Table of Contents

List of Figures & Tables

1.0 Introduction

In today’s competitive eCommerce landscape, customer reviews are a vital component of business success, directly influencing trust, visibility, and sales. For Nile, a leading South American eCommerce platform, maintaining a steady stream of positive reviews is essential for reinforcing customer confidence and driving conversions. This project aims to develop a predictive model to identify customers most likely to leave positive reviews, enabling targeted engagement strategies that optimise resource allocation while enhancing customer satisfaction.

The report provides a detailed overview of the methodology, including data preparation, exploratory analysis, model testing, and evaluation, and concludes with actionable recommendations aligned with Nile’s objectives. While the model shows considerable potential in predicting positive reviews, addressing key challenges such as improving data quality and ensuring seamless operational integration will be vital for achieving long-term effectiveness and business impact.

2.0 Methodology

The project has been approached using the CRISP-DM framework.

2.1 Business Understanding

Customer reviews are critical for Nile as they directly influence consumer trust, brand reputation, and sales. Positive reviews enhance visibility and act as social proof, driving competitiveness in the online retail space. However, managing reviews efficiently is resource-intensive, requiring a targeted approach to engage the right customers. This project aims to develop a predictive model to identify customers most likely to leave positive reviews, enabling Nile to optimize resource allocation, boost the volume of positive feedback, and strengthen its market position.

2.2 Data Understanding

Nile provided 9 datasets containing customers, orders, reviews and product details:

- Olist_Customers_Dataset
- Olist_Geolocation_Dataset
- Olist_Order_Items_Dataset
- Olist_Order_Payments_Dataset
- Olist_Order_Reviews_Dataset
- Olist_Orders_Dataset
- Olist_Products_Dataset
- Olist_Sellers_Dataset
- Product_Category_Name_Translation

Table 1, (See Appendices) illustrates the data dictionary for each of the datasets.

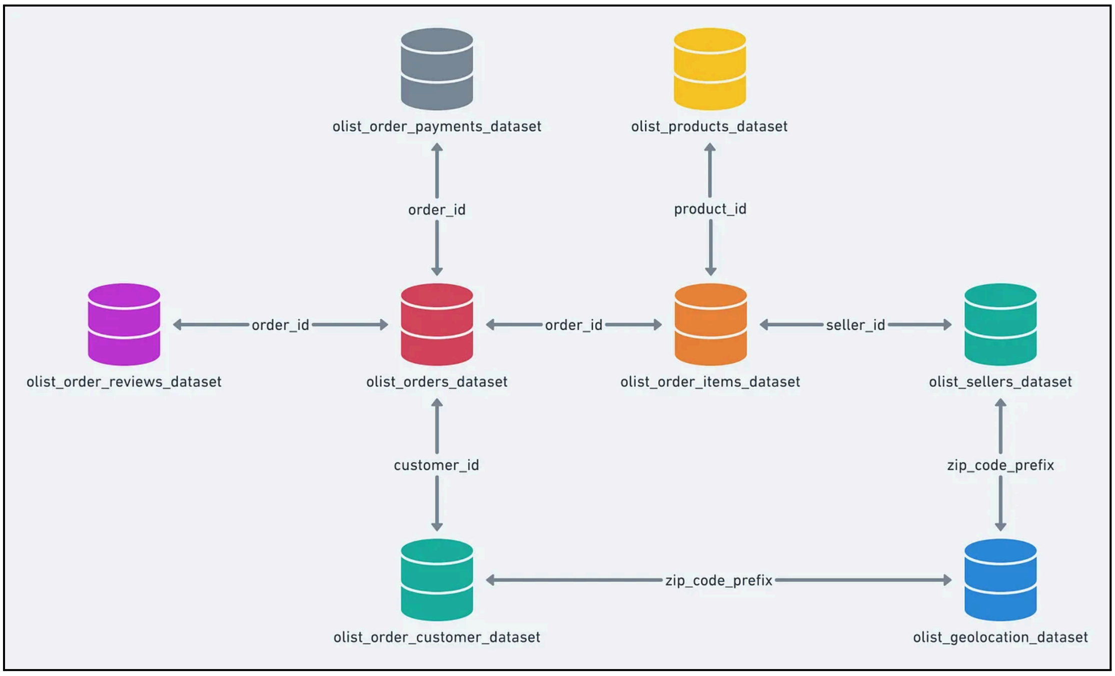


Figure 1. Dataset Overview

2.2.1 Outlier Analysis

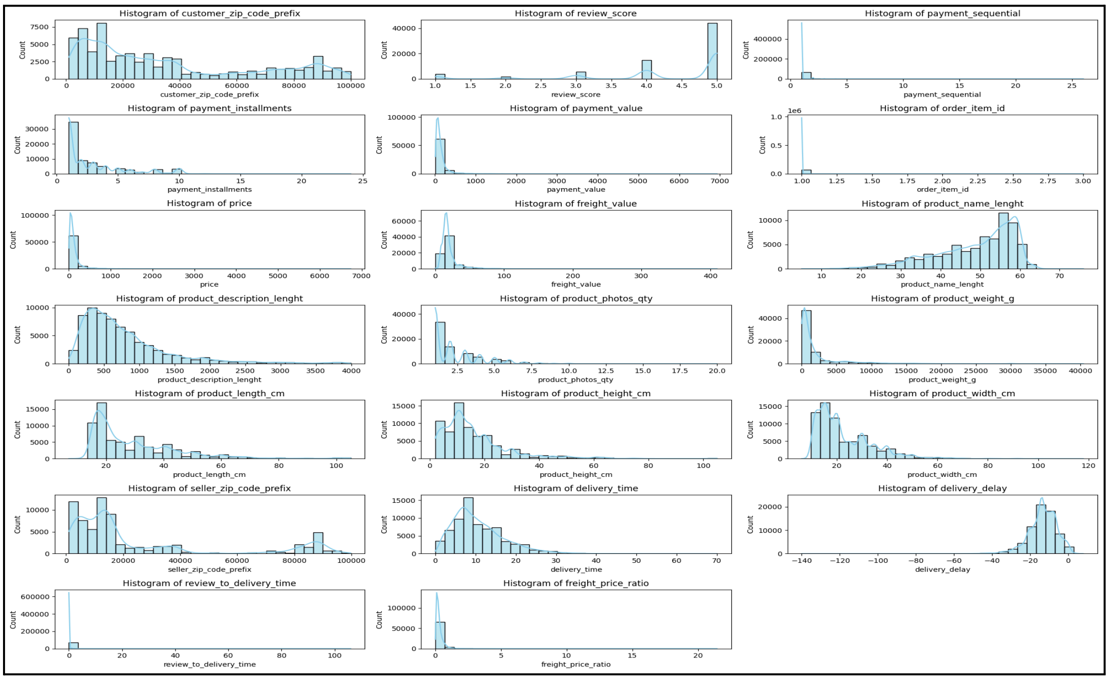

Histogram of customer_zip_code_prefix
Histogram of review_score
Histogram of payment_sequential
Histogram of payment_installments
Histogram of payment_value
Histogram of order_item_id
Histogram of price
Histogram of freight_value
Histogram of product_name_lenght
Histogram of product_description_lenght
Histogram of product_photos_qty
Histogram of product_weight_g
Histogram of product_length_cm
Histogram of product_height_cm
Histogram of product_width_cm
Histogram of seller_zip_code_prefix
Histogram of delivery_time
Histogram of delivery_delay
Histogram of review_to_delivery_time
Histogram of freight_price_ratio

Figure 2, Dataset Visualisations

Outliers were retained to capture real-world variability in Nile's operations, ensuring the model reflects critical edge cases like high payment values or delivery delays, crucial for understanding customer behaviour and operational challenges.

2.2.2 Initial EDA

Exploratory analysis was conducted to identify key patterns, distributions, and relationships within the datasets, guiding data cleaning and feature engineering. Key observations are summarised in Table 2 (see Appendices) and visualised in the following graphs, providing an overview to inform dataset merging and feature engineering.

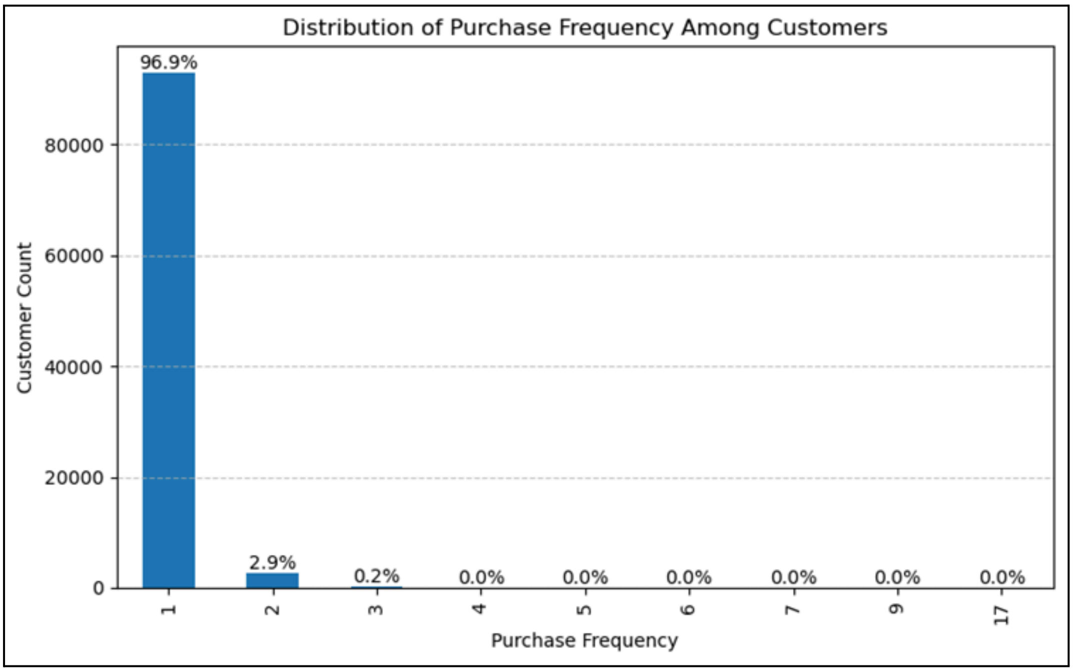


Figure 3, Customers: 96.9% of customers are one-time customers.

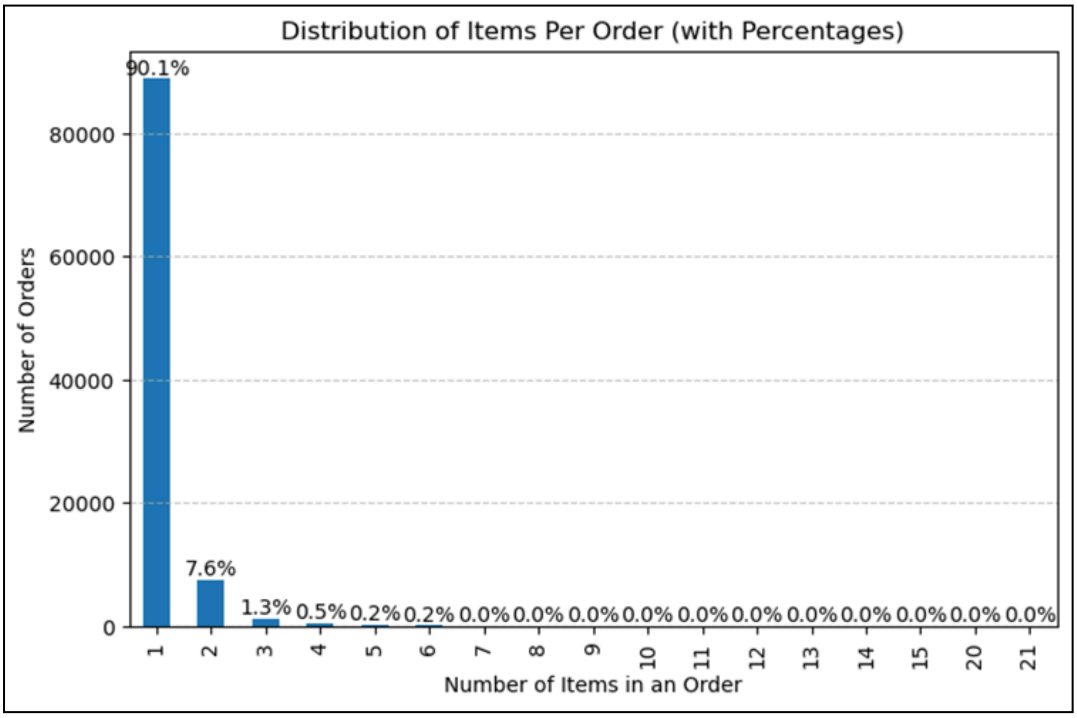


Figure 4, Items: 90.1% of orders contain only one item

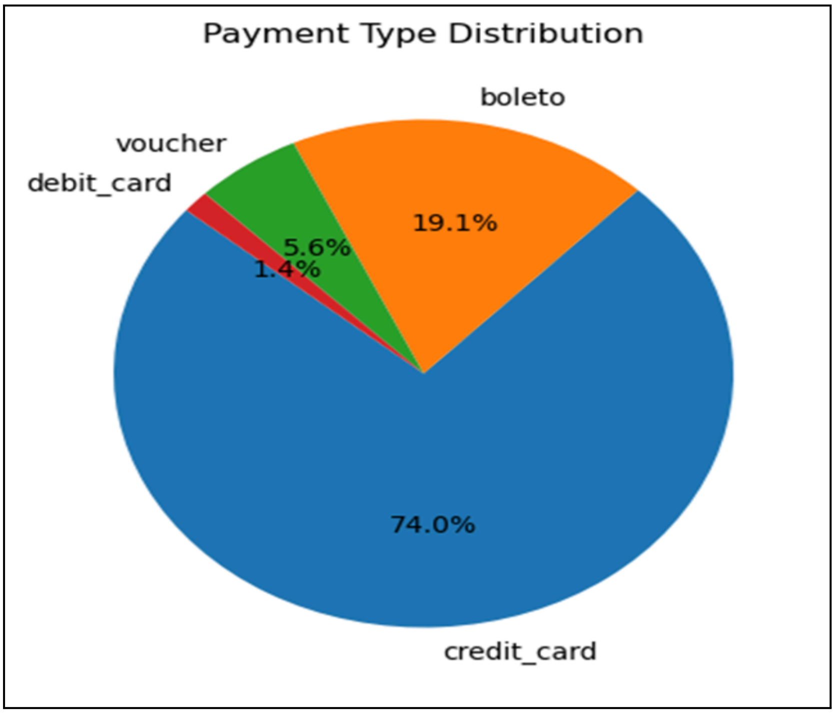


Figure 5, Payments: Credit card is the most frequent payment method, followed by boleto.

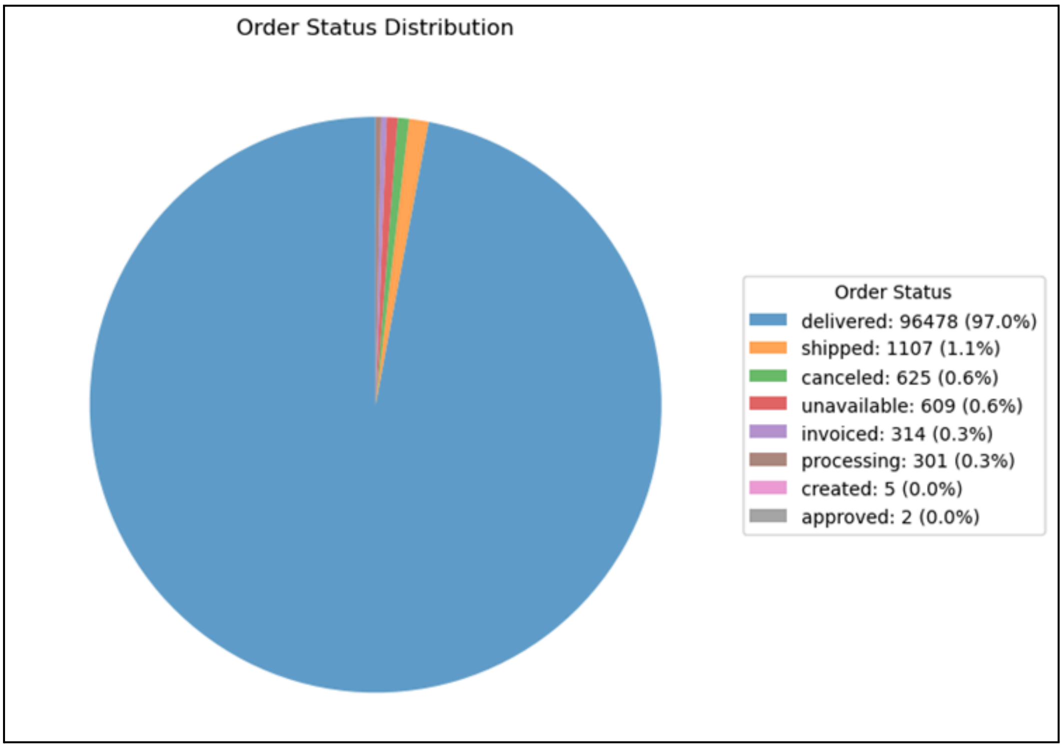


Figure 6, Orders: 96.5% of orders are marked as "delivered".

2.3 Data Preparation

2.3.1 Data Cleaning

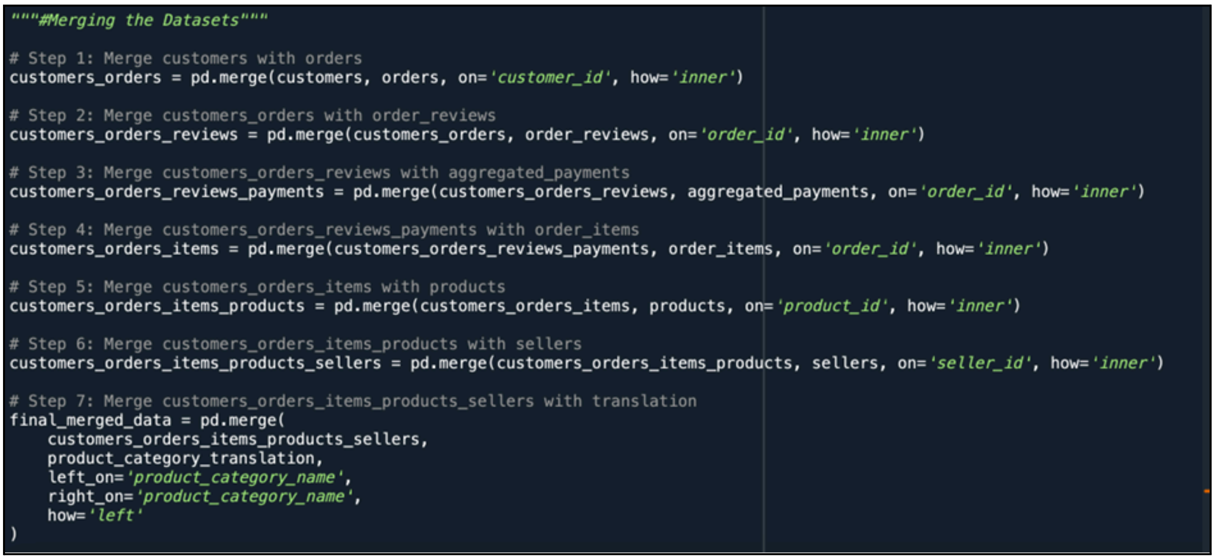
We performed data cleaning before and after merging the datasets to ensure the dataset's accuracy, consistency, and reliability by addressing missing values, resolving inconsistencies, removing duplicates, and filtering out irrelevant/invalid records, making the data ready for meaningful analysis and modeling.

Table 3, Datasets and Cleaning

Dataset	Key Cleaning Steps
Customers	Removed duplicates; Standardised customer_state (uppercase) and customer_city (lowercase); Validated states; No missing values
Geolocation	Removed duplicates; Standardised geolocation_state (uppercase) and geolocation_city (lowercase); Validated latitude/longitude ranges
Order Items	Removed duplicates; Converted shipping_limit_date to datetime; Removed non-positive price/freight_value; Stripped string whitespace
Order Payments	Removed duplicates; Eliminated invalid payment_value/installments/type; Standardised strings; Aggregated rows by order_id
Order Reviews	Removed columns with high missing values; Retained latest review per order_id; Dropped rows with missing critical fields; Converted dates
Orders	Removed duplicates; Dropped rows with missing critical fields; Converted dates; Validated event sequencing (e.g., delivery after purchase)
Products	Removed duplicates; Standardised product_category_name (lowercase); Dropped rows with missing fields; Validated positive numeric fields
Sellers	Removed duplicates; Standardised seller_city (lowercase) and seller_state (uppercase); Checked for missing values

2.3.2 Data Merging

Datasets were merged using key columns to ensure consistency and completeness. Inner joins were used to maintain accurate relationships between datasets, such as linking orders to customers and payments. A left join was applied to the translation dataset to preserve all product details, ensuring no critical information was lost, even if category translations were unavailable.



```python
"""#Merging the Datasets"""

# Step 1: Merge customers with orders
customers_orders = pd.merge(customers, orders, on='customer_id', how='inner')

# Step 2: Merge customers_orders with order_reviews
customers_orders_reviews = pd.merge(customers_orders, order_reviews, on='order_id', how='inner')

# Step 3: Merge customers_orders_reviews with aggregated_payments
customers_orders_reviews_payments = pd.merge(customers_orders_reviews, aggregated_payments, on='order_id', how='inner')

# Step 4: Merge customers_orders_reviews_payments with order_items
customers_orders_items = pd.merge(customers_orders_reviews_payments, order_items, on='order_id', how='inner')

# Step 5: Merge customers_orders_items with products
customers_orders_items_products = pd.merge(customers_orders_items, products, on='product_id', how='inner')

# Step 6: Merge customers_orders_items_products with sellers
customers_orders_items_products_sellers = pd.merge(customers_orders_items_products, sellers, on='seller_id', how='inner')

# Step 7: Merge customers_orders_items_products_sellers with translation
final_merged_data = pd.merge(
    customers_orders_items_products_sellers,
    product_category_translation,
    left_on='product_category_name',
    right_on='product_category_name',
    how='left'
)
```

Figure 7, Data Merging Code

Further Data Cleaning was done by:

1. Filtering Single-Item Orders: Focuses on single-item orders (90.1% of the data) to simplify and standardize the prototype analysis, avoiding complexities introduced by multi-item orders. Orders with multiple items will be included in the formal model for broader analysis.
2. Filtering Delivered Orders: Retains only successfully delivered orders (96.5% of the data) to focus on valid transactions for the prototype.

2.3.3 EDA

After merging the datasets, the unified dataset enabled a more comprehensive analysis of relationships between features and customer satisfaction (review scores).

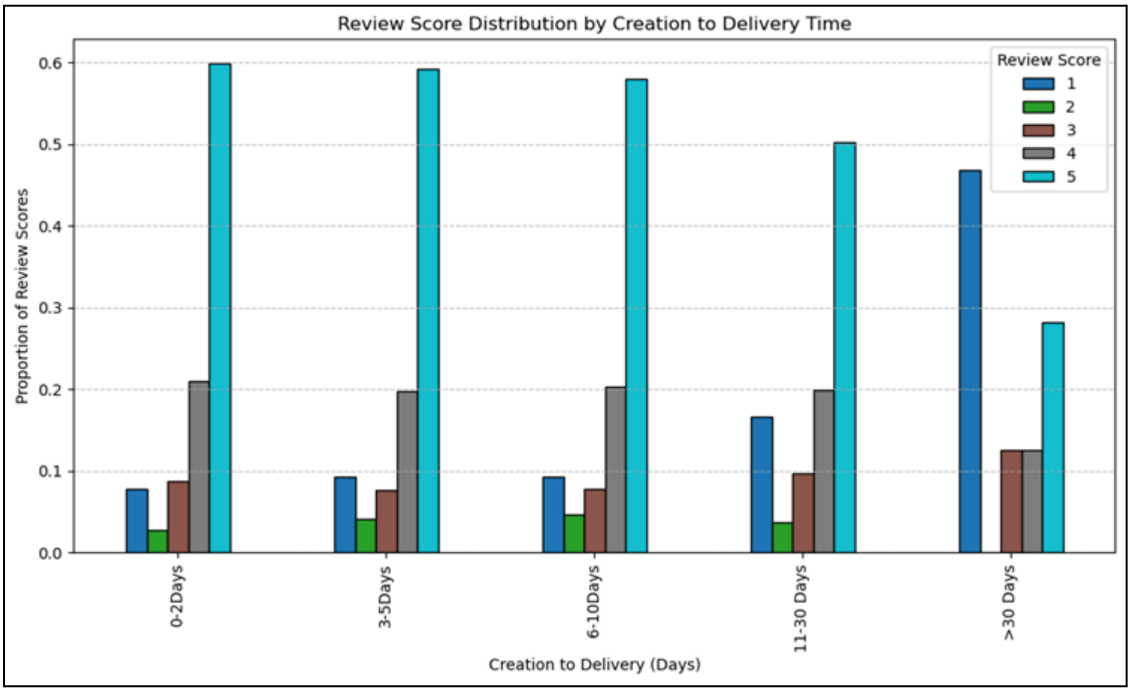


Figure 8, Delivery: The time gap between delivery and review creation impacts scores, with low scores increasing after 10 days.

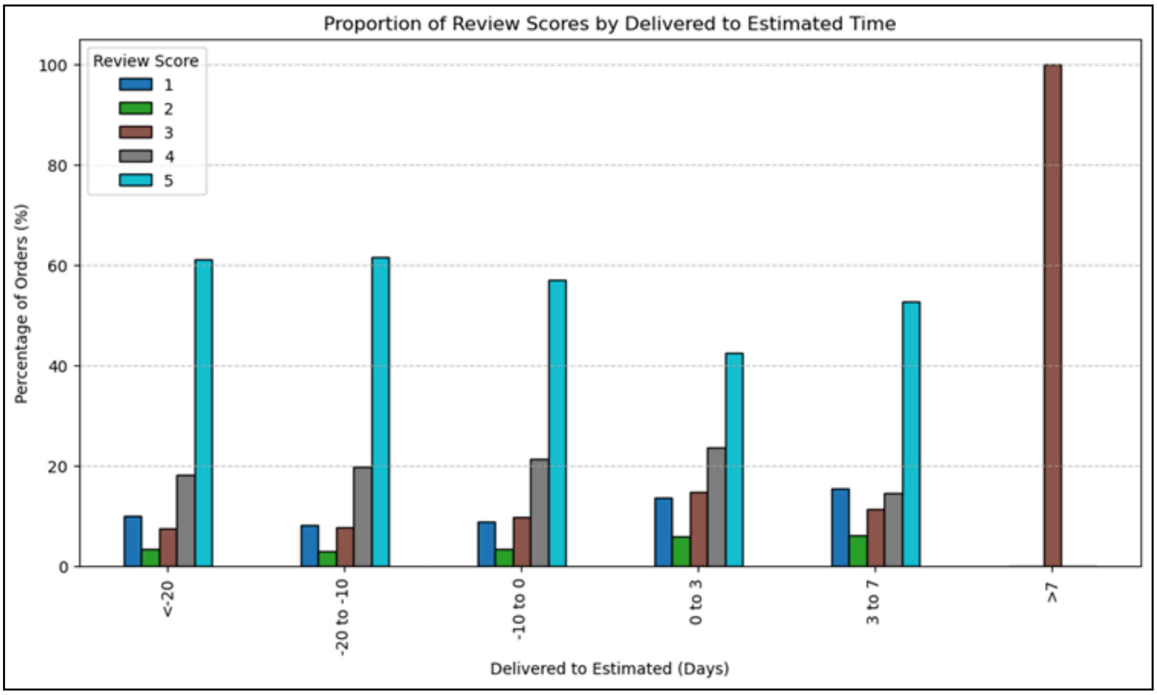


Figure 9, Delivery: Delivery delay has an impact on the rating.

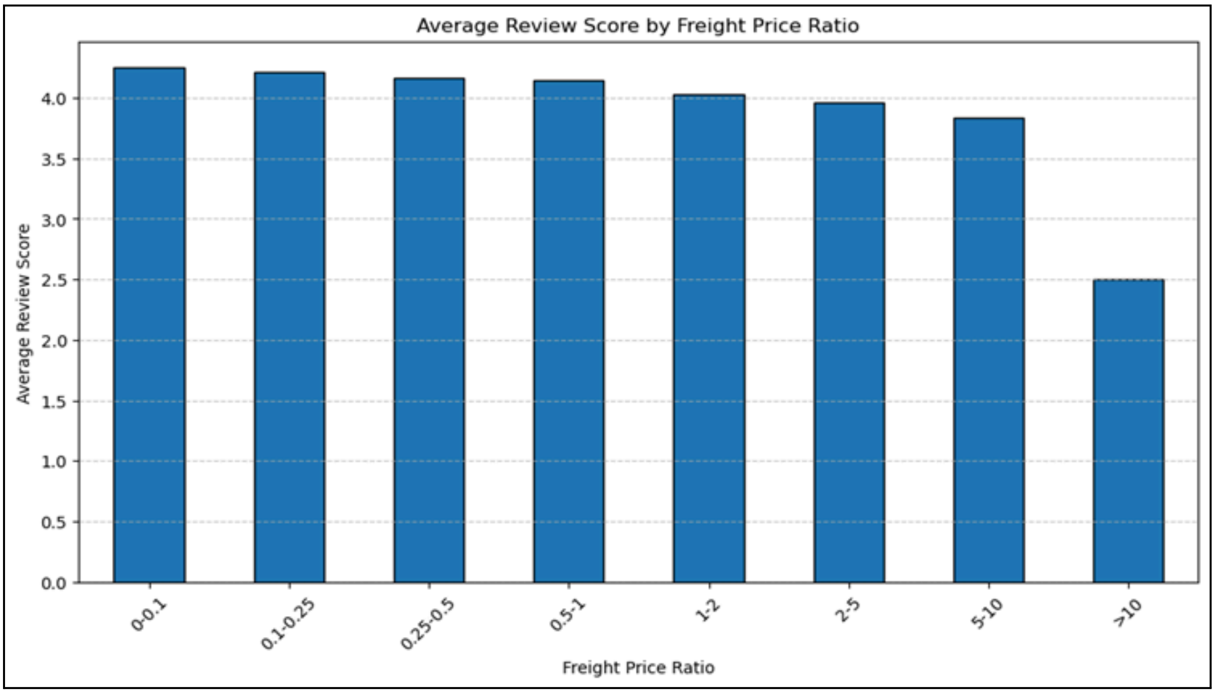


Figure 10, Price: The higher the freight price ratio, the lower the rating.

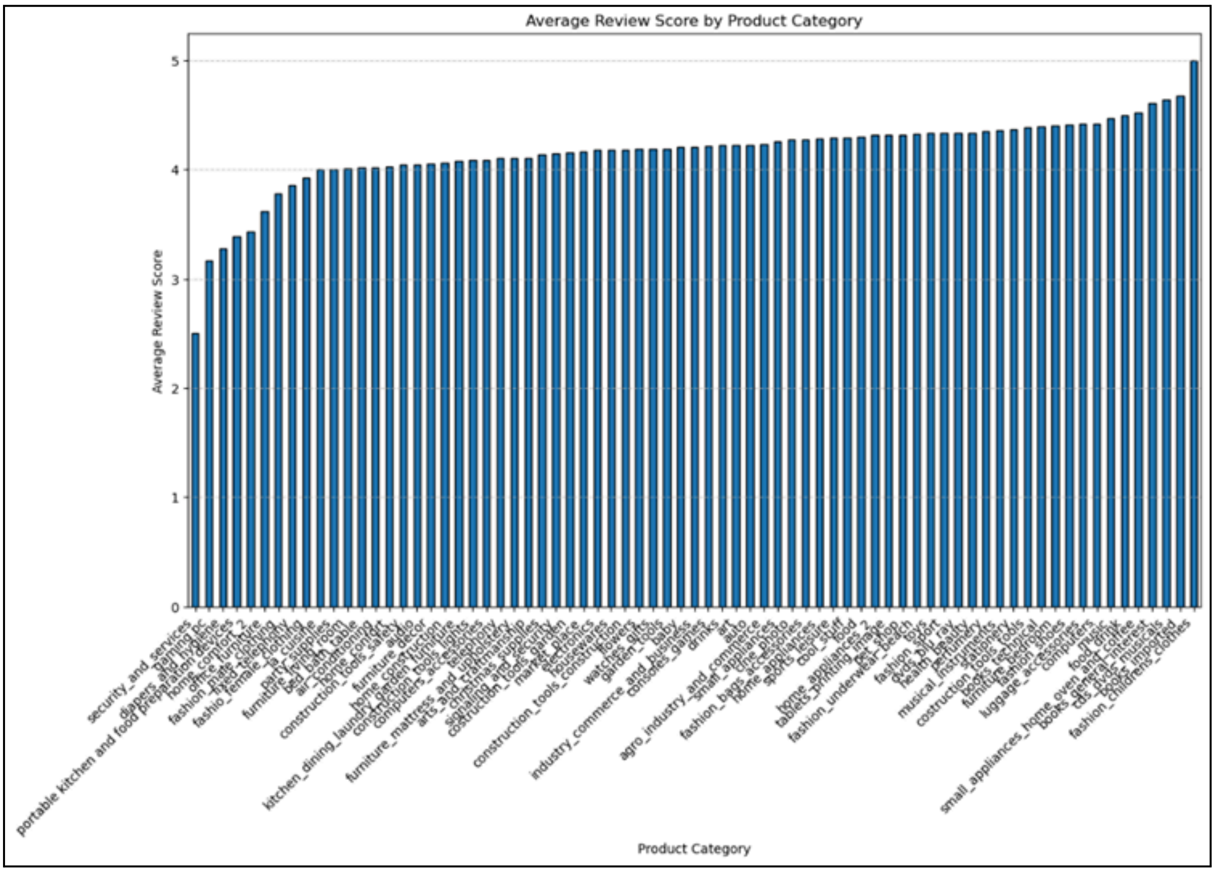


Figure 11, Product: There are differences in ratings across different product categories.

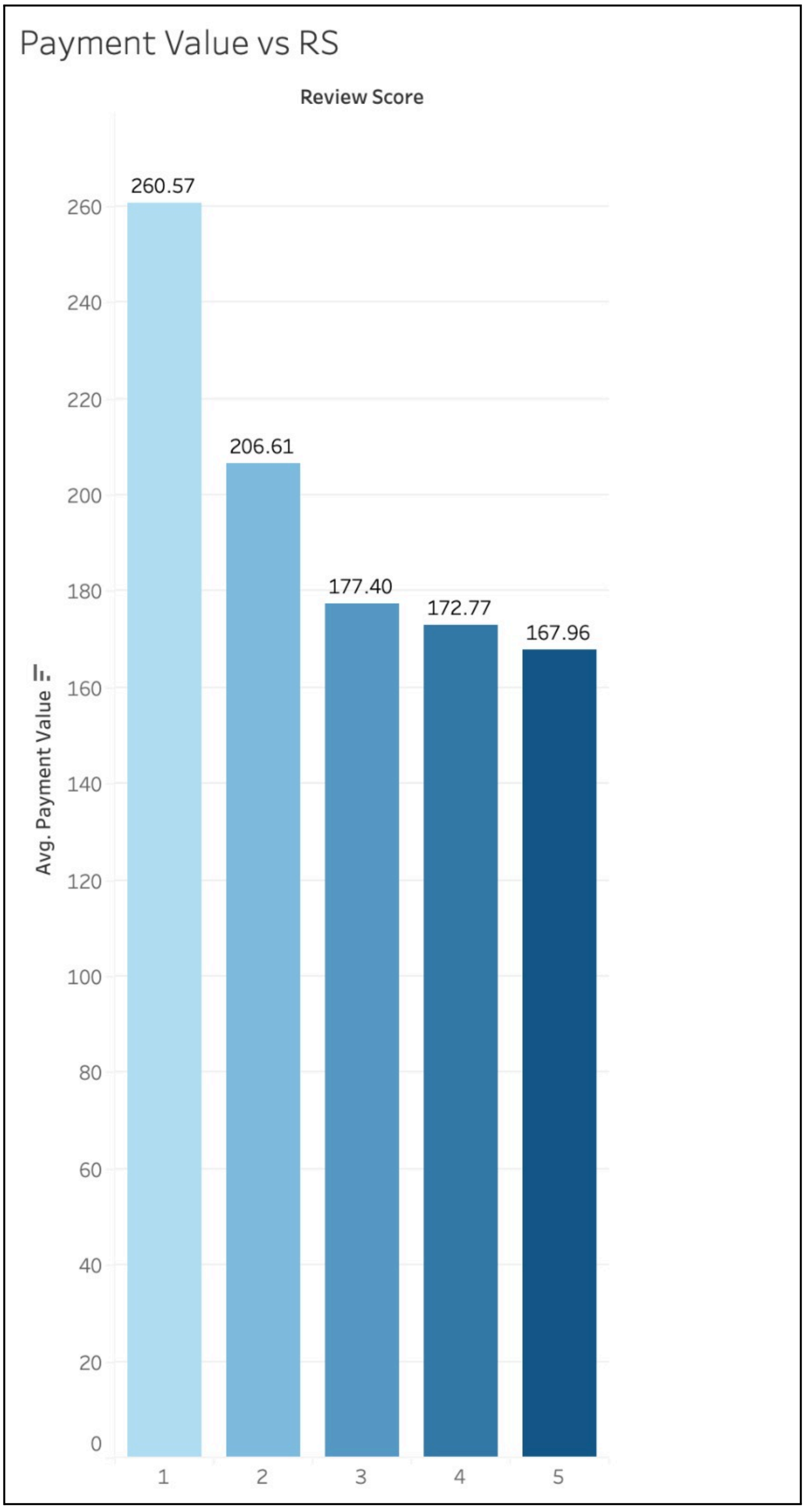


Figure 12, Payment Value vs Review Score: Higher payment values correlate with lower review scores

2.3.4 Feature Engineering

Several new features were engineered to capture key relationships:

- Delivery Time: Time between purchase and delivery
- Delivery Delay: Days the delivery exceeded the estimated delivery date
- Freight Price Ratio: Ratio of shipping cost to product price
- Review to Delivery Time: Time between delivery and review submission

These features provided insights into how costs, delivery performance, and customer responsiveness influenced satisfaction.

2.3.5 Bias-Variance Trade-Off

Feature engineering was designed to balance bias and variance by including predictive features such as freight_value and delivery_time while excluding redundant or highly correlated ones, like payment_value and price.

This approach ensured the model generalized well to unseen data, improving test accuracy while avoiding overfitting.

2.3.6 Feature Selection

Whilst correlation heatmaps and T-tests are primarily suited to identifying linear relationships, they have been utilised in this analysis as an initial feature selection technique along with EDA to highlight potential patterns and relationships, providing a foundation for the application of more complex, non-linear machine learning models.

The below heatmap has been generated to identify the relationships between numerical features as shown on Table 4 (See Appendices).

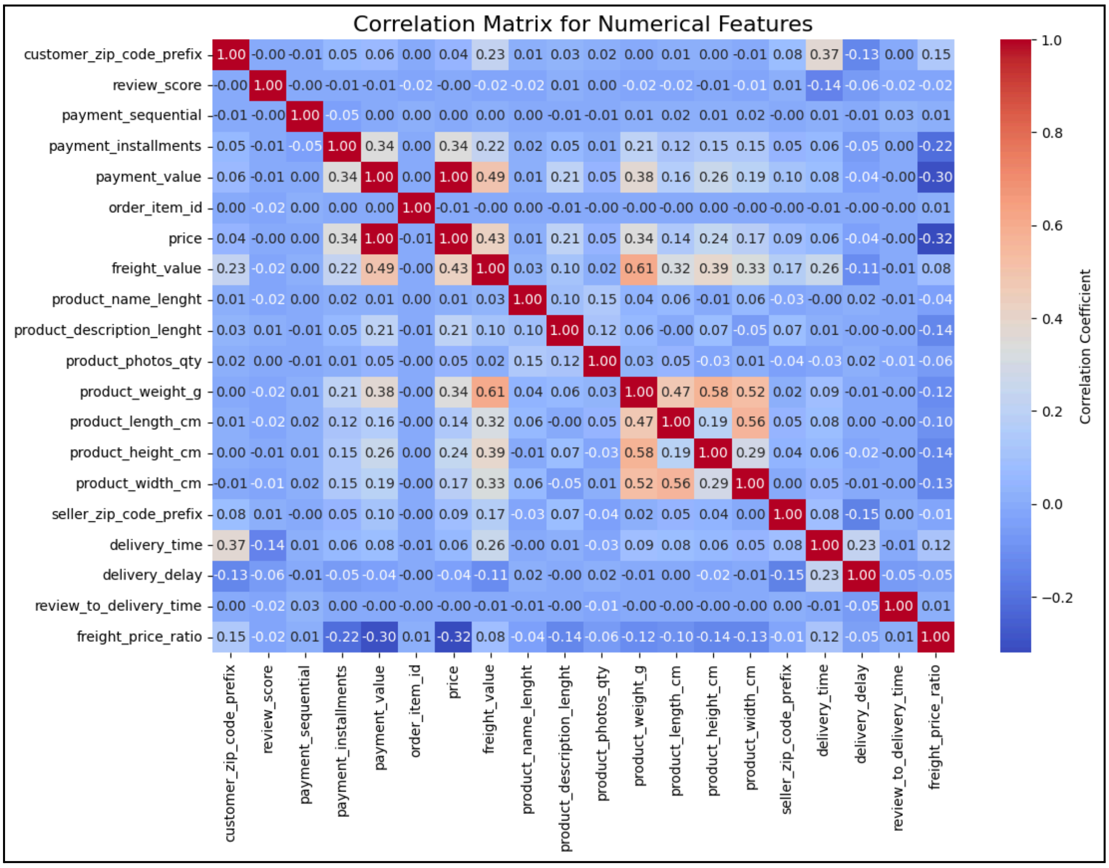


Figure 13, Correlation Matrix for Numerical Features

Table 5, (See Appendices) shows the 12 key features, provides a description of those features and states their significance/inclusion.

To evaluate feature relevance, some T-tests were conducted. To evaluate feature relevance, some T-tests were conducted.

A full summary of T-test results for all features is provided in Table 6 (See Appendices). Figures 14 and 15 illustrate T-tests for Delivery Time and Freight Value, respectively.

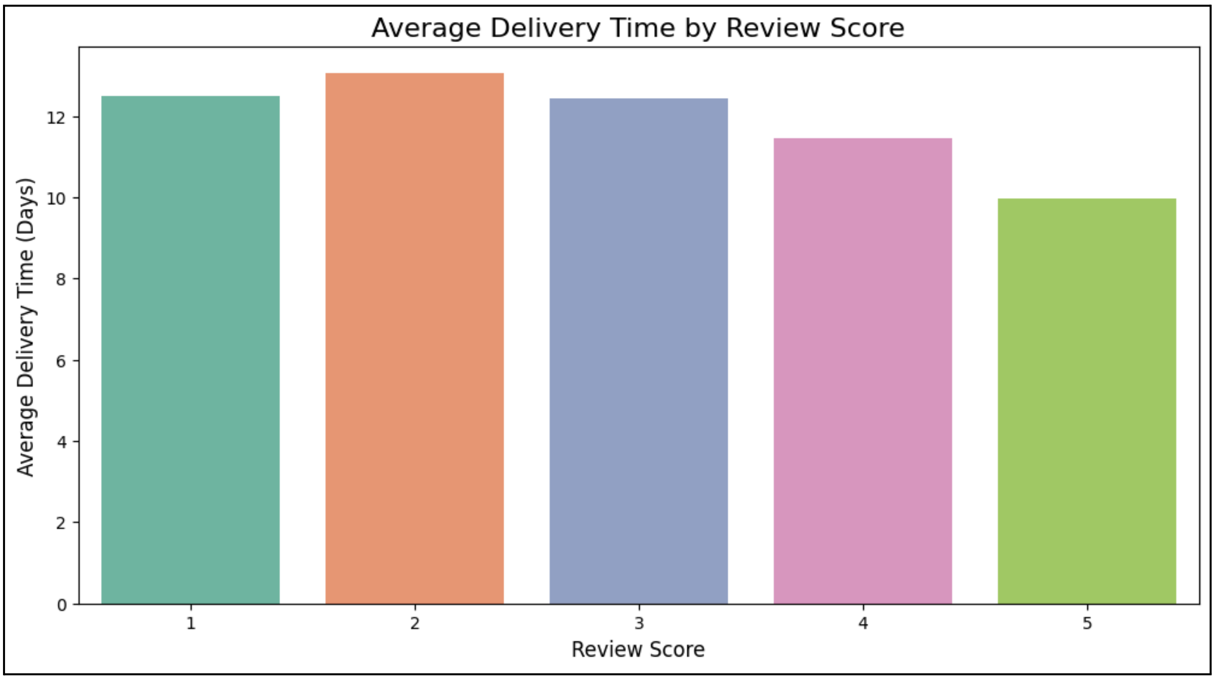


Figure 14, T-Test for Delivery Time

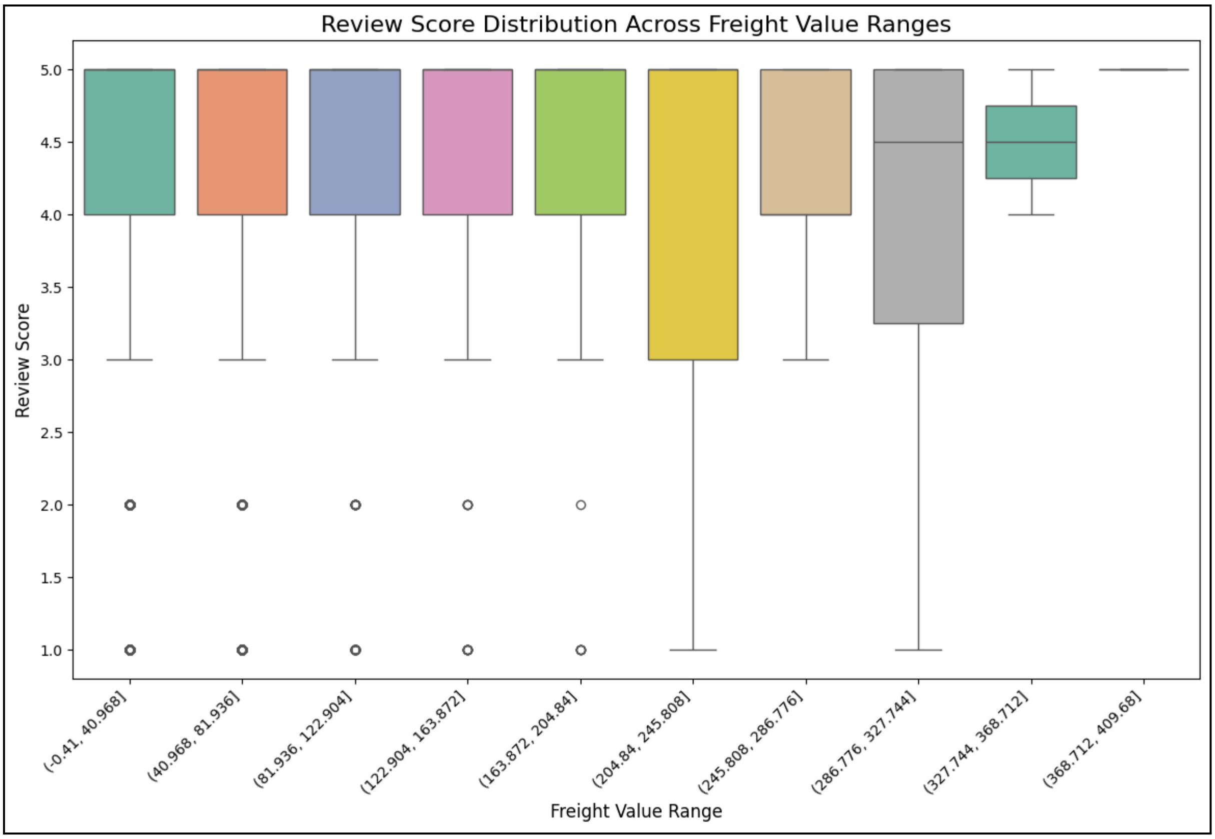


Figure 15, T-Test for Freight Value

After statistical analysis and EDA, the following features were included in the model:

- Customer State
- Payment Type
- Payment Installments
- Freight Value
- Product Category
- Seller State
- Delivery Time
- Delivery Delay
- Freight Price Ratio
- Product Name Length
- Product Weight
- Review to Delivery Time
- Price

The selected features were chosen for their predictive relevance, meaningful relationships with the target variable, and ability to capture key aspects of customer behavior, product characteristics, and transaction details.

Features excluded from the model were determined by:

- Negligible Correlation: Lack of relationship with the target variable.
- Redundancy: High multicollinearity with other features.
- Irrelevance: Unrelated to the business objective.
- Low Variance: Insufficient variability to provide predictive value

2.4 Modeling

2.4.1 Data Splitting

- Objective: To ensure that the model's performance is evaluated on unseen data, the dataset was split into:
    - Training Set: 80% of the data used for model training.
    - Test Set: 20% of the data used for validation.
- Binary Classification - The model uses two classes: positive (4-5 stars) and negative (1-3 stars).

2.4.2 Feature Preparation

- Feature Selection: Key features, including freight_value, delivery_time, and customer_state, were selected based on their statistical significance and business relevance.
- Categorical Encoding: payment_type, customer_state and product_category_name, seller_name were one-hot encoded to make them compatible with machine learning algorithms.
- Scaling: Numerical features were normalized using Min-Max scaling to ensure consistent ranges across features.

2.4.3 Baseline Model Training and Evaluation

Algorithms Used: The below algorithms were chosen to capture diverse patterns in delivery and review data, balancing interpretability and complexity while addressing non-linear relationships and optimizing predictive performance.

- Decision Tree
- Support Vector Machine (SVM)
- Logistic Regression (LogR)
- Random Forest (RF)
- Gradient Boosting Decision Trees (GBDT)
- Extreme Gradient Boosting (XGBDT)

Performance Metrics: Accuracy, precision, recall, and F1-score were calculated for each model to compare performance.

Confusion Matrices: Visualisations of the confusion matrices were generated for all of the ML Algorithms used.

Table 7, Model Performance Metrics

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
Decision Tree	0.7326	0.5154	0.5169	0.5157
SVM	0.8444	0.4222	0.5	0.4578
Logistic Regression	0.8441	0.6098	0.501	0.4604
Random Forest	0.8417	0.5953	0.5061	0.4747
GBDT	0.8444	0.6723	0.5006	0.4592
XGBDT	0.8427	0.6232	0.5076	0.4772

The table above shows the macro accuracy, precision, recall and F1-score for each model trained.

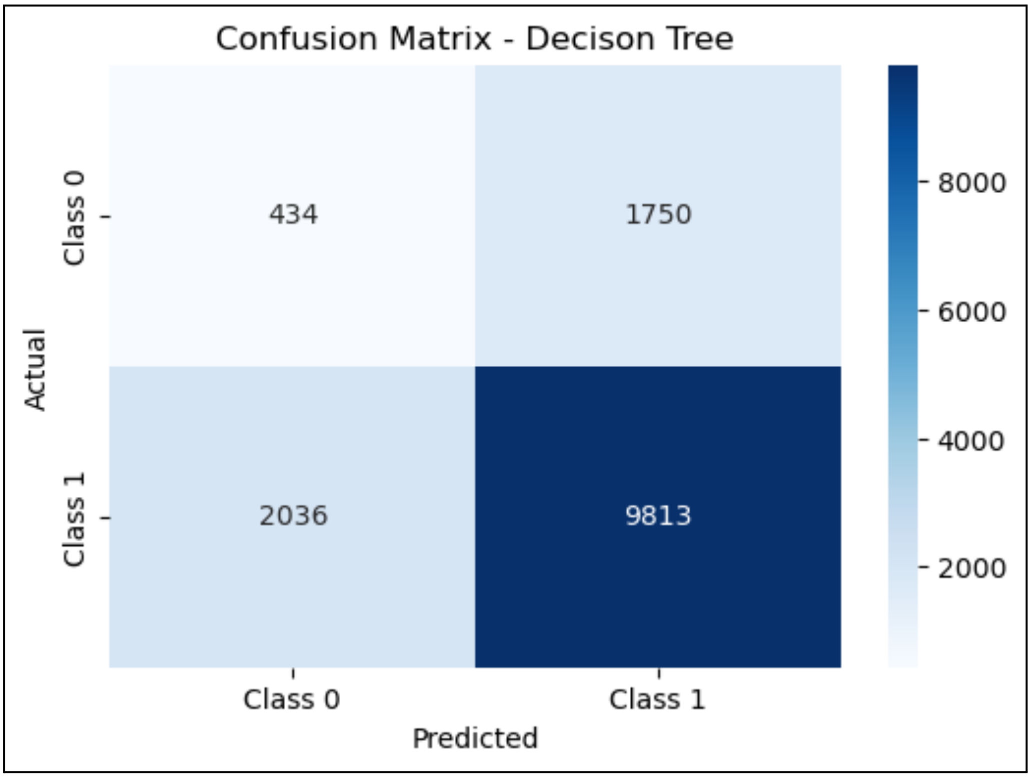


Figure 16, Confusion Matrix - Decision Tree

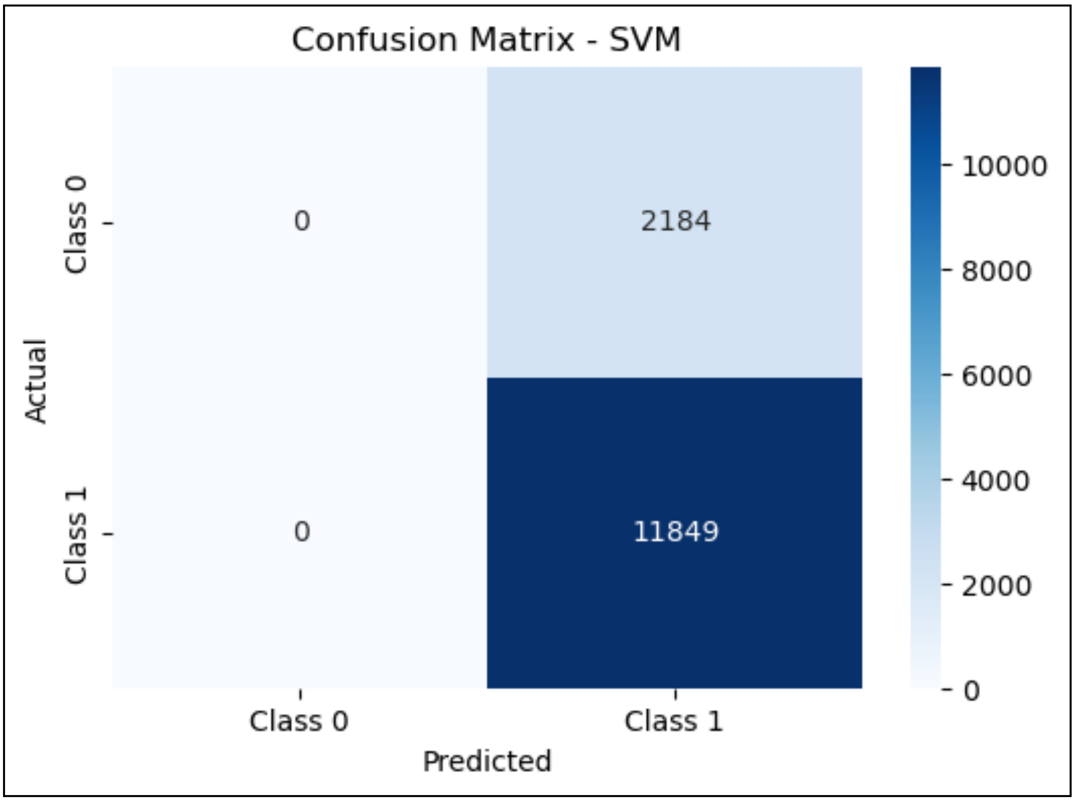


Figure 17, Confusion Matrix - SVM

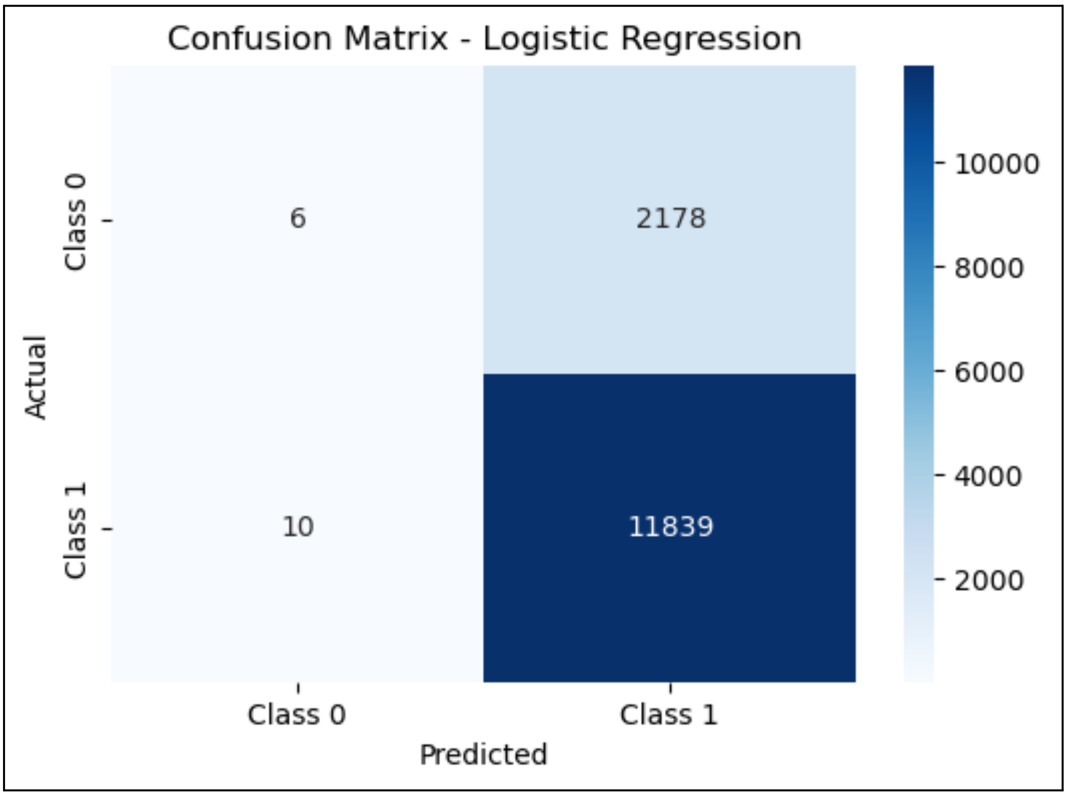


Figure 18, Confusion Matrix - Logistic Regression

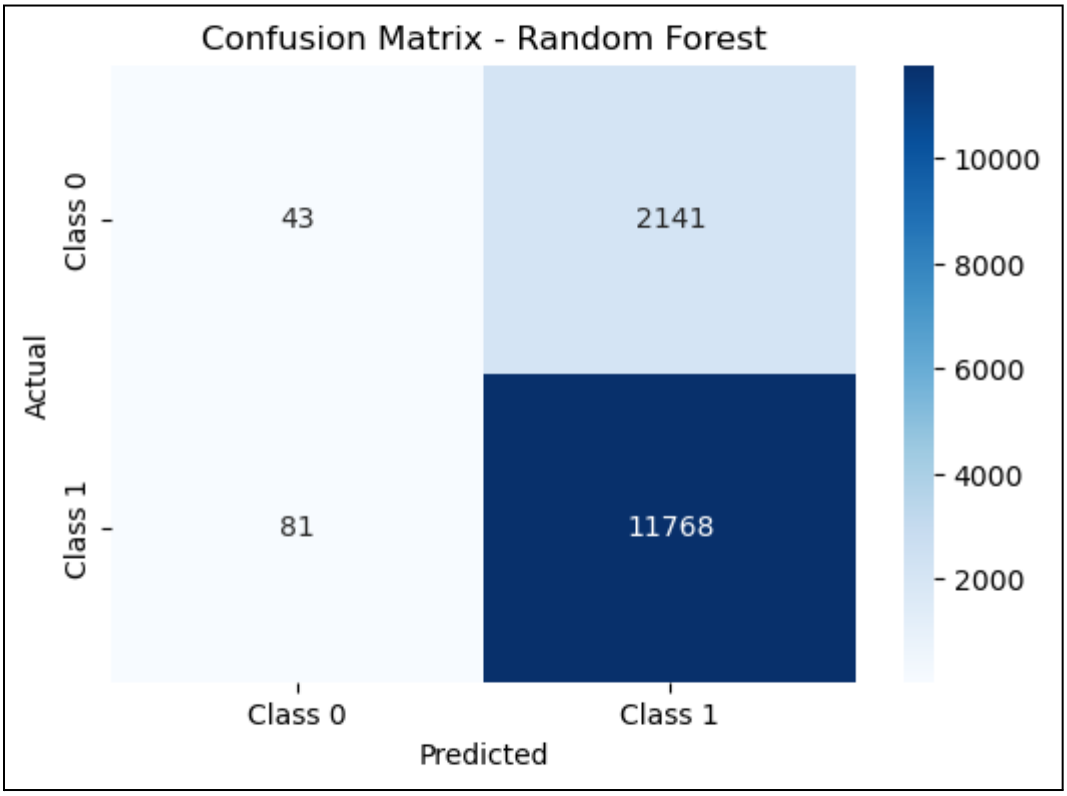


Figure 19, Confusion Matrix - Random Forest

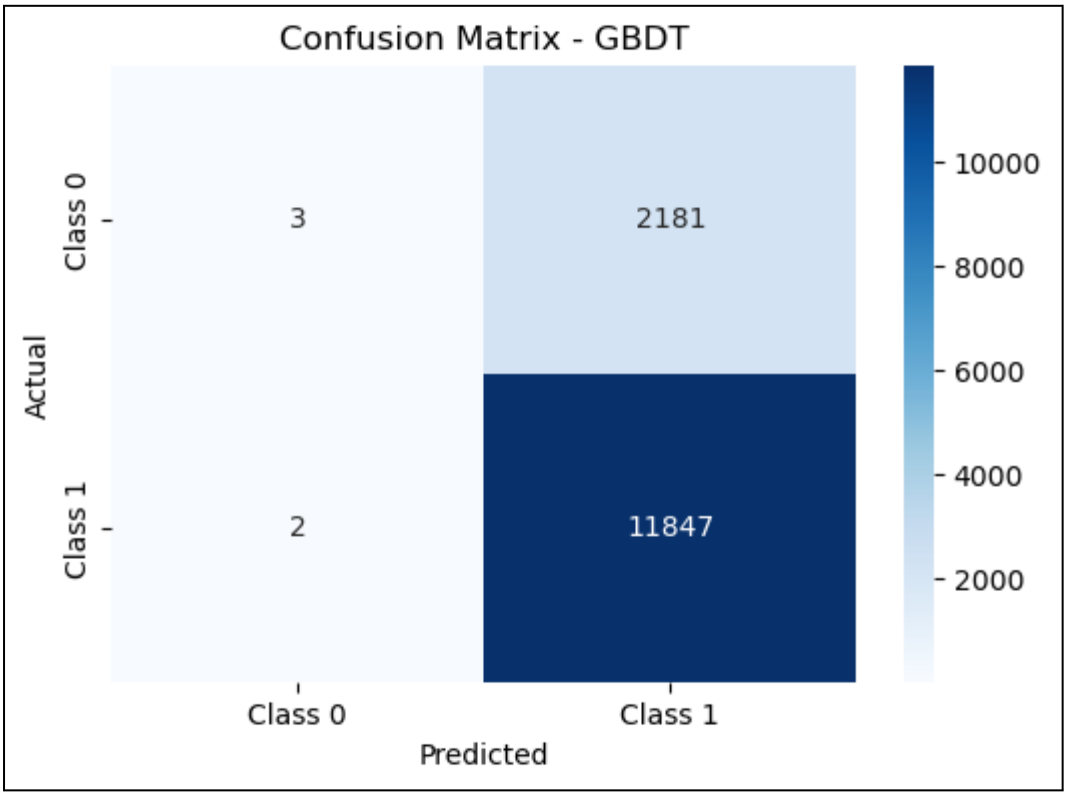


Figure 20, Confusion Matrix - GBDT

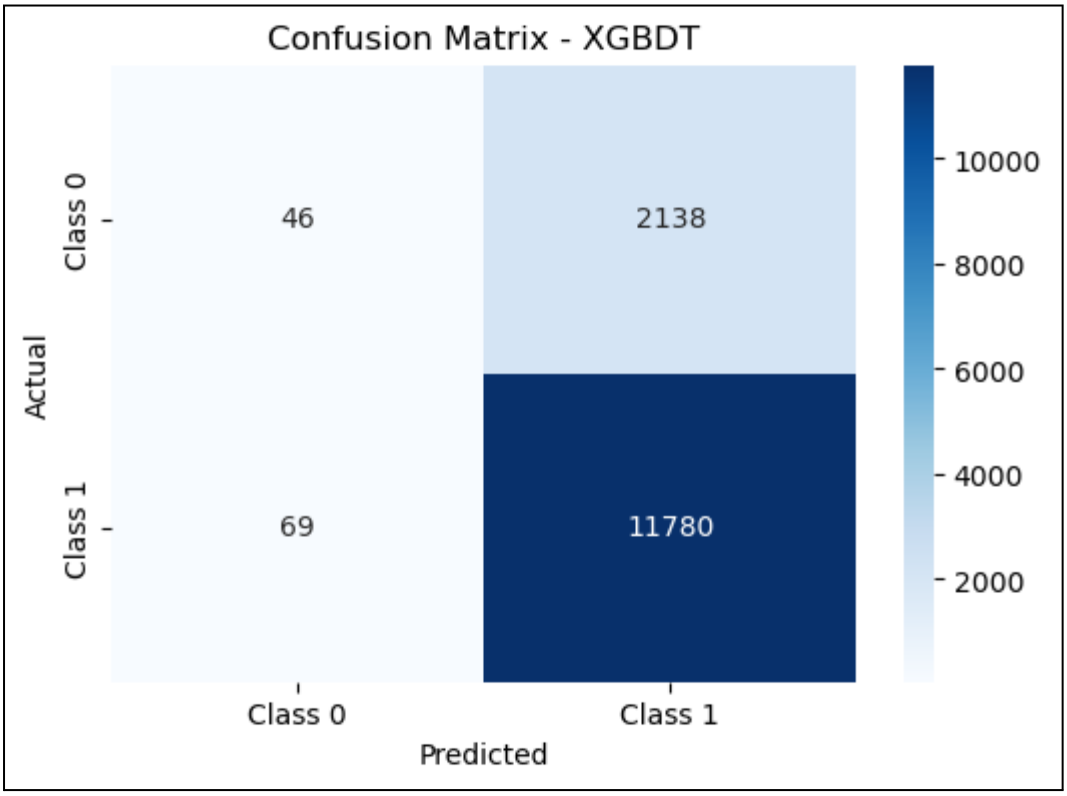
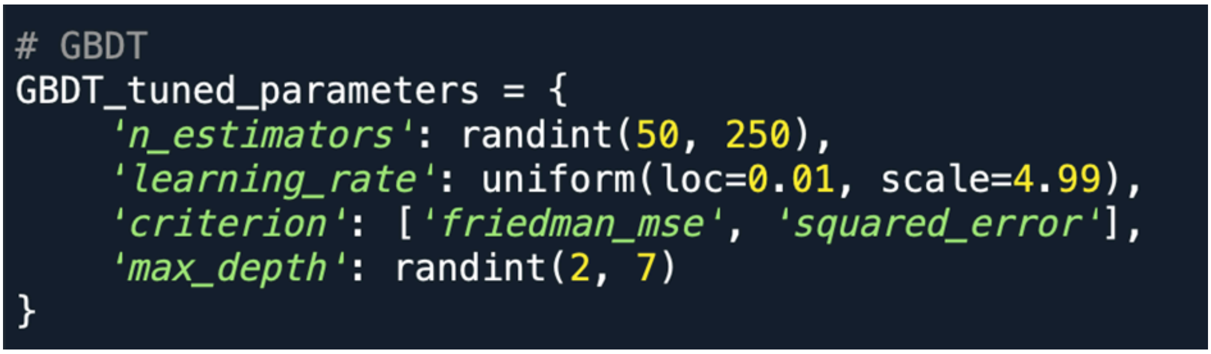


Figure 21, Confusion Matrix - XGBDT

To meet Nile's need for both high predictive accuracy and actionable insights, we selected GBDT. With a macro accuracy of 0.8444 and the highest precision of 0.6723 among tested models, GBDT captures complex relationships while offering interpretable feature importance scores to highlight key drivers of customer satisfaction. Its scalability ensures seamless integration for real-time predictions, making it a robust choice for Nile's objectives.

2.4.4 HyperParameter Tuning

Following the training and evaluation of baseline models GBDT was selected for hyperparameter tuning due to its superior performance compared to other models.



```
# GBDT
GBDT_tuned_parameters = {
    'n_estimators': randint(50, 250),
    'learning_rate': uniform(loc=0.01, scale=4.99),
    'criterion': ['friedman_mse', 'squared_error'],
    'max_depth': randint(2, 7)
}
```

Figure 22, Code for Hyperparameter Tuning

The performance of training data in the model was well, but the performance of the test data was much lower, indicating overfitting. This may be due to an imbalance in the targetvariable's classification (with class 1 representing 84.1% and class 0 representing 15.9%). To address this, we tried both class 1 oversampling using SMOTE and class 0 under sampling using RUS. SMOTE resulted in a slight improvement in test data performance.

2.5 Final Evaluation

Best Parameters: {'criterion': 'squared_error', 'learning_rate': 0.32738057062610587, 'max_depth': 2, 'n_estimators': 162}

Currently, after adjusting the hyperparameters, the model demonstrates strong performance on the training data and good accuracy on the test data. However, there is room for improvement in precision, recall, and F1 score on the test data.

Table 8, Model Performance Metrics after Hyperparameter Tuning

Metric	Training Value	Test Value
Accuracy	*0.8658*	*0.8443*
Macro Precision	*0.9142*	*0.6669*
Macro Recall	*0.7994*	*0.5039*
Macro F1-Score	*0.8286*	*0.4671*

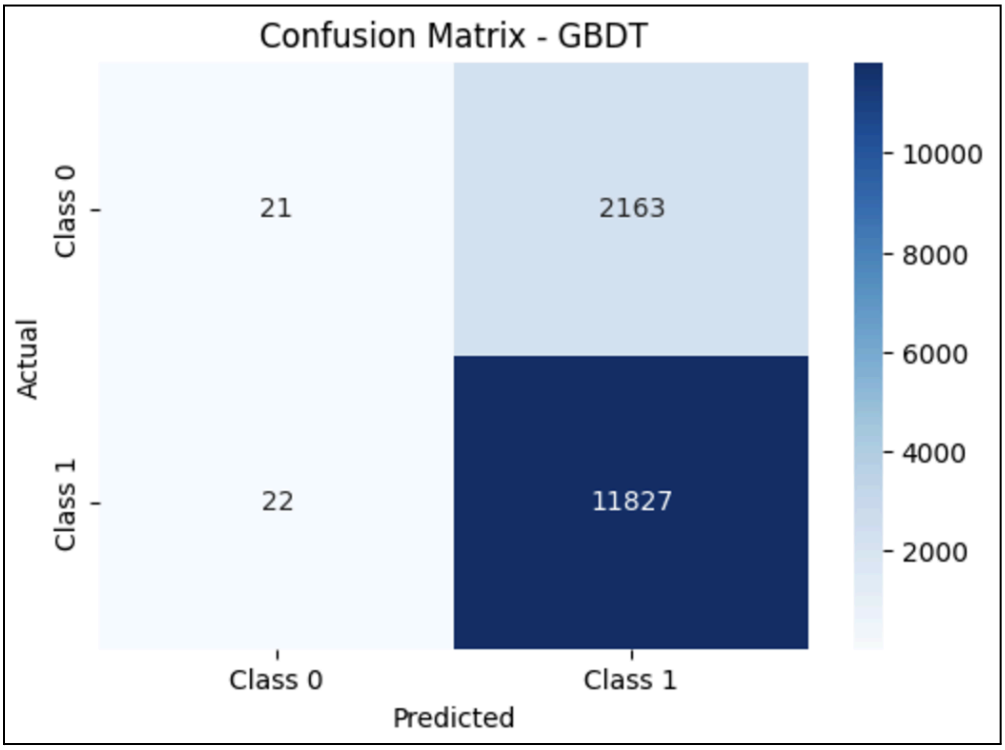


Figure 23, Confusion Matrix GBDT (Optimised)

3.0 Future Work

The future work section outlines potential enhancements and further directions to improve the model's performance and business impact.

3.1 Recommendations

Model Improvement:

- Collaborate with SME's to deepen understanding of the business context, enabling more effective feature engineering and selection.
- Incorporate NLP and deep learning models, such as CNNs, RNNs, and LSTMs, to enhance textual data analysis in future iterations.

Feature Expansion:

- Incorporate demographic data to enhance the model's ability to predict customer review scores.

Data Expansion:

- Increase the volume of the data for the model.

3.2 Challenges and Limitations

1. Data Quality: Missing values, outliers, and discrepancies in merged datasets required imputation, introducing potential bias.
2. Model Performance: Low precision and recall on test data and imbalanced classes (dominance of positive reviews) limited generalisability.

3.3 Deployment

Tableau Dashboard for Visualisation:

- A real-time analytics dashboard will display review prediction results and performance metrics, enabling business teams to monitor predictions and assess customer engagement outcomes effectively.

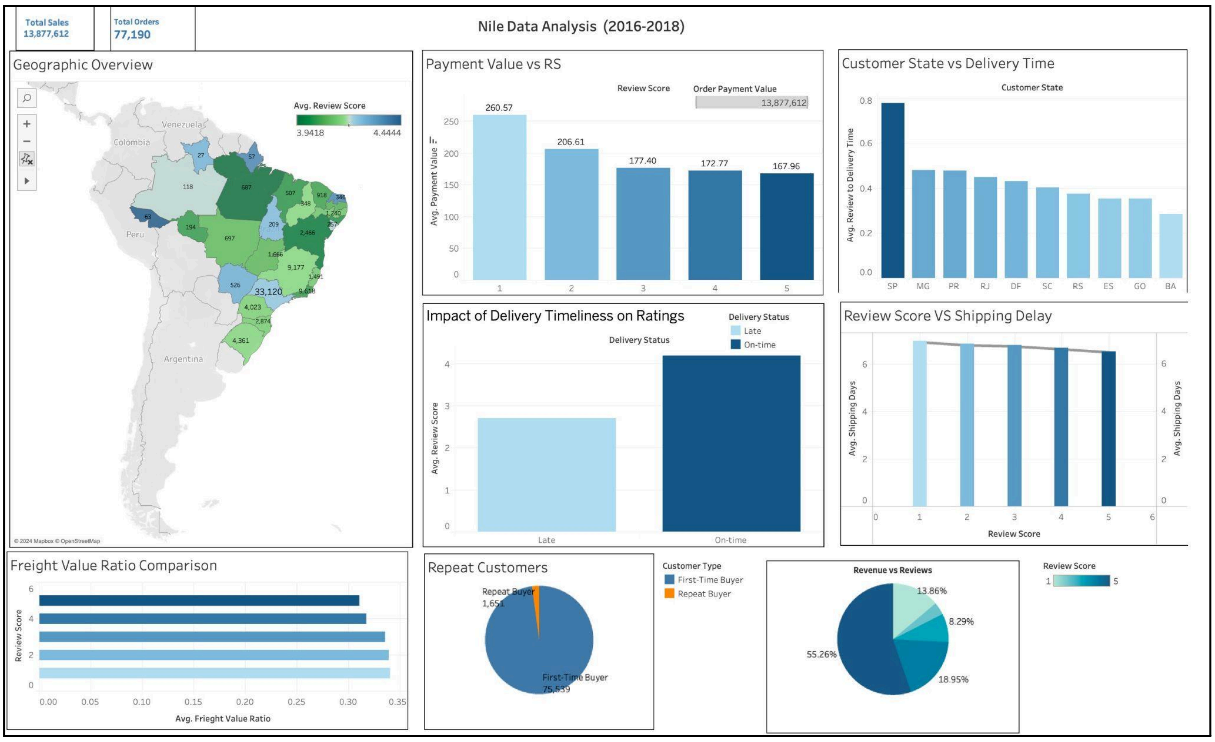


Figure 24, Analytics Dashboard - Tableau

4.0 Conclusion

The project successfully developed a predictive model that aligns with Nile’s objectives of improving customer review engagement and satisfaction. While the model demonstrates strong potential; addressing challenges like data quality, model refinement, and operational adoption will enhance its impact. Future work should focus on refining the model, enriching the dataset, and ensuring seamless integration into Nile’s operations for long-term success.

5.0 Appendices

Table(s) 1, Data Dictionary

olist_order_reviews_dataset	
Column Name	**Description**
review_id	Unique review ID
order_id	Unique order ID
review_score	1-5* (star rating given by the customer)
review_comment_message	Any comments left by the customer (in Portuguese)
review_creation_date	The date the survey (email notification) was sent to the customer
review_answer_timestamp	When the customer completed the review

olist_customers_dataset	
Column Name	**Description**
customer_id	Unique ID related to the order. Each new order generates a unique customer ID even if this customer has bought from the store before
customer_unique_id	Unique ID (one per customer) linked to their profile and order history
customer_zip_code_prefix	First five characters of their zip code (post code)
customer_city	The name of the city the customer lives in (Brazilian)

customer_state	The state which the customer lives in (Brazilian)

olist_geolocation_dataset	
Column Name	**Description**
geolocation_zip_code_prefix	The first five digits of the zip code
geolocation_lat	Latitude
geolocation_lng	Longitude
geolocation_city	The name of the city the customer lives in (Brazilian)
geolocation_state	The state which the customer lives in (Brazilian)

olist_order_dataset	
Column Name	**Description**
order_id	Unique ID of the order
customer_id	An ID unique to each order. This provides access to the unique customer ID in the customer dataset
order_status	Delivered, shipped, etc.
order_purchase_timestamp	The time of the purchase.
order_approved_at	The time the payment was authorised
order_delivered_carrier_date	When the order was posted

order_estimated_delivery_date	Show the date the customer was advised that the order would be delivered at the time of purchase
order_delivered_customer_date	When the order was delivered

olist_order_items_dataset	
Column Name	**Description**
order_id	Unique ID of the order
order_item_id	Each item purchased in the same order
product_id	The unique ID of each product purchased
seller_id	The unique ID of the seller
shipping_limit_date	The date the seller will send the order to the logistic partner
price	Item price
freight_value	The item freight value (if an order has multiple items the freight value is split between them)

olist_payments_dataset	
Column Name	**Description**
order_id	Unique ID of the order
payments_sequential	A customer may pay with more than one method. This is the sequence of payments.

payment_type	Method of payment
payment_installments	If the payment is in instalments (multiple payments over time) this is the number of instalments.
payment_value	Transaction value

olist_product_dataset

Column Name	Description
product_id	Unique ID of the product
product_category_name	Name of the product category
product_name_lenght	Number of characters extracted from the product name
product_description_lenght	Number of characters extracted from the product description
product_photos_qty	Number of product photos online
product_weight_g	Product weight in grams
product_length_cm	Product length in centimetres
product_width_cm	Product width in centimetres

olist_sellers_dataset

Column Name	Description

seller_id	Unique ID of the seller
seller_zip_code_prefix	The first five digits of the seller’s zip code (post code)
seller_city	City where the seller is based
seller_state	State where the seller is based

Table 2, Initial EDA - Key Observations

Category	Key Insights
Customers	96.9% are one-time buyers, indicating limited repeat business opportunities. São Paulo dominantes customer concentration, followed by Rio de Janeiro. São Paulo has the highest number of customers.
Items	90.1% of orders contain only a single item. Prices are typically under 200, and freight costs are under 60. No clear linear relationship between price and freight cost, with some freight costs exceeding item price.
Payments	Credit cards are the most frequent payment method, followed by boleto. Most payments are made in one installment, with values typically under 300.
Orders	96.5% of orders are marked as "delivered," ensuring focus on completed transactions.
Products	Product names are generally between 30 to 65 characters. Descriptions vary widely, with most being under 2,000 characters. Most products weigh less than 1.5 kg, reflecting typical eCommerce goods.
Sellers	Seller distribution mirrors customer distribution. São Paulo is the leading location for sellers.

Table 4, Feature Selection Insights

Category	Feature	Correlation Coefficient	Interpretation
Weak Correlation with Review Score	Delivery Time	-0.14	Longer delivery times tend to result in lower review scores.
Weak Correlation with Review Score	Delivery Delay	-0.11	Delays beyond the estimated delivery date negatively impact review scores.
Strong Correlations Between Features	Freight Value and Price	0.61	Higher-priced products often incur higher shipping costs.
Strong Correlations Between Features	Product Dimensions (e.g., product_length_cm, etc.)	High Correlation	These features are highly correlated with one another, suggesting multicollinearity.
Insights for Feature Selection	Delivery Time, Delivery Delay, Freight Price Ratio	Low Correlation	Included in the model due to logical connection to customer satisfaction.
Insights for Feature Selection	Product Dimensions	High Correlation	May require dimensionality reduction or careful selection to avoid redundancy.

Table 5, Feature Significance

Feature Name	Description	Included	Reasoning
Customer State	State of the customer for the order.	No	Weak correlation with review scores despite being significant in t-test.
Payment Type	Type of payment used (e.g., credit card, boleto).	No	Statistically significant impact on review scores (p-value < 0.05).
Payment Installments	Number of payment installments for the order.	No	Significant relationship with review scores in t-test (p-value < 0.05).
Freight Value	Cost of shipping for the order.	Yes	Key cost metric with a strong logical relationship to customer satisfaction.
Product Category	Category of the purchased product.	No	Minimal impact on review scores and weak statistical significance.
Seller State	State where the seller is located.	No	No significant relationship with review scores.
Delivery Time	Time taken to deliver the order (in days).	Yes	Significant negative correlation with review scores and strong business relevance.
Delivery Delay	Days the delivery was delayed beyond the estimated delivery date.	Yes	Statistically significant impact on review scores, indicating customer dissatisfaction.

Freight Price Ratio	Ratio of freight value to product price.	Yes	Captures the relative shipping cost, which strongly influences customer satisfaction.
Product Name Length	Length of the product name (in characters).	No	No logical or statistical connection to review scores.
Product Weight	Weight of the product (in grams).	No	High correlation with other product dimensions; excluded to reduce multicollinearity.
Review to Delivery Time	Time between delivery and review submission.	Yes	Strong logical connection to customer response behavior.

Table 6, T-Test Results

Feature	T-Statistic	P-Value	Significance
Payment Type	2.2517	0.0243	Yes
Payment Installments	2.0226	0.0431	Yes
Price	0.4355	0.6632	No
Freight Value	3.6114	0.0003	Yes
Product Category	4.1599	0.0000	Yes
Seller State	-2.6754	0.0075	Yes
Delivery Time	25.5840	0.0000	Yes
Delivery Delay	11.6150	0.0000	Yes
Freight Price Ratio	-2.8628	0.0042	Yes
Product Name Length	2.6111	0.0090	Yes
Product Weight	2.5718	0.0101	Yes
Review To Delivery Time	3.8301	0.0001	Yes