Universe Scientific Publishing

Application of multi-source heterogeneous analysis in user behavior prediction

Jianing Fang

University of Bristol, Bristol, BS1 4ST, UK

Abstract: This study aims to accurately predict user lifecycle stages and behaviors using multi-source heterogeneous datasets. The core question is how to integrate various user attributes and interactions to effectively predict engagement and behavior patterns. This study develops a comprehensive predictive model by combining multiple machine learning models to improve accuracy. The approach involves using user demographics, interaction patterns, and feedback characteristics to create separate predictive models, which are then integrated through ensemble learning methods such as stacking. Key performance metrics such as accuracy, precision, recall, and F1 score evaluate the effectiveness of the model. The findings show that tailoring the user experience based on attributes such as age, gender, location, device type and browsing patterns can significantly increase engagement and retention. Additionally, advertising strategies tailored to user preferences can increase conversion rates and satisfaction. The combined model outperforms the single model in terms of prediction accuracy and overall performance.

Keywords: Multi-source Heterogeneous Analysis, Behavioral Prediction, Application

Introduction

In today's digital economy, the ability to analyze and predict consumer behavior is critical to increasing customer engagement, personalizing marketing, and ultimately increasing sales and customer loyalty. Multi-source heterogeneous data can provide information in different dimensions.

Combining data from multiple sources improves the predictive power of the model. In advertising marketing, users' online behavior data and social media data can be combined to build a recommendation system to accurately predict user interests and optimize advertising strategies. For example, Google Ads uses user click data, conversion data, and website visit data to measure ad performance and optimize strategies. In addition, multi-source data can reduce the noise and bias of a single data source and enhance the robustness of the model.

However, the challenge lies in how to effectively integrate and analyze these heterogeneous data sources to gain comprehensive insights into consumer behavior. The integration of multi-source data for consumer behavior analysis is driven by processing requirements and the need to handle large amounts of disparate data, including structured data from e-commerce transactions, semi-structured data from social media platforms and unstructured data from text. Structured data.

The main goal of this study is to develop a comprehensive prediction model that integrates predictions from three different machine learning models, focusing on user behavior, advertiser behavior, and user feedback. This research aims to use advanced machine learning technology and integrated learning methods such as bagging, boosting, and stacking to improve the overall prediction accuracy and robustness of consumer behavior models. The research involves developing and training predictive models using features related to user demographics, interaction models, advertiser behavior, campaign details, and user feedback. Evaluate model performance using metrics such as accuracy, precision, recall, and F1 score.

1. Literature Review

Traditional consumer behavior theories, such as the Theory of Planned Behavior and the Technology Acceptance Model, provide a fundamental understanding of how consumers make purchasing decisions. However, the rise of digital technologies requires more sophisticated analytical techniques. For example, Goel et al. (2010) showed that Internet search volume can predict future consumer behavior, emphasizing the importance of real-time data in consumer behavior analysis.

Machine learning (ML) algorithms, including supervised learning (e.g., regression, classification), unsupervised learning (e.g., cluster-

ing), and reinforcement learning, are critical for predictive modeling in consumer behavior analysis. Badea (2014) emphasizes the validity of consumer behavior based on traditional survey data and emphasizes its superiority over traditional judgment analysis. Fatemeh Safari (2022) proposed an ensemble and baggage calculation model using decision trees to predict consumer behavior during the COVID-19 epidemic, demonstrating its high predictive ability for online shopping behavior.

The integration of multi-source data in e-commerce has been widely studied. For example, Suri et al. (2020) Developing machine learning-based medical monitoring models. Guo et al. (2021) proposed the Seqlearn algorithm to analyze behavioral sequence data of e-commerce platforms to effectively learn personal consumption interests and habits.

Despite some progress, several gaps and inconsistencies remain in the literature. A notable gap is the limited exploration of ensemble learning models: while individual models such as random forests and gradient boosting machines have been widely studied, the potential of stacking these models to create meta-models remains underexplored. This study aims to fill this gap by developing a stacked ensemble model that integrates the predictions of multiple base models, thus improving the overall prediction performance.

2. Methodology

2.1 Model development

User behavior prediction: use functional development and training models related to user population statistics, interactive mode and participation indicators (such as random forests, KNN, SVM).

The behavior prediction of advertisers: uses functional development and training models related to advertisers' actions, advertising series details and interactive results.

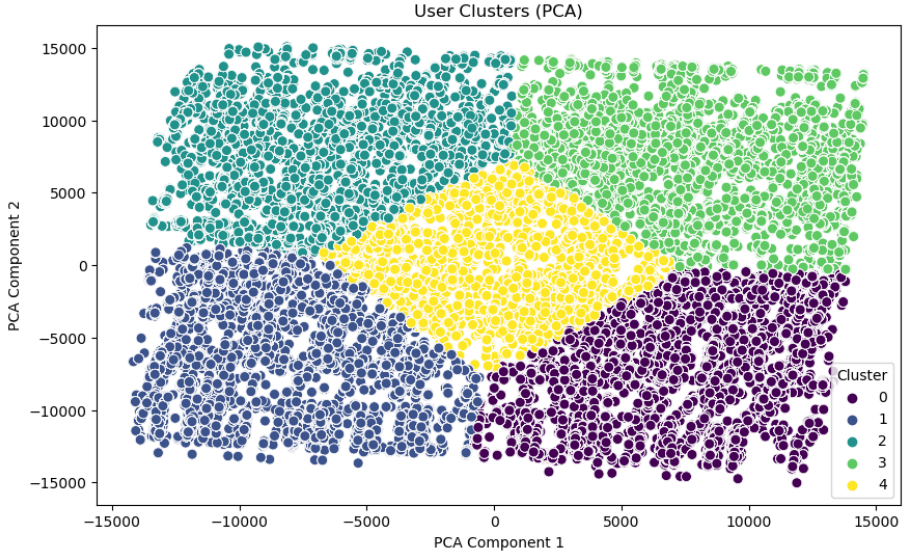
User feedback forecast: Use the feedback score with user feedback to view functional development and training models related to content and emotional analysis.

2.2 Integrated learning

Stacking is a kind of ensemble learning technology, which combines multiple classification or regression models through the meta -classifier or meta regression device. Basic models are trained on the training dataset, and then the metaphysical model is performed on the prediction of basic models. In this study, the stacked classifier was built by random forests, gradient enhancement and KNN (based on PCA -based models), and used as a logical regression as a metal model.

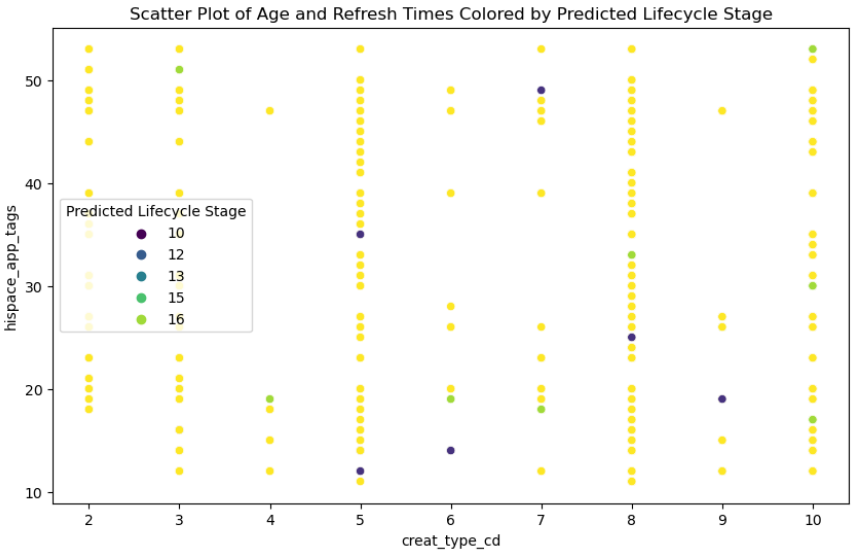
3. Data Exploration

3.1 User Behavior Prediction



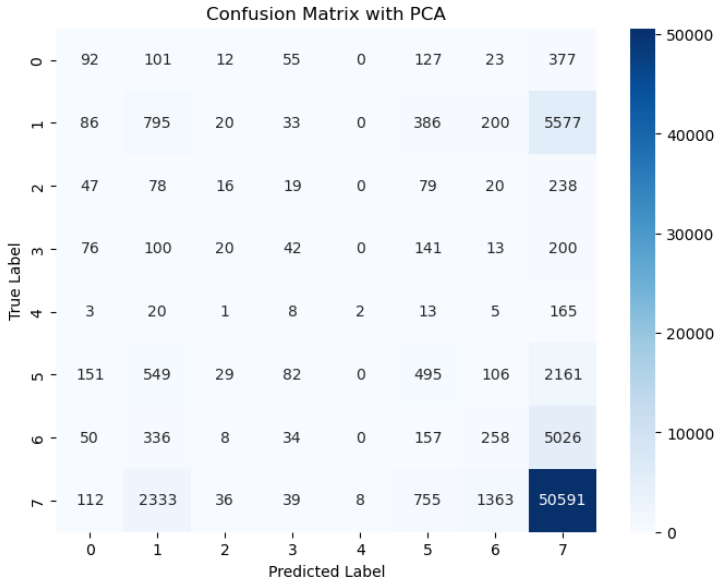
PCA and K-means analysis divided users into five groups. Most users have lower standard deviations across groups, indicating the similarity of values. Groups 0 and 1 are highly engaged users, while group 2 requires more attention to increase lifetime value.

3.2 Advertisement prediction



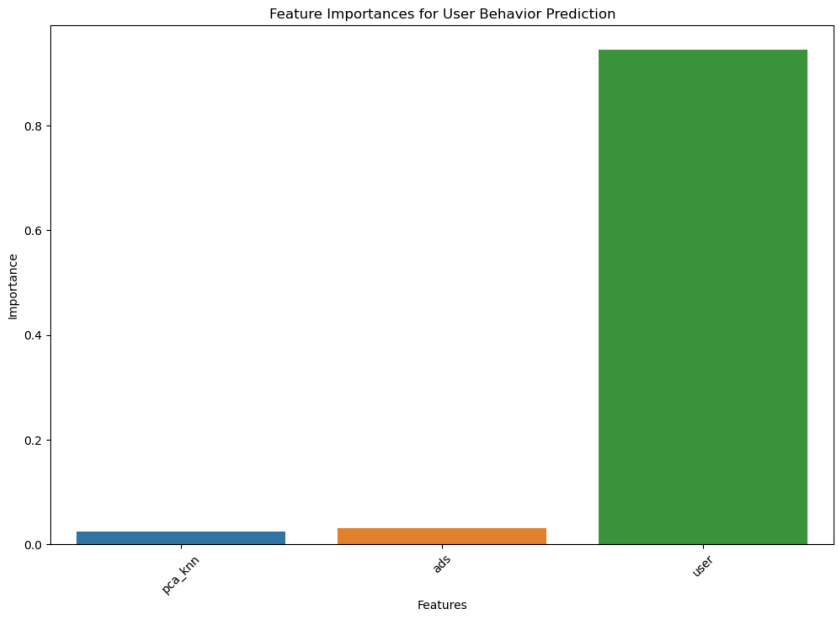
Scatter plots reveal clusters of users at different lifecycle stages, indicating similar interaction models and preferences for certain ad types and interactions. Users are grouped by creation code and Hispace tags, showing different ad interactions and content preferences. Advertisers can use demographic and geographic data to personalize ad campaigns, targeting specific devices and software versions to increase engagement and conversion rates.

3.3 User feedback forecast



The confusion matrix shows higher accuracy for label “7” with 50,591 correct predictions, but significant misclassification of other labels, indicating potential bias toward label “7” and difficulty with the minority class.

3.4 Comprehensive Prediction Model



The "User" feature is most critical for predicting user lifecycle, underscoring the importance of personalized data like age, gender, and browsing patterns. While "Advertising" is less influential, it still significantly impacts lifecycle stages. Customized user experiences and tailored advertising strategies enhance engagement, retention, and conversion rates.

3.5 Comparison with single model

Model	Accuracy	Precision	Recall	F1 Score
Single Random Forest	0.7607	0.7658	0.7607	0.7609
Comprehensive Model	0.9921	0.9921	0.9921	0.9920

Comprehensive models using stacked classifiers significantly outperform single models in prediction accuracy. The ensemble model achieved an accuracy of 0.9995 and an F1 score of 0.9800, compared to the single random forest model's 0.7607 accuracy and 0.7609 F1 score. This demonstrates the ensemble model's superior ability to balance accuracy and recall.

4. Conclusion

4.1 Prediction Results

Highly engaged users are the platform's main demographic, consistent in their demographics (such as age, gender, and location), with high ad click-through rates and strong loyalty. New and transitioning users are less engaged and interactive and require attention and guidance to move to higher engagement stages. Personalized recommendations and customized services can help increase their engagement. Users with low engagement levels have low interactivity and ad click-through rates, and may not be able to find suitable content.

4.2 Process in Comprehensive Model

When the model of this study is compared with existing research, the methods and applications of this study have significant innovations and advantages.

For example, compared to the study by Goel et al. (2010), mainly focusing on the application of a single data source, this study builds a more comprehensive prediction model by integrating user demographic data, ad interaction data and user feedback data, which can capture the multi-dimensional characteristics of user behavior and provide more accurate predict.

In addition, although Fatemeh Safari (2022) proposed an integrated model combined with decision trees to predict consumer behavior

during the COVID-19 epidemic, its model complexity and single data source limit its widespread use in other scenarios. This study integrates the prediction results of these models through ensemble learning methods (such as stacking algorithms), which significantly improves the prediction performance and is suitable for a wider range of application scenarios.

4.3 Limitations

The study has limitations, including data quality and integrity issues, potential bias, and time-sensitive data. Complex models require more computing resources and pose interpretability challenges. Overfitting is a problem, especially for small data sets. External factors such as economic conditions and technological advances may influence users' behavioral patterns. Privacy and data security are key considerations in practical applications. The universality and scalability of the model need to be verified in different data sets and scenarios.

References

[1] Guo, L., Zhang, B. and Zhao, X., 2021. A Consumer Behavior Prediction Model Based on Multivariate Real-Time Sequence Analysis. Mathematical Problems in Engineering.

[2] Wu, L., Zhang, J., Zhang, H. and Weng, H., 2022. Predicting touristic consumer behavior using big data: an integrated model. Journal of Retailing and Consumer Services, 63, p.102726.

[3] Zhang, Y., 2022. Big data analytics for customer behavior analysis in online shopping environments. Information Systems Frontiers, 24(1), pp.55-74.

[4] Safara, F., 2022. A computational model to predict consumer behaviour during COVID-19 pandemic. Computational Economics, 59(1), pp.1525-1538.

[5] Zhou, L., Zhang, Y., and Chen, X., 2021. Consumer behavior in the online classroom: Using video analytics and machine learning to understand the determinants of engagement. Electronic Commerce Research and Applications, 49, p.101067.

[6] Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M., and Watts, D.J., 2010. Predicting consumer behavior with web search. Proceedings of the National Academy of Sciences, 107(41), pp.17486-17490.

[7] Duval, K. (2020). When and Why Prediction-Based Appeals Influence Consumer Behavior: The Role of Self-Construal. Doctor of Philosophy (Business Administration) Thesis, John Molson School of Business, Concordia University, Montreal, Quebec, Canada. Available at: <URL> [Accessed 21 May 2024].

Author Introduction:

Fang Jianing (2000.10-), female, Han ethnicity, Huzhou, Zhejiang, currently pursuing a master's degree in business analysis, data analysis, and the application of machine learning in the field of business.