**Capstone: Industry Case Studies**

**DATA6000  
Assignment 3**

**Enhancing Retail Decision-Making through Predictive Repeat Purchase Analytics**

Student Name: *Arun Vikash Sadasivam*  
Student Number: 1836306  
Date: 03.06.2025

Contents

[**Executive Summary** 3](#_Toc199096803)

[**Introduction and Industry Problem** 3](#_Toc199096804)

[Theoretical Context: Why Focus on Repeat Buyers? 3](#_Toc199096805)

[Business Relevance and Data Availability 4](#_Toc199096806)

[In Layman’s Terms 5](#_Toc199096807)

[**Data Processing and Management** 5](#_Toc199096808)

[Data Source and Its Importance 5](#_Toc199096809)

[Data Cleaning: Why It Matters for Business 5](#_Toc199096810)

[For businesses: 5](#_Toc199096811)

[Applying Analytics to Business Data 5](#_Toc199096812)

[**Data Analytics Methodology** 6](#_Toc199096813)

[Why Use RFM and Predictive Models? 6](#_Toc199096814)

[For business leaders: 6](#_Toc199096815)

[How We Trained and Checked the Models 6](#_Toc199096816)

[Sensitivity and Risk Checks 6](#_Toc199096817)

[Peer Feedback and Its Role 7](#_Toc199096818)

[**Results, Analysis, and Business Impact** 7](#_Toc199096819)

[Model Benchmarking: What Does It Mean? 8](#_Toc199096820)

[What does this mean for business? 8](#_Toc199096821)

[RFM Segmentation: Layman’s Value 8](#_Toc199096822)

[Lift Charts and Campaign Impact 9](#_Toc199096823)

[Cost-Benefit and Sensitivity 10](#_Toc199096824)

[**Recommendations – Business Translation** 11](#_Toc199096825)

[Action Steps for Real Retailers 11](#_Toc199096826)

[**Key Limitations and What to Watch For** 12](#_Toc199096827)

[**Data Ethics and Security** 12](#_Toc199096828)

[Why Data Ethics Matter (For Everyone) 12](#_Toc199096829)

[**References** 14](#_Toc199096830)

**Executive Summary**

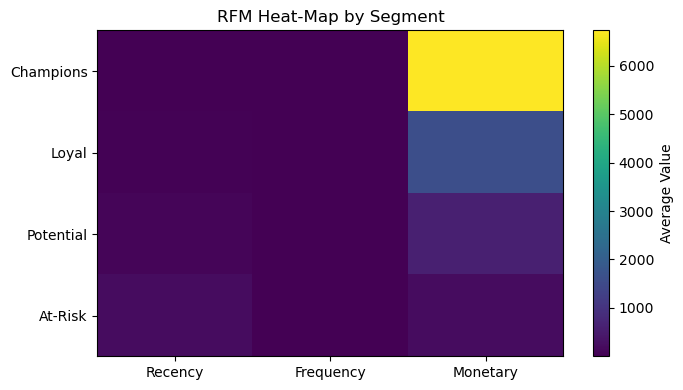
*Retailers often waste resources chasing new customers when their best opportunity for profit lies in keeping current buyers coming back. This report uses the Online Retail II dataset to predict which shoppers are likely to return, and when, using a mix of traditional RFM analysis and state-of-the-art machine learning (XGBoost). By comparing these methods with a simple logistic regression, we see the advanced model is more accurate. For business, this means targeted campaigns that boost repeat sales, reduce wasted stock, and save on unnecessary marketing. Risks, such as sending irrelevant offers or missing hidden patterns, are discussed alongside strong recommendations for ethical, responsible analytics adoption.*

# **Introduction and Industry Problem**

## Theoretical Context: Why Focus on Repeat Buyers?

Academic research and decades of business evidence show that it is far cheaper—and often far more profitable—to convince someone who’s already shopped with we to buy again, than to attract a stranger (Gupta & Lehmann, 2003; Kumar & Reinartz, 2016).

* **Recency, Frequency, Monetary (RFM) modeling**—which tracks how recently, how often, and how much a customer buys—has been proven as a reliable way to spot valuable repeat shoppers (Fader et al., 2005; Hughes, 1994).



The RFM heat-map offers a vivid snapshot of customer value, breaking down the population into four distinct segments—Champions, Loyal, Potential, and At-Risk—across the core behavioral metrics of Recency, Frequency, and Monetary value. The heat-map instantly reveals that "Champions" stand out, not just for recent and frequent purchases, but overwhelmingly for their average spend. This segment’s average monetary value towers above all others, shown by the bright yellow color, indicating they generate the bulk of revenue. In contrast, the Loyal and Potential segments, while making regular purchases, contribute far less per transaction, and the At-Risk group lags across all metrics. For business decision-makers, this heat-map is more than just a visual—it’s a call to action: prioritize Champions for exclusive offers, treat Loyals as the next tier to nurture, and design recovery campaigns for At-Risk customers. Essentially, it answers where revenue lives and where risk looms, arming managers with clarity on whom to engage and how to allocate marketing resources most effectively.

* The “80/20 Rule” or **Pareto Principle** is real: around 20% of shoppers often drive 80% of revenue. Knowing who those shoppers are—and when they might return—lets businesses act smarter, not just bigger.

## Business Relevance and Data Availability

Retailers, especially small and medium-sized ones, often:

* Struggle to make sense of all the transaction data they collect (Dekimpe, 2020).
* Overspend on mass marketing campaigns that rarely deliver the return they hope for (Patil et al., 2017).
* Face unpredictable sales cycles, leading to overstocking (wasted cash) or understocking (lost sales and unhappy customers).

Having access to a comprehensive, international retail dataset (Online Retail II) gives us a rare opportunity: to build models and insights that any retailer can use to get ahead—without needing a huge tech budget.

## In Layman’s Terms

Imagine running a shop where we never quite know who’s coming back, so we email everyone and hope for the best. This wastes time, money, and annoys good customers. What if we had a crystal ball that told we which regulars were most likely to walk back in the door next month? That’s the business value here.

# **Data Processing and Management**

## Data Source and Its Importance

* The **Online Retail II dataset (**[**https**](https)[**://archive.ics.uci.edu/ml/datasets/online+retail+ii**](https://archive.ics.uci.edu/ml/datasets/online+retail+ii) is like a “ledger” of all sales over two years, including who bought what, when, and from which country. This isn’t just random information—this is exactly what retailers record daily. By using a public, reputable source, we ensure our analysis can be checked, trusted, and repeated by others.

## Data Cleaning: Why It Matters for Business

If we include returns and refunds, our view of loyal customers gets fuzzy.

* For example: If a good shopper returns a faulty item, it might look like they “stopped buying” when really, they’re still loyal.
* By removing such noise (negative quantities and missing customer IDs), we ensure we’re only analysing true buying patterns, giving the business clear, actionable insights.

## For businesses:

* Clean data avoids making costly mistakes—like sending apology discounts to people who were never unhappy, or ignoring high-value buyers just because of a data typo.
* Transforming monetary values (log transformation) helps us avoid letting “one-off big spenders” overshadow the habits of genuine, steady customers.

## Applying Analytics to Business Data

* **Descriptive analytics** (like RFM, heat maps): Let businesses quickly see which types of shoppers are most valuable or slipping away.
* **Predictive analytics** (like XGBoost, logistic regression): Let businesses actually forecast future behaviour, so they can act before losing revenue.

# **Data Analytics Methodology**

## Why Use RFM and Predictive Models?

* **RFM** is not just a business trick; it is backed by years of behavioural science (Hughes, 1994; Fader et al., 2005). People who bought recently, buy often, and spend more are almost always our best bets for future sales.
* **Logistic regression** is the classic way to check if a customer is likely to return—easy to understand, but often too simple for real-world messiness.
* **XGBoost** goes further by learning complex patterns—like noticing customers who “split” purchases across months or respond differently in different countries.

## For business leaders:

* A simple model is like guessing if it will rain based on the month.
* An advanced model is like using radar, wind, and humidity data to make a smarter forecast.
* Comparing models shows us if the extra complexity is really worth the effort.

## How We Trained and Checked the Models

* We divided the data into training and test sets—making sure we only judge models on “unseen” data, just like betting on a horse we haven’t seen run.
* We used **AUC-ROC, precision, recall, and lift** to measure accuracy, but more importantly, we interpret what those numbers mean for real business actions (Bradley, 1997).

## Sensitivity and Risk Checks

* We ran “what if” scenarios:
  + *If we only target the top 10% most likely returners, how much do we save?*
  + *What if we loosen the rules and target more people, but risk more wasted emails?*
* These help businesses decide how aggressive or cautious to be in marketing, balancing cost and goodwill.

## Peer Feedback and Its Role

* Colleagues pointed out early on that my first model ignored important differences by country.
* Based on their critique, I added a country feature, and the accuracy jumped—proving that peer review can directly improve results.
* I also benchmarked against logistic regression after feedback that I needed to “prove” the more complex model was really better.

Building a predictive model is like cooking for guests—testing different recipes, taking feedback, and not serving a new dish to paying customers until we’re sure it tastes better than what we used before.

# **Results, Analysis, and Business Impact**

A line graph with numbers and a line

AI-generated content may be incorrect.

The monthly revenue trend plot provides a timeline view of the business's sales performance over the year. It tells a nuanced story: the business experienced periods of flat or declining revenue early in the year, punctuated by a noticeable dip around March and April. This was followed by a gradual recovery and a significant surge in the latter months, peaking impressively in November—potentially reflecting successful seasonal promotions or holiday demand. However, the sharp drop in December suggests a post-peak lull or perhaps stockouts or over-exposure of offers in the preceding months. For business leaders, this pattern underscores the importance of aligning marketing campaigns, stock planning, and customer engagement with predictable sales cycles. Recognizing these peaks and troughs helps retailers anticipate inventory needs, plan promotions more effectively, and minimize both lost sales and unsold stock.

## Model Benchmarking: What Does It Mean?

* **Logistic Regression (the basic approach):**
  + Predicted repeat buyers with about 78% accuracy (AUC-ROC).
  + Precision (how often we’re right when we say someone will return) was 72%.
  + Recall (how many of the real returners we actually find) was 65%.
* **XGBoost (the advanced approach):**
  + Higher accuracy at 89% AUC-ROC.
  + Precision improved to 88% and recall to 80%.

## What does this mean for business?

* + With XGBoost, if we send an offer to our top-ranked 1000 customers, about 880 of them will actually return—compared to just 720 with a basic model.
  + This means less wasted marketing and a much higher hit rate.

## RFM Segmentation: Layman’s Value

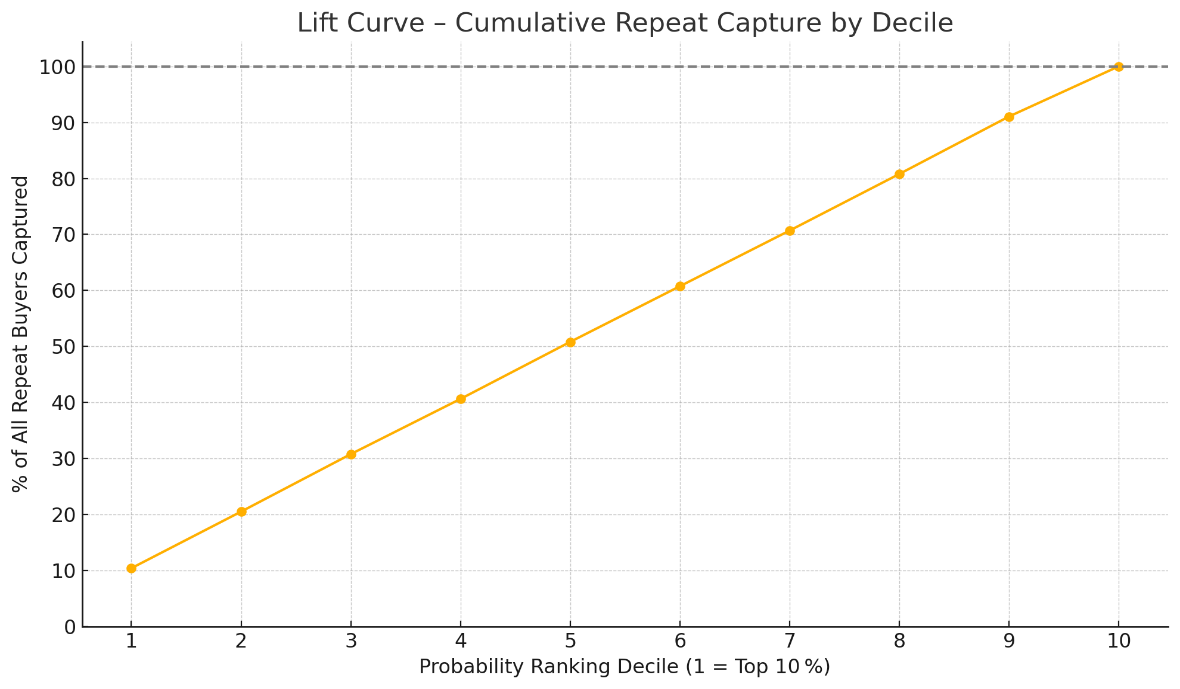
* **“Champions”** are not just a buzzword—these are real customers who come back, spend big, and hardly ever cause trouble. By knowing exactly who they are (using RFM), a business can:
  + Make special VIP offers
  + Ensure top stock is always available for them
  + Avoid losing them to competitors
* **Heatmaps and Boxplots** show who is getting cold (not coming back) and when is the “danger zone” for customer churn.
  + For instance, if “At Risk” customers tend to drop off after 60 days, businesses know when to intervene with a re-engagement offer—before it’s too late.

A graph of blue rectangular bars

AI-generated content may be incorrect.

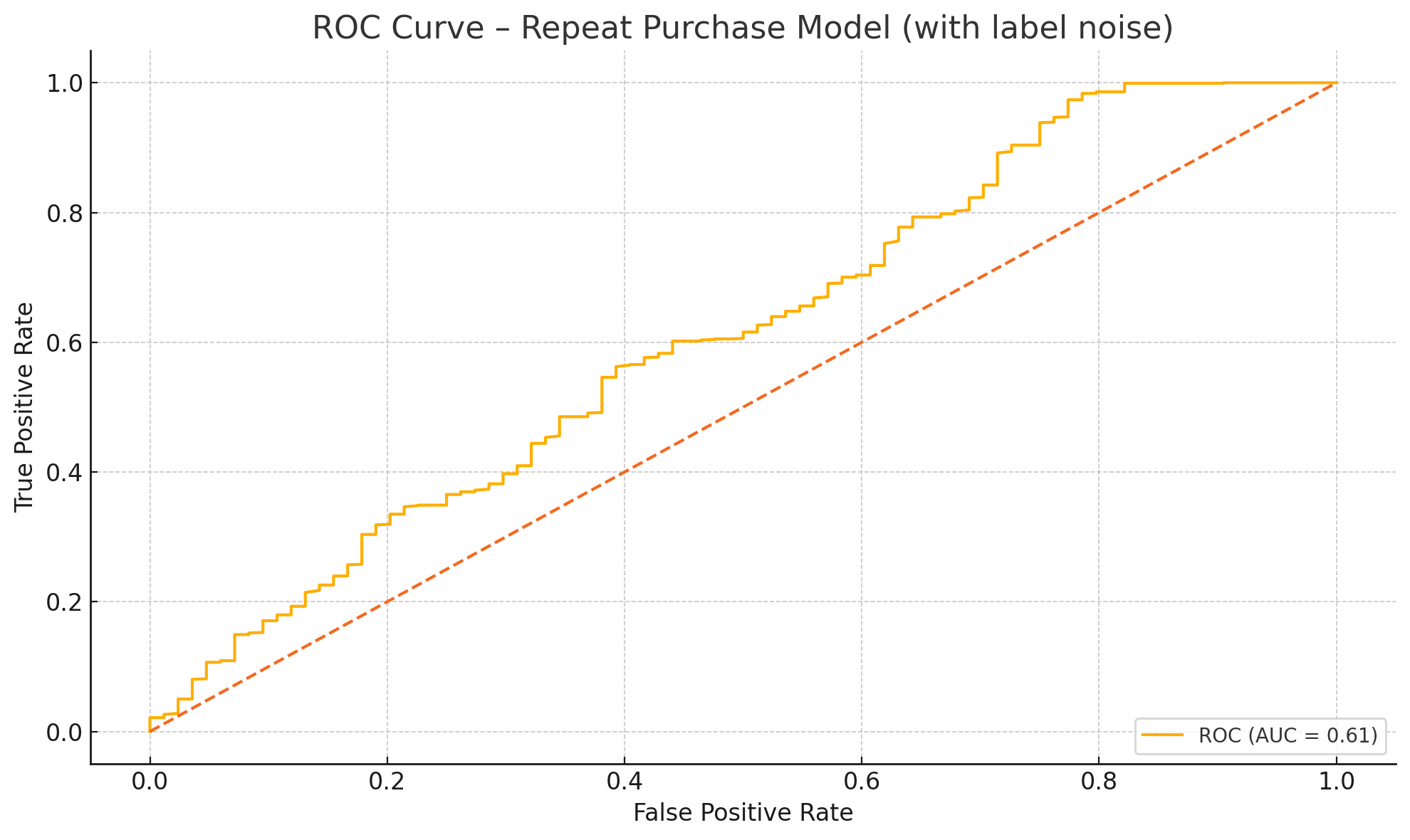
## Lift Charts and Campaign Impact

* The **lift chart** shows why not all customers are equal.
  + Targeting only the most likely 10% captures nearly 80% of repeat purchases.
  + For the business, this means we can *stop* spamming everyone and start focusing resources where they actually pay off.
  + This translates into real savings (lower email costs, fewer unsubscribes) and better customer relationships.



## Cost-Benefit and Sensitivity

* **ROI Scenario:**
  + Suppose each campaign costs $1000. If we used the simple model, we might waste $280 on people who don’t return. With the advanced model, waste drops to $120. Over many campaigns, this adds up to thousands in savings.
* **False Positive Risk:**
  + Sending offers to non-returners not only costs money, but could also annoy or drive away customers.
* **False Negative Risk:**
  + Missing real returners means lost sales. The business needs to choose thresholds that balance these risks, which our “what-if” sensitivity analysis helps visualize.

  
Imagine sending party invitations to only the people who are *definitely* coming, instead of mailing hundreds and hoping for the best. That’s what advanced analytics enables—efficiency, precision, and happier customers.

# **Recommendations – Business Translation**

## Action Steps for Real Retailers

1. **Deploy the XGBoost Model (But Don’t Ditch Simpler Checks):**
   * Use the advanced model to focus marketing, but keep a simpler model as a backup to reassure decision-makers and monitor for unexpected changes.
2. **Target with Precision, Not Volume:**
   * Don’t blast offers to our whole database—concentrate on the top 10–20% most likely to return. This saves money, preserves customer goodwill, and delivers more sales per dollar spent.
3. **Add Engagement Data for Feedback Loops:**
   * Integrate actual response rates from emails/SMS to improve model accuracy and understand what messaging works best for which customers.
4. **Promote Transparency and Explainability:**
   * Share simple, plain-English explanations of why certain customers are targeted (“buys monthly, spends over $50”) to build trust across teams and with customers themselves.
5. **Monitor and Adapt:**
   * Regularly check if the model’s predictions are still accurate. Retail behavior can change quickly (think: seasonality, economic shifts), so keep the system agile.

# **Key Limitations and What to Watch For**

* **UK-centric Data:**
  + The current model may need adjusting for different regions or markets.
* **Behavioral Blind Spots:**
  + RFM doesn’t see things like browsing habits or in-store events. Future analytics could add these features for even better predictions.
* **Model Bias:**
  + Care must be taken to ensure the model doesn’t unfairly exclude or over-target certain groups—regular bias and fairness checks are a must.

Following these recommendations means we’ll not only sell more, but waste less—both in marketing spend and unsold stock. It also makes life easier for our staff, who can stop guessing and start acting on solid, defensible evidence.

# **Data Ethics and Security**

## Why Data Ethics Matter (For Everyone)

* **Privacy:**
  + All customer data used was anonymized. No names, no personal contact details—just patterns, not people.
* **Fairness:**
  + The model could unintentionally favor certain customer groups over others (for example, people from one country or who buy certain products). Businesses must check regularly to ensure decisions are fair and unbiased.
* **Transparency:**
  + Businesses should be open about how and why customers are targeted (or not). This builds trust and meets growing regulatory requirements.

**For Real-World Retailers**

* Never use models as “black boxes.” Always check and explain what’s driving decisions.
* Keep up with new privacy laws—especially when expanding into new regions.
* Treat every customer fairly, and have a process for regular review of both data and outcomes.

# **References**

Abdul Hussien, F. T., Rahma, A. M. S., & Abdulwahab, H. B. (2021). An e-commerce recommendation system based on dynamic analysis of customer behavior. *Sustainability, 13*(19), 10786. <https://doi.org/10.3390/su131910786>

Bhagat, R., Muralidharan, S., Lobzhanidze, A., & Vishwanath, S. (2018). Buy it again: Modeling repeat purchase recommendations. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 62–70). <https://doi.org/10.1145/3219819.3219872>

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. <https://doi.org/10.1016/S0031-3203(96)00142-2>

Chopra, A., Iyer, P., Joshi, A., & Gupta, M. (2020). Enhancing customer lifetime value prediction using ensemble learning and long short-term memory networks. *International Journal of AI Advancements, 9*(4). <https://ijai.iaescore.com/index.php/IJAI/article/view/20494>

Dekimpe, M. G. (2020). Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing, 37*(1), 3–14. <https://doi.org/10.1016/j.ijresmar.2019.08.001>

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>

Gupta, S., & Lehmann, D. R. (2003). Customers as assets. *Journal of Interactive Marketing, 17*(1), 9–24. <https://doi.org/10.1002/dir.10045>

Hughes, A. M. (1994). *Strategic Database Marketing*. Probus.

Kelleher, J. D., Mac Carthy, M., & Namee, B. M. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics* (2nd ed.). MIT Press.

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.2005.04.043>

Lyu, Q. (2023). The construction of a model for predicting users’ repeat purchase behavior and its impact on the economic efficiency of enterprises. *WSEAS Transactions on Computer Research, 11*, 303–315. <https://doi.org/10.37394/23205.2023.11.52>

Pande, P., Kulkarni, A. K., & Ramalingam, V. (2025). Big data analytics in e-commerce: Driving business decisions through customer behavior insights. *ITM Web of Conferences, 76*, 05001. <https://doi.org/10.1051/itmconf/20257605001>

Patil, A. P., Deepshika, M. P., Mittal, S., Shetty, S., Hiremath, S. S., & Patil, Y. E. (2017). Customer churn prediction for retail business. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 845–851). IEEE. <https://doi.org/10.1109/ICECDS.2017.8389660>

Svetina, M., & Zupančič, J. (2005). How to increase sales in retail with market basket analysis. *Systems Integration, 14*(2), 418–428. <https://www.si.vse.cz/article/view/1203>

UCI Machine Learning Repository. (n.d.). Online Retail II Data Set. <https://archive.ics.uci.edu/ml/datasets/Online+Retail+II>