


Data-Driven Consumer Behaviour Prediction: Key Factors and Machine Learning Approaches

Yiling Wang ^a

University College London School of Management, University College London, London, U.K.

Keywords: Consumer Behaviour Prediction, Machine Learning, Big Data Analytics, Marketing Strategy, Data Privacy.

Abstract: In today's data-driven market environment, consumer behaviour prediction has become an indispensable tool for companies seeking to develop more targeted marketing strategies, optimise supply chains, and enhance customer experience. This article systematically explores the internal and external factors that influence consumer behaviour prediction, including psychological characteristics, demographic factors, market trends, advertising campaigns, and social media influences. It also examines the intricate interplay between these elements, highlighting how they can collectively shape purchasing decisions. In addition, the article analyses in detail different prediction methods, ranging from traditional statistical models (e.g., linear regression and decision trees) to emerging machine learning techniques (e.g., integrated learning and clustered classification models), and assesses their respective strengths and weaknesses. Addressing challenges such as data privacy, data imbalance, model interpretability, and real-time performance, this paper proposes a series of practical solutions while looking ahead to future trends in consumer behaviour prediction. Ultimately, the research provides valuable insights for enterprises, marketers, and researchers to optimise data-driven marketing strategies and business decisions.

1 INTRODUCTION


In the digital age, understanding and predicting consumer behaviour has become an important part of marketing, business planning and economic forecasting. Accurate prediction of consumer preferences and purchasing decisions help companies to optimise marketing strategies, product offerings and enhance customer satisfaction. Consumer behaviour prediction is influenced by factors such as psychological characteristics, demographics, social influences and market dynamics. With the development of big data and artificial intelligence, prediction methods have evolved from traditional statistical models to sophisticated machine learning techniques to improve accuracy.

This paper explores the key factors affecting the prediction of consumer behaviour, which are divided into two dimensions: internal (emotions, values, demographic characteristics) and external (market trends, advertising strategies, social media, macroeconomics). Meanwhile, different prediction methods from linear regression and decision trees to

integrated learning and cluster classification are analysed and their strengths and weaknesses are evaluated to provide practical guidance.

In addition, this paper discusses the challenges of consumer behaviour prediction, such as data privacy, data imbalance, model interpretability and real-time prediction challenges, and explores solutions such as privacy protection, data augmentation and interpretable AI. The outlook emphasises the combination of ethical AI, real-time data processing and innovative machine learning frameworks to enhance the accuracy and reliability of predictive models.

Through multi-dimensional analysis of consumer behaviour and the development of prediction technologies, this paper aims to deepen the understanding of the consumer decision-making process, provide data-driven strategic support for enterprises, marketers and policymakers, and help build a more efficient and personalised consumer interaction system.

^a <https://orcid.org/0009-0004-0086-5108>

2 FACTORS INFLUENCING THE PREDICTION OF CONSUMER BEHAVIOUR

In the field of consumer behaviour forecasting, the study of influencing factors is crucial. These factors not only determine consumers' purchasing decisions, but also influence market trends and corporate marketing strategies. From the view of internal factors, consumer behaviour is driven by psychological characteristics, social identity, personality preferences, etc. From the view of the external environment, market supply and demand, advertising, social media influence, etc. also play a key role. In addition, with the development of big data and artificial intelligence, data such as consumers' historical behaviours, search records and product evaluations have become an important basis for predicting consumer behaviour.

2.1 Internal Factors

Consumers' emotions play an important role in shopping decisions. Research has shown that pleasurable moods typically motivate consumers to make faster, more impulsive purchasing decisions, while negative moods reduce the desire to buy or prompt more cautious choices. For example, during the festive season, consumers tend to be willing to spend more due to pleasurable moods, while stress or anxiety may increase the demand for comfort or emotional comfort products.

In addition, demographic characteristics significantly influence consumer behaviour. Age, gender, income, occupation and education level determine consumption preferences, e.g. younger consumers favor personalised products, while older consumers pay more attention to practicality and quality assurance. Highly educated groups tend to consume rationally, and there are also differences in product choices between consumers of different genders.

Values also influence consumption decisions, especially in terms of environmental protection, sustainable development and social responsibility. For example, environmentally conscious consumers tend to choose recyclable or eco-friendly products, while socially responsible consumers are more likely to support brands with good CSR performance. As the concept of sustainable consumption becomes more popular, companies are also starting to introduce products that are in line with consumer values.

Family structure is also a key factor. Family size and life cycle influence consumption patterns, such as families with children spending more on food and education, while single consumers are more inclined to spend on personal interests and experiences. Newly married families prioritise household products, while retired groups are more concerned with health and retirement products.

Income level directly restricts consumers' purchasing power. Low-income groups are more concerned about value for money and tend to prefer discounted goods, while high-income groups place more emphasis on brand, quality and personalised experience. For example, in the luxury goods market, high-income earners are more likely to accept high premiums, while low and middle-income earners tend to favor promotions or substitutes.

Price sensitivity affects how responsive consumers are to price changes. High price sensitivity consumers tend to buy promotional items or wait for discounts, while low price sensitivity consumers are more concerned with product value. For example, in the electronics market, some consumers will wait for price reductions to buy, while loyal users are willing to pay a premium to get new products. This suggests that price works in conjunction with other psychological and social factors to influence the final purchase decision.

2.2 External Factors

In the digital age, the Internet and social media profoundly influence consumer behaviour. The popularity of online shopping has made product information readily available to consumers, and social media advertising and recommendation algorithms have enhanced this trend. Users often refer to reviews, watch review videos and discuss product experiences on social media platforms before making purchases, and this socialised shopping model enhances the interaction between brands and consumers.

Advertising and promotional strategies directly shape consumer decisions. Precise advertising, discount promotions and membership offers can effectively stimulate the desire to buy. For example, promotions such as limited-time discounts and buy-one-get-one-free boost sales, while brand precision marketing attracts target users and influences their brand choice through personalised recommendations.

Brand awareness, image and reputation affect consumer trust and loyalty. Well-known brands are more likely to win favour by virtue of their market accumulation and word-of-mouth, such as Apple and

Nike, which shape a high degree of loyalty and remain competitive even at higher prices. In addition, a brand's social responsibility, innovation and after-sales service also affect long-term loyalty.

Public emergencies change consumer demand and behaviour. During major disasters or public health crises, demand shifts from non-essentials to necessities and health products. For example, during the New Crown epidemic, demand for protective gear and telecommuting equipment surged, while the travel and entertainment sectors were hit. In addition, such events prompt consumers to pay more attention to food safety, health protection and emergency stockpiling, influencing long-term consumption trends.

3 CONSUMER BEHAVIOUR FORECASTING METHODS

3.1 Conventional Prediction Methods

3.1.1 Linear Regression

Linear Regression (LR) is a basic statistical method for consumer behaviour prediction, which assumes that there is a linear relationship between the independent variables (influencing factors) and the dependent variable (consumer behaviour). For example, it can be used to analyse the relationship between consumers' purchase amount and frequency of purchase and factors such as income level, advertising investment and brand loyalty. Multiple linear regression improves predictive accuracy by introducing multiple independent variables (Thalji, 2022).

Thalji's study used multiple linear regression to explore the impact of after-sales service information on consumer purchase decisions on e-commerce platforms (Thalji, 2022). Based on 533 questionnaires from the city of Dammam, Saudi Arabia, the study analysed the effect of return policy, maintenance, communication, and warranty on purchasing decisions, and the results showed that these factors were significantly and positively related to purchasing behaviour. In addition, the study used hierarchical clustering and K-mean clustering methods to group consumers by behavioural and demographic characteristics to reveal the attention of different groups to after-sales service information.

Multiple linear regression has the advantage of being able to analyse the effects of multiple variables on consumer behaviour simultaneously, capturing

complex relationships. For example, in addition to after-sales service factors, the study can incorporate variables such as consumers' demographic characteristics (age, gender, income level) as well as psychological characteristics (brand loyalty, purchase motivation) to provide a more comprehensive forecast. The method is particularly suitable for consumption scenarios where linear relationships are more obvious, such as pricing sensitivity analyses and evaluation of the effectiveness of promotional activities.

3.1.2 Classification Decision Trees

However, Gkikas et al. pointed out that decision trees are susceptible to overfitting and have weak generalisation ability. For this reason, the study combines Genetic Algorithm (GA) to optimise the decision tree structure, reduce redundancy through feature selection and pruning, and improve model robustness. In addition, the study suggests the use of integrated learning methods (e.g., Random Forest, Gradient Boosting Tree) to enhance the classification performance and improve the prediction stability and accuracy (Gkikas et al., 2022)

3.1.3 Nearest Neighbour Algorithm

The KNN algorithm is an instance-based learning method that selects the nearest K neighbors by calculating the distance between a new sample and the samples in the training dataset and performs classification or regression based on the categories of these neighbors. In consumer behaviour prediction, KNN is widely used to analyse user preferences. For example, Putra & Ilmi used KNN algorithms for consumer behaviour identification and product personalisation recommendation (Putra & Ilmi, 2024).

Putra & Ilmi's study shows that KNN can effectively capture consumer purchase patterns and improve the accuracy of personalised recommendations in a big data environment (Putra & Ilmi, 2024). The study uses KNN for consumer classification and product matching based on data such as social media interactions, purchase history and product browsing behaviour. The results show that consumers with similar behaviours tend to prefer similar products, which improves the effect of precision marketing. In addition, KNN does not require a complex training process and is suitable for real-time recommendation systems on e-commerce platforms, providing support for dynamic adjustment of marketing strategies.

However, KNN also has the disadvantages of high computational complexity and sensitivity to data size. Computing distance on large-scale datasets consumes more resources, and noisy or redundant features may affect prediction accuracy. The researchers suggest optimising data quality through principal component analysis (PCA) dimensionality reduction, feature selection, and accelerating the computation by combining it with KD trees or ball trees.

Another study PLOS ONE extends the application of KNN in online shopping behaviour prediction Azad et al. (Azad et al., 2023). The study analysed consumers' browsing history, shopping cart behaviour, purchase records and other data, and the results showed that KNN can effectively identify consumers' potential needs, accurately recommend products, and improve the effect of personalised marketing.

3.2 Novel Prediction Methods

In the field of consumer behaviour prediction, with the continuous advancement of data analytics, researchers have developed a variety of novel methods to improve the accuracy and reliability of predictions. The following are several common novel prediction methods.

3.2.1 Integrated Learning Methods

Integrated learning improves prediction by combining the results of multiple models. The method reduces the bias and variance of a single model and improves prediction robustness. Common methods include Bagging, Boosting, and Stacking. e.g., Bagging can train multiple models by sampling the dataset multiple times and reduce the risk of overfitting by averaging or voting.

Liu et al. showed that integrated learning demonstrated strong comprehensive capabilities in consumer behaviour prediction (Liu et al., 2020). The study combines models such as decision trees, support vector machines, and neural networks to construct an integrated learning framework to improve prediction accuracy. The results show that Bagging effectively reduces overfitting and improves model generalisation, especially for complex and non-linear data.

Verma further explored the application of Bagging method in the field of e-commerce (Verma, 2020). The study trains multiple decision tree models through random sampling and uses a voting mechanism to synthesise the results, reducing variance and enhancing the robustness of the model

to noisy data, especially when dealing with high-dimensional data.

The application of Boosting methods in consumer behaviour prediction has also attracted much attention. Alizamir et al. used Gradient Boosting Tree (GBT) to analyse consumers' purchase history, browsing behaviour and demographic characteristics to predict their future purchase propensity (Alizamir et al., 2022). It was found that Boosting performs well with unbalanced data (e.g., few class samples) and can effectively identify potential high-value customers.

3.2.2 Hybrid Method of Clustering and Classification

The hybrid method of clustering and classification improves prediction accuracy by first analysing consumers by clustering and then applying classification or regression models to each group. Common clustering algorithms include K-means and hierarchical clustering, which can be used by companies to identify different groups of consumers and develop precise marketing strategies. For example, Cai & Rodavia (2022) used K-means to perform a cluster analysis of consumer behaviour and applied classification models across the clusters to optimise purchase prediction.

Cai & Rodavia's study shows that the approach combines the pattern recognition capabilities of cluster analysis with the accurate prediction capabilities of classification models to improve the accuracy of the models. The study used the K-means algorithm to cluster consumers based on purchase history, browsing behaviour and demographic characteristics, and applied decision tree or logistic regression models within each cluster to predict purchase propensity.

In addition, Cai & Rodavia further validated the effectiveness of the method by using K-means clustering and applying a random forest classification model within each cluster. The results show that the method can handle the heterogeneity of consumer behaviour data more effectively and improve the prediction performance. For example, the study identifies high-value customers and price-sensitive customers and develops differentiated marketing strategies for different clusters, which significantly improves marketing effectiveness.

Hierarchical clustering eliminates the need to preset the number of clusters, and can be applied to exploratory analyses by showing the hierarchical structure of data through a tree diagram. For example, Anitha & Patil (2022) used the K-Means algorithm

combined with the RFM (Recency, Frequency, Monetary) model to perform cluster analysis of consumer purchase behaviour, identify different consumer groups, and develop targeted marketing strategies.

4 EXISTING LIMITATIONS AND FUTURE OUTLOOK

4.1 Current Status of Data Privacy

With the in-depth analysis of consumer behaviour by predictive models, the risk of user privacy leakage increases significantly. For example, unauthorised data collection and data misuse issues are frequent, making user information security a serious challenge. Therefore, future development requires further improvement of laws and policies to ensure legal and compliant use of data, and enhancement of data security through encrypted storage, access control and other technical means.

4.2 Challenge of Data Imbalance Problem

The data imbalance problem is a common challenge in consumer behaviour prediction, especially when dealing with a few types of data (e.g., customer churn, uncommon purchasing behaviours), the model often struggles to learn efficiently, resulting in a degradation of prediction performance. To address this problem, a synthetic minority class oversampling technique (SMOTE) can be used to balance the data distribution by generating new minority class samples. In addition, the WOE (Weight of Evidence) method can be used for variable binning and transformation to compute the ratio of minority classes to majority classes in each feature bin to improve model performance on unbalanced data (Hambali & Andrew, 2024).

4.3 The Need for Model Interpretability

In the application of predictive models, it is crucial to understand the logic behind the prediction results, especially in highly sensitive fields such as healthcare and finance, and improving model interpretability can help to enhance decision transparency and user trust. Flowchart visualisation techniques can be used to show the decision path of the model, for example, in

decision tree models, using branch nodes to visually present the prediction process. In addition, interactive visualisation tools (e.g. Tableau, Power BI) can dynamically display the model's prediction process and feature contributions, enabling users to gain a deeper understanding of the model's decision-making mechanism (Kovalerchuk et al., 2023).

4.4 Real-time challenges of models

The problem of latency in data processing has been a challenge for consumer behaviour prediction models, mainly due to the long time consuming data preprocessing and feature engineering, which affects the real-time performance of the model. To improve the real-time performance of the model, the computational cost can be reduced by streamlining the feature set and selecting key features without significantly reducing the model performance. In addition, optimising the real-time data acquisition process and adopting incremental learning methods can reduce the redundant steps in data processing and improve the real-time prediction capability of the model.

4.5 Future Outlook

With the continuous progress of data analysis technology and machine learning algorithms, future consumer behaviour prediction models will be continuously optimized in terms of data security, data quality, interpretability and real-time performance. The parallel promotion of data privacy protection through legal and technical means, the improvement of fairness in data collection and processing, and the combination of efficient model optimisation strategies will enable predictive models to play a greater role in business decision-making and user experience enhancement.

5 CONCLUSIONS

Consumer behaviour forecasting is becoming increasingly important in the modern business world. Through in-depth analysis of the internal and external factors that influence consumer behaviour, companies are able to identify market needs more accurately and optimise their marketing strategies. This study examines the key role of factors such as consumer psychology, demographic characteristics, social media influence, and advertising in predicting consumer behaviour. In addition, with the advancement of data analytics and artificial

intelligence technologies, the combination of traditional statistical methods and modern machine learning algorithms significantly improves the accuracy and usefulness of forecasting.

Despite significant progress in accuracy and applicability, current prediction methods still face many challenges. For example, data privacy protection is a growing concern, and organisations need to follow strict compliance requirements when using consumer data for forecasting. In addition, the problem of data imbalance makes the model limited in dealing with niche consumer behaviours, and how to optimise data sampling and model training becomes an important direction for future research. Meanwhile, model interpretability and real-time performance are also key factors that affect the effectiveness of consumer behaviour prediction, and companies need to find a balance between increasing model complexity and transparency.

Looking ahead, with the further development of big data, deep learning and personalised recommendation technologies, the accuracy of consumer behaviour prediction will continue to improve. Enterprises need to continuously optimise their data collection and analysis methods and combine them with the latest AI technologies to ensure the reliability and feasibility of prediction results. In addition, strengthening interdisciplinary cooperation, such as combining research in the fields of behavioural economics and psychology, will help to understand consumer behaviour patterns more comprehensively. Ultimately, while utilising data-driven strategies, companies should also focus on ethical and legal norms in order to achieve sustainable development and long-term competitive advantage in the market.

REFERENCES

- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, 34(5), 1785-1792.
- Azad, M. S., Khan, S. S., Hossain, R., Rahman, R., & Momen, S. (2023). Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights. *PLOS ONE*, 18(12), e0296336. <https://doi.org/10.1371/journal.pone.0296336>
- Cai, K., & Rodavia, M. R. (2022, October). K-Means Cluster Analysis Based on Consumer Behavior. In 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM) (pp. 143-146). IEEE.
- Gkikas, D. C., Theodoridis, P. K., & Beligiannis, G. N. (2022, May). Enhanced marketing decision making for consumer behaviour classification using binary decision trees and a genetic algorithm wrapper. In *Informatics* (Vol. 9, No. 2, p. 45). MDPI.
- Hambali, M. A., & Andrew, I. (2024). Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis. *Qeios*, Mar.
- Kovalerchuk, B., Dunn, A., Worland, A., & Wagle, S. (2023). Interactive decision tree creation and enhancement with complete visualization for explainable modeling. *arXiv preprint arXiv:2305.18432*.
- Liu, X., Zhang, Y., Chen, W., & Li, H. (2020). Ensemble learning-based consumer behavior prediction: A hybrid model approach. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), 3725-3738. <https://doi.org/10.1109/TNNLS.2020.2995362>
- Putra, S. E., & Ilmi, M. (2024). Application of K-Nearest Neighbor Algorithm for Consumer Behaviour Identification and Product Personalisation Based on Big Data Analysis. *Jurnal Ecotipe (Electronic, Control, Telecommunication, Information, and Power Engineering)*, 11(2), 205-213.
- Thalji, Z. J. (2022). Using multiple linear regression model to predict the customers' purchase decision based on after-sales services. *Research Square*.
- Verma, A. (2020). Consumer behaviour in retail: Next logical purchase using deep neural network. *arXiv preprint arXiv:2010.06952*.