THE UNIVERSITY OF SYDNEY

QBUS3600 - Group Report

Shelby Narborough	**510467010**
Max Harper	**510471039**
Thomas Hay	**510510730**
James Dwyer	**510518183**
Eva Wright	**520405938**

L'ORÉAL

Dermatological Beauty

The University of Sydney Business School

Contents

List of Figures

1 Business Understanding - Phase I

The aim of this report is to develop a predictive model for estimating the transaction value of customers, referred to as "patsumers," who shop on L'Oreal Dermatological Beauty's (LDB) La Roche-Posay eCommerce platform. To achieve this, the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework will be harnessed to guide the development of a system for predicting the transaction value of a patsumer over the next 6 months. CRISP-DM offers a more iterative and adaptable process compared to traditional linear approaches, which often suffer from rigidity and lack of flexibility in handling new insights that emerge during analysis (Huber et al., 2019). By allowing for a dynamic flow between phases, this methodology enhances the depth of understanding and enables continuous refinement of the model as new patterns and insights are uncovered (Appendix A.1, Chapman et al., 2000). The initial stage requires an understanding of the business context, as described in the problem description below.

1.1 Background

La Roche Posay is a subsidiary brand of LDB, offering a range of skincare product solutions to various skin conditions in what is referred to as the medically based skincare industry ("L'OrÃľal Group", n.d.). Historically, La Roche Posay, alongside all other LDB products, primarily retail through pharmacies, including large chain pharmacies such as Chemist Warehouse and Priceline ("Assessment information and resources: QBUS3600 Business Analytics in Practice", n.d.)

LDB is expanding their online presence by launching brand-specific eCommerce platforms, including the La Roche-Posay website that allows patsumers to purchase La Roche-Posay products online. Operating in the highly competitive Australian cosmetic industry, characterised by low market concentration and intense rivalry among skincare retailers, online shifts have proven effective in rapidly capturing market share by offering greater product accessibility, competitive pricing, and enhanced customer incentives (Richardson, 2024). Moreover, the introduction of eCommerce platforms has afforded LDB unprecedented access to information about their patsumers purchasing behaviours, including the transaction data utilised in this report.

1.2 Business Objectives & Success Criteria

LDB's business objectives are centered around enhancing customer engagement and driving revenue growth through data-driven strategies. The primary goal is to leverage advanced data analytics to improve their personalisation capabilities, allowing them to more accurately predict the future transaction value of their patsumers. This objective is complemented by the aim to gain deeper insights into customer behavior and preferences, enabling LDB to anticipate and meet patsumer needs more effectively. ("Assessment information and resources: QBUS3600 Business Analytics in Practice", n.d.)

The primary aim of this report is to develop an accurate and interpretable model for predicting the future transaction value of a patsumer. The resulting model will be utilised to inform a data-driven project that meets the following success criteria:

No.	Objective	Success Criteria
1	Develop an accurate and interpretable model for predicting the future transaction value of patsumers	Achieve significant improvements in prediction accuracy in the modelling phase, quantified by evaluation metrics (See Section 5).
2	Increase in overall volume (units sold)	Achieve a measurable increase in the total number of products sold following the project intervention.
3	Increase in total gross profit	Achieve an overall increase in total gross profit, contingent on the price of the units sold and cost of the project.
4	Enhance personalisation strategy	Improve effectiveness of targeted marketing campaigns, quantified by hightened conversion rates and increases in purchase quantities following project intervention.

Table 1: Business Objectives and Success Criteria

2 Data Understanding - Phase II

2.1 Data Description Report

The dataset provided by LDB describes the transactional history of each patsumer who shopped on the La Roche Posay website up until November of 2024, with a total of 6400 observations. This is broken into three main groupings: transactional data, purchase data, and location data. Transaction data refers to data that describes both the total amount spent and the number of transactions over a given period. The purchase data classifies the specific product types and the number of these purchases. The location data is the postcode of the patsumer. More specific information can be found in the Data Dictionary (Appendix A.4).

2.2 Data Quality Report

Quality Concern	Description
Exclusion of non-website purchasers	The dataset is sourced only through transactions that occurred on the La Roche Posay Website, which accounts for less than 20% of all patsumers ("Assessment information and resources: QBUS3600 Business Analytics in Practice", n.d.). Transactions through the website comprised only a fraction of the overall volume sold by La Roche Posay. For this reason, this patsumer segment may behave differently to patsumers who shop in-store.
Exclusion of non-recent Purchasers	Only patsumers who purchased between November of 2023 and May of 2024 are included. This exclusion of patsumers who have a transactional history yet did not make a purchase in the dependent variable period limits the scope of the type of model that can be created to predict the amount spent, not whether or not a purchase will be made.
Patsumers with 0 Purchases	Despite all customers having at least one transaction in the months prior to November 2023, 638 customers had not made any purchases across the available categories (roughly 10% of observations). When querying LDB about this, it is discovered that cookies cannot be collected from these customers. This creates issues for future data analysis and model building. Furthermore, inspecting La Roche Posay and other LDB brands' websites shows that this is not a result of products being left out of the data collection.
Missing Postcodes	There were 5,326 missing values in the "Post Code" variable, likely due to customers not providing this information during data collection, resulting in a limited sample of patsumers eligible for location analysis. A Mann-Whitney U test revealed that the difference in the mean spending between groups who provided a postcode and those who did not was not statistically significant (Appendix A.6, Table 16, however, the group who did not provide postcodes tend to have higher and more frequent outliers (see Appendix A.8).
Invalid Postcodes	It was disclosed that the postcodes had not been previously verified for legitimacy. To identify any invalid postcodes in the customer dataset, they were compared with the postcodes provided in the Mosaic dataset. Two discrepancies were identified: [96766, 1730].
Duplicate Columns	There were 2 columns with similar labels, storing different values: "Category_Face Care", & "Category_Face Care " (noting the additional whitespace). After consulting with L'oreal, this was found to be a input error, where the same category had been counted as two different variables.
Cumulative Time Periods	The spending and transaction count variables measure cumulative behaviors rather than the value of occurrences in specific time frames, which is problematic as it may lead to high multicollinearity because the cumulative nature of these variables creates overlapping information across time periods. This redundancy can obscure the distinct contribution of each period, inflate the correlations between predictors, and ultimately reduce the reliability of regression estimates, making it difficult to identify the true impact of each time frame on the outcomes being analysed.

2.3 Data Exploration Report

Prior to conducting exploratory data analysis, the data cleaning steps detailed in Section 3.2 were completed. This report aims to gain insight into the major properties, characteristics, patterns and statistics of the dataset in order to inform variable selection, possible transformations, and modelling.

2.3.1 Distribution of Variables

All transactional and purchase data variables exhibited a strong positive skew with heavy or relatively normal tails. This distribution pattern indicates that the majority of transactions or purchases are of relatively low value, while a smaller number of high-value transactions create a pronounced long tail to the right of the distribution. Figure 1 illustrates this positive skew, depicting the distribution of amounts spent in both recent and all-time periods. This skewed distribution has several implications:

- **Concentration of spending:** While most patsumers have moderate spending habits, a small segment of high-spending individuals significantly influences the overall distribution.
- **Discrepancy between mean and median:** Due to the influence of high-value outliers, the mean spending becomes notably higher than the median. This suggests that the average (mean) spending may not accurately represent the typical patsumer's behavior.
- **Potential for segmentation:** The clear distinction between the majority of moderate spenders and the minority of high spenders presents an opportunity for targeted marketing strategies or personalised approaches.
- **Model Considerations:** When analysing this data, it's crucial to use statistical methods that account for the skewed nature of the distribution, such as log transformations or non-parametric tests, to avoid misinterpretations based on assumptions of normality.

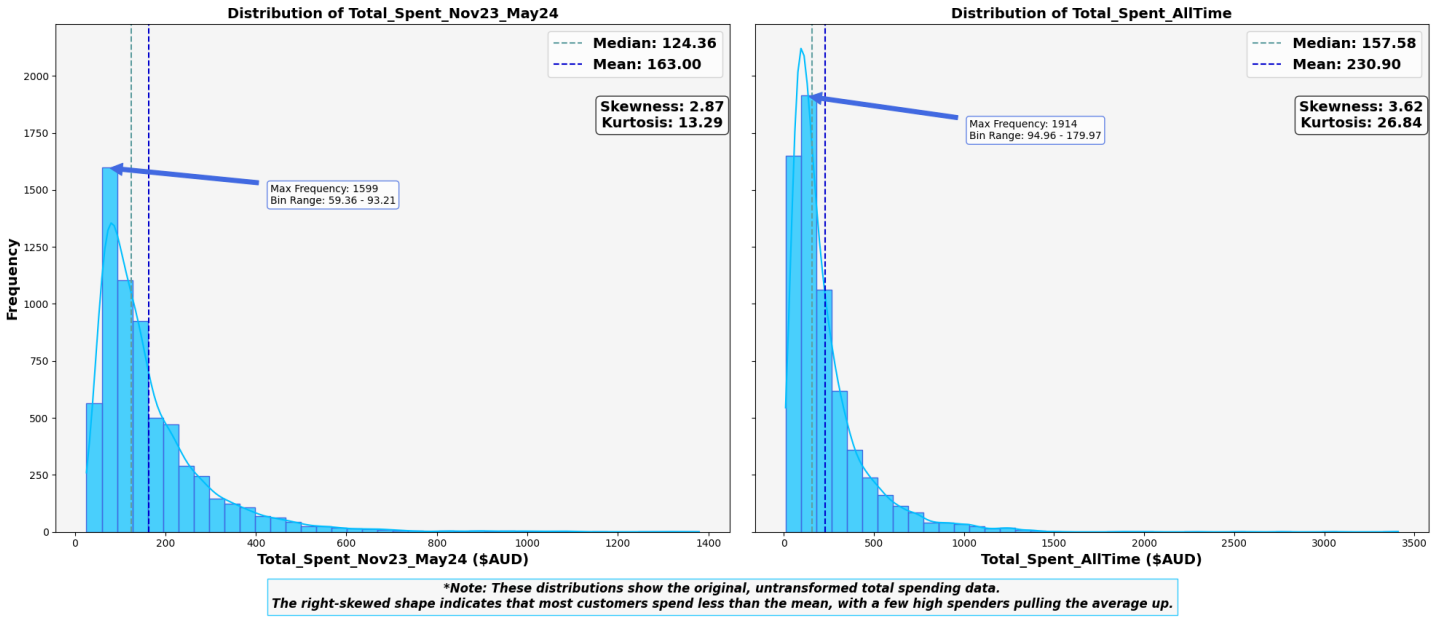


Figure 1: Histogram of Total Spent in Recent and All Time Periods

2.3.2 Variance Inflation Factor Analysis

Variance Inflation Factor (VIF) analysis was conducted to assess the extent of multicollinearity amongst variables (See Appendix A.5). The key findings are as follows:

- **Product Type - Infinite VIF:** Variables including brand descriptions, class descriptions, categories, sub-categories, and skin concerns showed a VIF of infinity, indicating perfect multicollinearity. This is due to the overlap in how these variables capture similar aspects of patsumer behavior, as individual products often belong to multiple categories and sub-categories. Notably, Class Description is recognised as a type likely to have high multicollinearity within its own subset. For example, Class Description Face Care would likely have overlap with Class Description Anti–Ageing, as it is possible for Anti–Ageing products to also fall under Face Care.

- **Time Periods - High VIF:** Cumulative spending and transaction count variables exhibited high VIF values (>10). The high VIFs suggest a strong correlation among these variables, which is expected given that they measure cumulative behaviors over overlapping time frames leading up to November 2023.
- **EANs - Modearte VIF:** Some specific EAN codes appear to have moderate to high VIF values, suggesting that they are strongly correlated with other variables in the dataset. This could be because these products are particularly popular or frequently purchased together with other items, leading to similar purchasing patterns across different patsumers. Postcodes also exhibit moderate VIF values; however, since postcodes are not on an ordinal scale but rather represent location data, the multicollinearity associated with them may be less concerning and can potentially be ignored.

2.3.3 Correlation between Variables & Total Spent in Recent Period

Figure 2 depicts the correlation coefficient between variables in the dataset with the total amount spend in the most recent 6 month period. The strongest relationships for transaction count and total amount spent are the 12 month and All Time cumulative periods, indicating that historical spending and purchasing patterns may be stronger predictors of recent spending patterns, compared to more recent cumulative periods exhibiting lower correlations.

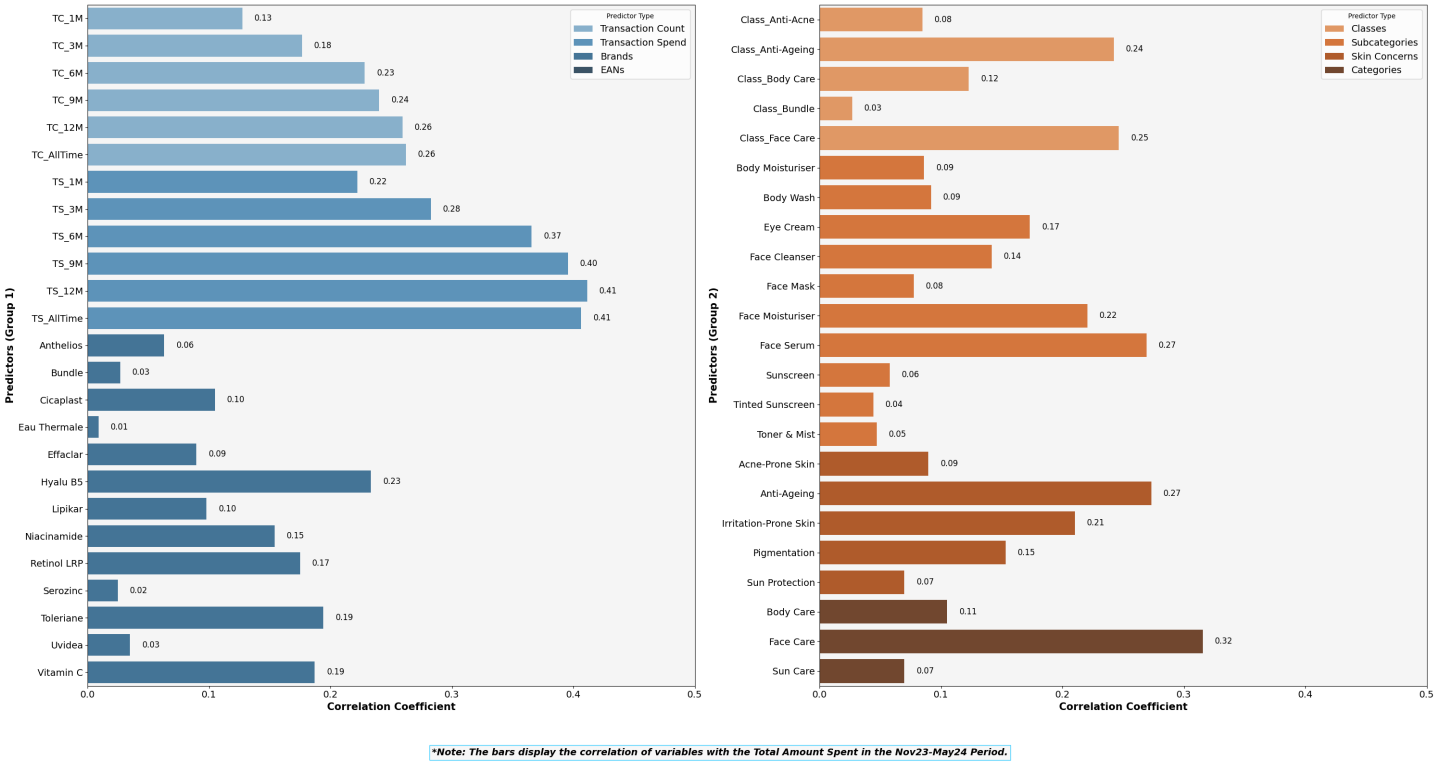


Figure 2: Correlation Coefficient between Predictors and Total Spent (Nov23-May24)

Figure 3 highlights the proportion of customers who purchased products across various brands, skin concerns, categories, and sub-categories, revealing intriguing insights about the relationship between product popularity and its impact on customer spending. Notably, several products demonstrate a disproportionate influence on spending despite lower purchase volumes. For instance, Hylau B5, while only the 4th most popular brand, exhibits the highest correlation (0.23) with customer spending. Similarly, Vitamin C products show the second-highest correlation (0.19) with spending, yet were purchased by only 16.5% of customers, likely attributed to the higher average price of products belonging to these brands (Table 2). This pattern extends to Anti-Ageing products and Face Serums, which display high correlations with spending but are not the most frequently purchased in their respective categories. These observations suggest that certain products, despite lower market penetration, exert a strong influence on overall customer spending.

Brand:	Anthelios	Cicaplast	Hylau B5	Lipikar	Toleriane	Vitamin C	Effaclar
Average Price:	40.25	29.12	74.46	38.17	42.22	68.28	41.08

Table 2: Average Price of Product by Brand (Septmember 2024)(“La Roche-Posay Skincare Official Site | La Roche-Posay Australia”, n.d.

)

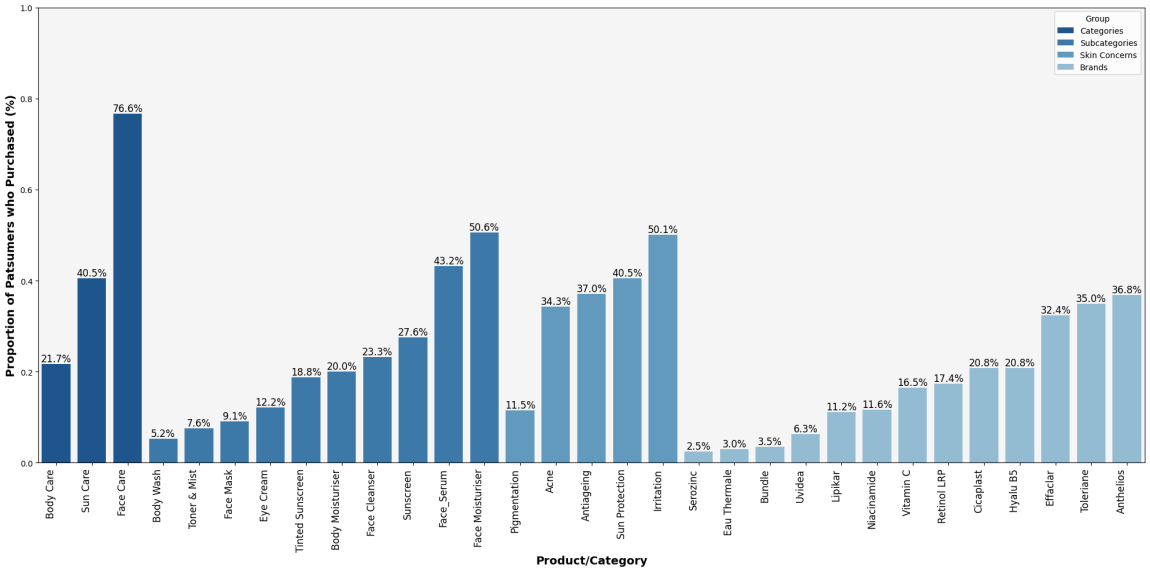


Figure 3: Proportion of Patsumers who Purchased by Category/Product

2.3.4 Interaction Effects

Among all product groupings, face care emerges as a significant driver of customer spending, with the highest categorical grouping correlation (0.32) and the highest market penetration of 76.6% of patsumers purchasing at least one face care product. To delve deeper into the impact of key product types on total spending, particularly those with the strongest correlations, Figure 4 visualises the interaction between the number of face care purchases and specific product category purchases. The visualisation reveals a strong linear relationship between face care purchases and amount spent, indicating that each additional face care product acquired contributes to an increase in overall spending. Binary indicators were employed to identify patsumer purchases within four critical categories: Face Serum, Face Moisturiser, Irritation-Prone Skin products, and Anti-Ageing products. Consistent with the correlations presented in Figure 2, the data demonstrates that spending tends to increase more substantially when a patsumer purchases a Face Serum or Anti-Ageing product. In contrast, the effect on spending is less pronounced for Face Moisturiser or Irritation-Prone Skin product purchases. This disparity suggests a hierarchy of product impact on customer spending within the face care category.

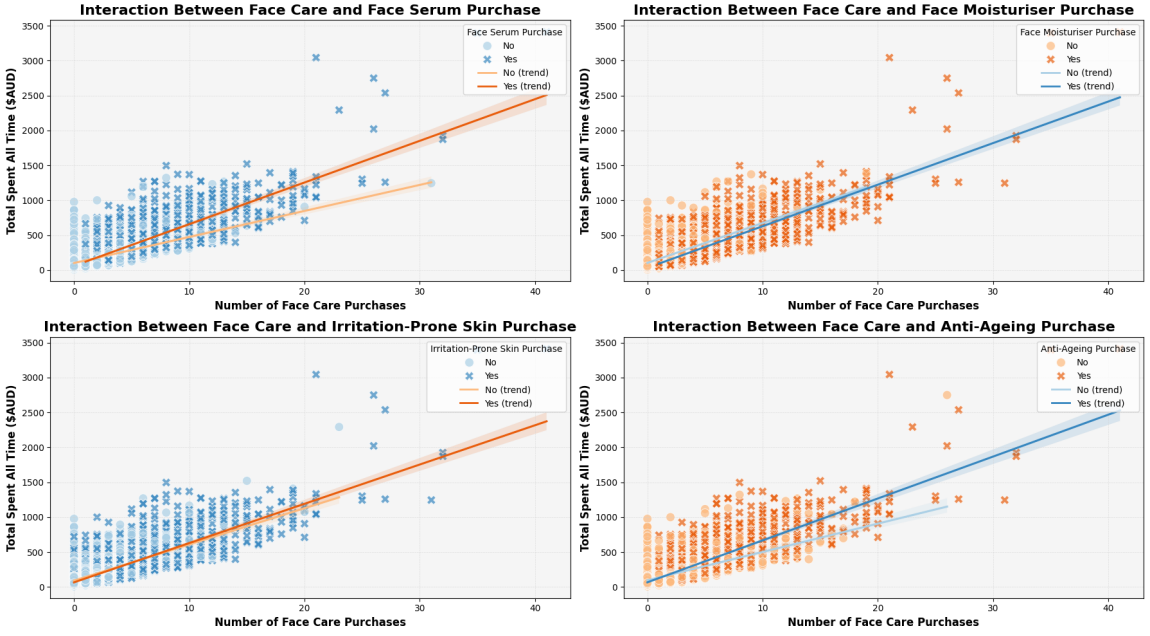


Figure 4: Interaction between Face Care Purchases and Total Spending for Key Product Types

To further explore these interactions, a preliminary log-log MLR analysis was conducted, chosen due to the strong positive skew present in all variables. Note all coefficients were deemed statistically significant (See Appendix A.6, Table 17). The

key insights from the analysis are as follows:

- **Face Serum:** Patsumers who purchase Face Serum products tend to spend 2.13% less overall. However, each additional Face Care purchase increases their spending by 5.16%, indicating that they may concentrate their spending on this product type. The interaction term suggests that as the quantity of Face Care purchases increases, the spending on Face Serum products also increases proportionally.
- **Face Moisturiser:** Patsumers who purchase Face Moisturisers generally spend 4.77% less overall. However, for each additional % increase in Face Care purchases, spending increases by 5.56%. This suggests that while Face Moisturiser buyers spend less overall, their spending increases with additional Face Care purchases.
- **Irritation-Prone Skin:** Patsumers purchasing products for irritation-prone skin spend 1.82% more overall. Each additional % increase in Face Care purchase increases spending by 0.17%, indicating a slight interaction effect. This suggests that customers with irritation-prone skin may spread their spending more evenly across various products.
- **Anti-Ageing:** Patsumers who buy Anti-Ageing products spend 1.51% less overall. However, each additional % in Face Care purchases increases their spending by 4.47%. This interaction suggests that customers focused on anti-ageing products tend to increase their spending as they purchase more Face Care items, indicating a focused spending pattern on these products.

2.3.5 Covariance between Brand Pairings

Figure 5 illustrates the covariance between different brand pairings, quantifying how the probability of purchasing one brand changes in relation to the purchase of another. This analysis reveals several notable patterns in customer buying behavior:

- **Strong positive correlations:** Toleriane and Cicaplast exhibit the highest covariance (0.31), indicating a strong tendency for customers to purchase these brands together. This is closely followed by the pairings of Toleriane with Anthelios (0.22), and Hyalu B5 with both Retinol (0.18) and Vitamin C (0.18). The top brand pairings suggests a strong customer preference for certain brand combinations, implying a complementary relationship in the minds of LDB's patsumers
- **Top performers:** Interestingly, these highly correlated brands also tend to rank among the most frequently purchased products, suggesting a synergistic relationship in their popularity.
- **Weak Correlations:** Many brand pairs show covariance values close to zero, implying that purchases of these brands are largely independent of each other.

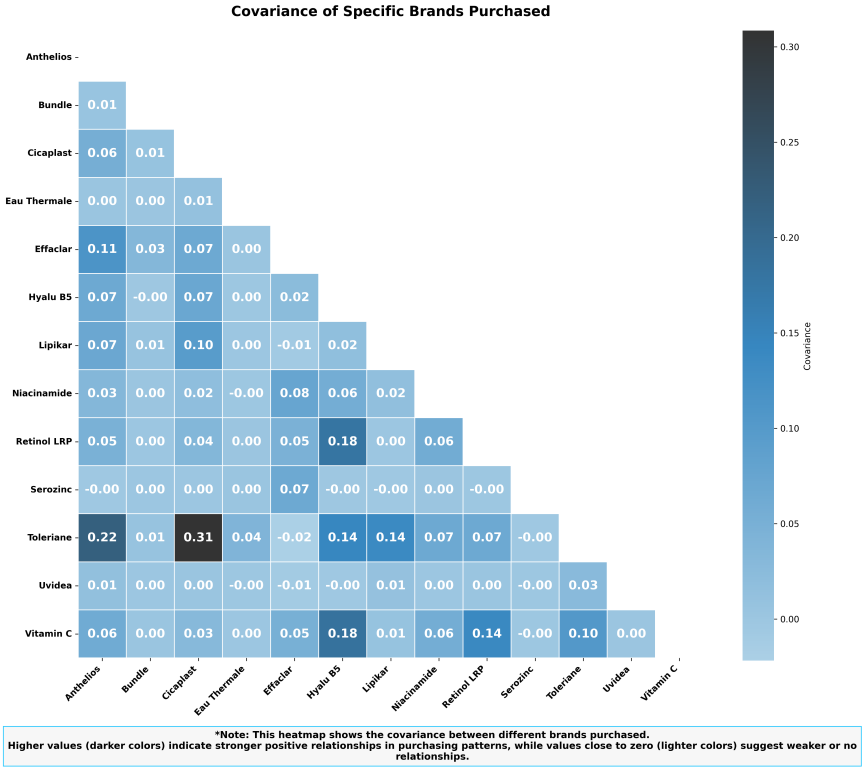


Figure 5: Covariance Heatmap

3 Data Preparation - Phase III

3.1 Selected Variables for Modelling

The table below highlights the variables excluded from the modelling process and outlines the control principles applied. Importantly, only a few variables have been removed despite potential multicollinearity. Instead of simply eliminating highly correlated variables, control principles have been implemented to prevent potentially collinear variables from being included in the same model. This approach allows for a more comprehensive exploration of models and variable combinations. The aim is to preserve the richness of the dataset while ensuring statistical validity, thereby uncovering meaningful relationships that might otherwise go undetected.

Excluded Variables	Rationale for Exclusion
Class Descriptions	Highly ambiguous grouping with potential significant overlap between products. Inclusion would lead to difficult results to interpret and trace back to specific causes, ultimately compromising the model's explanatory power.
Postcodes	Not intended for numerical use in analysis. To preserve valuable geographical insights, postcode data will be merged with a customer segmentation dataset later, ensuring that potential information is not lost while avoiding direct numerical interpretation.

Table 3: Removed Variables

Variable Controls	Implementation Rationale and Specific Applications
Temporal Aggregation Control	Restrict inclusion to prevent multicollinearity: 1) Only one cumulative time period should be added to the model. 2) Cumulative and differenced periods must not coexist in the model. This applies to spending and transactions within various time periods.
Hierarchical Integrity	Adhere to the hierarchy principle: If an interaction effect is included in the model, the parent effects must also be included. This ensures proper model specification and facilitates accurate main and interaction effects interpretation.
Transformation Exclusivity	Prevent redundancy and maintain clear interpretation: 1) If the log of a variable is included, its original form may not be included. 2) If the original form is included, its log transformation may not be included. This principle applies to all transformed variables in the model.
Granularity Consistency	Avoid nested hierarchical data to prevent multicollinearity: 1) If a brand description is included, associated EANs belonging to that brand cannot be included. 2) If a category is included, subcategories belonging to that category cannot coexist in the model. This ensures a consistent level of analysis and prevents redundancy in the model.

Table 4: Variable Selection Criteria and Control Principles for Predictive Modelling

3.2 Data Cleaning Report

Data Cleaning	Description
Missing Post-codes	To ensure that the missing data did not disrupt the analysis, the missing values were replaced with 0 as a neutral placeholder. Additionally, a binary indicator, "Has_Postcode", was created to distinguish between customers based on postcode provision to allow for segmentation analysis of the subset who have postcodes provided.
Data Types	The "Post Code" variable was initially stored as a float, resulting in postcodes being represented with a decimal point (e.g., 2010.0). To ensure the data accurately reflected postcodes, the variable was converted to an integer type (e.g., 2010).
Invalid Post Codes	**96766**: This postcode belongs to Hawaii, which is irrelevant for Australian customers. The La Roche-Posay website states, "All orders must be sent to a registered Australian Postal Address." (La Roche Posay, n.d.). Therefore, this is likely an input error made by an Australian customer. As a corrective measure, this entry was replaced with a placeholder value of 0. **1730:** This postcode corresponds to a PO box in Seven Hills, NSW. PO box postcodes are not typically included in the Mosaic dataset. While it is possible that the customer's PO box is located away from their residential area (e.g., near their workplace), it was decided to replace this entry with the corresponding residential postcode for Seven Hills, 2147.
Duplicate Columns	To resolve the duplicate columns for Face Care purchases, the columns were summed together as a replacement for both.

3.2.1 Differenced Time Periods

To better understand the influence of purchasing behaviours in specific time periods, different spending and transactions were calculated to isolate them within defined month-range intervals. Figure 6 visualises the average spending and average number of transactions within these differenced intervals. Where appropriate, the time frames were aggregated into 6-month periods to align with the most recent November 23 to May 24 period.

Among these 6-month periods, the most recent time frame showed the highest average spend of $163, suggesting that customers are spending more in the current period. Interestingly, the average spend from November 22 to May 23 compared to the subsequent May 23 to August 23 period decreased from $70.51 to $52.06, potentially indicating a seasonal pattern where customers spend more during the summer months (possibly due to events such as Christmas). However, the dataset lacks sufficient long-term data to confirm this pattern, highlighting the potential value of longer-term data to uncover spending trends across different seasonal periods.

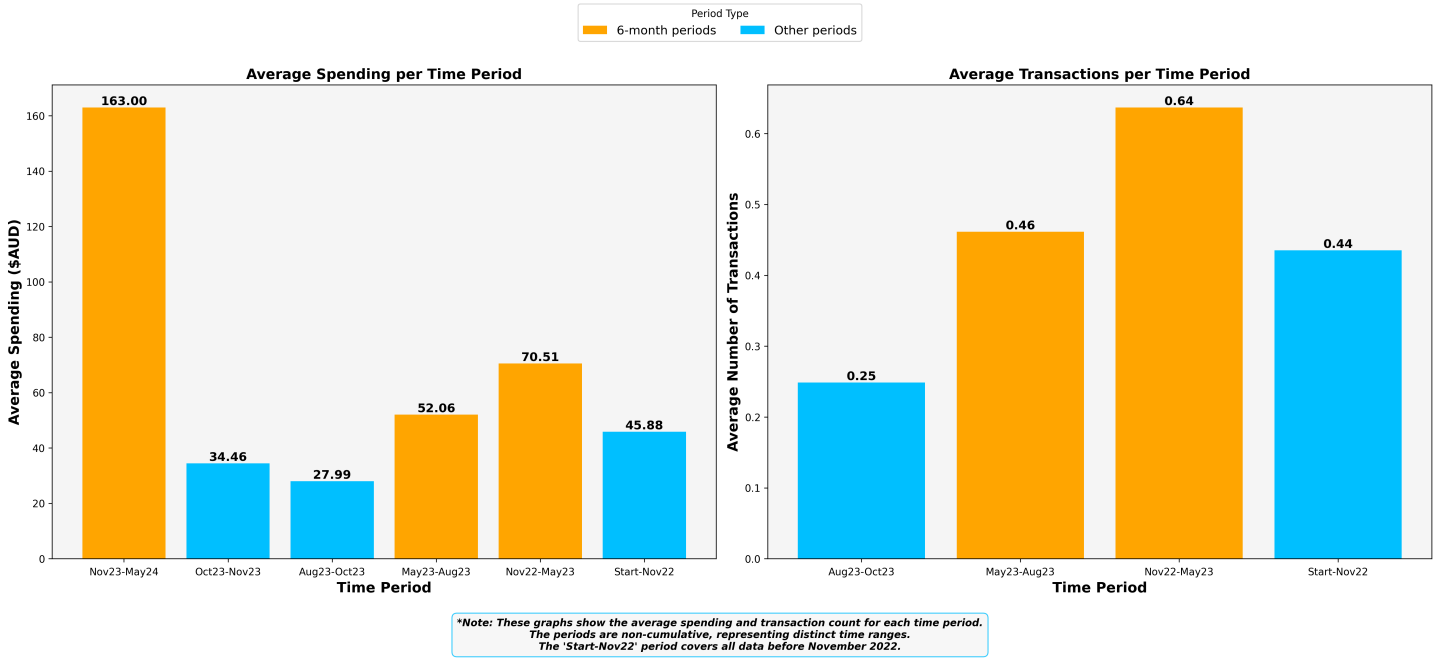


Figure 6: Average Spending and Transaction Count by Differenced Time period

A preliminary log-linear multiple linear regression (MLR) model demonstrated that differenced time periods are better predictors of future spending than cumulative periods. When analysing cumulative time periods, only the 6-month and All Time periods showed statistical significance (Appendix A.6, Table 18). In contrast, all differenced periods proved to be statistically significant, indicating their stronger predictive power. Interestingly, the periods aggregated into 6-month intervals (May 2023 - August 2023, November 2022 - May 2023) exhibited the highest correlations with recent spending, indicating that aggregating data into comparable time frames to the dependent variable may potentially enhance predictive power.

Unlike cumulative periods, The β coefficients for the differenced periods decreased as the periods extended further back in time. The most recent period (October 2023 to November 2023) demonstrated the strongest effect out of all variables tested in both differenced and cumulative models, with $\beta = 1.7 * 10^{-3}$ indicating that a 1 unit increase in spending during this period corresponds to a 1.7% increase in spending over the most recent 6 months. These findings highlight the importance of recent changes in spending behaviour, suggesting that patsumers who have increased their spending in recent periods are more likely to maintain higher transaction values in the near future. This pattern underscores the value of using differenced, aggregated time periods for more accurate predictions of future spending behaviour.

3.2.2 Log Transformations

The variables listed in Table 5 show that the log transformation significantly reduced skewness, enhancing the robustness of the regression analysis. A notable improvement was defined by the criterion that the log transformation reduced absolute skewness to below 0.5, indicating a more balanced and symmetric distribution. A $log(x+0.1)$ transformation was applied to these variables. This transformation is visualised in Figure 7, which sees the skewness of the dependent variable drop from 2.87 in Figure 1 to 0.15, with the mean only \$0.06 higher than the median indicating a more symmetric distribution. The Yeo-Johnson transformation was also considered, effectively reducing skewness to 0.05 while aligning the mean and median (Appendix A.7); however, it was ultimately less interpretable than the log transformation due to its complex non-linear adjustments, making it harder to relate results back to the original scale.

Variable	Skew	Description	Log Skew	Description
Total_Spent_AllTime	3.62	Strong Positive Skew	0.24	Weak Positive Skew
Total_Purchases_AllTime	2.86	Strong Positive Skew	0.08	Approximately Symmetric
Category_Face Care	2.73	Strong Positive Skew	0.23	Weak Positive Skew
Total_Spent_Nov23_May24	2.87	Strong Positive Skew	0.15	Weak Positive Skew

Table 5: Variables Improved by Log Transformation

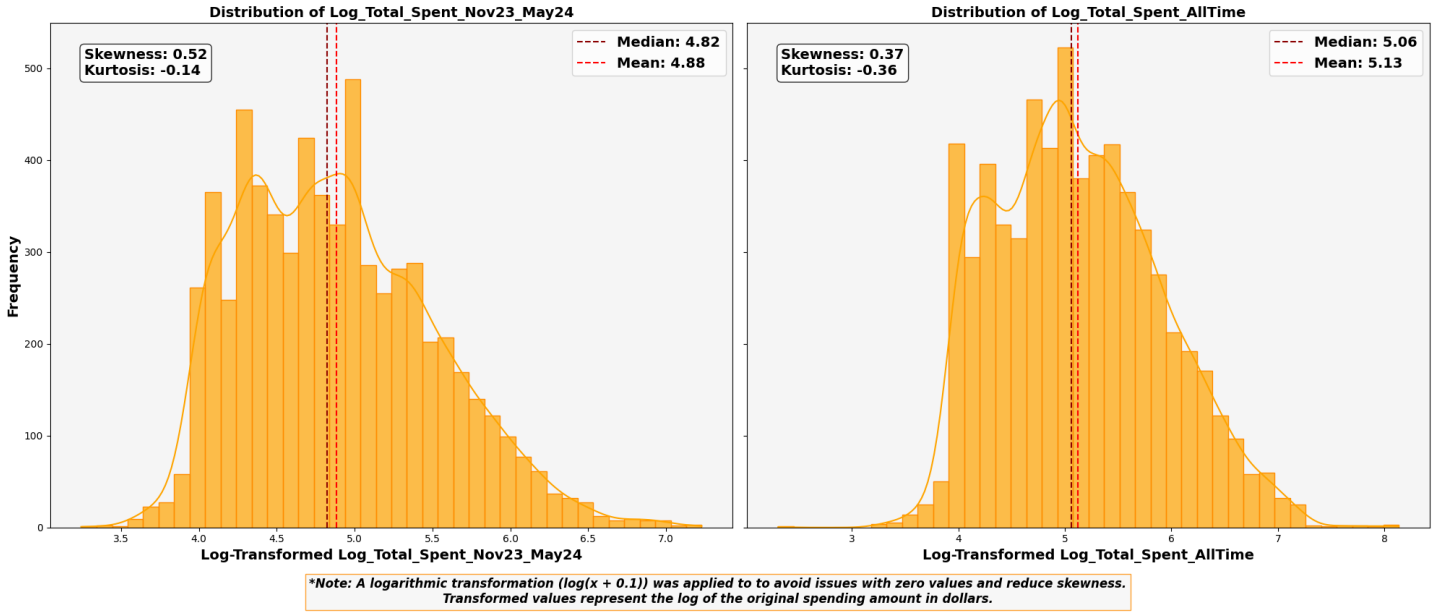


Figure 7: Log Transformed Distribution of Amount Spent

3.3 Derived Attributes & Generated Records

Variable	Description
"Total Purchases AllTime"	Describes the number of purchases made by each Patsumer, calculated by summing transaction counts across Face Care, Sun Care and Body Care categories (assuming all purchases fall into these main categories).
"Has_Purchased"	Binary indicator to indicate whether the patsumer has made a purchase
"Has_Postcode"	Binary indicator to indicate postcode provision
"Bool [Variable]"	Binary indicator to indicate whether a patsumer purchased a product belonging to a specific brand
"Purchased [Variable]"	Binary indication of whether a patsumer has purchased a product belonging to a specific category, sub-category, or skin concern.
"Log [Variable]"	Variables that have a $log(x + 0.1)$ transformation applied
"Total Spent [Month, Year] - [Month, Year]"	The total amount spent by a patsumer in a specific time frame
"Total Trans [Month, Year] - [Month, Year]"	The total number of transactions made by a patsumer in a specific time frame
"[Group] Spender"	The spending group which the patsumer belongs to according to the interquartile range of Total Spent All Time (Low Spender being Minimum-Q1, Moderate Spender being Q1-Q2, High Spender being Q2-Q3, Very High Spender being Q3-Maximum, and Top 10% Spender if they are in the top 10% of spenders. This was split into K-1 categorical dummy variables, with the **Baseline Level: Low Spender**.
"[Categorical Var] times [Numerical Var]"	The interaction effect between 2 variables, indicating that the effect of the numerical variable depends on the level of the categorical variable.

3.4 Data Merging

3.4.1 Mosaic Data

The Mosaic Postcode Dataset is a nationwide postcode-level breakdown of customer segmentation developed by Experian. Experian's Mosaic dataset classifies each household in Australia into 14 groups, which can be subdivided into 51 types. Each group/type has been developed based on attributes such as demographic characteristics, lifestyle, behaviour and attitudes, detailed in Appendix A.9 ("Experian Mosaic", n.d.). Prior to joining the data, the dominant segment present in each postcode was determined by finding the segment index where the maximum frequency was present. For the subset of customers who have a postcode provided, their dominant segment was assigned by matching the postcode indexes among the two data sets. It is important to note that the accuracy of segment allocation at the postcode level is limited, as segments are assigned based on the dominant segment within that postcode, hence a patsumer may be misplaced into a segment that does not accurately capture their characteristics.

Figure 8 compares the dataset demographics to those of Australia, revealing significant deviations between group representations. Groups C, F, and G, who are characterised by young age, average to high income, and urban residence, are over-represented. Conversely, groups H, L, M, and N, who are typically older age, lower income, and have manual labor occupations, are under-represented. This distribution underscores the segment-dependent nature of LDB patsumers, suggesting that marketing strategies should target young, high-earning demographics. However, this analysis alone does not account for segment-specific spending patterns, a crucial factor for revenue optimisation.

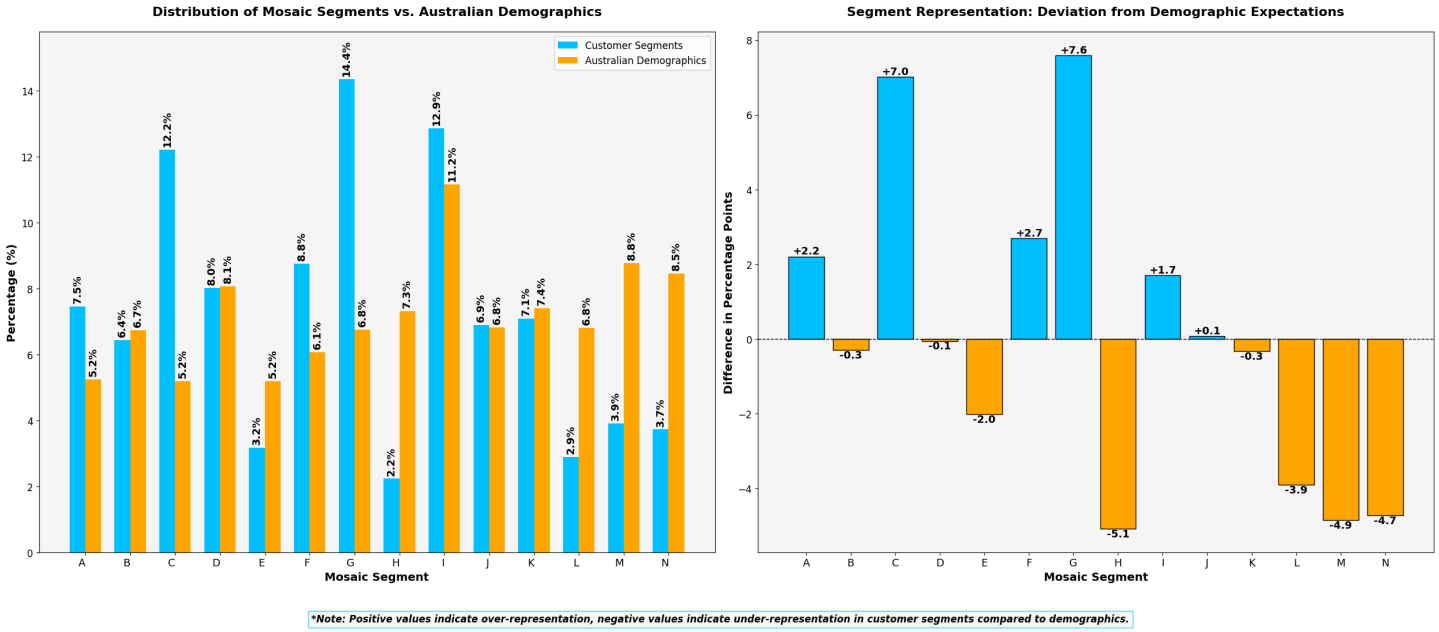


Figure 8: Patsumer Distribution in Dominant Segment Postcodes and Deviations from Expected Values

Figure 9 illustrates divergent spending patterns across segments, contrasting with the representation trends observed earlier. Notably, over-represented segments like F and G exhibit lower spending, while under-represented segments such as N demonstrate higher expenditure. The highest-spending segments are characterised by older age and below-average income, suggesting age as a potential key determinant of customer spending behavior.

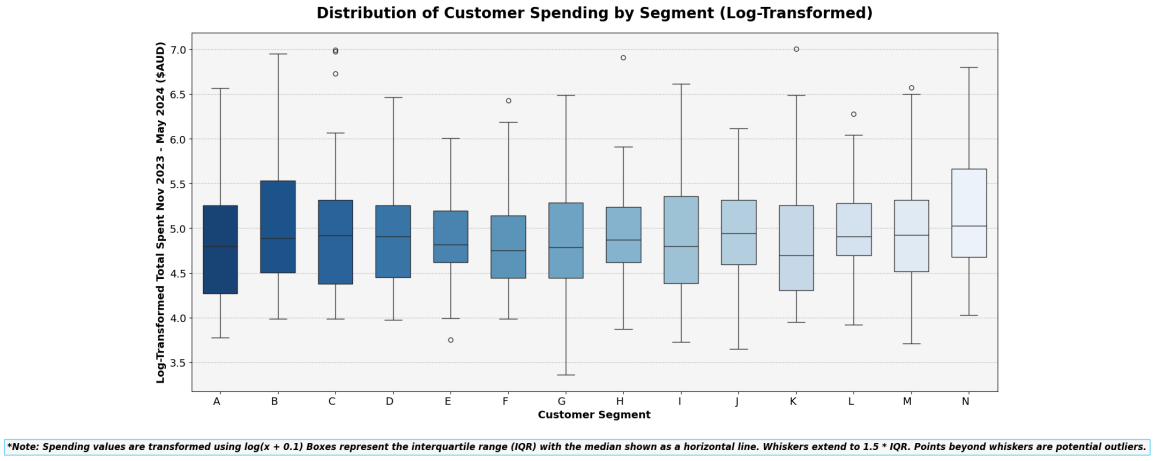


Figure 9: Transformed Spending Distribution Across Customer Segments

The relationship between segments and brand descriptions is explored in Figure 10, employing boolean categorical data to mitigate individual purchase volume bias. The heatmap depicts the probability of specific brand description purchases within each segment, revealing both broad trends and nuanced patterns. For instance, Anthelios consistently outperforms Serozinc across all segments, with segment L showing a 52% likelihood of Anthelios purchases. The accompanying stacked bar chart offers an alternative perspective, illustrating the expected quantity of brand descriptions purchased per capita in each segment. Segments B and N emerge as the highest per-capita purchasers, despite their contrasting demographic profiles: B represents high-income urban dwellers, while N comprises below-average income rural residents. However, the limited data points for these segments (69 for B and 40 for N) necessitate further postcode data collection for more robust analysis.

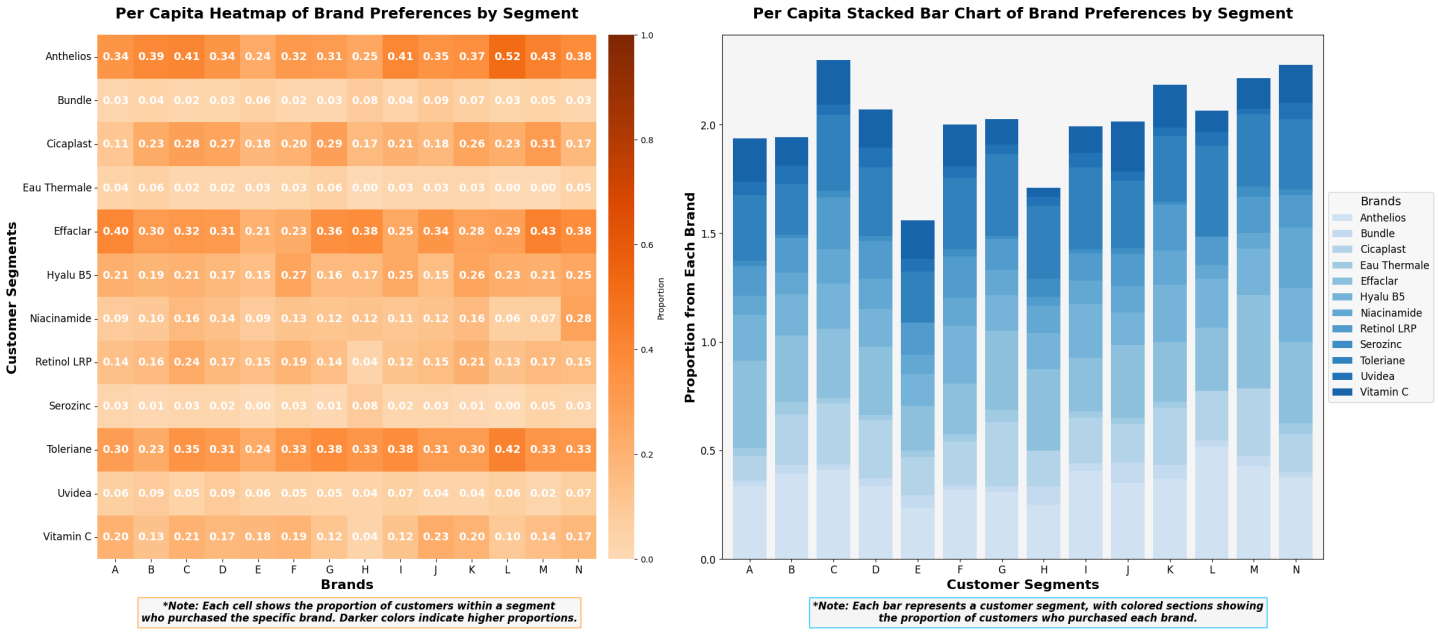


Figure 10: Segment-Based Brand Description Purchase Analysis

3.4.2 Postcode Data

To accurately map the geographic location of customers, a postcode locality database was utilised to retrieve comprehensive information, including latitude, longitude, locality, state, and SA4 names for each postcode (Australia, n.d.). Duplicate postcodes were removed, keeping only the first occurrence, resulting in a slightly generalised locality approximation. The SA4 field was standardised by removing notations or suffixes (For instance "Brisbane - South" was replaced by "Brisbane"). Following these cleaning steps, these columns were joined to the data using the postcode index. After checking for missing values, one missing SA4 name was updated manually for postcode 3336, which is located in Melbourne. The overall geographic distribution aligns closely with population expectations, with the majority of patsumers concentrated in urban areas (See Appendix A.2. Notably, the top 10% of spenders are predominantly located in cities, although a small number are scattered in rural areas of New South Wales and northern Queensland.

The table below describes the variables added to the data. The Geo-Plotting variables were added for geographic visualisation only, and were not included in the modelling. The merged variables were split into $K-1$ dummy variables, with the baseline level as described.

Merged Variable	Description
"Dominant Segment"	The Dominant Segment Group of the Postcode in which a patsumer is located. **Baseline Level: Segment A**
"electoraterating"	Federal Government Demographic Rating (e.g., Inner Metropolitan, Rural). **Baseline Level: Inner Metropolitan**
"sa4name"	The name of the city which the postcode resides in. **Baseline Level: Adelaide - Central and Hills**
Geo-Plotting Variable	**Description**
"sa4name general"	The general name of city which the postcode resides in
"Postcode Latitute"	The latitude of the postcode
"Postcode Longitude"	The longitude of the postcode
"locality"	The name of the postcode in which a patsumer is located

4 Modelling - Phase IV

4.1 Modelling Techniques and Assumptions

The modeling phase employed various analytical techniques to predict customer spending, as described in Table 6. These methods were selected based on their ability to address specific business objectives and provide insights into consumer

behavior patterns.

Technique	Description	Assumptions
Multiple Linear Regression	A statistical method that uses multiple explanatory variables to predict the outcome of a response variable. It attempts to model the linear relationship between the explanatory variables and the response variable.	Linearity, independence of errors, homoscedasticity, absence of multicollinearity, normality of residuals.
Lasso	A regression analysis method that performs both variable selection and regularisation to enhance the prediction accuracy and interpretability of the statistical model. It adds a penalty term to the loss function, which can force certain coefficients to zero. (Emmert-Streib and Dehmer, 2019)	Assumes some coefficient values are exactly zero, leading to sparse models. Assumes linearity between predictors and response.
Gradient Booster	A machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion and generalises them by allowing optimisation of an arbitrary differentiable loss function.	No strict distributional assumptions. Assumes the underlying pattern can be approximated by an ensemble of weak learners. Can handle non-linear relationships. Sensitive to outliers.
Random Forest	The random forest method is a supervised learning algorithm that creates many decision trees to form a forest. Each tree is built upon random subsets of the data and then combines them to create a single model based on the weighted averages of all of the trees.	Assumes independence between trees. No strict distributional assumptions. Assumes features have some predictive power. Can handle non-linear relationships and interactions.
Neural Networks	Neural Networks are computational models consisting of linked nodes called artificial neurons which receive inputs and output to other nodes, similar to how neurons in the brain process information. Deep Learning is the process of training these models, adjusting input and output calculations based on the accuracy of the final answer.	Assumes the problem can be represented by a series of hierarchical feature transformations. No strict distributional assumptions. Assumes sufficient data for training.

Table 6: Comparison of Different Modeling Techniques

4.2 Test Design

Data Division: Prior to modeling, the data was split into a training set (80%) and a test set (20%). The training set was utilised for model training and tuning, while the test set was reserved for final evaluation. This separation ensures that the model's performance is assessed on unseen data, providing an unbiased estimate of its ability to generalise, as the test set remains untouched during training and tuning.

Cross Validation: To mitigate the risk of overfitting within the training set, a 5-fold cross-validation approach during model selection was adopted. The training data was divided into five subsets, with each model being iteratively trained on four subsets and validated on the remaining one. This technique enhances the model's ability to generalize across different data splits.

Test Set: Final evaluation metrics will be based on the total amount spent during the Nov23-May24 period in the test set. This data was excluded from all training phases to prevent data leakage, ensuring that the model accurately predicts future transaction values solely based on historical purchasing behavior.

Evaluation Metric: Each model is evaluated using root mean squared error (RMSE), given by the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4.3 Parameter Settings

4.3.1 Dependent Variable Transformation

Section 3 of this report detailed the logarithmic transformation applied to the dependent variable for normalisation purposes. Consequently, all model evaluations in the modeling stage utilise logarithmic RMSE. While logarithmic RMSE is not directly interpretable, it facilitates consistent model comparisons. The final section of Phase IV will present the RMSE of each model in its original scale, enabling more intuitive interpretation of the results and supporting informed decision-making.

4.3.2 Feature Selection

In any predictive model, selecting the optimal subset of independent predictors is crucial for improving model performance and generalisability. To achieve a balance between computational feasibility and model efficacy, a forward selection algorithm was implemented. This algorithm strategically selects the feature that provides the most significant improvement in a given evaluation metric at each iteration, ensuring a locally optimal subset of features.

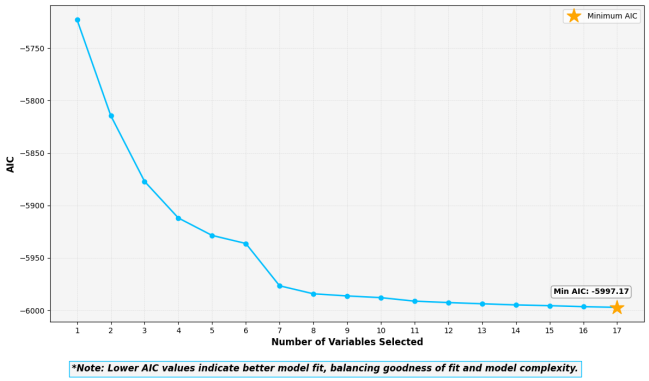
The algorithm incorporates two key enhancements:

- **Improvement Threshold:** The algorithm terminates when the marginal improvement of adding a new feature falls below a predefined threshold, preventing overfitting and unnecessary complexity.
- **Feature Exclusivity Constraints:** To maintain model integrity and ensure logical consistency, specific rules were enforced, as outlined in Table 4. These include adhering to the hierarchy principle and avoiding the inclusion of multiple cumulative time periods as predictors.

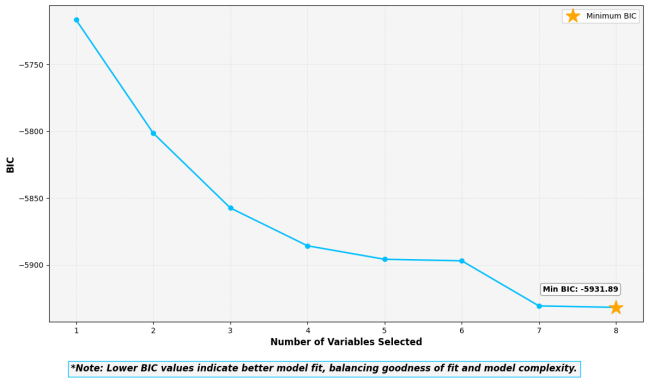
The RMSE generated from 5-fold cross-validation was selected as the primary evaluation metric when considering a feature for addition. Additionally, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were explored as alternative model selection criteria. AIC and BIC are model performance criteria that assess model accuracy while simultaneously penalising over-complexity (Chakrabarti and Ghosh, 2011). Although both criteria penalise complexity, BIC imposes a heavier penalty relative to AIC. Figure 11 shows that with an increasing number of variables, BIC's accuracy in identifying the true model improves. Consequently, given the large number of variables in this study, the results from BIC provide a stronger indication of the optimal model compared to AIC.

For both AIC and BIC, a forward selection loop was constructed, whereby the most beneficial variable is added based on the respective criterion. The selection process terminates once no further improvements can be made. This refined approach to feature selection ensures a rigorous and systematic method for identifying the most relevant predictors, ultimately enhancing the model's predictive power and interpretability.

These different criteria further emphasised the importance of an improvement threshold to limit excessive feature addition. However, the decision was made to continue using RMSE as the primary selection metric, as the computation of AIC and BIC cannot be performed for all models, such as ensemble learning methods



(a) AIC Progression as Number of Variables Increase



(b) BIC Progression as Number of Variables Increase

Figure 11: AIC and BIC progression

4.3.3 Geographical Data

Initial exploration involved a multi-stage model approach, where different models were fitted depending on whether a person provided their postcode. The feature selection algorithm, when applied to the training split of the data containing postcode information, yielded an interesting yet problematic feature set. This approach revealed two significant issues:

- **Overfitting:** The selected coefficients were overfitted to small subsets of the dataset, compromising the model's ability to generalise effectively to the broader population. For instance, a strong negative coefficient was assigned to the SA4 region "Riverina". However, this region accounts for only 0.58% of the consumers who provided their postcode, indicating a lack of representativeness.
- **Regional Bias and Discrimination:** This approach raised potential ethical concerns, particularly regarding regional bias. The model's structure could potentially lead to unfavourable treatment of individuals from certain areas, such as the Riverina, due to the negative coefficient associated with that region. This introduces significant issues of fairness and discrimination in model outcomes.

In light of these findings, the decision was made to amalgamate the dataset and exclude geographical information. This strategic adjustment aims to ensure both the generalisability of the model and its ethical integrity. By removing potentially biased geographical indicators, the model can provide more equitable predictions across all regions, aligning with principles of fairness and non-discrimination in predictive modelling.

Feature	Coefficient	% of Patsumers in Region
sa4name_Riverina	-0.58	0.58%
state_WA	0.19	8.28%

Table 7: Coefficients for Regional Features & % of patsumers in Region

4.4 Model Descriptions

4.4.1 Multiple Linear Regression

To establish a robust baseline model, a series of tests were conducted using multiple linear regression. These tests employed various improvement thresholds to optimise model performance and feature selection. The analysis focused on finding an optimal balance between model complexity and predictive accuracy, as illustrated in Figure 12. This revealed the following insights:

- **Threshold Sensitivity:** As the threshold decreases (moving right to left on the x-axis), the number of variables included in the model increases, as shown by the blue line.
- **Error Reduction:** The orange line represents the MSE, which generally decreases as more variables are added, indicating improved model performance.
- **Optimal Point:** A clear "elbow" point is observable where the rate of MSE reduction diminishes significantly, despite the addition of more variables. This occurs at an improvement threshold of 10^{-3}.

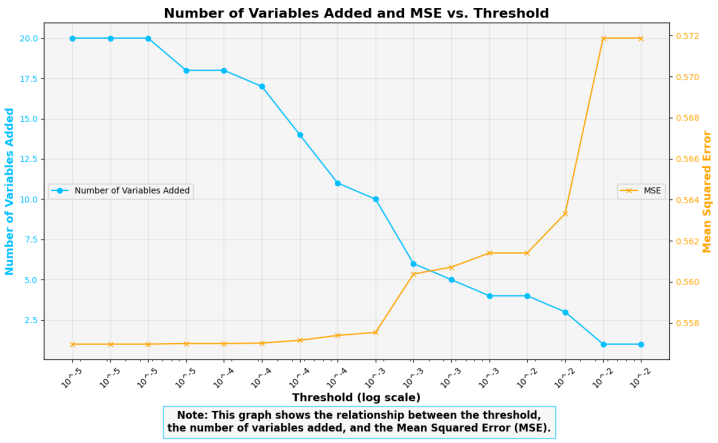


Figure 12: Threshold Trade-Off: Number of Variables Added and MSE vs. Threshold

Based on this analysis, an improvement threshold of 10^{-3} was selected for subsequent regression models. This threshold provides an optimal balance between model complexity and performance improvement. Applying this threshold to the multiple linear regression model resulted in the selection of four key variables: Log_Total_Spent_AllTime, Total_Spent_6M, Transaction_Count_6M, Purchased_Antiageing.

This parsimonious set of predictors achieved a log RMSE of 0.56 across 5-fold cross-validation, establishing a strong baseline for comparison with more complex models. The identified threshold and selected variables provide valuable insights into the most influential factors affecting the target variable. This baseline model not only serves as a performance benchmark but also offers interpretable results that can guide further analysis and decision-making processes in subsequent modelling efforts.

4.4.2 LASSO

Given the high dimensionality of the initial variable set, the LASSO (Least Absolute Shrinkage and Selection Operator) model was applied to the selected variable lists from the multiple linear regression models previously identified using AIC, BIC, and RMSE criteria. Additionally, a subset was created using the RMSE threshold method with a maximum improvement threshold of 0.01. The optimal alpha parameter for the LASSO model was determined by maximising the BIC, a common approach that balances model fit with complexity and interpretability (Emmert-Streib and Dehmer, 2019).

As shown in Figure 13, the non-zero alpha term indicates a new model configuration. This model yielded an RMSE of 0.56, suggesting a slight decrease in overall performance compared to the unmodified multiple linear regression model. However, it addresses overfitting issues by reducing the number of selected variables. The LASSO model identified 14 out of 20 features as significant predictors (see Appendix A.10). The most influential predictors, based on coefficient magnitudes, were: Log_Total_Spent_AllTime (0.233), Total_Spent_AllTime (0.150), Transaction_Count_AllTime (-0.131), and Total_Spent_6M (0.124). Six features were excluded by the model, with their coefficients set to zero or near-zero: Transaction_Count_6M, Brand Description_Retinol LRP, Transaction_Count_3M, EAN_AntheliosInvisibleSunscreen50ml, Sub-Category_Sunscreen, and Purchased_Tinted Sunscreen. This suggests these variables have minimal impact on the target variable when considering the other predictors.

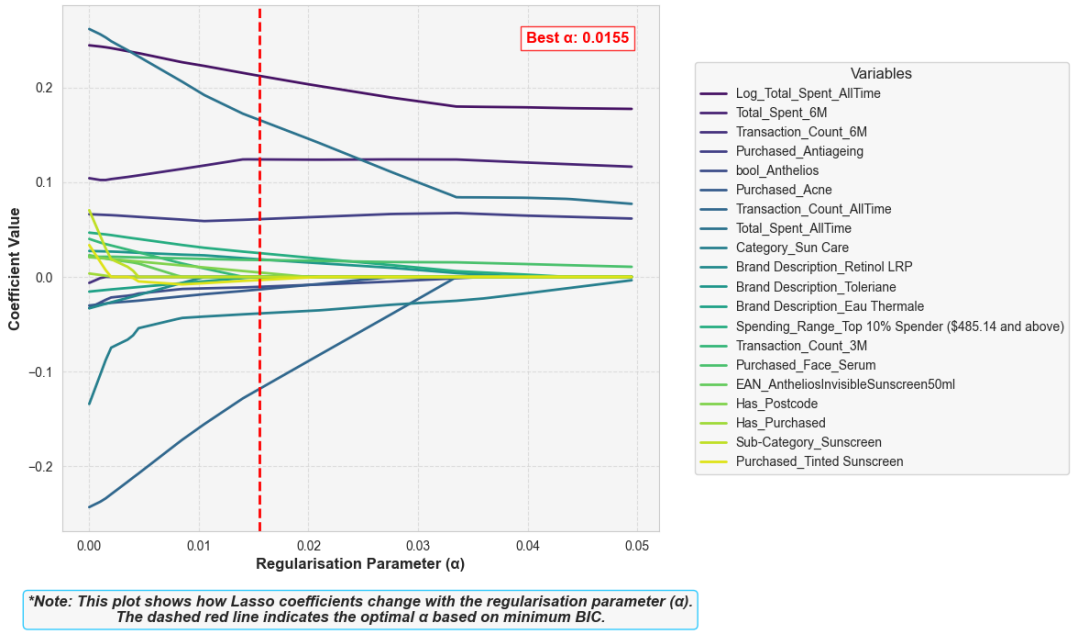


Figure 13: LASSO Model Performance with Zero Threshold RMSE Subset

In contrast to the 20-variable LASSO model, the 4-variable model derived from the 0.01 threshold RMSE forward selection yielded an optimised alpha of 0, indicating all predictors are significant. Table 8 summarises the results of both LASSO models, alongside the LASSO-transformed AIC and BIC models:

Original MLR Model	LASSO Alpha	RMSE
RMSE Criterion (No Threshold)	0.02	0.56
RMSE Criterion (0.01 Threshold)	0.00	0.56
AIC Criterion	0.02	0.56
BIC Criterion	0.00	0.56

Table 8: LASSO Summary Table

The analysis reveals that the LASSO models do not provide significant performance improvements over the original multiple linear regression models. Moreover, the standardisation and iteration processes inherent in LASSO modelling introduce additional complexity, potentially compromising the interpretability of the results. Given the importance of deriving clear business insights from the model, the marginal benefits of LASSO do not justify the loss in interpretability. Consequently, further exploration of LASSO techniques will not be pursued in this analysis.

4.4.3 Gradient Boosting

Gradient boosting combines multiple weak learners into an ensemble, iteratively fitting each model to the residuals of the previous model. This allows gradient boosting to capture complex patterns in the data but also makes it prone to overfitting. The feature selection algorithm, using a fixed threshold, selected only two features: 'Total_Spent_12M' and the interaction term 'Category_Face Care:Purchased_Face_Serum'. These features were then passed into a grid search to optimise hyperparameters, including learning rate, number of estimators, and tree depth for each weak learner. The resulting 5-fold cross-validated log RMSE was 0.58.

4.4.4 Random Forest

The Random Forest method was initially applied to the entire dataset to explore its potential for improving model performance. This approach yielded an optimum forest with a RMSE of 0.55. However, two significant issues emerged: the inclusion of all independent variables in the model raised substantial concerns about overfitting, and while the Random Forest model showed a marginal improvement of 0.01 in RMSE compared to the best linear regression model, this came at the cost of significantly reduced interpretability.

To address these issues, an alternative approach combining forward selection with grid search optimisation was considered. However, this approach was deemed computationally infeasible due to the exponential increase in required fits. For instance, a 5-fold grid search with 180 candidates per fold would result in 900 fits. Incorporating forward selection would escalate this to tens of thousands of fits, likely without proportional improvement in model performance.

Given the computational constraints, the focus shifted to assessing the relative importance of variables. The performance of reduced models, each missing one predictor variable, was compared to evaluate feature importance. The top 20 variables by relative importance are visualised in Figure 14.

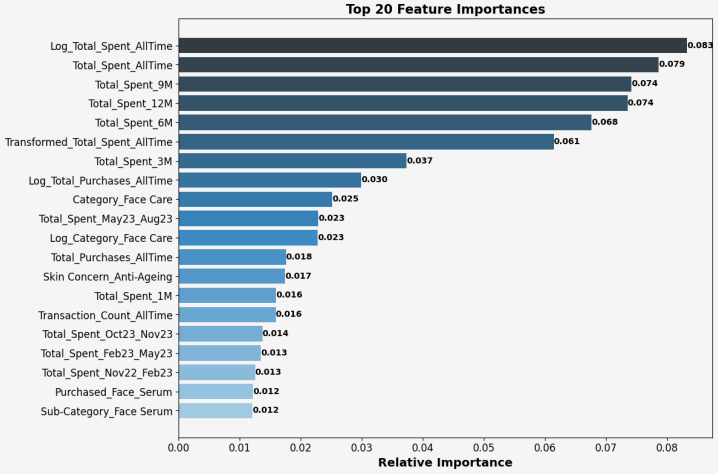


Figure 14: Random Forest Feature Importance

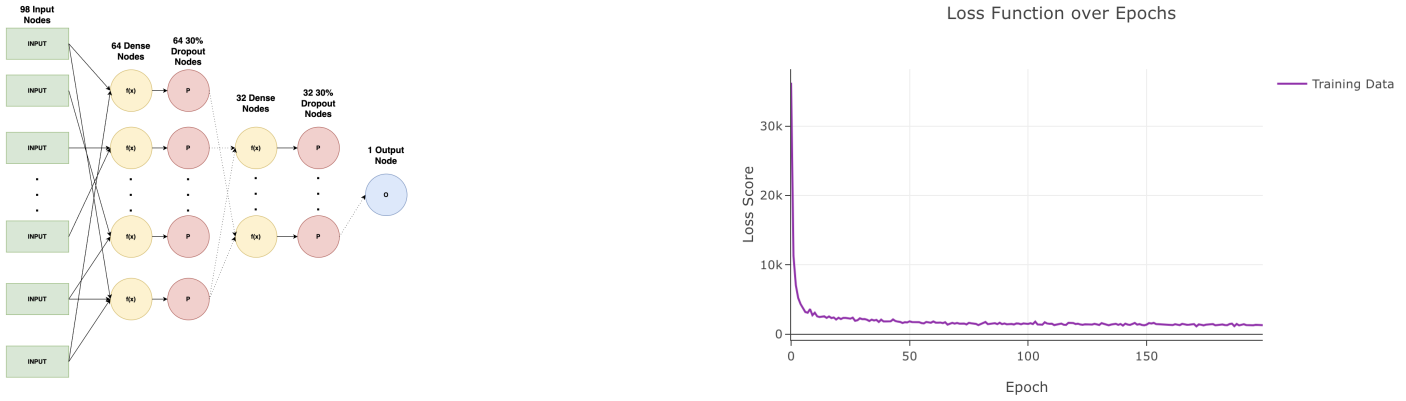
To mitigate overfitting, a refined model was constructed using only the top 10 variables identified in the importance analysis. This refined model achieved an RMSE of 0.56. However, this performance did not surpass the full-variable Random Forest

model (RMSE 0.55) or show improvement over the linear regression model.

The Random Forest method, despite its sophistication, did not demonstrate significant advantages in this context. The full-variable model risked overfitting and lacked interpretability, while the refined model, although more interpretable, did not offer performance improvements over simpler methods. Given these findings, the Random Forest method will not be utilised for modelling consumer spending in this analysis. The focus will remain on more interpretable models that offer a better balance between performance and insights derivable for business decision-making.

4.4.5 Neural Network

As Neural Networks heuristically "learn" which features are significant through Deep Learning, a neural network was constructed on the entire dataset. While this method can often produce highly accurate models, it is generally impossible to interpret the solution process and is computationally expensive. For the structure of the neural network, the input was fed into a 64 neuron layer, which was reduced to a second 32 neuron layer before being outputted as a single neuron (the final score). As can be seen in Figure 14.a, these layers are separated by dropout layers. These function by randomly eliminating 30% of connections each epoch in order to reduce overfitting.



(a) Neural Network Structure

(b) Loss function score over time during Deep Learning

Figure 15: Neural Network and Deep Learning

After the Deep Learning was preformed on the Neural Network, the trained model returned a low RMSE of 0.21. While this was a significant and positive result, a more sophisticated model could likely have better "learned" the relationships in the data and produced a lower RMSE. However, the simplicity of the model was bounded by the limitation of our computational resources. Greater computer power would have potentially allowed for a longer training period, which is necessary as accuracy improves (the loss function decreases) slowly after the initial few epochs, as seen in Figure 14.b.

While the RMSE was substantially low, even compared to the other models, the Neural Network has little to no interpretability in a business case. This is due to the fact that the neural network "learns" patterns by adjusting the weights of various inputs via a non-linear function to produce a correct answer. Therefore, it is impossible to map individual features linearly to the output of the model, and thereby derive business insights.

Overall, due to the loss of interpretibility in using a Neural Network and the limited improvement in RMSE metrics with the given computational power, the decision was made to not pursue a Neural Network as a potential solution.

4.4.6 Second Multi-Stage Linear Model

Given the marginal increases in predictive accuracy with more complex models, a linear regression model was selected. This choice allows for direct interpretation of coefficients, providing valuable insights for revenue estimation.

During the data exploration phase, a significant skew in the expenditure data was identified. To address this and increase the model's predictive capacity, a multi-stage model was developed, segmenting consumers based on their historical expenditure across all periods. After experimenting with various splits, such as a model for the top 25%, the focus was ultimately placed on the top 50% of all-time spenders. This group represents 62.51% of the total expenditure in the target period of the training data, suggesting it would be most effective to target higher spenders.

To improve the model's handling of skewed data, a log-transformation was applied to the dependent variable. A forward selection algorithm with the previously discussed threshold was then executed, yielding the coefficients for the high-spending model as shown in Table 9. The coefficients for the low-spending model are presented in Table 10.

Variable	Coefficient
Constant	4.83
Total Spent 12M	0.01
Purchased Anti-ageing	0.09
Category Sun Care	-0.03
Bool Toleriane	0.07
Bool Hyalu B5	0.09
Has Purchased	-0.12
EAN_EffaclarMoisturiser40ml	-0.03

Table 9: Coefficients for High-Spend Linear Regression Model

Variable	Coefficient
Constant	4.40
Total Spent All Time	0.00
Category Sun Care	-0.07
Sub-Category Face Moisturiser	-0.05
Bool Eau Thermale	-0.18
Bool Niacinamide	0.11
Purchased Anti-ageing	0.12
Bool Retinol LRP	-0.09
Has Postcode	0.05
Sub-Category Eye Cream	-0.09
Skin Concern Irritation-Prone Skin	0.02

Table 10: Coefficients for Low-Spend Linear Regression Model

5 Evaluation - Phase V

In selecting an appropriate model, a critical trade-off exists between accuracy and interpretability (Brownlee, 2014). While accuracy is crucial for producing reliable predictions and insights, interpretability is equally important for understanding how the model arrives at its solutions. Models with high accuracy but low interpretability, often referred to as "black box" models, have limited practical applications and may pose ethical concerns, potentially discriminating based on inappropriate metrics ("A Survey Of Methods For Explaining Black Box Models", n.d.).

Appendix A.3 illustrates the spectrum of models ranging from high interpretability to high accuracy. Linear regression, LASSO regression, and ridge regression models offer high interpretability, while neural networks and gradient boosting models prioritise accuracy at the expense of interpretability. This investigation explores models across this spectrum, starting with highly interpretable models and progressing towards those with higher accuracy but lower interpretability. The evaluation of model success incorporates key quantitative and qualitative metrics:

- **Quantitative Metric:** Root Mean Square Error (RMSE) score, an industry-standard measure of model performance.
- **Qualitative Metric:** Interpretability, assessing whether the model's conclusions and predictive processes are easily explainable, crucial for practical business applications and stakeholder understanding.

5.1 Selecting Final Model

Model	log RMSE	Dollar RMSE	Interpretability
Base Linear Regression	0.56	118.54	High
LASSO	0.56	118.50	High
Gradient Booster	0.58	119.85	Low
Random Forest	0.56	118.89	Low
Neural Network	0.36	123.77	Low
Multi-Stage Linear Regression	0.56	119.35	High

Table 11: Results from Model Selection Investigation

Table 11 presents the results of the model selection investigation. Most models produced consistent RMSE values within a similar range. Although the Neural Network model achieved a significantly lower log RMSE of 0.36, its poor interpretability made it unsuitable for a business solution.

Given the lack of clear differentiation in performance among the highly interpretable models, the multi-stage linear regression was selected as the final model. This choice was based on its combination of low RMSE and high interpretability, which aligns well with the project's requirements. In business contexts, stakeholders require a clear understanding of the factors driving model predictions to align decisions with business strategy. While models like gradient boosters and neural networks can capture complex patterns, their lack of transparency limits their utility for generating actionable insights.

5.2 Final Model:

The multi-stage linear regression model strikes an optimal balance between low RMSE and high interpretability, making it the most suitable choice for predicting consumer transaction value. Table 12 summarises the key advantages of the chosen solution.

Aspect	Comment
Low RMSE	Multi-Stage Linear Regression achieved a dollar RMSE of 119.35, demonstrating strong predictive accuracy.
Flexibility	The model adapts well to different consumer segments, capturing nuanced relationships and offering more tailored insights.
High Interpretability	As a linear model, it provides transparency, allowing businesses to understand how individual features influence predictions clearly. This is essential for informed decision-making.
Actionable Business Insights	The model's feature selection highlighted specific product categories, offering clear opportunities for targeted revenue growth and strategic planning.

Table 12: Positive Aspects of the Final Model

5.3 Final Model Evaluation on Unseen Test Data

Metric	Low Spenders	High Spenders	Overall
RMSE	89.27	140.66	116.25

Table 13: Final Model Test Results on Unseen Test Data

In order to validate the choice of linear regression, and specifically a multi-stage linear regression method, the model was evaluated on unseen test data using RMSE scores. The results of this can be seen in Table 12. Two insights immediately become clear analysing these output RMSE. For one, the weighted RMSE of 116.25 is relatively low suggesting that the model has strong predictive power for pastumer transaction value. Additionally, the test RMSE is smaller than the train

RMSE, which suggests that the model is not over-fitting the training data, indicating that the model generalises well and will offer reliable insights into customer spending behaviour.

6 Deployment - Phase VI

6.1 Project & Recommendations

The model results indicate that purchases of specific brands and products belonging to skin-concern categories are strong predictors of increased customer spending. Leveraging this insight, a strategic initiative to boost sales in these specific product categories has the potential to significantly enhance overall customer spending.

The proposed project involves optimising La Roche Posay's existing skin care quiz. This interactive tool, which currently assesses individual skin concerns and types to recommend suitable products, can be enhanced by integrating the model's parameters. This integration will allow for targeted recommendations of products known to correlate with higher spending patterns. Notably, the model's coefficients reveal that spending is not solely driven by product price. For instance, Vitamin C products, despite being the second most expensive brand, does not appear in the model. This suggests that factors such as customer satisfaction, leading to repurchases, or positive experiences encouraging further brand exploration, may be related to purchases from the products included in the model.

Given these insights, the enhanced quiz will prioritise recommendations from the selected brands and product categories, where appropriate for the customer's skin profile. This data-driven approach aims to optimise product recommendations, potentially increasing both customer satisfaction and overall revenue. The project is broken down into several key recommendations, aimed at meeting the success criteria described in Section 1, with detailed discussion on their impact following.

No.	Project Recommendation	Related Objective
1	**Optimise Product Recommendations:** Reprogram the Skincare Quiz algorithm to recommend products that increase spending while maintaining the integrity and relevance of suggestions.	Increase purchase volume/profit (Section 6.1.1)
2	**Enhance Email Timing** Timing email sending based on the periods in the model with positive beta values to maximise the effectiveness of the emails	Increase purchase volume/profit (Section 6.1.2)
3	**Incorporate User Feedback System:** Develop a rating system allowing customers to evaluate the effectiveness of purchased products. Utilise historical purchase data to refine future recommendations and avoid suggesting products that received negative feedback.	Enhance Personalisation (Section 6.1.3)
4	**Leverage Quiz Data for Business Intelligence:** Analyse data collected from the quiz to gain insights into customer needs and skin concerns. Use these insights to enhance personalisation efforts and inform future product development initiatives.	Enhance Personalisation (Section 6.1.4)

6.1.1 Optimise Product Recommendations

Description:
The deployment of the model will occur in two stages. In the first stage, customers will be segmented into two broad spending groups: the top 50% and bottom 50%. Based on these segments, two differently programmed quiz types will be tailored to provide personalised product recommendations. These will be sent to the patsumers recorded in the data set via email.

- **Recommendations target group "top 50% of Spenders:"** When multiple product "solutions" address a customer's skincare concerns, the recommendation will consistently prioritise products from the brands Toleriane and Hyalu B5, as well as products targeting the "Anti-Ageing" skin concern. At least one product from these categories will be recommended, while still ensuring the recommendations genuinely align with the customer's specific needs. Hyaluronic Acid serum, a key product in the Hyalu B5 range, is suitable for all skin types and effectively addresses a wide range of common skincare concerns, will always be recommended (Sharkey, 2020).
- **Recommendations for target group "bottom 50% of Spenders:"** Similarly, for the bottom 50%, recommendations will prioritise products belonging to the Niacinamide brand, and Anti-Ageing skin concern, which exhibited the strongest positive coefficients in the final model.

Rationale:
The forward-selection model identified binary indicators for these specific categories, demonstrating their strong predictive power in influencing customer spending. The positive coefficients associated with these variables suggest that purchasing products in these categories increases overall spending, compared to other categories like sun care, acne-prone skin, and body moisturises, which exhibit negative coefficients, indicating a downward impact on spending.

Projected Impact on Revenue:
Based on industry average of a 9.89% click-to-conversion rate (Richardson, 2024), an estimated 316 out of 3,200 targeted customers from each group ($3200 \times 0.0989 = 316$) will purchase a recommended product. The model predicts a spending increase for customers buying from specific categories. According to the model, the average increase in spending is approximated by the mean of the recommended brands coefficients:

- An 8% increase in spending for the top 50%, resulting in their average spend of \$194.65 in the most recent period to increase to \$210.87. This groups projected revenue contribution will become $\$210.87 \times 316 = \$66,736.40$.
- A 15.5% increase in spending for the bottom 50%, resulting in their average spend of \$129.00 to increas to \$143.84. This groups projected revenue contribution will become $\$143.84 \times 316 = \$45,520.90$.

Varibales:	Top 50%	Bottom 50%
Number of Patsumers	3200.00	3200.00
Conversion Rate:	9.89%	
Number of Converted Patsumers:	316	316
Average Spend (Nov23-May24)	$194.65	$129.00
Mean of Coefficients:	0.08	0.115
Increase in Spend:	$16.22	$14.84
Total Average Spend:	$210.87	$143.84
Revenue change (Converted Patsumers):	$ 66,736.40	$ 45,520.90
Revenue (Nov23-May24)	$1,043,191.00	
% of Revenue Made from Group	62.51%	37.49%
% of Non Converted Customers	0.90	0.90
Revenue Contribution (Non Converted):	$ 587,606.13	$ 352,413.28
Total Revenue:	$1,052,276.71	
Increase in Revenue:	$9,085.71	

Figure 16: Projected Impact of Recommendation Optimisation

The total revenue from online sales in Nov23-May24 is \$1,043,191. Adding the increased spend of converted patsumers to the expected spend of non-converted patsumers (calculated by multiplying the revenue in the previous period by the % of non-converted patsumers and their respective % revenue contributions), the increase in revenue is expected to be \$9,085.71 higher than that of the previous period.
Note: Analysis assumes uniform spending patterns; actual results may vary.

Expected Costs:
Given La Roche Posay's existing email marketing infrastructure and established skin care quiz, the primary costs associated with this project are focused on the strategic reprogramming of the quiz. The estimated breakdown is as follows:

- Reprogramming to implement product prioritisation: 2 hours
- Integration with existing systems: 2 hours
- Quality assurance and testing: 2 hours

Utilising the standard software developer rate of \$150 per hour ("Software Developer Salary in AU (October, 2024)", n.d.), the projected cost is calculated as: Total Cost = 6 hours × \$150/hour = \$900. *Note: This estimate assumes no unforeseen technical challenges and is based on the current understanding of La Roche Posay's systems. Any additional requirements or complexities identified during the implementation phase may affect the final cost.*

Impact on Profit:
The expected increase in profit is equivalent to the increase in revenue subtracting associated costs: $\$9,085.71 - \$900 = \$8,185.71$.

Impact on Number of Units Sold:
Assuming that every converted patsumer makes a purchase, an additional 632 additional units will be sold following the project intervention.

6.1.2 Enhance Email Timing

Description:
The quiz will be sent to patsumers via email based on the time after the most recent purchase. At 6 and 12 months after the recent purchase, the email containing the personalised skincare quiz will be sent to the past patsumer. This aims to capitalise on periods of increased likelihood of patsumer repurchasing, further encouraging a repurchase. In conjunction with this, increased mailing will occur between October and November, as this has been highlighted as a period of increased repurchase. Note that this enhancement will target all patsumers.

Rationale:
Firstly, positive betas for time-series predictors were added in the forward selection loop throughout the modelling phase. Although the variable controls restricted the inclusion of multiple time-series predictors due to their high collinearity, it still indicates a high correlation between when a patsumer has last purchased. Secondly, when using a differences time series, the EDA highlighted an uptick of purchasing at 6 months into the historical data. This can be interpreted as either a result of the relative time after purchase or the specific time of the year.

Projected Impact on Revenue:
Taking the conservative value of 10% and applying it to the existing 9.89% click-to-convergence rate gets an increase to 10.88%. This results the number of converted patsumers in each group to rise to 384, and an $20,311.44 increase in revenue.

Varibales:	Top 50%	Bottom 50%
Number of Patsumers	3200.00	3200.00
Conversion Rate:	9.89%	
Adjusted Conversion Rate:	10.88%	
Number of Converted Patsumers:	348	348
Average Spend (Nov23-May24)	$194.65	$129.00
Mean of Coefficients:	0.08	0.115
Increase in Spend:	$ 16.22	$ 14.84
Total Average Spend:	$210.87	$143.84
Revenue change (Converted Patsumers):	$ 73,410.04	$ 50,072.99
Revenue (Nov23-May24)	$1,043,191.00	
% of Revenue Made from Group	62.51%	37.49%
% of Non Converted Customers	0.90	0.90
Revenue Contribution (Non Converted):	$ 587,606.13	$ 352,413.28
Total Revenue:	$1,063,502.44	
Increase in Revenue:	$20,311.44	
Cost of Reporgramming	$900.00	
Increase in Profit:	$19,411.44	

Figure 17: Projected Impact of Email Timing Optimisation

Expected Costs:
Beyond the expected costs to reformat the quiz, there are no additional expenses associated with this aspect of the deployment.

Impact on Profit:
The expected increase in profit for this recommendation specifically can be described as the resulting profit subtracted by the profit of the previous recommendation: $\$20,311.44 - \$900.00 = \$19,411.44$.

Impact on Number of Units Sold:
Assuming that every converted patsumer makes a purchase, an additional 696 units will be sold, which is a 64 unit increase from the baseline optimisation.

6.1.3 Incorporate User Feedback System

Description:
The user feedback system will help improve product recommendations by allowing customers to rate the products that they have previously bought. These ratings can then be incorporated into future recommendations to boost customer satisfaction and avoid recommending those with negative feedback. Coupling this product feedback with historical purchase data will create a more refined personalisation system that will help boost sales.

Rationale:
When conducting the exploratory data analysis, it became very clear that more customer-specific data was required. Through the implementation of this system, more data will be collected on customer preferences and purchase patterns. This will

help create a more accurate model that can better predict customer satisfaction and therefore increase personalisation.

Projected Impact on Revenue:
The main expected impact from this recommendation is the increase in conversion rates through more accurate recommendations. A study completed by McKinsey has indicated advanced personalisation techniques increase revenue by 10-15%. However, this assumes the revenue of customers who have already clicked through. Based on previous calculations, an increase of 15% would result in the change in revenue being boosted to $31,324.10. This is also a large increase compared to just optimising the product recommendations, as it is able to more specifically target customer preferences. Also not captured in these calculations is the expected increase in customer loyalty, further increasing revenue over time.

Varibales:	Top 50%	Bottom 50%
Number of Patsumers	3200.00	3200.00
Conversion Rate:	9.89%	
Adjusted Conversion Rate:	10.88%	
Number of Converted Patsumers:	348	348
Average Spend (Nov23-May24)	$194.65	$129.00
Mean of Coefficients:	0.08	0.115
Increase in Spend:	$ 16.22	$ 14.84
Total Average Spend:	$210.87	$143.84
Revenue change (Converted Patsumers):	$ 73,410.04	$ 50,072.99
Revenue (Nov23-May24)	$1,043,191.00	
% of Revenue Made from Group	62.51%	37.49%
% of Non Converted Customers	0.90	0.90
Revenue Contribution (Non Converted):	$ 587,606.13	$ 352,413.28
Increase in Revenue:	15%	
Revenue change (Converted Patsumers):	$84,421.55	$ 50,074.14
Total Revenue:	$1,074,515.10	
Increase in Revenue:	$31,324.10	
Cost of Reporgramming	$1,800.00	
Increase in Profit:	$29,524.10	

Figure 18: Projected Impact of User Feedback System

Expected Costs:
The additional costs associated with this recommendation follow the expected costs from optimising the product recommendations. As an approximation, it would take the same amount of time to implement and as such cost around $1,800.

Expected Impact on Profit:
Accounting for additional increases in costs, the expected profit is $29,524.10, which is a $10,112.66 increase from the previous recommendation.

Expected Impact on Number of Units Sold:
While increase in spending implies additional products have been purchased, the varying price of products results in difficulty quantifying the additional units sold. Assuming an average product cost of $47.65 (Taken as the mean of average brand pricing in Table 2), the number of additional units can be approximated by the change in revenue between the previous recommendation: $\$29,975.02 - \$11,263.4 = 393$ additional units.

6.1.4 Leverage Quiz Data for Business Intelligence

Description:
The newly acquired data, specifically from the repeat patsumers, can be linked to the existing patsumer database, allowing for more complex feature engineering. Information such as skin type, ratings of products used and patsumers, skin concern and regularity of breakouts can be used to create a more precise model.

Rationale:
The rationale stems from the difficulties faced when creating the model in phase IV. Issues of high multicollinearity and a lack of variety in the variable types indicated a more complex dataset would allow for more insightful modelling.

Projected Impact:
This implementation strategy aims to enable ongoing model optimisation over time. While not immediately measurable by quantifiable metrics, the long-term vision is to enhance model performance as deployment progresses. Additionally, deeper insights into customer behavior will inform future product development, offering valuable perspectives on La Roche-Posay's customer base. For instance, identifying a significant percentage of customers struggling with dehydrated skin could drive

the expansion of the Hyalu B5 product line. Although the impact on profit and units sold is beyond the scope of this project, such insights can provide valuable strategic direction for the brand.

6.2 Limitations and Drawbacks

As with any recommended strategies, there will always be drawbacks and limitations with the solutions. Rather than focusing on specific issues for each recommendation, overall issues will be addressed. When creating highly personalised marketing strategies, many issues can arise. Personalised marketing strategies emphasise to the patsumer that data has been collected and used to influence them to spend and buy more. It was found by Learmonth (Dhanya and Jaidev, 2019) that consumers will often find personalised advertisements annoying and intrusive, creating negative feelings towards the company. This is demonstrated by the 2012 case where Target would predict personal events of customers based off of purchasing data, such as pregnancies (John et al., 2018), and then recommend products based off of the predictions. This represents a large breach of a consumers privacy and contributes to a sense of the customer being surveilled. In the projects put forward in this report, it is important to ensure that the data is being used ethically, transparent data processes are used, and informed consent is received from the patsumer before any data is used. For example, this can be achieved by putting an acknowledgement at the beginning of the skincare quiz when re-programming it.

Another issue arising from an analytically driven marketing strategy is patsumer trust and confidence. It was found that if the customer dislikes how their information is shared, purchase interest drops (John et al., 2018). So, leveraging too much of the customer data can therefore decrease the effectiveness of the marketing campaign and drive patsumers away. Again, this issue can be avoided through clear and transparent data processes, and not using sensitive data. Therefore, many of the risks that these recommendations bring can be avoided using ethical and transparent data methods that do not take advantage of the patsumer.

6 References

A Survey Of Methods For Explaining Black Box Models. (n.d.). Retrieved October 5, 2024, from https://ar5iv.labs.arxiv.org/html/1802.01933

Assessment information and resources: QBUS3600 Business Analytics in Practice. (n.d.). Retrieved September 25, 2024, from https://canvas.sydney.edu.au/courses/59822/pages/assessment-information-and-resources

Australia, M. P., Melbourne. (n.d.). Matthew Proctor. Retrieved October 20, 2024, from https://www.matthewproctor.com

Brownlee, J. (2014, July). Model Prediction Accuracy Versus Interpretation in Machine Learning. Retrieved October 5, 2024, from https://machinelearningmastery.com/model-prediction-versus-interpretation-in-machine-learning/

Chakrabarti, A., & Ghosh, J. K. AIC, BIC and Recent Advances in Model Selection (P. S. Bandyopadhyay & M. R. Forster, Eds.). In: *Philosophy of Statistics* (P. S. Bandyopadhyay & M. R. Forster, Eds.). Ed. by Bandyopadhyay, P. S., & Forster, M. R. Vol. 7. Handbook of the Philosophy of Science. Amsterdam: North-Holland, 2011, January, pp. 583–605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0. Retrieved September 29, 2024, from https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf

Dhanya, D., & Jaidev, U. P. (2019). Perceived Personalization, Privacy Concern, e-WOM and Consumers' Click Through Intention in Social Advertising. *International Journal of E-Services and Mobile Applications*, *11*(4), 39–55.

Duval, A. (2019, April). *Explainable Artificial Intelligence (XAI)*. https://doi.org/10.13140/RG.2.2.24722.09929

Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Machine Learning and Knowledge Extraction*, *1*(1), 359–383. https://doi.org/10.3390/make1010021

Experian Mosaic. (n.d.). Retrieved September 26, 2024, from https://mosaic.experian.com.au/compare

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications âĂŞ a holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

John, L., Kim, T., & Barasz, K. (2018, January). Ads That DonâĂŹt Overstep. Retrieved October 19, 2024, from https://hbr.org/2018/01/ads-that-dont-overstep

La Roche-Posay Skincare Official Site | La Roche-Posay Australia. (n.d.). Retrieved October 20, 2024, from https://www.laroche-posay.com.au/?gad_source=1&gclid=CjwKCAjw1NK4BhAwEiwAVUHPULF3rtlhXwxSYLRnha4zaIPwr2drIIpYUCyaP-nct9Q79UBoCNgkQAvD_BwE&gclsrc=aw.ds

L'OrÃľal Group: Dermatological Beauty Division. (n.d.). Retrieved September 25, 2024, from https://www.loreal.com/en/division-beaute-dermatologique/

Richardson, A. (2024, April). *Cosmetic and Toiletry Retailing in Australia* (Industry Report No. G4271B). IBISWorld. https://my.ibisworld.com/au/en/industry/G4271b/about

Sharkey, L. (2020, October). How to Use Hyaluronic Acid âĂŤ and Why You Should. Retrieved October 19, 2024, from https://www.healthline.com/health/beauty-skin-care/how-to-use-hyaluronic-acid

Software Developer Salary in AU (October, 2024). (n.d.). Retrieved October 21, 2024, from https://www.seek.com.au/career-advice/role/software-developer/salary

A Appendix

A.1 CRISP-DM Framework

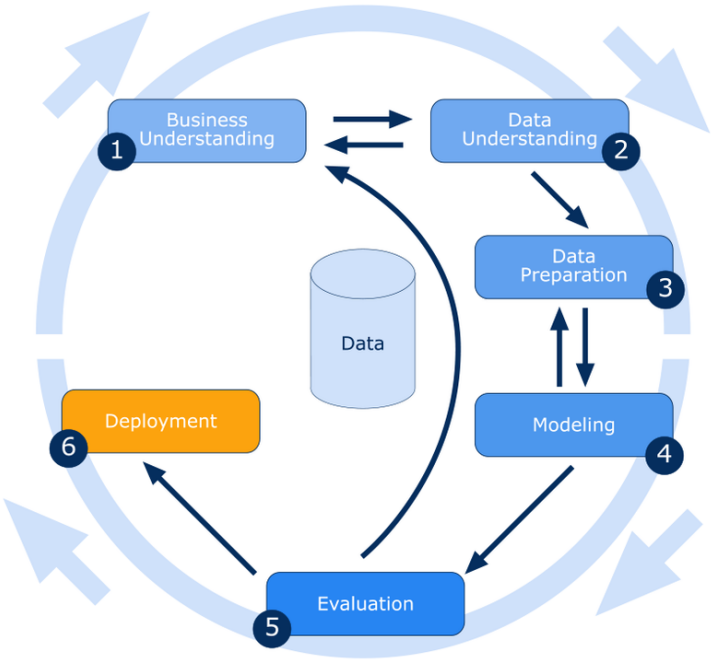


Figure 19: CRISP-DM Framework (Wainaina, 2024)

A.2

Patsumer Map

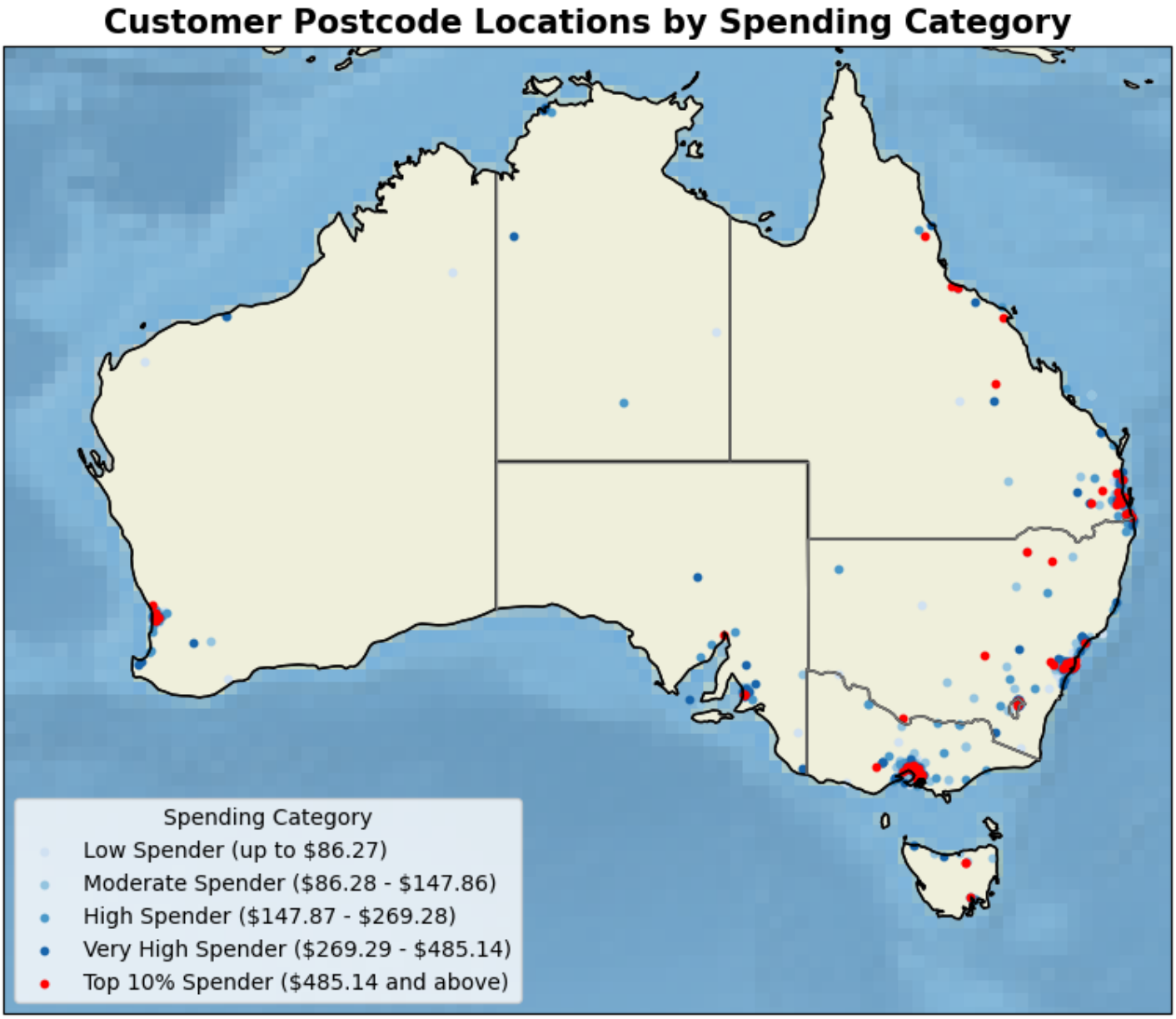


Figure 20: Map of All Patsumers

A.3 Accuracy-Interpretability Trade Off

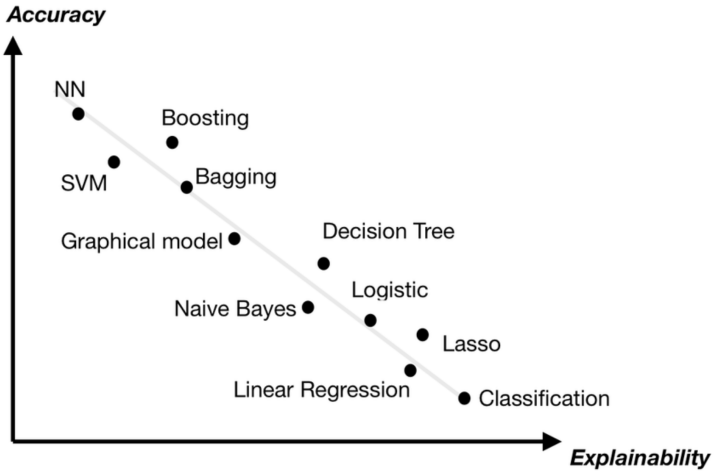


Figure 21: Model Trade Off (Duval, 2019)

A.4 Data Dictionary

Variable Name	Description
Total Spent Nov ’23 to May ’24	The total amount spent 6 months after 20th November 2023, inclusive
Transaction_Count_[X]	How many transactions were made by a customer in the X period prior to the 20th November 2023
Total_Spent_[X]	The value of transactions in the X period prior to the 20th November 2023. X="AllTime" is an unclear timeframe, supposedly dating 2–3 years prior.
Brand-Description_X, Class-Description_X, Category_X, Sub-Category_X, Skin-Concern_X, EAN_X	The number of times that the customer purchased X product type before November 20th 2023

Table 14: Key Variables in Data Set

A.5

Variance Inflation Factors

VIF Category	Variables
VIF = Inf	Brand Description_Anthelios, Brand Description_Bundle, Brand Description_Cicaplast, Brand Description_Eau Thermale, Brand Description_Effaclar, Brand Description_Hyalu B5, Brand Description_Lipikar, Brand Description_Niacinamide, Brand Description_Retinol LRP, Brand Description_Serozinc, Brand Description_Toleriane, Brand Description_Uvidea, Brand Description_Vitamin C, Class Description_Anti-Acne, Class Description_Anti-Ageing, Class Description_Body Care, Class Description_Bundle, Class Description_Face Care, Category_Body Care, Category_Face Care, Category_Sun Care, Sub-Category_Body Moisturiser, Sub-Category_Body Wash, Sub-Category_Eye Cream, Sub-Category_Face Cleanser, Sub-Category_Face Mask, Sub-Category_Face Moisturiser, Sub-Category_Face Serum, Sub-Category_Sunscreen, Sub-Category_Tinted Sunscreen, Sub-Category_Toner & Mist, Skin Concern_Acne-Prone Skin, Skin Concern_Anti-Ageing, Skin Concern_Irritation-Prone Skin, Skin Concern_Pigmentation and Dark Spots, Skin Concern_Sun Protection, Total_Purchases_AllTime
High VIF (>10)	Total_Spent_1M (10.718669), Transaction_Count_1M (10.660767), Total_Spent_3M (21.135946), Transaction_Count_3M (22.262152), Total_Spent_6M (43.459969), Transaction_Count_6M (43.961597), Total_Spent_9M (118.663370), Transaction_Count_9M (112.176746), Total_Spent_12M (155.189239), Transaction_Count_12M (141.306464), Total_Spent_AllTime (68.670741), Transaction_Count_AllTime (56.915049), EAN_Pure10NiacinamideSerum (13.932451), EAN_CicaplastB5BaumeBothSKUs40ml (10.841174)
Moderate VIF (5-10)	EAN_VitaminC10Serum30ml (6.993062), EAN_RetinolB3Serum30ml (8.941791), Post Code (8.039612), Has_Postcode (8.247139)
Low VIF (<5)	EAN_TolerianeMoisturiser40ml (1.800810), EAN_HyaluB5Serum30ml (4.602475), EAN_EffaclarMoisturiser40ml (2.924149), EAN_EffaclarSerum30ml (2.200149), EAN_AntheliosInvisibleSunscreen50ml (3.217108), Brand Description_Anthelios_Output (1.881601), Brand Description_Cicaplast_Output (1.520241), Brand Description_Effaclar_Output (1.952670), Brand Description_Hyalu B5_Output (1.320182), Brand Description_Lipikar_Output (1.422161), Brand Description_Toleriane_Output (1.943037), Brand Description_Vitamin C_Output (1.408710), Has_Purchased (4.407912)

Table 15: Categorisation of Variables Based on Variance Inflation Factor (VIF)

A.6

Significance Tests

Time Period	**All Time**	**Nov23-May24**
U **Stat:**	2,773,205.50	2,773,205.50
p-value	0.13	0.13
Significant ($p < 0.05$)	No	No

Table 16: Mann-Whitney U Test Assessing Differences in Mean Spending Across Postcode Provision

Variable	**Coefficient (β) as %**	**p-value**	**Significant ($p < 0.05$)**
Model 1: Interaction between Face Care and Face Serum			
Category_Face Care	73%	0.00	Yes
Purchased_Face_Serum	-2.1%	0.00	Yes
Interaction_FaceCare_Serum	5.2%	0.00	Yes
Model 2: Interaction between Face Care and Face Moisturiser			
Category_Face Care	80%	0.00	Yes
Purchased_Face_Moisturiser	-4.8%	0.00	Yes
Interaction_FaceCare_Moisturiser	5.6%	0.00	Yes
Model 3: Interaction between Face Care and Irritation-Prone Skin			
Category_Face Care	6.3%	0.00	Yes
Purchased_Irritation	1.8%	0.00	Yes
Interaction_FaceCare_Irritation	0.17%	0.00	Yes
Model 4: Interaction between Face Care and Anti-Ageing			
Category_Face Care	65%	0.00	Yes
Purchased_Antiageing	-1.5%	0.00	Yes
Interaction_FaceCare_Antiageing	4.5%	0.00	Yes

Table 17: Coefficients & Significance for Interaction Models with Face Care Purchases on Total Spent (All Time)

Variable	**Coefficient (β) as %**	**p-value**	**Significant ($p < 0.05$)**	**Correlation**
Model 1: Cumulative Time Periods				
Total_Spent_1M	0.02%	0.21	No	0.22
Total_Spent_3M	0.02%	0.14	No	0.28
Total_Spent_6M	0.04%	0.00	Yes	0.37
Total_Spent_9M	0.01%	0.71	No	0.40
Total_Spent_12M	0.02%	0.19	No	0.41
Total_Spent_AllTime	0.06%	0.00	Yes	0.41
Model 2: Differenced Time Periods				
Total_Spent_Oct23_Nov23	0.17%	0.00	Yes	0.22
Total_Spent_Aug23_Oct23	0.15%	0.00	Yes	0.18
Total_Spent_May23_Aug23	0.13%	0.00	Yes	0.26
Total_Spent_Nov22_May23	0.09%	0.00	Yes	0.25
Total_Spent_Start-Nov22	0.06%	0.00	Yes	0.14

Table 18: Coefficients, Significance, and Correlations for Time Periods with Total Spent (Nov23-May24)

A.7

Yeo-Johnson Transformation

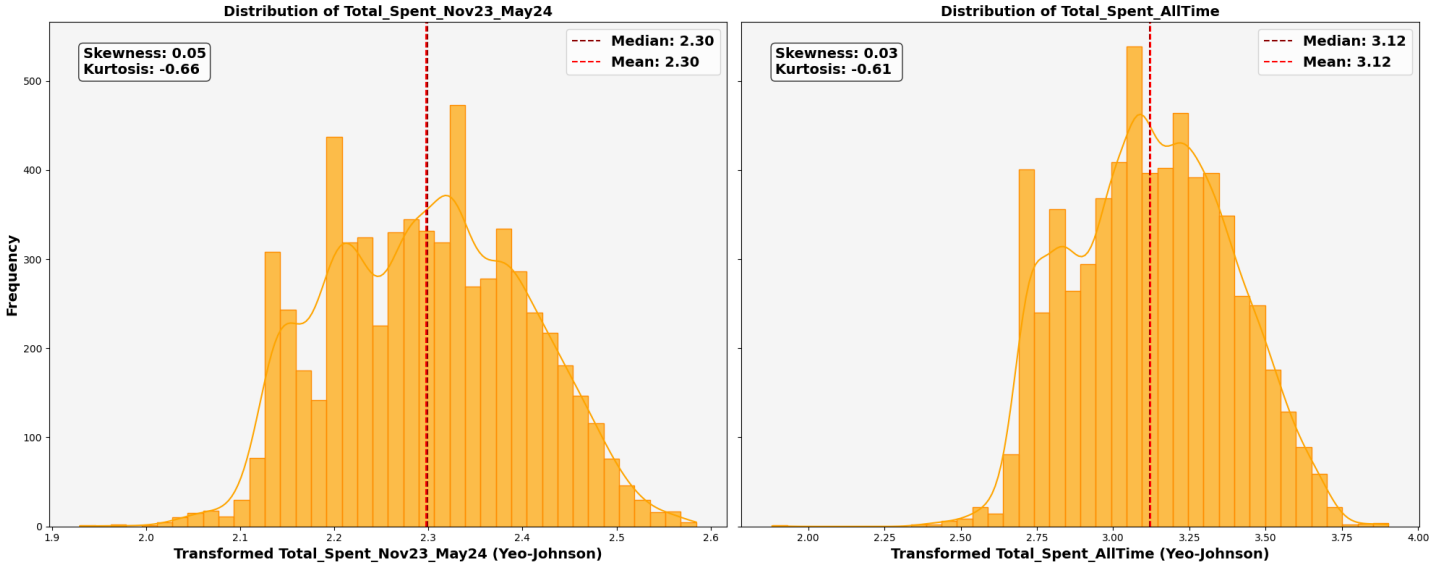


Figure 22: Yeo-Johnson Transformed Histogram of Total Spent

A.8

Postcode Provision Box Plot

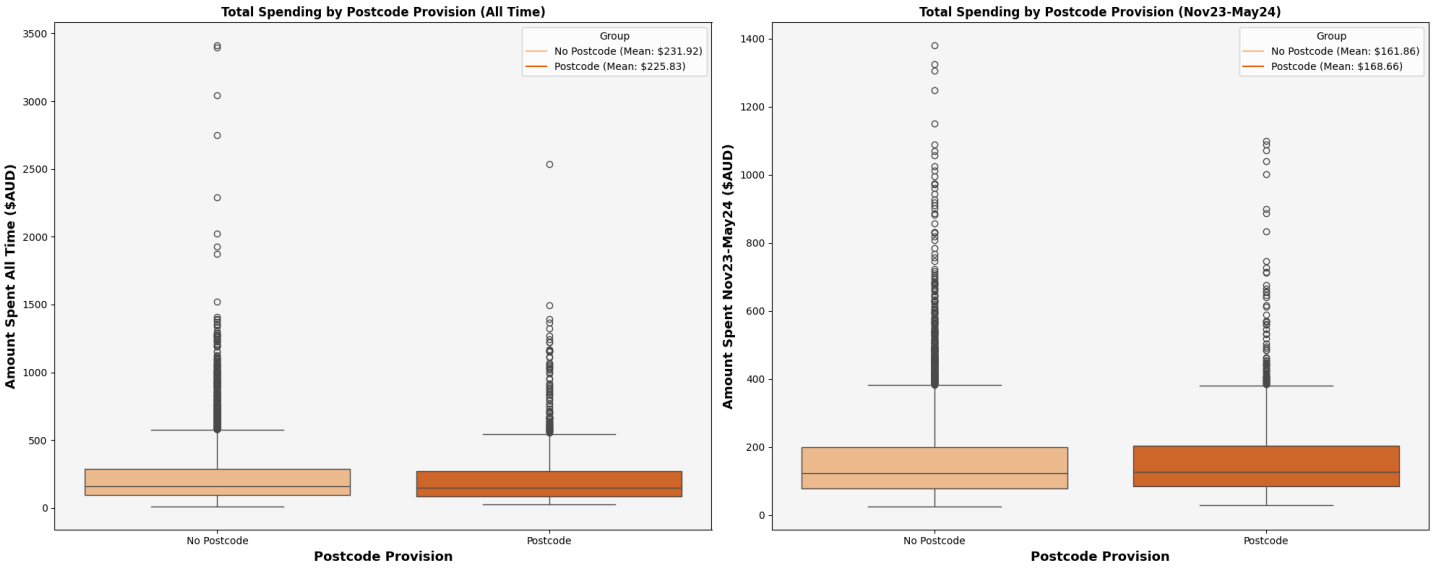


Figure 23: Box Plot of Total Spent by Postcode Provision

A.9

Mosaic Data Index

Index	Segment Name	Segment Description
A	First Class Life	Wealthiest Group in Australia, typically older middle-aged families with significant assets and income.
B	Comfortable Foundations	Gen X families with school-aged children, working in white collar professions and living in suburban areas.
C	Striving for Success	Young, successful, careerdriven professionals living in central city areas with high income and no children
D	Secure Tranquility	Affluent retirees living in higher valued properties in desirable areas
E	Family Fringes	Middle-aged traditional families living on large outer-suburban plots, with comfortable incomes and long commutes
F	Establishing Roots	Millennial first home buyers, living 10km+ from the city centre with above average income
G	Growing Independence	Educated millennials at the start of their careers, renting apartments close to city-centres.
H	Middle BlueCollars	Younger bluecollar workers renting far away from city centres, with below average income.
I	Traditional Pursuits	Average income traditional families & single parents with school-aged children living in outersuburban and regional locations
J	True Grit	Bluecollar households in gainful employment, residing in locations across outer suburban, regional and mining towns.
K	Mature Freedom	Gen X couples without children, renting apartments and terraces in high growth suburbs.
L	Hardship & Perseverance	Unemployed and blue collar workers living in units and flats on low incomes.
M	Graceful Ageing	Older retirees with below average income, living in owned properties or retirement villages.
N	Rural Commitment	Rural people working in agriculture, living on large plots of land far from main roads and main towns.

Table 19: Mosaic Data Index Descriptions

A.10

Lasso Coefficients

Feature	Coefficient	Selected by Lasso
Log_Total_Spent_AllTime	0.232776	Yes
Total_Spent_6M	0.124149	Yes
Transaction_Count_6M	-0.000000	No
Purchased_Antiageing	0.053360	Yes
bool_Anthelios	-0.011742	Yes
Purchased_Acne	-0.018134	Yes
Transaction_Count_AllTime	-0.130657	Yes
Total_Spent_AllTime	0.150140	Yes
Category_Sun Care	-0.031686	Yes
Brand Description_Retinol LRP	-0.000000	No
Brand Description_Toleriane	0.016983	Yes
Brand Description_Eau Thermale	-0.005386	Yes
Spending_Range_Top 10% Spender ($485.14 and above)	0.017646	Yes
Transaction_Count_3M	0.000000	No
Purchased_Face_Serum	0.026493	Yes
EAN_AntheliosInvisibleSunscreen50ml	-0.000000	No
Has_Postcode	0.005991	Yes
Has_Purchased	-0.002043	Yes
Sub-Category_Sunscreen	-0.000000	No
Purchased_Tinted Sunscreen	-0.000000	No

Table 20: Lasso Coefficients and Feature Selection

A.11

Presentation Slides

Drafted Presentation slides - to be updated.

Enhancing LDB's E-commerce Strategy

A PROJECT FOR LA ROCHE-POSAY

QBUS3600 Group Presentation
Group 10

OUR TEAM

Shelby Narborough
510467010

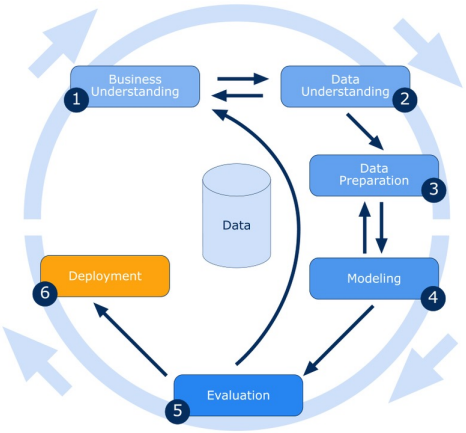
Thomas Hay
510510730

Max Harper
510471039

Eva Wright
520405938

James Dwyer
510518183

DATA STRATEGY – CRISP DM



BACKGROUND

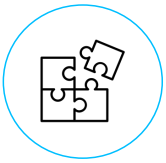
La Roche Posay
A multinational skin-care brand historically retailed through large-chains pharmacies and supermarkets.

Business Analytics Principles
Developing a regression model to predict patsumer transaction values.

Expanding Online Presence
Aiming to move beyond traditional distribution channels to reach new patsumers.

Discovering Market Patterns
LDB can prioritise marketing efforts by uncovering key factors that influence spending patterns.

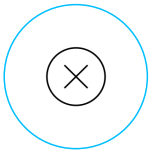
Personalisation Strategy
Leverage data insights to understand patsumer behaviours and preferences better.



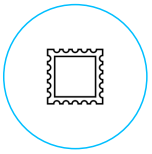
Our Solution
We propose an automated, customer-centric and scalable marketing solution that targets ideal patsumers.

DATA QUALITY

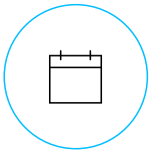
Exclusion of non-website purchases
The dataset is sourced only through transactions that occurred on the La Roche Posay Website



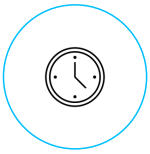
Patsumers with zero purchases
10% of observations have no recorded categorical purchase history



Invalid postcodes
There are discrepancy postcodes when querying against a legitimate postcode database

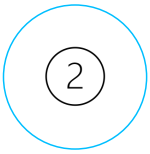


Cumulative periods
The transactional history data is stored cumulatively



Exclusion of non-recent purchases
Only patsumers who purchased between November of 2023 and May of 2024 are included

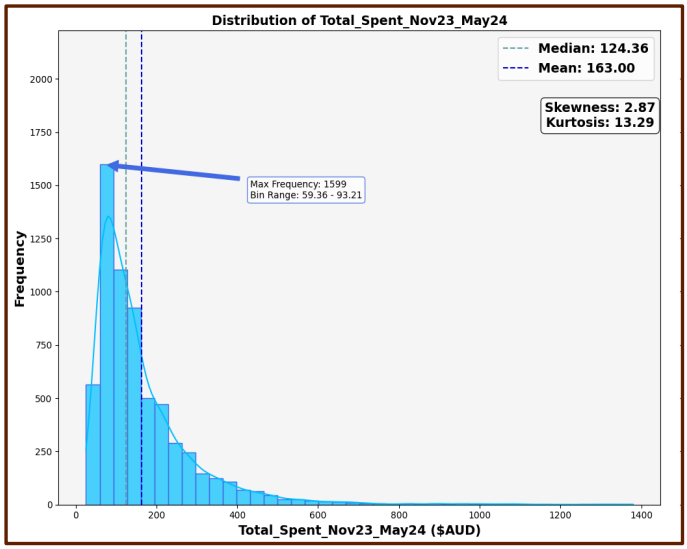
Missing postcodes
5326 patsumers have no postcode data



Duplicate columns
There are two columns with an identical name, representing the same data

DATA EXPLORATION

Variable Distribution



KEY INSIGHTS:

- All variables exhibited a strong positive skew
- The majority of patsumers are making. transactions/purchasing products of lower value/quantity.
- A smaller number of high value transactions and high purchase quantities pull the mean up higher than the median, causing a positively-skewed distribution.

DATA EXPLORATION

Exploring Multicollinearity

Variance Inflation Factor	Variables
Infinite VIF: $VIF = inf$	Brands, Classes, Categories, Sub-Categories, Skin Concerns
High VIF: $VIF > 10$	Total Spent, Transaction Count, EAN (Niacinamide & Cicaplast Baume)
Moderate VIF: $5 < VIF < 10$	EAN(Vitamin C Serim, Retinol Serum), Postcode
Low VIF: $VIF < 5$	Remaining EANs

KEY INSIGHTS:

- Variables including brands, classes, subcategories, and skin concerns indicated perfect multicollinearity, as they capture similar aspects of patsumer purchasing behaviour.
- Transaction counts and total spent displayed high VIF, as expected given they measure behaviours over overlapping time frames.

DATA EXPLORATION

Exploring Correlation vs Popularity

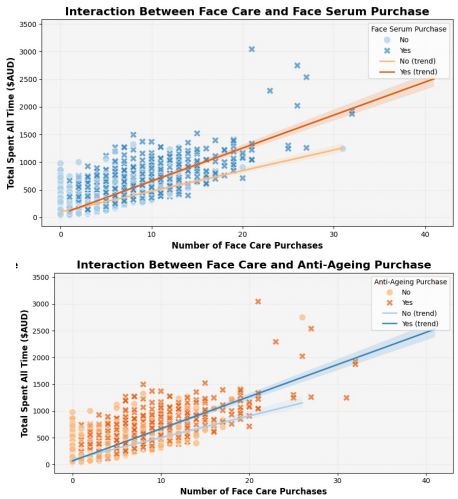
Variable Set	Highest Correlation Coefficient	Most Popular (Proportion Purchased)
Brands	• Hylau B5 (0.23) • Toleriane (0.19) • Vitamin C (0.19)	• Anthelios (36.8%) • Toleriane (35%) • Effeclar (32.4%)
Sub-Categories	• Face Serum (0.27) • Face Moisturiser (0.22)	• Face Moisturiser (50.6%) • Face Serum (43.2%)
Skin Concerns	• Anti-Ageing (0.27) • Irritation-Prone (0.21)	• Irritation (50.1%) • Sun Protection (40.5%)

KEY INSIGHTS:

- Brands like Hylau B5 and Vitamin C exhibited highest correlations with spending, however, are only the 4th and 6th most popular brands
- Anti-Ageing is only the 3rd most popular concern, yet, exhibits the strongest correlation.
- **Inference:** Higher-priced products have a strong influence on spending, despite lower popularity

DATA EXPLORATION

Exploring Interaction Effects

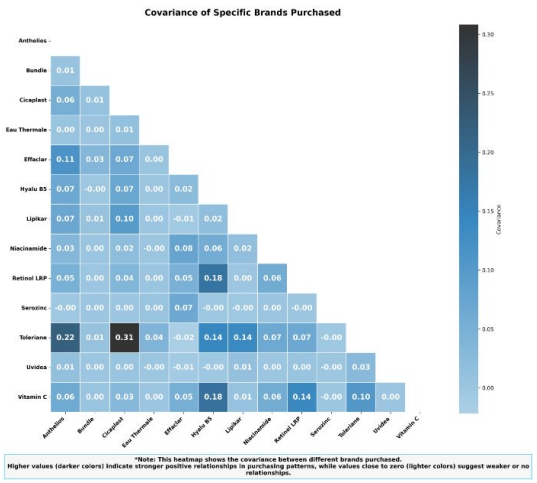


KEY INSIGHTS:

- A significant boosting effect can be seen on the effect of number of face care purchases on spending, given if either a face serum or anti-ageing product is purchased.
- **MLR Analysis:** For each % increase in face care products, patsumer spending increases by 5.16% if they purchase a Face Serum, and 4.47% if they purchase an Anti-Ageing product.

DATA EXPLORATION

Exploring Covariance

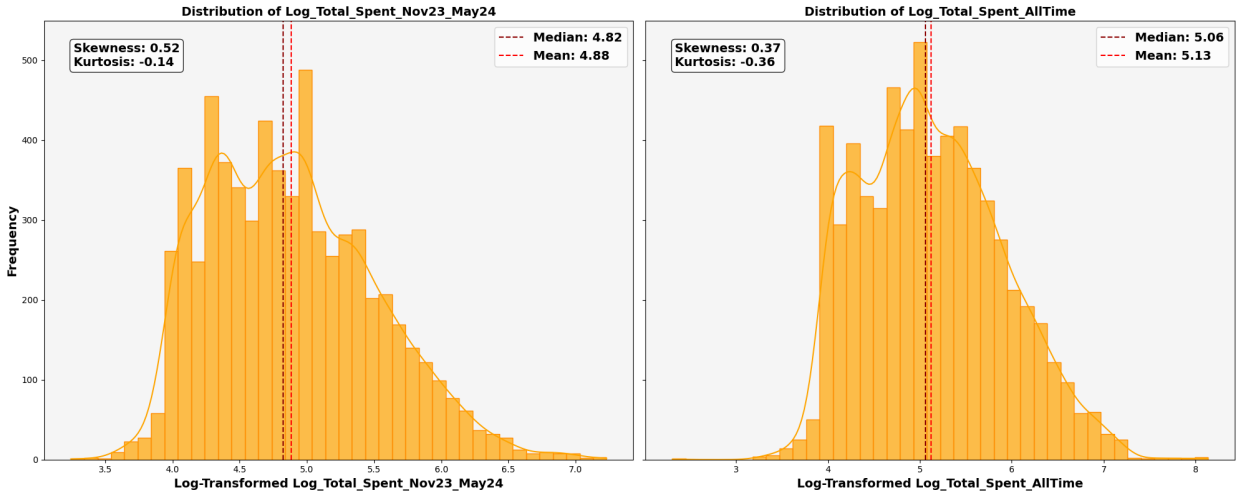


KEY INSIGHTS:

- A significant boosting effect can be seen on the effect of number of face care purchases on spending, given if either a face serum or anti-ageing product is purchased.
- **MLR Analysis:** For each % increase in face care products, patsumer spending increases by 5.16% if they purchase a Face Serum, and 4.47% if they purchase an Anti-Ageing product.

LOG TRANSFORMATION

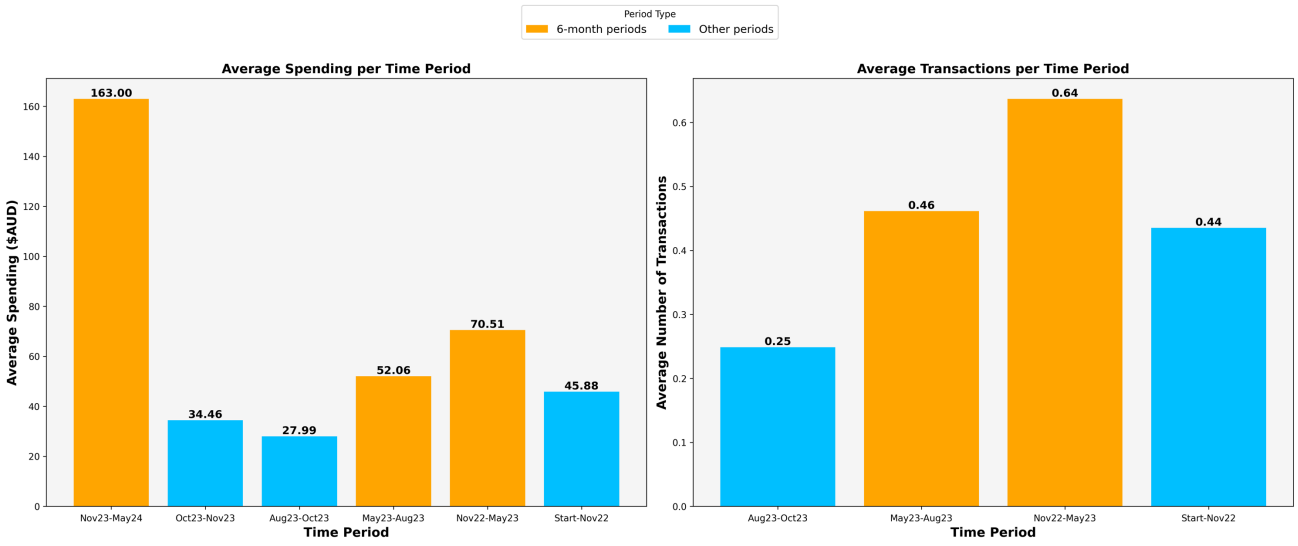
Variable Distribution



**Note: A logarithmic transformation (log(x + 0.1)) was applied to to avoid issues with zero values and reduce skewness. Transformed values represent the log of the original spending amount in dollars.*

TIME INTERVALS

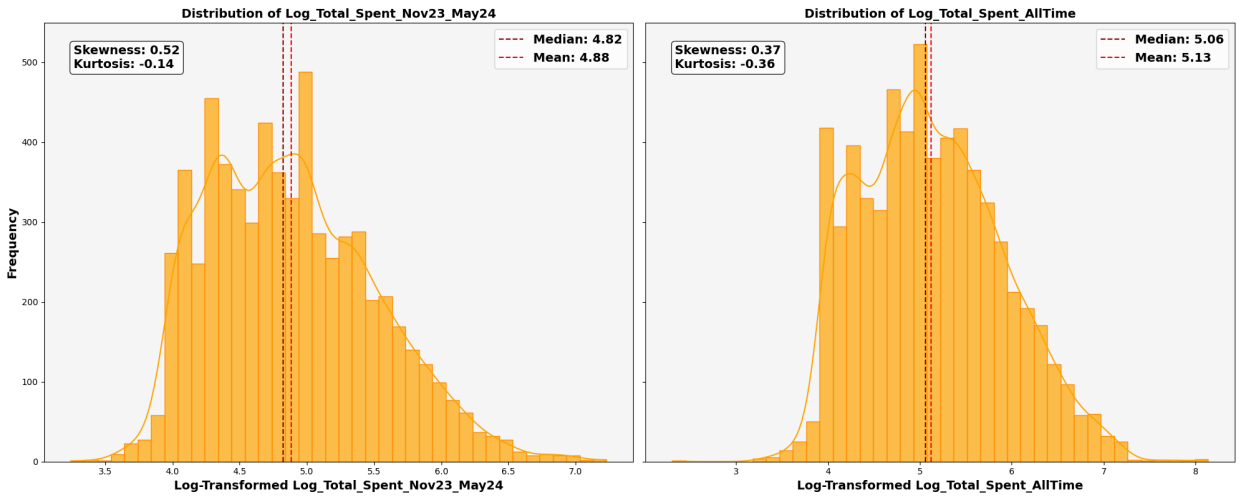
Differenced



**Note: These graphs show the average spending and transaction count for each time period. The periods are non-cumulative, representing distinct time ranges. The 'Start-Nov22' period covers all data before November 2022.*

Time IN

Variable Distribution

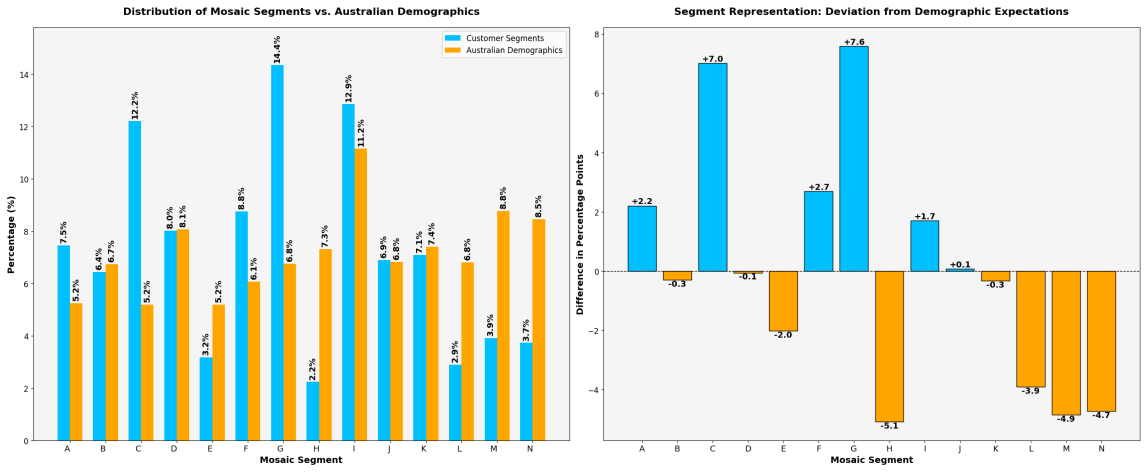


**Note: A logarithmic transformation (log(x + 0.1)) was applied to to avoid issues with zero values and reduce skewness. Transformed values represent the log of the original spending amount in dollars.*

DATA MERGING

Mosaic

mosaic
SMART DATA

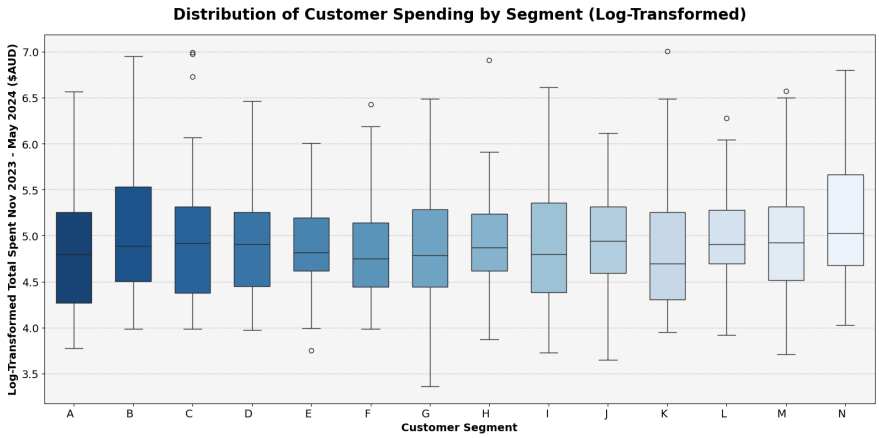


*Note: Positive values indicate over-representation, negative values indicate under-representation in customer segments compared to demographics.

DATA MERGING

Mosaic

mosaic SMART DATA



*Note: Spending values are transformed using log(x + 0.1) Boxes represent the interquartile range (IQR) with the median shown as a horizontal line. Whiskers extend to 1.5 * IQR. Points beyond whiskers are potential outliers.

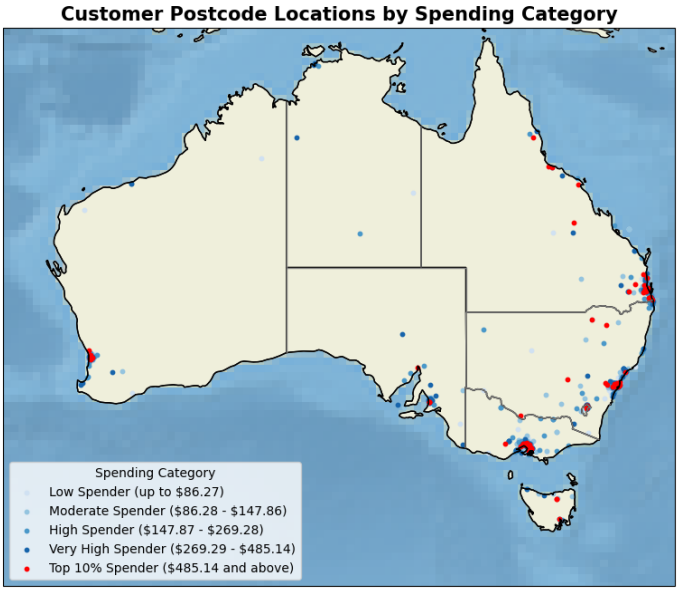
DATA MERGING

Postcode

Dominant Segment
The dominant segment group of the postcode in which a patsumer is located

Electorate Rating
Federal Government Demographic Rating

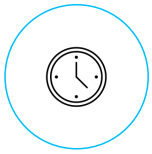
SA4 Name
Name of the city in which the postcode resides in



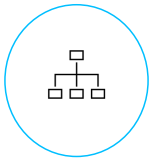
VARIABLE SELECTION

Excluded Variables	Reason For Exclusion
Class Descriptions	Highly ambiguous grouping with potential significant overlap between products.
Postcodes	Not intended for numerical use in analysis

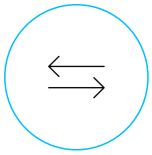
Variable Controls



Temporary Aggregation Controls
Restrict inclusion to prevent multicollinearity



Hierarchy Integrity
Strictly Adhere to the hierarchy principle.



Transformation Exclusivity
Prevent redundancy and maintain clear interpretation:

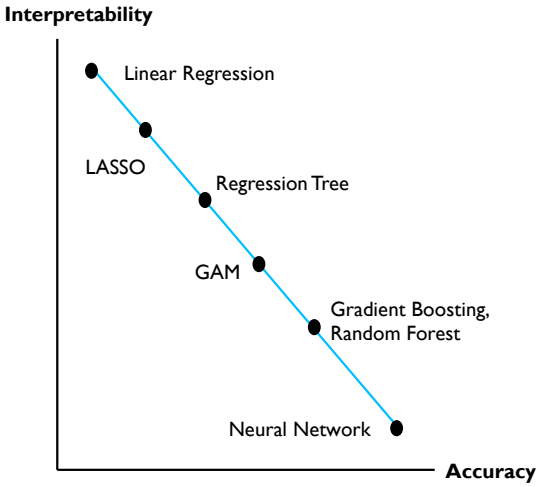
Granularity Consistency
Avoid nested hierarchical data to prevent multicollinearity

INITIAL MODELLING

Trade Off

Accuracy
More complex models can produce more accurate predictions

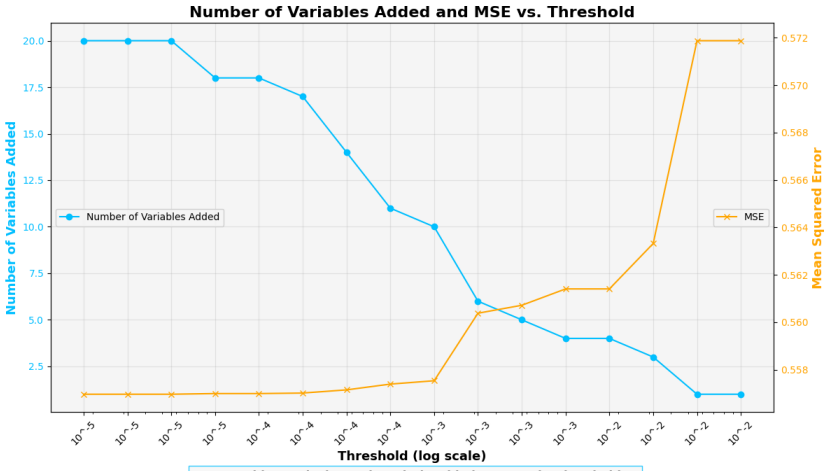
Interpretability
Less complex models can be more understandable, explainable and attributable



FEATURE SELECTION PROCESS

Forward Selection

- We employed a forward selection algorithm to find at least a **locally optimum** feature set
- This terminated after a RMSE reduction of less than 10^{-3} which is the elbow of the plot to the right.

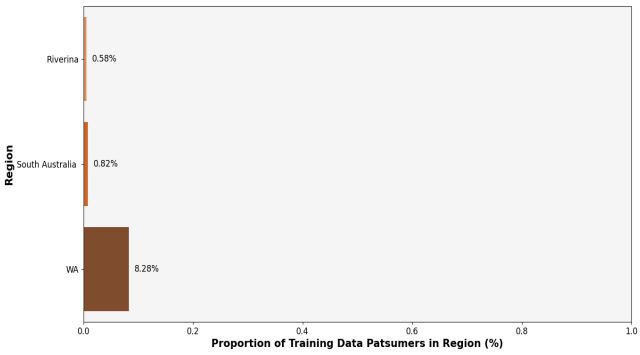


INITIAL GEOGRAPHICAL MODEL

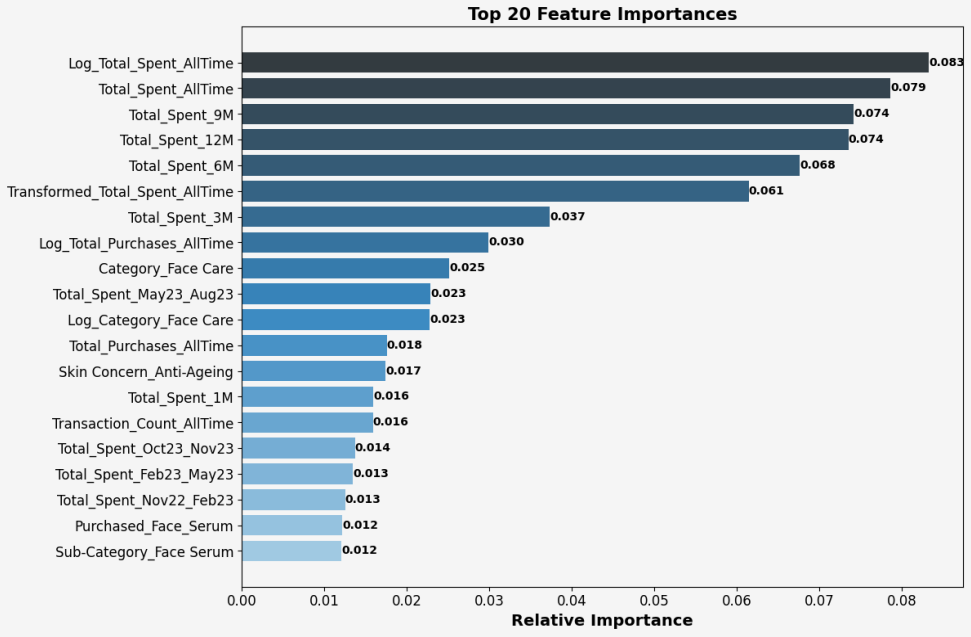
Multi-Stage Model

- We initially tested a multistage linear model which performed separated feature selection for people with and without postcodes
- This appeared to overfit heavily and potentially introduce discriminatory measures hence we decided to **exclude all geographic information**

Feature	Coefficient
Intercept	3.385
Log_Total_Spent_AllTime	0.310
bool_Anthelios	-0.153
sa4name_Riverina	-0.584
Brand Description_Effaclar	-0.0261
Sub-Category_Body Wash	0.131
state_WA	0.186
Brand Description_Hyalu B5	0.069



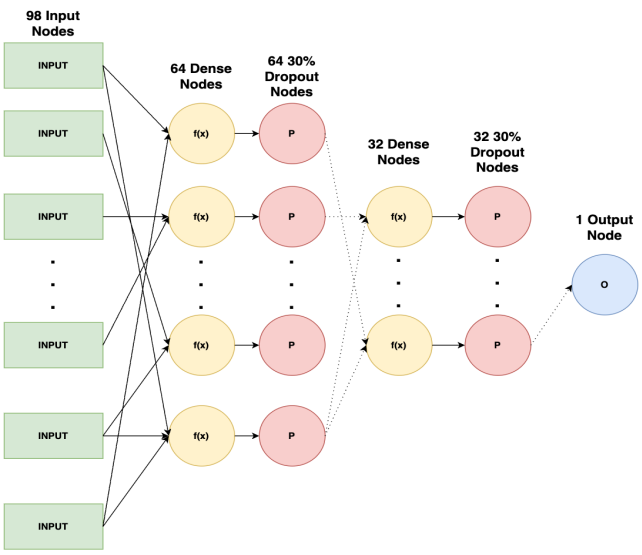
Random Forest

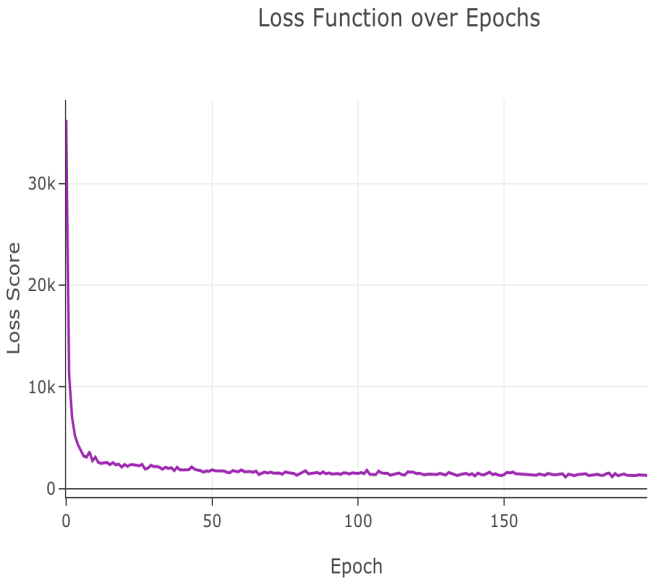


Using top 10
RMSE = 0.560

INITIAL MODELLING

Neural Network

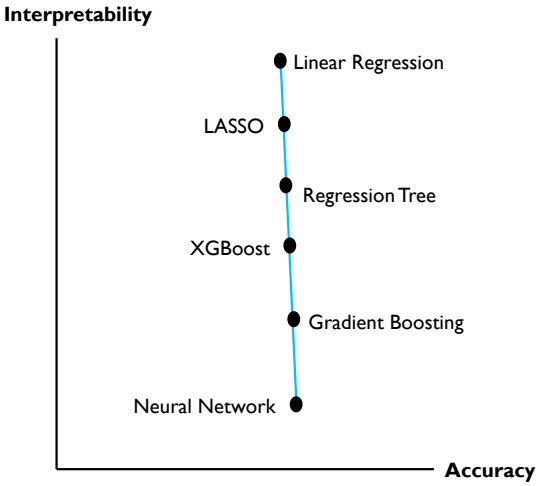




INITIAL MODELLING

Trade Off Findings

Model Type	5 Fold Cross-Validated RMSE
Linear Regression	0.572
LASSO	0.564
Ridge	0.571
GAM	0.622
K-Nearest Neighbours	0.593
Random Forest	0.560
Neural Network	0.79



FINAL MODEL

Multi-Stage Linear Regression

- Model for Top 50% of spenders:

Variable	Coefficient
Constant	4.83
Total Spent 12M	0.01
Purchased Anti-ageing	0.09
Category Sun Care	-0.03
Bool Toleriane	0.07
Bool Hyalu B5	0.09
Has Purchased	-0.12
EAN_EffaclarMoisturiser40ml	-0.03

- Model for Bottom 50% of spenders:

Variable	Coefficient
Constant	4.40
Total Spent All Time	0.00
Category Sun Care	-0.07
Sub-Category Face Moisturiser	-0.05
Bool Eau Thermale	-0.18
Bool Niacinamide	0.11
Purchased Anti-ageing	0.12
Bool Retinol LRP	-0.09
Has Postcode	0.05
Sub-Category Eye Cream	-0.09
Skin Concern Irritation-Prone Skin	0.02

MODEL EVALUATION – MODEL REVIEW

Model	Log RMSE	Dollar RMSE	Interpretability
Base Linear Regression	0.56	118.54	High
LASSO	0.56	118.50	High
Gradient Booster	0.58	119.85	Low
Random Forest	0.56	118.89	Low
Neural Network	0.36	123.68	Low
Mutli-Stage Linear Regression	0.55	117.94	High

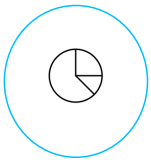
COMPARING SOLUTIONS:

- While the **Neural Network** had the lowest train RMSE, all models produced similarly low scores.
- Multi-Stage Linear Regression produced the lowest Dollar RMSE.
- Given the similarities between the different models, the team choose the model with the highest degree of interpretability, the **Multi-Stage Linear Regression**.

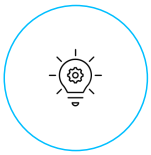
MODEL EVALUATION – FINAL MODEL

POSITIVE ASPECTS OF THE FINAL MODEL

Low RMSE
114.79 RMSE demonstrated strong accuracy.



Flexibility
Adapts well to different consumer segments.



Interpretability
Individual influence on predictions clear.

Actionable Insights
Feature selection highlights specific categories.

FINAL MODEL VALIDATION

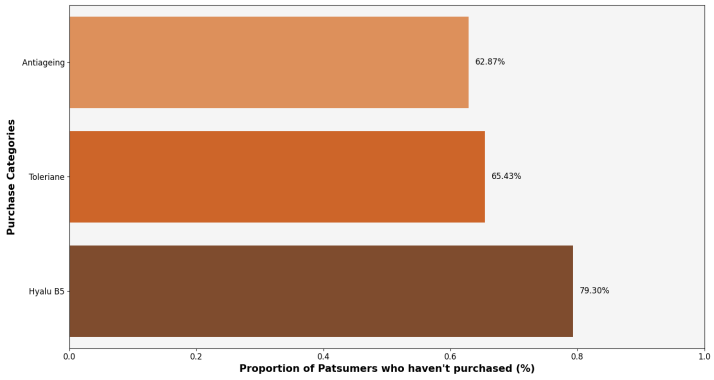
Metric	Low Spenders	High Spenders	Overall
RMSE	88.72	138.38	114.79

MODEL DEPLOYMENT PROJECT

SkinCare Quiz

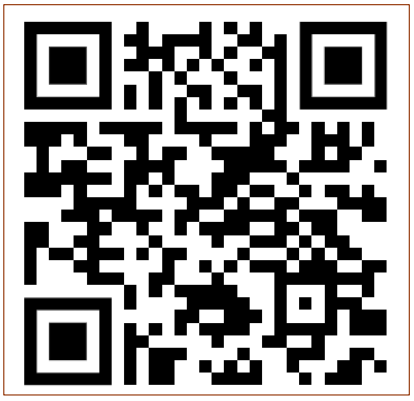
- Based off our linear regression for the top 50% of spenders there were 3 variables which stood out.
- Direct link to customer spend: if a customer first buys a Hyalu B5 product there is a 7.68% increase in their spend.
- Room for growth: 79.3% of training dataset hasn't purchased Hyalu B5.
- We propose an email-distributed personalised skincare quiz which gives tailored recommendations targeted at Anti-Ageing, Toleraine and Hyalu B5 products.
- Potential scope for discounts to incentivise.

Variable	Coefficient
Purchased Anti-ageing	0.0702
Bool Toleriane	0.0506
Bool Hylau B5	0.0768



MODEL DEPLOYMENT PROJECT

Skincare Quiz



MODEL DEPLOYMENT SUB-PROJECTS

Email Timing
- $19,400 Increase in revenue

Using Quiz Data
- $29,500 Increase in revenue

User Feedback
- $31,500 increase in revenue

Limitations and Drawbacks
- Concerns with privacy
- Potential to drive customers away