.1509/jmkg.2005.69.4.167

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Reichheld, F. F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the web. Harvard Business Review, 78(4), 105–113. https://hbr.org/2000/07/e-loyalty-your-secret-weapon-on-the-web

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing, 67(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589

Simester, D., Timoshenko, A., & Zou, T. (2020). Targeting prospective customers: A field experiment. Marketing Science, 39(3), 529–549. https://doi.org/10.1287/mksc.2019.1195

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. Journal of Retailing, 85(1), 31–41. https://doi.org/10.1016/j.jretai.2008.11.001

Wiering, M., & Van Otterlo, M. (Eds.). (2012). Reinforcement learning: State-of-the-art. Springer. https://link.springer.com/book/10.1007/978-3-642-27645-3

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs. https://www.publicaffairsbooks.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395700/

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., & Van Kenhove, P. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research, 156(2), 508–523. https://doi.org/10.1016/S0377-2217(03)00160-1

Berson, A., Smith, S., & Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill.

Bose, R. (2002). Customer relationship management: Key components for IT success. Industrial Management & Data Systems, 102(2), 89–97. https://doi.org/10.1108/02635570210419636

Chatterjee, S., Nguyen, B., Ghosh, S. K., Bhattacharjee, K. K., & Chaudhuri, R. (2021). Adoption of artificial intelligence-integrated CRM systems in the B2B sector: An empirical study. Journal of Business Research, 131, 40–52. https://doi.org/10.1016/j.jbusres.2021.03.038

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Davenport, T. H., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48, 24–42. https://doi.org/10.1007/s11747-019-00696-0

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. Decision Support Systems, 55(1), 359–363. https://doi.org/10.1016/j.dss.2012.05.044

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Ghosh, S., & Reilly, D. L. (1994). Credit card fraud detection with a neural-network. In System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on (Vol. 3, pp. 621–630). IEEE. https://doi.org/10.1109/HICSS.1994.323314

Kim, M. J., Lee, C. K., & Jung, T. (2020). Exploring consumer behavior in the context of social media commerce: Focusing on the roles of social support and social presence. Sustainability, 12(10), 4371. https://doi.org/10.3390/su12104371

Kotler, P., Kartajaya, H., & Setiawan, I. (2017). Marketing 4.0: Moving from traditional to digital. Wiley.

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 517–526). https://doi.org/10.1145/1557019.1557075

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Rust, R. T., & Huang, M.-H. (2021). The feeling economy: How artificial intelligence is creating the era of empathic business. Journal of the Academy of Marketing Science, 49, 5–21. https://doi.org/10.1007/s11747-020-00707-2

Appendix

GitHub Link

https://github.com/Karanbohara01/thesis.git

Project Plan

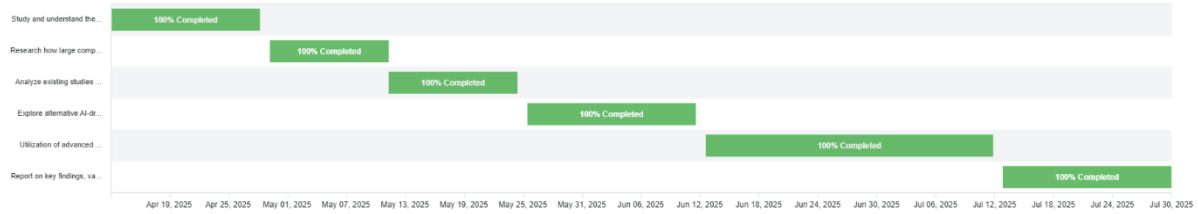


Figure 20: Project Plan

Risk Analysis

Rank	Name	Occurance	Impact	Plan B
1	Loss of motivation when encountering challenges or obstacles.	rare	mild	Maintain strong focus and seek support when necessary.
2	Time management	rare	**fatal**	Use a time tracker for effective planning.
3	Lack of knowledge in searching for datasets	often	low	Learn more and examine others' approaches for deeper insight
4	Handling multiple tasks at once	often	**fatal**	Prioritize tasks, break them down, focus individually.
5	Personal health	frequently	mild	Adjust routines and be careful for next time
6	Difficulty in understanding complex AI models	Often	**High**	Take online courses, consult experts, and break down models into simpler concepts.
7	Limited access to quality datasets	Frequently	**Fatal**	Explore open-source datasets, collaborate with peers, and use data augmentation techniques.

Figure 21: Risk Analysis

Figure 22: SWOT Analysis

Dashboard Screenshots

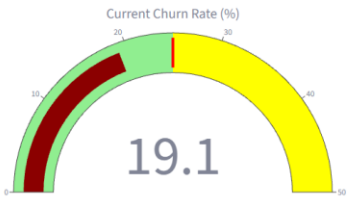
🚀 AI-Powered Retail Analytics Dashboard

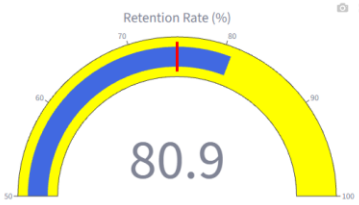
A comprehensive overview of customer segmentation, sales performance, and predictive churn insights.

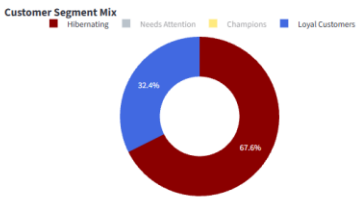
Key Performance Indicators

Total Revenue	Total Customers	Churned Customers	Churn Rate
$8,832,003	4,312	825	19.1%
			↓ -5.9% vs benchmark

Customer Health & Segmentation







📈 Monthly Performance Snapshot

Summary for December, 2010

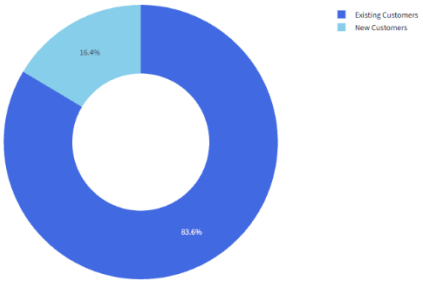
Performance Summary Table

Category	Customers	Revenue Impact ($)
Existing Active	2,914	$0
New	573	$311,878
Lost	186	$0

- **Existing Active:** Customers acquired in previous months who made a purchase this month.
- **New:** Customers who made their very first purchase this month.
- **Lost:** Customers whose inactivity period crossed the 180-day churn threshold this month. Revenue impact shows their total lifetime value.

This Month's Customer Base

Composition of Active Customers This Month



Customer Flow Waterfall

Monthly Customer Gain vs. Loss

Retention by Customer Segment

Customer Retention by Segment

Segment	Churned	Retained	Retention_Rate
Champions	0.000000	841.000000	100.0%
Loyal Customers	11.000000	819.000000	98.7%
Needs Attention	66.000000	843.000000	92.7%
Hibernating	748.000000	984.000000	56.8%

Actionable Insights by Customer Segment

Click on each segment to understand their characteristics and recommended marketing actions.

Champions - Your Best & Most Loyal Customers

Segment	Number of Customers	Avg. Days Since Last Purchase	Avg. Number of Purchases	Avg. Total Spent ($)	Churn Rate (%)
Champions	841	14.1	12.6	6818.3	0

Characteristics:

- Highest spending customers
- Frequent purchases
- Recently active

Recommended Actions:

- Reward with exclusive offers
- Seek testimonials and case studies
- Avoid discounts (they'll pay full price)
- Offer VIP customer experiences

Loyal Customers - Your Consistent Supporters

Segment	Number of Customers	Avg. Days Since Last Purchase	Avg. Number of Purchases	Avg. Total Spent ($)	Churn Rate (%)
Loyal Customers	830	42.9	4.8	1972	1.3

Characteristics:

- Good purchase frequency
- Moderate spending
- Somewhat recently active

Recommended Actions:

- Upsell higher-value products
- Offer loyalty program memberships
- Keep them engaged with regular content
- Request product reviews

Hibernating - Lapsed or Low-Value Customers

Segment	Number of Customers	Avg. Days Since Last Purchase	Avg. Number of Purchases	Avg. Total Spent ($)	Churn Rate (%)
Hibernating	1732	162.2	1.3	340.6	43.2

Characteristics:

- Lowest spending
- Haven't purchased in a long time
- Highest churn rate

Recommended Actions:

- Launch a 'win-back' campaign
- Offer special reactivation discounts
- Focus only on those with high historical value
- Consider sunsetting unresponsive customers

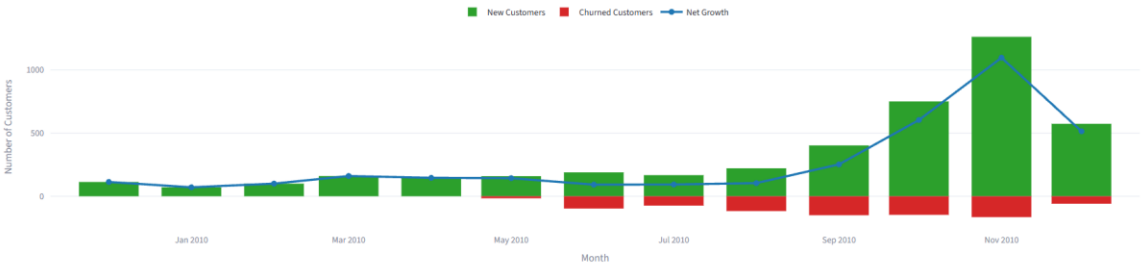
Performance Trends & Deep Dives

Deploy

Customer Lifecycle | Sales & Products | Geographical Performance | Churn Drivers & Demographics | Churn Drivers | Attrition Analysis

Monthly Customer Acquisition vs. Churn

Customer Acquisition vs. Churn with Net Growth



Average Monthly Net Growth
268.2

Latest Month Net Growth
514.0
↑ 245.8 vs average

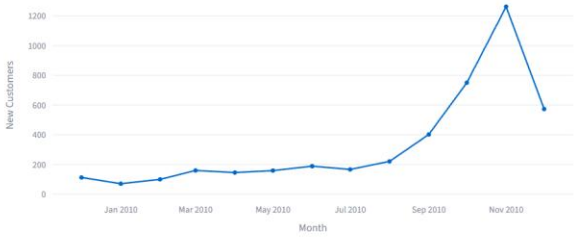
Customer Lifecycle | Sales & Products | Geographical Performance | Churn Drivers & Demographics | Churn Drivers | Attrition Analysis

Deploy

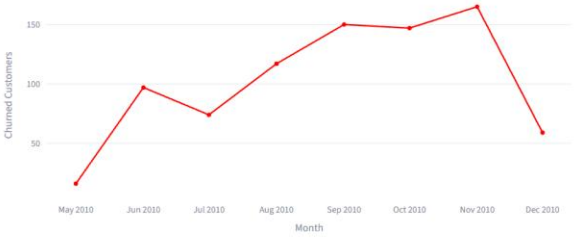
Performance Trends & Deep Dives

Customer Acquisition & Churn Trends

Monthly Acquisitions



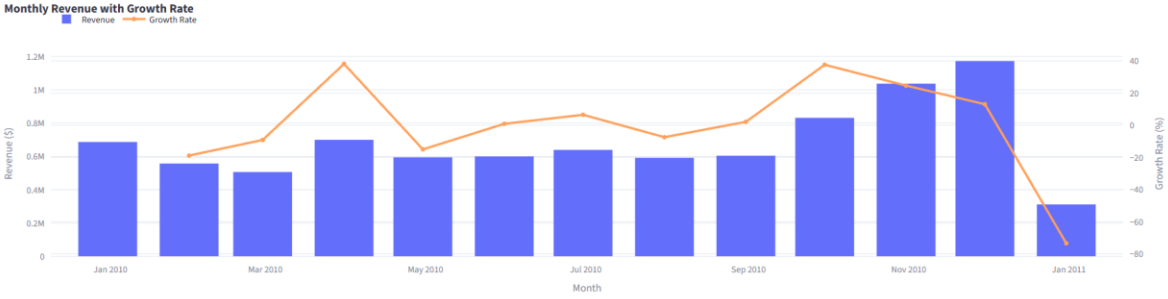
Monthly Churn



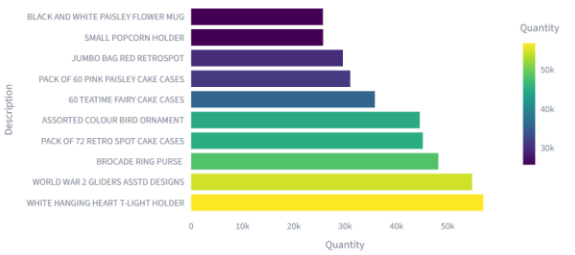
Sales Performance

Deploy

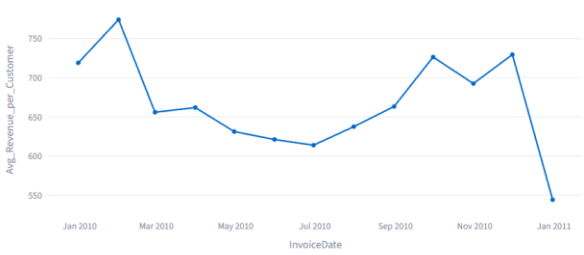
Monthly Revenue with Growth Rate



Top 10 Products by Volume



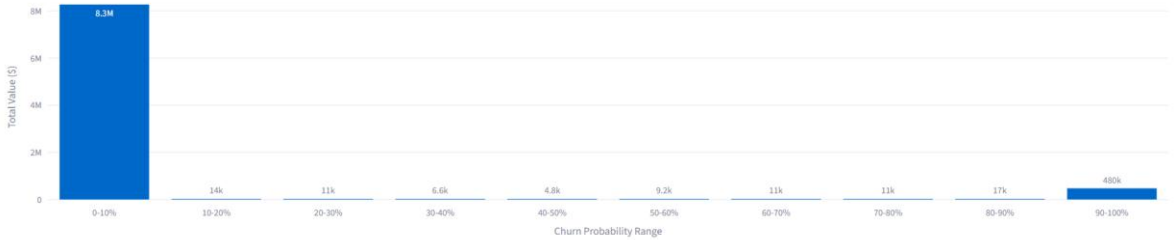
Average Revenue per Customer



Strategic Churn Insights

Value at Risk by Churn Probability

Total Customer Value by Churn Risk Level



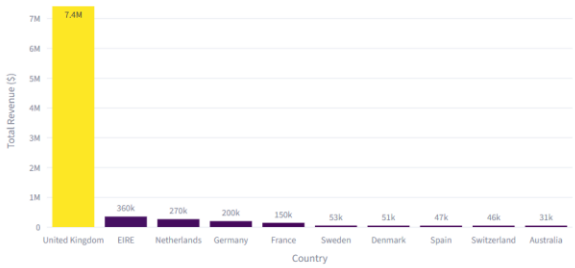
Identifies revenue at risk from potential churn, helping prioritize retention efforts.

Customer Lifecycle | Sales & Products | Geographical Performance | Churn Drivers & Demographics | Churn Drivers | Attrition Analysis

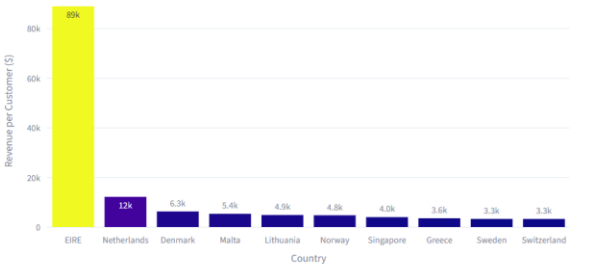
Deploy

Geographical Performance Analysis

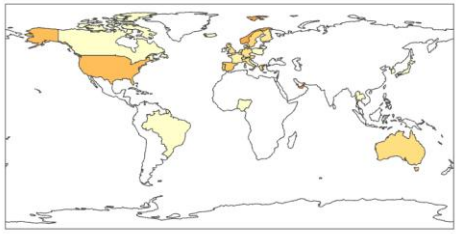
Top Countries by Revenue



Revenue per Customer by Country



Worldwide Customer Churn Rate Heatmap

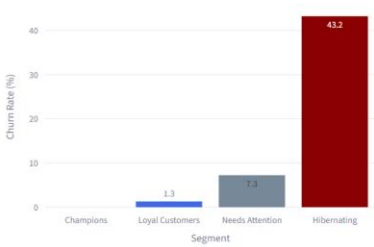


🤔 Churn Drivers & Demographic Insights

This section breaks down churn by key customer attributes to understand who is most likely to leave.

Churn Rate by Customer Segment

Percentage of Churned Customers in Each Segment



Average Tenure by Customer Segment

Average Customer Tenure by Segment

Demographics: Churn Breakdown by Country

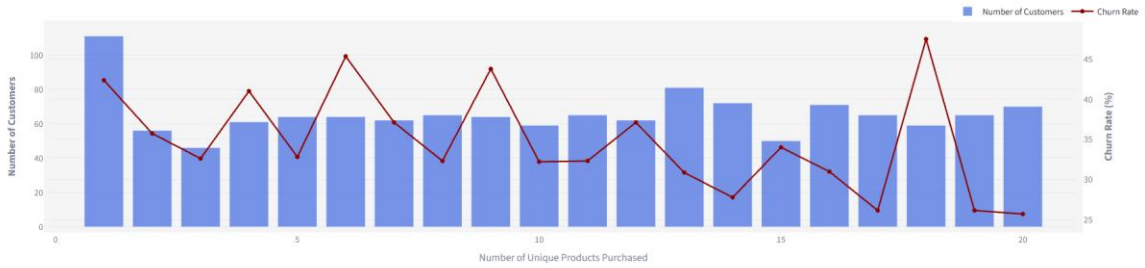
Country	Total_Customers	Accounts Ceased	Active Accounts	Accounts Churned (%)
RSA	1	1	0	100.0%
Bahrain	2	1	1	50.0%
United Arab Emirates	4	2	2	50.0%
Cyprus	7	3	4	42.9%
Norway	5	2	3	40.0%
USA	6	2	4	33.3%
Spain	24	7	17	29.2%
Greece	4	1	3	25.0%
Unspecified	4	1	3	25.0%
Sweden	16	4	12	25.0%

Customer Lifecycle | Sales & Products | Geographical Performance | Churn Drivers & Demographics | Churn Drivers | Attrition Analysis | Deploy

🔬 Strategic Churn Driver Analysis

Does purchasing a wider variety of products increase loyalty?

Churn Analysis by Product Variety



How to Read This Chart: The blue bars show how many customers purchased a certain number of unique products (e.g., exactly 1, 2, 3, etc.). The red line tracks the churn rate for each of those groups.

Key Insight: This chart reveals a powerful trend: **the more varied a customer's purchases are, the less likely they are to churn.** The churn rate is highest for customers who only buy one unique product and drops significantly as customers engage with a wider range of your inventory. This suggests that cross-selling and encouraging product discovery are powerful customer retention strategies.

Dynamic Attrition Trend Analysis

Deploy

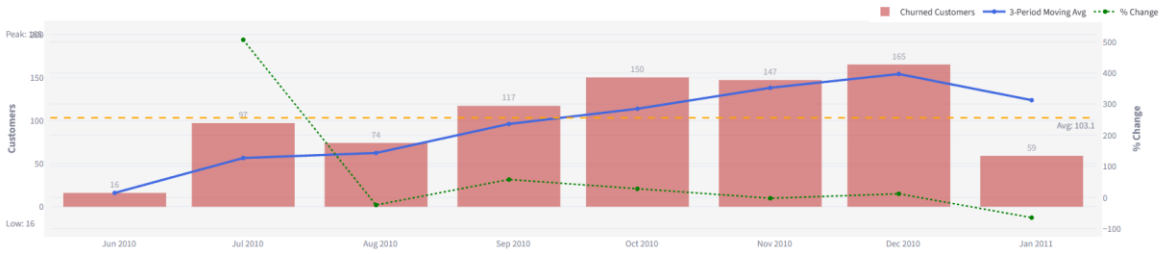
Analyze the trend of churned customers over different time periods with advanced insights.

Select Analysis Period:

○ Daily ○ Weekly ● Monthly ○ Quarterly

Current Monthly Churn	Average Monthly Churn	Change from Previous Period
59	103.1	-64.2%
	↓ -44.1 vs current	

Monthly Attrition Trend with Moving Average and % Change



Key Insights

Trend Analysis

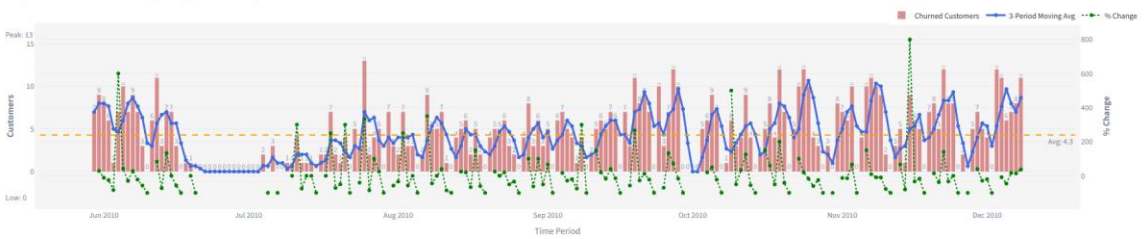
- The decreasing trend based on 3-period moving average
- Recent 3 periods are higher than previous 3 periods
- Highest churn: 165 on 2010-11-30
- Lowest churn: 16 on 2010-05-31

Statistical Summary

- Average churn: 103.1 customers per period
- Standard deviation: 51.5
- Coefficient of variation: 49.9%
- Total churned customers shown: 825

Daily Attrition Trend with Moving Average and % Change

Deploy



Key Insights

Trend Analysis

- The increasing trend based on 3-period moving average
- Recent 3 periods are lower than previous 3 periods
- Highest churn: 13 on 2010-07-25
- Lowest churn: 0 on 2010-06-10

Statistical Summary

- Average churn: 4.3 customers per period
- Standard deviation: 3.6
- Coefficient of variation: 83.3%
- Total churned customers shown: 825


Correlation Heatmap of Key Customer Metrics
Recency
Frequency
MonetaryValue
Churn_Probability
1.00
-0.25
-0.12
0.89
-0.25
1.00
0.66
-0.17
-0.12
0.66
1.00
-0.08
0.89
-0.17
-0.08
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Source Code

04_dashboard > app.py > ...

```python
# Import libraries
import streamlit as st
import pandas as pd
import plotly.express as px
from pathlib import Path
import numpy as np
import plotly.graph_objects as go
from plotly.subplots import make_subplots


# --- 1. PAGE CONFIGURATION ---
st.set_page_config(...

# --- 2. DATA LOADING & ENRICHMENT ---
Tabnine | Edit | Test | Explain | Document
@st.cache_data
def load_and_prepare_data():
    # Define paths using a robust method
    project_root = Path(__file__).parent.parent
    customer_data_path = project_root / "01_data" / "processed" / "final_dashboard_data.csv"
    transaction_data_path = project_root / "01_data" / "processed" / "cleaned_retail_data.csv"

    try:
        # Load the core datasets
        customer_df = pd.read_csv(customer_data_path)
        transaction_df = pd.read_csv(transaction_data_path, parse_dates=['InvoiceDate'])

        # --- Data Enrichment ---
        # Add Country to each customer from their last transaction
        customer_country_map = transaction_df.sort_values('InvoiceDate').drop_duplicates('Customer ID', keep='last')[['Customer ID', 'Country']]
        customer_df = pd.merge(customer_df, customer_country_map, on='Customer ID', how='left')

        # Create 'Segment' Column from RFM_Score
        score_bins = [0, 6, 8, 10, 12]
        score_labels = ['Hibernating', 'Needs Attention', 'Loyal Customers', 'Champions']
        customer_df['Segment'] = pd.cut(customer_df['RFM_Score'], bins=score_bins, labels=score_labels)

```

Figure 23: Dashboard

```python
# Import the pandas library for data manipulation
import pandas as pd

# --- Load the Dataset ---

file_path = '../01_data/raw/online_retail_II.xlsx'

# Load the Excel file into a pandas DataFrame
try:
    retail_df = pd.read_excel(file_path, sheet_name='Year 2009-2010')
    print("✅ Dataset loaded successfully!")
except FileNotFoundError:
    print(f"❌ Error: The file '{file_path}' was not found. Please ensure the path is correct.")



# 1. Display the first 5 rows of the dataframe
print("\nFirst 5 rows of the dataset:")
display(retail_df.head())

# 2. Display the shape of the dataframe (rows, columns)
print(f"\nDataset shape (rows, columns): {retail_df.shape}")

# 3. Display a concise summary of the dataframe, including column data types and non-null counts
print("\nDataset Info:")
retail_df.info()
```

Figure 24: Data Loading

```python
# ============================================================================
# Step 2: Data Cleaning
# ============================================================================
import pandas as pd

# We are assuming 'retail_df' is already loaded from the previous step.

# --- 2.1 Handle Missing Values ---

# First, let's see how many missing values are in each column.
print("--- 2.1 Handling Missing Values ---")
missing_values = retail_df.isnull().sum()
print("Missing values per column:\n", missing_values)
print("-" * 30)

# The 'Customer ID' is crucial for our goal of personalized marketing.
# If the ID is missing, we cannot attribute the purchase to any specific customer.
# Therefore, the best strategy is to remove rows where 'Customer ID' is null.
rows_before = retail_df.shape[0]
retail_df.dropna(subset=['Customer ID'], inplace=True)
rows_after = retail_df.shape[0]
print(f"Removed {rows_before - rows_after} rows with missing 'Customer ID'.")
print(f"Dataset shape after dropping rows: {retail_df.shape}")
print("-" * 30)


# --- 2.2 Correct Data Types ---

# 'Customer ID' should be treated as a whole number (integer) but since pandas
# might have it as a float due to the previous nulls, let's convert it.
print("\n--- 2.2 Correcting Data Types ---")
retail_df['Customer ID'] = retail_df['Customer ID'].astype(int)
print("Converted 'Customer ID' to integer type.")
```

Figure 25: Data Cleaning

```
# --- 3.1 Create a 'TotalPrice' Column ---
# This is a required feature for calculating the Monetary value.
print("--- 3.1 Engineering 'TotalPrice' Feature ---")
retail_df['TotalPrice'] = retail_df['Quantity'] * retail_df['Price']
print("Created 'TotalPrice' column.")
display(retail_df.head())
print("-" * 30)


# --- 3.2 Calculate RFM Values for each Customer ---
print("\n--- 3.2 Calculating RFM Values ---")

# To calculate Recency, we need a "snapshot" date. This will be the day after
# the last transaction in the dataset.
snapshot_date = retail_df['InvoiceDate'].max() + dt.timedelta(days=1)
print(f"Snapshot date for Recency calculation: {snapshot_date}")

# Group data by each customer
# We will calculate Recency, Frequency, and Monetary value for each Customer ID
rfm_df = retail_df.groupby('Customer ID').agg({
    'InvoiceDate': lambda date: (snapshot_date - date.max()).days, # Recency
    'Invoice': 'nunique',                                          # Frequency (count of unique invoices)
    'TotalPrice': 'sum'                                            # Monetary (sum of all purchases)
})

# Rename the columns to be more descriptive
rfm_df.rename(columns={'InvoiceDate': 'Recency',
                       'Invoice': 'Frequency',
                       'TotalPrice': 'MonetaryValue'}, inplace=True)

print("\nCalculated RFM values for each customer.")
print("Displaying the first 5 rows of the RFM DataFrame:")
display(rfm_df.head())
print("-" * 30)
```

Figure 26: Feature Engineering

```
# ==============================================================================
# Step 6: Machine Learning Modeling with K-Means Clustering
# ==============================================================================
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns

# --- 6.1 Load the Processed RFM Data ---
# Load the RFM data we saved in the previous notebook.
try:
    rfm_df = pd.read_csv('../01_data/processed/rfm_customer_data.csv')
    # Set Customer ID as the index if it's not already
    if 'Customer ID' in rfm_df.columns:
        rfm_df.set_index('Customer ID', inplace=True)
    print("✅ RFM data loaded successfully!")
except FileNotFoundError:
    print("❌ Error: Could not find 'rfm_customer_data.csv'. Please ensure the previous steps were run.")

# Select the features for clustering
rfm_features = rfm_df[['Recency', 'Frequency', 'MonetaryValue']]


# --- 6.2 Scale the Data ---
# K-Means is sensitive to the scale of the data. For example, MonetaryValue has a much larger
# range than Frequency. We need to scale the data so all features have a similar importance.
print("\n--- 6.2 Scaling RFM features ---")
scaler = StandardScaler()
rfm_scaled = scaler.fit_transform(rfm_features)
rfm_scaled = pd.DataFrame(rfm_scaled, index=rfm_features.index, columns=rfm_features.columns)
```

Figure 27: K-Means-Clustering

```python
# ============================================================================
# Step 7: Predictive Modeling - Churn Prediction
# ============================================================================
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

print("--- 7.1 Preparing Data for Predictive Modeling ---")
# We are assuming 'rfm_df' is loaded from the previous steps in this notebook

# Define the Churn Status (our target variable 'y')
churn_threshold = 180
rfm_df['Churn_Status'] = (rfm_df['Recency'] > churn_threshold).astype(int) # 1 for Churned, 0 for Active

# Define the features we will use for prediction ('X')
features = ['Recency', 'Frequency', 'MonetaryValue', 'RFM_Score']
X = rfm_df[features]
y = rfm_df['Churn_Status']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
print(f"Data split into {len(X_train)} training samples and {len(X_test)} testing samples.")

print("\n--- 7.2 Training a Logistic Regression Model ---")
# We use Logistic Regression because it's simple, interpretable, and gives probabilities.
model = LogisticRegression(random_state=42)
model.fit(X_train, y_train)
print("Model training complete.")
```

Figure 28: Predictive Modelling

Risk Research Ethics Approval

Project Information

Project Ref	74
Full Name	Karan Bohara
Faculty	Faculty of
Department	School of
Supervisor	Manoj Shrestha
Module Code	ST6047CEM
EFAAF Number	EFAAF
Project Title	Prediction of Retail Purchase Behavior through Customer Engagement Data
Date(s)	Date(s)
Created	Created

Project Summary

This study explores the use of AI, machine learning, and predictive analytics to enhance customer engagement and improve retail decision-making. By integrating and analyzing customer data, businesses can optimize inventory, personalize marketing, and increase customer satisfaction. This research is solely desk-based and does not involve primary data collection from human participants.

Names of Co-Investigators and their organisational affiliation(place of study /employer)	N/A-This is an individual research project.
Is this project externally funded?	No
Are you required to use a Professional Code of Ethical Practice appropriate to your discipline?	No
Have you read the Code?	No

Project Details

What are the aims and objectives of the project?	Analyze AI-driven customer engagement techniques in the retail sector. Examine case studies of AI applications in retail (Amazon, Walmart, Nike). Develop a framework for implementing predictive analytics for retail businesses.
Explain your research design	This study analyzes publicly available secondary data sources, including peer-reviewed academic journals, industry reports, and case studies. A thematic analysis will be used to identify trends and patterns in AI-driven customer engagement techniques. Predictive models such as decision trees and random forests will be evaluated based on prior studies, using criteria such as accuracy, precision, and recall. The analysis will focus on assessing their effectiveness in predicting retail purchase behavior.
Outline the principal methods you will use	1. Conduct a literature review to evaluate AI-driven predictive analytics applications in retail. 2. Analyze case studies from companies like Amazon and Walmart to understand real-world implementation of AI in customer behavior prediction. 3. Compare machine learning models (e.g., decision trees, random forests, neural networks) to determine their accuracy in predicting purchasing behavior, based on previous research findings.

Are you proposing to use a validated scale or published research method / tool?	Yes
Does your research seek to understand, identify, analyse and/or report on information on terrorism or from terrorist organisations, require access to terrorist groups or those convicted of terrorist offences or relate to terrorism policies in other international jurisdictions?	No
Does your research seek to understand, identify, analyse and/or report on information for other activities considered illegal in the UK and/or in the country you are researching in?	No
Are you dealing with Secondary Data? (e.g. sourcing info from websites, historical documents)	Yes
Is this data publicly available?	Yes
Could an individual be identified from the data? e.g. identifiable datasets where the data has not been anonymised or there is risk of re-identifying an individual	No
Are you dealing with Primary Data involving people? (e.g. interviews,questionnaires, observations)	No

Are you dealing with personal data?		No
Please specify what personal data you will be collecting.	This study does not involve the collection of personal data. It relies entirely on publicly available secondary sources, such as industry reports and academic publications.	
Are you dealing with sensitive data (special category data)?		No
Will the Personal or Sensitive data be shared with a third party?		No
Will the Personal or Sensitive data be shared outside of the European Economic Area(EEA)?		No
Is the project solely desk based? (e.g. involving no laboratory, workshop or offcampus work or other activities which pose significant risks to researchers or participants)		Yes
Will the data collection, recruitment materials or any other project documents be in any language other than English?		No
Are there any other ethical issues or risks of harm raised by the study that have notbeen covered by previous questions?		No

DBS (Disclosure & Barring Service) formerly CRB (Criminal Records Bureau)

Question		Yes	No
Does the study require DBS (Disclosure & Barring Service) checks?		×	
If YES, Please give details of the level of check, serial number, date obtained andexpiry date (if applicable)	N/A		
If NO, does the study involve direct contact by any member of the research team with children or young people under 18 years of age?			×
If NO, does the study involve direct contact by any member of the research team with adults who have learning difficulties, brain injury, dementia,degenerative neurological disorders?		×	×
If NO, does the study involve direct contact by any member of the research team with adults who are frail or physically disabled?			×
If NO, does the study involve direct contact by any member of the research team with adults who are living in residential care, social care, nursing homes, re - ablement centres, hospitals or hospices ?			×
If NO, does the study involve direct contact by any member of the research team with adults who are in prison, remanded on bail or in custody?			×
If you have answered YES to any of the questions above please explain the nature of that contact and what you will be doing			

External Ethics Review

Question		Yes	No
Will this study be submitted for ethical review to an external organisation ? (e.g.Another University, Social Care, National Health Service, Ministry of Defence, Police Service and Probation Office)			×
If YES, name of external organisation	N/A – The study does not require external ethics review as it does not involve human participants, medical research, or personal data collection. The research is entirely based on publicly available secondary data sources.		
Will this study be reviewed using the IRAS system?			×
Has this study previously been reviewed by an external organisation?		×	×

Confidentiality, security and retention of research data

Question		Yes	No
What data are you collecting / using / recording?	This study only uses secondary data from publicly available sources, such as academic journals, industry reports, and case studies. No primary data is collected.		
Are there any reasons why you cannot guarantee the full security and confidentiality of any personal or confidential data collected for the study?			×
Please provide an explanation	No personal or confidential data is collected. However, all research materials and data sources will be securely stored using university-approved cloud storage with encrypted access. This ensures compliance with data protection standards and prevents unauthorized access to research materials.		
Is there a significant possibility that any of your participants, and associated persons, could be directly or indirectly identified in the outputs or findings from this study?			×
Please provide an explanation	N/A		
Is there a significant possibility that a specific organisation or agency or participants could have confidential information identified, as a result of the way you write up the results of the study?			×
Please provide an explanation	N/A		
Will any members of the research team retain any personal of confidential data at the end of the project, other than in fully anonymised form?			×
Please provide an explanation	No personal or confidential data is retained beyond the research study, as only secondary sources are used.		
Will you or any member of the team intend to make use of any confidential information, knowledge, trade secrets obtained for any other purpose than the research project ?			×
Please give an explanation	Justification: This research exclusively analyzes publicly available secondary data from academic journals and published case studies. No confidential business information, trade secrets, or proprietary data is accessed or analyzed. All materials used are properly cited and available through legitimate public sources.		
Have you taken necessary precautions for secure data management, in accordance with data protection and CU Policy		×	
Specify location (physical and electronic) where data will be stored	All research materials will be stored in university-approved cloud storage with end-to-end encryption. Backup copies may also be stored on a secured local drive and an encrypted laptop. Access will be strictly limited to the researcher, with role-based permissions and multi-factor authentication (MFA) enabled. The data will be retained for the duration of the study and securely deleted from all storage locations upon project completion, ensuring full compliance with data protection regulations.		
Will you be responsible for destroying the data after study completion?		×	
If NO, who will be responsible for this?			
Please explain how any identifiable and anonymous data will be destroyed	No identifiable or personal data is collected; therefore, no specific data destruction process is required.		
Planned disposal date	30 days after final submission		

Participant Information and Informed Consent

Question		Yes	No
Will all the participants be fully informed BEFORE the project begins why the study is being conducted and what their participation will involve ?			×
Please explain why			
Will every participant be asked to give written consent to participating in the study, before it begins ?			×
If NO, please explain how you will get consent from your participants.If not written consent, explain how you will record consent	This research does not involve human participants, so no consent is required.		
Will all participants be fully informed about what data will be collected, and what will be done with this data during and after the study ?			×
If NO, please specify	No recordings will be made, as this study is based on secondary research.		
Please explain what recordings (audio, visual or both) will be made and how you will gain consent for recording participants			
Will all participants understand that they have the right not to take part at any time, and/or withdraw themselves and their data from the study if they wish?			×
If NO, please explain why			
Will every participant understand that there will be no reasons required or repercussions if they withdraw or remove their data from the study?			×
If NO, please explain why			
Does the study involve deceiving, or covert observation of, participants?			×
Will you debrief them at the earliest possible opportunity?			×
If NO to debrief them, please explain why this is necessary			

Risk of harm, potential harm and disclosure of harm

Question		Yes	No
Is there any significant risk that the study may lead to physical harm to participants or researchers ?			×
If you have answered Yes, please explain how you will take steps to reduce or address those risks. If you have answered No, explain why you believe this is the case	No risk of physical or psychological harm exists as the study does not involve human participants, experiments, or interventions. The research relies solely on publicly available data sources, eliminating any direct impact on individuals or organizations.		
Is there any risk that your study may lead or result in harm to the reputation of the University Group, its researchers or the organisations involved in the study?			×
If you have answered Yes, please explain how you will take steps to reduce or address those risks. If you have answered No, explain why you believe this is the case	Justification: The study examines general retail industry practices using published case studies and academic research. No specific companies will be criticized, and all findings will be presented objectively. The research maintains academic neutrality and does not make evaluative judgments about any organization's practices.		
Is there a risk that the study will lead to participants to disclose evidence of previous criminal offences, or their intention to commit criminal offences?			×
If you have answered Yes, please explain how you will take steps to reduce or address those risks. If you have answered No, explain why you believe this is the case	Justification: As this is a desk-based study analyzing published materials, there is no interaction with human participants who might disclose criminal activities. The research methodology does not create any circumstances where such disclosures could occur.		
Is there a risk that the study will lead participants to disclose evidence that children or vulnerable adults are being harmed, or at risk or harm?			×
If you have answered Yes, please explain how you will take steps to reduce or address those risks. If you have answered No, explain why you believe this is the case	Justification: The research does not involve any direct engagement with children, vulnerable adults, or at-risk populations. The analysis focuses solely on retail customer behavior patterns from aggregated, anonymized datasets available in published research.		
Is there a risk that the study will lead participants to disclose evidence of serious risk of other types of harm ?			×
If you have answered Yes, please explain how you will take steps to reduce or address those risks. If you have answered No, explain why you believe this is the case	Justification: The study's secondary data analysis approach eliminates risks of physical, psychological, or social harm. No experimental interventions are conducted, and all data examined is from previously published studies that have undergone their own ethical reviews.		
Will participants be made aware of the circumstances in which disclosure has implications for confidentiality?			×

Payments to participants

Question	Yes	No
Do you intend to offer participants cash payments or any kind of inducements, or reward for taking part in your study ?		×
If YES, please explain what kind of payment you will be offering(e.g.prize draw or store vouchers)		
Is there any possibility that such payments or inducements will cause participants to consent to risks that they might not otherwise find acceptable ?		×
If YES, please explain)		
Is there any possibility that the prospect of payment or inducements will influence the data provided by participants in any way ?		×
If YES, please explain)		
Will you inform participants that accepting payments or inducements does not affect their right to withdraw from the study at any time ?		×

Capacity to give valid consent

Question	Yes	No
Do you propose to recruit any participants?		×
Do you propose to recruit any participants who are children or young people under 18 years of age?		×
Do you propose to recruit any participants who are adults who have learning difficulties, mental health conditions, brain injury, advanced dementia, degenerative neurological disorders ?		×
Do you propose to recruit any participants who are adults who are physically disabled and cannot provide written and/or verbal consent		×
Do you propose to recruit any participants who are with adults who are living in residential care, social care, nursing homes, reablement centres, hospitals or hospices ?		×
Do you propose to recruit any participants who are with adults who are in prison, remanded on bail or in custody?		×
If you have answered YES to any of the questions above please explain overcome any challenges to gaining valid consent	This study does not require participant recruitment, as it relies entirely on secondary data.	
Do you propose to recruit any participants with possible communication difficulties, including difficulties arising from limited use of knowledge of the English language ?		×
If YES, please explain how you will overcome any challenges to gaining valid consent		
Do you propose to recruit participants who may not be able to fully understand the nature of the study, the foreseen implications or cannot provide consent?		×
If YES, please explain how you will overcome any challenges to gaining valid consent		

Recruiting Participants

Question		Yes	No
Who are the participants?	No participants involved in the study		
How are participants being recruited? Please provide details on all methods of recruitment you intend to use			
Do you foresee any conflict of interest?			✕
Please explain how will this conflict of interest be addressed	No conflicts of interest exist. The research is independent and does not involve funding, sponsorships, or affiliations with commercial retail companies that could influence findings.		

Online and Internet Research

Question	Yes	No
Will any part of your project involve collecting data via the internet or social media?		×
If YES, please explain how you will obtain permission to collect data by these means		
Will this require consent to access?		
If NO, please explain how you will get permission/ consent' to collect this information?		
Will you be collecting data using an online questionnaire/ survey tool? (e.g. BoS, Filemaker)?		×
If YES, please explain which software and how you are ensuring appropriate data security		
Is there a possibility that the study will encourage children under 18 to access inappropriate websites, or correspond with people who pose risk of harm ?		×
If YES, please explain further		
Will the study incur any other risks that arise specifically from the use of electronic media ?		×
If YES, please explain further		

Information gathered from human participants

Question	Yes	No
Primary		
Does your project involve primary data collection from human participants via questionnaires, focus groups, interviews, psychological tests, photography/videography etc.?		×
If YES, Please detail the information to be collected and methods that will be used.		
Is there the possibility of physical or psychological harm to the researcher(s) or the participants?		×
If YES, please explain the possible harm and action taken to reduce/remove the risk		
Are any specific exclusions needed to prevent possible harm to participants (e.g. excluding people with known mental health problems)?		×
If YES, please explain exclusions needed and how these will be carried out		
Are any of the questionnaires or other tests being used in the research diagnostic for specific clinical conditions?		×
If YES, Please explain how you will take steps to reduce or address these risks		