Table of Contents

1 Abstract

As eCommerce has become more ubiquitous in the retail industry, customer reviews have become vital for sustaining company revenue. This report discusses the process taken by the analytics team in creating a classification model to satisfy the request of the eCommerce platform, Nile. The company put forth a contract to develop a model which predicts whether their customers will leave positive reviews, hoping to better target buyers for positive feedback and reduce operational inefficiencies related to obtaining reviews. Employing the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, the team will systematically clean and choose data for model training. Then, models will be developed and evaluated based on key performance metrics and their fit to Nile's goals. The use and integration of the chosen model for the business will be discussed as a part of the deployment phase before conclusions and recommendations are given. Justifications for the approaches taken and methods used will be included for each phase in the project.

2 Introduction

Nile, an eCommerce platform selling a variety of products in South America, has requested a model to predict which customers will leave positive reviews. The goal is to increase favourable reviews while minimizing resources spent on incentives to obtain them.

The proposed solution addresses two pain points which form the current business problem Nile is facing–poor reviews and operational inefficiency (Wu & Buyya, 2015). Adopting the CRISP-DM methodology, a framework first introduced in 1999 and considered to be the standard for various data science projects, the team systematically developed a classification model to satisfy the needs of the client (Martinez-Plumed, et al., 2019). This report presents the approach the team took as a part of the CRISP-DM process, including data understanding, data preparation, modelling, model evaluation, and deployment. Conclusions and recommendations were also considered in conjunction with the team findings (Saltz, 2021).

3 Data

The data understanding phase focuses on exploring data related to the business problem to assess the structure, quality, and relevance of the data, identifying any issues like missing or inconsistent values. The data preparation phase follows data understanding and involves cleaning, transforming, and organizing the data to ensure accurate analysis and modelling. (Bokrantz, et al., 2024)The data understanding phase focuses on exploring data related to the business problem to assess the structure, quality, and relevance of the data, identifying any

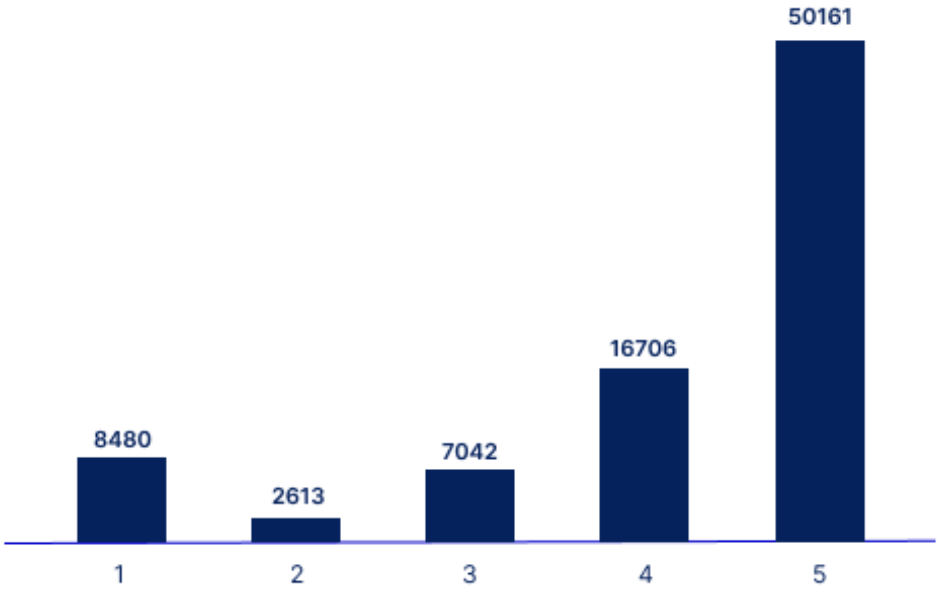
issues like missing or inconsistent values. The data preparation phase follows data understanding and involves cleaning, transforming, and organizing the data to ensure accurate analysis and modelling. (Bokrantz, et al., 2024)

3.1 Data Understanding

As part of the data understanding phase, the team aimed to uncover patterns, detect anomalies, and validate assumptions about variables related to the business problem. Exploratory Data Analysis (EDA) was conducted on the raw dataset comprising of eight tables and the following key insights were derived:

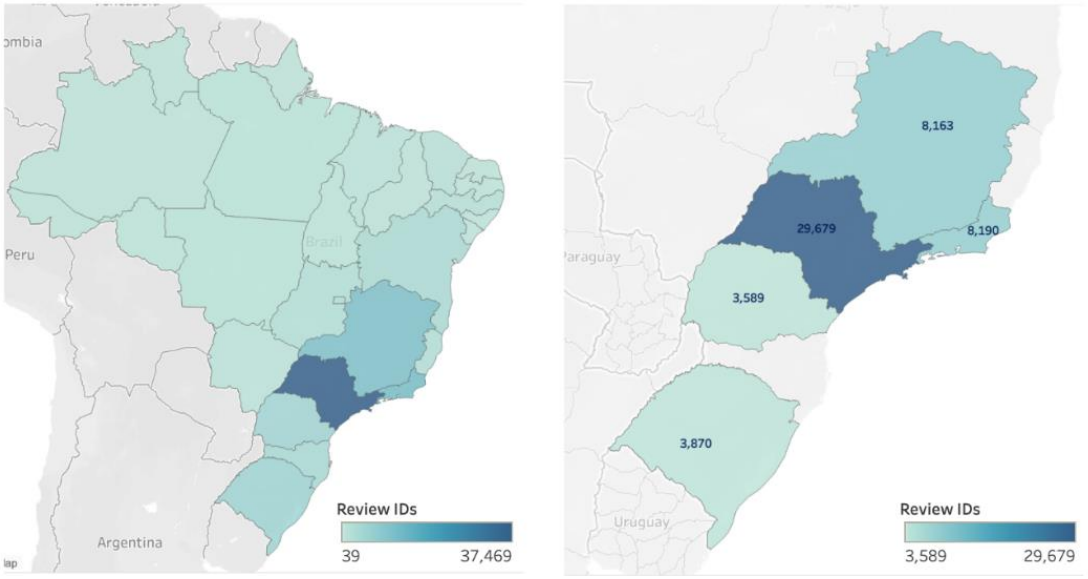
Figure 1 shows that the data is significantly skewed towards higher scores (4 and 5).

Figure 1. The distribution of review scores based on the number of orders



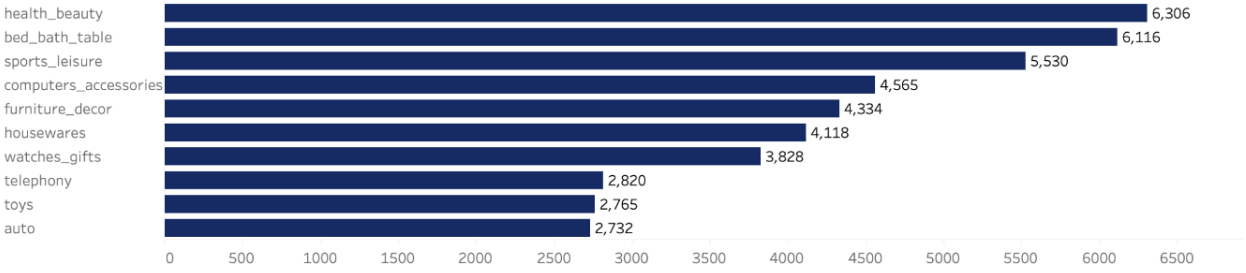
There is significant variation in the number of reviews across the 27 states of Brazil as shown in *Figure 2*.

Figure 2. All customer states by reviews (left) and Top states by positive reviews (right)



The most reviewed categories with high ratings are shown in *Figure 3*. These categories tend to have smaller basket sizes and lower average order values, indicating frequent, smaller purchases that contribute to a higher volume of reviews.

Figure 3. Top categories by review count



The geolocation and customers datasets both contained location related information. Latitude and longitude were the only additional provided by the geolocation dataset. We determined that this was unneeded for our analysis, making the geolocation dataset redundant. For a list of all anomalies identified in the data during this stage, refer to Table 1.

Table 1. Anomalies in data understanding

Feature	Anomalies
Payment Type	Rows were deleted where payment type was not defined.
Product Category Name	Two categories with no English translation were manually translated.
Product Weight	Few instances of the product weight being 0 or 2, these rows were removed.
Order Processing Dates	Few instances where the product was delivered to the customer before being delivered to the carrier were removed.

3.2 Data Preparation

To maintain data integrity, only orders that were delivered were considered in modelling, ensuring that reviews corresponded to completed transactions. Orders containing multiple unique products were filtered out, shifting our focus to a unique product per order. Additionally, for instances where customers left multiple reviews for the same product, only the latest review was retained.

Relevant features, such as delivery times and payment values were retained, while features like timestamps and identifiers were excluded or engineered to reduce noise. Additionally, features such as delivery times, number of items per order and number of orders per customer were calculated to understand customer purchase behaviour.

To enhance feature interpretability, the categorical variables of customer state, payment type, and product category were transformed by applying one-hot encoding, which turns each category into a binary variable. For more information detailing all the feature transformations and feature selection, refer to appendix A.

4 Model

Following data understanding and preparation, the team began selecting modelling techniques and testing models based on the business problem. Results of these models were checked against Nile's business objectives and the most suitable model was selected. These actions constitute the modelling and evaluation phase of CRISP-DM. (Schröer, et al., 2021)

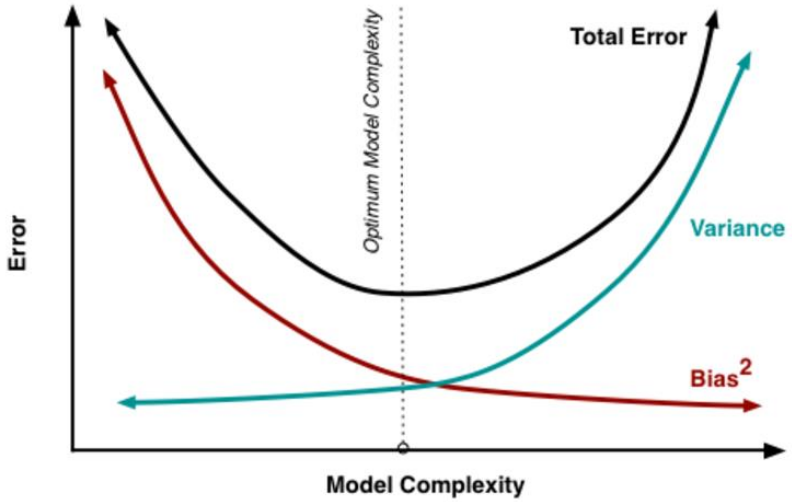
4.1 Modelling

We implemented two ensemble classification methods. The first was the Random Forest Classifier (RF) which is a high-performance classification tool and is resistant to overfitting (Breiman, 2001). The second method is the Gradient-Boosted-Decision-Tree Classifier (GBDT). As a 'TreeBoost procedure' the GBDT classifier is characterized by its robustness, which

includes low sensitivity to outliers, non-normal distributions, and a reduced need for input transformation (Friedman, 2001).

For each ensemble method, we created six different models, four models to classify review scores as a binary variable, where a review score of 4 or 5 is considered as the positive class and two models to classify review scores as a multi-class variable. It was observed that when the data complexity is simple, undersampling the majority class leads to a significant increase in the area under the precision-recall curve (AUPRC), while keeping loss in information to a minimum, as shown in Figure (Kim & Jung, 2023).

Figure 4. Undersampling and feature selection to achieve the ideal model complexity (Fortmann-Roe, 2012)



Since our final model only utilised five numerical features and established that the class distribution was heavily skewed, the team implemented 1:1 under-sampling of the positive class for all our binary classification models. Furthermore, all models made use of an 80/20 train-test split. To find the best set of hyper-parameters for all models, we made use of the random search function as it increases efficiency in identifying important parameters and reduces computational cost, when compared to grid search or manual hyper-parameter selection (Bergstra & Bengio, 2012).

For each method, we split the model between using recall (positive class) or macro recall as the scoring criteria to pick the model, shifting focus between identifying as many true positives as possible and finding a balance between identifying true positives and true negatives.

4.2 Model Evaluation

We initially selected fifteen features impacting review scores to train our models. After running the preliminary models and using the feature importance function, we identified key features based on the mean and standard deviation of overall impurity decrease. (Scikit-learn, 2024). Based on these values, the top five of fifteen features were selected to construct simpler models. For more information, refer to Appendix B.

Additionally, after comparing our pilot models with the recall scoring criteria versus the models with the macro recall scoring criteria, we identified that the models created using recall as our scoring approach led to better overall performance. *Figure 5* and Table 2 show the confusion matrices and metric performances of the final RF and GBDT models.

Figure 5. Confusion matrices for final RF (left) and GBDT (right) models

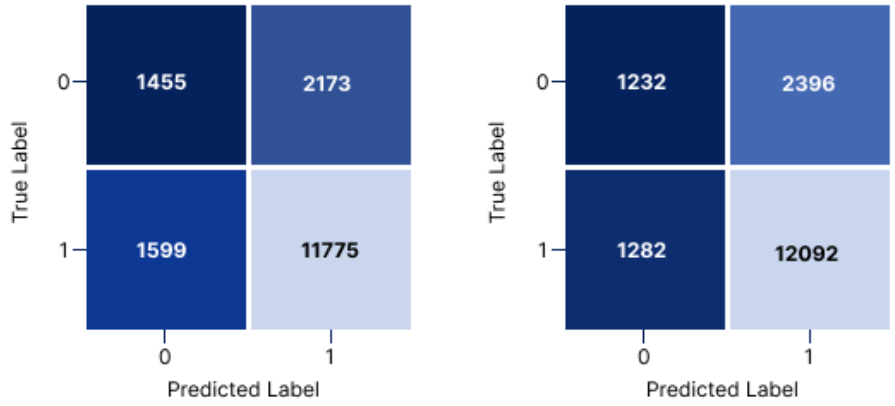


Table 2. Final model evaluation metrics

Model	Scoring Criteria	F1 Macro	F1 (Positive Class)	Precision Macro	Precision (Positive Class)	Recall Macro	Recall (Positive Class)	Accuracy
RF	Recall	0.65	0.86	0.66	0.84	0.64	0.88	0.78
GBDT	Recall	0.62	0.87	0.67	0.83	0.61	0.92	0.79

Between the 5-feature RF and GBDT models, there is a 0.59% difference in model accuracy. The F1 score measures the relationship between the predictions of the model and the true class in the data. Recall measures how well the model can identify a specific class, while precision measures the agreement between our predicted label and the true label (Sokolova, 2009). The RF model achieved a higher macro F1 score, and a higher macro recall score compared to the GBDT. The higher macro scores in the RF model indicate that it strikes a better balance

between precision and recall for both classes, making the RF model more effective for identifying customers likely to leave positive reviews while minimizing false positives and associated costs.

5 Deployment

Following the successful selection of a model, the team moved to the deployment phase of the CRISP-DM process, which concentrates on the integration of solutions into organizations. This can include planning for deployment, monitoring, and maintenance, producing a final report, and conducting a final project review depending on the individual needs of the project (IBM, 2021).

Brackmann et al. (2023) categorized machine learning (ML) applications in retail into decision-oriented solutions, aiding business decisions, and economic-operative solutions supporting customer transactions. Our created classification model as a decision-oriented tool can help Nile optimize resource allocation and minimize costs by feeding real-time data to our model that consequently sends out automated e-mails to customers predicited to leave a postive review. One method of integrating the model into the firm's systems is the use of a cloud-native application which would split components into a series of inter-related microservices (Deng, et al., 2024). The use of microservices, allows for a scalable system where the model could be easily integrated and maintained independent of other components (Torvakar & Game, 2019).

6 Conclusion

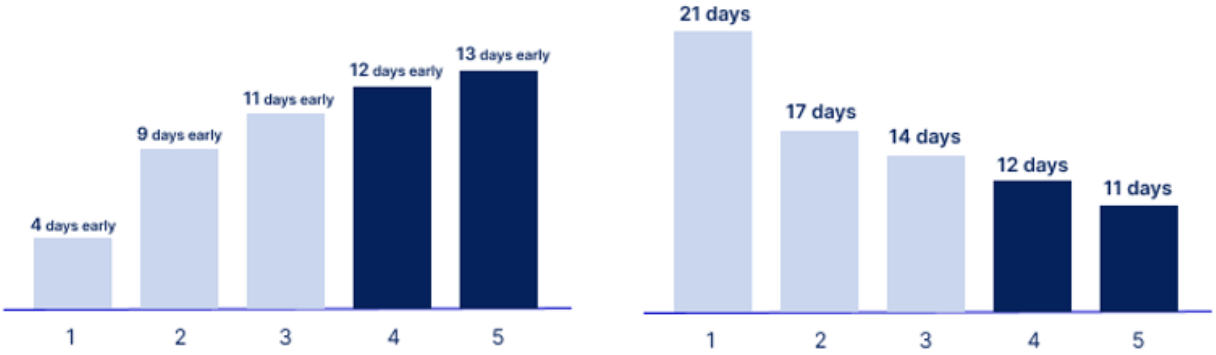
Following the CRISP-DM framework, the team sought to understand the data by exploring patterns, relationships, and anomalies in variables. Data was then prepared and used in different combinations for modelling. Upon evaluation, we found that the RF model performed best, offering an optimal combination of precision and recall for both classes. By testing the model with a smaller set of key features, the increased simplicity helped the team yield an effective solution tailored to Nile's needs, while minimizing resource usage.

Throughout the modelling process, the team experienced issues with overfitting, which may cause bias and limit the performance of the model. Additionally, the data was heavily skewed toward higher review scores. The action of under-sampling the majority class to remedy this, while reducing bias, resulted in a very minor loss of information. The model is also trained on data from 2016 to 2018, which is outdated and may affect deployment success. Nevertheless, this project highlights the importance of data-driven solutions for boosting customer engagement and improving efficiency. A successful implementation of our model will help Nile grow and meet its goals.

7 Recommendations

Based on our model findings, as seen in Figure 6, there is a significant impact of delivery performance on review scores. Faster delivery and ahead-of-schedule orders lead to higher customer ratings. Nile may focus on optimizing supply chain processes to improve delivery times and operational efficiency to maximise positive reviews.

Figure 6. Average days overdue (left) and delivery time (right) by review score



Furthermore, during the analysis, we identified anomalies in the data, such as the presence of seventy-one overlapping product categories. We recommend consolidating these categories to streamline the data as seen in Appendix C. Additionally, we suggest targeted incentives such as redeemable coins for writing reviews or a frequent buyers discount for customers likely to leave positive reviews, as this can boost future purchases and engagement. Other strategies, such as basket coupons, discount vouchers on impulsive buys, and cashback on prepaid transactions, can also add to positive reviews and Nile's brand image.

As a final recommendation, we suggest focusing on three key deliverables. Firstly, capturing customer demographic data will provide a deeper understanding of their behaviour, enabling targeted strategies and effective segmentation. Secondly, by performing sentiment analysis on review content, combined with advanced AI tools, valuable insights can be derived about customer behaviour, helping Nile-eCommerce to make improvements accordingly. Lastly, future-proofing model by regularly updating the features and incorporating evolving data needs will keep it adaptable to changing trends, ensuring Nile-eCommerce remains ahead of the competition.

8 References

Bergstra, J. & Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research,* 13(1), pp. 281-305.

Bokrantz, J., Subramaniyan, M. & Skoogh, A., 2024. Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM.. *Production Planning & Control,* 35(16), pp. 2234-2254.

Brackmann, C., Hutsch, M. & Wulfert, T., 2023. Identifying Application Areas for Machine Learning in the Retail Sector: A Literature Review and Interview Study. *SN Computer Science,* 4(5), pp. 1-17.

Breiman, L., 2001. Random Forests. In: R. E. Schapire, ed. *Machine Learning.* Dordrecht: Kluwer Academic Publishers, pp. 5-32.

Deng, S. et al., 2024. Cloud-Native Computing: A Survey From the Perspective of Services. *Proceedings of the IEEE,* 112(1), pp. 12-46.

Fortmann-Roe, S., 2012. *Understanding the Bias-Variance Tradeoff: An Overview,* s.l.: s.n.

Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics,* 29(5), pp. 1189-1232.

IBM, 2021. *IBM.* [Online]
Available at: https://www.ibm.com/docs/en/spss-modeler/18.3.0?topic=deployment-overview
[Accessed 28 11 2024].

Kim, A. & Jung, I., 2023. Optimal selection of resampling methods for imbalanced data with high complexity. *PLOS ONE,* Issue https://doi.org/10.1371/journal.pone.0288540.

Martinez-Plumed, F. et al., 2019. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering,* 33(8).

Saltz, J. S., 2021. *CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps.* Orlando, IEEE.

Schröer, C., Kruse, F. & Gómez, J. M., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science,* 181(1), pp. 526-534.

Scikit-learn, 2024. *Sckikit-learn.* [Online]
Available at: https://scikit-learn.org/dev/auto_examples/ensemble/plot_forest_importances.html
[Accessed 30 11 2024].

Sokolova, M., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management,* Issue DOI: 10.1016/j.ipm.2009.03.002, pp. 45(4):427-437.

Torvakar, N. & Game, P. S., 2019. Microservices and It's Applications : An Overview. *International Journal of Computer Sciences and Engineering,* 7(4), pp. 803-809.

Wu, C. & Buyya, R., 2015. Business Needs. In: C. Wu & R. Buyya, eds. *Cloud Data Centers and Cost Modeling: A Complete Guide to Planning, Designing and Building a Cloud Data Center.* s.l.: Morgan Kaufmann Publishers , pp. 43-95.

9 Appendix

9.1 Appendix A: Table of Data Preparation Actions

This appendix provides information on the data cleaning and feature selection process for the project. Tables A1 to A9 account for each individual dataset that was given to the analytics team and records the actions taken to prepare each dataset for modelling. Features were kept, dropped, or created by the team, and may have also been cleaned based on what the team discovered about each respective variable. Table A10 represents the features that were ultimately chosen for modelling. These features could have been used for the 15-feature RF and GBDT models, the 5-feature RF model, or the 5-feature GBDT model. Some features were used in multiple models.

Appendix Table A1. Customers Dataset

Feature	Status
Customer ID	Kept
Customer Unique ID	Kept
Customer Zip Code Prefix	Kept
Customer City	Dropped
Customer State	Kept
Multiple Orders	Created
Order Frequency	Created

Appendix Table A2. Geolocation Dataset

Feature	Status
Geolocation Zip Code Prefix	Dropped
Geolocation Latitude	Dropped
Geolocation Longitude	Dropped
Geolocation City	Dropped
Geolocation State	Dropped

Appendix Table A3. Order Items Dataset

Feature	Status
Order ID	Kept
Order Item ID	Kept
Product ID	Kept
Seller ID	Kept
Shipping Limit Date	Dropped
Price (total)	Kept
Freight Value	Dropped
Product Count in order	Created
Product unit price	Created
Product min unit price	Created
Product max unit price	Created

Appendix Table A4. Order Payments Dataset

Feature	Status
Order ID	Kept
Payment Sequential	Dropped
Payment Type	Kept & Cleaned
Payment Instalments	Kept
Payment Value	Kept
Payment Type Count	Created

Appendix Table A5. Order Reviews Dataset

Feature	Status
Review ID	Kept
Order ID	Kept
Review Score	Kept
Review Comment Title	Dropped
Review Comment Message	Dropped
Review Creation Date	Dropped
Review Answer Timestamp	Dropped
Review Title Length	Created

Review Comment Length	Created
Has Reviewed	Created
Time to Review	Created

Appendix Table A6. Orders Dataset

Feature	Status
Order ID	Kept
Customer ID	Kept
Order Status	Kept & Cleaned
Order Purchase Timestamp	Dropped
Order Approved Timestamp	Dropped
Order Delivered to Carrier Timestamp	Dropped
Order Delivered to Customer Timestamp	Dropped
Order Estimated Delivery Date	Dropped
Delivery to Carrier vs. Customer Difference	Created & Cleaned
Delivery to Customer vs. Purchase Difference	Created
Actual vs. Expected Delivery to Customer	Created

Appendix Table A7. Products Dataset

Feature	Status
Product Category Name	Kept & Cleaned
Product Name Length	Dropped
Product Description Length	Dropped
Product Photo Quantity	Dropped
Product Weight (g)	Kept & Cleaned
Product Length (cm)	Dropped
Product Height (cm)	Dropped
Product Width (cm)	Dropped
Product Volume (cm^3)	Created & Cleaned

Appendix Table A8. Seller Dataset

Feature	Status
Seller ID	Kept
Seller Zip Code Prefix	Kept
Seller City	Kept
Seller State	Kept

Appendix Table A9. Product Category Name Translation

Feature	Status
Product Category Name	Dropped
Product Category Name Translated	Kept

Appendix Table A10. Modelling Features

Feature	Model
Delivery Days Overdue	15 Feature, 5 Feature (RF), 5 Feature (GBDT)
Purchase to Delivery Difference	15 Feature, 5 Feature (RF), 5 Feature (GBDT)
Product Count per Order	15 Feature, 5 Feature (RF), 5 Feature (GBDT)
Payment Instalments per Order	15 Feature, 5 Feature (RF)
Payment Value per Order	15 Feature, 5 Feature (RF)
Customer Order Frequency	15 Feature, 5 Feature (GBDT)
Product Weight	15 Feature, 5 Feature (GBDT)
Carrier Customer Delivery Difference	15 Feature
Customer State	15 Feature
Return Customers	15 Feature
Payment Type	15 Feature
Total Payment Types Used	15 Feature
Unit Price	15 Feature
Product Volume	15 Feature
Product Category	15 Feature

9.2 Appendix B: Model Evaluation Metrics

This appendix details evaluation metrics and data relevant to evaluation metrics for all machine learning models created for the project during the modelling phase in the CRISP-DM process. Table B1 shows expected model accuracy using a simple calculation involving the number of classes employed and their share of values in a dataset. Assuming that a model randomly assigns a value to one class, the chance of it assigning the value correctly is dependent on the split of the data, and summing the rate of accuracy for each class will give the expected accuracy of the model. For example, if class one and class zero values have a 1:1 split for a dataset, expected accuracy for the model will be 0.5 since for each class, the model will be correct half of the time, and each class makes up half of the dataset.

Appendix Table B1. Expected model accuracy

Class	Expected Accuracy
Binary	0.66
Multiclass	0.41

Table B2, B3, and B3 provide the evaluation metrics for all models evaluated for the project. Model algorithms include Random Forest (RF) and Gradient Boosted Decision Tree (GBDT). The RF method uses bagging, training multiple models independently on different subsets of the training data. The GBDT method uses boosting, where succeeding models are trained based on previous models. Scoring methods for choosing models in the algorithms were either recall or recall macro, which looked at the proportion of true positives that were identified and the mean of recall for each individual class, respectively. The evaluation metrics included the F1, precision, recall, and accuracy metrics. Precision measures the proportion of positives that were correct, while recall measures the proportion of true positives that were correctly identified. F1 takes both the precision and recall measure of a model into consideration to calculate model performance. The team split each of these measures further into positive class or macro metrics, where they were calculated for just the positive class or as an average of the scores for all classes. Accuracy measures the number of correct predictions made against incorrect predictions made overall by the model.

Appendix Table B2. Model evaluation metrics for 15-feature multiclass models

Model	Scoring	F1 Macro	F1 (Positive Class)	Precision Macro	Precision (Positive Class)	Recall Macro	Recall (Positive Class)	Accuracy
RF1	Recall Macro	0.33	0.39	0.37	0.28	0.35	0.67	0.48
RF2	Recall	0.29	0.67	0.29	0.66	0.30	0.68	0.50
GBDT1	Recall Macro	0.26	0.67	0.35	0.23	0.32	0.84	0.58
GBDT2	Recall	0.24	0.76	0.32	0.62	0.27	0.99	0.62

Appendix Table B3. Model evaluation metrics for 15-feature binary models

Model	Scoring	F1 Macro	F1 (Positive Class)	Precision Macro	Precision (Positive Class)	Recall Macro	Recall (Positive Class)	Accuracy
RF1	Recall Macro	0.63	0.82	0.62	0.85	0.65	0.79	0.73
RF2	Recall	0.65	0.87	0.68	0.84	0.64	0.90	0.79
GBDT1	Recall Macro	0.64	0.83	0.64	0.86	0.65	0.81	0.74
GBDT2	Recall	0.53	0.86	0.57	0.80	0.54	0.92	0.77

Appendix Table B4. Model evaluation metrics for 5-feature binary models

Model	Scoring	F1 Macro	F1 (Positive Class)	Precision Macro	Precision (Positive Class)	Recall Macro	Recall (Positive Class)	Accuracy
RF1	Recall Macro	0.62	0.81	0.62	0.85	0.64	0.78	0.71
RF2	Recall	0.65	0.86	0.66	0.84	0.64	0.88	0.78
GBDT1	Recall Macro	0.63	0.83	0.63	0.85	0.64	0.82	0.74
GBDT2	Recall	0.63	0.87	0.66	0.83	0.62	0.90	0.78

9.3 Appendix C: Recommended Product Categorisation

This appendix relates to the recommendation the team proposed to Nile for scaling up their product categories. The objective of this suggestion is to make the product categories more usable for modelling purposes as this variable is currently too complex for modelling and would result in overfitting. Additionally, many product categories are not significantly distinct from one another, so the information loss from scaling up would be minimal. Table C1 shows some combined categories the team suggested and the number of original categories that would be absorbed in these new categories.

Appendix Table C1: Recommended Product Categorisation

Categorisation	Number of Categories Combined
Arts and Crafts	2
Education	4
Electronics & Technology	12
Entertainment & Media	3
Fashion & Accessories	8
Food & Beverages	4
Health & Personal Care	4
Home & Living	22
Professional & Industry	4
Specialty Items	5
Sports & Outdoor	3