# Project Report: Analysing Customer Satisfaction in Brazilian E-Commerce

## Problem Statement

E-commerce has seen exponential growth over recent years, and understanding the factors that influence customer satisfaction is vital for any business aiming to optimize customer experience and increase future sales. Online reviews are a key part of the customer journey, offering insight into how satisfied customers are with products and services. Analysing customer satisfaction, as reflected in ratings and reviews, is critical for businesses to refine their offerings and boost customer loyalty.

### Why is it Important?

Customer satisfaction has a direct impact on a company’s sales and revenue growth. High satisfaction leads to better customer retention, positive word-of-mouth, and higher future sales. By understanding the factors influencing customer satisfaction, businesses can take proactive measures to address issues and improve their service offerings, thus enhancing their competitive edge in the marketplace.

## Project Objective

The objective of this project is to analyse the Brazilian E-Commerce Public Dataset available on Kaggle to identify and determine the key factors influencing customer satisfaction. The insights derived from this analysis can guide businesses in improving their services, products, and overall customer experience.

## How Data Science Can Solve the Problem

Data science plays a crucial role in uncovering patterns and relationships within data that may not be immediately obvious. By applying machine learning techniques to the dataset, we can analyse the impact of various factors (e.g., delivery time, product price, product quality, customer service) on customer satisfaction, which is represented through ratings and reviews.

Techniques such as logistic regression, decision trees, and random forests can help:

* Identify the key features that influence customer satisfaction.
* Develop a predictive model that classifies customers into satisfied and dissatisfied categories based on their reviews.
* Provide actionable insights for improving customer satisfaction.

## Dataset

The dataset used for this analysis is the Brazilian Olist E-Commerce dataset, available on Kaggle. This dataset contains information on:

* olist\_customers\_dataset: Customer details.
* olist\_sellers\_dataset: Seller details.
* olist\_products\_dataset: Product details.
* olist\_orders\_dataset: Order details.
* olist\_order\_reviews\_dataset: Customer reviews.
* olist\_order\_payments\_dataset: Payment details.
* olist\_order\_items\_dataset: Order items.
* olist\_geolocation\_dataset: Geolocation details.

Each dataset includes multiple variables that can be used to analyze customer satisfaction.

## Technology Stack

* Pandas: Data manipulation and analysis.
* Numpy: Numerical operations and array handling.
* Seaborn/Matplotlib: Data visualization to explore the dataset and model results.
* Scikit-learn: Machine learning models, including logistic regression, decision trees, and random forests.
* Statsmodels: Statistical modeling and hypothesis testing.

1. Data processing and analysis
   1. Data preprocessing

The first step in preparing the data for analysis was to clean it by addressing outliers and handling any inconsistencies that could negatively impact model performance. Specifically, **outliers in delivery days** were identified as having unusually large values that did not seem realistic. These extreme values could skew the analysis and affect the model’s predictive accuracy. To mitigate this, we applied the **Interquartile Range (IQR) method** to remove these outliers. This method works by calculating the IQR (the difference between the 75th and 25th percentiles) and removing data points that fall outside the acceptable range (i.e., 1.5 times the IQR above the 75th percentile or below the 25th percentile). This process helped improve the quality of the data and ensured that only valid records were used for further analysis and modeling.

* 1. Data analysis

Once the data was cleaned, we conducted several analyses to better understand the relationships between different features and customer satisfaction.

* **Delivery Days and Customer Satisfaction**: We specifically analyzed how the number of delivery days impacted the customer satisfaction score. A longer delivery time can potentially lower satisfaction, and this relationship was examined using various visualization techniques.
* **Histogram Plot for Review Scores**: A histogram plot was created to examine the distribution of customer review scores. This allowed us to identify patterns, such as skewness or clustering of scores, and assess how reviews were distributed across different satisfaction levels.
* **Correlation Matrix**: A correlation matrix was generated to identify which features have the most significant relationships with customer satisfaction. This analysis highlighted which variables (e.g., product price, delivery time, payment value) were strongly correlated with higher or lower satisfaction scores, guiding us toward the most impactful features for model development.



A graph with green lines

Description automatically generated

Correlation between features and output

A screenshot of a calculator

Description automatically generated

## Overview of Methods and Models

### Models Used:

* Logistic Regression: Suitable for binary classification tasks (satisfied vs. dissatisfied customers). It helps in identifying linear relationships between customer satisfaction and the influencing factors.
* Decision Tree: A non-linear model that works well for capturing complex relationships in the data, making it useful for identifying which factors (e.g., delivery time, price) impact customer satisfaction the most.
* Random Forest: An ensemble method that aggregates predictions from multiple decision trees to increase model accuracy and robustness. It handles noisy data well and provides feature importance, helping businesses understand the key drivers of customer satisfaction.

### Justification for Choosing These Methods:

* These models are well-suited for classification tasks and can handle complex, non-linear relationships in the data.
* They are robust to missing and noisy data, which is common in real-world datasets.
* They provide feature importance, allowing businesses to identify the most significant factors influencing customer satisfaction.
* Random Forest and Decision Tree models perform well on high-dimensional datasets, making them suitable for this analysis.

## Project Goals

The goal is to develop a predictive model that accurately classifies customer satisfaction based on various factors such as:

* Payment Value
* Delivery Time
* Product Price By understanding these factors, businesses can make informed decisions to improve the customer experience and increase satisfaction, leading to higher sales.

## Success Measurement

The success of the project will be measured using the following metrics:

* Accuracy: The overall percentage of correctly classified instances.
* Precision: The proportion of positive predictions that are actually correct.
* Recall: The proportion of actual positives that are correctly identified by the model.
* F1-Score: The harmonic mean of precision and recall, offering a balance between the two.

## Model Performance and Results

## Model Metrics:

The following metrics were achieved using the Random Forest model on the test dataset:

| Metric | Value |
| --- | --- |
| Accuracy | 0.68 |
| Precision | 0.71 |
| Recall | 0.81 |
| F1-Score | 0.76 |

Classification Report:

| Class | Precision | Recall | F1-Score |
| --- | --- | --- | --- |
| 0 (Dissatisfied) | 0.60 | 0.46 | 0.52 |
| 1 (Satisfied) | 0.71 | 0.81 | 0.76 |
| Accuracy |  |  | 0.68 |
| Weighted avg | 0.67 | 0.68 | 0.67 |

### Cross-Validation:

* Best Parameters: {'learning\_rate': 0.1, 'max\_depth': 5, 'max\_iter': 100, 'min\_samples\_leaf': 20}
* Best Cross-Validation Score: 0.67
* Test Accuracy: 0.68

### Insights:

* The model achieved an accuracy of 68%, which is a reasonable result for this type of binary classification task.
* The recall score for satisfied customers (class 1) is higher than for dissatisfied customers, indicating the model is better at identifying satisfied customers.
* Further optimization, such as hyperparameter tuning or exploring other machine learning techniques, could improve the model's performance.

### Different models used for predication:

1. HistGradient Boosting Classifier:

| Metric | Value |
| --- | --- |
| Accuracy | 0.68 |
| Precision | 0.67 |
| Recall | 0.68 |
| F1-Score | 0.67 |

1. Decision Tree Classifier:

| Metric | Value |
| --- | --- |
| Accuracy | 0.63 |
| Precision | 0.63 |
| Recall | 0.63 |
| F1-Score | 0.63 |

1. Random Forest Classifier:

| Metric | Value |
| --- | --- |
| Accuracy | 0.68 |
| Precision | 0.67 |
| Recall | 0.68 |
| F1-Score | 0.66 |

1. Voting Classifier:

| Metric | Value |
| --- | --- |
| Accuracy | 0.64 |
| Precision | 0.64 |
| Recall | 0.64 |
| F1-Score | 0.6 |

## Conclusion

This project demonstrates how machine learning can be applied to understand customer satisfaction in Brazilian e-commerce. By analyzing various factors like delivery time, product price, and payment value, the model provides actionable insights that can guide businesses in improving customer experience. The results show that the model performs reasonably well, but there is room for improvement, especially in identifying dissatisfied customers.

Next Steps:

* Explore additional features that could influence satisfaction, such as customer support interactions.
* Deploy the model in a real-world setting to provide continuous feedback to businesses on customer satisfaction.