Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	5614252, 5646570, 5652824, 5662617
Module Code:	IB9BW0
Module Title:	Analytics in Practice
Submission Deadline:	2nd Dec at 12:00:00
Date Submitted:	2nd Dec
Word Count:	1,967
Number of Pages:	12
Question Attempted: *(question number/title, or description of assignment)*	Review Prediction: eCommerce Platform
Have you used Artificial Intelligence (AI) in any part of this assignment?	Yes

Content

1.Introduction

Customer reviews play significant role in the success of online retail platforms, influencing customer trust and purchasing decisions. However, effectively generating and improving these reviews remains a challenge. This report outlines a targeted model approach to predict customers most likely to leave positive reviews. The proposed solution entails the development and deployment of prediction model of which customers will leave positive reviews using advanced data mining in the dataset of 89,999 review scores. Through this approach, the report aims to uncover factors influencing review scores and strategically target customers while minimizing unnecessary costs.

2.Methodology

This report used "Cross-Industry Standard Process for Data Mining (CRISP-DM)" to structure the analytics project.

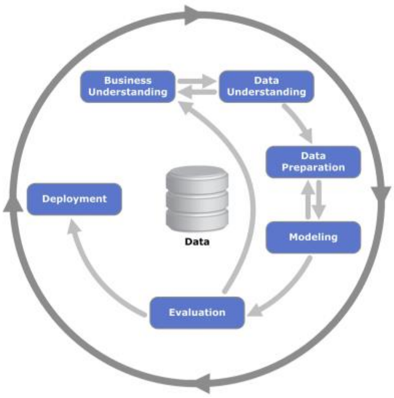


Figure 1

2.1 Data Understanding and Preparation

Initially, the dataset was analyzed to identify columns relevant to review score. The report focuses on analyzing how customer characteristic impact review score. The sellers and geolocation dataset were excluded from the analysis, as they primarily measure seller specific data and regional data which are not directly related to review scores. Key datasets, including customers, orders, order items, products, product category name, payments, were combined in the initial stages.

We used an inner join to merge dataset such as orders, customers, and reviews, and left join to merge payment and product category, to prevent data loss from main data, allowing comprehensive analysis and focusing on review score.

Subsequently, missing value and duplicate entries were addressed because of the low quantity and insignificant impact to the result. Incomplete records are removed to prioritize the data suitability and integrity.

To enhance dataset usability, key data engineering techniques were applied. The payment value feature was added to compare prices with total payments, and delivery time days was calculated to reflect actual durations. Delivery statuses were classified as early, on time, or delayed to analyze discrepancies with estimated dates. Product categories were translated into English and grouped into broader classifications for better interpretability.

Attribute	**Description**
Customer unique_id	A unique identifier for each customers
Review score	Review score given by the customer (ranging from 1 to 5)
Customer state	The region where the customer resides
Price	The price of the product ordered
Payment value	The amount paid by the customer
Delivery time days	The number of days taken for the product to be delivered
Delivery status	The status of the delivery (e.g., early, on time, delayed)
Product category name english	The name of the product category

Figure 2

2.2 Data Visualisation

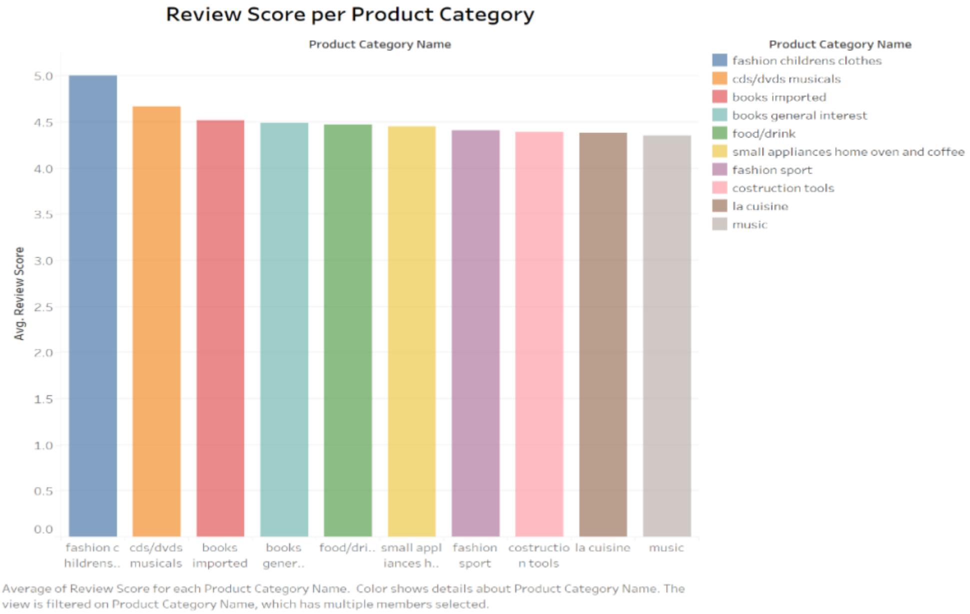


Figure 3

Figure 3 implies an analysis of top 10 review scores across product categories. Fashion children's clothes emerge as the highest rated category boasting an impressive average score approaching 5.0, which indicates exceptional customer satisfaction. Also, categories including CDs/DVDs musicals, Books imported, Food/drinks, and small appliances also showcase strong performance with scores ranges from 4.0 to 4.5 indicating overall positive customer sentiment.

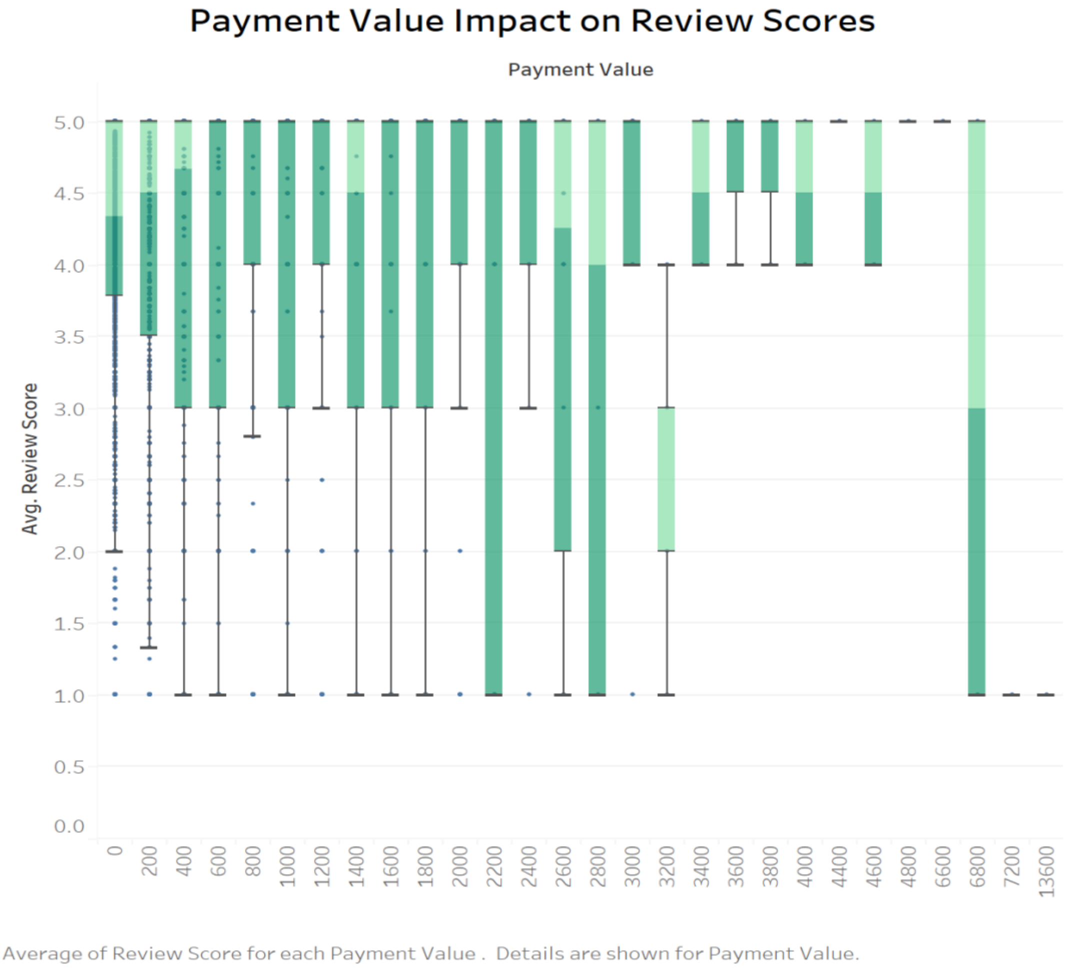


Figure 4

Figure 4 demonstrates the influence of payment value on review scores. It reveals that higher payment values generally correspond with more consistent and higher review scores (4-5), illustrating increased customer satisfaction. On the other hand, lower payment values exhibit greater variability (1-5), depicting mixed customer experiences. Especially outliers at lower values with exceptionally low scores may indicate specific concerns associated with lower prices products or services.

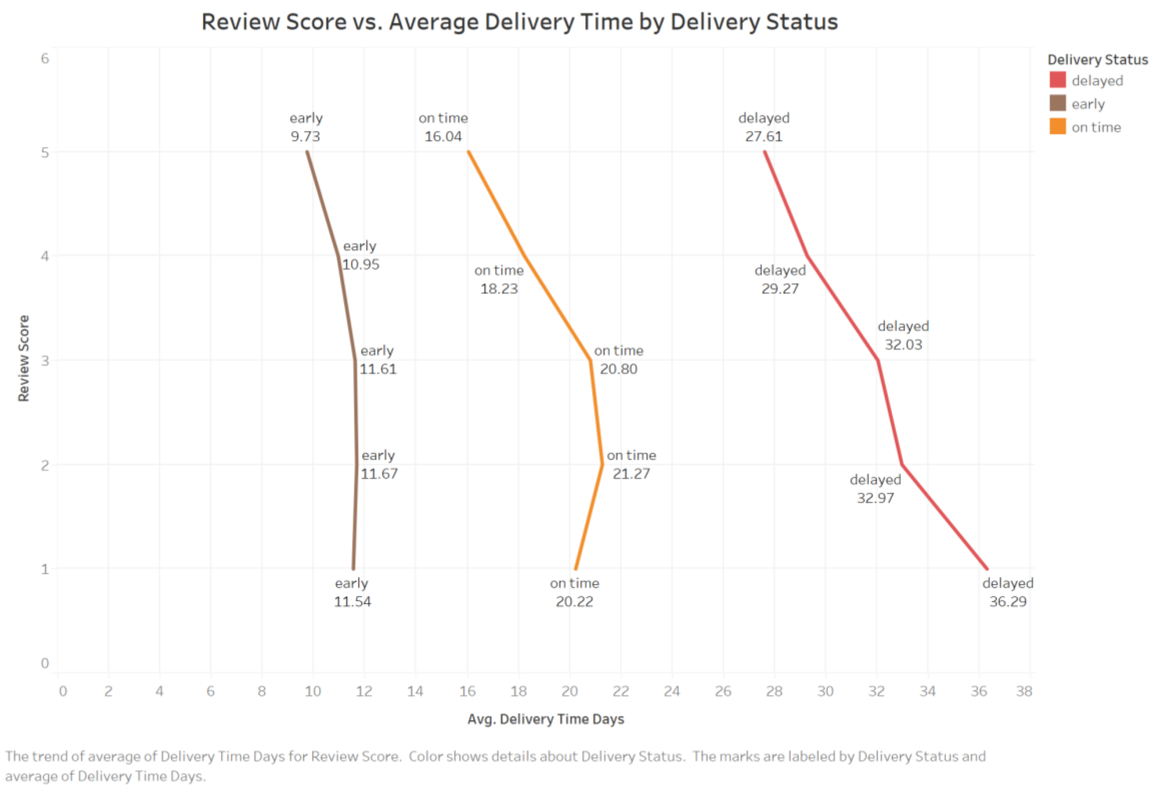


Figure 5

This Figure shows us the correlation between delivery time and review scores, categorized by delivery status. Early deliveries which are represented in brown consistently garner higher scores (4-5), suggesting greater customer satisfaction. Deliveries that arrive on time indicating in orange receives slightly lower yet positive scores around 4. In contrast delayed deliveries depicted in red exhibit a marked decline in customer satisfaction with scores dropping below 3 when delays surpass 30 days. This underscores the negative impact of extended delivery times on customer satisfaction.

2.3 Modelling Techniques

Using these metrics, a confusion matrix is created to compare various evaluation measures. This includes precision, recall, accuracy, and F1-score. Precision measures how accurate the model is when model predicts a positive outcome helping reduce unnecessary costs by focusing on correct predictions. Recall measures that the model ability to identify all positive value, ensuring that as many true positive as possible are captured.

The purpose of this analysis is to predict the likelihood of a customer leaving a positive review which is defined as a review score of 4 or 5. To achieve this objective various machine learning models are employed, trained, and assessed to identify the most effective model in predicting positive customer reviews.

For the modeling, the report uses a binary classification approach to predict whether customers would leave positive reviews and three algorithms are used: Random Forest, Gradient Boosted Decision Trees (GDBT), and XGBoost.

True positives (TP) refer to cases that the model correctly predicts a positive outcome, such as customer leaving a positive review. In contrast, False Positive (FP) means when the model incorrectly predicts a positive outcome for customer who doesn't actually give positive review. Similarly, True Negative (TN) and False Negative (FN) describe situations that the model correctly or incorrectly predicts a negative outcome, respectively.

$$Precision \ = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Rrecision * Recall}{Precision + recall}$$

$$Auccuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.4 Model Selection

The models that we have selected for this task includes: Random Forest Classifier (Classification), Random Forest Regressor (Regression), Gradient Boosted Decision Trees (GBDT) and XGBoost (Extreme Gradient Boosting)

Model Accuracy and Performance Metrics

Model	Accuracy	Macro F1 Score	Macro Precision	Macro Recall	MSE	R2
XGB	0.8109	0.6316	0.7679	0.6123		
GBDT	0.8109	0.6177	0.7879	0.6018		
Random Forest-classification	0.8013	0.6428	0.7181	0.6237		
Random Forest-regression					0.1556	0.1029

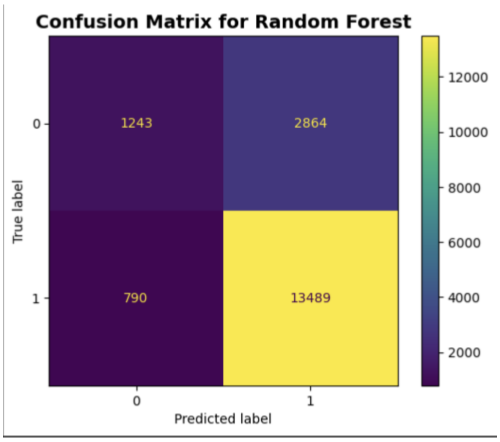
Figure 6

1) Random Forest- regression

The report first conducted regression analysis using Random Forest as it was more suitable for predicting review scores. Regression analysis is particularly effective for continuous variables (e.g., score, prices). The model performance is good on the training data (Precision:0.991, Recall:0.979, F1-score:0.984, Accuracy:0.989), indicating potential overfitting. However, its performance on the test data significantly drops 18.9%-35.6%, which means the model can't generalize all the data. Additionally R^2 value (0.1) and MSE (0.16) highlight that the model struggle to explain data variance, making it less suitable for regression model.

To address the limitation of regression model, the report uses binary classification approach. Review scores are categorized into binary classes: 1-3 as negative, and 4-5 as a positive. Confusion matrix and feature importance are used for visualization and comparison.

2) Random Forest - classification



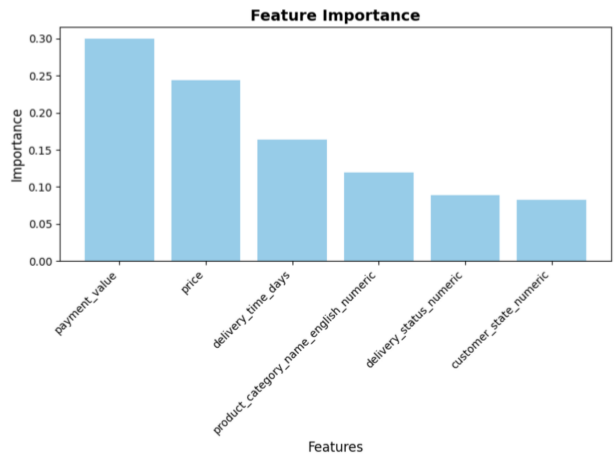
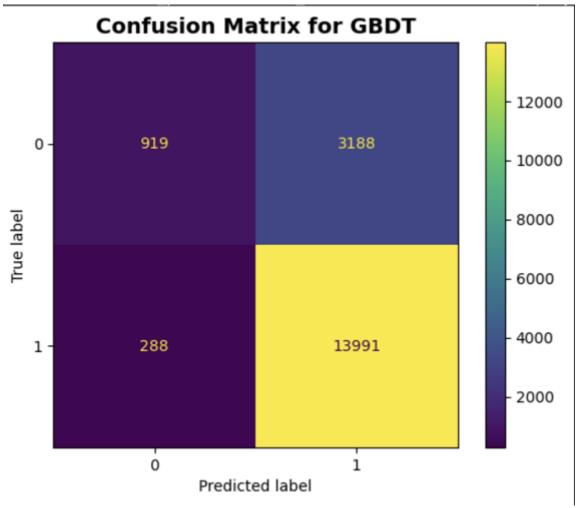


Figure 7 **Figure 8**

The Random Forest Classifier demonstrates a substantial discrepancy between its training and test performance which indicates a potential overfitting issue. Although the model achieved exceptional results on the training dataset, with an accuracy of 0.990, macro precision of 0.990, and macro F1- score of 0.985, its performance significantly declines on the test data resulting in an accuracy of 0.8013, macro precision of 0.718, and macro F1- score of 0.643. This suggests the model is overfitting to the training data, limiting its ability to generalize. From the confusion matrix, the model correctly predicts 13,489 TP but misclassifies 2,864 as FP. Additionally, it misclassifies 790 as FN, which impacts its recall. The feature importance shows that payment value (0.3) is the most influential feature, followed by price (0.24) and delivery time days (0.16), which play key roles in predicting review scores.
While it captures training patterns effectively, its poor generalization highlights the need for further refinement or regularization before production deployment.

3) GDBT



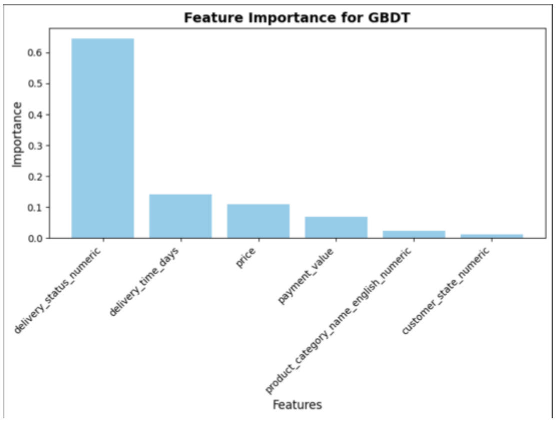


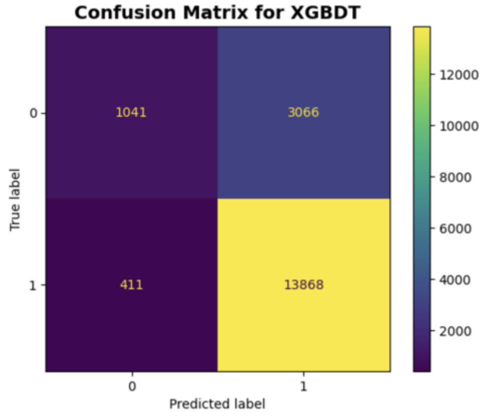
Figure 9 **Figure 10**

Another approach is the GDBT model, like Random Forest model, which is a tree-based ensemble learning algorithm. However, GDBT builds trees sequentially, which helps reduce overfitting and improve accuracy. Additionally, it updates the model during the training process by using the gradient of the loss function, making it effective for relatively small or complex dataset.

This GBDT model exhibits consistent performance across training and test datasets, achieving an accuracy of 0.811 in both. It demonstrates high precision 0.788 on test data, effectively predicting positive reviews while minimizing FP.

However, the lower macro recall value of 0.602 indicates that the model may overlook certain positive reviews, which could be a disadvantage in identifying all satisfied customers.
From the confusion matrix, the model correctly predicts 13,921 TP but misclassified 3,188 as FP, affecting precision. The most key feature for GDBT model is delivery_status_numeric (0.646), indicating that delivery timing strongly influences review score predictions.

4) XGB



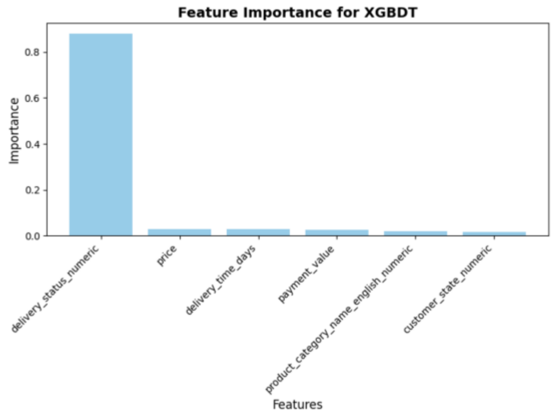


Figure 11 **Figure 12**

From Tianqi and Carlos (2016), XGBoost incorporates a regularized model to prevent overfitting. XGBoost is a high-performance tree-based learning model that extends GDBT. Since GDBT has limitation in handling large datasets, the report proposed using XGBoost to improve F1-score and accuracy.

XGBoost exhibits robust training performance achieving a training accuracy of 0.803 and demonstrating consistent outcomes across both training and test datasets. The test performance characterized by an accuracy of 0.811 and a macro precision of 0.768 indicates XGBoost's effectiveness in identifying TP while maintaining a balance between FP and FN. Notably, the less pronounced discrepancy between training and test performance, in comparison to Random Forest, underscores XGBoost's superior generalizability, which can be attributed to its regularization techniques that effectively mitigate overfitting.

Furthermore, the confusion matrix analysis reveals a reduction in misclassifications, particularly a decrease in FN, when compared to Random Forest. This enhances performance renders XGBoost more reliable in accurately identifying positive reviews. This critical insight highlights XGBoost's superior generalizability and balanced performance, positioning it as the most dependable model for predicting customers with a high likelihood of leaving positive reviews, especially when the objective is to minimize both FP and FN.

3.Model Evaluation

This model is designed to reduce unnecessary costs by targeting prospective customers review scores.

In this analysis four models are utilized to predict the likelihood of customers leaving positive reviews. Each model is meticulously evaluated using various metrics like accuracy, macro F1 Score, macro precision, macro recall, mean squared error (MSE), and R Squared to ascertain their effectiveness.

Metric	Value
Ture Positives (TP)	13,492
False Positives (FP)	2,850
True Negatives (TN)	1,257
False Negatives (FN)	786

Figure 13

Among the 4 models XGBoost demonstrates exceptional performance achieving an impressive accuracy of 0.8109 and a commendable macro precision of 0.7679. XGB effectively identifies genuine positive instances while minimizing the occurrence of FP. However, its macro recall (0.6123) is moderate, indicating potential oversights of certain positive instances.

Metric	Value
Ture Positives (TP)	13,490
False Positives (FP)	2,860
True Negatives (TN)	1,247
False Negatives (FN)	788

Figure 14

The GBDT model attains an accuracy of 0.8109 and the highest macro precision of 0.7879. This signifies its remarkable proficiency in accurately predicting positive reviews. However, its relatively lower recall 0.6018 tells us that GBDT may overlook a substantial number of TP instances.

Metric	Value
Ture Positives (TP)	13,489
False Positives (FP)	2,864
True Negatives (TN)	1,243
False Negatives (FN)	790

Figure 15

The Random Forest Classifier exhibits an accuracy of 0.8013, but it is overfitting, resulting in a discrepancy between training and test performance. While it demonstrates superior recall 0.6237 compared to GBDT and XGBoost, its precision is lower indicating a higher likelihood of FP.

Alternatively, the Random Forest Regressor is evaluated using regression metrics attaining a MSE of 0.1556 and an R-Squared of 0.1029. This model exhibits moderate predictive capabilities for continuous outcomes but was less effective than classification models in predicting positives reviews.

4.Conclusion and Recommendation

Among the four models evaluated, XGBoost stands out as the best choice due to its well-balanced performance across key metrics. It achieves strong accuracy, satisfactory precision, and adequate recall, making it the most reliable model for predicting positive customer reviews. XGBoost effectively minimizes false positives and false negatives, crucial for maintaining high customer satisfaction and accurately capturing positive experiences.

To boost customer satisfaction and review scores, Nile should implement a comprehensive strategy targeting key improvement areas. Marketing efforts should focus on high-value product categories, like children's fashion, using customer testimonials, while addressing dissatisfaction in lower-performing categories to build trust.

Loyalty programs and exclusive benefits should target top customers to retain satisfaction, while improving product quality and delivery times for lower-value purchases. Optimizing logistics, addressing supply chain bottlenecks, and offering proactive communication and compensation for delays will further enhance satisfaction and foster long-term loyalty.

References

1. Shawn Chumbar (2023). “The CRISP-DM Process: A Comprehensive Guide”. Available at:

https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151

2. Schwenke, C. and Schering, A.G., 2014. True positives, true negatives, false positives, false negatives. Wiley StatsRef: Statistics Reference Online.

3. Riyanto, S., Imas, S.S., Djatna, T. and Atikah, T.D., 2023. Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. International Journal of Advanced Computer Science and Applications, 14(6).

4. Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).