Customer Satisfaction Prediction in Online Goods Delivery through Interpretable Predictive Models and Sentiment Analysis

Akula Venkata Satya Sai Gopinadh[1*], SVSN Sarma[2], Gudipudi Radhesyam[3]

[1]Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India
satyaakula707@gmail.com

[2]Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India
svsnsarma@sssihl.edu.in

[3]Charter Global Technologies Pvt. Ltd, Hyderabad, India
gudipudi.radhesyam@gmail.com

Abstract. In this paper, we explore the effectiveness of machine learning (ML) in forecasting customer satisfaction scores based on the sales dataset curated from Olist, a prominent Brazilian e-commerce enterprise. Customer satisfaction score is classified into four distinct categories: Poor, Average, Good, and Excellent, with a prevalence of Excellent ratings among the majority of sales orders. Motivated by the recognition that delivery duration and product, seller rating scores, derived from previous customer transactions, play pivotal roles in shaping customer satisfaction, we embark on a comprehensive analysis. In our investigation, we leverage advanced machine learning techniques, specifically Random Forest (RF), XGBoost (XGB), and Decision Tree (DT) to forecast customer satisfaction scores. Additionally, we incorporated sentiment analysis of review comments into our research. Notably, the XGBoost (XGB) model emerged as the top performer, achieving an average precision, recall, and macro F1-score of 0.53,0.52 and 0.53 respectively. This underscores the effectiveness of incorporating sentiment analysis alongside traditional ML models in predicting customer satisfaction in e- commerce settings.

Keywords: E-Commerce**,** Customer Satisfaction Score Prediction, Classification, Sentiment Analysis, Interpretable Machine Learning.

1 Introduction

E-commerce, short for electronic commerce, has revolutionized the way businesses operate and consumers engage in transactions. This digital paradigm involves buying and selling goods or services over the internet, eliminating traditional geographical barriers and fostering a global marketplace. E-commerce platforms encompass a diverse array of entities, from online retailers and marketplaces to digital payment systems. The convenience of browsing, selecting, and purchasing products or services from the comfort of one's home has driven the exponential growth of e-commerce. This industry thrives on technological innovations, secure payment gateways, personalized shopping experiences, and efficient supply chain management. As the digital landscape continues to evolve, e-commerce remains a dynamic force, reshaping the retail landscape and influencing consumer behavior on a global scale.

Customer satisfaction is vital in e-commerce, influencing brand loyalty and success [2]. Sellers use it to improve product quality, shipping, and support, enhancing the overall user experience. Prioritizing customer satisfaction drives successful transactions and builds a positive brand reputation. Customer satisfaction relies on a diverse array of factors, including the initial purchase encounter, repeat transactions, product returns, product attributes, inventory management, logistics, and the quality of customer support [3]. It presents a significant challenge to businesses, as the negotiating power of customers compels them to offer top-notch products and exceptional service to meet customer expectations. This approach not only enhances customer trust but also fosters loyalty [4]. Companies need to enhance operational performance to address negative feedback, identify key satisfaction drivers, retain customers, and optimize resources [5]. Customer satisfaction scores are vital indicators for vendors, representing customers' overall impressions of online retail stores. Using machine learning (ML), these scores can be predicted, aiding vendors in understanding factors contributing to both poor and strong satisfaction. ML models have been widely applied in industries like e-commerce, telecom, and hotels for this purpose [5], [6]. Most models primarily focus on accurately predicting high customer satisfaction scores, which overlooks insights from lower scores that could boost vendor sales. Sentiment analysis of user reviews in e-commerce is crucial for understanding customer satisfaction and preferences. Analyzing sentiments expressed in reviews helps businesses gain insights into customer opinions, preferences, and experiences, enabling them to identify areas for improvement, optimize product offerings, and enhance overall satisfaction.[8].

The structure of this paper unfolds as follows: Section 2 delves into related works pertaining to the prediction of online customer satisfaction scores through the ML approach. Following this, Section 3 outlines the methodology employed, while Section 4 presents the experimental results, followed by Section 5 presents the Interpretability. Finally, Section 6 encapsulates the paper with a conclusion.

2 Related Work

Numerous studies delve into predicting online customer behavior through the application of machine learning models. This study [1] assesses machine learning's efficacy in predicting e-commerce customer satisfaction scores, categorizing them based on features like delivery duration and average product rating. Compared to Logistic Regression and KNN, Random Forest outperforms with an average precision of 0.34, recall of 0.36, and macro F1 of 0.32. Key features include mean and standard deviation of product ratings, each scoring 0.313 and 0.087, respectively. Researchers at Asia Pacific University [3] utilize Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), and RF models to examine the factors impacting customer satisfaction in Brazil's e-commerce landscape. Their findings highlight RF as the most accurate model, showcasing superior performance. Significant insights underscore the pivotal role of estimated delivery date and the duration for goods delivery in shaping customer satisfaction. Furthermore [7], researchers utilized supervised machine learning to mine opinions from 39,976 traveler reviews on Vietnamese hotels from Agoda.com. Logistic Regression (LR), Support Vector Machines (SVM), and Neural Network (NN) emerged as top-performing models for effectively extracting opinions in the Vietnamese language. This study stands as a valuable reference for deploying opinion mining in business applications. This study assesses the predictive efficacy of various document embedding methods for sentiment analysis in Brazilian Portuguese, using a unified dataset with predefined partitions. The analysis aims to evaluate model generalization across different contexts and assess the feasibility of employing a single model for diverse scenarios [8].

3 Methodology

This study aims to predict e-commerce customer satisfaction scores by identifying key influencing features. The process begins with dataset acquisition, followed by a thorough analysis and extraction of insights. Data files are merged, cleansed, normalized, and split into training and testing sets.

Subsequently, algorithms are implemented, models built, and performance evaluated. Balancing techniques address training dataset imbalances, while feature engineering and selection enhance model effectiveness. Hyperparameter tuning optimizes each algorithm's performance, and results are compared with those in a referenced study [1].

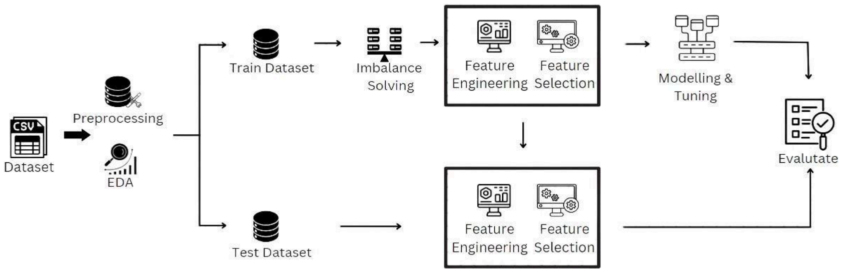


Fig.1: Prediction Process Chart

3.1 Dataset

This study utilized datasets from Olist Shops, a leading Brazilian e-commerce company, covering sales from 2016 to 2018, available on Kaggle [11]. The datasets include 52 columns and 100,000 records, compiled from nine individual files: products, product category name translation, orders, order items, order reviews, sellers, order customers, payment, and geolocation. These files were merged to form a unified dataset for analysis. The geolocation data is used exclusively for Exploratory Data Analysis (EDA), while the order reviews dataset is employed for sentiment analysis to better understand customer feedback and satisfaction.

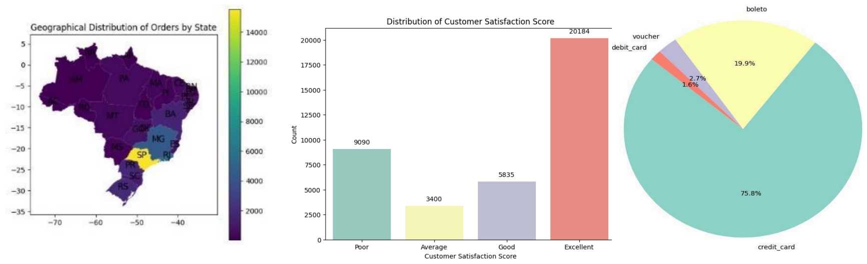
3.2 Data Pre-processing

The dataset underwent an extensive cleaning process to eliminate incomplete or erroneous records, ensuring the highest data integrity. Initially, the 'order_items_dataset' was merged with the 'order_payments_dataset' based on 'order_id.' This intermediary dataset was subsequently merged with the 'order_reviews_dataset' and 'orders_dataset' to form a consolidated dataset, which was thoroughly cleansed of duplicates and null values.Subsequent integration steps involved merging the refined dataset with the 'customers_dataset' using 'customer_id' and the 'products_dataset' using 'product_id,' with meticulous removal of records containing null values at each stage. The final integration was with the 'product_category_name_translation' dataset based on 'product_category_name,' ensuring the elimination of all remaining null values.

Initially comprising 94,445 records and 33 columns, the dataset was rigorously refined through these processes. A final merge with the 'order_reviews_dataset,' including the removal of unwanted columns and any remaining null values, resulted in a curated dataset of 38,509 records and 37 columns.

3.3 Exploratory Data Analysis (EDA) and Visualization

The analysis of order distribution in Brazil exposes regional disparities, notably in the dominance of the south and southeast regions, led by states such as São Paulo and Rio de Janeiro. Factors like larger populations and economic development contribute to their higher e-commerce activity. Conversely, the northeast region exhibits lower order volumes, indicating a less active e-commerce is in Fig.2(a).



(a) (b) (c)

Fig.2: Geographical Distribution of Orders by State, Distribution of Customer Satisfaction Score and Distribution of Payment mode in Fig.2 (a), (b) and (c) respectively.

Customer satisfaction scores are classified into four categories: Poor, Average, Good, and Excellent. Most orders receive an Excellent rating, followed by Poor, Good, and Average scores, as shown in Fig. 2(b). Regarding payment methods, credit cards dominate with a 75.8% share due to their convenience and security, as depicted in Fig. 2(c). Boleto Bancário, a Brazilian cash-based option, holds 19.9%, while vouchers and debit cards account for 2.7% and 1.6%, respectively. This distribution highlights the preference for credit cards in terms of flexibility.

3.4 Dataset Split

The dataset, containing 38,509 records, was partitioned into training (75%) and test (25%) sets, resulting in 28,881 training and 9,628 test records. For modeling, numerical columns were scaled using Min-Max Scaling, and categorical columns were transformed with Label Encoding. This rigorous data preparation ensures a robust foundation for effective analysis and modeling.

3.5 Imbalance Solving

To address the imbalance in the training dataset, this study applies the Synthetic Minority Over-Sampling Technique (SMOTE) to augment the minority class. By generating synthetic data points, SMOTE mitigates class imbalance, leading to improved model performance and predictive accuracy. This approach ensures a more balanced and representative training dataset, enhancing the reliability of model evaluation and predictions.

3.6 Feature Engineering

To enhance the predictive capabilities of machine learning models for customer satisfaction scores, two distinct sets of features are engineered from the datasets.

Delivery process duration:

Customers generally tend to be happier when their packages arrive sooner than anticipated. To capture this sentiment, we derive features from the dataset that implicitly reflect customer satisfaction scores.

- **Estimated Number of Days for Delivery:** It is crafted by calculating the duration between the approval of an order and its delivery to the customer.
- **Arrival Punctuality:** It signifies the temporal alignment between the confirmed delivery by the carrier and the initially estimated delivery date.

Average Product and Seller rating:

Acknowledging the significant impact of product and seller ratings on customer satisfaction as in [1], we adopt a thoughtful approach to feature engineering. This entails calculating seller and product rating scores based on the average historical review scores linked to specific seller IDs and product IDs. These resulting features capture nuanced insights from cumulative customer feedback, elevating the predictive capacity of our machine learning model.

- **Product rating:** It calculates the average review score for each product by grouping the data based on unique product identifiers.
- **Seller rating:** It calculates the average review score for each seller by grouping the data based on unique seller identifiers.

As per [1], to prevent data leakage, the average rating is computed exclusively from the training dataset. This calculated average will subsequently replace values in the test dataset based on matching seller and product IDs.

Review Sentiment scores:

We developed a method for enhancing customer feedback analysis using the DistilBERT-based multilingual model [9] for sentiment analysis. By processing each review comment, the model generates three key features: **review_positive_score, review_negative_score,** and **review_neutral_score.** This approach provides valuable insights into customer sentiment, aiding businesses in understanding satisfaction levels and making informed improvements to products or services. Utilizing Hugging Face's advanced DistilBERT model demonstrates our commitment to leveraging top-tier tools for accurate and efficient sentiment extraction.

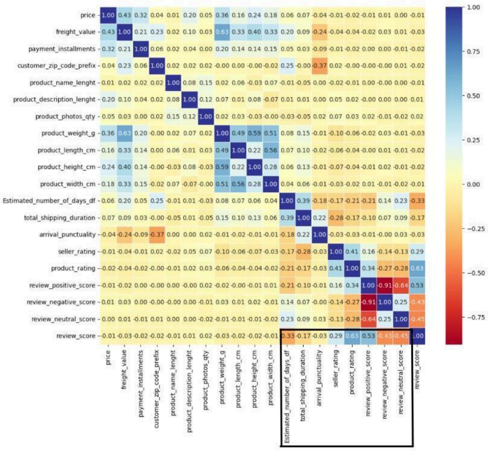


Fig.3: Correlation Heatmap of features

Insights from Fig. 3 reveal that key factors significantly correlate with customer review scores. Notable contributors include product ratings, seller ratings, sentiment scores from user reviews, estimated delivery duration, and total shipping duration. These aspects play a crucial role in shaping customer assessments and overall satisfaction scores.

3.8 **Feature Selection**

Feature selection enhances predictive model efficiency. The Random Forest classifier, an ensemble of decision trees, effectively captures complex relationships and handles categorical and numerical features. It resists overfitting, but the choice of method depends on dataset characteristics, with alternatives like Recursive Feature Elimination (RFE) or LASSO regression also being viable.

3.9 Modeling

RF, XGB, and DT models are widely used for predicting customer satisfaction scores due to their simplicity, robust performance, and versatility. RF combines decision trees for accuracy, XGB iteratively refines performance, and DT offer interpretability. These models are valuable in e-commerce satisfaction score forecasting for their straightforward implementation and strong predictive capabilities.

3.10 Model Tuning & Evaluation

In enhancing machine learning models for customer satisfaction prediction, GridSearchCV was used to explore different hyperparameter combinations for Random Forest, XGBoost, and Decision Tree models. Parameters such as estimators, depth, learning rate, and minimum child weight were adjusted to improve accuracy. The F1-Score, considering both precision and recall, guided the selection of optimal hyperparameters. Metrics like recall, precision, and F1-Score were used to evaluate model performance, resulting in state-of-the-art findings.

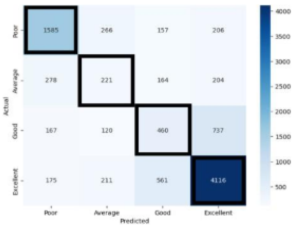
4 Experimental Results

In addressing data imbalance, we implemented SMOTE (Synthetic Minority Over-sampling Technique) to enhance recall before initiating hyperparameter tuning for each model. Since the data is limited, we opted not to use undersampling techniques to avoid further reducing the dataset size.. Subsequent to the tuning process, Table I provides insights into the optimal imbalance treatment method alongside the finest hyperparameter adjustments, aiming to achieve a balanced enhancement in both precision and recall.

Table.1: Optimal Outcomes from Fine-Tuning Model Parameters and Imbalance Treatment Technique

Model	Imbalance Treatment	Name of the Parameter	Value of the Parameter
XGB	SMOTE	n_estimators	150
		max_depth	15
		learning_rate	0.2
		min_child_weight	1
RF	SMOTE	n_estimators	150
		max_depth	15
		min_samples_split	2
		class_weight	None
DT	SMOTE	min_samples_split	2
		max_depth	15
		class_weight	"balanced"

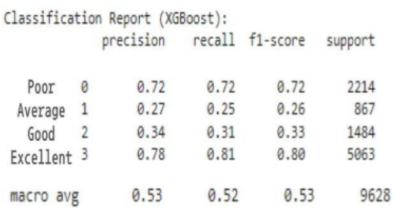
By utilizing the optimized hyperparameters outlined in Table I, the performance evaluation of all machine learning models is comprehensively depicted in Fig. 4. The figures denoted as a1, a2, and a3 present the confusion matrix for the Random Forest (RF), XGBoost (XGB), and Decision Tree (DT) models, respectively. Additionally, figures b1, b2, and b3 showcase the precision, recall, and F1 score for each class within the RF, XGB, DT models, respectively. This holistic representation offers a detailed insight into the classification capabilities of each model across different evaluation metrics.



	Poor	Average	Good	Excellent
Poor	1585	266	157	206
Average	278	221	164	204
Good	167	120	460	737
Excellent	175	211	561	4116

(a1) XGB

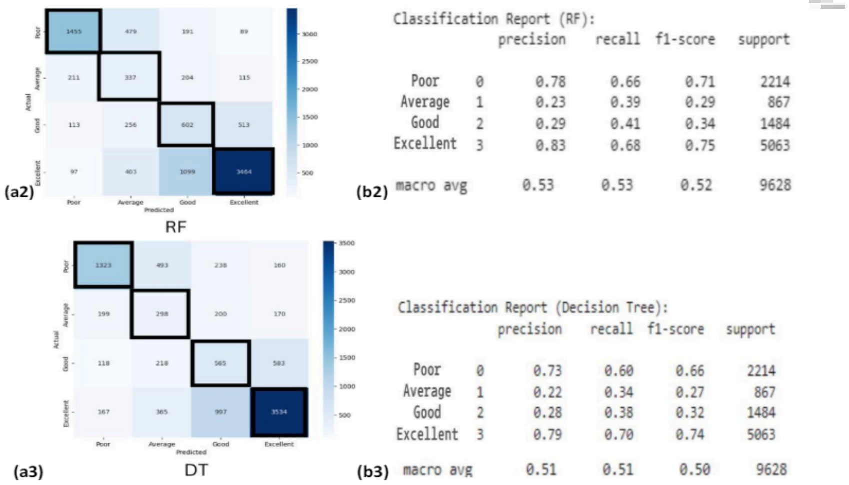
(b1)



```
Classification Report (XGBoost):
              precision    recall  f1-score   support

     Poor  0       0.72      0.72      0.72      2214
  Average  1       0.27      0.25      0.26       867
     Good  2       0.34      0.31      0.33      1484
Excellent  3       0.78      0.81      0.80      5063

  macro avg        0.53      0.52      0.53      9628
```



```
Classification Report (RF):
                 precision    recall  f1-score   support

    Poor     0        0.78      0.66      0.71      2214
 Average     1        0.23      0.39      0.29       867
    Good     2        0.29      0.41      0.34      1484
Excellent    3        0.83      0.68      0.75      5063

macro avg             0.53      0.53      0.52      9628
```

```
Classification Report (Decision Tree):
                 precision    recall  f1-score   support

    Poor     0        0.73      0.60      0.66      2214
 Average     1        0.22      0.34      0.27       867
    Good     2        0.28      0.38      0.32      1484
Excellent    3        0.79      0.70      0.74      5063

 macro avg            0.51      0.51      0.50      9628
```

Fig.4: Result of confusion matrix; (a1) XGB (a2) RF (a3) DT and classification report; (b1) XGB (b2) RF (b3) DT

From Fig. 4, Random Forest achieved the highest precision and recall rates for the "Poor" and "Excellent" classes, with rates of 0.78, 0.83, and 0.66, 0.68, respectively. XGBoost demonstrated the best precision for the "Good" class at 0.34 and optimal precision rates for the "Poor" and "Excellent" classes at 0.72 and 0.78, respectively, along with superior recall values for these classes at 0.72 and 0.81. Overall, XGBoost outperformed with a mean F1 score of 0.53. Random Forest followed closely with average values of 0.53 for precision, recall, and F1-score. Decision Tree achieved mean values of 0.51 for macro F1-score, recall, and precision.

5 Interpretability

Interpretable machine learning methods are essential for understanding complex algorithms, particularly in transparent contexts. Categorized into Global Model Agnostic and Local Model Agnostic Methods, they illuminate feature effects and individual model decisions, enhancing interpretability and trustworthiness. These methods facilitate informed decision-making and ethical deployment of machine learning systems. In our research paper, we utilized Accumulated Local Effects (ALE) from the global model agnostic category and Local Interpretable Model-agnostic Explanations (LIME) from the local model agnostic category.

Accumulated Local Effects (ALE): ALE provides global explanations in black-box ML. Challenges include reliability, effect characterization, and robust inference; addressed through innovative statistical tools [10]. Centered ALE plots refine interpretation by centering at zero, showing deviations from the average prediction, especially useful for features with many values.

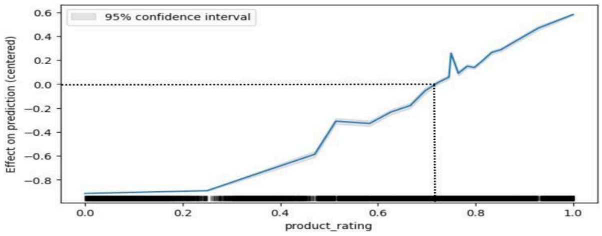


Fig. 5: ALE plot of product_rating feature for XGB

Analysis of the `product_rating` feature shows an average effect on the target feature around 0.7, indicating a positive correlation. Higher product ratings are associated with greater influence on target feature classification, suggesting higher customer satisfaction with positively rated products. This highlights the significance of `product_rating` in predicting or classifying outcomes, indicating that customers tend to express higher satisfaction when products receive favorable ratings.

Local Interpretable Model-agnostic Explanations (LIME): LIME provides transparent explanations for individual predictions by approximating the local decision boundary of complex ML models. In LIME, the intercept represents the baseline prediction probability, prediction local indicates the model's forecasted probability for the instance, and the right value reflects the probability of the true class label, aiding in classification.

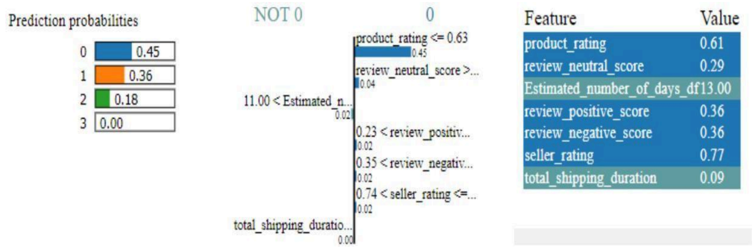


Fig. 6: LIME plot of DT Model

By analyzing these plots, we can observe how variations in feature values influence the model's prediction for each data point individually. This granular analysis allows us to understand the relative importance of different features in predicting the target class for each instance, providing valuable insights into the model's decision-making process at the individual level.

6 Conclusion

This study aims to improve the accuracy of predicting customer satisfaction scores by identifying influential features within the e-commerce sector. Establishing strong correlations between these features and satisfaction levels is crucial for refining strategies, services, logistics, and customer support.

Through the integration of these features into XGB, RF, and DT machine learning models, our research aims to enhance predictive accuracy. Notably, XGB demonstrates superior performance in utilizing these features, achieving excellence across various metrics such as confusion matrix, average precision, recall, and F1-score.

Furthermore, our analysis highlights the pivotal role of specific factors like product and seller ratings, estimated delivery duration, and review sentiment scores in shaping customer satisfaction. Supported by correlation matrix assessments and visual representations, our study not only improves predictive models but also uncovers the fundamental drivers of satisfaction in e-commerce.

In conclusion, this research significantly contributes to understanding customer satisfaction dynamics in e-commerce, offering valuable insights for optimizing operations and enhancing customer experiences. By leveraging advanced machine learning techniques and analyzing key factors, businesses can better anticipate and meet customer expectations, ultimately fostering long-term success in the competitive e-commerce market.

References

[1] P. Wangkiat and C. Polprasert, (2023) "Machine Learning Approach to Predict E-commerce Customer Satisfaction Score".

[2] Anastasia Griva, (2022) “I can get no e-satisfaction”.

[3] Wong, Ann-Nee & Poolan Marikannan, Booma. (2020). Optimising e-commerce customer satisfaction with machine learning.

[4] Chinomona, Richard. (2014). The Influence of E-Service Quality on Customer Perceived Value, Customer Satisfaction and Loyalty in South Africa.

[5] Wu, Tong & Liu, Xinwang. (2020). A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce.

[6] P. Hamsagayathri and K. Rajakumari, (2020) "Machine learning algorithms to empower Indian women entrepreneur in E-commerce clothing".

[7] T. Kim Phung, N. An Te and T. Thi Thu Ha, (2021) "A machine learning approach for opinion mining online customer reviews".

[8] F. D. Souza and J. Baptista de Oliveira e Souza Filho, (2021) "Sentiment Analysis on Brazilian Portuguese User Reviews".

[9] Lik Xun Yuan, (2023) distilbert-base-multilingual-cased-sentiments-student.

[10] Chitu Okoli, (2023) “Statistical Inference using machine learning and classical techniques based on accumulated local effects (ALE)”.

[11]“Brazilian E-Commerce Public Dataset by Olist”, Kaggle.com,2021.