Enhancing Online Purchase Prediction: A Comparison of Supervised Learning and Behavioural Clustering

Ane Novrup Larsen
Nathasja Skov Fink Nielsen

1 Abstract

E-commerce platforms attract large volumes of visitors, yet only a small fraction complete a purchase. This project is a comparative study focusing on whether segmenting visitors based on their navigational engagement can improve the accuracy of online purchase intent prediction using machine learning on session-level behavioural data. Using the Online Shoppers Purchasing Intention dataset, Logistic Regression, Decision Tree, and Random Forest classifiers are evaluated using standard classification and probability-based metrics. Finding the most promising model (Random Forest), we tune it before comparing it with the tuned model with cluster labels. The behavioural segmentation using clustering was done with K-Means. Through analysis and comparison, we found that clustering did not improve the accuracy, and we actually saw a minimal decrease in multiple metrics. In conclusion, a tuned Random Forest was the best prediction model of those we compared.

2 Introduction

E-commerce platforms attract many visitors, yet only a small proportion of sessions result in a completed purchase [3]. This makes early identification of potential buyers valuable for supporting personalisation and targeted marketing. As a result, predicting online purchase intent has become an important application of machine learning.

Modern e-commerce systems collect session-level behavioural data, such as navigation patterns, page visits, and time spent on content, which can be used by supervised learning models to predict purchase outcomes [1]. However, user behaviour varies significantly between sessions, and raw behavioural features may not fully capture these differences.

One way to deal with this is to use unsupervised learning to group sessions that show similar patterns of interaction. Previous work by Huseynov et al. (2016) demonstrated that online shoppers can be segmented into distinct behavioural groups (e.g., 'browsers' vs. 'buyers') based on session metrics.[2]. The idea in this project is to test whether using these behavioural groups as additional input to supervised models leads to better predictions of purchase intent.

3 Problem Formulation

E-commerce platforms collect detailed behavioural data from visitors, and this data is widely used in supervised machine learning models to predict whether a session will result in a purchase. While such models often achieve strong performance, they are primarily based on raw behavioural features and may therefore struggle to reflect meaningful differences in visitor behaviour.

In practice, not all visitors interact with an online store in the same way. In some sessions, visitors primarily browse the website, whereas other sessions show more purchase-focused navigation. These differences suggest that visitor behaviour may contain underlying patterns that are not fully captured by individual session metrics alone.

This project therefore evaluates whether behavioural segmentation using unsupervised clustering actually improves the accuracy of purchase intention prediction, compared to relying on raw behavioural features alone.

Research question:

Does segmenting online visitors based on their navigational engagement (Unsupervised Learning) improve the accuracy of predicting purchase intention (Supervised Learning) compared to using raw traffic metrics alone?

4 Methods

This study uses a comparative experimental design. We evaluate supervised classification models against a hybrid approach that integrates unsupervised clustering features, measuring performance differences.

4.1 Dataset

The project is based on the *Online Shoppers Purchasing Intention* dataset from the UCI Machine Learning Repository [6]. The data consists of records from individual visits to an e-commerce website, where each record describes how a user interacted with the site during a visit. In this work, the task is to compare whether predicting if a visit resulted in a purchase or not, which is indicated by

the Revenue variable, will have improved accuracy with the use of clusters to segment the visitors.

The data is divided into training and test sets. Since sessions resulting in a purchase are relatively rare, stratified sampling is applied based on the target variable (Revenue) to ensure that both sets contain a similar proportion of purchasing and non-purchasing sessions. The data is split before any preprocessing or training is done.

4.2 Preprocessing

Numerical features are standardised with StandardScaler to ensure comparable scales across features, while categorical features are transformed using One-Hot Encoding. Preprocessing is integrated into the model pipelines so that all transformations are learned exclusively from the training data and applied consistently to the test data.

4.3 Supervised Learning

Three supervised learning models are used in this project: logistic regression, decision tree, and random forest. Logistic regression is included as a simple and interpretable baseline model, while decision trees allow the model to capture non-linear relationships in user behaviour. Random forest is included as a more robust ensemble model, as it combines multiple decision trees to improve generalisation and handle tabular data effectively.

To allow for a fair comparison, all models are trained using the same preprocessing pipeline and evaluated on the same held-out test set with identical metrics. Based on this initial comparison, random forest showed the strongest performance and was therefore selected for further hyperparameter tuning before the final evaluation.

4.4 Unsupervised Learning (Behavioural Segmentation)

For the clustering we used K-Means. We first started with selecting the behavioural features such as pages visited (Administrative, ProductRelated etc), ProductRelated_Duration, BounceRates, and PageValues. These columns were then used for the clustering algorithm and, after the preprocessing, we used the Elbow method[5] and silhouette score[4] to find the optimal number of clusters for our dataset.

We also evaluated DBSCAN to investigate if the data contained non-convex (concave) clusters that the spherical assumption of K-Means might miss. However, DBSCAN resulted in 33.86% noise with an epsilon of 0.5, while increasing epsilon merged the points into one big cluster, leading us to proceed with K-Means.

After the clustering, we took the tuned model from supervised learning and added the K-Means clusters to the categorical features. Using the same tuned pipeline and evaluation function, we could then compare the resulting metrics with that of the tuned model without clusters.

5 Analysis

5.1 Evaluation Metrics

To analyse whether the clustering improves the accuracy of predicting purchase intention, we made use of the metrics precision, recall, and F1-score to evaluate the model's ability to correctly identify even minority customers with purchase intention, balance false positives and false negatives, and assess overall classification performance beyond accuracy.

We utilized ROC-AUC to see how well the models differentiate between buyer and non-buyer independent of the decision threshold, and log-loss to quantify the uncertainty of the predictions, penalising confident but incorrect predictions.

Finally, with the use of Youden's Index[7], we found the optimal decision threshold for each model that maximizes the separation between the True Positive Rate and False Positive Rate. This optimization was important for handling the class imbalance (85/15 buy/no-buy ratio), since the default threshold of 0.5 often struggles to correctly identify the minority class.

5.2 Baseline and Tuned Supervised Model Evaluation

We compared baseline models of Logistic Regression, Decision Trees, and Random Forest and found that Random Forest has the highest ROC-AUC (0.92) and Precision (0.79), but only a Recall of 0.40 meaning that it is more conservative and does not accept buyers unless it is very sure. Although Logistic Regression found more buyers (Recall 0.79), it had too many false alarms (Precision 0.43). While the Decision Tree achieved higher recall (0.53) than the baseline Random Forest (0.40), it had a significantly higher log-loss (5.36) and lower ROC-AUC (0.71), indicating poor probability esti-

Table 1: Performance comparison of Baseline and Tuned Models

Metric	Logistic Reg.	Decision Tree	Random Forest	Tuned RF
Class 1 (Shoppers) Performance				
Precision	0.43	0.52	**0.79**	0.53
Recall	0.79	0.53	0.40	**0.84**
F1-Score	0.56	0.52	0.53	**0.65**
Overall Model Performance				
Optimal Threshold	0.4375	1.0000	0.5500	0.3961
ROC-AUC	0.8932	0.7195	0.9188	**0.9224**
Log-Loss	0.4559	5.3642	**0.2575**	0.2988

mation and stability. For comparison metrics, see Table 1.

We chose Random Forest because it is the most robust and accurate model, which aligns with our research question that focuses on the accuracy of predicting purchase intention.

After tuning the Random Forest model, we found that the tuned model has lower precision (0.53) and log-loss (0.2987) than the untuned, the recall jumped from 0.40 to 0.84, and the threshold got much lower. Now, the model does not turn away as many difficult-to-spot buyers as it did before. The f1-score (from 0.53 to 0.65) also improved, which means that the model is mathematically better at balancing precision and recall. The ROC-AUC also saw a small increase of 0.0037.

Comparing with the Logistic Regression model, the tuned Random Forest model was better in all metrics, showing that it has made up for the low recall before tuning while still being close to the untuned precision.

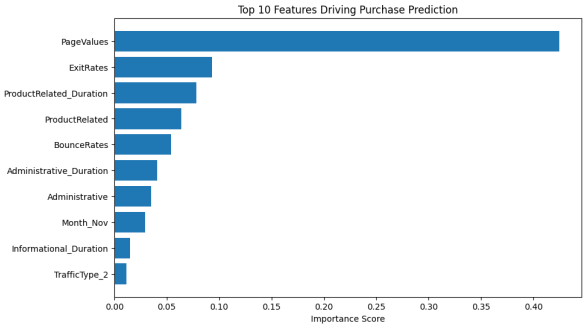

Figure 1: Tuned RF Feature Importance

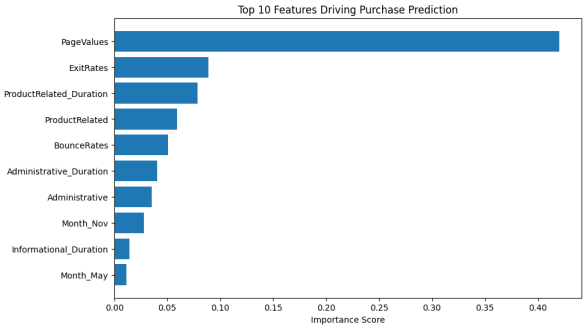

Figure 2: Tuned RF + Clusters Feature Importance

5.3 Impact of Behavioural Segmentation

After adding the clusters to the data and running it through the tuned Random Forest pipeline, precision, recall, f1-score, and ROC-AUC decreased while the threshold and log-loss increased. Table 2 compares the scores from the model before and after adding clusters.

Besides comparing the metrics, we also looked at feature importance in the two models. Here we found that top 9 was the same, with PageValues having a score of 0.42 and the rest below 0.10, while the cluster features ranked 75-77 out of 78 features.

6 Findings

As detailed in the Analysis section, integrating the cluster labels into the Tuned Random Forest resulted in a decrease in performance across all key metrics: Recall dropped from 0.84 to 0.83, F1-score from 0.65 to 0.64, and ROC-AUC from 0.9224 to 0.9209.

It seems that by adding the clusters that are based on the same behavioral metrics (e.g. PageValues, ProductRelated_Duration) that the model already knows, we inject noise instead of helping the model as there was no information gain for Random Forest.

Through our feature importance analysis, we found that the clustering has so little importance that

Table 2: Performance comparison: Tuned Random Forest vs. Tuned Random Forest with Clusters

Metric	Tuned RF	Tuned RF + Clusters
Class 1 (Shoppers) Performance		
Precision	**0.53**	0.52
Recall	**0.84**	0.83
F1-Score	**0.65**	0.64
Overall Model Performance		
Optimal Threshold	0.3961	0.3949
ROC-AUC	**0.9224**	0.9209
Log-Loss	**0.2988**	0.3028

the model is almost ignoring them. In a Random Forest, which selects a random subset of features for each split, the presence of "noise" variables (the clusters) increases the probability of excluding strong predictors like PageValues at critical splits, leading to the observed decline in predictive power.

Based on these findings, we reject the model with clusters in favour of the Tuned Random Forest (without clusters). This model is able to identify most purchasing sessions (recall of 0.84) while still clearly separating purchasing and non-purchasing behaviour (ROC-AUC of 0.92).

7 Conclusion

Our research question was whether unsupervised behavioural segmentation (clustering) would capture latent user patterns and thereby improve the predictive accuracy of the supervised model.

From our findings, we can conclude that clustering based on the same behavioural metrics does not improve the predictive accuracy. Rather, it introduced feature redundancy as it did not find any new relevant information in the clustering (low feature importance).

We therefore select the Tuned Random Forest (without clusters) as the best of these models. This model effectively addresses the business goal of identifying rare purchase events, capturing 84% of potential buyers (Recall). It also achieves this high coverage with significantly better Precision (0.53) than the Logistic Regression baseline (0.43). This means that the model is not simply guessing; it is correctly identifying high-intent users while minimizing the "spam" factor associated with less precise models.

By rejecting the complex clustering step, we deliver a simpler but efficient prediction engine. This allows the e-commerce platform to target the right users with real-time interventions while avoiding wasted marketing resources on non-buyers.

References

[1] Christian Allenbrand. "Clicking through the Clickstream: A Novel Statistical Modeling Approach to Improve Information Usage of Clickstream Data by E-Commerce Entities". In: *Intelligent Information Management* 15.3 (2023). URL: https://www.scirp.org/journal/paperinformation?paperid=125364.

[2] Farhad Huseynov and Selçuk Yıldırım. "Behavioural segmentation analysis of online consumer audience in Turkey by using real e-commerce transaction data". In: *International Journal of Economics and Business Research* 14.1 (2016), pp. 12–28.

[3] William C. McDowell, Rex C. Wilson, and Craig O. Kile. "An Examination of Retail Website Design and Conversion Rate". In: *Journal of Business Research* 69.11 (2016), pp. 4837–4842. DOI: 10.1016/j.jbusres.2016.04.040.

[4] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[5] Robert L Thorndike. "Who belongs in the family?" In: *Psychometrika* 18.4 (1953), pp. 267–276.

[6] UCI Machine Learning Repository. *Online Shoppers Purchasing Intention Dataset*. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset. Accessed: 2025-12-10. 2018.

[7] William J Youden. "Index for rating diagnostic tests". In: *Cancer* 3.1 (1950), pp. 32–35.