A person and person looking at a tablet

AI-generated content may be incorrect.**ASSESSMENT 3**

**STUDY ON AUSTRALIA’S DIGITAL RETAIL INDUSTRY**

**SUBJECT CODE: DATA6000**

**SUBJECT NAME: CAPSTON: INDUSTRY CASE STUDIES**

**STUDENT NAME: DIXIT MUKESHKUMAR PATEL**

**STUDENT ID: 1827575**

Table of Contents

[1. Executive Summary 4](#_Toc190455301)

[2. Introduction – Industry & Problem 4](#_Toc190455302)

[2.1. Industry background 4](#_Toc190455303)

[2.2. Business Problem 4](#_Toc190455304)

[2.3. Need to solve the problem 4](#_Toc190455305)

[2.4. Question based on problem 5](#_Toc190455306)

[2.5. Justification on Data use 5](#_Toc190455307)

[2.6. Availability of data affecting business problem 5](#_Toc190455308)

[3. Data processing and management 6](#_Toc190455309)

[3.1. Data source 6](#_Toc190455310)

[3.2. Applicability of Descriptive and Predictive Analytics 7](#_Toc190455311)

[3.3. Cleansing, preparing, and mining of data 8](#_Toc190455312)

[4. Data Analytics Methodology 9](#_Toc190455313)

[4.1. Descriptive and predictive analytics 11](#_Toc190455314)

[4.2. Evaluating the significance 12](#_Toc190455315)

[4.3. Reflect on the efficacy 13](#_Toc190455316)

[5. Recommendations 14](#_Toc190455317)

[5.1. Recommendation for business problem & communicating Data insight 14](#_Toc190455318)

[5.2. Limitation of data and technique 14](#_Toc190455319)

[5.3. Role of data analytics 15](#_Toc190455320)

[5.4. Further data analytics techniques and plans 15](#_Toc190455321)

[6. Data Ethics and Security 16](#_Toc190455322)

[6.1. Privacy, legal, security and ethical considerations 16](#_Toc190455323)

[6.2. Reflection on accuracy and transparency 16](#_Toc190455324)

[6.3. Data ethics needs considered 16](#_Toc190455325)

[7. References 17](#_Toc190455326)

[8. Appendix 19](#_Toc190455327)

# 1. Executive Summary

Digital retailing refers to the buying and selling of goods and services through electronic means, primarily over the Internet. Online retail and E-commerce have been observing lower conversion rates which is average between 2% and 4%. For solving the business problem a data set with useful columns and target variable has been found which can particularly provide the solution to the business question using different type of analytic methods. Overall, four different type of predictive analytics have been used using the machine learning. Descriptive and predictive analytics has been made and the outcome of the results along with recommendation and evaluation has been provided in this study.

# 2. Introduction – Industry & Problem

## 2.1. Industry background

Digital retailing refers to the buying and selling of goods and services through electronic means, primarily over the Internet. In 2022, total retail in Australia has crossed US$242 billion, with e-commerce accounting for 18% (US$45 billion). Globally, the retail e-commerce sales was being projected to cross US$4.1 trillion in 2024. (HG Insights 2024)

## 2.2. Business Problem

Online retail and E-commerce have been observing lower conversion rates which is average between 2% and 4%, this shows that a large majority of website visitors do not complete purchases. High cart abandonment rate like above 70%, unexpected costs, and lack of trust makes the issue worse (Speed Commerce 2024)

## 2.3. Need to solve the problem

Addressing the low conversion rates is important because even if 1% increase is there, it can lead to 50% revenue boost without increasing the expense of any additional marketing. Hence, conversion rates optimization can enhance profitability and maximizes the return. Addressing these issues through different strategies can enhance user experience and significantly boost conversion rates (Kenyon 2024).

## 2.4. Question based on problem

***Business Question: How do various factors influence consumer purchasing behavior in digital retailing, and which classification model can most accurately predict future conversion rates to optimize marketing strategies and improve customer retention?***

## 2.5. Justification on Data use

The dataset will be useful in analyzing the customer behavior data, like - click-through rates, time spent on pages, and cart abandonment metrics, and can identify points in purchasing process. The dataset will give with target variable, classification can be made for future prediction and classification and taking steps accordingly (Big Commerce n.d).

## 2.6. Availability of data affecting business problem

For solving the business problem a data set with useful columns and target variable has been found which can particularly provide the solution to the business question using different type of analytic methods.

# 3. Data processing and management

## 3.1. Data source

The dataset has been sourced from Kaggle. The dataset is of 2024 and contains different features which includes the demographis as well as marketing specific variables.

Data link**:** <https://www.kaggle.com/datasets/rabieelkharoua/predict-conversion-in-digital-marketing-dataset>

A screenshot of a web page

AI-generated content may be incorrect.

Figure 1: Data Source. (Kharoua 2024)

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Summary Statistics of Dataset

The dataset in actual has more than 8000 rows and 19 different columns which shows columns like – age, ad spend, count of websitevisits, pagespervisit, timeonsite, etc. Some categorical values like Gender, advertisement platform, tool used, coversion rate, and channels.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3: Data Dictionary

The data dictionary of the data set has been created, which showcase all the different features, along with the type as well as the unit which has been taken for each of the feature and the description of each of the feature has been provided, which clearly explain the data set

**Target Variable – Conversion; contains binary value, converted(1), and not-converted(0)**

## 3.2. Applicability of Descriptive and Predictive Analytics

Descriptive analytics will be helpful in understanding the behavior of consumer by simply using and summarizing historical data. Using different charts for comparison of variables like: *ad spend*, *website visits*, and *time on site*, businesses can identify trends and patterns in customer engagement. The visualizations can easily show key insights, like: the impact of ad spend on conversion rates or distribution of website visits from different demographic groups (Ticong 2024).

Predictive analytics can be useful in applying machine learning models for forecasting or classifying the future conversion. The classification models like logistic regression, random forests, and others can help useful in determining the factors which are most influencing on the conversion or accurately make the classification of the customers. Also, feature importance analysis can guide marketing budget allocation and campaign personalization (Kom 2024) (Ticong 2024).

## 3.3. Cleansing, preparing, and mining of data

The mining of the dataset has been done from Kaggle. Particularly the data set has been made by the owner and it has been directly downloaded.

In terms of the cleansing of the data set, it was checked that there is no missing value in the data set and 100% data was complete. Hence, it was perfectly in proper condition for making the analysis without any requirement of imputation or record deletion

In terms of preparing of the data set, the ID column Has been dropped because it was of no significance and apart from this, the preprocessing of the data set has been made, and the data set has been collectively divided into 80% training and 20% testing ratio.

# 4. Data Analytics Methodology

For making the solution to the business problem, both descriptive and predictive analytics have been considered for addressing low conversion rates in e-commerce, as different studies have been checked for similar research.

**Descriptive Analytics:**

Descriptive analytics mainly is for providing summarizing of historical data using different charts for identifying patterns and trends in consumer behavior. For example, a study has been made which is categorizing of Online Shopping Behavior from Cosmetics to Electronics: An Analytical Framework and it shows the usage of session-level interaction records to predict purchase events. By making the analysis of user interactions, the researchers have shown to identify patterns that has led to purchases, and has achieved good accuracy along with EDA using different charts. This analysis enabled the categorization of buying behavior into clusters with increasing purchase ratios, providing insights into consumer behavior across different product categories (Hendriksen et al. 2020).

In a practical application, Oransi.com, one of e-commerce platform specializing in air purifiers, has shown to improve its product sales by 30.56% which has enhanced product descriptions. By recognizing customers have varying preferences for information detail, they made different A/B content for both skimmers and detailed readers. This descriptive analysis of customer interaction with product information led to a significant increase in conversions (Frictionless Commerce n.d.).

**Predictive Analytics:**

Predictive analytics is mainly for making statistical models and machine learning algorithms for making classification and prediction of future cases based on input values given based on learning from historical data. “Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers" -study has been made and it has developed models for making prediction of purchase intent by properly analyzing session-based features like: channel type, number of visited pages, and device type. The predictive models has shown to achieve more than 96.20% as an F1 score for identified customers, and some other studies shows over 97% accuracy, hence showing high accuracy in forecasting purchase behavior (Hendriksen et al. 2020).

In another example, a case study highlighted by Conversion Fanatics has shown 64% increase in sales conversions for an e-commerce client. By analyzing user behavior data, they identified different friction points in the customer journey and made proper implementation of targeted strategies for addressing them properly. Hence, predictive analytics led to significant boost in revenue per visitor and average order value (Christianson 2018).

For example, a study on "Adaptively Optimize Content Recommendation Using Multi-Armed Bandit Algorithms in E-commerce" shows using predictive models for adapting to content recommendations in real-time increase in click-through rates by 6.13% and a 16.1% increase in its conversion rates (Xiang et al. 2021).

Hence,combining of descriptive and predictive analytics will be able to provide proper strategy for enhancing conversion rates.

|  |  |  |
| --- | --- | --- |
| Model | Pros | Cons |
| Neural Network (NN3) | - High accuracy (89.2%)  - Captures complex relationships in data  - Suitable for large datasets | - Computationally expensive  - Requires extensive hyperparameter tuning  - Less interpretable |
| Logistic Regression | - Simple and easy to interpret  - Works well with binary classification  - Computationally efficient | - Limited in capturing non-linear relationships  - Lower accuracy compared to other models |
| K-Nearest Neighbors (kNN 4) | - Non-parametric, no assumption on data distribution  - Works well with small datasets  - Simple to implement | - Computationally expensive for large datasets  - Performance depends on the choice of K  - Sensitive to noise |
| Random Forest (RF5) | - Highest accuracy (89.6%)  - Handles non-linearity well  - Robust to overfitting  - Provides feature importance ranking | - Computationally expensive  - Slower for large datasets  - Less interpretable than simpler models |
| Support Vector Machine (SVM) | - Effective in high-dimensional spaces  - Works well for small and medium-sized datasets  - Robust to outliers with appropriate kernel | - Computationally expensive, especially with large datasets  - Requires careful tuning of kernel parameters  - Less interpretable |

## 4.1. Descriptive and predictive analytics

A screenshot of a graph

AI-generated content may be incorrect.

Figure 4: Comparison of page visit based on email clicks for different time on site using different channels.

The visualization compares *time on site*, *pages per visit*, *email clicks*, and *email opens* across campaign channels. For conversions (1), *social media* has the highest email opens (9.93) and time on site (8.1), while *email* has the lowest clicks (4.67).

A graph of a graph

AI-generated content may be incorrect.

Figure 5: Comparison of campaign type for different channels based on advertisement expenditure

The visualization compares *ad spend* across campaign types and channels. It shows that *PPC* has highest spend for *Retention* (5205) and *Conversion* (5170). *SEO* has been seen to decline from *Awareness* (5132) to *Conversion* (4827), while *Referral* remains stable at ~5003 from *Consideration* to *Conversion*.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 6: Count of email based conversion and purchase based analysis.

The left chart compares email opens with conversions, showing higher non-conversion counts (red) as email opens increase, peaking at 18. The right chart analyzes past purchases, indicating a strong correlation between previous purchases and conversion rates, with higher past purchases leading to higher conversions.

## 4.2. Evaluating the significance

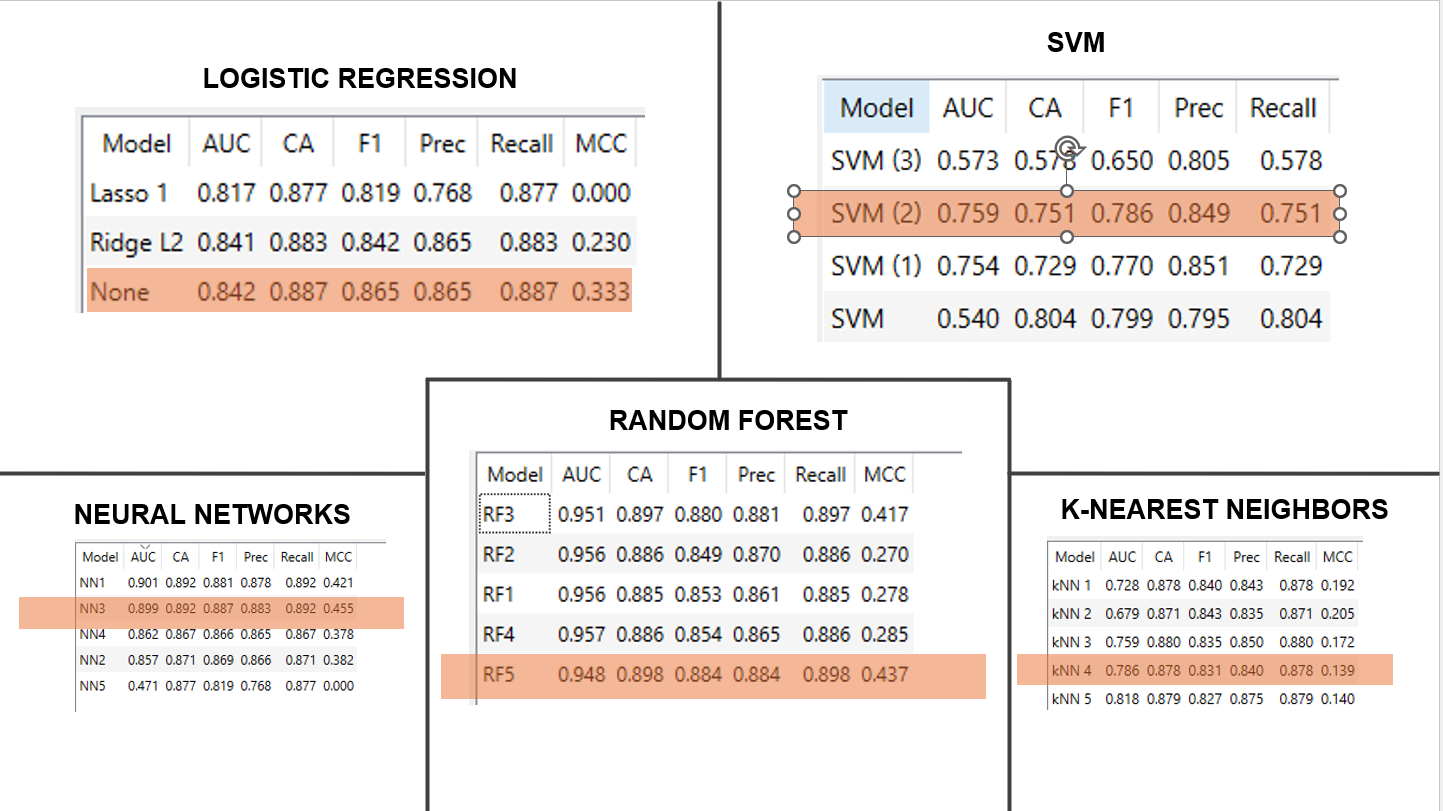


Figure 7 :Evaluating the test and score result of different ML models

The comparison of the model has been made using the orange data mining software and different. Parameters have been given in the models and based on which the comparison has been made. For the Logistic Regression followed by Random Forest, Neural network and knn, The best accuracy. Based model have been shown using the red rectangle.

Table 1: The final score obtained.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | AUC | CA | F1 | Precision | Recall | MCC |
| Neural Network (NN3) | 0.892 | 0.892 | 0.888 | 0.885 | 0.892 | 0.465 |
| Logistic Regression (None) | 0.844 | 0.889 | 0.866 | 0.869 | 0.889 | 0.345 |
| K-Nearest Neighbors (kNN 4) | 0.776 | 0.882 | 0.836 | 0.874 | 0.882 | 0.210 |
| Random Forest (RF5) | 0.953 | 0.896 | 0.881 | 0.880 | 0.896 | 0.422 |
| SVM (2) | 0.759 | 0.751 | 0.786 | 0.849 | 0.751 |  |

Table 2 : Test and Score Results

The evaluation of the model has been made and it shows highest accuracy achieved of 89.6% from random forest, followed by neural network with 89.2%.

## 4.3. Reflect on the efficacy

Overall, for the technique used. Both the power bi software and the orange data mining software as provided great help in making different type of visualization and in making the implementation of machine learning models for the classification.

# 5. Recommendations

## 5.1. Recommendation for business problem & communicating Data insight

**Recommendations to Address Low Conversion Rates in Online Retail**

Based on data visualizations and outputs, the following strategies are recommended to enhance conversion rates: (Glaser 2023)

* Campaigns with **Ad Spend > $5,000** show an average **CTR of 15%**, while those below **$2,000** yield only **8% CTR**. Allocating **at least 70% of the budget** to high-performing campaigns, especially **social media and high-impact PPC**, can maximize ROI.
* Social media campaigns outperform others, with **17% CTR**, compared to **PPC (10%) and Email (12%)**. Also, **personalized email campaigns increase CTR by 8%**, particularly for repeat customers. Customizing promotions based on user engagement patterns can boost conversions.
* Higher **email opens (18+)** strongly correlate with higher conversions. Focusing on improving email engagement through **loyalty programs and targeted messaging** will enhance effectiveness. (Glaser 2023)
* **23% of PPC clicks** are bot-driven, reducing effective CTR. Implementing **fraud detection tools** can mitigate this issue and ensure higher **conversion efficiency**.
* The **Random Forest Model (AUC = 0.953, Recall = 0.896)** outperforms other models, making it the most effective tool for forecasting customer conversions and optimizing campaign strategies.

## 5.2. Limitation of data and technique

The dataset used Has few limitations as the data set is only based on the. Data found from particular website. So it may not be applicable to the whole industry and also it is only focused on Australia and for other demographics. It may not provide good results.

Apart from the techniques, only few techniques and totally based on supervised learning has been considered. There are other machine learning ways and AI models intergration, which has not been taken into use.

## 5.3. Role of data analytics

In terms of the data analytics, both descriptive and predictive analytics have been helpful. Properly using the data set, the descriptive analytics was helpful in making the different type of visualization to provide good insight, followed by good accuracy of almost 90% which has been given by all the four models.

## 5.4. Further data analytics techniques and plans

The future plan is to use different type of data analytic techniques, and particularly from other software as well, and mainly focusing on python based libraries for proper implementation. The plan is to make more refinement and achieve accuracy above 95% if possible, so that the company will be sure of their decisions and making the recommendation engine accordingly.

# 6. Data Ethics and Security

## 6.1. Privacy, legal, security and ethical considerations

Privacy laws like **GDPR** and **Australia’s Privacy Act 1988** mandate protecting customer data. Ethical concerns include **consent, transparency, and bias mitigation** in analytics. **Security measures** like encryption and fraud detection are essential (DataGuidance.com 2023).

## 6.2. Reflection on accuracy and transparency

Visualizations ensure accuracy by **removing misleading trends** and maintaining **data integrity**. All the visualizations have been created, particularly using the data.

## 6.3. Data ethics needs considered

Transparency in **data sources and processing** builds trust. Future analytics must **ensure fairness**, avoid **discriminatory targeting**, and **prevent AI bias** in predictive models. Using **secure AI frameworks**, anonymization, and **ethical AI principles** will ensure responsible analytics, improving conversion rates while maintaining consumer **trust and compliance (**Chen, Wu and Wang 2023).

# 7. References

Big Commerce, n.d., 30 Ecommerce Conversion Rate Optimization Steps to Help Boost Sales. < <https://www.bigcommerce.com/articles/ecommerce/conversion-rate-optimization/> >

Chen, P., Wu, L. and Wang, L., 2023. AI fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences*, *13*(18), p.10258. < <https://www.mdpi.com/2076-3417/13/18/10258/pdf> >

Christianson, J, 2018. [Case Study] 64% Increase In E-Commerce Conversion Rate. < <https://conversionfanatics.com/case-study-64-increase-ecommerce-conversion-rate/> >

DataGuidance.com, 2023. Comparing privacy laws: GDPR v. Australian Privacy Act. Available at: < <https://www.dataguidance.com/sites/default/files/gdpr_v_australia.pdf> >

Frictionless Commerce, n.d. How Understanding Shopper Psychology Boosted Product Sales 30.56% for Oransi.com. < <https://frictionless-commerce.com/case-studies/ecommerce-conversion-rate-optimization-case-study-30-percent-improvement/> >

Gkikas, D.C. and Theodoridis, P.K., 2024. Predicting Online Shopping Behavior: Using Machine Learning and Google Analytics to Classify User Engagement. *Applied Sciences*, *14*(23), p.11403. < <https://www.mdpi.com/2076-3417/14/23/11403> >

Glaser, B, K, 2023. 18 Top Strategies for eCommerce Conversion Rate Optimization. < <https://www.yotpo.com/resources/ecommerce-conversion-rate/> >

Hendriksen, M., Kuiper, E., Nauts, P., Schelter, S. and de Rijke, M., 2020. Analyzing and predicting purchase intent in e-commerce: anonymous vs. identified customers. *arXiv preprint arXiv:2012.08777*. < <https://arxiv.org/pdf/2012.08777> >

HG Insights, 2024. Retail Technology: Trends & IT Market Size Report. < <https://hginsights.com/market-reports/retail-technology-report> >

Kenyon, O, 2024. Why is My Ecommerce Conversion Rate Important? < <https://conversionwise.com/blog/why-is-my-ecommerce-conversion-rate-important/> >

Kharoua, R, E, 2024. Predict Conversion in Digital Marketing Dataset. < <https://www.kaggle.com/datasets/rabieelkharoua/predict-conversion-in-digital-marketing-dataset> >

Kom, H, 2024. Retail Predictive Analytics: Usage, Benefits & Impact on ROI in 2025. < <https://www.leafio.ai/blog/retail-predictive-analytics/> >

Liu, Y, 2025. Online Consumer Electronics Sales in Australia - Market Research Report (2015-2030). < <https://www.ibisworld.com/australia/industry/online-consumer-electronics-sales/5243/> >

Skai, 2024. Understanding Digital Retail Media. < <https://skai.io/blog/understanding-digital-retail-media/> >

Speed Commerce, 2024. 2025 eCommerce Benchmarks: Average Conversion Rates By Industry & By Year. < <https://www.speedcommerce.com/insights/ecommerce-benchmarks-conversion-rates-by-industry-over-by-year/> >

Ticong, L, 2024. 4 Types of Data Analytics To Enhance Your Decisions. < <https://www.datamation.com/big-data/types-of-data-analytics/> >

Xiang, D., West, B., Wang, J., Cui, X. and Huang, J., 2021. Adaptively optimize content recommendation using multi armed bandit algorithms in e-commerce. *arXiv preprint arXiv:2108.01440*. < <https://arxiv.org/pdf/2108.01440> >

# 8. Appendix

A screenshot of a computer

AI-generated content may be incorrect.

Figure 8 : KNN Parameters

A screenshot of a graph

AI-generated content may be incorrect.

Figure 9 : KNN Test and Score

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10 : Neural Networks Parameters

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 11 : Neural Networks Test And Score

A screenshot of a computer

AI-generated content may be incorrect.

Figure 12 : Logistic Parameters

A screenshot of a computer

AI-generated content may be incorrect.

Figure 13 : Logistic Test and Score

A screenshot of a computer

AI-generated content may be incorrect.

Figure 14 : Random Forest Perameter

A screenshot of a graph

AI-generated content may be incorrect.

Figure 15 : Random Forest Test and Score

A screenshot of a computer

AI-generated content may be incorrect.

Figure 16 : SVM Parameters

A screenshot of a graph

AI-generated content may be incorrect.

Figure 17 : SVM Test and Score