5663833, 5587796, 5630106, 5663354, 5635111, 5595277

WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

# Masters Programmes: Group Assignment Cover Sheet

| | |
|---|---|
| **Student Numbers:**<br>Please list numbers of all group members | **5663833, 5587796, 5630106, 5663354, 5635111, 5595277** |
| **Module Code:** | **IB9BW0** |
| **Module Title:** | **Analytics in Practice** |
| **Submission Deadline:** | **2 Dec 2024** |
| **Date Submitted:** | **2 Dec 2024** |
| **Word Count:** | **1990** |
| **Number of Pages:** | **11** |
| **Question Attempted:**<br>*(question number/title, or description of assignment)* | **Empowering Nile's Marketing:**<br>**Predictive Modelling for Customer Reviews** |
| **Have you used Artificial Intelligence (AI) in any part of this assignment?** | **No** |

**Academic Integrity Declaration**

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

**Upon electronic submission of your assessment you will be required to agree to the statements above**

## Contents

## 1. Introduction

The aim of this project is to develop a machine-learning model that predicts the likelihood of receiving positive customer reviews for an e-commerce company. Following the Cross Industry Standard Practices for Data Mining (CRISP-DM), the report ensures a systematic and practical approach to data mining and model development (Provost and Fawcett, 2013).

## 2. Methodology: CRISP-DM

### 2.1 Business Understanding

Nile, a leading e-commerce platform in Brazil, aims to develop a predictive model to identify customers likely to leave positive reviews. Since customer reviews influence purchasing decisions and the company's reputation, Nile's aim is to create a cost-effective model for encouraging positive feedback, which can help Nile improve its online reputation and drive sales growth. With key stakeholders including Nile's management, marketing department, customers, and potential product sellers, the successful completion of this project will improve customer reviews and support Nile's overall business goals.

### 2.2 Data Understanding

Since this is Nile's first attempt at using a predictive model, the strategy has been designed to test and deploy it as cost-effectively as possible, establishing a solid foundation for future investigations into more intricate data and advanced models. Given the data from Nile's database system, including 8x tables exported to CSV, and a file to help translate the product categories into English, the analysis at an early stage is centred on specific data variables (on time, delivery time, total value, freight percentage) directly affecting the review score.
As a result, for this project, only three datasets — order_reviews, order, and order_items are utilized and new features for further analysis have been developed.

| **Datasets Retained** | **Order_reviews** | **Orders** | **Order_items** |
|---|---|---|---|
| **Columns Retained** | order_id<br>review_answer_timestamp<br>review_score | order_status<br>order_id<br>order_purchase_timestamp<br>delivery_customer_date<br>order_estimated_delivery_date | order_id<br>frieght_value<br>price |

*Table 1. Datasets and Columns Utilisation for Project.*

During review, some NA values and incorrect data types are found which are to be fixed in the cleaning process. NA values need to be removed, while data types need to be converted from object to datetime for date-related data, and to category data types for other variables. Furthermore, it is confirmed that all datasets contain the crucial column "order_id," which is needed to merge the datasets.

## 2.3 Data Preparation

### 2.3.1 Data Cleaning

Figure 1 shows that most of our data correspond to delivered orders, thus only this subset of the data is retained as it would provide the most accurate and reliable data. The unneeded variables for the analysis are eliminated to guarantee clarity when using the data in the modelling.

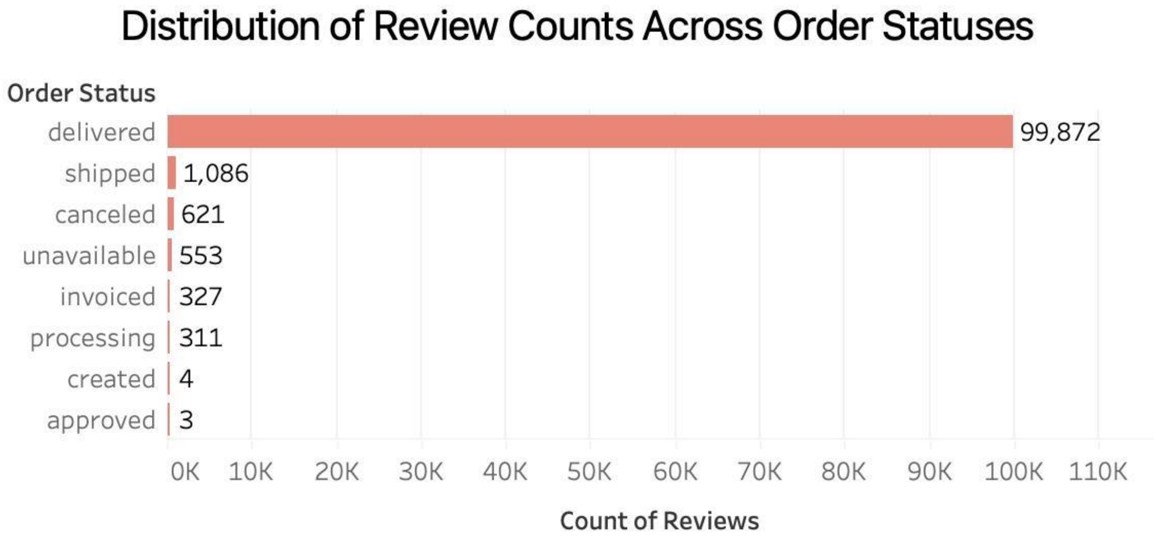

*Figure 1. Distribution of Review Counts Across Order Status.*

All date-related fields are converted to DateTime format and the order_id to a categorical type. Data aggregation is done by using ‘groupby’ function, to ensure that no duplicate data exist in order_id. After grouping and removing duplicates, it was found that less than 1% of the dataset had missing values, so all NA entries were confidently removed without concern.

### 2.3.2 Feature Engineering & Further Exploratory Data Analysis

The calculated fields are shown in Table 2. The ‘total value’ replicated the payment_value from the excluded payment dataset, ensuring consistency and minimal redundancy. freight_percentage enabled the sensitivity and impact of a high freight proportion on the review score to be analysed.

| Calculated Fields | Description |
|---|---|
| on_time | if estimated _delivery >= delivered_customer_date (True / False) |
| delivery_time (in days) | delivered_customer_date - purchase_timestamp |
| total_value | freight_value + price |
| freight_percentage | freight_value / total value * 100 |

*Table 2. New Features for Modelling*

'Delivery_time' and 'on_time' proved to be critical in identifying delays and trends in customer satisfaction. Once the new features are created, the unnecessary variables are removed, but the necessary columns are retained to keep the dataset clean for modelling.

Also, the outliers are identified by distributing a histogram of delivery time (as shown in Figure 2). To maintain data quality and improve the model's reliability, extreme delivery times — likely due to data entry errors or extraordinary cases — are excluded from the analysis.

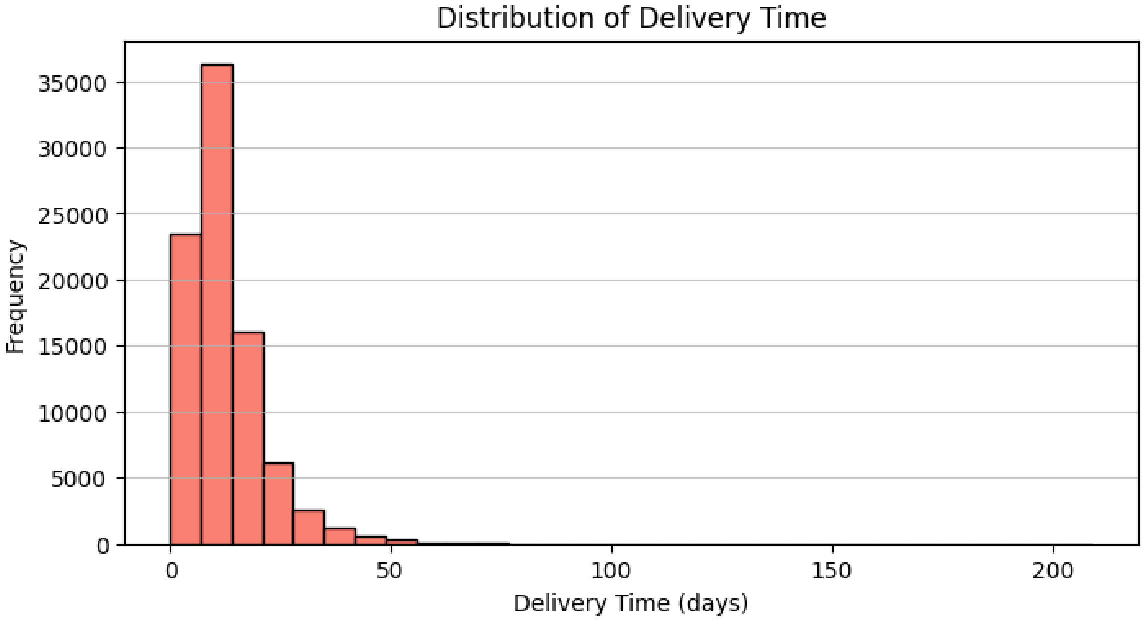

*Figure 2. Distribution of Delivery Time.*

The correlation coefficients to understand the relationships between independent variables and the review scores are calculated due to which features significantly impacting customer satisfaction are identified. Additionally, a heatmap created to visualize these features and their correlations, made it easy to identify the relationships.

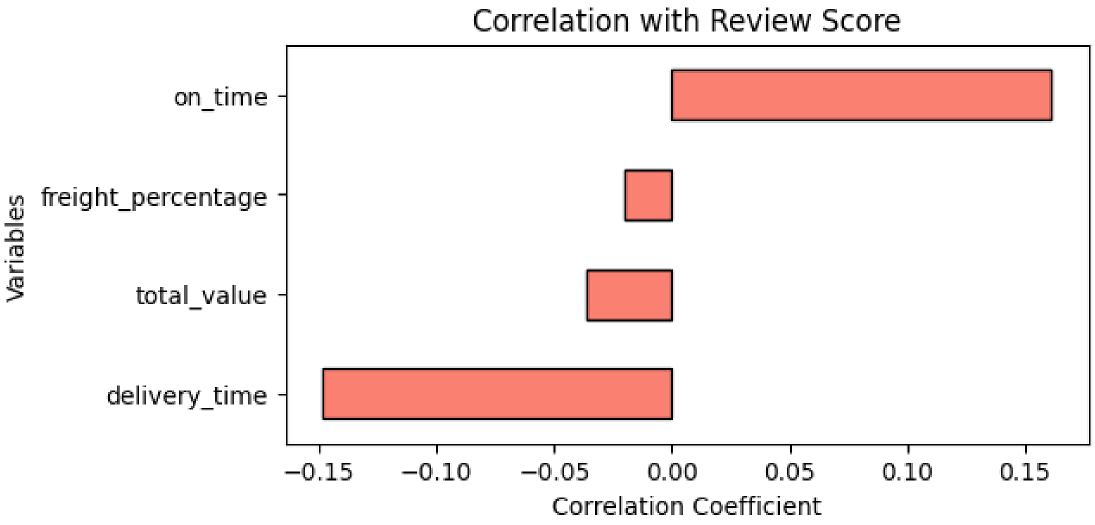

*Figure 3. Correlation with Review Score.*

By preparing these valuable features and conducting exploratory data analysis (EDA), a solid foundation for modelling has been established, enabling valuable metrics to be derived and actionable insights to be provided.

## 2.4 Modelling

### 2.4.1 Model Selection & Design

Four ML algorithms - Decision Tree, Random Forest, Gradient Boosting (GBDT), and XGBoost (XGBT) are tested to determine the most effective approach. The multi-class (review score 1, 2, 3, 4, 5) model, binary model (review scores 1-3, 4-5) and three-class model (review scores 1-2, 3, 4-5) were explored for the analysis.

### 2.4.2 Model Performance Assess & Comparison

The findings in Table 3 show that the binary classification consistently outperformed the other approaches, achieving the highest accuracy of 82%, particularly with the GBDT and XGBT models. Despite having high accuracy, the low recall scores indicate the challenge in identifying true positives which may have resulted from class imbalance. The binary GBDT model was ultimately chosen for its balanced performance across key metrics showing reliability and resilience in predicting positive reviews. This model has moderate precision in predicting positive and negative classes, a relatively low ability to capture all positive and negative samples with balanced performance and room for improvement.

| Model | Class | Accuracy | Macro train | Macro test |
|---|---|---|---|---|

| | | (test) | Precision | Recall | F1 Score | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | 1-3,4-5 | 0.82 | 0.71 | 0.53 | 0.51 | 0.53 | 0.53 | 0.53 |
| Random Forest | 1,2,3,4,5 | 0.49 | 0.94 | 0.85 | 0.91 | 0.22 | 0.21 | 0.20 |
| | 1-2,3,4-5 | 0.75 | 0.96 | 0.9 | 0.93 | 0.36 | 0.35 | 0.35 |
| | 1-3,4-5 | 0.76 | 0.91 | 0.93 | 0.95 | 0.55 | 0.53 | 0.53 |
| Gradient Boosting (GBDT) | 1,2,3,4,5 | 0.61 | 0.80 | 0.22 | 0.19 | 0.28 | 0.21 | 0.18 |
| | 1-2,3,4-5 | 0.81 | 0.80 | 0.36 | 0.35 | 0.44 | 0.35 | 0.33 |
| | 1-3,4-5 | 0.81 | 0.72 | 0.53 | 0.51 | 0.70 | 0.53 | 0.51 |
| XGBT | 1-2,3,4-5 | 0.81 | 0.37 | 0.37 | 0.45 | 0.45 | 0.35 | 0.34 |
| | 1-3,4-5 | 0.82 | 0.54 | 0.54 | 0.69 | 0.69 | 0.53 | 0.51 |

*Table 3. Model Performance Findings.*

## 2.5 Evaluation

The confusion matrix shows that the GBDT model is underpredicting the class 0. There are 3072 examples in the test, but it only predicts 399.

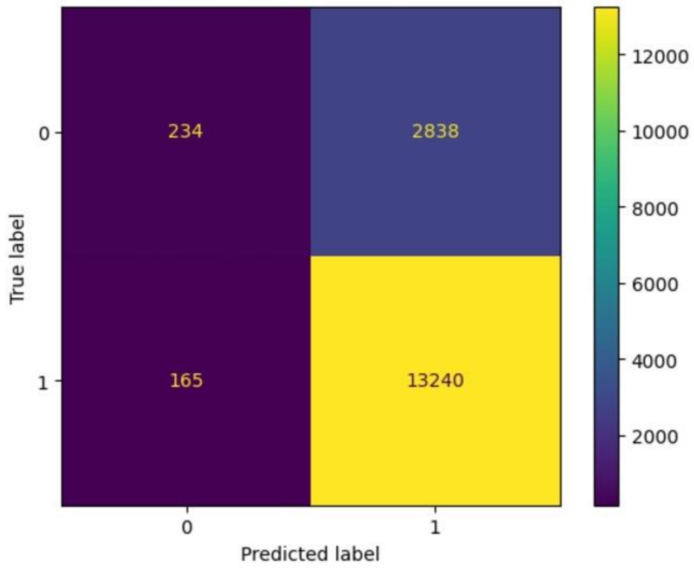

*Figure 4. Confusion Matrix for GBDT Model.*

Observation of training and testing metrics (accuracy, macro recall, precision and F1-score) in Table 3 indicates minimal variance, which suggests that the model may be experiencing underfitting (as shown in Figure 5). This warrants further investigation into model complexity.

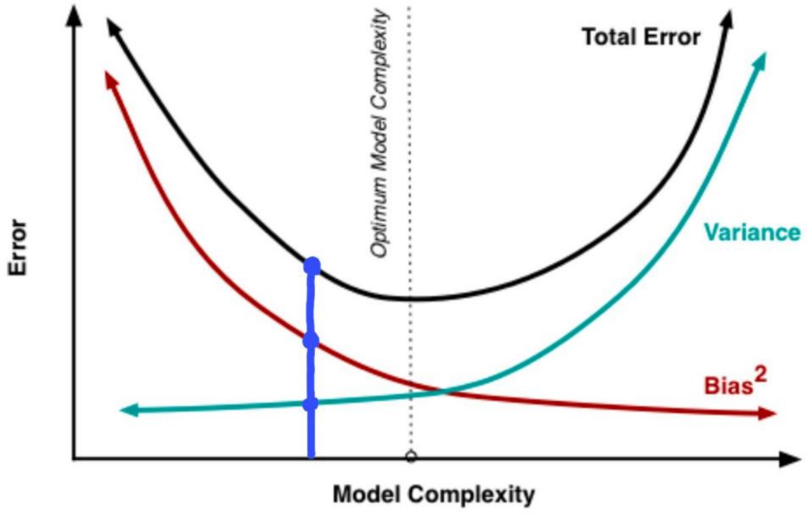

*Figure 5. The Bias-Variance Trade-Off.*

As shown in Table 4, the analysis indicates that despite implementing all four strategies, the recall and F1 score for the minority class (class 0) remains significantly low and the troubleshooting method has not yielded improvements for underprediction.

| Troubleshooting | Class 0 | | | Class 0 | | |
|---|---|---|---|---|---|---|
| **Options** | **Precision** | **Recall** | **F1 Score** | **Precision** | **Recall** | **F1 Score** |
| Hyperparameter Tuning | 0.25 | 0.01 | 0.02 | 0.81 | 0.99 | 0.89 |
| Biased Threshold | 0.16 | 0.30 | 0.21 | 0.80 | 0.62 | 0.70 |
| Undersampling | 0.50 | 0.10 | 0.17 | 0.82 | 0.97 | 0.89 |
| Oversampling | 0.59 | 0.07 | 0.12 | 0.82 | 0.98 | 0.89 |

*Table 4. Troubleshooting Findings of GBDT Model.*

Given the underfitting limitations of this model, there are some suggestions for future analyses. Firstly, the model can be refined by extracting features from existing databases that have a significant impact on reviews (as shown in Table 5). Once the data engineering is reconstructed, the degree of model fit will determine whether or not it is necessary to proceed with adding or removing attributes from the existing datasets. Additionally, the constraints in the data filtering stage can be lifted to expand the detection scope; for instance, in this prediction, only the order status of delivered is considered, whereas it can be later extended to all the categories in this item by one hot encoding. Furthermore, in addition to the existing dataset, extra data related to review scores, such as whether a product meets customer expectations, can be requested from Nile. Eventually, the adoption of two more modelling approaches, stacking, and voting, can better cope with the complexity of the ensemble model, consolidating the strengths of all the models and thus boosting the prediction accuracy.

| Features | Dataset |
|---|---|
| Payment Type | Payments_dataset |
| Payment Installments | Payments_dataset |
| Average Review Score by Product Category | Order_review_dataset, Products_dataset, Product_category_name_translation |
| Product Dimensions (length and breadth) | Products_dataset |
| Product Photo Quantity | Products_dataset |

*Table 5. Features to be Added in the Future.*

Analyses of Figure 6 detailing performance for a model predicting weekly aggregated binary review scores over 2017 and into early 2018 reveal significant insights into the model's performance in terms of accuracy and behaviour. First of all, there is a huge contrast between actual scores, which vary between 0.75 and 0.9, and the model's predicted scores, almost unnaturally high and stuck close to 1 — this shows a big mismatch. That implies that this model has a bias to overestimate positive reviews, and such happens probably due to overfitting or its inability to grasp complexity in data. Large differences show for the middle of 2017 where actual scores fall but predictions remain high. However, for the period between January to August of 2018, predicted scores began to descend into the range of actual scores and, therefore became more pattern recognition capable. The sudden decline in September 2018 of both scores may reflect either the actual trend of the data or anomalies, hence the need for further calibration and selection of features to add to the model for more accurate prediction. In general, analyses indicate that further improvement in the model is important in capturing the real behaviour of the data.

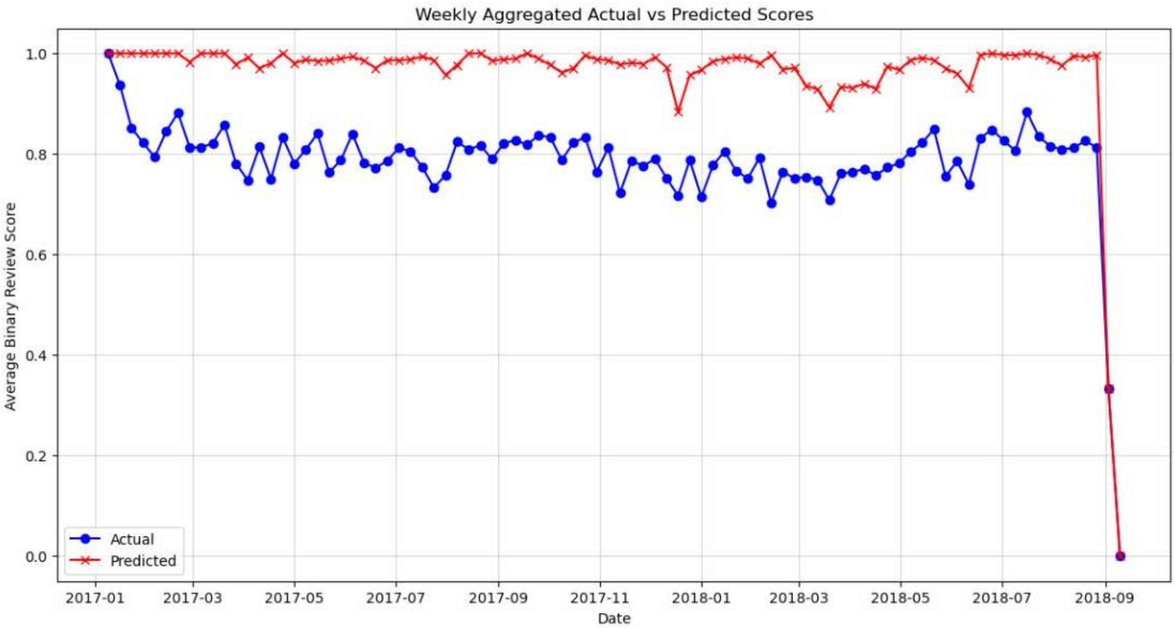

*Figure 6. Weekly Aggregated Actual Versus Predicted Scores.*

Class imbalance can increase the variance of the train-test split due to the inconsistent representation of minority classes. The application of n-fold cross-validation repeated 50 times helps reduce this variance by giving a variety of training and validation subsets for the extraction of more robust and reliable performance evaluations of machine learning models (Aliferis and Simon, 2024).

## 2.6 Deployment

A comprehensive deployment plan will be developed to implement the predictive model within the Nile platform that integrates with existing management (CRM) and enterprise resource planning (ERP) systems to facilitate seamless data exchange. This plan will encompass several key components, starting with training sessions for marketing and customer service teams to ensure effective utilization of the model and a support system to address any issues that may arise during its operation. Additionally, implementing monitoring systems not only continuously tracks the model's performance but also conducts regular maintenance to ensure its accuracy over time. A feedback mechanism will also be established to gather users' insights, allowing for continuous improvement of the model and its deployment.

This process will prioritize customer privacy, adhering to data protection regulations such as the General Data Protection Regulation (GDPR), which requires explicit user consent for data collection. Robust security measures will be implemented to protect customer data from unauthorized access. Furthermore, by integrating with other systems and business processes, the predictive model can be used to enhance cross-functional collaboration within the Nile platform, ultimately achieving greater efficiency, better decision-making, and increased profitability. The model will aid in building customer profiles and gathering feedback, thereby enhancing customer satisfaction, increasing sales revenue, and reducing churn rates. However, it is essential to acknowledge that implementing this model may raise several ethical concerns, particularly regarding customer privacy. Ensuring that customers are fully informed about data collection practices and that the platform's privacy policy explicitly addresses these practices will be critical. Other important ethical considerations include maintaining the authenticity of reviews generated through incentivized methods and ensuring robust data security measures are in place.

## 3. Conclusion

Four methods were evaluated in this study, which centred on a dataset generated from order_id. Consequently, troubleshooting techniques (Options 1 through 4) were used to enhance the model's functionality. The gradient boosting model was ultimately selected, however, underprediction and underfitting were discovered. It is advised to enhance the data balance and add a number of features in order to fix this. In virtue of these discoveries, more initiatives will be recommended to improve the model's predictive power.

5663833, 5587796, 5630106, 5663354, 5635111, 5595277

## References

Aliferis, C. and Simon, G. (2024) ‘Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI’, in Simon, G.J. and Aliferis, C. (eds) *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences*. Health Informatics. Switzerland: Springer, Cham, pp. 477-524.

Provost, F. and Fawcett, T. (2013) *Data science for business: what you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.