DOI: 10.17148/IJARCCE.2023.121118

Predicting the Customer Behaviour Utilizing Tree Based Machine Learning Algorithms

Hind Khalid Alghamdi[1], Salma Mahjoub Omar[2], Hanaa Namankani[3]

Student, Facility of Computing and Information Technology, King Abdulaziz Iniversity, Jeddah, Saudi Arabia[1]

Doctor, Facility of Computing and Information Technology, King Abdulaziz Iniversity, Jeddah, Saudi Arabia[2]

Doctor, Facility of Computing and Information Technology, King Abdulaziz Iniversity, Jeddah, Saudi Arabia[3]

Abstract: This project examines a tree-based machine learning approach to predict customer behavior outcomes in e-commerce, using a large dataset. The project compares different classification methods to solve three Customer Relationships Management problems: predicting customer satisfaction, churn modeling, and the next product to buy modeling. The analysis is fully automated, making it easy for small e-retailers to implement. The study employs decision tree, random forest, and gradient boosting techniques.

Keywords: machine learning, churn, decision tree, random forest, gradient boosting.

I. INTRODUCTION

Consumer behavior is the process of searching, selecting, purchasing, or evaluating a good or service. It is influenced by a variety of internal and external factors, including personal characteristics, psychological factors, social and cultural influences, economic and environmental factors, marketing and advertising, product quality, and the overall shopping experience, and can be analyzed and predicted through various machine learning techniques.

Customer satisfaction evaluation is important for businesses in developing effective marketing strategies and meeting the needs and wants of their customers. Therefore, predicting customer behavior can help businesses understand the needs and wants of their customers, develop effective marketing strategies, and optimize their operations. [1][2] Also, customer churn is the termination of the customer's relationship with the company or loss of customers, which can lead to marketing costs and revenue loss for the company. Otherwise, recent studies focus on machine learning and data mining techniques to analyze customer behavior [3].

II. BACKGROUND

A. Machine Learning Algorithms

Machine learning is a field of AI that involves creating systems that improve performance based on data. It is often used in businesses to make interactions more efficient and secure and has the potential to predict future outcomes. There are two main types of machine learning algorithms: supervised and unsupervised, with supervised learning being the most widely used. Examples of supervised machine learning include linear regression, layered classification, and support vector devices [4][5].

B. Problem Statement

The problem addressed in this text is small online retailers failing and withdrawing from the market due to accumulated losses and a lack of customers. Where, predicting customer behavior can help e-retailers enhance sales and target a large number of customers, but sometimes they cannot handle the financial cost. Therefore, the goal is to develop an automated classification modeling framework for predicting critical behavioral outcomes for existing customers, including customer satisfaction, churn modeling to increase retention, and next product to buy modeling to increase cross-selling.

C. Aims and Objectives

Before To Study the effectiveness of predicting consumer behavior using Machine learning Tree-Based techniques to enhancing sales to small retailers in e-commerce field. Evidence regarding which methods and features are especially helpful in predicting customer behavior is scattered, and existing literature does not offer e-retailers a clear guide regarding feature engineering for analytical CRM.

DOI: 10.17148/IJARCCE.2023.121118

Therefore, the objectives of this project are as follows:

1.Engineer standard (based on thorough analysis of existing literature) and novel features from purchase history data commonly available to e-retailers.
2.Find the optimal parameter for each modeling phase using hyperparameter tuning and cross-validation techniques. The f-score cross-validation procedure is used to estimate the performance of machine learning models when making predictions on data not used during training.
3.Assess which feature is important to predict the customer behavior.
4.Provide additive explanations of model predictions with apply an approach to quantifying a prediction’s expected quality.

D. Study Contribution

The goal of this project is to predict customer behavior from a large real-world e-commerce dataset using tree-based machine learning modeling techniques that will employ decision tree, random forest, and gradient boosting. Each of the models will be evaluated and compared to determine which of the three is the best model for predicting customer behavior.

E. Related Work

This study[6] focused on predict the customer purchase behavior in e-commerce context. The method involved quantify the strength of these factors: (1) using associations between products to predict the needs of customers; (2) combining collaborative filtering and a hierarchical Bayesian discrete choice model to learn preference of customers; (3) building a support vector regression-based model, called Heat model, to calculate the popularity of products; (4) developing a crowdsourcing approach based experimental platform to generate train set for learning Heat model. Combining these factors, a model, called COREL. Moreover, this study [7] proposed an application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis. The method involved develop an analytic tool which can support online vendors to predict behaviors of the patrons according to Dentsu’s AISAS perspective. The proposed model was developed by the results from 75 specialists who evaluated the behavior that will likely occur after the comments have been posted. The results, hence, were collected and prepared for the data modelling process using the Naïve Bayes probability concept, afterwards, testing for the model’s accuracy with 10-fold cross validation technique. Naïve Bayes technique gives the best result for the behavior analysis. The predictive model for AISAS behavior from this study can give average accuracy higher than 86 percent.Furthermore, this paper [8] the researchers aim to investigate the relationship between consumer behavior parameters and readiness to buy to build a model to predict consumer behavior as changing criteria such as environmental and personal parameters have a direct impact on consumer behavior and thus affect the purchase of the product. A random forest algorithm from machine learning algorithms was used on this model with a public database from kaggle.com, which was divided into three groups based on priority, region, and product category. Finally, the results showed that the accuracy of the algorithm reached 94%, which indicates the dependence of the client's behavior on personal relationships and the surrounding environment.

Moreover, this study [9], discuss how to take advantage of customer behavior data to predict customer churn using customer segmentation and misclassification cost. The model used the data of a telecom company. The main process of model is segmenting customers first, then combining decision tree algorithm with misclassification cost factor to predict customer’s status on different customer groups. The results show that the proposed model is able to enhance the prediction accuracy of customer’s status better than those models without customer segmentation and misclassification cost in terms of the accuracy and coverage of model. Also, this paper [10] aims to give prophetic frame to prognosticate client churn. They used comparison of churn vaticination models grounded on different performances of machine- literacy styles and named variables. The results show the mainly superiority of boosting performances in vaticination compared with simple and bagging models. Finally, this paper [11] proposed a machine learning model to predict customer return visits in airline services. They applied the model to learn from the previous feedback comments and satisfaction ratings of the customers on the previous usage of the service. The experimental results show an accuracy of 83.42% for predicting the customers’ return visits.

III. METHODLOGY

The methodology involves phases as shown in Figure 1 such as defining and describing data, pre-processing it to remove errors and missing values, selecting features using an evaluation test, after that using machine learning algorithms which are: Random Forest, Decision Tree, and Gradient Boost to evaluate customer satisfaction, churn rate, and next product to buy. In addition, confusion matrix outputs are evaluated to improve the accuracy and reliability of the outcome models.

DOI: 10.17148/IJARCCE.2023.121118

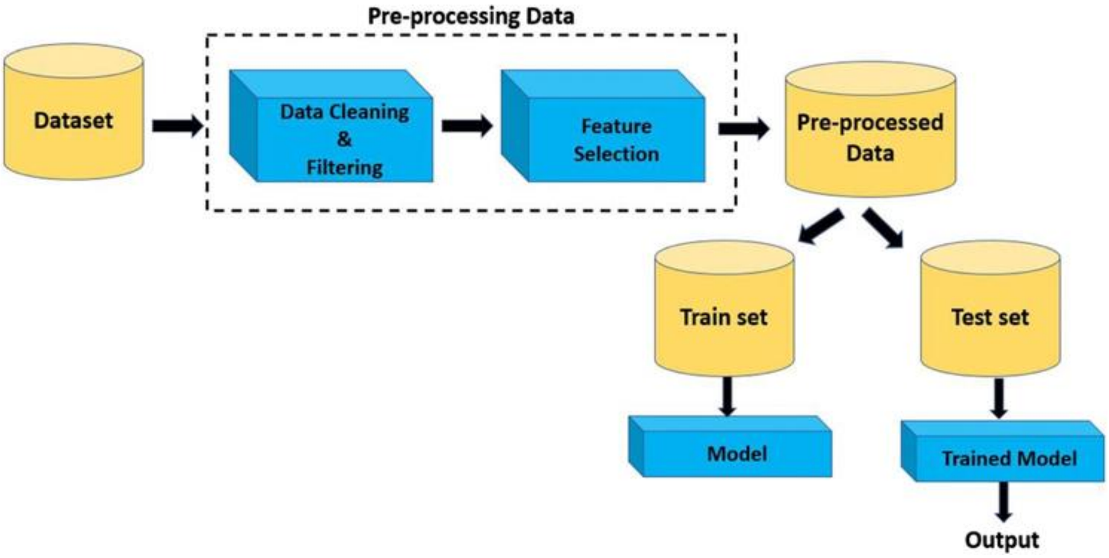

Figure 1 The Methodology Phases

A. Dataset Description

This is an ecommerce public dataset of orders made at Olist Store. The dataset has information of 115,689 orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.

B. Exploratory Data Analysis

First, explore the customers from where they order. As Shows in Figure 2, the greatest number of customers come from 'SP', 'RJ', and 'MG which are the abbreviations for the states of São Paulo, Rio de Janeiro, Minas Gerais, respectively. These three states are all in the Southeast of Brazil and are the most populous economically important states the country. Paulo is the populous state in Brazil is known for its vibrant culture bustling cities and diverse economy. Rio de is famous for its beautiful, vibrant nightlife, and rich heritage. Minas Gerais known for its natural beauty, architecture, and thriving mining industry

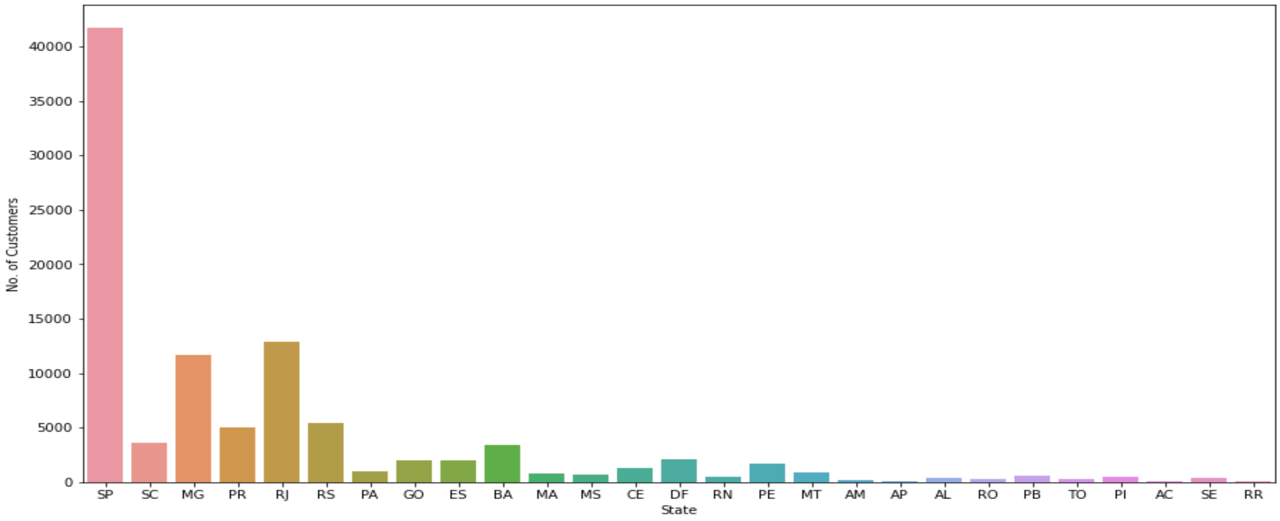

Figure 2 The Customer State

Next, the distribution of price, As shows in Figure 3 by plot chart that presents most orders are under 1000 Brazilian Reals.

DOI: 10.17148/IJARCCE.2023.121118

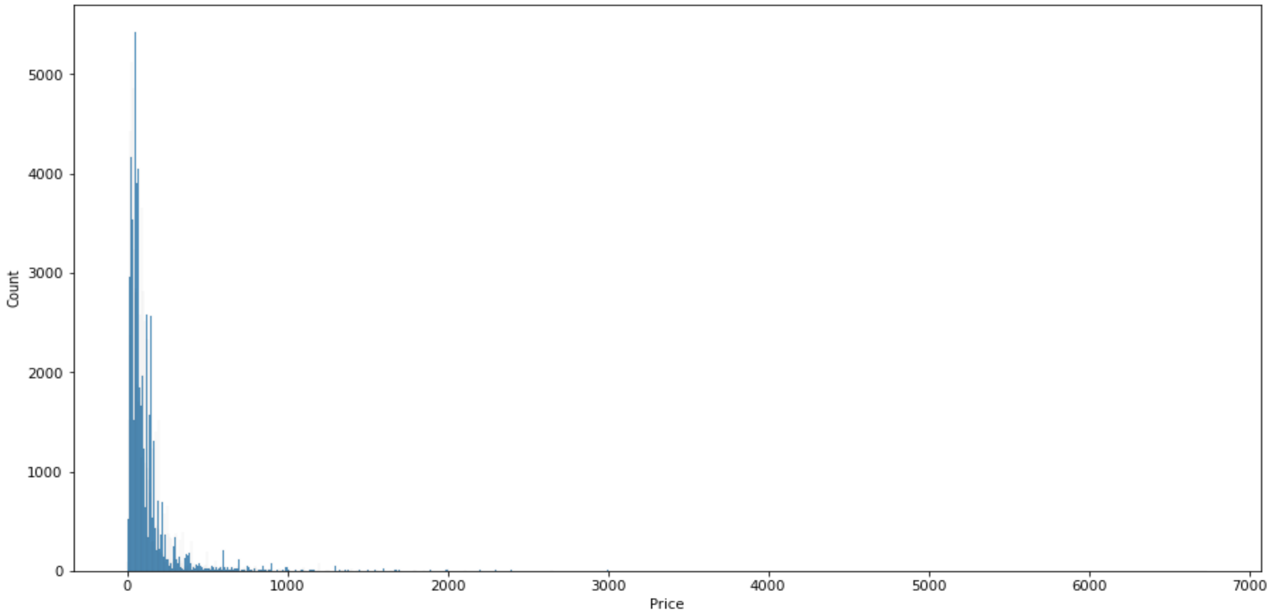

Figure 3 The Price Distribution

After that, the distribution of freight value where, as shown in Figure 4, most values are under 100 kg.

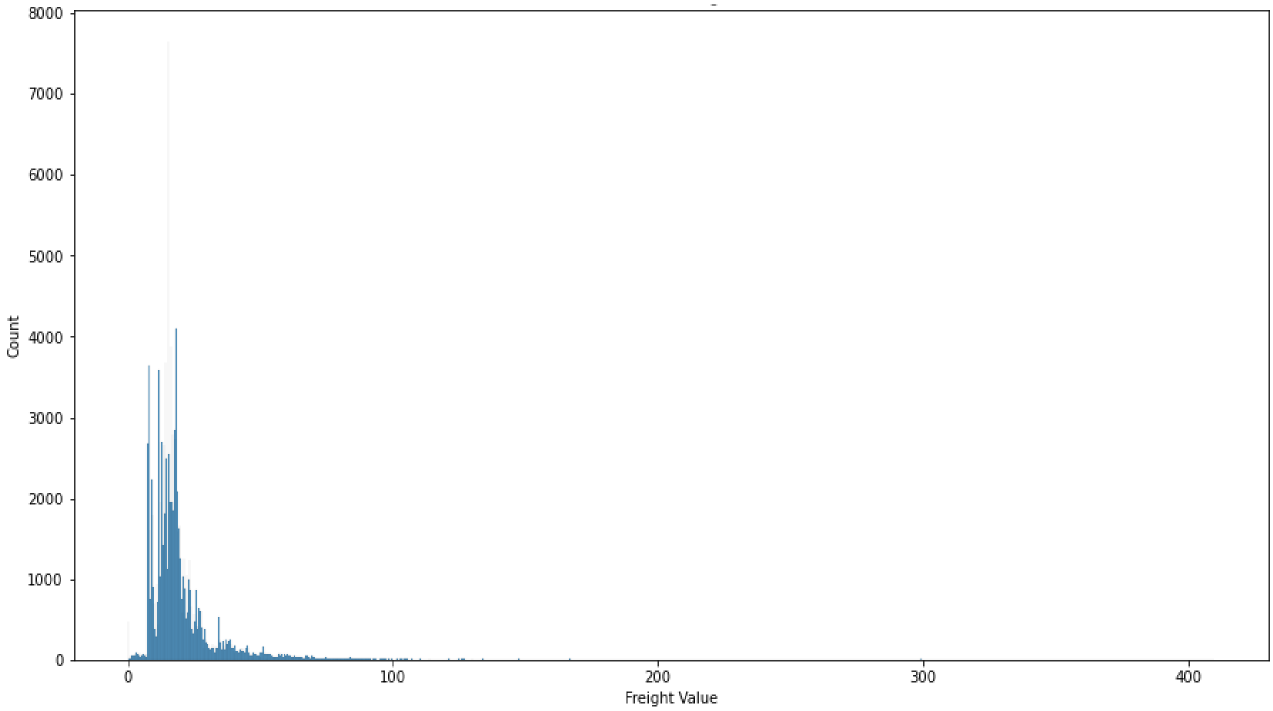

Figure 4 The distribution of freight value

For more insights, here the analysis of the payment types. As shown in Figure 5 the most customers prefer credit cards, followed by boleto, voucher, and then debit cards, also the presence of a 'not_defined' value which is analogous to missing values which is 0.

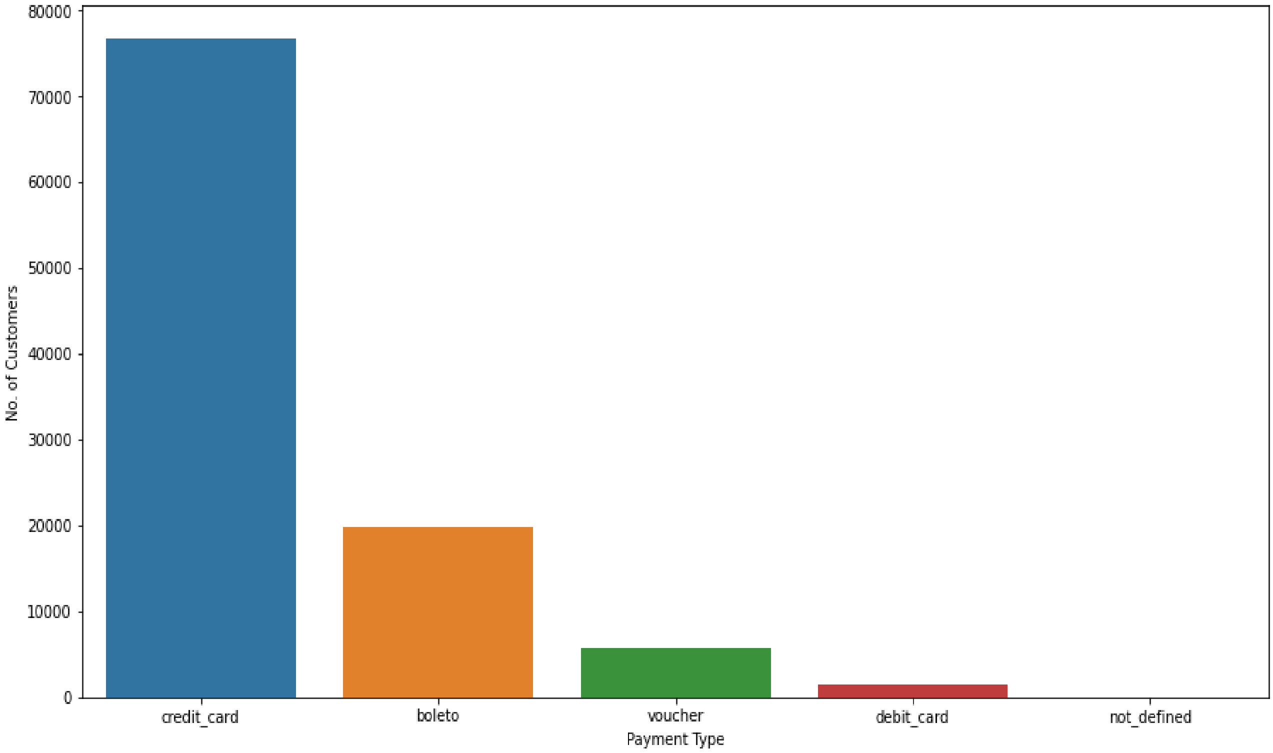

Figure 5 Payment Types

Next, the amount paid by each payment method. As shown in Figure 6, most amount is paid by credit card, followed by debit card and boleto, and at the end voucher.

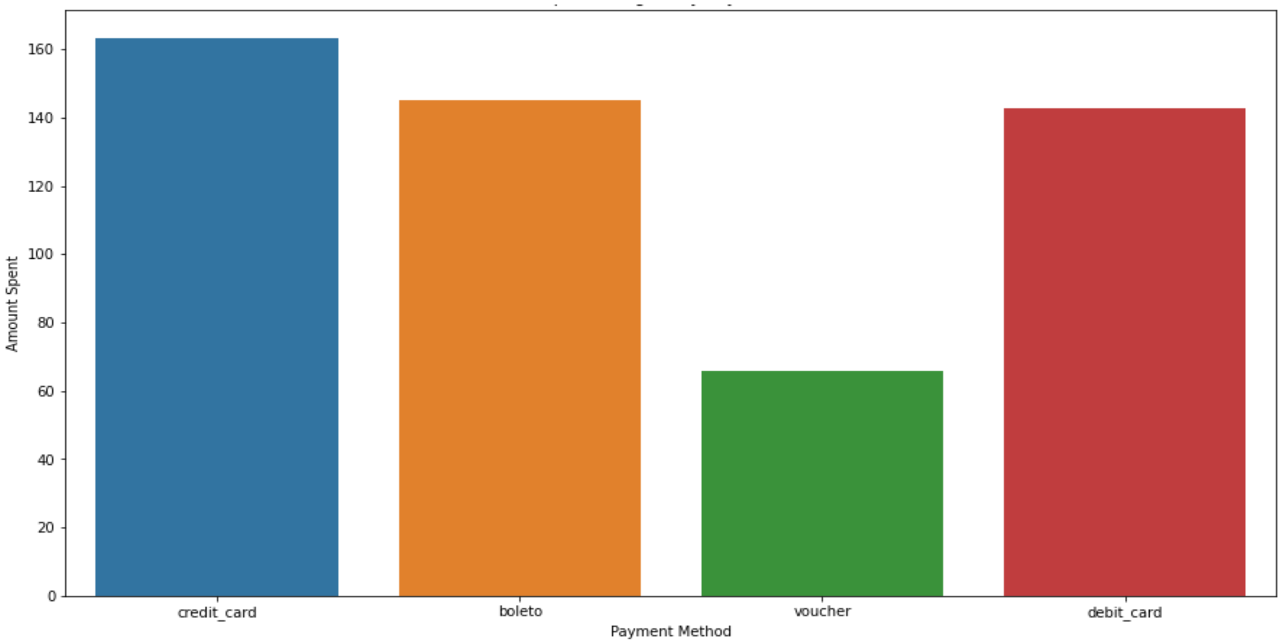

Figure 6 Payment Methods

Also, the distribution of a number of installments, As shown in Figure 7, most customers prefer payment via only one installment. However, customers also opt for more than one installment, the number is not insignificant.

DOI: 10.17148/IJARCCE.2023.121118

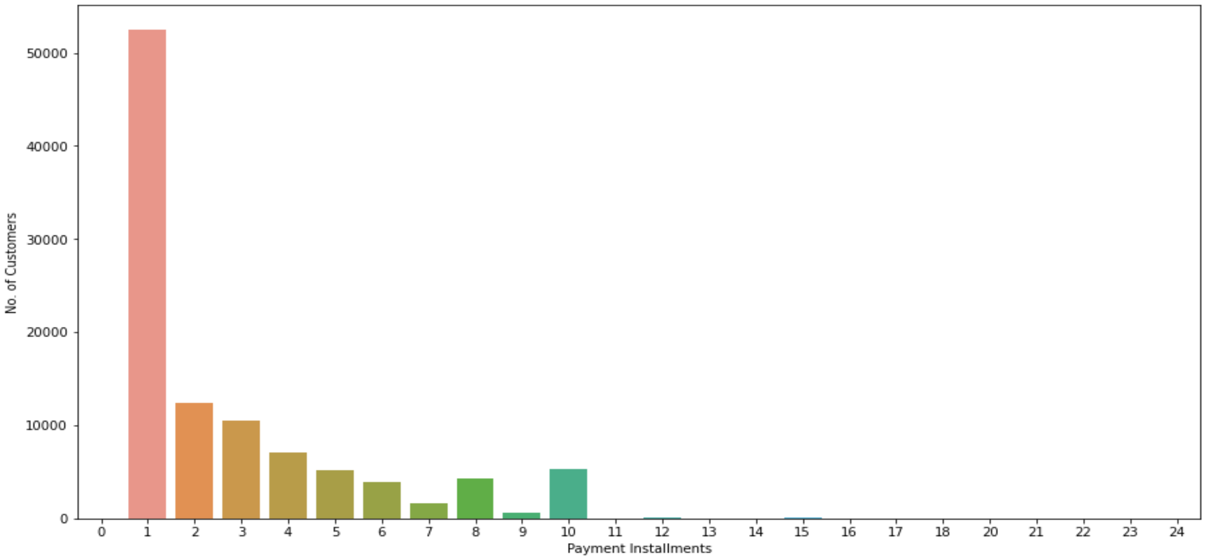

Figure 7 payment installment

Finally, the product category distribution of products, As shown in Figure .8, most products are from the cool_stuff and pet_shop categories and the least from the la_cuisine category.

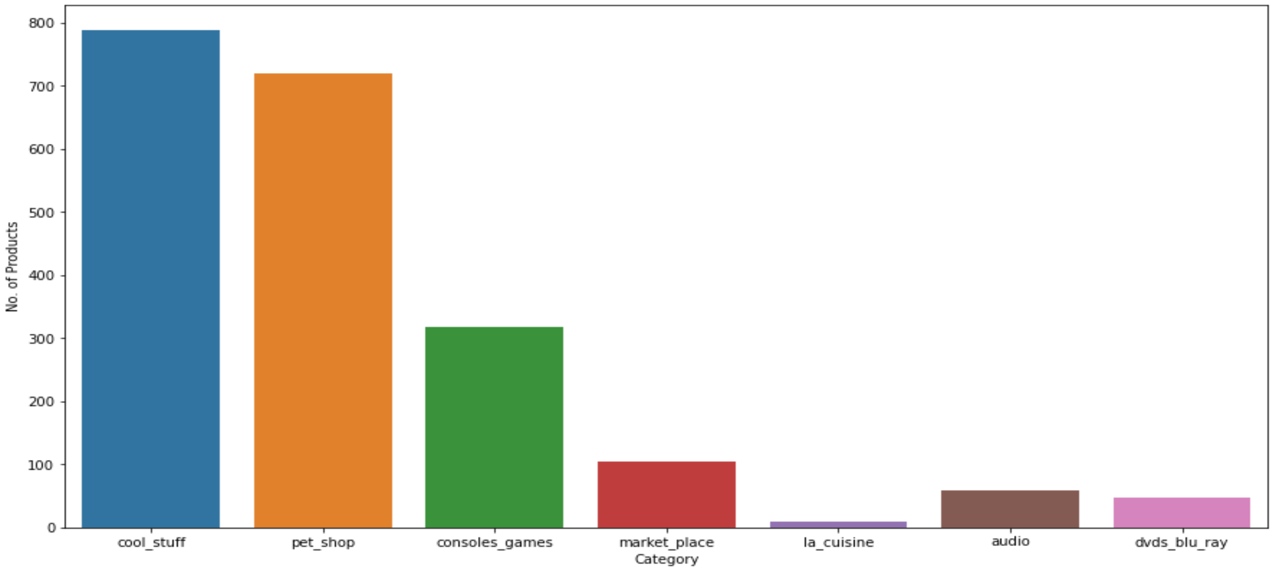

Figure 8 Product Category Distribution

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.