The Application of Machine Learning in E-commerce

Xinran Feng

Ulink High School of Suzhou Industry Park, Nantong City, Jiangsu Province, China
Email: fxr2568226391@163.com (X.R.F.)

Manuscript received October 5, 2024; revised December 14, 2024; accepted January 22, 2025; published March 24, 2025.

***Abstract*—With the help of advanced computer technology, big data, and algorithms, AI has developed from expert systems to deep learning and machine learning. This transition has opened up opportunities for AI to handle complex tasks and has revolutionized various industries. This paper presents and illustrates three e-commerce tasks: predicting purchases, segmenting customers, and analyzing baskets. Multiple algorithms, such as decision trees, cluster analysis, association rules, etc., are introduced for these tasks. Additionally, we explain how to use machine learning techniques to assess value and performance and provide context for the analysis findings. It also describes the findings of the analyses and explains how to use machine learning techniques to assess performance and value. The decision tree is the most accurate of the nine classifiers used in the same period and is more accurate in processing and predicting the results when it comes to car purchases. To ascertain the features of particular purchases and provide businesses with tailored advice to boost sales, clustering analysis was employed in the second customer segmentation case. In the third basket analysis example, association rules are employed to ascertain the frequency of purchases made by various customers as well as the quantity of goods they plan to buy simultaneously. This information helps managers better arrange the goods to increase sales.**

***Keywords*—e-commerce, machine learning, purchase prediction, customer segmentation, basket analysis**

I. Introduction

The development of Artificial Intelligence (AI) has a long history (Mehta, 2018), but it has received more and more attention in recent years. With the continuous improvement and innovation of technologies, AI has been widely applied to all fields of the national economy and people's livelihood, bringing convenience to people's lives and more business opportunities for the development of enterprises. Machine learning, natural language processing (Hirschberg, 2015), computer vision (Voulodimos, 2018), deep learning, and other technologies have made great breakthroughs, making AI gradually become an important strategic emerging industry around the world.

AI has various applications in e-commerce that bring significant value to businesses (Pallathadka *et al.*, 2023) Some key applications include personalized recommendations, chatbots and virtual assistants, image and voice recognition, predictive analytics, fraud detection and security, dynamic pricing, supply chain optimization, and sentiment analysis. A real-world example that demonstrates the value of AI in e-commerce is Amazon's personalized product recommendation system. By analyzing customer data, such as browsing and purchase history, Amazon's AI algorithms generate highly targeted recommendations. This enhances the customer experience, increases sales, and allows Amazon to gather valuable data for continuous improvement. According to a study by McKinsey, personalized recommendations can drive up to 35% of Amazon's revenue. Overall, AI in e-commerce improves customer satisfaction, optimizes operations, and drives revenue. It enables businesses to provide personalized experiences, streamline processes, and make data-driven decisions, ultimately shaping the future of the industry. With the continuous development of technology, the application of machine learning in e-commerce will continue to expand, bringing more opportunities and challenges to the e-commerce industry.

This paper focuses on three typical applications of AI in e-commerce—customer purchase prediction (Bughin *et al.*, 2017; Martínez *et al.*, 2020), customer segmentation (Qin, 2023; Yamprayaml *et al.*, 2023), and basket analysis (Aguinis *et al.*, 2013; Ünvan, 2021). These applications leverage machine learning techniques and previous customer data to enhance companies' understanding of their target market, improve marketing effectiveness, and drive expansion and profitability. Customer purchase prediction, also known as customer buying prediction or customer churn prediction, involves using machine learning algorithms and historical customer data to predict the likelihood of a customer making a purchase or discontinuing the use of a product or service. This helps businesses anticipate customer behavior and tailor their marketing strategies accordingly. Customer segmentation plays a vital role in a company's competitive positioning by dividing the customer base into distinct groups or segments based on shared characteristics, behaviors, or preferences (Andaleeb, 2016). This micro-segmentation approach strengthens a brand's values and enables a better connection with the target market, leading to more effective marketing campaigns. Basket analysis, also known as market basket analysis or association analysis, is a technique used by merchants and businesses to uncover relationships between products that are frequently purchased together. By identifying patterns and correlations among products in the same shopping basket, businesses can optimize product placement, cross-selling, and recommendations (Li *et al.*, 2005). This not only increases revenues but also enhances the overall customer shopping experience. In summary, basket analysis, customer segmentation, and purchase prediction are essential techniques that enable businesses to gain insights into their target markets, develop effective marketing strategies, and drive expansion. By leveraging these strategies, companies can gain a competitive edge, increase customer satisfaction and loyalty, and make data-driven decisions.

This paper begins by providing an overview and analysis of previous studies on the three applications mentioned. It then proceeds to demonstrate how machine learning methods can be effectively utilized for these tasks using real-world data. The performance of the models will be evaluated to gauge their effectiveness. The findings of this study hold

doi:10.18178/joebm.2025.13.1.858

significant managerial implications for companies considering the future application of AI in their operations.

II. Customer Purchase Prediction

Customer purchase prediction refers to predicting customers' future purchases based on historical behavioral data (Peker *et al.*, 2017). Businesses can forecast a customer's propensity to purchase by examining a variety of data, including demographics, browsing history, engagement patterns, and historical purchase behavior. By using this data, marketing plans can be made more individualized, client retention campaigns can be made more effective, and overall business performance may be raised. Also, businesses can estimate client purchases to anticipate customer behavior and make well-informed decisions. Businesses can take proactive steps to retain consumers, save acquisition costs, and maximize marketing efforts by precisely forecasting whether a client is likely to make a purchase or churn. Customizing marketing messages, offers, and recommendations enables firms to enhance customer satisfaction, boost conversion rates, and strengthen customer loyalty. Furthermore, consumer purchase prediction aids in the identification of prospective high-value clients and the efficient use of resources by organizations to optimize earnings. It helps to capture the dynamic preferences of customer groups for products and, in turn, recommends customers' preferred products to facilitate purchase behavior.

Many machine learning methods have been applied to predicting customers' purchase intention, such as SVM, Random-Forest (Biau & Scornet, 2016), and Gradient-Boosted Trees. The first machine learning model introduced in marketing is SVM. In the context of marketing, Cui and Curry compared SVM with the multinomial logit model and presented the result that SVM had better predictive performance. Specifically, the multinomial logit model is more suitable for presenting implications, but they found that SVM is more suitable for environments dealing with large-scale data (Lee *et al.*, 2021). Moreover, Random-Forest (Cutler *et al.*, 2012) and Gradient-Boosted Tree (Xu *et al.*, 2014) are two common ensemble models. Both make use of boosting and bagging models.

To lessen correlation, each split tree in the random forest is randomly allocated input variables. The individual trees in the random forest are built from bootstrap samples of the original data. Lastly, the final prediction is computed by averaging the prediction outcomes of each tree. In GBM, many trees are trained one after the other, and each tree increases accuracy by lowering mistakes in earlier applications (Zhang & Haghani, 2015). XGB is a model that understands scarcity and has won multiple Kaggle data science challenges.

Customer Purchase Prediction has many applications in life as well. Customers may find it overwhelming to find new makeup and beauty products, but thanks to Sephora's extensive technological offerings, they can shop with confidence knowing that the products they choose will suit their skin tone and lifestyle. Using information from Sephora's data, individual customer profiles are created based on their preferences and past purchases. The products that AI determines customers need from that data are then sorted through and displayed on the company's homepage in a personalized "Recommended for You" section. Additionally, Sephora sends tailored rewards and marketing messages to its customers based on their spending history from that year, predicting their level of loyalty. The systems are in place, and 80% of Sephora's clientele is wholly devoted to the retailer. Moreover, customers of Progressive Insurance can choose to share their driving information with them. The business has gathered billions of miles' worth of driving data, which it then feeds into an algorithm to identify the variables that influence particular driving-related problems. Data indicates which clients are more likely to have mishaps. Accurate policies result in cost savings for every customer. Progressive can eliminate bottlenecks and concentrate on delivering prompt, accurate service by better understanding the insurance market and modifying its offerings based on predictive trends.

III. Customer Segmentation

In marketing, market segmentation is the process of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers (known as segments) based on shared characteristics. Numerous factors, including purchase history, psychographics, geographies, demographics, and customer lifetime value, can be used to segment customers. To be more precise, the data were classified as either purchasing or non-purchasing clients, and the averages of the two groups were compared to determine whether the explanatory variables that made up the session differed. The majority of the variables were found to differ between the non-conversion session and the conversion session as a consequence of the analysis. Particularly, there were notable variations in visit quality metrics such as length, page view, and inflow channel. In marketing strategy and customer relationship management, consumer segmentation is essential. Mark Johnson states, "Customer Segmentation is one of the most important strategic marketing tools available to companies today." Businesses can gain a deeper understanding of their heterogeneous client base and more effective marketing strategies by segmenting their consumer group based on common attributes. This makes it possible for companies to create product offerings, tailored promotions, and targeted marketing efforts that appeal to particular clientele groups. Better consumer involvement, greater response rates, higher customer happiness, and more steadfast brand loyalty are the outcomes. Additionally, customer segmentation aids in the discovery of new consumer categories (Wu & Chou, 2011), the identification of unexplored market prospects, and the optimization of resource allocation for optimal return on investment (Jonker *et al.*, 2002). The most classical is the Clustering method, of which the most commonly used is the k-means method. The main goal of the cluster analysis algorithm K-means is to iteratively identify the k clusters in a dataset such that the total squared distance between each data point and the cluster centroid is as small as possible (Ikotun *et al.*, 2023). The K-means method has the benefits of being straightforward, simple to understand, and easy to apply. It can also handle big datasets with reasonable scalability.

Customer Segmentation has many applications in life, for example, for the National Grid. As China's biggest supplier and service provider of electricity, the State Grid must satisfy the demands and requirements of various users. Users' preferences and behaviors vary when it comes to electricity services. Algorithms for cluster analysis are used to group user electricity data, categorize users into groups based on shared electricity behavior and preferences, and create customized marketing and service plans. Customized power supply solutions and dependable power guarantee services, for instance, can be offered to industrial users; preferential tariffs and smart home solutions can be offered to residential users. The State Grid can use cluster analysis algorithms to profile users, group users with similar electricity consumption behaviors and preferences, and provide precise and customized services to various user categories to better meet the needs and expectations of various users.

IV. Basket Analysis

Market Basket Analysis is a type of data mining that identifies patterns of consumer behavior in any retail environment. It is based on transactional data, usually from point-of-sale systems or e-commerce platforms. A retailer may learn from a market basket analysis that customers frequently buy shampoo and conditioner together. In this case, offering a promotion for both products at the same time would not significantly increase revenue; however, offering a promotion for just one of the items would probably increase sales of the other. Rezende and Ladeira (2019) investigated a financial institution's market basket analysis and implemented some São Paulo state personal consumer association regulations. A detailed explanation of the data processed, including all filters and procedures, was provided. The study also included examples of various algorithms and modeled reporting on association rule algorithms. They were able to ascertain the financial institution's shopping basket based on the results, and they tested the outcomes under various regulations and circumstances. Product associations and customer purchase behavior can be better understood through the use of basket analysis (Leppänen *et al.*, 2017). This example demonstrates several possible advantages of market basket analysis via association rules, including the results' actionability, simplicity, and understandability, and a type of unsupervised technique in which the user hasn't adjusted the rule's consequent. Businesses may improve the variety, positioning, and pricing of their items by knowing which products are frequently bought in tandem. For instance, they could group related products or place them adjacent to one another to promote cross-selling and raise the average order value. Additionally, basket analysis aids in identifying popular product combinations for enterprises.

Basket analysis is widely used in business. A clothing company named "Zara" analyzed the data in their shopping baskets using basket analysis. They discovered via this analysis that a significant portion of consumers who bought particular clothing styles also bought matching accessories. Customers purchasing trench coats, for instance, would frequently purchase scarves, and those purchasing pants, belts, and so on. Based on this discovery, Zara would market and sell these accessories alongside the matching clothing to encourage customers to purchase them both at once. Sales of the accessories were successfully raised by this display, which also increased sales for Zara. Sales of the accessories were successfully raised by this display, which also increased sales for Zara. This example demonstrates how a correlation between a customer's purchasing behavior and the products in their shopping basket can be found by analyzing the information in the basket, allowing a sales strategy to be tailored appropriately.

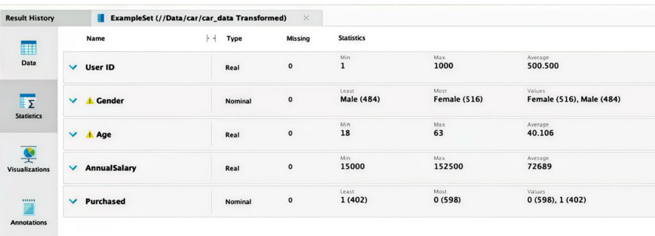
V. Results and Discussion

A. Application 1: Purchase Behavior Prediction

The target of this application is to predict car purchases using the basic features of customers. The government needs to enact appropriate transport policies and environmental protection policies by forecasting the volume of purchases, which may affect consumer demand for cars. For example, the government may introduce measures such as purchase restriction policies or emission standard restrictions to control the number of cars and environmental protection requirements, which may affect consumer demand for cars. So the government needs to forecast the purchase of cars.

Data

Our first dataset (Dataset 1) for application1 has 1000 samples, consisting of four features. The target of classification is whether the customers purchased or not purchased. The sample and summary statistics are presented in Fig. 1.



ExampleSet (//Data/car/car_data Transformed)

Name	Type	Missing	Statistics		
User ID	Real	0	Min 1	Max 1000	Average 500.500
⚠ Gender	Nominal	0	Least Male (484)	Most Female (516)	Values Female (516), Male (484)
⚠ Age	Real	0	Min 18	Max 63	Average 40.106
AnnualSalary	Real	0	Min 15000	Max 152500	Average 72689
Purchased	Nominal	0	Least 1 (402)	Most 0 (598)	Values 0 (598), 1 (402)

Fig. 1. Summary statistics of dataset 1.

The distribution of the features of Annual Salary, Age, Gender, and the target Purchased are demonstrated in Figs. 2–4. In this group, there are about the same number of men and women, and the average age is around 40. This paper also checked the association between the features and the target (see Figs. 5 and 6).

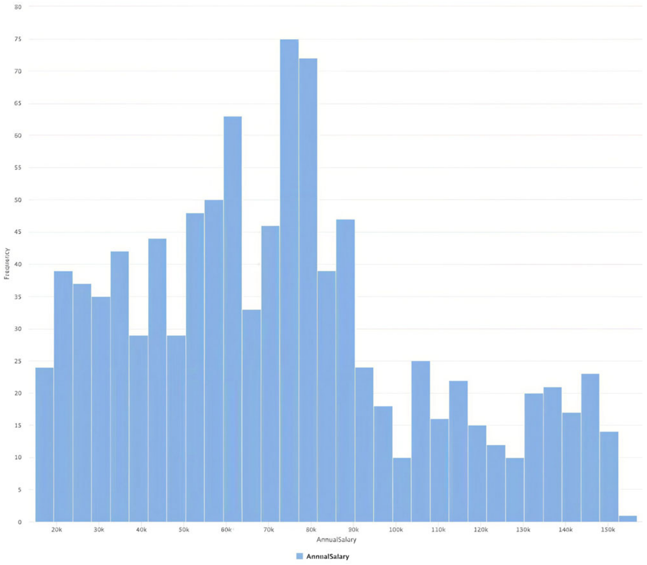


Fig. 2. Histogram of annual salary.

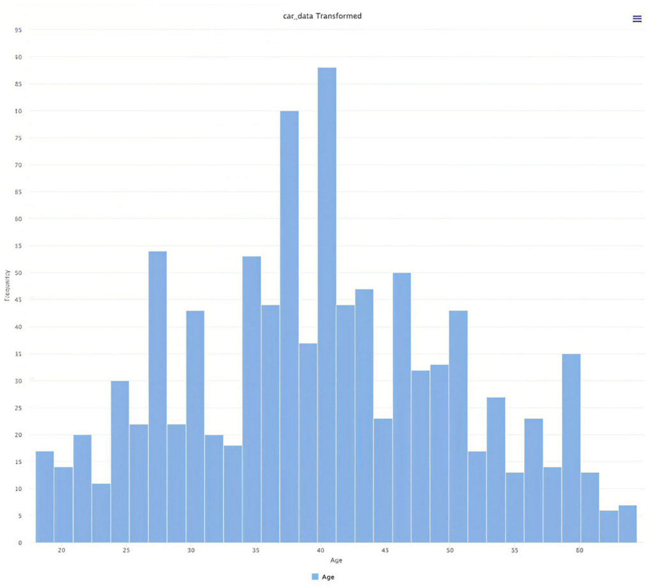


Fig. 3. Histogram of age.

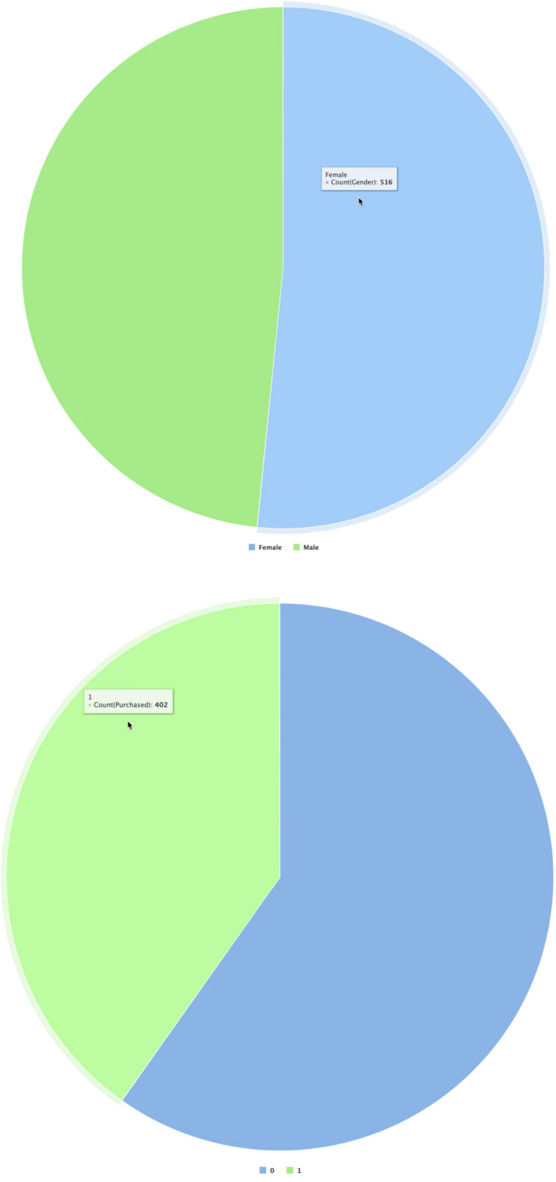


Fig. 4. Pie chart of gender and purchased count.

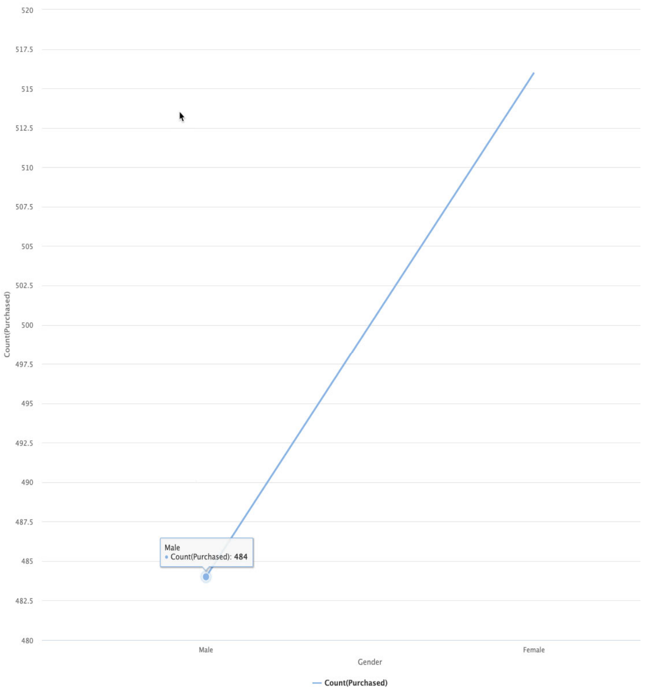


Fig. 5. Correlation of gender and purchased.

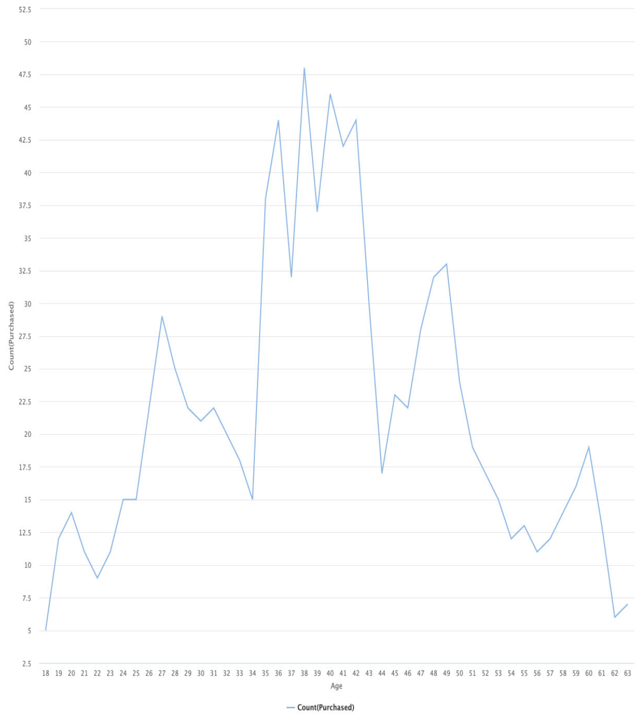


Fig. 6. Correlation of age and purchased.

Method

This paper utilized nine classifiers to predict car purchases. The classifiers include Naïve Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine. Naive Bayes classifiers are a family of linear "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features (see Bayes classifier). The term "General" Linear Model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors. Logistic Regression, also known as Logit

Regression or Logit Model, is a mathematical model used in statistics to estimate (guess) the probability of an event occurring having been given some previous data. Logistic Regression works with binary data, where either the event happens (1) or the event does not happen (0). This operator is a fast learning method for large-margin optimizations. The Fast Large Margin operator applies a fast margin learner based on the linear support vector learning scheme. Deep learning is a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data. The prototype will use a combination of deep learning, natural language processing, and dynamic network analysis to detect and examine the cross-platform spread of disinformation. A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. Random forest is a combination of decision trees that can be modeled for prediction and behavior analysis. The decision tree in a forest cannot be pruned for sampling and hence, prediction selection. Gradient Boosting is a machine learning algorithm, used for both classification and regression problems. Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks.

Results

The performance of the classifiers is measured with two metrics: accuracy and runtimes. From Figs. 7 and 8, we can see that Decision Tree has the highest accuracy of 88.8%. Second is Gradient Boosted Trees around 88.1%. Random Forest is 87.4%. Then are Naive Bayes and Deep learning, which are 83.9%. Next is Logistic Regression, then with a Generalized Linear Model. Fast Large Margin and Support Vector Machines are the least accurate, with around 59.8%. Meanwhile, the runtime of the decision tree is the shortest. We can conclude that decision tree has the best performance in car prediction applications. The possible reason could be that the simple structure of the decision tree fits our dataset with only three features.

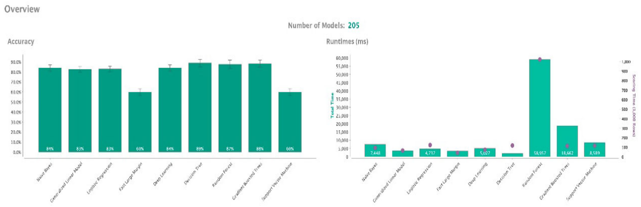
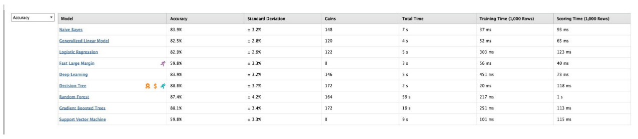


Fig. 7. Model performance of classifiers.



Accuracy ▾

Model	Accuracy	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Naive Bayes	83.9%	± 3.2%	148	7 s	37 ms	93 ms
Generalized Linear Model	82.5%	± 2.8%	120	4 s	52 ms	65 ms
Logistic Regression	82.9%	± 2.9%	122	5 s	303 ms	123 ms
Fast Large Margin	59.8%	± 3.3%	0	3 s	56 ms	40 ms
Deep Learning	83.9%	± 3.2%	146	5 s	451 ms	73 ms
Decision Tree	88.8%	± 3.7%	172	2 s	20 ms	118 ms
Random Forest	87.4%	± 4.2%	164	59 s	217 ms	1 s
Gradient Boosted Trees	88.1%	± 3.4%	172	19 s	251 ms	113 ms
Support Vector Machine	59.8%	± 3.3%	0	9 s	101 ms	115 ms

Fig. 8. Accuracy of classifiers.

We then look into the tree constructed in Fig. 9. There are three rules according to the decision tree. When the annual salary amount is larger than 92000, the event of purchasing the car is 1, which means those people have a good chance of buying a car. For those people who have salaries below 92000 per year, their purchases also depend on age. Those older than 45.5 are more likely to buy a car, and the event of buying a car is 1. Those younger than 45.5 are 0, and they may not have enough to buy a car.

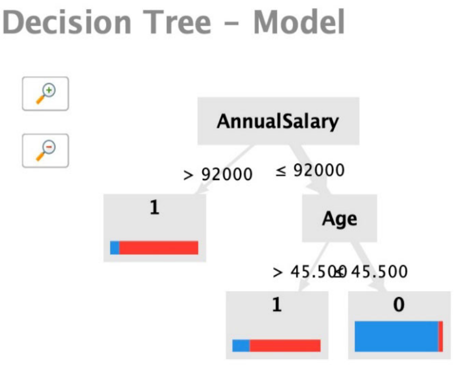


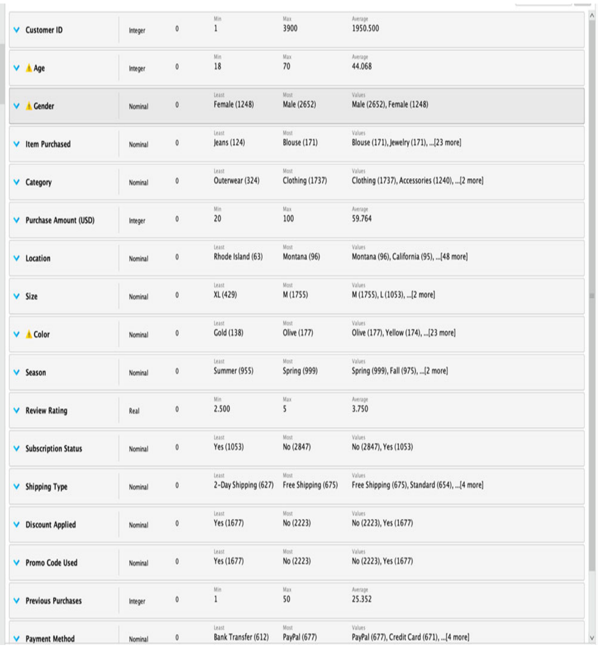
Fig. 9. Decision tree.

B. *Application 2: Customer Segmentation*

The basic information of shopping trends is used here and its customer data is categorized and then the purchase amount of each is evaluated in order to provide different personalized advice to different customers.

Data

The dataset (Dataset 2) for application1 has 1000 samples, consisting of sixteen features, including customer ID, age, gender, item purchased, category, purchase amount (USD), location, size, color, season, review rating, subscription status, Shipping type, discount applied, promo code used, previous purchases, payment method. Some of them may be useless in this case, like consumer ID, and location, which need to be deleted before calculation. (See Fig. 10)



Customer ID	Integer	0	Min 1	Max 3900	Average 1950.500
⚠ Age	Integer	0	Min 18	Max 70	Average 44.068
⚠ Gender	Nominal	0	Least Female (1248)	Most Male (2652)	Values Male (2652), Female (1248)
Item Purchased	Nominal	0	Least Jeans (124)	Most Blouse (171)	Values Blouse (171), Jewelry (171), ...[23 more]
Category	Nominal	0	Least Outerwear (324)	Most Clothing (1737)	Values Clothing (1737), Accessories (1240), ...[2 more]
Purchase Amount (USD)	Integer	0	Min 20	Max 100	Average 59.764
Location	Nominal	0	Least Rhode Island (63)	Most Montana (96)	Values Montana (96), California (95), ...[48 more]
Size	Nominal	0	Least XL (429)	Most M (1755)	Values M (1755), L (1053), ...[2 more]
⚠ Color	Nominal	0	Least Gold (138)	Most Olive (177)	Values Olive (177), Yellow (174), ...[23 more]
Season	Nominal	0	Least Summer (955)	Most Spring (999)	Values Spring (999), Fall (975), ...[2 more]
Review Rating	Real	0	Min 2.500	Max 5	Average 3.750
Subscription Status	Nominal	0	Least Yes (1053)	Most No (2847)	Values No (2847), Yes (1053)
Shipping Type	Nominal	0	Least 2-Day Shipping (627)	Most Free Shipping (675)	Values Free Shipping (675), Standard (654), ...[4 more]
Discount Applied	Nominal	0	Least Yes (1677)	Most No (2223)	Values No (2223), Yes (1677)
Promo Code Used	Nominal	0	Least Yes (1677)	Most No (2223)	Values No (2223), Yes (1677)
Previous Purchases	Integer	0	Min 1	Max 50	Average 25.352
Payment Method	Nominal	0	Least Bank Transfer (612)	Most PayPal (677)	Values PayPal (677), Credit Card (671), ...[4 more]

Fig. 10. Sample of dataset 2.

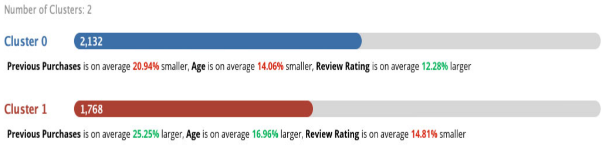
Method K-means

This data on shopping trends is grouped by the well-known clustering technique K-means according to purchase amount or frequency. Each item is assigned to the cluster with the nearest center of mass, using k centers of mass as the centroids of the clusters. For instance, when it comes to buying amounts, the center of mass is recalculated by averaging the purchase amounts of all the items in each cluster once the data has been assigned to them. Until the center of mass stabilizes or the maximum number of iterations is reached, this process is repeated. The k-means technique is easy to implement, effective, and beneficial for large datasets with known numbers of clusters. The k-means algorithm is frequently extended and modified for market segmentation purposes.

Results

Clustering is divided into two categories. In the first category, the purchase count is around 2100. Previous purchases are on average 20.9%, age is on average 14.1% smaller, and review rating is on average 12.3% larger. (See Fig. 11)

In the second category, the purchase count is around 1800. Previous purchases are on average 25.3%, Age is 16.9% larger, and review rating is on average 14.8% smaller. From this graph, it can be analyzed that the second category is more valuable and customers have a higher propensity to buy, which can be a key marketing target so that sales may rise.

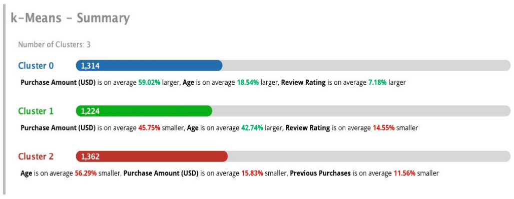


Previous Purchases is on average 20.94% smaller, **Age** is on average 14.06% smaller, **Review Rating** is on average 12.28% larger

Previous Purchases is on average 25.25% larger, **Age** is on average 16.96% larger, **Review Rating** is on average 14.81% smaller

Fig. 11. Two categories of data.

Clustering can also be divided into three categories. In the first category, the purchase count is around 59.0%, age is on average 18.5% larger, and review rating is on average 7.2% larger. In the second category, the purchase count is on average 45.8%. Age is on average 42.8%, review rating is on average 14.6% smaller. In the third category, Age is on average 56.3%, purchase count is around 15.8%, and Previous purchases are on average 11.6% smaller. Despite having a lower rating and a higher purchase amount, the first category may indicate that the buyer is dissatisfied with the goods. These clients may be older and have more purchasing power, but they are not as happy with their previous shopping experiences. Therefore, businesses want to maximize this category of customers' purchasing experiences to increase customer satisfaction and encourage repeat business. Businesses can target consumers in the second category, where ratings are higher but purchases are lower, with marketing efforts in the hopes of boosting sales. (See Fig. 12)



Purchase Amount (USD) is on average 59.02% larger, **Age** is on average 18.54% larger, **Review Rating** is on average 7.18% larger

Purchase Amount (USD) is on average 45.75% smaller, **Age** is on average 42.74% larger, **Review Rating** is on average 14.55% smaller

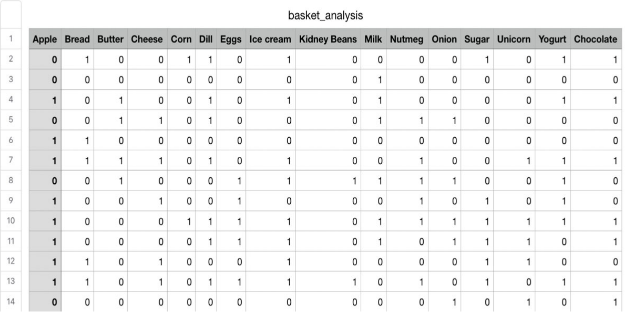
Age is on average 56.29% smaller, **Purchase Amount (USD)** is on average 15.83% smaller, **Previous Purchases** is on average 11.56% smaller

Fig. 12. Three categories of data.

C. *Application 3: basket analysis (cross-selling)*

Data

The third dataset (Dataset 3) for application 3 has 1000 samples, consisting of sixteen categories, including apple, bread, butter, cheese, corn, dill, eggs, ice cream, kidney beans, milk, nutmeg, onion, sugar, unicorn, yogurt, chocolate. The target of classification is to observe which two products are best sold together. The sample and summary statistics are presented in Fig. 13.



basket_analysis

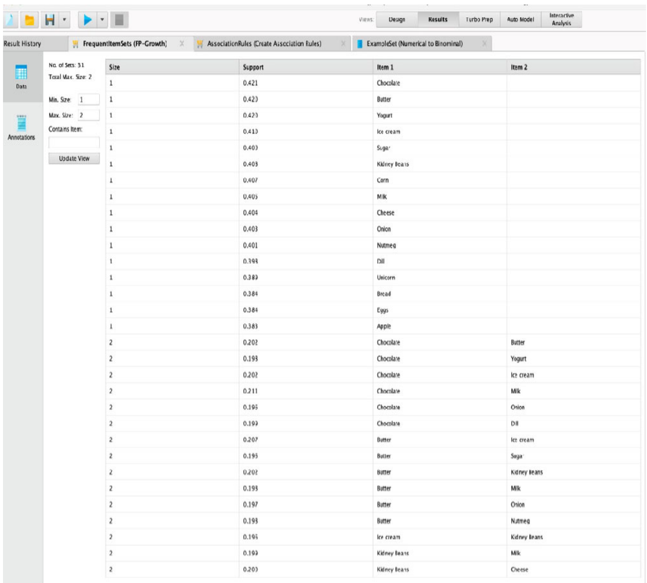
Apple	Bread	Butter	Cheese	Corn	Dill	Eggs	Ice cream	Kidney Beans	Milk	Nutmeg	Onion	Sugar	Unicorn	Yogurt	Chocolate
0	1	0	0	1	1	0	1	0	0	0	0	1	0	1	1
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	1	0	1	0	1	0	0	0	0	1	1
0	0	1	1	0	1	0	0	0	1	1	1	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	0	1	0	0	1	0	0	1	1	1
0	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0
1	0	0	1	0	0	1	0	0	0	1	0	1	0	1	0
1	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1
1	0	0	0	0	1	1	1	0	1	0	1	1	1	0	1
1	1	0	1	0	0	0	1	0	0	0	0	1	1	0	0
1	1	0	1	0	1	1	1	1	0	1	0	1	0	1	1
0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1

Fig. 13. Sample of dataset 3.

Method Association rules

Association rules are a tool for discovering patterns in large data sets. Here they reveal relationships between purchased goods that often occur with some regularity. Association rules are usually expressed in the form of "X->Y", e.g., here chocolate and milk, i.e., if one buys chocolate, one also buys milk, with a certain level of confidence. Common metrics for evaluating association rules include support, confidence, lift, and leverage. Association rules can be mined by algorithms such as Apriori, FP-Growth, and Eclat. By identifying items or events that frequently occur together, association rules provide decision-makers with valuable information in a variety of areas such as marketing, medical, and social science research.

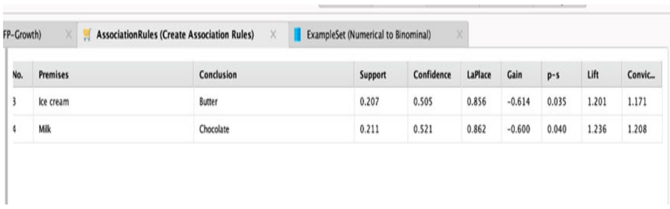
Results


Views:
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
Result History
FrequentItemSets (FP-Growth)
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)
Data
Annotations
Update View

No. of Sets: 31
Total Max. Size: 2
Min. Size: 1
Max. Size: 2
Contains Item:

Size	Support	Item 1	Item 2
1	0.421	Chocolate	
1	0.420	Butter	
1	0.420	Yogurt	
1	0.410	Ice cream	
1	0.409	Sugar	
1	0.408	Kidney Beans	
1	0.407	Corn	
1	0.405	Milk	
1	0.404	Cheese	
1	0.403	Onion	
1	0.401	Nutmeg	
1	0.398	Dill	
1	0.389	Unicorn	
1	0.384	Bread	
1	0.384	Eggs	
1	0.383	Apple	
2	0.202	Chocolate	Butter
2	0.193	Chocolate	Yogurt
2	0.202	Chocolate	Ice cream
2	0.211	Chocolate	Milk
2	0.195	Chocolate	Onion
2	0.199	Chocolate	Dill
2	0.207	Butter	Ice cream
2	0.195	Butter	Sugar
2	0.202	Butter	Kidney Beans
2	0.193	Butter	Milk
2	0.197	Butter	Onion
2	0.193	Butter	Nutmeg
2	0.195	Ice cream	Kidney Beans
2	0.199	Kidney Beans	Milk
2	0.209	Kidney Beans	Cheese

Fig. 14. Support of items.


FP-Growth)
AssociationRules (Create Association Rules)
ExampleSet (Numerical to Binominal)

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convic...
3	Ice cream	Butter	0.207	0.505	0.856	-0.614	0.035	1.201	1.171
4	Milk	Chocolate	0.211	0.521	0.862	-0.600	0.040	1.236	1.208

Fig. 15. Support of 2 items.

From Fig. 14, it can be seen that at size 1, chocolate has the highest support of about 0.421. The next highest support is butter and yogurt at 0.420, then ice cream and sugar, milk and cheese, etc. are in the third place. When the size is 2, chocolate and butter have the highest support of 0.202, followed by chocolate and yogurt, chocolate and ice cream, and so on. The most special is chocolate and onion, their support is 0.196 is also relatively high, it can be seen that people often buy these two items together.

In Fig. 15, it can be seen that customers often buy ice cream and butter together and milk and chocolate together and their support is 0.207 and 0.211 respectively and confidence is 0.505 and 0.521 respectively. So if ice cream and butter are placed together and milk and chocolate are placed together, it can increase the sales of both items.

VI. Conclusion

This paper presents the application of machine learning methods to 3 types of e-commerce tasks. Based on three real commercial data, several machine learning methods are used to show how the application can be done, including Decision Trees, gradient-boosted trees, K-means, Association rules, and other methods. The study had some specific findings, for example, decision trees worked best for predicting car purchases, which had the highest accuracy. In the second group of applications, the second category is more valuable has a higher propensity to buy, and can be used as a key marketing target. In basket analysis, the frequency of buying two items together is analyzed, and selling butter and ice cream together and placing chocolate and milk together can boost sales.

There are some limitations in this paper, for example, in the first task, too few features were used. For example, in the second task, there is no evaluation of how many classes are best for clustering. The clustering algorithm needs to be constantly adjusted for sample classification, which can take a very long time to compute when the amount of data is very large. In the third application, the results of association rules are generally judged by visual observation to determine whether they are reasonable or not, rather than by explicit evaluation metrics, which may be wrong. In future research, it could be evaluated how well the results of these data analyses work in real applications.

Conflict of Interest

The author declares no conflict of interest.

References

Aguinis, H., Forcum, L. E., & Joo, H. 2013. Using market basket analysis in management research. ***Journal of Management***, 39(7): 1799–1824.

Andaleeb, S. S. 2016. Market segmentation, targeting, and positioning. ***In Strategic marketing management in Asia: case studies and lessons across industries:*** 179–207. Emerald Group Publishing Limited.

Biau, G., & Scornet, E. 2016. A random forest-guided tour. ***Test***, 25: 197–227.

Bughin, J., Hazan, E. *et al.* 2017. Artificial intelligence is the next digital frontier.

Cutler, A., Cutler, D. R., & Stevens, J. R. 2012. Random forests. ***Ensemble Machine Learning: Methods and Applications***. 157–175.

Hirschberg, J., & Manning, C. D. 2015. Advances in natural language processing. ***Science***, 349(6245): 261–266.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. ***Information Sciences***, 622: 178–210.

Jonker, J. J., Piersma, N., & Van den Poel, D. 2004. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. ***Expert Systems with Applications***, 27(2): 159–168.

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. 2021. A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. ***Journal of Theoretical and Applied Electronic Commerce Research***, 16(5): 1472–1491.

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. 2017, September. Data-driven news generation for automated journalism. ***Proceedings of 10th International Conference on Natural Language Generation***: 188–197).

Li, S., Sun, B., & Wilcox, R. T. 2005. Cross-selling sequentially ordered products: An application to consumer banking services. ***Journal of Marketing Research***, 42(2): 233–239.

Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C., & Haltmeier, M. 2020. A machine learning framework for customer purchase prediction in the non-contractual setting. ***European Journal of Operational Research***, 281(3): 588–596.

Mehta, N., & Devarakonda, M. V. 2018. Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey? ***Journal of Allergy and Clinical Immunology***, 141(6): 2019–2021.

Pallathadka, H., Ramirez-Asis, E. H., Loli-Poma, T. P., Kaliyaperumal, K., Ventayen, R. J. M., & Naved, M. 2023. Applications of artificial intelligence in business management, e-commerce, and finance. ***Materials Today: Proceedings***, 80: 2610–2613.

Peker, S., Kocyigit, A., & Eren, P. E. 2017. A hybrid approach for predicting customers' purchase behavior. ***Kybernetes***, 46(10): 1614–1631.

Qin, F. 2023. Essays on Market Segmentation and Retailers' Competing Strategies (Doctoral dissertation, Purdue University).

Rezende, F., & Ladeira, M. 2019. Market basket analysis in a financial institution. Singular. ***Engenharia, Tecnologia e Gestão***, 1(1): 6–12.

Ünvan, Y. A. 2021. Market basket analysis with association rules. ***Communications in Statistics-Theory and Methods***, 50(7): 1615–1628.

Voulodimos, A., Doulamis, N., Bebis, G., & Stathaki, T. 2018. Recent developments in deep learning for engineering applications. ***Computational Intelligence and Neuroscience, 2018***.

Wu, R. S., & Chou, P. H. 2011. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. ***Electronic Commerce Research and Applications***, 10(3): 331–341.

Xu, Z., Huang, G., Weinberger, K. Q., & Zheng, A. X. 2014. Gradient boosted feature selection. ***Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining***: 522–531.

Yamprayaml, K., Choemprayong, S., & Taiphapoon, T. 2023. Current practices, challenges, and opportunities for lifestyle-based market segmentation of older consumers in Thailand. ***Proceedings of 2023 International Conference on Cyber Management and Engineering (CyMaEn)*** IEEE: 530–535.

Zhang, Y., & Haghani, A. 2015. A gradient boosting method to improve travel time prediction. ***Transportation Research Part C: Emerging Technologies***, 58: 308–324.