A Machine Learning Approach in Classifing Consumer Behaviour for La Roche-Posay

Shuyu Bai, Chengjie Deng, Zhecheng Zhang, Bill Nguyen and Kess Ngo

University of Sydney

Abstract

This report primarily focuses on analysing the online sales data for La Roche-Posay through advanced machine-learning models. The report starts with describing and understanding the business problem and relates it to relevant machine learning concepts. It then summarised several significant exploratory data analysis (EDA) findings, which assist in formulating relevant business strategies. It also gives a strong focus on developing relevant machine learning models such as logistic regression, random forest, neuron network and XGBoost model. By applying these tools, the underlying consumer behaviour pattern could be better captured and utilised in proposing relevant strategic business projects aiming at increasing consumer satisfaction and retention rates while enhancing sales performance. Additionally, the report introduces a marketing concept, i.e., the RFM framework, in clustering customer base according to their purchase behaviour. The report ends with demonstrating two relevant strategic business projects, with corresponding analysis in quantifying the project's costs and revenues.

Key words: Classification Models; Predictive Analytics; Customer Segmentation; Model Optimization; Consumer Behavior Analysis; Personalized Marketing Strategies

1. Problem description

La Roche-Posay, the leading subsidiary of L'Oréal, is currently facing a significant business transformation targeted at enhancing its online sales performance. The primary objective for this transformation is to increase consumer satisfaction and retention rates while enhancing sales performance. In this project, patsumers[1]' data will be analysed to detect and comprehend their purchase behaviour and preferences.

The Australian skincare market is highly competitive, with CeraVe, La Roche-Posay, and Garnier as the leading brands (Statista, 2024). La Roche-Posay, in particular, has established a concrete client base among female consumers in Australia. To sustain its market leadership, it is essential for the brand to continuously adapt to shifting consumer preferences and digital innovations, which will be key to maintaining its competitive edge (Statista, 2024).

Hence, this report aims to understand "patsumers" purchasing patterns and discover crucial factors that play significant roles in influencing and classifying which brand the customers would purchase. More importantly, the project will use advanced machine learning models to classify consumers' behaviour and predict the probability for each consumer to purchase a particular brand, represented by a binary value (with 0 indicating not purchase and 1 suggesting purchase). Statistically speaking, this project could be formulated as a classification problem, and several machine learning models, including logistic regression, random forest, XGBoost, and neuron networks will be developed. Ultimately, effective and customised marketing strategies will be developed through interpreting the model results and combining other relevant qualitative research.

Additionally, the RFM framework is referenced in this project for clustering consumer segmentations. According to Blattberg, Kim, and Neslin (2008), the RFM framework leverages three key behavioural dimensions in understanding consumer behaviour,

[1] Patsumers: customers with specific skin care issue/concern.

Table 1. Summary Statistics of Continuous Variables

	Total_ Spent_ 1M	Total_ Spent_ 3M	Total_ Spent_ 6M	Total_ Spent_ 9M	Total_ Spent_ 12M	Total_ Spent_ AllTime
Mean	34.36	62.45	114.51	156	185.01	230.9
Median	0	0	84.72	115.1	135.47	157.59
Std	69.57	94.73	135.9	166.67	187.46	224.22
Skewness	2.79	2.36	2.68	3.18	3.45	3.62
Kurtosis	12.3	9.39	15.09	23.49	28.47	26.87

including recency, frequency, and monetary value, making this model a simple and comprehensive tool for segmenting and capturing consumer behaviour. The rest of the report will apply this framework to cluster all consumers according to these three mentioned dimensions, and corresponding marketing strategies will be recommended for each consumer segment.

2. Summary of potential insights from EDA

This section will summarise insights obtained from the individual exploratory data analysis (EDA) project, which mainly consists of 1. Dataset properties and characteristics; 2. Pattern & Relationships; 3. Insights towards models & links to business strategies; and 4. RFM model analysis.

2.1. Dataset Properties & Characteristics

The dataset contains online sales information for La Roche Posay before May 2024. Comprehensive information regarding total spending and transaction counts across every 3 months for each customer, the total number of transactions for each product, and whether a consumer has purchased a particular brand (represented by binary value), are provided in the dataset. This subsection will illustrate the distribution, relevant statistics, and major properties of continuous and discrete variables, respectively.

2.1.1. Continuous variables

All continuous variables are rather skewed in this dataset, with skewness and kurtosis ranging from 2.36 to 3.62 and from 9.39 to 28.47, respectively (Table 1). Over time, as the period extends further from Nov 23, the number of 0 values in total spending decreases, indicating that more customers are making purchases. Additionally, the standard deviation of these continuous variables increased over time, suggesting that some customers were increasing their spending while others remained inactive.

Due to a large proportion of zero values existing in those continuous variables, the corresponding non-zero portion is also analysed. As the period extended, the total consumer spending increased gradually. The raised skewness also indicates higher spending volatility while increased kurtosis represents occasional higher spenders (Table 2).

2.1.2. Discrete Variables

All discrete variables are very skewed, which mainly represents the number of times a consumer purchases a specific product. Also, some discrete variables, such as "*Transaction_Count_1M*", have low cardinality, with only six unique values, while others, including "*Skin Concern_Anti-Ageing*" have even more than 15 categories. Proper encoding methods will be applied for those discrete variables with corresponding justification in the following section. However, it is noticed that there exists a large proportion of zero values in many discrete variables (even greater than 80%), proper feature engineering approaches will be applied and explained in the following section.

2.2. Patterns & Relationships

2.2.1. Consumer cyclical patterns

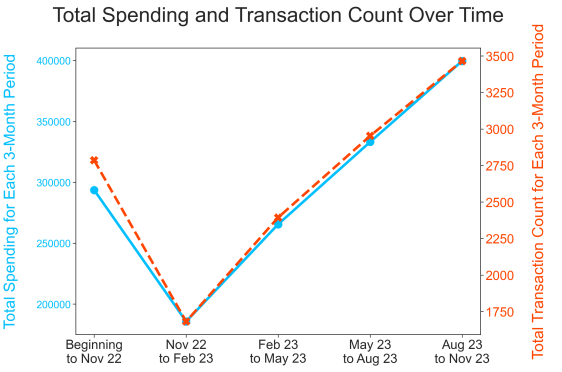


Fig. 1. Consumer cyclical patterns

We observed a clear consumer cyclical pattern in both total spending and transaction count across all three-month periods, both reached their lowest point on Nov 22 and started to recover gradually from Feb 2023. This can be attributed to the end of the discount season, leading to temporary consumption fatigue. Also, skincare products generally possess a repurchase

Table 2. Summary Statistics of Continuous Variables (YJ Transformation)

	Total_ Spent_ 1M_ YJ	Total_ Spent_ 3M_ YJ	Total_ Spent_ 6M_ YJ	Total_ Spent_ 9M_ YJ	Total_ Spent_ 12M_ YJ
Non-zero samples	1710	2813	4317	5201	5714
Mean	128.96	142.08	169.76	191.96	207.22
Median	110.56	115.12	132.64	143.8	145.96
Std	77	95.4	134.16	165.16	186.43
Skewness	2.42	2.52	3.12	3.55	3.69
Kurtosis	11.89	10.97	18.8	27.25	31.08

cycle, which observing a cyclical pattern in customer behaviour is normal.

2.2.2. Multi-collinearity

Total spending and transaction count related variables all present a strong multi-collinearity effect in our dataset, as shown in the correlation matrix below (Figure 2, Figure 3). This is likely because total spending and transaction count in the later period would include earlier information, forcing these variables to be non-independent. Albeit multi-collinearity would not be a concern in these non-linear models such as random forest and XGBoost, appropriate feature engineering techniques will be implemented in the next section to better capture consumer behaviour across each period.

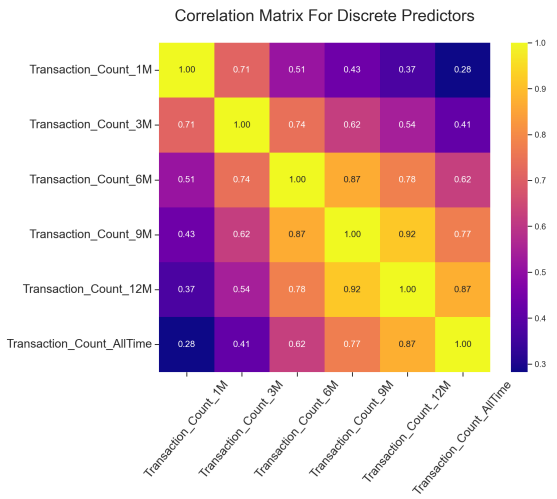


Fig. 2. Correlation Matrix for Transaction

2.2.3. Product Property Discovery - Mean Test Result Revisit

The Tukey HSD test assesses whether an increase in the number of purchases of a product will also lead to an increase in that consumer's total spending. For

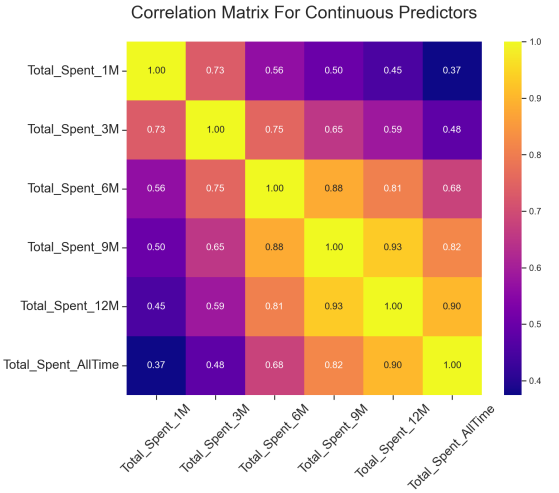


Fig. 3. Correlation Matrix for Spending

products such as body moisturisers, where most null hypotheses of Tukey HSD test results are rejected, the average total spending does not deviate from product consumption (Figure 4). This indicates these products might have many substitutes in the market. However, products like Vitamin C and eye creams, on the other hand, display a clear trend of higher average spending with increased customer purchases (Figure 5, Figure 6). Hence, La Roche-Posay could offer promotions for these products to simulate consumers' purchases while boosting revenue from these high-potential products.

2.2.4. Correlation between products & Association rule mining

A high correlation between two products indicates they are often purchased by the same consumers or serve complementary needs. As illustrated in Table 3, products related to similar skin concerns or brands have very high correlations, meaning customers interested in a particular skin concern tend to purchase its associated brand. Additionally, association rule mining is applied to explore which brands or products are frequently

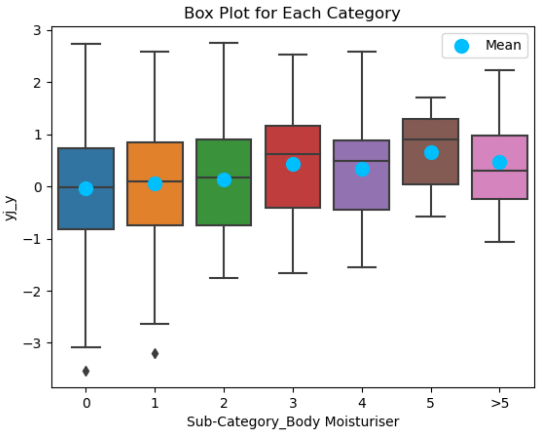


Fig. 4. Boxplot for Body Moisturiser

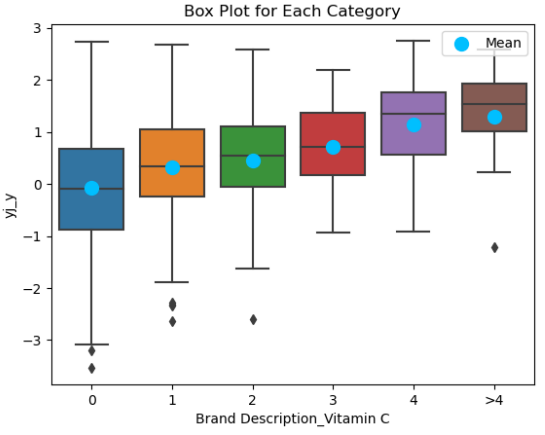


Fig. 5. Boxplot for Vitamin C

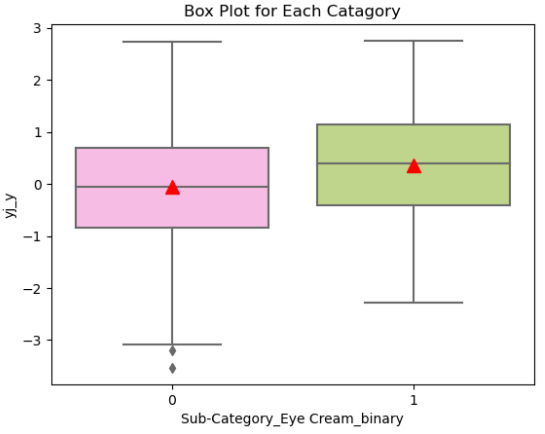


Fig. 6. Boxplot for Eye Cream

purchased together and how products from different brands are grouped in customers' baskets.
According to Table 4, about 10% of transactions include both Cicaplast and Toleriane, and there is 48% of customers who purchase Cicaplast also buy Toleriane. Hence, these two brands should be stocked together to cater to customer demand.

2.3. Insights towards the modelling section & Links to business strategy

2.3.1. Brand popularity

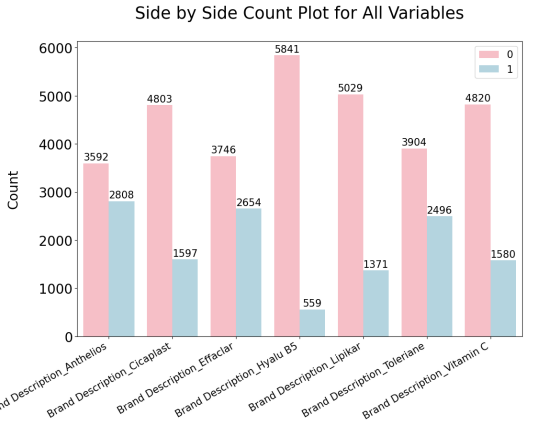


Fig. 7. Class Imbalance for Classification Label

According to the side-by-side count plot, there presents varying degrees of class imbalance among different categories (Figure 7). Some variables, like "Brand Description_ Anthelios_ Output", demonstrate relatively balanced classes, while other variables, such as "Brand Description_ Hyalu B5_ Output", are denominated by a single class. This imbalanced class existed in the dataset indicates that appropriate weight adjustment needs to be implemented to enhance the model performance.

2.3.2. Cross-selling opportunities

Based on all analysis mentioned above, brands "Cicaplast" and "Toleriane" are always purchased together, showing that sensitive skin consumers seek skin repair products (Table 4). Additionally, bundling face moisturizers and serums could further boost brand loyalty (Table 4). Conversely, products such as body moisturizers present a weak correlation with increased total spending, possibly hinting at high market substitutability (Figure 4).

Table 3. Correlation Between Products

Pair	Skin Concern	Product with Highest Corrl	Brand with Highest Corrl
0	Acne-Prone Skin	Face Moisturiser (0.61)	Effaclar (0.96)
1	Anti-Ageing	Face Serum (0.73)	Hyalu B5 (0.70)
2	Irritation-Prone Skin	Face Moisturiser (0.6)	Toleriane (0.85)
3	Pigmentation and Dark Spots	Face Serum (0.49)	Niacinamide (1.00)
4	Sun Protection	Sunscreen (0.73)	Anthelios (0.91)

Table 4. Association Rule Mining

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
Brand-Cicaplast	Brand-Toleriane	0.21	0.35	0.10	0.48	1.38
Face Serum	Face Moisturiser	0.43	0.51	0.25	0.57	1.13
Face Moisturiser	Face Serum	0.51	0.43	0.25	0.49	1.13

Table 5. RFM Segmentation

Segment	Recency Score	Frequency Score	Monetary Score	Criteria Summary	Count
Soulmates	Any	≥ 2	≥ 3	Loyal and frequent spenders	1054
Ex-Lovers	≤ 5	≥ 2	Any	Recently disengaged, used to be active	947
Potential Lovers	Any	≤ 2	High (RFM ≥ 7)	Low frequency, high potential	1987
Other	Any	Any	Any	Doesn't meet other segment criteria	2412

2.4. RFM framework analysis

As mentioned previously, the entire consumer base will be segmented into three parts, including soulmates, potential lovers, and ex-lovers, based on R (recency), F(frequency), and M (monetary value) scores, below are explanations for these three measurement benchmarks:

- Recency: How recently the customer made a purchase (Depending on the transaction count in each period)
- Frequency: How often the customer made a purchase (Depending on the log-transaction count in each period)
- Monetary: Value of the customer's purchase (Depending on the total amount spent for each customer)

By applying the clustering benchmarks mentioned above, the corresponding consumer segments have been filtered out, below are the customer profiles for each segmentation.

- Soulmates: The most loyal clients for our brand, with the highest total spend and transaction frequency.
- Ex-lovers: This group of clients also spend a relatively large amount on our products; however, their transaction frequency is relatively low compared to soulmates.
- Potential Lovers: This group of customers spend relatively frequently, however, the total amount spent for this group is slightly lower compared to soulmates and ex-lovers.

3. Machine Learning Section

3.1. Data Pre-processing

In this section, basic operations of the raw dataset, including merging relevant columns and dropping unnecessary columns, deleting columns containing a large proportion of missing values, and managing outliers, will be conducted.

Merge relevant columns and drop unnecessary columns: Columns 'Category_ Face Care' and 'Category_Face Care ' are merged due to their same meaning. Columns 'Has_Transaction_Nov23_May24', 'CustomerID', 'Unnamed: 0' are excluded from our analysis as instructed in the data dictionary.

Handling columns with missing values: After iterating through all columns, we noticed that the

variable "Post Code" contains 5326 missing values which consists of 83% of its total observations. Hence, this column is deleted from the dataset since a large proportion of missing values makes it impractical to apply any imputation or filling methods, which would potentially lead to misleading modelling results.

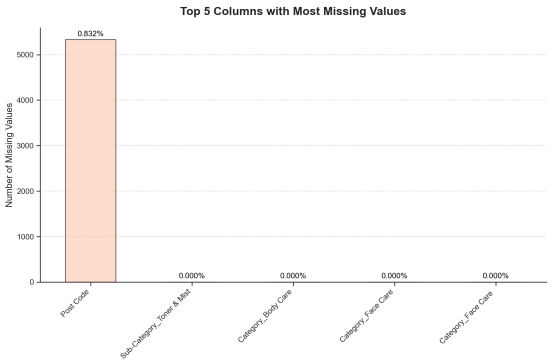


Fig. 8. Missing Value Visualization

Grouping variables: All attributes in our dataset are classified as discrete, continuous, and binary. Variables with a "float" data type, such as "Total_Spent_1M", are considered continuous variables since they represent consumers' spending nature which can be interpreted with a continuous nature. For the rest of the variables, which mainly illustrate count-based values, we noticed that many of them contain a rather larger proportion of 0 values (even greater than 80%). We classify these variables into binary variables since most of the consumers do not purchase them during the data-collection period. Variables that are not strictly continuous or binary are considered discrete variables.

Managing outliers: We noticed a right-skew in the distribution of discrete variables, indicating potential outliers. To address this, we grouped values with fewer than 10 occurrences into a "more" category to prevent the model from capturing infrequent trends, effectively reducing the risk of overfitting.

3.2. Feature Engineering

The feature engineering processes involved exploring variable distributions and detecting relationships to better comprehend the dataset. Different encoding methods for discrete and binary variables are applied and new features for time-series data, such as total transaction counts and spending over various periods, are created to better understand consumers' purchasing behaviour across different time periods.

Variable pattern discovering: We noticed the continuous variables are highly skewed with a significant right tail and multicollinearity (Figure 9; Table 6). Since the logistic regression model is rather sensitive to variable scale, we applied YJ transformation for normalising these variables and this could potentially enhance model performance. However, models including random forest and XGBoost are less sensitive to variable scale and multicollinearity, so we used the original data as inputs to better capture the true underlying trend. After YJ transformation, distributions for all continuous variables better align normality (Figure 10).

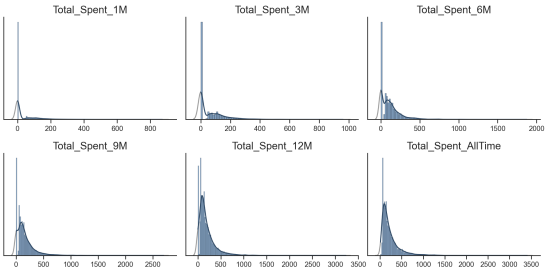


Fig. 9. Distribution of Total Spending

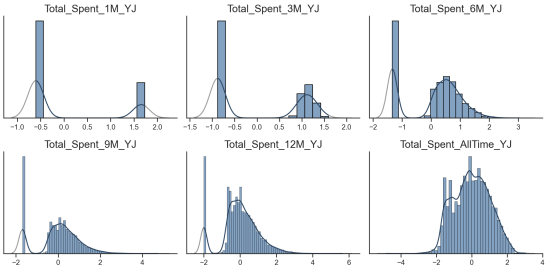


Fig. 10. Distribution of Total Spending (YJ Transformation)

Table 6. VIF for Spending Variables

Variable Name	VIF
Total_Spent_1M	2.69
Total_Spent_3M	4.91
Total_Spent_6M	10.38
Total_Spent_9M	23.34
Total_Spent_12M	28.40
Total_Spent_AllTime	11.03

Encoding discrete variables: Ordinal encoding is applied for the pre-processed discrete variables. Since all machine learning models cannot process string inputs, ordinal encoding could convert them to numerical format. Also, this encoding method is

appropriate as variables have a meaningful order due to their count-based nature.

Encoding binary variables: We conducted dummy encoding for binary variables, since it effectively captures their two-category nature without imposing any order. Additionally, it does not significantly inflate the dataset or introduce unnecessary complexity since it only creates 0 and 1 values, keeping the model efficient and easy to interpret.

Creating new features: Multicollinearity effect is identified among variables, such as total spendings and transaction counts across different 3-month periods. Although most models, except the logistic regression, are not sensitive to multicollinearity, including these variables could increase unnecessary model complexity. Hence, we created new variables

To address this, we created new variables by subtracting the values from earlier periods from the later ones to calculate the net total spending and transaction count for each 3-month period. This approach not only resolves the multicollinearity issue but also allows the model to better capture customer purchasing patterns within each time period.

3.3. Machine Learning Model Development

3.3.1. Machine-learning models overview & optimisation matrix

There are four models built in this project, namely logistic regression, Neural Networks, Random Forest, and XGBoost. Logistic regression model is selected as the reliable benchmark model due to its interpretability and low computational cost, making it a reliable starting point for comparison with other complex non-linear models like XGBoost and random forest.

During the model development process, when building the binary classification model, we focused on optimizing the group 1 F1 score, which represents customers who are likely to purchase. This choice is driven by our business objectives of better detecting those high-value customers, rather than predicting clients who are unlikely to purchase our products. However, when fitting the multi-label classification models, our primary objective is to optimize the f1_macro score. This metric ensures balanced consideration of each label's F1 score, ensuring the models also perform reasonably well for labels with fewer samples including Hyalu B5 and Lipikar (Figure 7). Additionally, a five-fold cross-validation is applied to ensure that the model performs well not only on the training data but also generalised well.

Accuracy is not chosen as the optimising benchmark since there exists a large number of zero values in the dataset, which optimising accuracy would train the model predicting more zero values and creating a false sense of higher accuracy while performing poorly in classifying actual buyers. Additionally, the AUC (area under the curve) is utilised to further evaluate the model performance and capability in distinguishing between positive and negative samples, even though this is not the primary optimising metric.

Therefore, our main goal remained to maximise the F1 score for group 1 to ensure models effectively classify customers with purchase behaviour, aligning with our primary business objective.

3.3.2. Modelling workflow

In the following model development process, we applied both one-to-one and one-to-many classification approaches to comprehend and classify consumer behaviours. One-to-one models involve developing a corresponding binary classification machine learning model for each output variable, which could potentially enhance model accuracy by inputting more resources in capturing underlying trends for each classification label. On the other hand, the one-to-many model applies multi-label classification techniques, which predicts all seven output labels using a single model. This ensures higher modelling efficiency and reduces computational costs. By comparing model results of these two approaches, we could offer more accurate and commercially valuable results and better understand underlying patterns between independent variables and the response. If purchasing patterns for each product are significantly different from each other, those one-to-one models would outperform those one-to-many models since separate models could better capture distinct brand's patterns. However, if both perform similarly, one-to-many modelling may be preferable due to its higher efficiency and lower maintenance costs in practical business scenarios.

3.3.3. Logistic Regression

3.3.3.1. Model Overview

A popular statistical model for binary or multi-class classification applications is logistic regression. It is especially well-liked because of how easily interpretable it is—the model's coefficients make it clear how each feature and the intended output relate to one another. The model's interpretability makes it an ideal choice for understanding consumer behaviour and product preferences. By interpreting its coefficients, we could comprehend those important features contributing to classifying our response and better capturing consumer behaviours intuitively.

3.3.3.2. Model Development (one-to-one classification)

One important step in building the logistic regression model is hyperparameter tuning. We applied both L1 (Lasso) and L2 (Ridge) regularization here. L2 regularization reduces the coefficients to assist prevent overfitting, while L1 regularization performs feature selection by pushing the coefficients of unimportant characteristics to zero. Additionally, max_iter=1000

was used in the model's configuration to guarantee correct convergence, particularly considering the difficult nature of multi-label classification tasks. Additionally, **GridSearchCV** is also utilized for hyperparameters optimisation, it basically searches through different regularization strengths and penalties to find the best-performing configuration.

3.3.3.3. Model Development(Multi-label classification)
To handle multi-class classification, we used a One-vs-Rest (OVR) approach in combination with logistic regression. This method creates separate binary classifiers for each label, enabling the model to treat each class independently. The hyperparameter tuning was performed using GridSearchCV as well, which allowed us to find the optimal regularization parameters for L1 and L2 penalties, along with different regularization strengths.

3.3.4. Neural Network
3.3.4.1. Model Overview (one-to-one classification)
A one-to-one neural network model for each label is developed to predict consumer purchases. Neural networks, commonly used for non-linear models for classification and regression tasks, consist of input, hidden and output layers. These networks process data through interconnected layers using weights and activation functions. The ReLU (Rectified Linear Unit) activation function is used for the input and hidden layers. It effectively addresses the vanishing gradient problem and improves training speed by maintaining positive values unchanged and setting negative values to zero. Additionally, the sigmoid activation function is utilized for the output layer, which maps the output to a value between 0 and 1 and effectively generates the probability of a purchase (1) and not purchase (0), making it an appropriate choice for this classification task.

3.3.4.2. Model Setup
The following section will illustrate how the neuron network model is built. The model development process starts with a customized function to optimise the F1 score for group 1 (consumers have a purchase history). This function ensures the model focuses on those customers with a purchase history during training and tuning, rather than optimizing the F1 score for the entire dataset.

3.3.4.3. Managing the class imbalance
Since there exists a significant class imbalance (i.e., a large proportion of zero values with each group), the class_weight package is referenced to calculate weights for each class to adjust the model's focus. This ensures that the model could give more focus on the minority class (group 1 - customers with purchase behaviour) during the model training process, rather than over-predicting the majority (group 0 - customers without purchase behaviour). This approach could improve the overall performance of the model.

3.3.4.4. Neural network development
A function is defined for tailoring the best predictive neural network model for classifying each label. This function optimizes the network's performance through hyperparameter tuning. Specifically, the following hyperparameters have been tuned:
Number of hidden layers and neurons: This step is to ensure our model effectively balances computational complexity and computational resources, while capturing nonlinear features underlying the data. To achieve these goals, the maximum number of hidden layers is set to be 3, with neurons ranging from 64 to 256 per layer in steps of 32.
Dropout rate: To avoid overfitting, a dropout rate of up to 30% is introduced to randomly "drop" 0% - 30% neurons in each iteration. This step could effectively improve the model's generalisation by reducing reliance on noises in the training data.
Learning rate: A proper range of learning rate is defined as 1e-4, 1e-3, and 1e-2. In particular, a smaller learning rate (e.g., 1e-4) allows for more cautious weight updates, which is suitable for complex data. However, larger learning rates such as 1e-2 enable faster updates, which is ideal for simpler data or early training stages.
Batch size: Different batch sizes (8, 16, 32, 64) are tuned in our model. Generally, a smaller batch size enables the model to better adapt to complex decision boundaries but it is less efficient. However, larger batch sizes are more efficient but may miss subtle data differences. Tuning this hyperparameter helps retrieve the best balance between training efficiency and model performance.

3.3.4.5. Random Search and hyper-parameter optimisation

After building the basic architecture and specifying the hyper-parameters tuned in the neural network model, a random search method is referenced for hyperparameter tuning. Compared to other methods like grid search, the random search approach randomly selects several hyperparameter combinations, allowing tuning hyperparameters under relatively limited computational resources.
Under each training iteration, different combinations of hyperparameters are tried, aiming to maximise the F1 score for group 1. This approach helped to find the optimal neural network architecture for each product, improving model performance in real business scenarios.

3.3.4.6. Neural network model for multi-label classification

We further build a multi-label classification neural network model. The basic model architecture is roughly the same as the one-to-one neural network model.

However, the design of the output layer and the loss function is different, specifically:

Output Layer Design: In the design of the output layer for the multi-label classification model, it functions as a multi-label classifier where each node corresponds to an individual product. These nodes output binary predictions indicating whether a consumer will purchase each specific product. The sigmoid function is applied at each node to predict the probability of purchase, ensuring that the output is appropriately scaled between 0 and 1.

Loss Function (weighted binary cross-entropy): For the one-to-one neural network models, the loss function employed is binary cross-entropy, suitable for binary classification tasks. However, for multi-label classification scenarios where the purchase behaviour for each product is independent, a custom weighted binary cross-entropy loss is designed. This loss function effectively addresses class imbalance prevalent in scenarios where most customers do not purchase the products, enhancing the model's sensitivity towards predicting the minority class of purchasers.

L2 Regularization: To prevent our model from overfitting, the L2 regularisation is applied to the hidden layer weights. This approach could also limit model complexity and improve the model's generalization ability.

3.3.5. Random Forest

3.3.5.1. Model overview

The random forest model is a good fit for classifying consumer behaviour, especially given the imbalanced nature of our dataset, which contains a lot of zero values. The random forest model could effectively handle this imbalance by building multiple decision trees, reducing overfitting, enhancing the model's generalization, and improving prediction accuracy. It does well in managing categorical and continuous variables, while it is also resilient to outliers and multicollinearity. Another significant advantage of applying the random forest model is due to its built-in feature importance analysis. By analyzing the contribution of each feature to the prediction, insights regarding product importance could be concluded, which provides valuable guidance for business decision-making.

3.3.5.2. Model development

Given the significant class imbalance within the dataset, the SMOTE (Borderline Synthetic Minority Over-sampling Technique) technique is applied prior to model training to oversample the training set. This approach synthesizes new minority class samples, particularly targeting those near the decision boundary, to enhance the model's ability to balance the distribution of classes. By doing so, SMOTE allows the model to learn more relevant features from the minority class, enhancing the overall prediction accuracy.

Optuna is utilised for hyperparameter tuning to optimize the performance of the random forest model. Specifically, Optuna is an efficient automated framework that uses Bayesian optimization to find the optimal hyperparameter combination. Key hyperparameters tuned for the random forest model are as follows:

n_estimators (number of trees): This hyper-parameter determines the number of decision trees fitted in the random forest model. Generally, having more trees in the model would improve the model's stability and generalization, whereas this also requires higher computational costs. Hence, a range of 100-1,000 trees is set to balance the model's efficiency and performance.

max_depth (maximum depth of each tree): This hyper-parameter limits the complexity of each tree within the model. Trees that are too deep might lead to model overfit, while shallow trees might fail to capture important patterns underlying the data. Hence, a range from 3 to 20 is set here so the model could utilise both simple and deeper trees, allowing identifying complex features and interactions.

min_samples_split (minimum samples required to split a node): This hyper-parameter controls the minimum number of samples needed to split a node. Smaller values of sample split generally increase model complexity, while a larger value helps prevent overfitting. A range of 2 to 10 for this hyper-parameter could effectively ensure the model maintains moderate complexity without learning from extra noises.

min_samples_leaf (minimum samples required per leaf): This hyper-parameter limits the minimum number of samples required in each leaf node. A higher number of samples could reduce model complexity and improve generalization. Hence, a range of 1 to 10 is defined to ensure each leaf node has enough samples while avoiding overfitting.

max_features (maximum number of features): This hyper-parameter defines the proportion of features used for each split, with a range of 0.1 to 1.0. Lower values for this hyper-parameter could increase tree diversity and reduce variance, while a higher value may increase bias by capturing too many features within the model. The range of 0.1 to 1.0 is defined for this hyper-parameter aiming at balancing the model's variance and bias.

class_weight (class weights): To address data imbalance, class weights are adjusted to force the model to focus more on the minority class (i.e., customers with a purchase history). By assigning higher weights to the minority class, we ensured better prediction performance for that group, which effectively improved performance metrics such as F1 score and AUC for the model.

3.3.5.3. Random Forest Model for multi-label classification

The one-to-many random forest model generally shares a similar basic framework to the one-to-one model. However, OneVsRestClassifier (OVR) is used here to generate the multi-label classification output. It treats each classification label as an independent binary classification while sharing the same base random forest model. Key differences between the one-to-one and one-to-many classification tasks include output layer design, hyperparameter tuning, and evaluation methods. The following part will explain these differences in detail:

Output Layer Design: In the one-to-many architecture of the model, the output layer features multiple nodes, each independently predicting the likelihood of purchasing a specific product. Employing the OVR, this setup constructs a separate binary classification model for each label, with all predictions made using the same underlying base model. The final output is multi-dimensional, where each dimension represents the predictive outcome for an individual product.

Hyperparameter Tuning: Hyperparameters are tuned for the entire model, rather than targeting an individual product. By wrapping a random forest model with OVR, we employ multi-label cross-validation using cross_val_score to identify the optimal parameter settings that best predict multiple product purchases simultaneously. The f1_macro is specifically optimised to evaluate the overall performance across all products. This choice ensures a balanced model performance across all labels and does not neglect the performance of the minority class (those likely to purchase a product).

Evaluation Method: The function evaluate_ multi_ label_ model is designed to calculate performance measurement metrics such as AUC, accuracy, and F1 score for each product label, based on the multi-label classification model. This allows easier comparison of prediction performance across different products.

3.3.6. XGBoost

3.3.6.1. Model Overview

XGBoost is a powerful decision-tree-based ensemble learning algorithm that uses gradient-boosted decision trees to iteratively build trees. Unlike random forest, which builds multiple independent decision trees in parallel and makes predictions through a voting mechanism, XGBoost builds decision trees sequentially, with each tree aimed at correcting the prediction errors of the previous tree. Moreover, its in-built features like regularisation could also prevent overfitting and ensure efficiency on large datasets, and it could also support weighted loss functions, allowing the model to handle class imbalance and focus more on minority class predictions.

3.3.6.2. Model Setup

Similar to the approach in building the Random Forest model, addressing class imbalance is essential before training. Since the majority of class labels are zeroes, indicating no purchase, the model tends to overpredict the 'no purchase' outcome. To address this, SMOTE is implemented again to oversample the training data, which effectively increases the number of positive (purchase) cases. This technique could thereby balance the ratio of positive to negative classes in the training set.

Besides, dynamic class weights (scale_ pos_ weight) are also set to give higher importance to the minority class. This technique ensures the model to focus more on customers with purchasing behaviour, by using Optuna, the dynamic class weights can be automatically adjusted.

Similarly, Optuna is used for hyper-parameter tuning for the XGBoost model; this approach could avoid the high computational cost compared to the traditional grid search methods. Key hyper-parameters tuned for XGBoost model include:

n_estimators: This hyper-parameter represents the number of trees. Here, the range is set to be 100 - 2,000, ensuring the model to balance computational cost and stability.

max_depth: This hyper-parameter controls the maximum depth of each decision tree. Similar to random forest, the range from 3 to 20 is set to balance complexity and prevent overfitting.

learning_rate: This range is set between 0.001 and 0.1, controlling the step size for each boosting iteration. A lower rate (0.001) ensures stability and avoids missing local optima, while a higher rate (0.1) speeds up convergence. This range of learning rate allows effective adjustments for different scenarios of our model.

lambda and alpha: These hyper-parameters control L2 (lambda) and L1 (alpha) regularization. L2 reduces model complexity to avoid overfitting, while L1 encourages feature sparsity to select important features. Generally, smaller values allow more flexibility for the model, while larger values better constrain the model. The range from 1e-8 to 10.0 defined allows the exploration of different regularization strengths to find the optimal model.

subsample and colsample_bytree: These hyper-parameters control the proportion of samples and features used in each iteration. A lower subsample value could enhance generalization, but it may lead to underfitting if this value becomes too low. colsample_bytree, on the other hand, reduces collinearity between features. A lower colsample_bytree value helps the model prevent overfitting, while a higher value for this hyper-parameter could enable the model to include information. Hence, these two hyper-parameters are all set in the range of 0.5 to 1.0.

min_child_weight: This hyper-parameter is set within the range from 1 to 20. This hyper-parameter controls

the minimum sample weight required for each leaf node. A larger value for this hyper-parameter could prevent the model from focusing too much on noises (overfitting), whereas a smaller value could allow more flexibility during model training.

scale_pos_weight: This hyper-parameter automatically adjusts class weights to balance the ratio of purchase vs. non-purchase behaviour.

Optimal hyperparameters are applied to the XGBoost model, and the trained model is used to make predictions on the validation set. The model's performance is further evaluated by calculating metrics such as AUC and F1 Score. Specifically, as mentioned previously, the F1 Score serves as the primary optimization target, measuring the model's effectiveness in capturing the minority class.

3.3.6.3. XGBoost Model for multi-label classification

A multi-classification XGBoost is further constructed via OVR to train the model and predict multiple products simultaneously. However, key differences between building the one-to-one and one-to-many XGBoost model will be explained as follows:

Objective Function: The structure of the one-to-many model in XGBoost is similar to the one-to-one model, but it utilizes the OVR to manage multi-label tasks. Although the loss function for each label remains binary, the model overall is optimized for the macro F1 score across multiple labels, rather than the F1 score for individual products. The macro F1 score is a performance metric that averages the F1 scores calculated for each label independently, providing a measure of the model's overall accuracy and balance across all categories, particularly in imbalanced datasets.

Hyperparameter Tuning Complexity: In the one-to-one XGBoost model, the hyperparameters are tuned independently for each label. However, when constructing the multi-label XGBoost model, a shared hyperparameter tuning process is implemented. This approach optimises the entire model across predictions for all products simultaneously, enhancing overall predictive performance and efficiency.

Data Processing and Label Dependency: In the multi-label XGBoost model, all product labels are simultaneously fed into the model, which utilizes the OVR to facilitate independent predictions for each product. Although each label is predicted independently, the model shares common features and a uniform hyperparameter tuning range. This shared setup can influence the importance assigned to features, as the predictions for one product might impact the assessments of others due to their interactions within the model.

3.4. Overall Evaluation

3.4.1. Overall Performance

The table below illustrates the F1 score for all models we fitted previously, including both binary classification and multi-label classification results. Noticeably, the F1 score deviates significantly among different models.

3.4.1.1. One-to-one VS Multi-label Classification - Horizontal Analysis

When comparing the table results horizontally, the binary classification models outperform those multi-label classification models significantly. For instance, in the XGBoost model, brands Anthelios and Hyalu B5 possess a higher F1 score in the corresponding one-to-one models, with an F1 score of 0.61 and 0.35 respectively. By contrast, the F1 score in the multi-label classification models are 0.47 (Anthelios) and 0.21 (Hyalu B5), which is way below the performance in those one-to-one models.

One potential explanation for this mentioned discrepancy between one-to-one and multi-label classification models could be that the binary classification simplifies the model tasks, as it only needs to focus on distinguishing a single label. However, the multi-label classification involves classifying several labels simultaneously. Since there might be features overlap between brands, the complexity of classifying the correct label increases and the likelihood of misclassification also rises.

3.4.1.2. Model Performances Across Each Brand - Vertical Analysis

When comparing the table results vertically, the F1 scores of logistic regression are consistently the lowest across all brands. In contrast, the XGBoost models achieve the highest F1 scores for nearly every brand, followed by the random forest model and the neuron network model.

Reasons for a poor F1 score performance of the logistic regression could be that this model could only capture a linear relationship between variables.

The poor performance of logistic regression may be due to its linear nature. This means the model itself could only capture the linear relationship between variables. For example, attributes such as 'Total_Spent_AllTime' and 'Brand Description_Vitamin C_Output' exhibit scattered patterns without a clear trend, restricting logistic regression's effectiveness.

XGBoost excels mainly due to its advanced boosting strategies (e.g., gradient boosting) and its ability to handle imbalanced data by adjusting sample weights. These adjustments assist the model to focus more on those classes with relatively small observations (group 1), which could enhance the model's overall performance. Additionally, XGBoost's built-in anti-overfitting mechanisms, such as regularization, optimise

Table 7. Comparison of F1 Score Across Different Models

F1 Score for Each Brand								
	1 to 1 Calssification Result				Multi-label Classification Result			
Brand	Logistic Regression	Neuron Network	Random Forest	XGBoost	Logistic Regression	Neuron Network	Random Forest	XGBoost
Anthelios	0.29	0.43	0.6	0.61	0.29	0.29	0.45	0.47
Cicaplast	0.06	0.47	0.46	0.42	0.03	0.11	0.22	0.29
Effaclar	0.49	0.59	0.61	0.63	0.49	0.56	0.54	0.55
Hyalu B5	0	0.32	0.27	0.35	0	0.07	0.2	0.21
Lipikar	0.01	0.43	0.41	0.43	0	0.05	0.33	0.33
Toleriane	0.51	0.64	0.65	0.66	0.51	0.6	0.61	0.61
Vitamin C	0.01	0.37	0.38	0.38	0.01	0.08	0.1	0.55

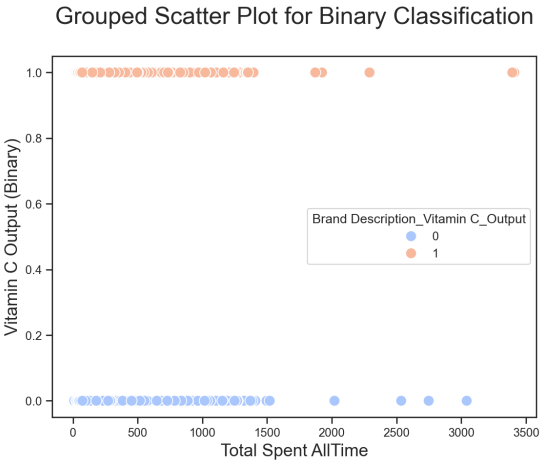


Fig. 11. Scatter Plot of Vitamin C Output VS Total Spending

its performance in complex classification tasks (please see model setup section for details).

The random forest model aggregates multiple decision trees to effectively reduce model variance while enhancing the model's robustness. While theoretically speaking, neural networks could also capture complex, non-linear relationships, they are relatively sensitive to hyperparameter tuning and the data, which often leads to local optima or overfitting. In contrast, the random forest model remains more stable and performs better in capturing non-linear interactions.

3.4.1.3. Brand Performance Across Each Model

When comparing brands across all models, Effaclar and Toleriane consistently show higher F1 scores in both binary and multi-label classifications. This indicates that these products are easier to classify, probably due to a more balanced 0 and 1 class for this label than others.

By contrast, Hyalu B5, Vitamin C, and Cicaplast generally have a lower F1 score compared to other brands, with logistic regression scores of nearly zero. This is probably due to a small number of purchases, resulting in many zero values in these labels. Such label imbalance further challenges the model to learn meaningful patterns from consumer behaviours, increasing the risk of misclassification and thus lowering the F1 score.

3.4.1.4. Selecting the best model

Fig. 12. Comparison of One-to-one VS Multi-label F1 Scores

When choosing our final best model, we prioritise the F1 score as it effectively balances precision and recall. We also consider AUC as a supportive benchmark, ranging from 0 to 1, with a value closer to 1 indicating better model performance, especially useful for imbalanced datasets.

From the AUC perspective, XGBoost and random forest models both achieve a higher AUC score compared to the neuron network. This finding further demonstrates these models' ability to address classification tasks. The low AUC for the logistic regression model also confirms the model's limitation in dealing with non-linear relationships.

Although the AUC score could assist us in assessing overall model performance, we ultimately base our decision on the F1 score due to its practical balance of precision and recall. Hence, XGBoost model is selected

Table 8. Comparison of AUC Score Across Different Models

AUC for Each Model								
	1 to 1 Calssification Result				Multi-label Classification Result			
Brand	Logistic Regression	Random Forest	XGBoost	Neuron Network	Logistic Regression	Random Forest	XGBoost	Neuron Network
Anthelios	0.56	0.59	0.61	0.59	0.56	0.6	0.58	0.54
Cicaplast	0.62	0.66	0.58	0.67	0.6	0.67	0.65	0.63
Effaclar	0.7	0.72	0.73	0.71	0.7	0.72	0.69	0.71
Hyalu B5	0.46	0.77	0.8	0.79	0.46	0.8	0.76	0.75
Lipikar	0.62	0.66	0.67	0.66	0.58	0.68	0.66	0.64
Toleriane	0.73	0.77	0.78	0.75	0.73	0.76	0.76	0.74
Vitamin C	0.54	0.53	0.57	0.57	0.53	0.58	0.59	0.56

as our final best model due to its lowest F1 scores across all brands.

3.5. Model Interpretation

The next section will focus on interpreting the XGBoost model we selected, which primarily focuses on identifying attributes or products that most influence label classification. The interpretation will be mainly based on the feature importance plot of the XGBoost model, which reflects how frequently each variable is used when fitting the model. A higher feature importance score indicates a greater influence of that variable. Generally, the most important features can be grouped into 2 parts: money-related features and product-related features.

3.5.1. Money-related Features

Total_ Spent_ Nov23_ May24 and Total_ Spent_ AllTime are consistently the top two contributive attributes across all brands, indicating that spending behaviour is the most important factor in determining brand preference. This pattern demonstrates that consumer expenditure plays a significant role in determining which brand to purchase, especially for brands Anthelios, Vitamin C, and Toleriane.

3.5.2. Product-related Features

Commonalities across all brands:

Face Care and irritation-prone skin-related products appear frequently across all brands, suggesting their strong influence on brand classification. Face moisturiser and face serum are also consistently ranked in the top five for nearly all brands, showing their key role in differentiation. This also suggests that these are the most popular product types for La Roche Posay. Additionally, toleriane and anti-ageing related products often appear, suggesting their relevance in distinguishing certain brands due to their targeted customer base.

Differences between each brand:

Brands Effaclar and Cicaplast focus heavily on face care and irritation-prone skin products, demonstrating their sensitivity-targeted approach. Hyalu B5 also stands out with a higher importance on anti-ageing products, reflecting its focus on anti-aging functions. Lipikar prioritises irritation-prone skin products more, indicating its emphasis on soothing skin concerns. Such differentiation strategies allow each brand to establish a unique identity among specific customer segments

Table 9. Feature Importance for Multi-Label XGBoost Model (Product-related)

XGBOOST Multi-label Feature Importance Table						
Anthelios	Cicaplast	Effaclar	Hyalu B5	Lipikar	Toleriane	Vitamin C
Face Care	Face Care	Irritation Prone Skin	Face Care	Face Care	Face Care	Face Care
Irritation Prone Skin	Irritation Prone Skin	Face Care	Irritation Prone Skin	Irritation Prone Skin	Irritation Prone Skin	Irritation Prone Skin
Face Moisturiser	Face Moisturiser	Face Moisturiser	Anti-Anging	Face Moisturiser	Face Serum	Face Moisturiser
Face Serum	FaceSerum	Face Serum	Face Serum	Face Serum	Face Moisturiser	Face Serum
Toleriane	Toleriane	Anti-anging	Irritation Prone Skin	Toleriane	Anti-Ageing	Anti-Ageing

Table 10. Feature Importance for Soulmate Segment (Product-related)

XGBOOST Multi-label Feature Importance Table(Soulmate)						
Anthelios	Cicaplast	Effaclar	Hyalu B5	Lipikar	Toleriane	Vitamin C
Face Care	Cicaplast	Effadcular	Toleriane	Anti-Ageing	Face Care	Anti-Ageing
Face Moisturiser	Irritation Prone Skin	Face Moisturiser	Hyalu B5	Irritation Prone Skin	Irritation Prone Skin	Irritation Prone Skin
Irritation Prone Skin	Face Care	Irritation Prone Skin	Face Moisturiser	Face Care	Face Serum	Face Care
Anti-Ageing	Face Serum	Face Cleanser	Face Serum	Face Moisturiser	Face Moisturiser	Face Moisturiser
FaceSerum	Face Moisturiser	Face Care	Irritation Prone Skin	Lipkar not buy	Anti-anging	Lipkar not buy

Table 11. Feature Importance for Ex-lover Segment (Product-related)

XGBOOST Multi-label Feature Importance Table (Ex-Lover)						
Anthelios	Cicaplast	Effaclar	Hyalu B5	Lipikar	Toleriane	Vitamin C
Face Care	Face Care	Face Care	Face Care	Face Care	Face Care	Irritation Prone Skin
Face Moisturiser	Face Moisturiser	Irritation Prone Skin	Anti-Ageing	Face Moisturiser	Face Moisturiser	Face Care
Irritation Prone Skin	Irritation Prone Skin	Effaclar	Irritation Prone Skin	Irritation Prone Skin	Irritation Prone Skin	Face Moisturiser
Face Serum	Toleriane	Acne Prone Skin	Face Moisturiser	Toleriane	Toleriane	Anti-anging
Anti-Ageing	Anti-Ageing	Toleriane	Face Serum	Lipkar not buy	Face Serum	Face Serum

Table 12. Feature Importance for Potential Lover Segment (Product-related)

XGBOOST Multi-label Feature Importance Table (Potential Lover)						
Anthelios	Cicaplast	Effaclar	Hyalu B5	Lipikar	Toleriane	Vitamin C
Face Care	Face Care	Face Care	Face Care	Face Moisturiser	Face Care	Face Care
Face Moisturiser	Face Moisturiser	Face Moisturiser	Hyalu B5	Face Care	Irritation Prone Skin	Face Moisturiser
Irritation Prone Skin	Irritation Prone Skin	Irritation Prone Skin	Anti-Ageing	Irritation Prone Skin	Toleriane	Irritation Prone Skin
Face Serum	Anti-Ageing	Anti-Ageing	Face Moisturiser	Lipkar not buy	Face Serum	Anti-anging
Anti-Ageing	Cicaplast	Face Serum	Face Serum	Acne_Prone	Face Moisturiser	Toleriane

3.6. Consumer Segmentation Behaviour Analysis Using XGBoost

To analyze consumer behavior within each segment (i.e., soulmates, ex-lovers, and potential lovers), we applied our best model (XGBoost) to each segment respectively. The next section will provide insights into consumer purchasing behaviour for each segment based on our findings.

The tables above summarise the top 5 most influential products for each brand across each consumer segment. Commonalities across all segments:

Again, face care and irritation-prone skin products are the most influential and popular product types across all consumer segments, highlighting their influence on brand differentiation. Additionally, face moisturiser, face serum, and anti-ageing products also frequently appear, indicating their significance in consumer preference.

Differences across all segments:

The Soulmate group focuses on basic care and anti-aging products across brands compared to other consumer segments. Ex-Lover, on the other hand, shows a unique emphasis on acne-prone skin (Effaclar),

suggesting a concern for specific issues. For the potential-lover group, the product focus is more diverse, highlighting multiple skincare needs. Overall, while there are shared influential features, each segment has distinct preferences, guiding targeted marketing strategies.

3.7. Model Limitation

We acknowledge the fact that certain factors, including the imbalanced dataset, potential issues raised by data leakage and the need to construct an extra two-stage model, could negatively influence our model performance. The following part will justify and criticise how these mentioned factors might affect our model performance.

3.7.1. Imbalanced Dataset

One of the main issues we encountered was the class imbalance in our dataset. There exists a significantly higher proportion of 0 values (customers who do not purchase this product) compared to the proportion of 1 value (customers who purchase the product). This may influence the model's tendency to predict more majority classes. However, proper techniques for handling class imbalances are implemented, as mentioned in the model building section, to avoid any undesirable misclassification.

3.7.2. Criticise the Data Leakage Concern

Data leakage is not a concern in our project since the target labels are based on transaction records up to November 2023. However, all predictive features come from records before the previous November, meaning this temporal separation ensures no future data was used during training.

3.7.3. Criticise the need for a Two-Stage Model

We noticed the importance and applicability of the two-stage model in real-world business scenarios. Typically, a two-stage model effectively conducts a binary classification to filter out customers with purchase records at the beginning, which allows the model to only focus on observations with valid purchasing records. However, both minimum total spending and transaction counts are non-zero in our dataset, indicating there is no need to formulate a two-stage model for this business problem.

4. 4. Proposal of the Strategic Project

4.1. 4.1 Strategic Project 1: Expanding the Flagship Product Lines

4.1.1. Strategy Overview & Rationale Behind It

Our first strategic recommendation primarily focuses on expanding our flagship product line. After analysing the feature importance plots of the XGBoost model, we identified the top 5 most influential products for each brand. According to Table 9, generally, face serums and face moisturizers are the most popular product types. Products that target irritation-prone skin and with anti-aging purposes also appear frequently on the table, meaning these are the most demanding product functions in the market. Besides, the brand "Toleriane" contributes significantly to classifying the corresponding brand in our model, indicating its strong market potential. Additionally, by integrating our findings from the EDA section, we also notice that the total sales for the brand Anthelios are the highest, meaning Anthelios and Toleriane's core products and target audience are more sensitive and responsive to changes in market demand.

Hence, based on the findings concluded above, we believe that the current market demand is specially focused on facial serums and moisturisers, with consumers showing a stronger preference for products that offer sensitive skin repair and anti-aging benefits. Hence, La-Roche Posay should promote products for brands Toleriane and Anthelios, with a target focus on facial serums and moisturisers. The following part will justify how this strategic recommendation could be implemented in the reality.

4.1.2. Strategic Action Plan

La-Roche Posay should expand its product lines for the brands Toleriane and Anthelios, with specific focuses on sensitive skin repair and anti-aging benefits. To implement this strategy effectively, comprehensive skincare sets can be introduced, with products that cover every step of a skincare routine—from cleansing and moisturising to repair and anti-aging treatment. This strategy could provide a complete solution for our customers. Additionally, high-end, limited-edition gift boxes will be introduced to our customers, such as repair serums and anti-aging moisturisers. This strategy could satisfy the needs of mid-to-high-end customers who seek a premium skincare experience.

Apart from introducing comprehensive skincare sets, La-Roche Posay could also introduce seasonal promotions and limited-edition marketing strategies to address customers' skincare requirements throughout the year. To implement this, seasonal product bundles tailored to specific skincare needs can be introduced. For example, face moisturiser and face repairing products can be bundled and introduced during winter, while oil control and sun protection products can be promoted during summer. Moreover, special holiday-themed gift boxes with limited quantities could also be introduced during festive periods like Christmas, New Year, and Valentine's Day. The limited product quantity offered could further create a sense of exclusivity for the brand, encourage quick purchasing decisions of the customers,

and ultimately, boost the total sales by leveraging the scarcity of the products.

By implementing these mentioned strategies, consumers with different skincare needs could all be covered, and these strategies could effectively expand our brand influence and increase total sales.

4.1.3. Quantify the Strategy Impact

We start our analysis by estimating costs and projected increase in revenue. Based on the tailored marketing strategies mentioned above, it is reasonable to assume that the total revenue will be increased to a certain extent. Here, the formula we applied in quantifying the project impact is as follows:

Estimated Increased Revenue = Customers' Numbers × Average Transaction Amount × Frequency × Estimated Increase Rate - Cost

Relevant assumptions made in this section include:

1. Assumption 1: The market growth rate after implementing our strategy is estimated to be 7.8According to the reports by Future Market Insights and Grand View Research, brands investing in customised skincare solutions could obtain an average market growth of 7.8% (Sudip, 2024; Grand View Research, 2023).
2. Assumption 2: 2,652 customers will purchase brands Toleriane and Anthelios in the next six-month period after implementing our proposed promotion strategy. This is because of the reason that in our dataset, 5,304 customers purchased Toleriane and Anthelios products over one year. By dividing this number by 2, we could assume that there will be 2,652 customers purchasing these two brands.
3. Assumption 3: It is assumed that each customer will purchase Toleriane and Anthelios once every six months. After consulting the data from La Roche-Posay's online store, the price range for Toleriane products is approximately 30−33, while Anthelios products range from $25-$35 (Byrdie, 2023; Dischem, 2024; La-Roche Posay, 2023). Hence, the Average Transaction Amount × Frequency can be calculated as (30 + 25) * 2 = $110
4. Assumption 4: The corresponding marketing expenses of our proposal are estimated to be 10% of sales. According to L'Oréal's financial reports, advertising and promotion expenses typically account for 31.5% to 32.4% of sales. Since our customised marketing activities are specifically focused on brands Toleriane and Anthelios, we assume the corresponding marketing expenses will be 10% of sales. This means the total marketing expenses can be calculated as 2652*55* 10% = $14586

By applying these mentioned numbers to the Estimated Increased Revenue formula, after implementing customised strategies and marketing activities, approximately $8,525 in revenue growth could be generated. However, we must acknowledge that this estimation is solely based on our existing dataset with 6,400 observations, meaning the proposed targeted marketing strategy could bring an additional $8,525 in revenue for every 6,400 customers focused.

4.2. Strategic Project 2: Tailored Marketing Strategies for Each Consumer Segment

4.2.1. Strategy Overview

In the previous section, the dataset has been clustered into three consumer segments: soulmates, ex-lovers, and potential lovers. Customised marketing strategies for each group will be recommended, aiming at boosting conversion rates and improving client lifetime value.

4.2.2. Strategic Action Plan (for Each Consumer Segment)

4.2.2.1. Soulmates

For the Soulmates group, the loyalty program is recommended and implemented through the reward points system. Customers could earn points for every AU$50 spent, which can be redeemed for rewards like cosmetic samples, encouraging repeat purchases to accumulate points faster. According to Shap (1997), setting a threshold for the reward helps to increase single purchase amounts. A practical implication for this reward plan could be that the customers can earn 1 point for every $50 spent or they could obtain 4 points for every $150 spent. This strategy could effectively motivate larger purchases. Additionally, during the major discount seasons (i.e., between November and February), special promotional offers such as double points rewards could be issued to encourage customers' spending, further boosting purchase frequency and amount.

Since Soulmates' average spending is around $138, the threshold for obtaining a reward point is set at $50. Hence, customers would be encouraged to spend an extra $150 to earn more reward points. This approach could stimulate the consumers to increase their spending in increments that are natural to them, thereby raising the average transaction value for this segment.

4.2.2.2. Ex-lovers

For ex-lovers, we use a targeted marketing approach to send tailored emails regarding their previously engaged products. In order to entice customers to return to our brand, relevant email content may contain time-limited incentives or discount vouchers. By appealing to their previous tastes and offering an extra incentive, this marketing strategy seeks to re-engage those former

loves, fostering recurring business and mending their relationship with the brand.

A double points program is also proposed for the ex-lover segment. When ex-lovers activate their accounts and make their first purchase, additional reward points are offered to them. By leveraging the appeal of earning extra rewards in a short timeframe, ex-lovers could be more encouraged and stimulated to spend more, which could motivate them to return and rebuild their loyalty to the brand.

4.2.2.3. Potential Lovers

For potential lovers, our primary marketing goal is to convert them into soulmates by providing highly customised product bundles. We intend to create carefully crafted bundles that include their best-loved product features, such as products with features including reduced irritation, anti-ageing, and anti-acne (Figure 13). These bundles might provide potential lovers with customised care that addresses their specific skin conditions. Furthermore, rewarding customers with loyalty points or introductory discounts (like 3%) on their first purchases can strengthen their bond with the company and promote long-term involvement.

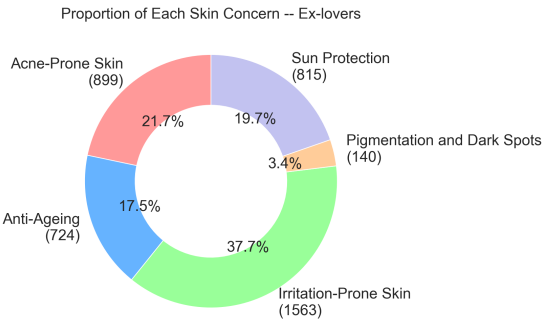


Fig. 13. Proportion for each skin concern - Ex-lovers

4.2.3. Quantify the Strategy Impact

The following part will quantify the project impact of this recommendation. The number of customers in the soulmate segment will increase by 10%, mainly due to the conversion of some “Potential Lovers” into “Soulmates”. Additionally, the average transaction value is also expected to increase by 10%, driven by promotional efforts encouraging higher spending. This 10% estimate is based on Palmatier’s analysis that similar loyalty programs can yield a 5% to 12% increase (2014). For the ex-lovers segment, it is estimated that personalised emails and attractive offers will achieve a reactivation rate of 7%. According to Mahajan (2023), a similar marketing strategy could lead to a 7% increase in the number of customers. Additionally, due to the tailored offers and points incentives provided, these reactivated customers are likely to increase both their purchase frequency and transaction value. After re-activation, both purchase frequency and transaction amount are estimated to increase by 5%, supported by relevant market assumptions. For the “potential lovers” segment, it is assumed that the bundling promotion will ultimately convert 10% of the original ”Potential Lovers” into ”Soulmates”. This also suggests that the number of customers in the “potential lovers” segment is expected to decrease by 10%.

4.2.3.1. Cost Analysis

According to McKinsey’s 2018 consumer sector research, depending on the size and target audience, marketing campaign setup costs normally range from $10,000 to $50,000. A $10,000 project cost could be a reasonable estimate for our soulmate sector, which has about 1,000 clients. This $10,000 cost also reflects the lower-end range of the report for smaller campaigns. For the ex-lovers and potential lovers segments, offering a 3% discount would incur a cost equivalent to 3% of their respective revenues. Based on McKinsey’s report (2018), bundling products could be an effective approach to raising perceived value without raising production costs too much. Hence, based on this argument, our proposed strategy could efficiently reduce costs and increase client interaction by grouping popular products. By applying this approach, all cost estimates mentioned in this section are grounded in real market data and supported by research, making our analysis robust and data-driven.

Calculation of the increased revenue

The following section illustrates the detailed calculation for quantifying the project impact.

The estimated increased revenue should be calculated through this formula: Estimated Increased Revenue= Customers’ Numbers × Average Transaction Amount × Frequency × Estimated increase rate - Cost

As illustrated in the formula above, the only unknown variable in this formula is “Estimated Increase Rate”, however, this variable can be inferred based on all analyses mentioned above. The table below calculates and summarises all variables mentioned in the formula for each consumer segmentation respectively.

By entering all figures into the formula from the two tables above, we successfully calculate the estimated increased revenue for each consumer segment, as explained below:

Soulmates:

Estimated Increased Revenue= (1054+1987 × 0.1) × 137.26 × 1.52 × (1.21-1) - 10000=57660

Ex-Lovers:

Estimated Increased Revenue=947 × 90.56 × 1.19 × (1.1025-1) × (0.97) = 14724

Potential Lovers:

Table 13. Estimated Increase Rate Calculation

	Customer's Numbers Increased Rate	Average Transaction Amount Increased Rate	Frequency Increased Rate	Estimated Increase Rate
Soulmate	1.1	1.1	1	1.21
Ex-Lover	1	1.05	1.05	1.1025
Potential Lover	0.9	1	1	0.9

Table 14. Statistical Summary for Each Consumer Segment

	Customer Amout	Average Transcation Amount	Frequency
Soulmates	1054	137.26	1.52
Ex-Lovers	947	90.56	1.19
Potential Lovers	1987	1.19	0.58

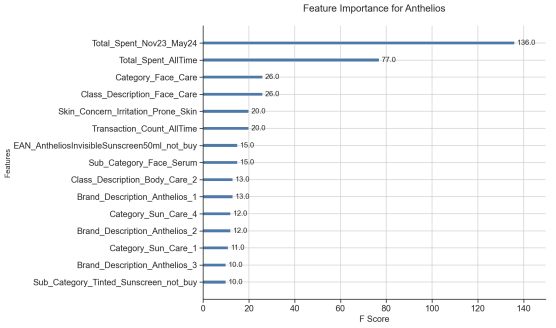
Estimated Increased Revenue=(1987-1987 × 0.1) × 126.26 × 0.58 × (0.9-1) × (0.97) = -18501
By adding up the total revenue of each consumer segment, the total estimated increase in revenue is: 57660 + 14724-18501 =53883, which would be an 11.53% increase compared to the current revenue.

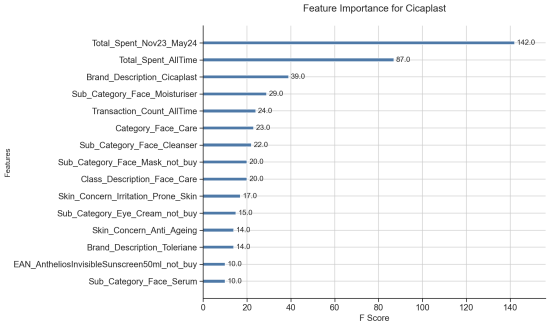
References

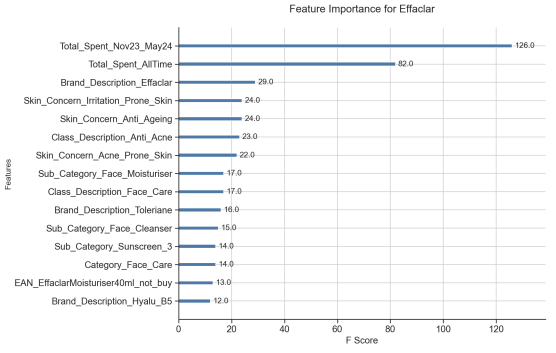
1. Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Database Marketing: Analyzing and Managing Customers (1st edition.). Springer New York.
2. Clichy. (2023, July 28). 2023 Half-Year Results. L'Oréal Finance. https://www.loreal-finance.com/eng/news-release/2023-half-year-results
3. Future Market Insights. (2024, March). Customized Skincare Market. Www.futuremarketinsights.com. https://www.futuremarketinsights.com/reports/customized-skincare-market
4. Grand View Research. (2023). Personalized Skin Care Products Market Size Report, 2030.
5. McKinsey & Company. (2015). The consumer sector in 2030: Trends and questions to consider. McKinsey & Company. https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/the-consumer-sector-in-2030-trends-and-questions-to-consider
6. L'Oréal Finance. (2023, February 9). 2022 Annual Results. L'Oréal Finance. https://www.loreal-finance.com/eng/news-release/2022-annual-results
7. Porcay, L. (2024). Toleriane Double Repair Face Moisturizer — La Roche-Posay. La Roche-Posay - Skincare, Sunscreen, Body Lotion - Official Site. https://www.laroche-posay.us/our-products/face/face-moisturizer/toleriane-double-repair-face-moisturizer-tolerianedoublerepair.html
8. Sharp, B., & Sharp, A. (1997). Loyalty Programs and Their Impact on repeat-purchase Loyalty Patterns. International Journal of Research in Marketing, 14(5), 473-486. https://doi.org/10.1016/s0167-8116(97)00022-0
9. Statista. (2024). Leading personal care brands among female consumers in Australia as of February 2024, by index score. https://www.statista.com/statistics/1466518/australia-leading-personal-care-brands-by-index-score/
10. Statista. (2024). Skin Care. https://www.statista.com/outlook/cmo/beauty-personal-care/skin-care/australia
11. Steinhoff, L., & Palmatier, R. W. (2014). Understanding loyalty program effectiveness: managing target and bystander effects. Journal of the Academy of Marketing Science, 44(1), 88-107. https://doi.org/10.1007/s11747-014-0405-6
12. Weingus, L. (2021). This Derm-Approved Moisturizer Kept My Skin Hydrated, Never Greasy. Byrdie. https://www.byrdie.com/la-roche-posay-toleriane-ultra-sensitive-skin-face-moisturizer-review-5183783

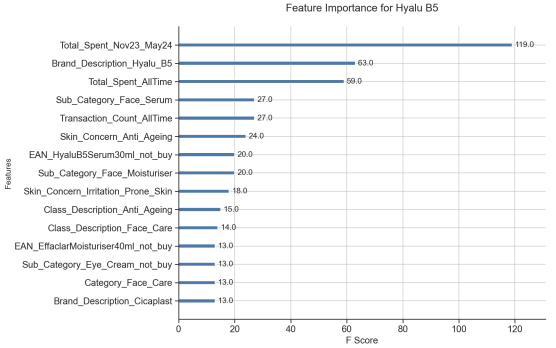
Appendix

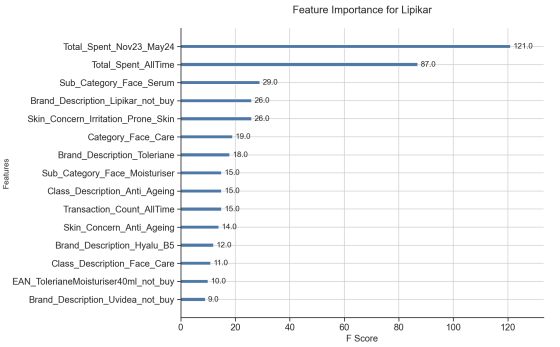
A. XGBoost Feature Importance Plots for Each Label (Entire Dataset)

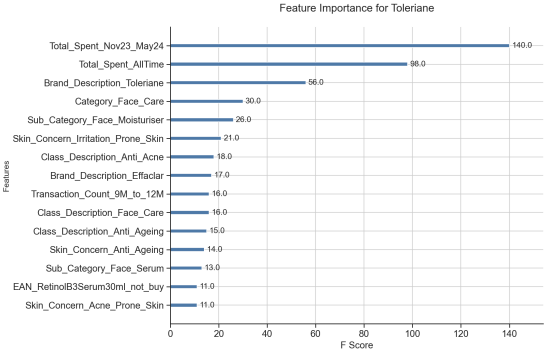


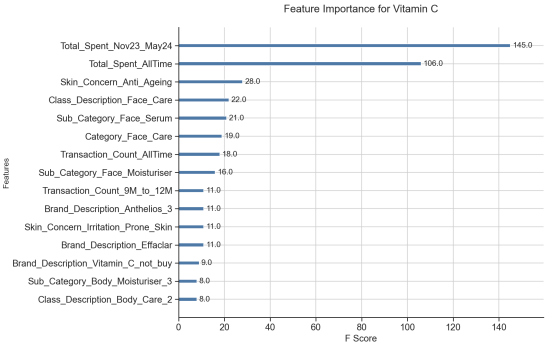




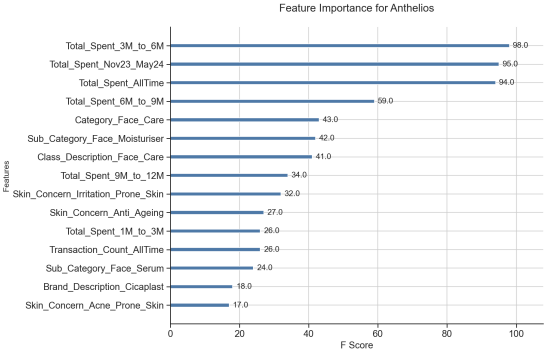


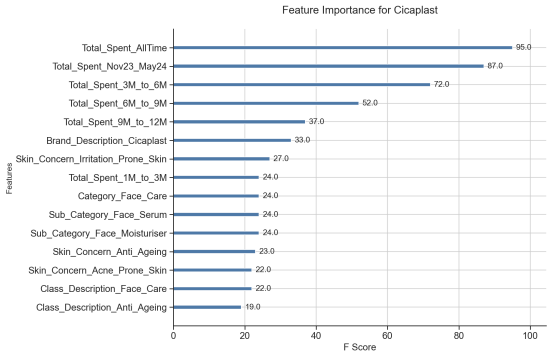


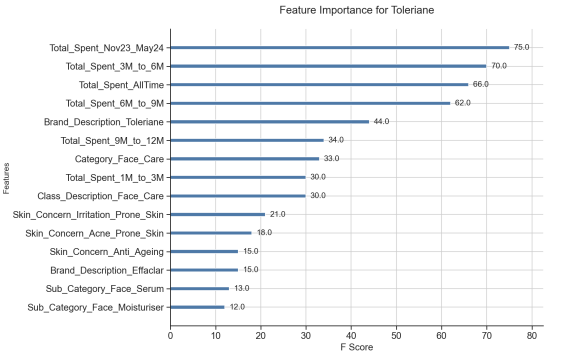


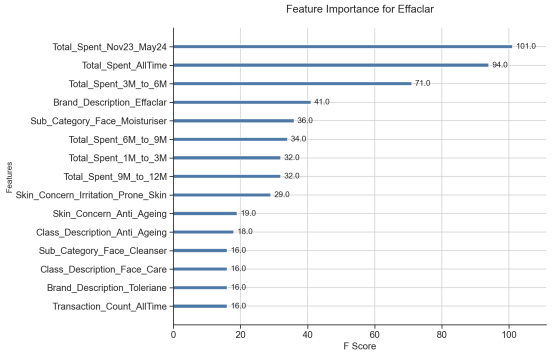


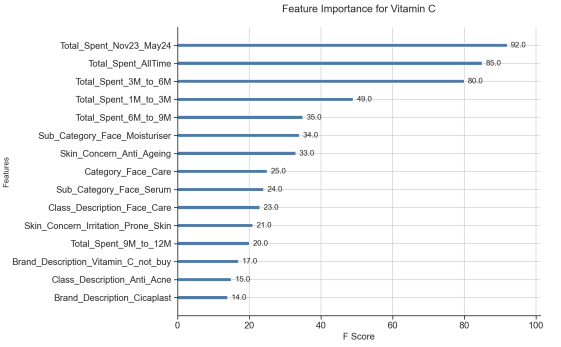
B. XGBoost Feature Importance Plots for Each Label (Soulmate Segment)



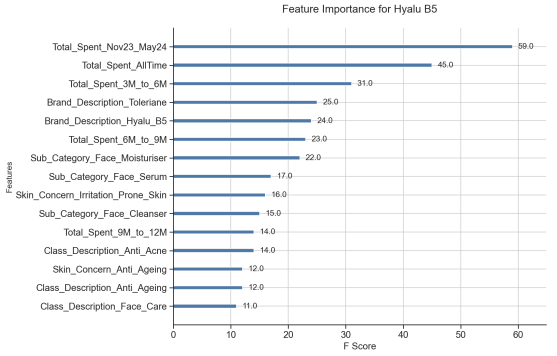


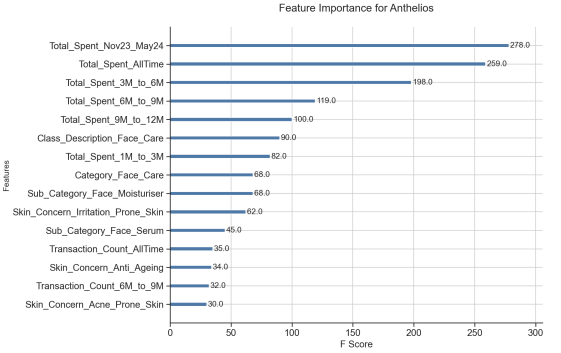


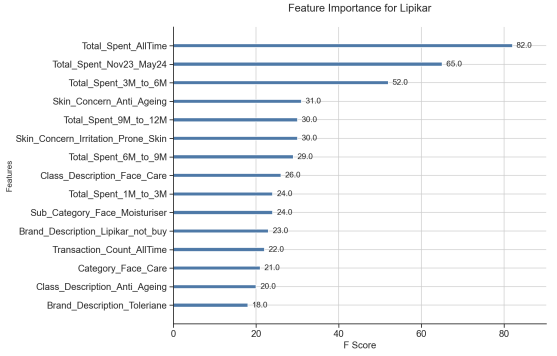


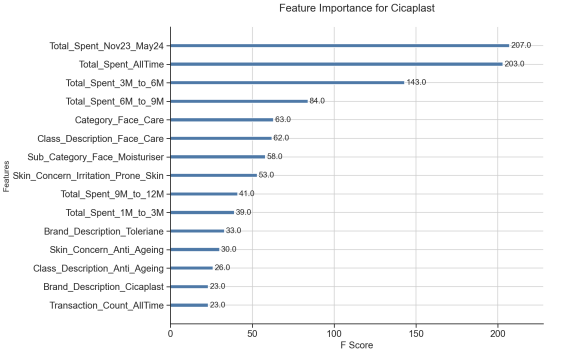


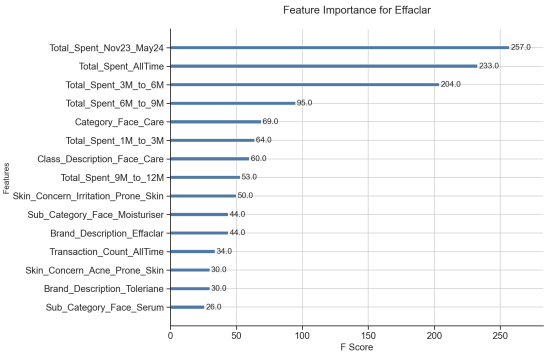
C. XGBoost Feature Importance Plots for Each Label (Ex-lovers Segment)

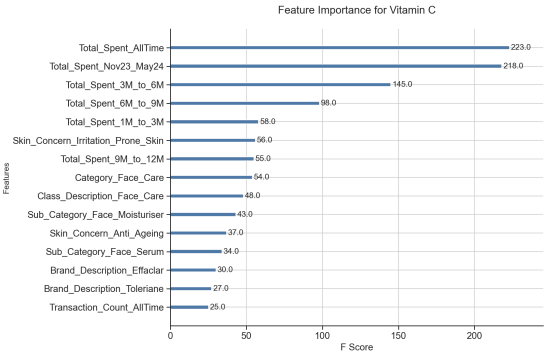


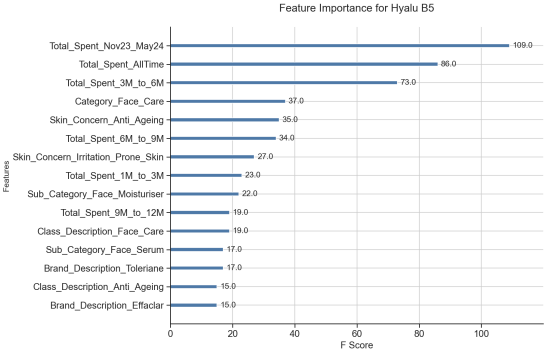




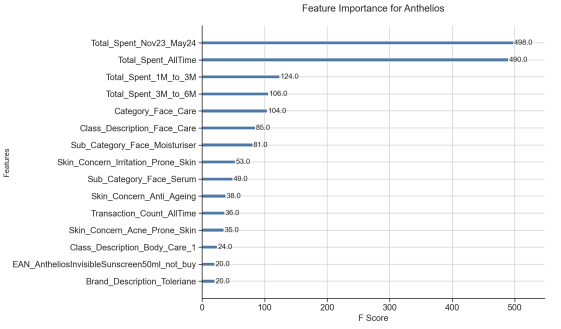


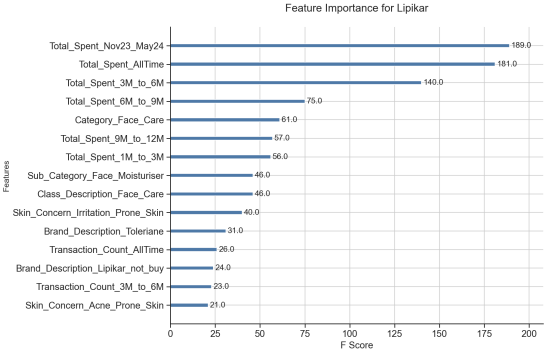


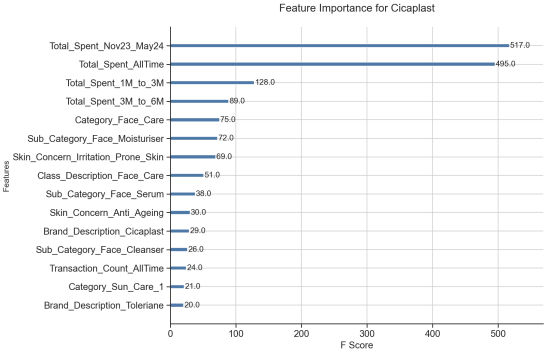


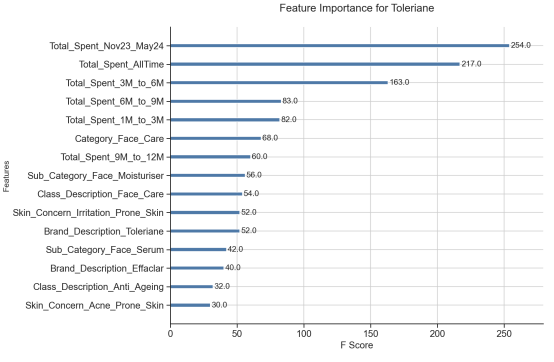


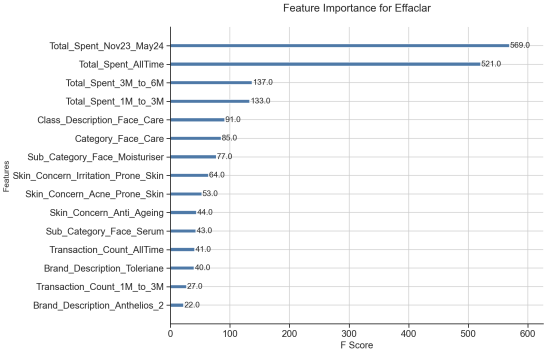
D. XGBoost Feature Importance Plots for Each Label (Potential Lovers Segment)

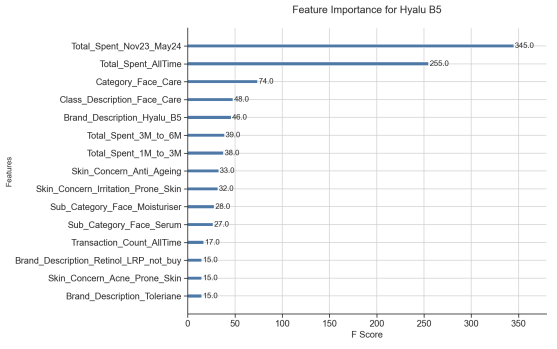


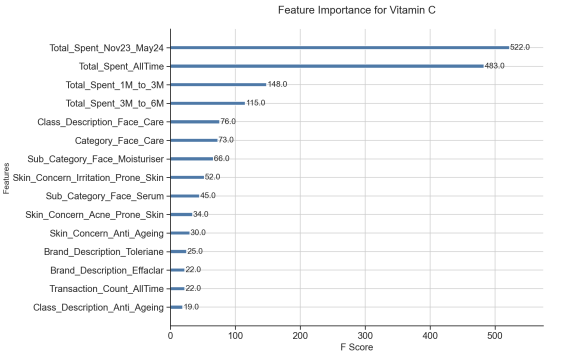


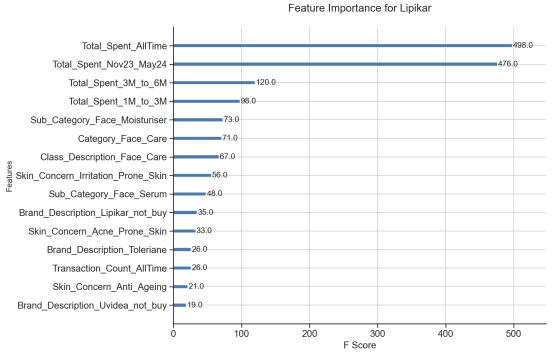


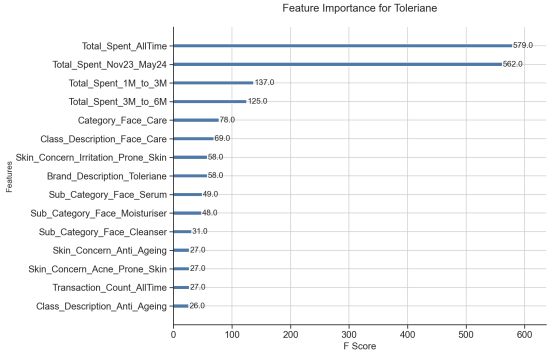





Feature Importance for Hyalu B5
Total_Spent_Nov23_May24
Total_Spent_AllTime
Category_Face_Care
Class_Description_Face_Care
Brand_Description_Hyalu_B5
Total_Spent_3M_to_6M
Total_Spent_1M_to_3M
Skin_Concern_Anti_Ageing
Skin_Concern_Irritation_Prone_Skin
Sub_Category_Face_Moisturiser
Sub_Category_Face_Serum
Transaction_Count_AllTime
Brand_Description_Retinol_LRP_not_buy
Skin_Concern_Acne_Prone_Skin
Brand_Description_Toleriane
345.0
255.0
74.0
48.0
46.0
39.0
38.0
33.0
32.0
28.0
27.0
17.0
15.0
15.0
15.0
Features
0
50
100
150
200
250
300
350
F Score


Feature Importance for Vitamin C
Total_Spent_Nov23_May24
Total_Spent_AllTime
Total_Spent_1M_to_3M
Total_Spent_3M_to_6M
Class_Description_Face_Care
Category_Face_Care
Sub_Category_Face_Moisturiser
Skin_Concern_Irritation_Prone_Skin
Sub_Category_Face_Serum
Skin_Concern_Acne_Prone_Skin
Skin_Concern_Anti_Ageing
Brand_Description_Toleriane
Brand_Description_Effaclar
Transaction_Count_AllTime
Class_Description_Anti_Ageing
522.0
483.0
148.0
115.0
76.0
73.0
66.0
52.0
45.0
34.0
30.0
25.0
22.0
22.0
19.0
Features
0
100
200
300
400
500
F Score


Feature Importance for Lipikar
Total_Spent_AllTime
Total_Spent_Nov23_May24
Total_Spent_3M_to_6M
Total_Spent_1M_to_3M
Sub_Category_Face_Moisturiser
Category_Face_Care
Class_Description_Face_Care
Skin_Concern_Irritation_Prone_Skin
Sub_Category_Face_Serum
Brand_Description_Lipikar_not_buy
Skin_Concern_Acne_Prone_Skin
Brand_Description_Toleriane
Transaction_Count_AllTime
Skin_Concern_Anti_Ageing
Brand_Description_Uvidea_not_buy
498.0
476.0
120.0
98.0
73.0
71.0
67.0
56.0
48.0
35.0
33.0
26.0
26.0
21.0
19.0
Features
0
100
200
300
400
500
F Score


Feature Importance for Toleriane
Total_Spent_AllTime
Total_Spent_Nov23_May24
Total_Spent_1M_to_3M
Total_Spent_3M_to_6M
Category_Face_Care
Class_Description_Face_Care
Skin_Concern_Irritation_Prone_Skin
Brand_Description_Toleriane
Sub_Category_Face_Serum
Sub_Category_Face_Moisturiser
Sub_Category_Face_Cleanser
Skin_Concern_Anti_Ageing
Skin_Concern_Acne_Prone_Skin
Transaction_Count_AllTime
Class_Description_Anti_Ageing
579.0
562.0
137.0
125.0
78.0
69.0
58.0
58.0
49.0
48.0
31.0
27.0
27.0
27.0
26.0
Features
0
100
200
300
400
500
600
F Score