Table of Contents

Table of figures

List of tables

Introduction

This report describes the technical considerations of the machine learning model created to predict which customers will leave positive reviews on Nile, a leading South American eCommerce platform. It addresses the technical requirements and challenges considered while formulating and solving the project, providing justifications for the methods, variables and models chosen. It outlines the process undertaken, including data preparation and engineering, modelling, and concludes with an overview of the overall approach.

Data Preparation

Data processing and manipulation was performed using Python with the panda and numpy libraries. The datasets were loaded into data frames and file names were standardised to provide consistency and easier usability. The necessary data frames were extracted to be merged, and the final data frame was saved as a new csv. The primary keys for merging datasets were customer id and order id, for instance order id was used when replacing product category names from Portuguese to English. The attributes were selected for their informativeness in predicting the target variable - positive reviews - and based on data mining techniques. In addition, new variables were included, as deemed necessary in predicting positive reviews. These are total delivery time[1], early/late status[2], total order price[3], category scores[4], seller scores[5], state scores[6], payment instalments, photo quantity, and product description length. The rationale for creating these variables is provided in Appendix 2.

Data quality and integrity checks were performed on the merged dataset. Both missing values and duplicates were removed. Considering the amount of missing data as insignificant compared to the overall data set, they were removed. Similarly, any information considered as false positive was eliminated, ensuring reliability of predictions and alignment with customer behaviour.

[1] (customer delivery date – order approval date)

[2] (estimated delivery date - customer delivery date)

[3] (item price + freight value)

[4] average review received by each product category

[5] average review received by each seller

[6] average review score given by each state

Modelling

In this stage, the form of the model and its attributes are determined, and the parameters are tuned to ensure the model fits the data. The final dataset was imported into Python and the variables were split into two groups: X[7] and Y[8]. By doing so, the model differentiates between the target variable and the independent variables. The next step involves choosing between building a regression model, a multi-class classification model or a binary-class classification model. Considering Nile's objective, a binary-class classification model was deemed most appropriate, as it distinguishes between positive and negative. It also reduces model complexity, making it faster and easier to train the model.

The XGBDT[9] algorithm was selected. This algorithm is an improvement of GBDT[10], as the addition of the XGBoost library results in higher performance, efficiency, and improved scalability with larger datasets, such as Nile’s (Kavlakoglu et al, 2024). This algorithm builds decision trees in parallel, uses the errors from previous week learners to update the sample weight value and then iterates round by round to obtain the optimal result. Its essence is the addition model and forward step algorithm, which improve its accuracy (Wang and Dong, 2024). The model’s performance was assessed with precision, recall and F1-score. Precision was prioritised due to the need of removing false positives, which could result in wasted resources and opportunity loss. Finally, the model was validated by splitting the data into training data[11] (80%) and test data[12] (20%). The training data was used to fit the model, while the test data was used to validate it.

The model assumptions are included in Appendix1.

[7] X are the independent variables.

[8] Y is the dependent variable.

[9] Extreme Gradient Boosting Decision Trees.

[10] Gradient Boosted Decision Trees.

[11] X_train & Y_train

[12] X_test and Y_test.

Overall approach

The data mining process followed the Cross Industry Standard Process for Data Mining (CRISP-DM) codification as illustrated in figure 1 below (Shearer's, 2000). This was the foundation in structuring the approach taken to solve the problem.

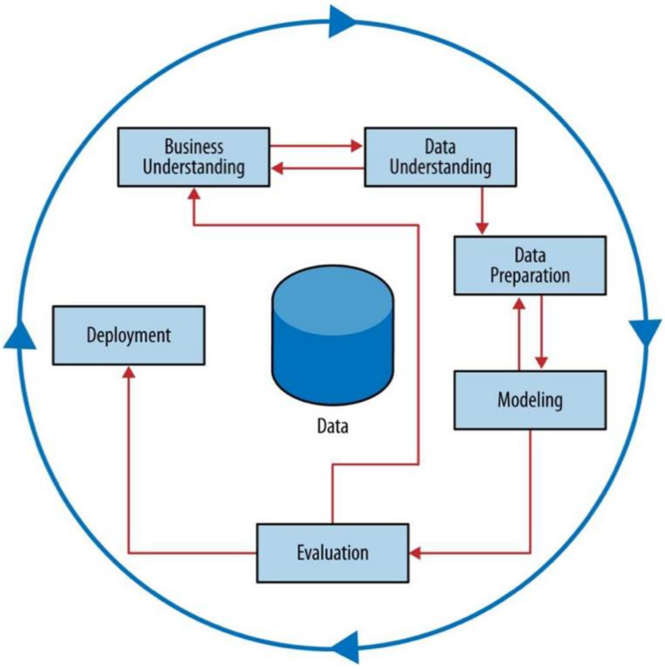


Figure 1: The CRISP data mining process

As shown above the process was not linear. The initial formulation was not sufficient, so multiple iterations were deemed necessary, until an ideal solution was formulated. This was an iterative process of discovery.

Business Understanding

The table below outlines how the business problem was decomposed into subtasks. The solution of each subtask was then composed to solve the overall problem.

Business problem: "Predicting which customers will leave positive reviews"		
Subtask	Description	Solution
Define Business Objective	Identifying desired outcomes, ensuring objectives are achievable, measurable and that project outcomes align with business needs.	Develop a predictive model that identifies customers that are likely to leave positive reviews.

Understand Business Context	Analysing the factors that influence positive reviews, ensuring strategies are aligned with the broader environment.	Chosen variables: • Total order price • Delivery time in days • Early/Late status • Amount of instalments • Average category score • Average seller score • Average state score • Quantity of photos • Length of description
Identify Stakeholders	Identifying stakeholders, understanding their interests, and assessing their influence and knowledge.	Focus on two key stakeholders: fellow analysts and clients.
Risk analysis	Identifying, assessing and prioritising potential risks.	Mitigate false positives, as they could have the worst impact on the business decision making.
Developing hypothesis	Formulating predictions and testing the relationship of chosen variables to the dependent variable.	Use machine learning to test whether the chosen variables can predict positive reviews.

Table 1: Business Understanding

Data Understanding

As mentioned by Wirth and Hipp's (2000), this stage involves data familiarisation and quality checks, along with the discovery of insights and identification of notable subsets to develop hypotheses on underlying patterns. Data understanding was gained by visualising single attributes and their combinations, as well as statistical and evidence-based analysis of relationships between attributes. This helped identify and characterise outliers and missing values. The relationship between average review score and delivery time was observed using a bar chart as shown below. Additionally, a geo-map was used to visualise the review scores in different areas.

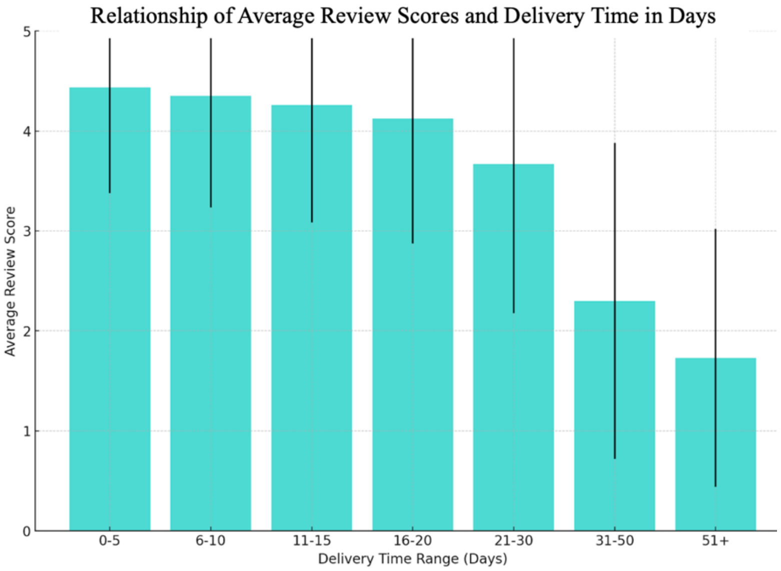


Figure 2: Pearson Correlation Relationship

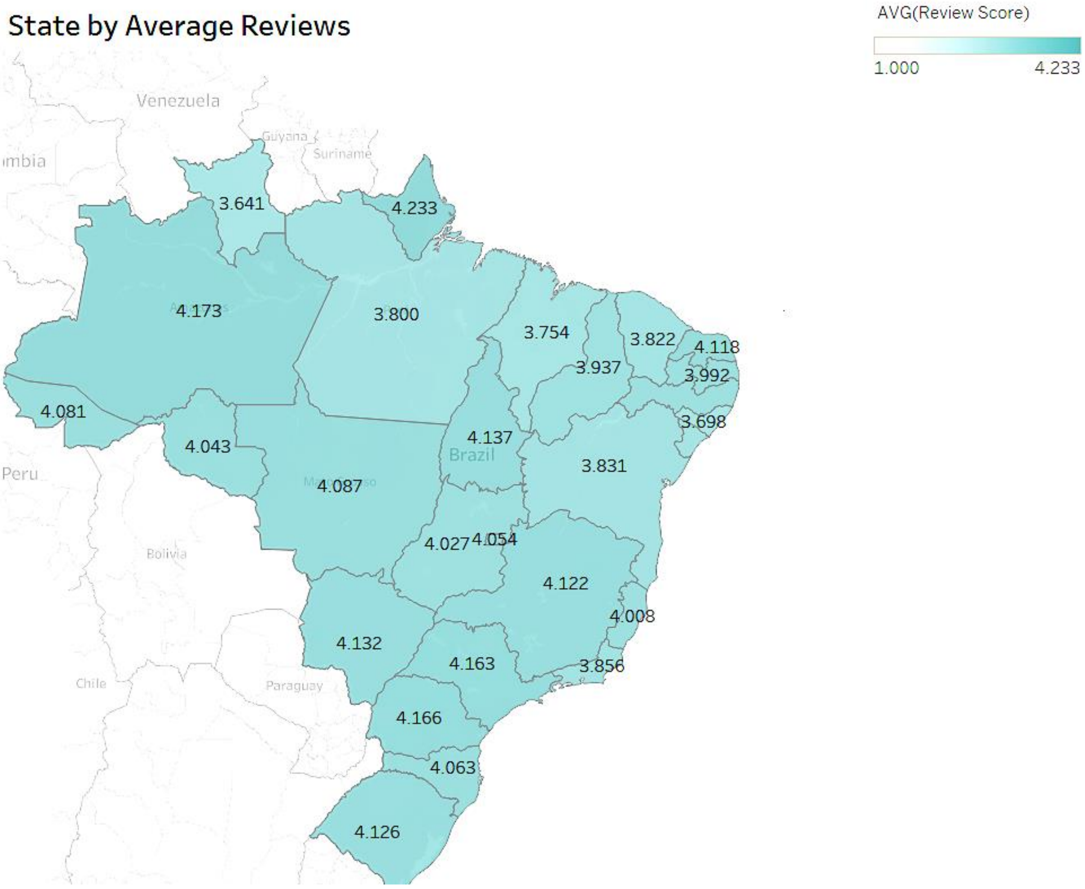


Figure 3: Average Review by State

This stage facilitated the formulation of the data mining problem and project plan. Ultimately, a clearer picture was formed on whether the assumptions made during stage one concerning informativeness, representativeness and quality were accurate and correspond with expectations (Berthold et al., 2010). Consequently, all attributes were selected, formulating our hypotheses (Appendix-2).

The table below describes data quality.

Quality	**Assessment**
Accuracy	The data is fairly accurate, although there were some missing values, errors and duplicates. A relatively small number of false positives identified.
Completeness	The data necessary to construct the model included a small number of missing values, which were removed.
Timeliness	The data is outdated as it reflects customer engagement behaviour prior to the COVID-19 pandemic.

Table 2: Data Quality Assessment

Data Preparation

In this stage, the dataset was modified ensuring the modelling techniques are best supported and least biased, maximising the model's predictive abilities. This stage consisted of four steps as described by Berthold et al. (2010): selection, correction, construction and integration.

Based on the results of the data understanding phase, sections of datasets were selected due to their relevance to the problem. Given that Nile aims to predict positive reviews, a binary variable was created and defined as: ‘Positive’ (1) for reviews with 4-5 stars and ‘Negative’ (0) for reviews with 1-3 stars. Introducing this variable reduced complexity, sped up training, and enabled easier interpretation of results. In the next step, individual errors were addressed to improve data quality. Errors, such as false positives, were identified and removed to ensure reliability of predictions. Missing data was deemed insignificant relative to the overall dataset and excluded to maintain consistency. New variables were created to provide the model with additional inputs such as ‘total price’. These engineered features served as helpful indicators, so that the model does not have to re-learn the usefulness of such attributes. Finally, the data was integrated into one table using horizontal integration, with customer and order id as identifiers.

Modelling

In this stage, a collection of models was developed to identify the most appropriate for the problem. To do so, a deeper understanding was gained on the underlying principles of the data analysis methods used within Berthold et al.'s (2010) four step procedure: selecting the model class, defining the score function, applying the algorithm, and validating the results. A binary classification approach was deemed most appropriate based on the company's objective. The dependent and independent variables were split as X and Y, their combinations are as follows:

Y	X
Review	Delivery time in days
	Total order price
	Early/Late status
	Number of instalments
	Average category score
	Average seller score
	Average State Score
	Quantity of photos
	Length of description

Table 3: Dependent and independent variable combination

It is important to note that, although the application of the selected score function enabled the comparison between models, it was not sufficient to identify the best one. To do so, different algorithms were applied, and feature importance and confusion matrixes (Appendix-3) were utilised to support the evaluation of the models.

Overall, Random Forest offers the most reliable performance, while GBDT and XGBDT provide good prediction in true positives. It was observed that Random Forest had the best balance in variable weighting, with XGBDT coming in second, and GBDT over-relying on one variable. The confusion matrixes demonstrate that Random Forest has most predictions falling within true positives and true negatives, while the other two still indicate mostly true positives, but also containing a notable number of false negatives, which indicate the models could be improved regarding identification of total positives. Considering that Random Forest achieves an unrealistically high accuracy, and its weak models are unable to learn from each other, XGBDT was selected. Decision trees are ideal for Boolean data and, when combined with the XGBoost library, this algorithm is scalable and allows the formation of decision trees in parallel, offering speed, efficiency, reducing bias and underfitting (NVIDIA, 2023).

Evaluation

This section evaluates the achieved performance metrics, providing recommendations for improvement.

Performance Metric	Score
Precision	0.906
Recall	0.716
F1 Score	0.762

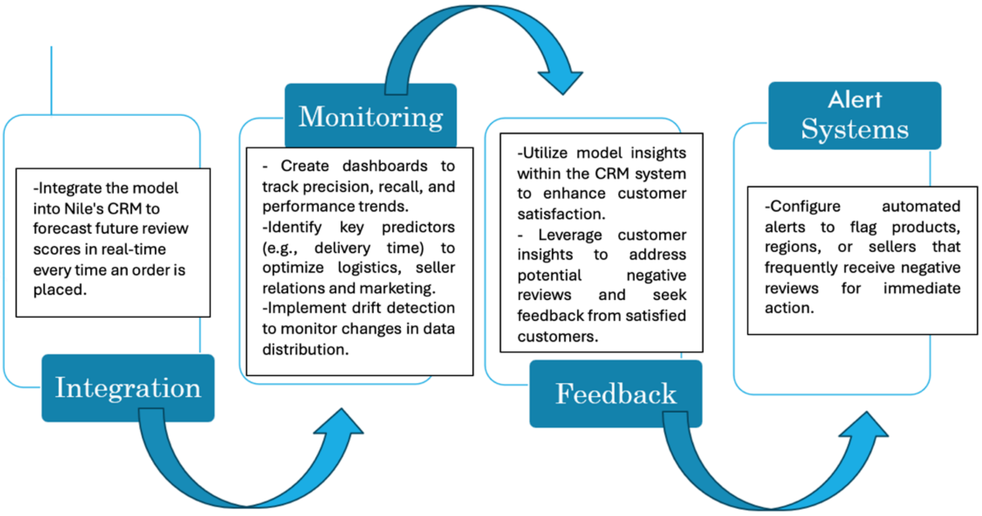
Table 4: Performance Metrics

Having identified false positives as being of higher risk, precision was chosen as the evaluation matrix. Basing predictions on false positives can result in missed opportunities, where the company wastes resources and neglects customers who would actually leave positive reviews. A higher precision score indicates that less false positives are identified by the model. In this case, a precision of 0.906 means that 90.6% of total positive predictions were correct.

Unlike precision, recall identifies all positive labels by incorporating false negatives in the equation. A recall of 0.716 signals that the model identified 71.6% of total positives, suggesting that 28.4% of positive instances (false negatives) were missed. Finally, the F1 score suggests that the model is 76.2% effective in predicting positives, considering both precision and recall.

Deployment

The model predicts which customers will leave positive reviews, providing valuable insights to support decision-making regarding customer satisfaction.


Integration

-Integrate the model into Nile's CRM to forecast future review scores in real-time every time an order is placed.

Monitoring

- Create dashboards to track precision, recall, and performance trends.
-Identify key predictors (e.g., delivery time) to optimize logistics, seller relations and marketing.
-Implement drift detection to monitor changes in data distribution.

Feedback

-Utilize model insights within the CRM system to enhance customer satisfaction.
- Leverage customer insights to address potential negative reviews and seek feedback from satisfied customers.

Alert Systems

-Configure automated alerts to flag products, regions, or sellers that frequently receive negative reviews for immediate action.

Figure 4: Deployment process

In the future, it is advised to integrate further attributes to the model, like customer purchase frequency. The company could also account for seasonal trends and past review patterns, which influence behaviour and could enhance model accuracy. Additionally, regional review scores can be used to understand local consumption patterns and preferences, and tailor marketing strategies targeting those customers.

Conclusion

Overall, the model focuses on key attributes deemed crucial in predicting future review scores. The binary classification approach ensures a clear distinction between positive and negative reviews, keeping the model and its interpretations simple. The results indicate the model’s potential in identifying positive reviews, serving as a valuable asset for Nile. Most importantly, the model focuses on the identification of true positives to ensure efficient use of resources. Finally, the report provides valuable insights that can be used in the future to enhance the model and strengthen Nile’s customer satisfaction and engagement.

Bibliography

Berthold M.R., Borgelt C., Höppner F., Klawonn F. (2010). Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data. 1st ed. Springer.

Grus, J. e. (2019). Data science from scratch: first principles with Python. Second. Sebastopol, CA: O'Reilly Media.

Harter, A., Stich, L. and Spann, M. (2024). The Effect of Delivery Time on Repurchase Behavior in Quick Commerce. Journal of Service Research, p.10946705241236961.

Indajang, K., Candra, V., Sianipar, M.Y., Sembiring, L.D. and Simatupang, S. (2023). The effect of service quality and price on customer satisfaction. Ekonomi, Keuangan, Investasi Dan Syariah (Ekuitas), 4(3), pp.942-950.

Kavlakoglu, E. and Russi, E. (2024). What is XGBoost? Available at: https://www.ibm.com/topics/xgboost (Accessed: 24 November 2024).

NVIDIA. (2023). XGBoost - What Is It and Why Does It Matter? Available at: https://www.nvidia.com/en-us/glossary/xgboost/ (Accessed: 27 November 2024).

Provost, F. and Fawcett, T. (2013). Data science for business: what you need to know about data mining and data-analytic thinking. 1st ed. Sebastopol, CA: O'Reilly Media.

Shavitt, S. and Barnes, A.J. (2020). Culture and the consumer journey. Journal of retailing, 96(1), pp.40-54.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

Wang, J.C. and Dong, L.J. (2024). Risk assessment of rockburst using SMOTE oversampling and integration algorithms under GBDT framework. Journal of Central South University, 31(8), pp.2891-2915.

Wirth, R. and Hipp, J. (2000). April. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).

Appendix 1 – Model Assumptions

	Assumption
1	The data is representative of the broader population for accurate predictions.
2	Missing data were adequately handled to prevent bias.
3	Positive reviews are sufficiently represented in the data set
4	The data does not contain significant outliers that could heavily influence predictions.
5	The relationship between the dependent and independent variables is stable over time.
6	Chosen attributes are accurately recorded.
7	Chosen attributes, influence customer satisfaction and thus review score.

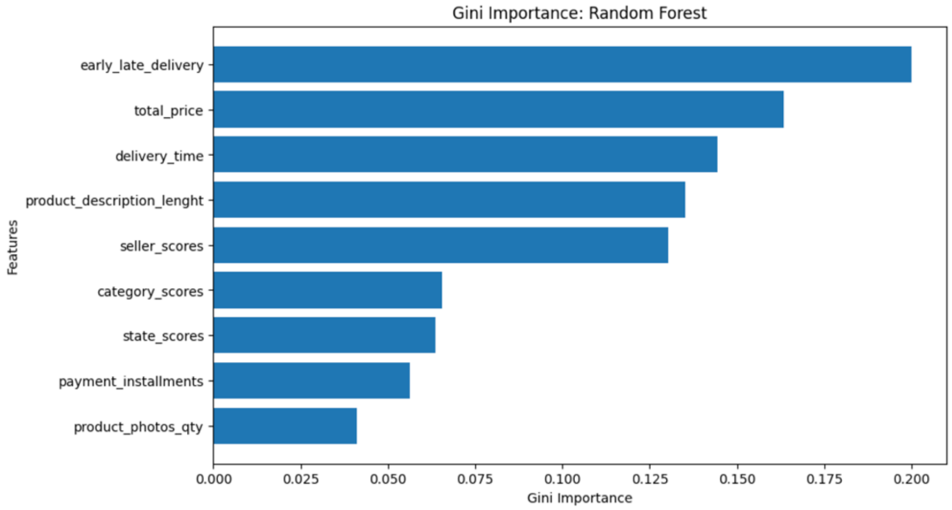
Appendix 2 - Hypotheses

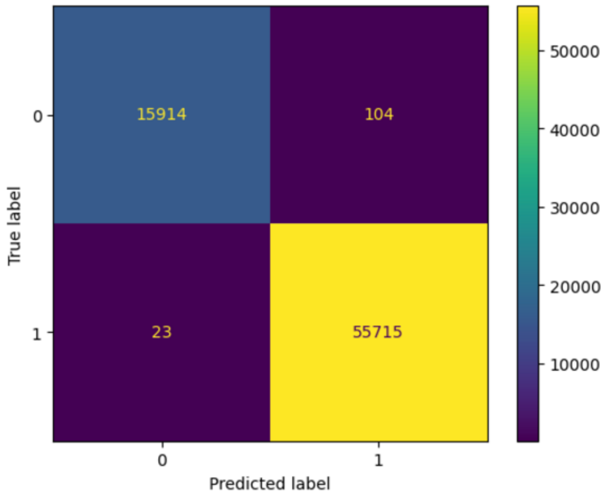
Hypothesis	Logical Explanation
Delivery time affects customer satisfaction.	Customers often expect timely deliveries. Delays or faster-than-expected deliveries can impact satisfaction.
Total cost affects customer satisfaction.	High cost may lead to dissatisfaction if not justified by quality, while reasonable costs can boost satisfaction (Indajang et al., 2023).
Early/Late delivery time affects customer satisfaction.	Late deliveries often result in frustration, while early deliveries may exceed expectations, improving satisfaction and encouraging repeat purchases (Harter et al., 2024).
Number of instalments made affect customer satisfaction.	Flexible instalment options can enhance satisfaction, while limited options can reduce it.
Some product categories have a stronger impact on customer satisfaction than others.	Some customers value certain categories more, associating them with higher quality or importance, leading to greater impact on satisfaction.
Some sellers have a stronger impact on customer satisfaction than others.	Some sellers have a greater impact on customer satisfaction due to factors like product quality, service and reputation.

Customer satisfaction varies across different states.	Customer satisfaction can vary across different states due to factors like culture/regional preferences (Shavitt and Barnes, 2020).
Photo quantity affects customer satisfaction.	A higher number of photos provides better visual information that can increase satisfaction, whereas too few to no photos may lead to dissatisfaction.
Description length affects customer satisfaction.	Detailed descriptions help customers make informed decisions, thus can enhance satisfaction, while short descriptions can have the opposite effect.

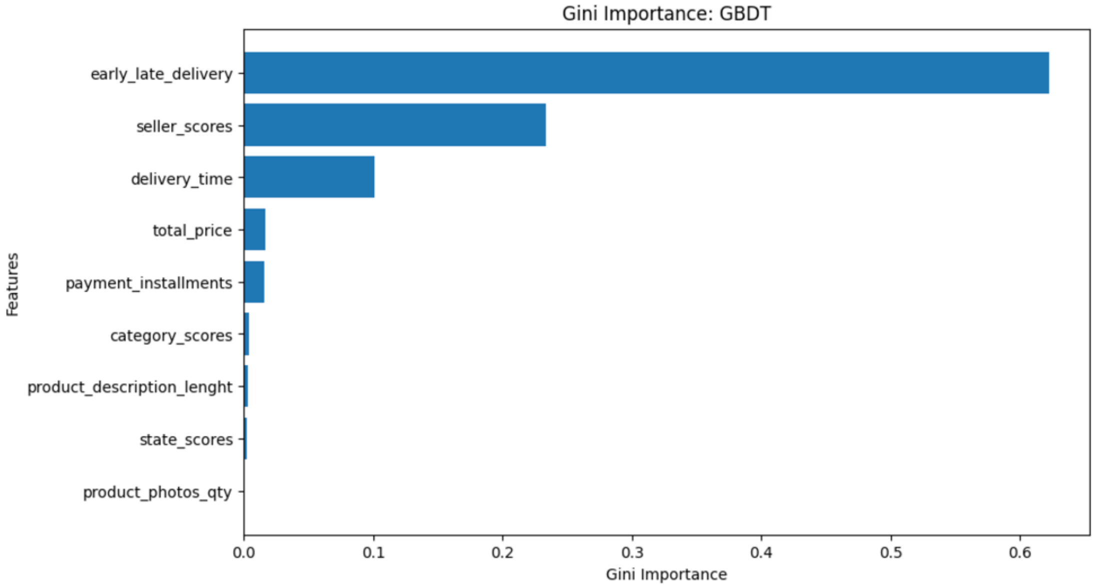
Appendix 3 – Feature importance and confusion matrixes

Random Forest:

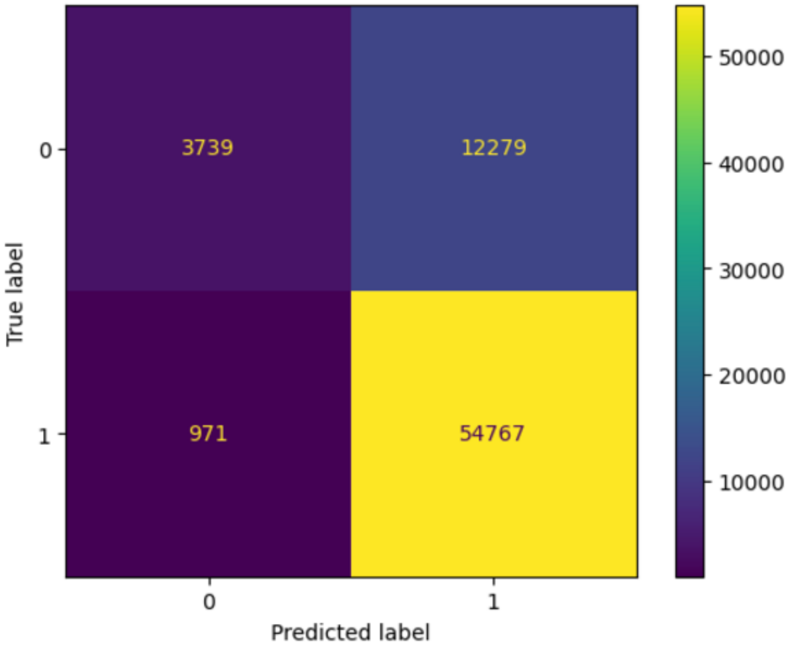




GBDT:

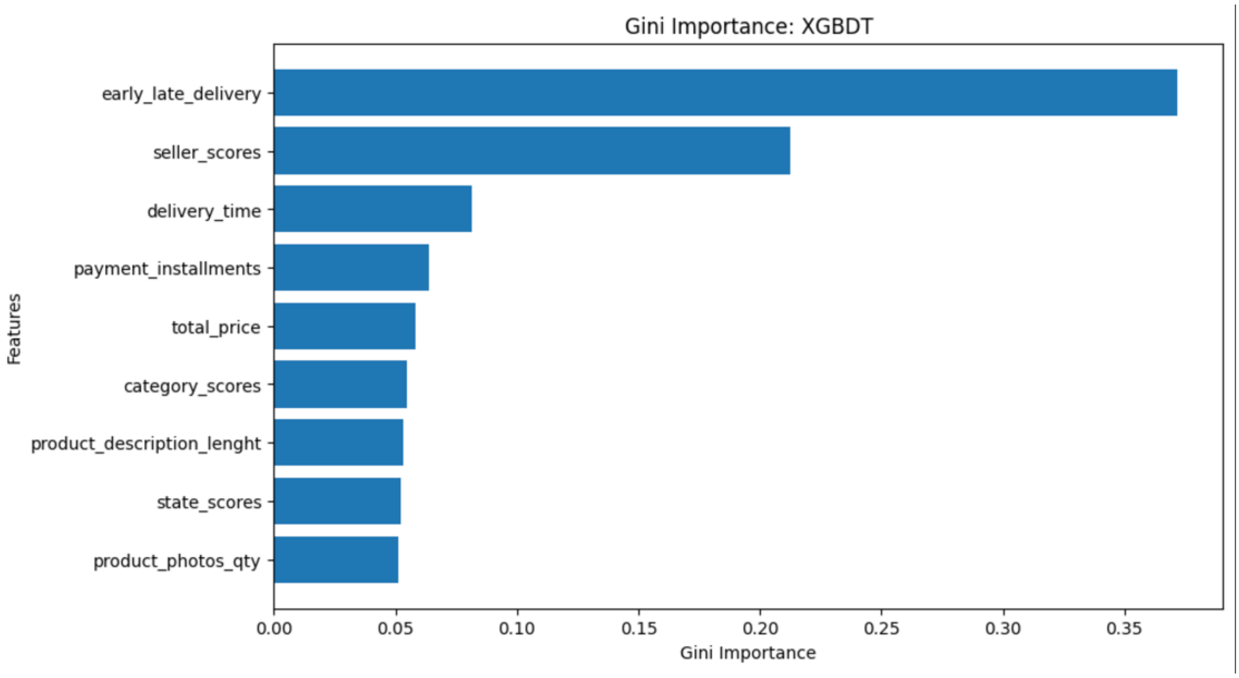


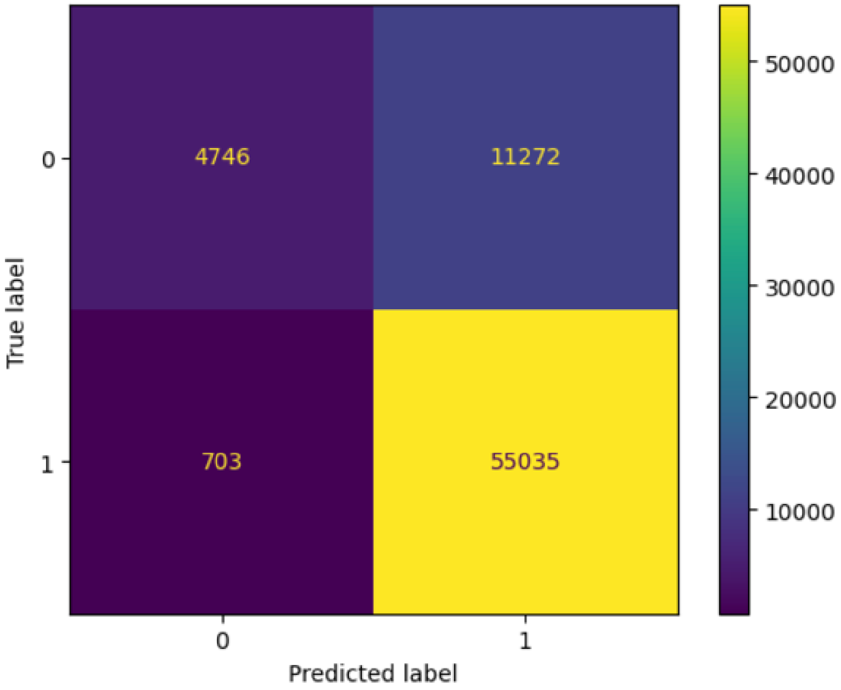
Gini Importance: GBDT
Features
early_late_delivery
seller_scores
delivery_time
total_price
payment_installments
category_scores
product_description_lenght
state_scores
product_photos_qty
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Gini Importance



3739
12279
971
54767
True label
0
1
Predicted label
0
1
50000
40000
30000
20000
10000

XGBDT:


Gini Importance: XGBDT
early_late_delivery
seller_scores
delivery_time
payment_installments
total_price
category_scores
product_description_lenght
state_scores
product_photos_qty
Features
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Gini Importance


0
1
True label
4746
11272
703
55035
0
1
Predicted label
50000
40000
30000
20000
10000