2025
OLIST E-COMMERCE ANALYSIS

Predicting Customer’s Satisfaction Through Customer’s Rating

SID:

540369759 | 540507328 | 530824099 | 540619768 | 540225237

1. Problem Formulation & Business Understanding

This section elaborates on the understanding of business context and formulates research questions.

1.1. Business Problem Definition

Customer satisfaction is a major factor in Olist's success in the context of Brazil's leading e-commerce company. That's why customer ratings directly influence future purchasing decisions, product visibility, and seller reputation.

The business question addressed in this project is "Can we predict whether a customer will be entirely satisfied with their purchasing experience, according to the information provided after their order is delivered?"

As a result, the model in this project was constructed as a binary classification problem of machine learning (ML) problems, in which:

- Customers who gave a rating score of four or more are considered "fully satisfied,"
- Customers who gave a rating score of less than four are considered "not fully satisfied."

This project can help Olist in identifying at-risk customers and implementing mitigation strategies to improve customer satisfaction outcomes by predicting customer satisfaction in advance (immediately after product delivery).

1.2. Justification of Machine Learning

Traditional rule-based approaches (which often use "if-then" statements) are insufficient due to the complexity and large amount of available data. Moreover, the non-linear and complex patterns of customer satisfaction necessitate data-driven approaches that are applicable in a variety of contexts.

The following reasons make machine learning (ML) appropriate for this problem:

- ML models can be used to evaluate large datasets and uncover hidden trends that increase prediction accuracy.
- These techniques are able to understand how hundreds of features combine to influence consumer satisfaction.
- Machine learning models can progressively enhance their performance through ongoing learning and adaptation to new data.

Business Benefits of Applying ML:

- Proactive Retention: By proactively resolving customer dissatisfaction issues, the company can increase customer retention.
- Preserve Reputation: Remove unfavourable reviews that undermine seller performance and product credibility.
- Improved Seller Management: Because dissatisfaction trends can point to inadequate seller performance. As a result, businesses can criticise and motivate sellers to raise the standard of their transactions.

Business Constraints and Limitations:

- Threshold Selection: Choosing this customer satisfaction threshold is a business decision, although in this project it is simplified to a binary (>= 4 and < 4). Model performance and commercial consequences may differ depending on the threshold (e.g., rating score < 5 as "Not Fully Satisfied").
- Changing Expectations of Customers: The model may need to be retrained and evaluated on a regular basis due to changing customer expectations from transactions or external influences.
- Intervention Cost Strategy: Outside of the company's fundamental cost structure, proactive interventions might have significant external expenditures. As a result, companies must create efficient and economical intervention plans as well as precise guidelines for when and how to respond to model predictions. When the model indicates that a customer will be "Not Fully Satisfied" when, in reality, they are "Fully Satisfied," there is a risk of "over-intervention."
- Intervention Timing: The interpretability and lead time of model prediction outputs are crucial since Customer Relationship Management (CRM) operational teams need to be able to respond swiftly to them. As a result, mitigating techniques may perform at their best.

1.3. Business Success Metrics

Table 1 Business Success Metrics

Business Success Metric	Definition	Strategic Business Value Goals
Average Rating for Customer Satisfaction	This indicator shows how satisfied customers are with their overall purchases. The business can take the necessary preventative measures if our prediction algorithm accurately detects customers which are at risk of becoming "not fully satisfied."	A higher average rating denotes better customer satisfaction and service quality, which may also be a sign of the intervention strategy's performance and the prediction model's efficacy.
Net Promoter Score (NPS)	NPS is a crucial indicator of loyalty among customers. Companies can decrease the number of possible "detractors" and convert them into "passives" or "promoters" by anticipating which consumers are most likely to be "not fully satisfied".	The success of the predictive model may indirectly affect organic growth through word-of-mouth and greater customer trust, which lowers acquisition costs through an increase in expected NPS.
Rate of Customer Retention	A “fully satisfied” customer's likelihood of making another purchase is highly predicted by their post-purchase experience. This technique keeps clients who might otherwise leave after a negative shopping experience by lowering disappointment through early intervention.	The predictive model's positive impact is the rising rate of customer retention, which benefits the business because keeping existing customers is less expensive than finding new ones.
Interval between next purchases	The average number of days between purchases by customers is taken into consideration by this metric. Consumers who are "fully satisfied" with their purchases typically make repeat purchases more quickly.	Intervention techniques can reduce the interval of customer purchases by increasing customer engagement and loyalty. Therefore, the company's long-term revenue and sales volume may rise as a result.

Trade-Off Considerations:

- Retention Value vs. Intervention Cost: Not every "flagged" situation will end in "not fully satisfied." As a result, companies need to weigh the importance of preserving consumer pleasure against the expense of taking preventative measures. Prioritising high-risk customers makes financial sense when interventions are inexpensive (such as sending preventative communications or special coupons).
- Precision vs. Recall: A model with high precision will have reduced recall since precision and recall cannot coexist (Upadhyay, 2020). As a result, while a model with a low recall may miss some "Fully Satisfied" customers, one with a high recall may contain more false positives.
- Short-Term Expenses vs. Long-Term Benefits: Discounts, coupons, or shopping cashback may end up in short-term costs for intervention efforts. Nonetheless, these approaches can raise lifetime value and long-term retention. Budgetary restrictions must be taken into account; However, investing more in retaining current customers can lead to increased loyalty and repeat business.

2. Exploratory Data Analysis (EDA)

This section describes the process of exploring the dataset and connecting with model selection.

2.1. Overview of the Dataset

This project utilizes the Olist e-commerce dataset, a comprehensive collection of real transaction records from a Brazilian online marketplace between 2016 and 2018. The dataset comprises multiple interconnected tables, capturing details about orders, customers, payments, products, delivery logistics, and customer reviews. To address the business objective, relevant tables and variables were selected based on the data dictionary, variable types, and relationship keys. After evaluating the structure of the dataset, the decision of table inclusion and exclusion in the analysis is summarised below:

Table 2 Decision of Variables Usage

Table	Decision of Usage	Reason
Customers Dataset, Order Reviews Dataset, Order Items Dataset, Products Dataset, Orders Dataset, Order Payments Dataset	Incorporate in the modelling process	Provides information that would be relevant to predict the Review Score variable, but needs feature engineering
Geolocation dataset, Product Category Name Translation, and Sellers Dataset	Not used	Focus on location granularity and seller details, which are outside the scope of the current analysis.

2.2. Data Merging Process & Key Descriptive Statistics

To ensure a consistent structure, the dataset merging process is shown by Figure 1 Dataset Merging Process. After merging the data, there are 99,331 observations with 31 variables (both numeric and categorical).

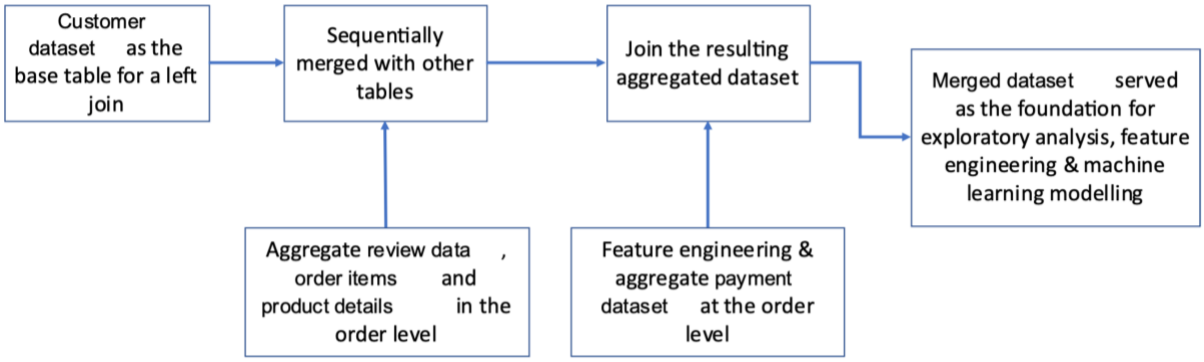

Figure 1 Dataset Merging Process

The original review score as the target variable ranges from 1 (very dissatisfied) to 5 (very satisfied). As seen in Figure XX, the distribution is heavily right-skewed, with a dominant concentration of 5-star reviews (57,328 instances). The second-largest group is 4-star reviews (19,142), followed by progressively fewer low ratings. This skew suggests a positivity bias, which is common in post-purchase survey data where satisfied customers are more likely to respond.

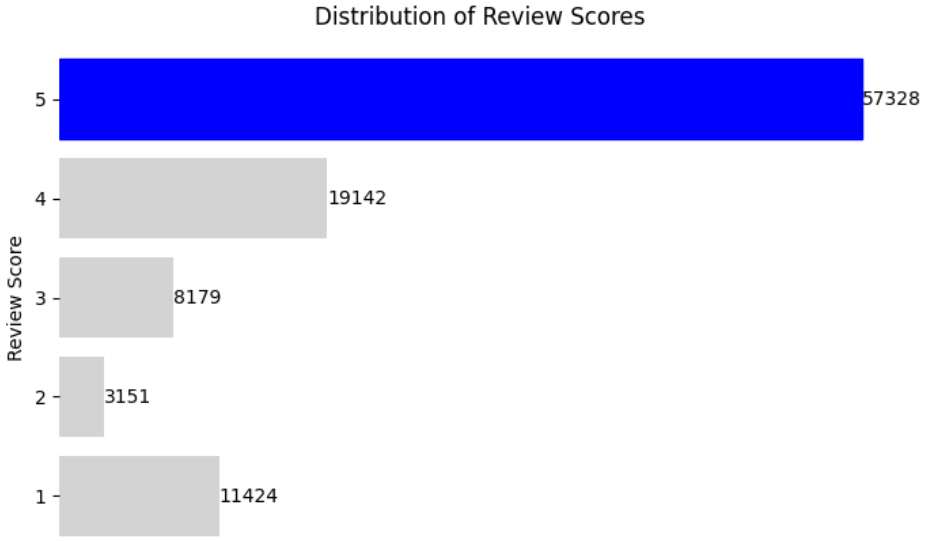

Figure 2 Bar Chart of Review Score

In addition to examining individual distributions and data quality, the relationship between the target variable and other variables was examined to identify which features would be influential for predicting customer satisfaction. For example, longer delivery durations are linked to lower satisfaction. The correlation matrix of numerical variables can be found in Figure 3.

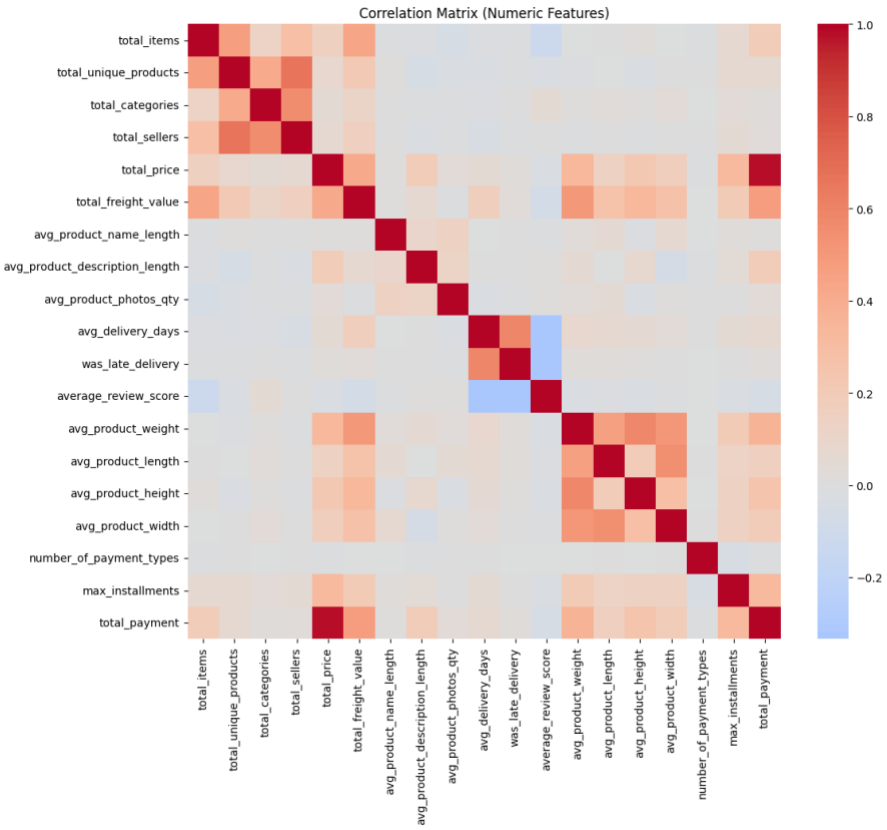

Figure 3 Correlation Matrix

2.3. Identification of Data Quality Issues and Biases

A critical part of the exploratory analysis involved identifying and addressing potential data quality issues and biases that could affect model performance or distort business insights. This was an iterative process, where initial discoveries during data merging and EDA guided multiple rounds of cleaning and feature engineering.

Table 3 Data Quality & Bias Identification

Issue	Method	Treatment
Missing Value	Four iterations of checking. 1st iteration of checking resulted in 17 variables having missing values, the highest is average delivery days (2.98%), and the lowest is total payment (only 1 observation)	Column dropping and inputting value (explained more in the next part)
Duplicate Rows	Checked for any duplicate observations.	Null duplicate value
Inconsistent Values or Outliers	Identify any inconsistent values or outliers from descriptive statistics and distribution plots. Five orders showed extreme delivery durations, where the actual delivery date was far later than the estimated date. These cases could reflect logistical errors or system delays.	Calculated derived features like was_late_delivery. Retained outliers as they may represent valuable edge cases, particularly for dissatisfaction prediction.
Non-normal distribution of variables	Perform descriptive statistics and a histogram plot for numerical variables to check the central tendency, skewness, and kurtosis	Perform feature engineering to standardise or scale variable if needed
Potential bias	The distribution of review scores is highly skewed toward 5-star ratings. Over 50% of reviews are	Transform the target variable into a binary classification

Issue	Method	Treatment
	perfect scores, indicating a likely positivity bias (Satisfied customers may be more likely to leave reviews)	

2.4. Link Between EDA Findings and Modelling Decisions

Key EDA Findings and the implications for the machine learning modelling process are displayed in Table 4.

Table 4 EDA Insight

EDA Insight	Modeling Decision
Orders and items span multiple rows across tables	Aggregated order-level data (e.g., total items, total price) during preprocessing.
Some orders lack product details or delivery dates	Filtered out orders without reviews and created flags or imputations for other missing values.
Some features need to be modified to fit the modelling required	Performed feature engineering
Some features have extreme value based on descriptive statistics	Use method that is robust to outliers
Some possible significant predictors are categorical variables	Use feature engineering to encode the variable and choose a modelling that can account for multiple categories, such as CatBoost
Several numerical variables show significant correlation with the target variable	These relationships guided feature selection and highlighted the importance of incorporating both domain knowledge and statistical evidence in choosing predictors
Review scores are right-skewed, with 77% of customers marked as satisfied	Transformed review_score into a binary classification problem and prepared to handle imbalance using evaluation beyond accuracy.
Review scores show class imbalance	Used precision, recall, F1-score, and AUC to evaluate model performance and fairness.

3. Feature Engineering

This feature engineering section focuses on combining and transforming the raw dataset from Olist, which is at the item or product level, into the order level. It will improve the prediction power in the model and improve the interpretability for every independent variable that will have an impact on the target variable.

3.1. Created Features

The features below are selected and engineered based on the insight from the EDA and align with the objective of the prediction, which is to have more information about customer satisfaction through their rating.

Table 5 Feature Engineering and Business Implications

Feature Name	Description	Business Implication
total_items	Total items in the order	A higher number indicates bigger purchases that affect the delivery process, which leads to a higher risk.

Feature Name	Description	Business Implication
total_unique_products	Count of distinct product in the order	Presents the variety of items purchased, which might be from different sellers. This might lead to different packaging standards that affect customer satisfaction.
total_categories	Number of unique categories	Relevant for marketing insights and delivery coordination because its diversity might signal bundling behavior across categories.
total_sellers	Number of sellers in the order	Different sellers in one order may have longer delivery times or coordination failures, impacting customer experience.
total_price	Total price of all items	High-value orders are significant to be perceived. Dissatisfied high spenders lead to significant revenue loss.
total_freight_value	Total freight cost per order	Detecting any overpriced shipping costs that will negatively impact customer satisfaction.
avg_product_name_length	Mean character length of product names	Showing the quality of product listing. Clearer names can increase customer understanding and expectations.
avg_product_description_length	Mean length of product descriptions	The transparency of the product that is reflected by well-written prediction will bring more understanding and better customer expectations.
avg_product_photos_qty	Average number of product photos	Visual presentation creates buyer confidence. Fewer photos could be linked to poor reviews or bad expectations.
avg_product_weight	Mean product weight in grams	Heavier products might lead to more problems and longer delivery times, which affect customers' satisfaction.
avg_product_length	Mean length of products in cm	Help assess the complexity level of packaging and the delivery times.
avg_product_height	Mean height of products	Help assess the complexity level of packaging and the delivery times.
avg_product_width	Mean width of products	Help assess the complexity level of packaging and the delivery times.
avg_delivery_days	Days between purchase and delivery	Directly correlated to customer experience, longer delays will lead to bad reviews and customer churn.
was_late_delivery	1 for delivery exceeded estimated date	Late delivery is the main consideration for customer satisfaction. Late deliveries will lead to bad reviews and customer churn.
average_review_score	Mean review score per order	This variable is a guide for businesses to know how their customer satisfaction is, which could be a baseline for evaluating their customer service.
review_score_binary	1 if review score ≥ 4, else 0	This is the target variable, which is created to predict whether a customer is satisfied or not for every order that they create.
number_of_payment_types	Number of distinct payment methods used	More payment methods might indicate complex payment behavior or bigger purchases. This can influence fraud and customer satisfaction.
main_payment_type	Most valuable payment type used per order	Having more insights into the main payment type that is used by the customer to pay a bigger proportion of the order.
max_installments	Maximum installments count	Higher installments might indicate to the customer having a price sensitive behavior. The number of installments available will impact the customer experience.
total_payment	Final payment value including all installments	Used pricing validation, loyalty strategy, and business revenue calculation.

3.2 Feature Transformations and Engineering Techniques

- Aggregation: Item-level data are aggregated to the order level. This clearly helps the business to have more concise data. It will bring benefits to the model prediction preparation and improve their operational decisions because most of them are related to the order level.
- Binary Transformation: Categorise the yes or no variable to simplify the classification task by focusing on more targeted business outcomes, which give more actionable insights for business stakeholders. The trade-off is that by using binary, it is easier to classify the target variable, but it's harder to get the full picture as loss of information is present.
- Averaging: Helping the business to have more insights, such as delivery, packaging, listing, etc, at the order level, which leads to better decision making.
- Labelling: Categorical encoding for categorical variables is a crucial step for dataset preparation before modelling.

3.3 Handling Missing Values and Outliers

The table below describes the missing value treatment.

Table 6 Outliers and Missing Values Handling

Variables	Method	Justification
'average_review_score', 'number_of_payment_types', 'main_payment_type', 'max_installments', 'total_payment'	Dropped	The number of missing values is below one percent.
'avg_delivery_days', 'total_items'	Filled with –1	The number of missing values is below three percent. The –1 is used to represent unknown and keep the data type as numerical as the same time.
'avg_product_name_length', 'avg_product_description_length', 'avg_product_photos_qty', 'avg_product_weight', 'avg_product_height', 'avg_product_length', 'avg_product_width'	Use Median	The number of missing values is one to two percent and it is numerical value. Therefore, median is the good method to replace missing values

3.4 Feature Selection Process

Table 7 below explains features that are excluded from the model.

Table 7 Excluded Features

Feature	Explanation
order_id', 'customer_unique_id'	Those are unique value that does not give any impact to the model
'order_purchase_timestamp', 'order_estimated_delivery_date'	Date format variables cannot be used for the modelling

Feature	Explanation
'total_sellers','total_unique_products' ,'total_payment', 'total_categories''avg_product_name _length', 'number_of_payment_types'	Excluded due to high correlation which is shown by high VIF (VIF score more than 10)
'average_review_score'	Excluded due to data leakage
'customer_city'	Used only in CatBoost modelling due to high cardinality. However, that model can handle the cardinality problem.

4. Recommended Models & Justification

This section describes the potential models for the classification problem and a model recommendation for deployment.

4.1. Overview of Recommended Models

To address the business objective, structured experimentations using a variety of machine learning models were conducted. The models evaluated include Logistic Regression, Random Forest, LightGBM, XGBoost, and CatBoost, with at least two iterations per model. Detailed performance metrics and changes between iterations, including tuning parameters, are documented in Appendix A.

4.2. Justification for the Recommended Models

Based on this experimentation, the following three models are recommended:

1. Baseline Model: Logistic Regression
   Logistic regression was selected as our baseline due to its simplicity, interpretability, and low computational cost. As discussed in Appendix A, this model provided a transparent starting point and helped us evaluate the added value of more advanced techniques. After refining the definition of the target variable, the model demonstrated reasonable performance but remained limited in capturing non-linear patterns and complex feature interactions.

2. Best Single Model: CatBoost
   CatBoost emerged as the best-performing individual model, offering superior performance across key metrics such as precision, recall, and F1-score. In addition to its predictive strength, CatBoost is particularly well-suited to our dataset as it handles categorical variables natively and is robust to class imbalance and overfitting. This makes it a strong candidate for deployment in a business context where interpretability and performance must be balanced.

3. Model Stack
   To further enhance performance, we developed a model stack, combining the predictions of multiple models to leverage their complementary strengths. The stacked model provided marginal improvements in performance metrics over the best single model. However, it introduces increased complexity and reduced interpretability, which must be weighed against the marginal gains. Details on the structure and results of the stacked model are provided in Appendix A.

The comparison summary of the three recommended models can be seen in Table 8.

Table 8 Comparison of Selected Models

Model	Type	Interpretability	Handling Categorical Features	Performance (F1)	Business Suitability	Recommendation
Logistic Regression	Linear, Baseline	High	Requires encoding	Moderate	Transparent, simple	Baseline
CatBoost	Tree-based Boosting	Moderate–High	Native handling	High	Strong balance	**Best Single**
Model Stack	Ensemble	Low	Varies	Slightly higher	Complex to maintain	Optional

4.3 Recommended Model for Deployment

While the model stack yielded slightly better performance, CatBoost is recommended as the final model for deployment due to its:

- High predictive accuracy,
- Native handling of categorical features,
- Strong balance between business interpretability and technical robustness,
- Lower implementation complexity compared to a stacked ensemble.

This recommendation ensures that the solution remains practical, scalable, and aligned with business needs, while maintaining strong performance in predicting customer satisfaction.

5. Model Evaluation and Performance Analysis

This section presents a comprehensive evaluation of the three recommended models: the baseline model (Logistic Regression), the best single model (CatBoost(Tuned)), and a Stacked Model (Benchmark version of CatBoost, LightGBM, and XGBoost with Logistic Regression as the meta-learner). The evaluation focused on the model's predictive accuracy, robustness, and business relevance. The goal is to determine the most suitable model not only based on technical performance but also on its interpretability, reliability, and suitability with practical business decision-making.

5.1 Performance Evaluation of Three Recommended Models

Table 9 summarises model performance using classification model key metrics. Accuracy measures the overall correctness of predictions, while precision reflects how many of the customers predicted as satisfied were truly satisfied. Recall captures how well the model identifies truly satisfied customers. The F1 Score provides a balanced measure of precision and recall. AUC indicates the model’s ability to distinguish between satisfied and unsatisfied customers, and log loss evaluates how confident and accurate the model’s probability predictions are. Together, these metrics ensure the selected model performs reliably and is suitable for business implementation. Figure 4 compares the performance of three classification models, plots the True Positive Rate (sensitivity) against the False Positive Rate, providing a visual representation of each model’s ability to distinguish between satisfied and unsatisfied customers.

Table 9 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
Logistic Regression	0.8179	0.8272	0.9652	0.8909	0.7279	0.457731
CatBoost (Tuned)	0.8201	0.8271	0.9691	0.8925	0.7436	0.4459
Stacked Model	0.8191	0.8257	0.9699	0.8920	0.7415	0.4481

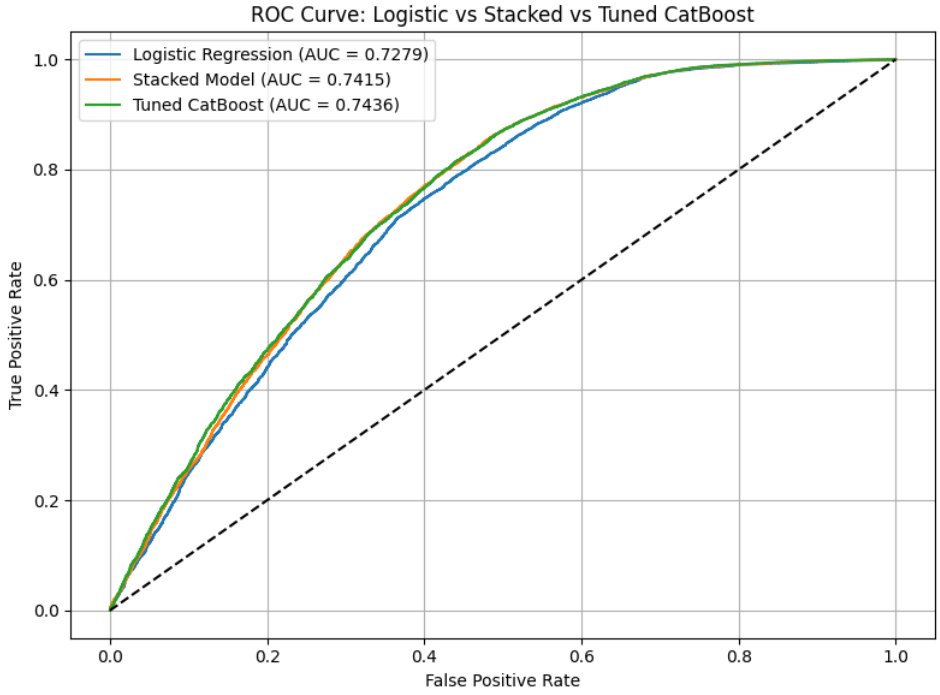

Figure 4 ROC Curve

Based on Table 9 and Figure 4, CatBoost outperforms the other two models in all evaluation metrics. It achieves the highest recall and AUC, with the lowest log loss, showing strong classification ability and high confidence in its probability estimates compared to other models. Closely following the CatBoost, the stacked model has its respective performance metrics, showing great potential for predictive ability, but it introduces significant interpretability challenges. Lastly, logistic regression has the lowest performance compared to other models, but its strength is the explainable nature of its prediction, providing transparency.

5.2 Robustness and Generalisation Analysis

To evaluate model stability and generalisability, 5-fold cross-validation was conducted. This approach ensures performance is not based on a single data split and provides a more reliable estimate of how each model would perform on unseen data. Table 10 evaluates model performance across 5-fold cross-validation based on performance matrices.

Table 10 Models Performances After Cross Validation

Model	Value	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	**Mean**	0.8183	0.8276	0.9653	0.8911	0.7248
	Std Dev	0.0030	0.0010	0.0033	0.0019	0.0053
Stacked Model	**Mean**	0.8199	0.8259	0.9711	0.8926	0.739
	Std Dev	0.0022	0.0009	0.0027	0.0014	0.0042
CatBoost (Tuned)	**Mean**	0.8202	0.8263	0.9707	0.8927	0.738
	Std Dev	0.0024	0.0011	0.0025	0.0015	0.0049

Based on the table, all three models maintained consistent performance across folds, as evidenced by the low standard deviations in key metrics. Logistic Regression showed the lowest AUC mean (0.7248) and the highest standard deviation (0.0053), highlighting its comparatively weaker and less stable ability to separate satisfied from unsatisfied customers. CatBoost (Tuned) delivered moderate results, with an AUC of 0.7380 and a standard deviation of 0.0049, demonstrating a strong balance between performance and reliability. The Stacked Model achieved the highest AUC (0.7390) and the lowest standard deviation (0.0042), indicating both superior predictive power and the most consistent generalisation across data folds. Overall, these results confirm that both CatBoost and the Stacked Model generalise well, with CatBoost showing strength in consistent performance in other key metrics and business interpretability. Logistic Regression remains a viable benchmark model with dependable results, but limited classification ability in comparison.

5.3 Error Analysis

Confusion matrices further highlight how each model handles prediction errors. Table 11 provides models for error in identifying False Positives and False Negatives.

Table 11 Confusion Matrices of Models

Model	False Positives	False Negatives
Logistic Regression	3,664	670
Stacked Model	3,736	549
CatBoost (Tuned)	3,698	563

While the Stacked Model captures more satisfied customers (lowest false negatives), it also produces the most false positives, potentially misguiding retention strategies. CatBoost provides the best balance, minimising both types of errors, making it the most reliable for business deployment.

5.4 Business Implications

Model evaluation metrics were further prioritized to be based on operational decision-making value. While traditional classification metrics (Accuracy, Precision, AUC) are important to see the general model performance. It is also important to look beyond technical metrics; business impact was evaluated through financial outcomes, such as retained customer value, intervention cost,

and ROI. These impacts were calculated through confusion matrix outcomes and cost assumptions.

Two cost assumptions were used for this calculation are :

- **Loss per dissatisfied:** 160 BRL. The expected loss of revenue when the model wrongly predicts a dissatisfied customer as satisfied. This value is derived from the average of the overall payment.
- **Recovery cost:** 80 BRL. This is the cost for following up with a dissatisfied customer (vouchers, support). This value is derived from 50% of the average payment. Kim (2019) found that price promotion can support customer retention, and Lee and Chen (2018) argued that the optimal discount percentage for customers is 50%.

Table 12 Business Implication Costs

Model	Retained Value (BRL)	Missed Loss (BRL)	Wasted Effort (BRL)	Total Intervention Cost (BRL)	Net Business Impact (BRL)	ROI (%)
LogReg	2.812.320	107.200	293.120	400.320	2.412.000	7,03
CatBoost(T)	2.829.440	90.080	295.840	385.920	2.443.520	7,33
Stacked	2.831.680	87.840	298.880	386.720	2.444.960	7,32

Table 12 shows that CatBoost has the best trade-off between business value, model performance, and interpretability. For the stacked model, it has a Higher Net Business Impact but is low on interpretability. The last place is Logistics, where it is the most interpretable, has lower model and financial performance. Based on the model evaluation, it is recommended to deploy the CatBoost model for operational use. It provides the strongest overall performance, robustness, business-aligned insights, and ROI. While the stacked model slightly improves the prediction power, its interpretability trade-off makes it less suitable for decision-making in a business context. Logistic Regression remains valuable as a benchmark and fallback model or a tool to explain predictions when needed.

6. Business Insights

After multiple iterations and evaluation of machine learning models to predict customer ratings using Olist's e-commerce dataset, this section will interpret those technical findings into more actionable business insights. Through performance metrics evaluation, the best-performing model is CatBoost, and hyperparameter tuning was done to ensure an optimal model was generated. The ultimate objective of this project is to provide a strategic decision that can help Olist, as an e-commerce platform, improve its customer satisfaction and experience. SHAP analysis was used to identify which features are the most impactful in predicting customers' ratings as shown in Figure . The following section will highlight key features and provide targeted, data-driven recommendations to improve customers' ratings and operational effectiveness.

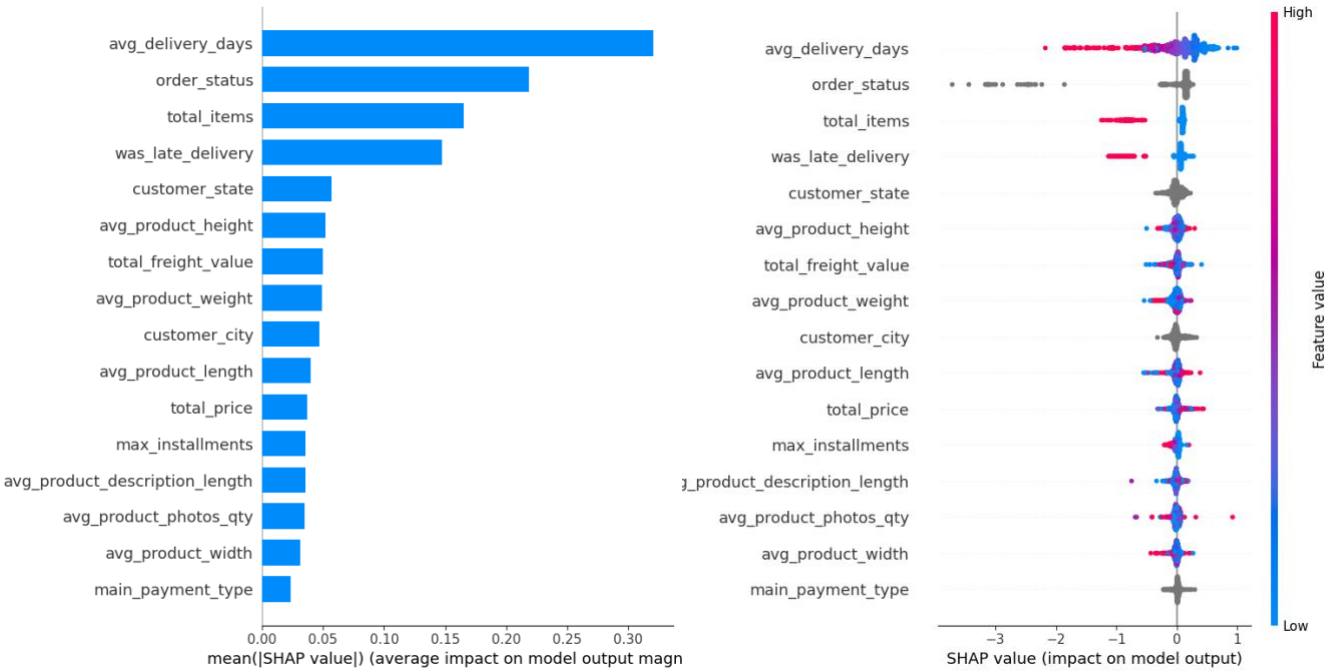

Figure 5 SHAP Value Mean (Left) and Impact on Target Variable (Right)

6.1 Key Business Insights

Some key business insights based on the analysis are elaborated below:

- **Avg_delivery_days**
  It is the most impactful feature in predicting customer ratings. Avg_delivery_days has the largest average SHAP value compared to other features, which indicates that delivery speed strongly impacts whether customers are satisfied or not satisfied, as shown through their order rating. In Figure 5, high avg_delivery_days has a strong negative value, which means long delivery times significantly increase the chance of a low rating. While low values of delivery time, shown in blue, will push the prediction towards a higher rating (satisfied). There are several fast delivery days that increase the likelihood of a low rating; It can happen in several cases where other parts of customers' expectations weren't met, such as a damaged item, a wrong item, and damaged packaging.

- **Order_status**
  In Figure 5, a certain status has a high negative impact on customer satisfaction. CatBoost processes categorical data without the need to encode. That is one of the reasons why, in this case, it's hard to interpret or look at which status negatively impacts the prediction of customer satisfaction. If we look at the data manually, most of the dissatisfied customer has either shipped, processed, or invoiced order_status.

- **Total_Items**
  A higher number of items purchased, as we can see, slightly decreases customer satisfaction. This can be due to the complexity in processing large orders or the increasing chances of issues occurring in large orders.

- **Was_late_delivery**
  This is a real indicator of whether the delivery is late or not, because longer delivery days don't always mean late delivery. In Figure 5, Late deliveries (1) are shown to have increased the likelihood of low customer ratings (not satisfied), while on-time or early deliveries can improve customers' ratings (satisfied). This aligns with the result from avg_delivery_days, which indicates that reliability matters as much as delivery speed.

- **Customer_state**
  From looking at the actual satisfaction across Brazilian states, four states (SP, MG, PR, MS, RS) have a high satisfaction rate.

- **Product attributes (avg_product_weight, height & length)**
  In Figure 5, we can see that heavier, larger products can increase the likelihood of negative SHAP values, which increases the likelihood of the customer not being satisfied (lower rating). This could potentially be due to issues with shipping/handling expectations, product damage, or higher fees.

6.2 Actionable Recommendations

1. **Prioritised Delivery Accuracy and Speed**
   With avg_delivery_days and was_late_delivery being the top features in predicting customer satisfaction, it is important to prioritise delivery accuracy and speed. This can be done through:
   - Identify sellers or customers' regions with a high rate of late deliveries. To isolate regions that are harder to reach or have a high number of delivery issues.
   - Partnering with reliable logistics providers, not just one, but with multiple, to ensure that all delivery areas are covered. This can help reduce delivery time and late arrivals.

2. **Improve Quality Control Standards for Large Orders and Products**
   Large orders with a high number of total items tend to increase customer dissatisfaction (low rating). This can be mitigated through additional quality control and packaging checks. Working with a logistics company to provide better status tracking for orders with multiple items from different sellers, all shipments are shipped correctly and delivered to the customer in great condition.

3. **Further Analysis on Region**
   Varied customer satisfaction with different regions can be an opportunity to further analyse the impact each region has on customer satisfaction. Olist can identify underperforming regions and look at which aspect can be improved for that region to increase customer satisfaction.

4. **Proactive Mitigation Strategies**
   Aside from improving delivery and quality control, Olist should take a proactive approach to mitigate customer churn because of dissatisfaction, especially for high-risk orders with the potential to receive a low rating. Based on the predictions, Olist can create mitigation strategies such as proactive customer service and preventive compensation (free voucher).

5. **SHAP & Feature Importance**
   Turn it into actionable insights in the business dashboard. For example, keep updating the average delivery time, as it has the highest effect on customer satisfaction.

6. **SOP (Standard Operating Procedure)**
   For sellers on the e-commerce platform, it is important to inform them about important features that impact customer satisfaction.

7. **Clustering analysis**
   To generate groups of customers with distinct behavioural and shopping patterns to make a more targeted marketing strategy.

8. **Comparison Analysis**
   If data from another timeframe is available, comparing the analysis of this data (before the pandemic) and after the pandemic would be interesting to see the change in customer behaviour.

6.3 Limitations of Insights and Assumptions

1. **Post Delivery Prediction:** The point of prediction of this model is after the order has been delivered to the customer, before they give their rating on the platform. This only allowed Olist to improve post-purchase outcomes and feedback, not to intervene on pre-delivered items. The downside of this method is that it can be used when the customer is already dissatisfied with the service and products, so they won't even consider placing another order. This case can be mitigated if the point of prediction is after the customer puts an order.

2. **Unobserved Factors:** This model only observed features from the customer side and their order features. Often, customer satisfaction can be reflected through the seller side of things, such as seller rating and seller location.

3. **The availability of a detailed review to perform text analysis with NLP**. Predicting value based on the text review would most likely more accurate.

4. We assumed that customer review scores really reflect what customers feel (whether satisfied or not).

7. Business Implementation and Impact

This section will outline how insights from the previous section can be operationalised and implemented within the business. To ensure this machine learning model can bring real-world business value, it must be smoothly integrated into Olist's daily operations, explained to appropriate stakeholders, continuously monitored over time, and improved. The following plan will address the step that needs to be taken in deploying this predictive model to create a long-term business impact.

7.1 Deployment Strategy

The CatBoost model will be deployed after the order has been delivered to the customer and will predict whether the customer is satisfied or not with their orders. This is so that Olist can engage with dissatisfied customers and improve aspects that impact customer satisfaction.

- Integration Point: After the order status has changed to "delivered", the model will instantly run and generate a status of either "satisfied" or "not satisfied".

- Departments Involved:
  - Customer Service Team: They can instantly follow up with unsatisfied customers to get feedback and provide incentives to ensure that customers are coming back to purchase at Olist.
  - Customer Experience Team: Create a proper program to improve customer satisfaction that can be implemented for further improvement.
  - Operations Team: Investigate delivery problems and pain points that are linked to dissatisfied customers.
  - Data Science & Engineering Team: Maintain model performance and infrastructure

- Real-World Constraints:
  - Required on-time status update and working closely with third parties that might be hard to control
  - Action must be immediately taken to avoid delaying post-delivery intervention or communication flow.

7.2 Risk, Challenges, and Mitigation Strategies

The risk mitigation plan is described in Table 13:

Table 13 Risk Mitigation Plan

Risk/Challenges	Description	Mitigation Plan
Model Drift	Customer preference or delivery process may change over time. A massive change happening in the country.	Quarterly monitor data relevancy, distribution, and performance.
Operational	Relying on performance from third parties which can be complex and hard to control. Such as, system integration and status update.	Use API to integrate with third party system, monthly review with third parties.

7.3 Monitoring and Continuous Improvement

A monitoring and continuous improvement plan to implement the model is described in Table 14.

Table 14 Continuous Improvement Plan

Key Monitoring Metrics	1. AUC for overall model performance 2. Accuracy & F2 Score for assessing classification capability 3. Improvement in Customer's Rating 4. Decrease in Complaint Rate 5. Change in NPS (higher score)
Tracking Tools	1. Dashboard for real-time monitoring 2. Notification for when model metrics dropped below certain numbers
Improvement Plan	1. Evaluate feature importance every quarter to ensure that Olist correctly monitors features that directly and significantly impact customer satisfaction 2. Feedback session from related stakeholders 3. Update model with new data every quarterly

7.4 Business Impact

Deploying this model to predict customer satisfaction through rating can help Olist to:

1. **Actively engage with customers**: Especially dissatisfied customers, to help reduce churn and complaint rates.

2. **Improve delivery accuracy and speed:** Working with logistic partners to improve overall delivery performance and flagging orders that are linked with dissatisfied customers or lower ratings.
3. **Increase operational efficiency:** Customer experience team can create a better program to improve customer satisfaction and rating through features that directly impact customer satisfaction.
4. **Overall improvement in customer trust and satisfaction:** If the model performs correctly and effectively, the long-term effect is an increase in customer trust through the NPS score and customer satisfaction through rating. Showing responsiveness in handling problems in e-commerce is one of the ways to show the customer that we care about their experience and feedback, which can turn a negative experience into an opportunity.

Overall, this model will support data-driven service excellence and decision making in the long run. It will allow Olist to create a more personalized customer experience and operationally optimized performance. With how fast customers can change between online platforms and even go back to offline stores because of a bad experience, this has made this an important and essential process in creating a sustainable e-commerce business.

References

Kim, J. (2019). The impact of different price promotions on customer retention. *Journal of Retailing and Consumer Services*, *46*, 95–102. https://doi.org/10.1016/J.JRETCONSER.2017.10.007

Lee, J. E., & Chen-Yu, J. H. (2018). Effects of price discount on consumers' perceptions of savings, quality, and value for apparel products: mediating effect of price discount affect. *Fashion and Textiles*, *5*(1), 1–21. https://doi.org/10.1186/S40691-018-0128-2/TABLES/4

Upadhyay, A. (2020, August 10). *Precision/Recall Tradeoff*. Analytics Vidhya. https://medium.com/analytics-vidhya/precision-recall-tradeoff-79e892d43134

AI Acknowledgement

We acknowledge the use of ChatGPT (https://chat.openai.com/) to evaluate and improve the academic language of our own work. We submitted our individual parts with the instruction to "Suggest improvements for our report's academic language and point out any grammatical errors that we need to address". The AI then pointed out any grammatical errors that we needed to address and suggested improvements for our academic language to better follow the style of writing that we wanted to be done. We've also used ChatGPT to help debug our code for this assignment.

Appendix A

A.1. Overview of Model Development Approach:

Predicting whether a customer would be fully satisfied (give a review score above 4) or not satisfied (give a score of 4 or lower) after receiving their order was the aim of this modelling process. This makes it possible for the company to put early intervention and mitigation plans into place, particularly when delivery performance poses a risk to customer satisfaction.

The modelling strategy adopted an iterative modelling strategy starting with more straightforward, interpretable models and working gradually up to more intricate ensemble-based models. The strategy is prioritising comprehension of the data and baseline performance over advanced algorithms. Every iteration included adjustments based on performance insights, guaranteeing that the finished model was useful and efficient.

Several models were created and assessed following a few iterations in order to investigate a variety of prediction capacities and interpretability levels:

Model	Reason(s)
Logistic Regression	It is used as a baseline because of its interpretability, simplicity, and good performance with linear connections.
Random Forest	An ensemble model that combines multiple decision trees with enhanced accuracy and resilience that effectively manages non-linearity and feature interactions.
XGBoost and LightGBM	High predictive performance and efficiency, which are particularly well-suited for structured tabular data.
CatBoost	In datasets with categorical characteristics, CatBoost frequently performs better than other boosting models due to its native handling of categorical variables.

Lastly, the technical and business relevance measures are used to assess the models. A balanced perspective of accuracy, risk, and operational viability is made possible by all multi-model methodologies employed, which eventually support well-informed recommendations for the top-performing models based on the following standards. First, the classification performance is evaluated using predictive metrics based on the model's accuracy, precision, recall, F1 Score, AUC, and log loss. Second, interpretability based on several model iterations, simpler models such as logistic regression, are preferable in order to produce outputs that business stakeholders can understand. Last but not least, business value is determined by how well the model supports proactive choices, including identifying possible dissatisfaction prior to review submission (predictive points).

A.2. Baseline Model

As a baseline model, we selected logistic regression, a widely used and interpretable classification algorithm. Logistic regression is suitable for binary classification tasks such as ours, where the target variable of review score binary takes on values of either 1 (fully satisfied, if the average review score is equal to 5) or 0 (not fully satisfied, if the average review score is less than 5).
Logistic regression was chosen as the baseline for several reasons:

- Simplicity and Interpretability: Logistic regression provides clear insights into the relationship between each predictor and the target variable. The model’s coefficients can

be directly interpreted as the effect of each input variable on the log-odds of customer satisfaction.
- Low computational cost: Logistic regression is computationally efficient and serves as a quick and reliable benchmark to compare against more complex models.
- Comparison for other models: As a basic classification method, logistic regression establishes a benchmark to evaluate whether more advanced models (e.g., random forests or gradient boosting) meaningfully improve predictive performance.

By starting with logistic regression, the value added by more complex algorithms can be assessed while maintaining a transparent understanding of baseline performance.

Initial Model Evaluation and Limitations

The initial logistic regression model, using the original definition of the target variable (*review score binary* = 1 if review score is exactly 5, else 0), yielded the following performance metrics:

Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
0.6256	0.6210	0.9015	0.7354	0.5967	0.6613

While the model achieved a high recall, it came at the cost of low precision and overall classification accuracy. This indicates that the model tended to overpredict the positive class (fully satisfied customers), likely misclassifying a significant number of dissatisfied customers as satisfied. Additionally, the relatively low AUC and high log loss suggested limited discriminative power and poor confidence in the predicted probabilities.

These limitations highlighted the need to revisit our problem framing, specifically the definition of customer satisfaction.

To better reflect real-world customer satisfaction patterns and improve model performance, we revised the target variable definition: transactions with a review score greater than or equal to 4 were now labelled as 1 (satisfied), and scores below 4 as 0 (not satisfied). This change is supported by the assumption that both scores 4 and 5 typically represent positive customer experiences.

The new class distribution became more imbalanced (77% satisfied vs. 23% not satisfied), but it aligned more closely with business expectations and customer sentiment. After applying this revised target, the logistic regression model showed significantly improved performance:

Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
0.7843	0.7837	0.9944	0.8766	0.6097	0.5165

This revision not only enhanced all performance metrics but also resulted in a model that is more reliable for downstream business applications, such as identifying likely satisfied customers or proactively addressing potentially negative experiences.

While logistic regression provides a solid and interpretable baseline, it has several limitations that prompted further experimentation, such as linearity assumptions, limited feature interactions, and lower performance with imbalanced data. These limitations led us to explore more advanced machine learning models, including Random Forest, XGBoost, LightGBM, and CatBoost. By experimenting with these models, the project aims to improve predictive performance, better capture complex customer behaviour patterns, and deliver more reliable business insights.

A.3. Model Experimentation & Refinement Process

Benchmark Model Results:

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
Logistic Regression	0.817921	0.827282	0.965200	0.890935	0.727911	0.457731
Random Forest	0.812727	0.825785	0.959336	0.887565	0.719073	0.488179
XGBoost	0.815725	0.826306	0.963336	0.889575	0.732833	0.455915
LightGBM	0.818723	0.824632	0.971283	0.891970	0.737750	0.448065
CatBoost	0.819145	0.826659	0.968324	0.891901	0.740117	0.448113

i. Logistic Regression:

The first iteration of logistic regression was conducted with the solver's default parameters ('lbfgs') and set maximum iterations of 1000. The model accurately identified the most delighted customers, as evidenced by its extremely high recall. The Precision and AUC of this model, however, indicate a poor distinction between the satisfied and unsatisfied classes. The results of the first iteration guide the second iteration by changing the solver to 'liblinear' with the same maximum iterations. The "liblinear" solver, which is better suited for smaller datasets and gives more flexibility over regularization. In comparison to the previous version, the model has improved in every metric. There were discernible improvements in precision and AUC, which improved the model's ability to differentiate between entirely satisfied and unsatisfied classes and decreased false positives. In summary, these two iterations of logistic regression do not provide significant improvement over the benchmark model (using the 'liblinear' solver) with the default maximum iteration of 100, because sometimes the default maximum iteration is better for generalisation.

Model	Accuracy	Precision	Recall	F1 Score	AUC Score	Log Loss
Logistic Regression (1st Iteration)	0.7975	0.8022	0.9784	0.8816	0.6762	0.4887
Logistic Regression (2nd Iteration)	0.8170	0.8275	0.9633	0.8902	0.7257	0.4585

ii. Random Forest:

To improve prediction, Random Forest is a good choice because of its ability to handle non-linear relationships and noisy data. The first iteration of the random forest used a number of estimators set to 100 and a max depth of 10, which is a good balance between performance and computational cost. The second iteration uses class_weight set to balanced and max depth of 10, which is to adjust the target features distribution as well as the depth of the tree. Even though our target features are not considered imbalanced, this step can ensure fairness and robustness for both classes. The second iteration performs better compared to the first one, with recall, AUC, and log loss slightly increased. Adjusting class weight can still impact performance even though the dataset is already balanced. Overall these two iteration performs better compared to the benchmark model of random forest.

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss

Random Forest(1st Iteration)	0.819103	0.821180	0.978243	0.892857	0.728248	0.454662
Random Forest(2nd Iteration)	0.784478	0.851323	0.872691	0.861875	0.730435	0.572936

iii. XGBoost:

To improve the prediction of customer review ratings, XGBoost is capable of handling high cardinality variables and correlated features. Following that capability, the first iteration uses customer city, which is a feature with a high cardinality problem, and the second iteration includes a total number of payments that are crucial for the ratings, even with a slightly high VIF score. Based on the table below, it seems that the first iteration improves the benchmark model and leads to higher AUC performance. However, the second iteration mildly decreases the performance, which slightly impacts the business to better not include a total number of payment information because it has high similarity with other variables in the model.

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
XGBoost(1st Iteration)	0.817963	0.825243	0.968926	0.891331	0.732864	0.451381
XGBoost(2nd Iteration)	0.817963	0.825638	0.968214	0.891260	0.730720	0.452692

iv. LightGBM

To optimise the classification model and improve predictive performance, the LightGBM algorithm was also selected due to its high performance on structured datasets, ability to handle categorical features, and support for efficient training with large datasets. In the first iteration, the LightGBM model was trained with only its numerical variables. The goal was to check the contribution of categorical variables to the benchmark. The model demonstrated strong recall and F1 score, indicating that it was very effective at identifying satisfied customers while maintaining a good balance between precision and recall.
In the second iteration, several hyperparameters (number of estimators, learning rate, max depth and number of leaves) were modified to potentially improve model performance. While performance remained relatively strong, the changes resulted in slightly lower scores across most metrics compared to the default model. Manual tuning (smaller learning rate, more trees, limited depth) slightly reduced performance, possibly due to underfitting or too conservative model complexity. This indicates that the default hyperparameters were already well-calibrated for the dataset, and the manual adjustments did not provide additional benefit in this case.

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
LightGBM(1st Iteration)	0.8187	0.8246	0.9713	0.8919	0.7377	0.4481
LightGBM(2nd Iteration)	0.8134	0.8206	0.9696	0.8889	0.7351	0.4547

v. CatBoost:

To further find the best model to predict customer review ratings, CatBoost was chosen for its capability to handle categorical features in the dataset. With this information, the first iteration of CatBoost used the same features as the other model without dummy encoding, for the second iteration, it includes customer_city features to test CatBoost's categorical capabilities.

Based on the table below it can be seen that the first iteration marginally improves upon the benchmark CatBoost model, and the second iteration marginally improves again, highlighting CatBoost strength in handling categorical features.

Model	Accuracy	Precision	Recall	F1-Score	AUC Score	Log Loss
CatBoost (1st Iteration)	0.818892	0.826549	0.968104	0.891744	0.741038	0.447542
CatBoost (2nd Iteration)	0.820032	0.826859	0.96942	0.892482	0.742195	0.446259

vi. Model Stacking:

To improve upon prediction accuracy, a meta-model was created through stacking the best-performing benchmark model. The base learners that are used in the meta model are LightGBM, CatBoost, and XGBoost, with Logistic Regression as the meta learner. Based on the table below, the stacked model improves the performance against other benchmark models.

Model	Accuracy	Precision	Recall	F1-Score	AUC Score	Log Loss
Stacked Model	0.819061	0.825697	0.969913	0.892014	0.741518	0.448149

A.4. Feature Engineering & Selection:

Feature Name	Description
avg_product_width	Average product width in centimeters
avg_product_length	Average product length in centimeters
total_items	Total number of items in the order
total_freight_value	Total freight/shipping fee for the order
avg_product_height	Average product height in centimeters
avg_delivery_days	Average number of days taken for delivery
avg_product_weight	Average weight of products in the order (grams)
avg_product_photos_qty	Average number of product photos
avg_product_description_length	Average number of characters in product descriptions
max_installments	Maximum number of payment installments for the order
total_price	Total price of all items in the order
was_late_delivery	Binary flag indicating if delivery was later than the estimated date

The table above portrays the final list of independent variables that are used for the modeling. All raw data sets are merged and aggregated into order level. This level clearly helps the model to get the prediction for the target variable because it predicts the satisfaction level of the customer per order. Some variables are excluded due to unnecessary reasons (ids and dates format variables), data leakage (anything related to review), and high VIF score. Moreover, there are 3 variables that are one hot encoded before they are used in the model, which are customer_state, order_status, main_payment_type, because they consist of categorical types of data. All this feature engineering that has been done has brought positive impact to the model because the data set is more compact and relevant with the target variable that wants to be predicted. In terms of the target variable, it is a binary variable that has been engineered from the average review score data. It will be 1 when the score is bigger or equal to 4 and 0 for less than 4.

Before After Comparison

Aspect	Before Feature Engineering	After Feature Engineering & Selection
Raw Features	Raw product info Timestamps, Redundant IDs	Aggregated order-level metrics, cleaned from useless variables, data leakage problem, and high correlations
Feature Count	Around 50 raw variables	Reduced to 15 meaningful features
Multicollinearity	High (confirmed via VIF)	Lowered using VIF filtering
Interpretability	Low (ambiguous column names)	High (renamed and domain-aligned features)
Top Model AUC (CatBoost)	Lower	Improved (due to more focused, relevant features)

A.5. Hyperparameter Tuning & Model Optimisation:

Based on the previous model experimentation, CatBoost has the highest single model performance. Hyperparameter tuning was conducted to further unlock CatBoost's potential in modelling the dataset. The method used to tune the hyperparameter is Optuna, an advanced optimisation framework that applies Bayesian optimisation strategies prioritising optimisation speed and performance. Hyperparameters and search ranges were selected for their impact on improving model performance based on theoretical best practices, empirical research, and computational considerations.

Hyperparameter	Search Range	Purpose of Tuning	Rationale for Range Selection
iterations	[300, 1000] (int)	Number of boosting rounds	Balances between underfitting (too few trees) and overfitting (too many); early stopping helps control
learning_rate	[0.01, 0.2] (float)	Controls the contribution of each tree	Lower range ensures stable learning; upper bound avoids overly aggressive updates
depth	[4, 10] (int)	Maximum depth of trees (model complexity)	Shallower trees generalize better; deeper trees capture complex patterns but risk overfitting

Hyperparameter	Search Range	Purpose of Tuning	Rationale for Range Selection
l2_leaf_reg	[1.0, 10.0] (float)	L2 regularization on leaf scores	Penalizes large weights to reduce overfitting while allowing flexibility
random_strength	[0.1, 10.0] (float)	Adds randomness to tree splits	Enhances model robustness and generalization
bagging_temperature	[0, 5] (float)	Controls stochasticity in bootstrap sampling	Higher values increase diversity among trees; 0 is deterministic
border_count	[32, 255] (int)	Number of splits (bins) for continuous features	Controls discretization granularity; higher bins offer precision at risk of overfitting
eval_metric	Fixed = 'AUC'	Performance metric to optimize	AUC is robust for imbalanced classification, especially where ranking quality matters

After conducting optimisation using optuna with 50 trials, the best parameters were presented below.

Hyperparameter	Best Value
iterations	512
learning_rate	0.1373275254053363
depth	4
l2_leaf_reg	8.258373988091536
random_strength	1.8912374586865055
bagging_temperature	3.295982964116213
border_count	150

Based on the table below, we can see that the optimisation marginally improves the model performance compared to the second iteration.

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
CatBoost (2nd Iteration)	0.820032	0.826859	0.96942	0.892482	0.742195	0.446259
CatBoost (Tuned)	0.818934	0.825795	0.969529	0.891908	0.743238	0.446844

A.6 Conclusion & Final Model Recommendations:

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
Logistic Regression (Benchmark Model)	0.817921	0.827282	0.965200	0.890935	0.727911	0.457731
Random Forest (2ndIteration)	0.784478	0.851323	0.872691	0.861875	0.730435	0.572936
XGBoost (1st Iteration)	0.817963	0.825243	0.968926	0.891331	0.732864	0.451381
LightGBM (2nd Iteration)	0.8134	0.8206	0.9696	0.8889	0.7351	0.4547
CatBoost	0.820032	0.826859	0.96942	0.892482	0.742195	0.446259

(2nd Iteration)						
Stacked Model	0.819061	0.825697	0.969913	0.892014	0.741518	0.448149

Metric	Top Performer
Accuracy	CatBoost (0.8200)
Precision	Random Forest (0.8513)
Recall	Stacked Model (0.9699)
F1 Score	CatBoost (0.8920)
AUC	CatBoost (0.7422)
Log Loss	CatBoost (0.4462)

Looking at the performance, CatBoost performs better than the rest of the models. It has the highest accuracy, F1 score, AUC, and the lowest score of log loss. This indicates that CatBoost is not only effective in identifying positive customer reviews but also in distinguishing between satisfied and unsatisfied customers. The Stacked Model follows the second-best performing model, but it introduces significant interpretability challenges, which is not desired for this project.

Model	Status	Justification
CatBoost	✓	Chosen Model. Has a consistent and good performance across metrics, and still has a good interpretability via SHAP. It also benefits from its ability to handle categorical features.
Stacked Model	X	Second best performing model but lack interpretability. In business setting, stakeholders want to understand what drives customer satisfaction, a black-box model like this was not the right choice.
LightGBM	X	Has strong recall but lower precision and F1. Looking from a business perspective, this could possibly mean overestimating customer satisfaction and missing areas for improvement.
XGBoost	X	Slightly worse than CatBoost in every metric. It offers the same benefits as CatBoost.
Logistic Regression	X	Has the best interpretability but the performance is worse compared to other boosting models. We still want a good balance between performance and interpretability. In addition, compared to CatBoost with a better balance between high Precision and AUC, it shows that the Logistic Regression model might tends to overestimate the positive class, which may lead to ineffective allocation mitigation approaches or misreading of customer shopping experiences.
Random Forest	X	Generates the highest precision but lowest accuracy, indicating high misclassification rate. We want a balanced model between predictive power and sentiment detection so this might not be the best model.

Appendix B:

B.1 Group Reflection

1. Biggest Challenge:

The biggest challenge that our team faced was understanding and cleaning the data process and problem formulation. It took us more than 2 weeks to finalise the target variable and have a cleaned dataset that is well prepared for the modelling. At first, we made a mistake by setting a target variable just based on what is the best solution or problem that we can solve for the company without having a clear understanding of the data. It turns out that we selected the target variable (customer churn) that has imbalanced data. It leads to an underperforming model. After realising it, all our group members decided to take the time individually to fully understand the data. After that, we start doing the EDA together and come up with a new target variable that we can set based on the data that we have and still have a great impact on the business in solving their problem or improving their performance. Our group handles this problem very well because we are very open to every group member's idea and come up with the finalised target variable after going through a long discussion and many inputs from each other. On the other hand, from a non-technical perspective, the biggest challenge is time availability; everybody has their own assignments and commitments during this busy week. However, we are very responsible, and we do not mind covering each other's tasks, so we still can manage the discussion well and do all the coding and modelling together.

2. Most Impactful Strategy:

Making a shared project timeline in Google Docs with specific tasks and due dates was one of the most effective tactics our team used. We are not only using Google Docs for storing Google Docs, but also using Trello to monitor task tracking and greatly decrease overlap by keeping everyone accountable and in sync for all tasks delegated to each person. This will two software also helped us to control our work progress and making sure that it followed the timeline that has been prepared in the initial meeting. For instance, each of us was tasked with working on a single model during the model-building process. Because the task boundaries were well-defined, this undoubtedly protected us from duplicating each other's work. In addition to improving our productivity, this approach enabled us to generate a superior final report. For everyone on the team to look over and assess our work, we also made a shared document in which we all wrote and presented the results of our research. Therefore, any mistakes or inconsistencies might be corrected right away. To foster cooperation and accountability in the team conditions, we will apply this strategy in subsequent projects by promptly establishing a shared task board and promoting open progress reports. Because the project manager or project lead can keep an eye on, guide, and reroute resources if there are any roadblocks that could slow down the team or project, this undoubtedly helps ensure that projects are finished on schedule.

3. Key Takeaway for Your Future Career

We think the biggest takeaway that we can take as a group for our future career, is knowing how to work on a machine learning project as a group. It is similar to a professional setting where you would not have to work on a big project alone, so you need to know how to work in a group setting.

This assignment has also shown us that building a good machine learning model is a highly iterative process, we're not just choosing one algorithm but keep refining the models as we go along. It's also about understanding the business process and problem which will shape our data preparation process. Working as a team made us realise how important it is to leverage diverse strengths, whether in technical or non-technical aspects of the assignment, because we can share knowledge and solve problems more effectively. It's a reflection of the real-world environment and projects. This experience highlights the role of clear communication, documentation, collaboration and balancing technical capability with non-technical, as well as other skills that are essential in the data analytics role that we all want to pursue.

B.2 Individual Reflection

1. Chris - 540369759

At the beginning of the team meeting, my group members assigned me to be the leader during this project. Therefore, the way I contribute to this group project is a little bit heavier on the non-technical side rather than the technical side. I always try to make a reminder for the meeting schedule, try to remind my group members regarding the agenda and start to open the conversation to initiate the discussion. Moreover, I got the opportunity from the team to decide critical things that needed to be decided after a long discussion and also assigned the writing part for every member. For example, in assigning the writing part, I was in charge of making the decision for everybody's task for the writing. My group trusts me to assign it because I am the leader, and all my decisions are aligned with their perspective. I assigned the task based on my observation during the discussion and see who was more confident when we were discussing a certain section of the report such as EDA, modeling, business context, etc. I believe having one person as a leader will bring the team's dynamic and working process to be more efficient because there will be one person who can decide when we had a long discussion about anything that is significant to the report.

2. Nanette Ridwan – 540225237

One of my assignments was to research the target features and the type of baseline model that we can use for this assignment. I ran three different churn prediction models in the beginning with a non-satisfactory result, and after discussing with the team, we decided to change the target features to customer reviews. Beyond completing my assigned task, one of my major contributions to the team was through my professional experience from working in one of the e-commerce companies back in Indonesia. My experience has allowed me to help the team in giving input and sharing knowledge that is useful for problem formulation that mirrors real challenges, and on how to effectively structure the data that is similar to the structure that e-commerce usually uses. I also help to share insight on key operational metrics, so that we can prioritize which features to explore. I was also assigned to do the business insight and implementation section where my experience can greatly help me and my team to make this part more realistic and applicable to actual industry scenarios.

3. Ferdinand Richard Wang – 530824099

I helped the team by offering crucial insights on significant topics such variable merging and selection, in addition to finishing the primary responsibilities provided in defining and outlining the business problem, and working on and assessing the model. For example, one of the tasks I did was to help formulate the problem and how to code the target variable for our first project objective, namely "will customers churn" even though in the end it was not used due to a very significant class imbalance. This communication and discussion approach certainly helps everyone on the team have a variety of perspectives to help us find the best ways to improve the standard of our final recommendations. A strong sense of support and unity was created inside our group as a result of my frequent check-ins with teammates who were having trouble meeting deadlines for their various assignments and my offer of help with research or other chores.

4. Aditya Parama Setiaboedi - 540507328

When drafting the project charter, I volunteered to take on the role of technical lead. Since roles were assigned at the beginning of the project, I was committed to fulfilling my part. For example, I set up the Python environment in Google Colab so that the team could collaborate using a shared platform, instead of relying on Jupyter Notebook, which only works on a single device. One challenge with Colab is that users need to upload the dataset every time the runtime resets. Alternatively, uploading it to Google Drive would restrict access to the file owner. Drawing on what I learned from another unit, I embedded the dataset via GitHub, allowing everyone to access it easily through Colab. I applied this solution to our project. As the technical lead, I also compiled the code into a single file, since each of us had initially worked on different scenarios separately. This helped make our testing and debugging process smoother. From this experience, I learned that assigning roles based on a person's skills and interests not only improves task distribution but also fosters a stronger sense of ownership and commitment within the team.

5. Bernardino Realino Ivan Pradipta – 540619768

For this assignment, I contributed to the group's overall work, from technical to non-technical, covering what was needed by the group. One of the examples of my contribution is optimising the model through hyperparameter tuning. One of the main challenges for the optimisation process was managing computational speed, which often takes a long time to optimise. Learning from the experience of working on other machine learning assignments, I advocate for tuning the best model using the Optuna optimisation method, as it is computationally efficient without affecting the performance. For the next part, I also volunteered to create a meta model, which I wanted to learn as I have no prior experience developing model stacking. This meta model gives insight into the group, which model was weighted more, hence focusing more on that particular model as a single model. From this project, I learn to strike a balance between learning from past experiences to create better result, but also experimenting to try to learn new things that we can carry on to the next assignment, project, or even our career. I realised that it is important to keep learning and experience new things, as our knowledge is very important even in little things.