MAST6100 Individual project

Predicting Online Shopper Purchase Intention Using Machine Learning

Kaviya Balasubramaniam kb726

Machine learning and deep learning

Module code: MAST6100

Abstract

This report examines the prediction of online purchase intention using the Online Shoppers dataset. The task is formulated as a binary classification problem with a highly imbalanced target variable. A complete machine learning pipeline was implemented in R, including exploratory data analysis, preprocessing, model training, and evaluation. Four models were compared: logistic regression, random forest, support vector machine (SVM), and a neural network. Model performance was assessed using ROC-AUC, PR-AUC, accuracy, F1-score, sensitivity, and specificity. Results indicate that random forest and neural network models achieve the strongest predictive performance, while logistic regression provides an interpretable baseline. The findings highlight the importance of appropriate evaluation metrics when modelling imbalanced e-commerce data.

Introduction

The rapid growth of e-commerce has increased the importance of understanding and predicting customer purchase behaviour. Accurately identifying whether an online browsing session will result in a purchase can support targeted marketing, personalised recommendations, and improved business decision-making. From a machine learning perspective, this problem can be framed as a binary classification task.

This report focuses on predicting online purchase intention using the Online Shoppers dataset, which contains behavioural, temporal, and technical features describing user sessions. A key challenge of this dataset is class imbalance, as only a small proportion of sessions result in a purchase. Consequently, evaluation metrics such as ROC-AUC and precision–recall AUC are prioritised over accuracy.

A complete machine learning pipeline was implemented in R, incorporating exploratory data analysis, preprocessing, and the comparison of four classification models: logistic regression, random forest, support vector machine, and neural network. The models were evaluated on a held-out test set to allow fair and consistent comparison.

Data description

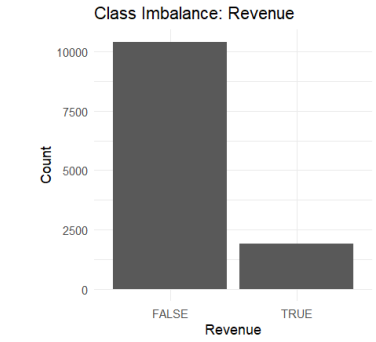
The analysis uses the Online Shoppers Purchasing Intention Dataset Originally published in the UCI Machine Learning Repository , consisting of 12,330 user sessions and 18 variables, including a binary target variable, Revenue, which indicates whether a session resulted in a purchase. The dataset contains behavioural features such as page visit counts and time spent on different page types, engagement metrics including bounce and exit rates, and contextual variables such as month, weekend, and visitor type.

The target variable is highly imbalanced, with approximately 15.5% of sessions resulting in revenue. This imbalance presents challenges for classification and motivates the use of class-sensitive evaluation metrics. An initial inspection confirmed that the dataset contains no missing values, eliminating the need for imputation. Overall, the dataset provides a realistic and suitable basis for modelling online purchase intention in an e-commerce setting.

Exploratory Data Analysis EDA

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the structure and characteristics of the Online Shoppers dataset, identify potential patterns associated with purchase

behaviour, and inform subsequent modelling decisions. Particular attention was paid to class imbalance, feature distributions, and relationships between key predictors and the target variable. The dataset contains 12,330 user sessions with 18 variables, including behavioural, technical, and temporal features, and a binary target variable (Revenue). The dataset contains no missing values.

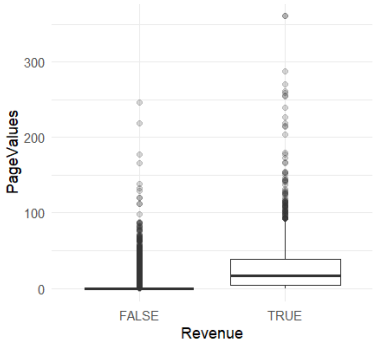


2.1 Target Variable Distribution

The target variable Revenue is imbalanced, with approximately 15.5% of sessions resulting in a purchase and 84.5% not resulting in a purchase. Figure 1 illustrates this imbalance. Why this matters: Class imbalance can bias classifiers toward the majority class. Therefore, performance metrics such as ROC-AUC, PR-AUC, sensitivity, and F1-score are prioritised over accuracy in model evaluation.

Figure 1: Class distribution of the revenue variable.

2.2 Key Numerical Predictors



2.2.1 PageValues PageValues shows a strong separation between purchasing and non-purchasing sessions. Sessions that result in revenue generally exhibit substantially higher PageValues.This suggests that PageValues

Figure 2: Distribution of PageValues by revenue outcome.

2.2.2 Product-Related Duration

Product-related session duration is notably higher for sessions that generate revenue. This indicates that user engagement with product pages is strongly associated with purchase likelihood.

Figure 3: Product-related duration by revenue outcome.

2.3 Temporal Effects

Monthly patterns reveal variation in conversion behaviour throughout the year. Certain months exhibit a higher proportion of purchasing sessions, suggesting seasonal effects in online shopping behaviour. This motivates the inclusion of Month as a categorical predictor in the models.

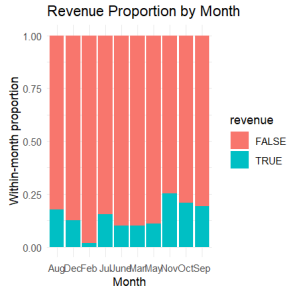


Figure 4: Proportion of revenue-generating sessions by month.

2.4 Summary of EDA Findings

The EDA highlights three key insights:

- The target variable is highly imbalanced, requiring appropriate evaluation metrics.

    - PageValues and product-related duration are strong predictors of revenue.
- Temporal effects suggest seasonal variation in purchasing behaviour.
- These findings directly inform the preprocessing strategy and model selection described in the next section

Methodology

This section describes the methodological framework used to build and evaluate the classification models. A structured machine learning pipeline was implemented to ensure reproducibility, fair model comparison, and robustness to class imbalance. The pipeline consists of data partitioning, preprocessing, model specification, cross-validation, and evaluation on a held-out test set.

Train–Test Split

To assess model generalisation performance, the dataset was divided into training and test sets using an 80/20 split. Stratified sampling was applied based on the target variable Revenue to preserve the original class distribution in both subsets. This resulted in 9,863 observations in the training set and 2,467 observations in the test set, with approximately 15.5% positive (revenue-generating) sessions in each.
The test set was held out and not used during any stage of model training or hyperparameter tuning. This ensures that final performance estimates reflect true out-of-sample predictive ability rather than optimisation on the evaluation data.

Preprocessing Strategy

A unified preprocessing strategy was applied across all models to ensure comparability and prevent data leakage. Preprocessing was implemented using the recipes framework in the tidymodels ecosystem, allowing all transformations to be estimated using only the training data and then consistently applied to both training and test sets.
The preprocessing steps included:

- Removal of zero-variance predictors: Predictors with no variability were removed, as they do not contribute to model learning and may cause numerical issues.
- Dummy encoding of categorical variables: Nominal predictors, including Month, VisitorType, and Weekend, were converted into binary indicator variables using one-hot encoding.
- Normalisation of numeric predictors: All numeric predictors were standardised to have zero mean and unit variance. This step is particularly important for models sensitive to feature scale, such as logistic regression, SVMs, and neural networks.

Using a consistent preprocessing pipeline across all models ensures that differences in performance can be attributed to the modelling approach rather than discrepancies in data preparation.

Handling Class Imbalance

The class imbalance observed in the dataset poses a significant challenge for classification. Rather than applying resampling techniques such as oversampling or undersampling, the imbalance was addressed through careful metric selection and interpretation.

In particular, metrics that account for ranking and class separation, such as ROC-AUC and precision–recall AUC (PR-AUC), were prioritised. These metrics are more informative than accuracy when the positive class is rare. Additionally, sensitivity, specificity, and F1-score were used to provide insight into the trade-off between identifying purchasing sessions and avoiding false positives.

The positive class was explicitly defined as Revenue = TRUE throughout the analysis to ensure consistent interpretation of evaluation metrics.

Cross-Validation

Cross-validation was employed on the training set to estimate model performance and tune hyperparameters while mitigating overfitting. Different cross-validation strategies were used depending on model complexity and computational cost:

- Logistic regression: 5-fold stratified cross-validation
- Random forest: 3-fold stratified cross-validation
- Support vector machine: 2-fold stratified cross-validation
- Neural network: 3-fold stratified cross-validation

Fewer folds were used for more computationally intensive models to balance performance estimation with practical runtime constraints, particularly given hardware memory limitations. All folds were stratified to maintain the class distribution of the target variable within each fold.

Model Evaluation Framework

After cross-validation and hyperparameter selection, each model was refit on the full training set using the optimal configuration. Final performance was then evaluated on the held-out test set.

Two types of predictions were generated for evaluation:

- Class probabilities, used to compute ROC-AUC and PR-AUC
- Class labels, used to compute accuracy, F1-score, sensitivity, and specificity

Confusion matrices were also produced to provide an interpretable summary of classification outcomes. This evaluation framework allows for a comprehensive comparison of models in terms of both discrimination and practical classification behaviour

Model Specification

To address the online purchase intention classification problem, four different classification models were implemented and compared. These models were selected to represent a range of methodological approaches, from classical statistical models to more flexible machine learning techniques. Using

multiple models allows for a balanced comparison between interpretability, predictive performance, and robustness under class imbalance.

Logistic Regression (Baseline Model)

Logistic regression was used as a baseline model due to its simplicity, interpretability, and widespread use in binary classification problems. The model estimates the probability of a session resulting in revenue by modelling the log-odds of the binary outcome as a linear combination of the predictor variables.

Despite its linear nature, logistic regression often performs competitively on structured tabular data and provides valuable insight into feature importance through model coefficients. In this analysis, logistic regression was fitted using a generalised linear model (GLM) with a binomial link function.

No hyperparameter tuning was required for this model. However, careful preprocessing, particularly normalisation and dummy encoding, was applied to ensure numerical stability and interpretability. Logistic regression serves as a strong reference point against which the performance of more complex models can be evaluated.

Random Forest

Random forest is an ensemble learning method that constructs a collection of decision trees and aggregates their predictions to improve predictive accuracy and reduce overfitting. Each tree is trained on a bootstrap sample of the data, and at each split, a random subset of predictors is considered. This randomness introduces diversity among the trees and enhances generalisation.

Random forest models are particularly well-suited to capturing non-linear relationships and interactions between variables, which are likely present in online browsing behaviour. Additionally, they are relatively robust to multicollinearity and outliers.

In this study, two key hyperparameters were tuned:

- mtry: the number of predictors considered at each split
- min_n: the minimum number of observations required in a terminal node

Hyperparameter tuning was performed using a small, space-filling grid to balance performance and computational efficiency. The final model was trained with an increased number of trees to stabilise predictions. Variable importance measures were extracted to identify the most influential predictors.

Support Vector Machine (SVM)

Support vector machines are margin-based classifiers that aim to find a decision boundary that maximises the separation between classes. In this analysis, an SVM with a radial basis function (RBF) kernel was used, allowing the model to capture non-linear decision boundaries.

The RBF kernel maps the input features into a higher-dimensional space, enabling complex separation between purchasing and non-purchasing sessions. However, SVMs are known to be sensitive to class imbalance and hyperparameter choices.

Two hyperparameters were tuned:
- Cost (C): controls the trade-off between margin width and classification errors
- Kernel width (σ): determines the smoothness of the decision boundary

Due to computational constraints, a small grid and a reduced number of cross-validation folds were used. The SVM results highlight the challenges of applying margin-based classifiers to imbalanced datasets, particularly when default classification thresholds are used.

Neural Network (Multilayer Perceptron)

A feed-forward neural network, implemented as a multilayer perceptron (MLP), was included to model complex non-linear relationships between predictors and the target variable. Neural networks are flexible function approximators capable of learning interactions that may not be explicitly captured by traditional models.

The neural network used in this study consists of a single hidden layer. Two key hyperparameters were tuned:
- Number of hidden units: controls the model's capacity
- Regularisation penalty (weight decay): reduces overfitting by penalising large weights

The model was trained for a fixed number of epochs using normalised inputs. Due to memory constraints, a small hyperparameter grid and limited cross-validation were employed. Despite these limitations, the neural network demonstrated strong discriminative performance on the test set.

Summary of Model Choices

The selected models represent complementary approaches to the classification task:
- Logistic regression provides interpretability and a strong baseline
- Random forest captures non-linear relationships and feature interactions
- SVM offers a margin-based perspective but is sensitive to imbalance
- Neural networks provide flexible, non-linear modelling capacity

Together, these models enable a comprehensive evaluation of how different classification techniques perform on an imbalanced, real-world e-commerce dataset.

Results and Model Comparison

This section presents and compares the performance of the four classification models on the held-out test set. The models are evaluated using multiple metrics to account for class imbalance and to capture different aspects of predictive performance. In particular, emphasis is placed on ROC-AUC and PR-AUC,

alongside accuracy, F1-score, sensitivity, and specificity. Confusion matrices and ROC curves are used to support the quantitative results.

Overall Performance Comparison

In my code we have a comparison table showing the overson performance comparisons
Across all models, random forest achieved the strongest overall performance, with the highest ROC-AUC (0.939) and PR-AUC (0.791). This indicates excellent discriminative ability and strong performance in identifying the minority class (revenue-generating sessions).

The neural network also performed well, achieving a ROC-AUC of 0.930 and PR-AUC of 0.725. While slightly weaker than the random forest, the neural network demonstrated a good balance between sensitivity and specificity, suggesting effective modelling of non-linear relationships.

Logistic regression, despite its simplicity, achieved a competitive ROC-AUC of 0.897. However, its PR-AUC and F1-score were lower than those of the more flexible models, indicating limitations in capturing complex interactions in the data.

The support vector machine (SVM) achieved a high ROC-AUC of 0.900 but exhibited unstable classification behaviour, particularly with respect to sensitivity and F1-score.

Sensitivity, Specificity, and Class-Level Behaviour

Sensitivity (true positive rate) and specificity (true negative rate) provide insight into how each model handles the trade-off between identifying purchasing sessions and avoiding false positives.

- The random forest achieved a sensitivity of 0.634 and specificity of 0.965, indicating a strong ability to correctly identify non-purchasing sessions while still capturing a substantial proportion of purchasing sessions.
- The neural network demonstrated a similar balance, with sensitivity of 0.665 and specificity of 0.951.
- Logistic regression showed high specificity (0.983) but relatively low sensitivity (0.374), meaning it was conservative in predicting purchases and missed many revenue-generating sessions.
- The SVM failed to predict any positive class instances on the test set, resulting in zero sensitivity and an undefined F1-score. This highlights the model's sensitivity to class imbalance and default decision thresholds.

These findings demonstrate that accuracy alone would be misleading in this context, as models with high accuracy may still perform poorly in identifying the minority class.

Confusion Matrix Analysis

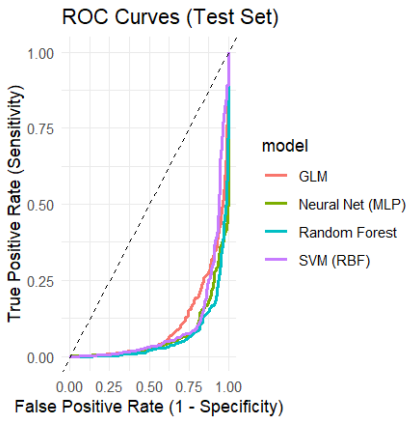
```
--- RF Confusion matrix ---
> print(rf_eval_kb726$confusion)
          Truth
Prediction FALSE TRUE
     FALSE  2013  140
     TRUE     72  242
> cat("\n--- RF Variable Importance Plot ---\n")
```

Figure 6: Confusion matrix for the random forest model on the test set

Confusion matrices further illustrate the differences between models. The random forest and neural network models correctly classified a larger number

of purchasing sessions compared to logistic regression, while maintaining relatively low false positive rates.
In contrast, the SVM classified all test observations as non-purchasing sessions. While this resulted in perfect specificity, it rendered the model ineffective for the practical goal of identifying purchase intention. This behaviour underscores the importance of model calibration and threshold adjustment when working with imbalanced datasets.

ROC Curve Comparison



ROC curves for all models are shown in Figure X. The curves confirm the numerical results, with the random forest and neural network dominating across most threshold values. Logistic regression follows closely but exhibits slightly weaker separation, while the SVM curve reflects its limited practical usefulness despite a high ROC-AUC value.

Figure 7: ROC curves for all models evaluated on the test set

The ROC analysis reinforces the conclusion that ensemble and neural approaches provide superior discriminative performance for this dataset.

Summary of Results

Overall, the results indicate that:

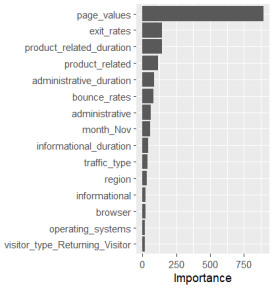
•Random forest is the strongest performer, offering robust discrimination and balanced classification behaviour.
•Neural networks provide competitive performance and effectively capture non-linear patterns.
•Logistic regression remains a reliable and interpretable baseline but underperforms in recall of the minority class.
•SVMs are highly sensitive to class imbalance and may require additional calibration or resampling strategies to be effective.

These findings motivate a deeper discussion of model trade-offs and practical implications, which is presented in the next section.

Discussion

This section discusses the results of the classification models in relation to the characteristics of the dataset, the exploratory data analysis, and the methodological choices made throughout the study. Particular emphasis is placed on understanding why certain models performed better than others and the implications of class imbalance for practical deployment.

Interpretation of Model Performance ~ The results demonstrate that models capable of capturing non-linear relationships and feature interactions, specifically random forest and neural networks, outperform linear models in predicting online purchase intention. This finding aligns with expectations given the complexity of user browsing behaviour, which is unlikely to follow strictly linear patterns.



The strong performance of the random forest model can be attributed to its ensemble structure, which allows it to model interactions between behavioural features such as PageValues and ProductRelatedDuration. These predictors were identified during EDA as strong indicators of purchase behaviour, and the random forest's ability to exploit such interactions contributes to its superior ROC-AUC and PR-AUC scores. Above is figure 5. **Figure 5: Variable importance from the random forest model**.

The neural network also demonstrated strong performance, suggesting that non-linear decision boundaries are beneficial for this task. However, its slightly weaker performance compared to the random forest may reflect the relatively small network architecture and limited hyperparameter tuning imposed by computational constraints.

Logistic Regression as a Baseline

Logistic regression provided a useful and interpretable baseline model. The estimated coefficients align well with the EDA findings, particularly the strong positive influence of PageValues and product-related engagement. However, the model's relatively low sensitivity indicates that a linear decision boundary is insufficient to capture the nuanced patterns required to identify purchasing sessions reliably.

Despite this limitation, the model's high specificity suggests that it could still be useful in scenarios where false positives are particularly costly, such as when promotions or incentives are expensive to deploy.

Challenges with SVM under Class Imbalance

The SVM model highlights the challenges of applying margin-based classifiers to imbalanced datasets. Although the SVM achieved a high ROC-AUC, it failed to predict any positive instances on the test set using the default classification threshold. This behaviour demonstrates that high ranking performance does not necessarily translate into effective classification when decision thresholds are poorly calibrated.

This finding underscores the importance of threshold adjustment, class weighting, or resampling techniques when deploying SVMs in imbalanced settings. Without such adjustments, SVMs may be unsuitable for operational use in purchase prediction tasks.

Role of Evaluation Metrics : The divergence between accuracy and class-sensitive metrics reinforces the importance of appropriate metric selection. Models such as logistic regression and SVM achieved relatively high accuracy but performed poorly in terms of sensitivity and F1-score. In contrast, random forest and neural networks achieved better balance between identifying purchasing sessions and avoiding false positives. This highlights that accuracy alone would be misleading for this problem and supports the use of ROC-AUC and PR-AUC as primary evaluation criteria in imbalanced classification contexts.

Practical Implications : From a practical perspective, the results suggest that ensemble and neural models are better suited for predicting online purchase intention in real-world e-commerce environments. Random forest, in particular, offers a strong combination of performance, robustness,

and interpretability through variable importance measures. However, practical deployment would require additional considerations, including threshold optimisation, cost-sensitive learning, and real-time inference constraints. Furthermore, interpretability remains an important consideration when models are used to inform business decisions.

Limitations and Future Work : Several limitations should be acknowledged. First, the analysis relied on a single dataset, limiting generalisability. Second, computational constraints restricted the size of hyperparameter grids and the complexity of the neural network architecture. Third, no explicit techniques such as resampling or class weighting were applied to address class imbalance. Future work could explore alternative imbalance handling strategies, more extensive hyperparameter tuning, and additional models such as gradient boosting. Moreover, incorporating cost-sensitive evaluation metrics could further align model performance with business objectives.

Conclusion

This study investigated the prediction of online purchase intention using the Online Shoppers dataset, framing the problem as a binary classification task under conditions of class imbalance. A complete machine learning pipeline was implemented in R, incorporating exploratory data analysis, preprocessing, model training, and evaluation across four classification models: logistic regression, random forest, support vector machine, and a neural network.

The results demonstrate that model choice plays a critical role in handling complex, imbalanced e-commerce data. Among the models considered, the random forest achieved the strongest overall performance, exhibiting the highest ROC-AUC and PR-AUC scores and a balanced trade-off between sensitivity and specificity. This indicates its effectiveness in identifying revenue-generating sessions while maintaining a low false positive rate. The neural network also performed well, capturing non-linear patterns in user behaviour and offering competitive predictive performance. Logistic regression provided a valuable and interpretable baseline, with coefficient estimates that aligned closely with insights from the exploratory data analysis. However, its limited sensitivity highlights the constraints of linear models in capturing complex behavioural interactions. The support vector machine, despite achieving high ranking performance, proved highly sensitive to class imbalance and default threshold selection, ultimately limiting its practical usefulness without further calibration. Overall, the findings suggest that ensemble and neural approaches are better suited for predicting online purchase intention in real-world settings, particularly when the cost of misclassification is asymmetric. From a practical perspective, the random forest model emerges as the most suitable candidate for deployment, offering a strong balance between predictive performance, robustness, and interpretability. This study highlights the importance of appropriate evaluation metrics, careful preprocessing, and model comparison when addressing imbalanced classification problems. Future work could extend this analysis by incorporating cost-sensitive learning, alternative imbalance handling techniques, and more advanced ensemble methods to further enhance predictive performance and business relevance.