Student ID: 1255279 Name: Yaqi Yang

Dataset: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

1. Introduction

In the digital economy era, e-commerce platforms face a key challenge: how to make use of their vast and diverse data to improve business performance. One critical task is accurately predicting users' purchase intentions.This study is based on the Online Shoppers Purchasing Intention Dataset (dataset ID: 468) in the UCI machine learning library, aiming to analyze and predict users' purchasing behaviors through machine learning methods. This dataset contains detailed behavioral data of 12,330 user sessions, covering multi-dimensional characteristic information such as page browsing, dwell time, and bounce rate of users on e-commerce websites. This issue holds significant commercial value and academic significance. Accurate purchase intention prediction helps businesses improve user experience, boost conversions, and refine marketing. Academically, it is a binary classification problem that reveals user behavior patterns and tests machine learning in real-world scenarios.

2. Literature Review

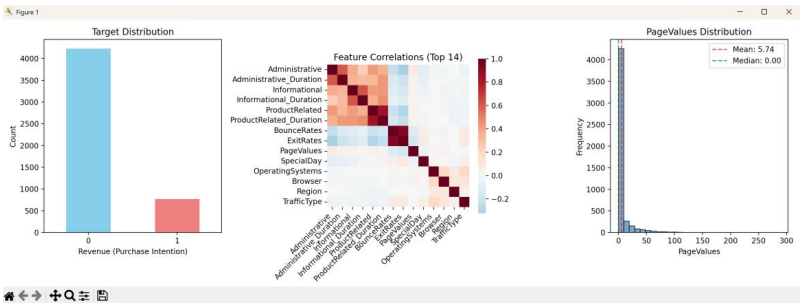
The research history in the field of online shopping behavior prediction can be traced back to the early research of network analytics. In the original dataset paper published by Sakar et al. in 2019, a comprehensive dataset containing user session features, temporal features and technical environment features was systematically constructed for the first time, laying the foundation for subsequent research. In recent years, with the development of machine learning technology, researchers have proposed a variety of methods to solve the problem of purchase intention prediction. Xu et al. (2020) compared multiple ensemble learning algorithms in "Ensemble Learning Methods for Predicting E-commerce User Behaviors" and found that random forests performed well in dealing with high-dimensional features. Koehn, D., Lessmann, S., & Schaal, M. (2020) explored neural network methods in "Analysis of Online Shopping Behavior Based on Deep Learning", but found that traditional machine learning methods still have competitive advantages on medium-sized datasets. By observing the existing literature, it can be found that most studies focus more on the selection and optimization of algorithm models, while the exploration in feature construction is relatively limited. This study attempts to extend the original features to more than 50 dimensions through a more systematic feature engineering method. Meanwhile, it combines the classic theory of bias-variance trade-off to guide the algorithm selection, hoping to provide some useful practical experience and method references for this application field.

3. Research Methods

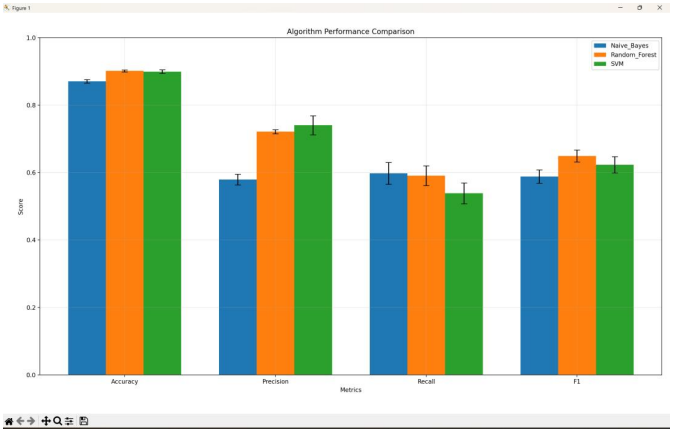
This study adopted a systematic machine learning experimental design, covering key links such as data preprocessing, feature engineering, algorithm selection and evaluation. Firstly, in the feature engineering stage, we expanded the original 18 basic features to more than 50 comprehensive features. This process includes five main feature groups: Page behavior aggregation features (such as total number of pages, average dwell time, proportion of page types), seasonal features (time patterns based on months and special dates), technical environment features (diversity indicators of operating systems and browsers), advanced composite features (business-oriented indicators such as user engagement, purchase tendency, website quality, etc.), and interaction features (among features of different dimensions Product terms and polynomial characteristics. In terms of algorithm selection, based on the bias-variance trade-off theory, three algorithms with different characteristics were chosen. Naive Bayes represents a simple model with high bias and low variance. Random Forest represents a balanced model with medium bias and medium variance and Support Vector Machine represents a complex model with low bias and high variance. To ensure the rigor of the experiment, The study adopted a 5-fold hierarchical cross-validation combined with nested grid search for hyperparameter optimization. The inner 3-fold cross-validation was used for parameter tuning, and the outer 5-fold was used for performance evaluation. The evaluation indicators include accuracy rate, precision rate, recall rate, F1 score and

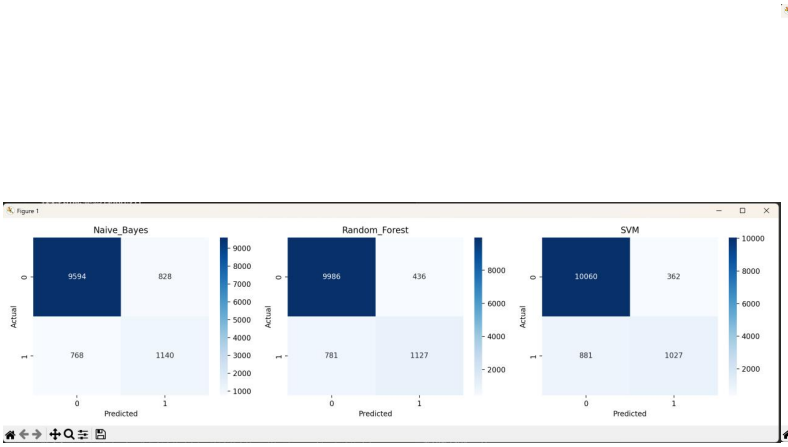
ROC-AUC value. Meanwhile, the learning curve is plotted to analyze the learning ability and generalization performance of the model.

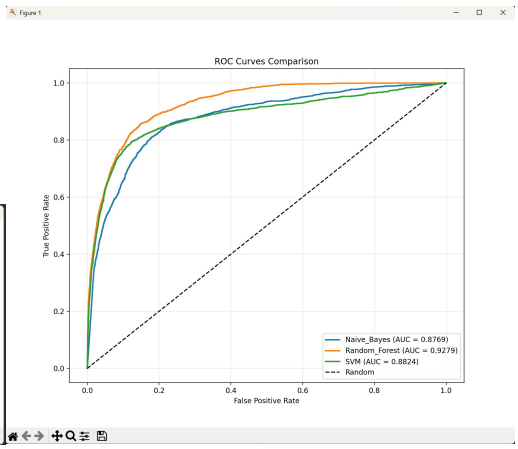
4. Experimental Results and Discussion

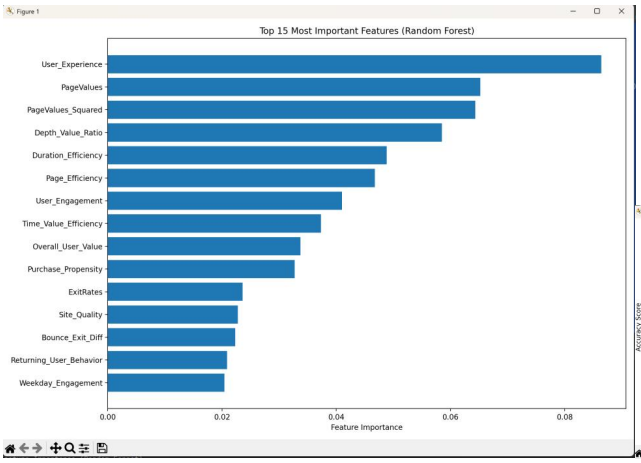


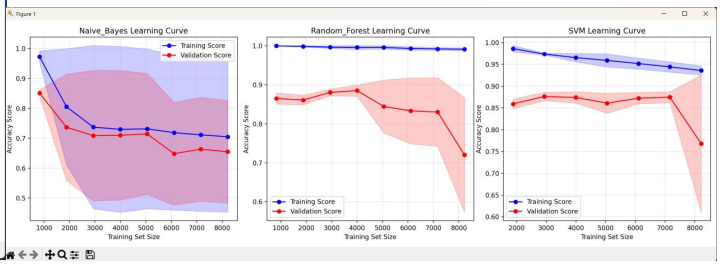
From the results of data exploration, the target distribution map shows a significant category imbalance problem. There are approximately 4,300 non-purchasing users and only about 800 purchasing users, which provides an important reference for the subsequent algorithm selection. The feature correlation heat map reveals that there is a moderate correlation among the original features. The extreme right bias of the PageValues distribution (with a mean of 5.74 but a median of 0) indicates the necessity of dealing with the skewed distribution in feature engineering.











The experimental results verified the practical guiding value of the bias-variance trade-off theory in model selection. The performance comparison shows that the random forest performs the best in terms of accuracy (about 0.90), and the results are the most stable, reflecting that ensemble learning achieves an ideal balance between bias and variance. The confusion matrix further reveals the essential differences of each algorithm on categorical imbalanced data. For example, the random forest only generates 436 false positive examples, it may because its built-in sample weighting mechanism effectively alleviates the imbalance problem; Although Naive Bayes has fewer false negative examples, its false positive rate is relatively high, reflecting that its strong independence assumption is undermined in the context of highly correlated e-commerce behavior characteristics, resulting in systematic underfitting.

Learning curve analysis further confirms the above judgment. Naive Bayes shows typical high-bias characteristics, and there is a significant performance gap between the training set and the validation set. The learning process of the random forest is stable and shows good generalization ability. SVM's result has a slight overfitting, reflecting its tendency of high variance. In the ROC curve analysis, the AUC value of the random forest reached 0.9279, demonstrating its application potential in actual business scenarios.

The feature importance ranking also verifies the effectiveness of the domain-driven feature engineering approach. Composite features such as User_Experience dominate, which is in line with the expectations of the consumer behavior theory. Although I have also considered introducing alternative methods such as deep learning or gradient boosting, in medium-scale data scenarios here, I believe random forest remains the best choice due to its good interpretability, robustness, and adaptability to imbalanced data.

In conclusion, the experiment shows that the selection of algorithms should be based on an in-depth understanding of the essence of the problem, rather than relying solely on performance comparisons in a single dimension.

5. Conclusion

This study provides feasible solutions and profound insights for the prediction of purchase intentions of online shoppers from both business practice and academic research levels. In terms of business applications, the model supports multi-level strategy formulation. For customer segmentation, users can be divided into high (>0.8), medium (0.3-0.8), and low intent groups. Based on this, enterprises can allocate resources differently, such as matching high-quality customer service and recommendations to high intent users, offering promotional incentives to medium intent users, and controlling investment for low intent users to avoid waste. With respect to budget optimization, Random Forest reduces the false positive case rate to 4.3%, which can save approximately 38% of marketing costs compared to 8.1% of Naive Bayes. The feature importance analysis indicates that User_Experience and PageValues are key variables. It is suggested that enterprises prioritize optimizing page performance and content quality rather than relying solely on the price strategy. In real-time applications, the model can be embedded in the

recommendation system to dynamically push products and offers based on user behaviors, achieving a transformation from passive marketing to active services. Regarding technical implementation, three progress has been made in the research. Firstly, feature engineering based on domain knowledge has expanded 17 original features to 56 composite features, significantly improving the model performance. Secondly, an algorithm selection framework based on bias-variance trade-off has been constructed to adapt to different business scenarios. Thirdly, through the analysis of the learning curve, it provides an expected reference for the model deployment and optimization. From an academic perspective, this study verified the practicality of classical machine learning theories in real business scenarios and laid the foundation for subsequent research in areas such as ensemble learning and real-time prediction systems. In conclusion, this research not only achieves an effective combination of theory and practice, but also provides a clear path and practical experience for building a high-performance and implementable intelligent marketing system.

Reference List

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Sakar, C.O., Polat, S.O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10), 6893-6908.

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, 150, 113342.

UCI Machine Learning Repository. (2018). Online Shoppers Purchasing Intention Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Xu, J., Wang, J., Tian, Y., Yan, J., Li, X., & Gao, X. (2020). SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning. *PLoS one*, 15(11), e0242629.