Predicting Online Shoppers Purchasing Intention

Michael Rayan

Department of Engineering, University College Roosevelt

[ENGDATA202 Machine Learning]

[Professor Marijn Jansen]

[May 10 2024]

Introduction

This paper will delve into the domain of online shopping analytics, focusing on a real-time shopper behaviour analysis system that calculates the likelihood of whether a shopper will make a purchase during a given session. The paper aims to provide a comprehensive exploration of binary classification in the context of online shopping analytics, with a focus on predicting user shopping session abandonment. By addressing this fundamental challenge, we aim to contribute to the advancement of data-driven strategies that optimise user engagement and enhance e-commerce performance

Understanding user behaviour and intentions is crucial in the realm of e-commerce, as it directly impacts conversion rates and user experience. Of particular importance is the identification of users who arrive at an e-commerce site with a clear intent to make a purchase. These users exhibit distinct behavioural cues that signify a readiness to complete a transaction without extensive browsing or hesitation (Ding et al., 2015). By harnessing the power of machine learning algorithms, we can uncover patterns in user behaviour that serve as indicators of purchasing intent.

This report will primarily explore the development of a predictive model aimed at identifying users with a direct purchasing intention upon their arrival at an e-commerce platform. Additionally, the model will be designed to provide personalised content to users who are at risk of abandoning their transactions before completion. By proactively addressing potential abandonment, our objective is to enhance conversion rates and elevate overall user satisfaction.

This paper will be focusing on predicting whether a user is likely to abandon their shopping session. This constitutes a binary classification problem, where our target variable is a binary categorical variable comprising the boolean values True and False. Here, True denotes that the user completes a purchase, while False indicates that the user does not.

Throughout this paper, we will delve into various aspects of binary classification, including data preprocessing, feature engineering, model selection, and evaluation metrics. Furthermore, we will discuss the practical implications of our predictive model within the e-commerce domain, emphasising its potential to drive actionable insights and improve business outcomes.

Dataset

In this section, the dataset will be described, and the approach to feature selection and preprocessing will be outlined. The dataset comprises feature vectors corresponding to 12,330 sessions, each uniquely attributed to a different user over a one-year period. This deliberate design choice aims to mitigate any biases stemming from specific campaigns, special days, user profiles, or temporal periods.

Of the 12,330 sessions in the dataset, 84.5% (10,422) are negative class samples, signifying sessions that did not culminate in a purchase, while the remaining 15.5% (1,908) are positive class samples denoting sessions that ended with a successful purchase. Notably, the dataset is heavily imbalanced, necessitating careful consideration when employing evaluation metrics. Given the smaller number of positive class targets, a robust true positive rate is essential.

Furthermore, it is advantageous for sellers to accurately predict instances where customers are likely to make a purchase, as opposed to instances where customers may abandon the session prematurely.

To evaluate the performance of the models, three key metrics will be utilised: Accuracy, Precision, and F1 score. These metrics offer insights into the overall effectiveness, the ability to identify positive class samples correctly, and the balance between precision and recall, respectively. Notably, the models will be ranked based on precision, prioritising the precision of positive class predictions.

The dataset encompasses a total of 10 numerical variables and 7 categorical variables. Among the numerical features, 'Administrative', 'Administrative Duration', 'Informational', 'Informational Duration', 'Product Related', and 'Product Related Duration' denote the number of different types of pages visited by the visitor in each session and the total time spent in each page category. These values are derived from URL information and are updated in real-time as users navigate through the website.

Additionally, features such as 'Bounce Rates', 'Exit Rates', and 'Page Values' are obtained from Google Analytics. 'Bounce Rate' quantifies the percentage of visitors who enter a webpage and subsequently exit without triggering any further interactions during the session. 'Exit Rate' represents the proportion of pageviews on a specific webpage that serve as the final interaction in the session. Lastly, 'Page Value' denotes the average value of web pages visited before a successful e-commerce transaction is completed.

The categorical variables present a challenge, particularly due to the common requirement of employing one-hot encoding. However, applying this technique to all categorical variables significantly inflates the number of predictors and often results in inaccurate models with high variance. While categorical variables are typically more interpretable, it's observed that many of them exhibit minimal correlation with the response variable.

Hence, among the variables "OperatingSystems", "Browser", "Region", "TrafficType", "VisitorType", "Weekend", and "Month", only the "VisitorType" variable will undergo one-hot encoding. This variable comprises three categories: 'New Visitor', 'Returning Visitor', and 'Other'. This selection was made through a trial and error process on the data, where the removal of most categorical variables did not noticeably impact model performance.

In the subsequent sections, the process of feature selection and preprocessing will be delved into, elucidating the methodologies for optimising model performance and mitigating the effects of imbalanced data. Through rigorous analysis and strategic feature engineering, the aim is to develop a robust predictive model capable of accurately discerning user behaviour and predicting purchase outcomes in the dynamic landscape of e-commerce.

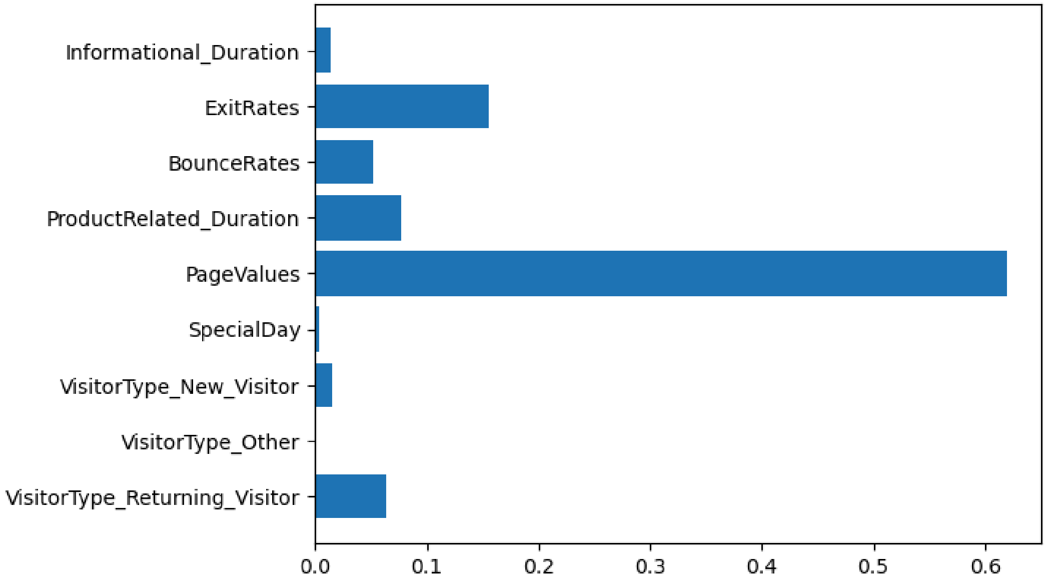
Methods

Feature Selection

As discussed previously, the dataset comprises 17 predictors, including 7 categorical variables with various types of categorization. Feature selection is deemed necessary to ensure the accuracy and precision of the models. This process is approached through two methods: a tree-based method and the mRMR algorithm used in relevant literature such as in an article by Sakar and colleagues (2018) .

In the tree-based method, a Random Forest algorithm is employed to rank the feature importances, thereby identifying the predictors utilised for the initial major splits. Subsequently, these important features are incorporated into the models. However, a limitation of this approach is its failure to inherently perform subset selection, as the precise number of features must be specified. Nevertheless, this method serves as an intuitive means to comprehend which predictors the model predominantly focuses on, owing to the highly interpretable nature of tree-based algorithms.

Figure 1. ***Barplot of feature importances.***



Conversely, in the mRMR algorithm, the objective is to maximise the relevance between the selected set of features and the class variable while simultaneously avoiding redundancy among the selected features.The algorithm does this by finding an optimal set of features that is mutually and maximally dissimilar and can represent the response variable effectively. The algorithm quantifies the redundancy and relevance using the mutual information of variables—pairwise mutual information of features and mutual information of a feature and the response. This aims to achieve maximum classification accuracy with a minimal subset of features. The algorithm generates a list of features, the length of which is specified by the user, arranged in descending order of relevance, from the most to the least relevant predictor.

In the findings concerning model performance, it becomes evident that mRMR demonstrates exceptional effectiveness, yielding models with the highest accuracy and precision. Upon specifying the features "Informational_Duration", "ExitRates", "BounceRates", "ProductRelated_Duration", "PageValues", "SpecialDay", and "VisitorType", and indicating a subset of length 5, the resulting output is as follows: 'PageValues', 'ProductRelated_Duration', 'ExitRates', 'VisitorType_Returning_Visitor', and 'SpecialDay'. Notably, the "VisitorType" variable is a component of the one-hot encoding scheme, indicating whether the visitor is returning or not. Specifically, a value of 0 denotes 'New' or 'Other', while a value of 1 signifies 'Returning'.

Sampling of Data

On the original imbalanced data, model performance is notably poor. To address this issue, a new dataset was created by combining all positive observations with a randomly selected subset of negative observations. This facilitated a deeper understanding of the data and improved model performance.

Subsequently, the SMOTE sampling method was employed on the original dataset. SMOTE is specifically designed to tackle imbalanced datasets by generating synthetic samples for the minority class. By addressing bias and capturing crucial features of the minority class, SMOTE enhances prediction accuracy and overall model performance.

This was implemented through a pipeline that involved oversampling the positive class using SMOTE and randomly undersampling the negative classes to a specified extent. Given the substantial number of negative classes, undersampling was necessary to balance the dataset. The data was then processed through this pipeline, resulting in a resampled dataset conducive to improved model performance.

Models

The data was initially split into training and testing sets to ensure consistency throughout the analysis. Cross-validation was employed during hyperparameter optimization. The models utilised included K-Nearest Neighbors (KNN), Logistic Regression, Bagged Trees, Boosted Trees, and a Support Vector Machine (SVM) with a radial kernel.

The KNN algorithm, chosen for binary classification, is non-parametric and offers interpretability. However, in high-dimensional settings, it can be impacted by nuisance features. Feature selection was performed prior to mitigating this issue. The training data was scaled, followed by Principal Components Analysis (PCA) with a maximum of three components. Twenty-fold cross-validation was conducted to determine the optimal number of neighbours (K) ranging from one to twenty-five, with the F1 score used for scoring.

The Bagged Trees method employed the RandomForests algorithm as its base estimator, with the maximum number of features set to the total number of predictors. Attention was paid to ensure that the increase in bias caused by bootstrapping did not overshadow the decrease in variance. The RandomForests Classifier was chosen for its ability to balance bias and variance effectively.

For Boosted Trees Classification, 10,000 estimators with a learning rate of 0.0005 were used to achieve optimal predictions. The maximum depth of the trees was set to three to maintain manageable tree sizes that compound in boosting. Gradient Boosting was preferred over AdaBoost due to its robustness and lesser sensitivity to outliers, achieved through weight updates based on gradients.

A Support Vector Machine classifier with a radial kernel was employed, with the cost parameter tuned using fifteen-fold cross-validation. The range for the cost parameter spanned from one to ten, and the F1 score served as the scoring metric during cross-validation.

Results

On checking performance of our models on the metrics of F1 score,Precision score and Accuracy it is seen that the Boosted Trees classifier is the best performing model. This makes sense as they are some of the best methods used in many different problems. In second is the Bagging Trees classifier. Though Logistic Regression may have a higher precision than the Bagged Trees it's ranking is harmed and brought to question by it's very low F1 score

Table 1. *Evaluation of models*

Model	**Precision**	**F1 Score**	**Accuracy**
GB Trees	0.8483	0.8161	0.8742
Logistic Regression	0.8259	0.5969	0.8267
Bagged Trees	0.7965	0.8161	0.8774
KNN classifier	0.7326	0.7576	0.8458
SVM - Radial	0.4744	0.7879	0.7891

Note that both the boosted trees and bagging trees have a similar f1 score but it is seen that the Boosted trees have a better true positive rate as they focus on correcting the errors made by the previous trees. Displaying the Confusion matrices for both models helps further analyse both top models' performance.

Figure 2. *Confusion matrix for the performance of the Gradient Boosted Trees on the test set*

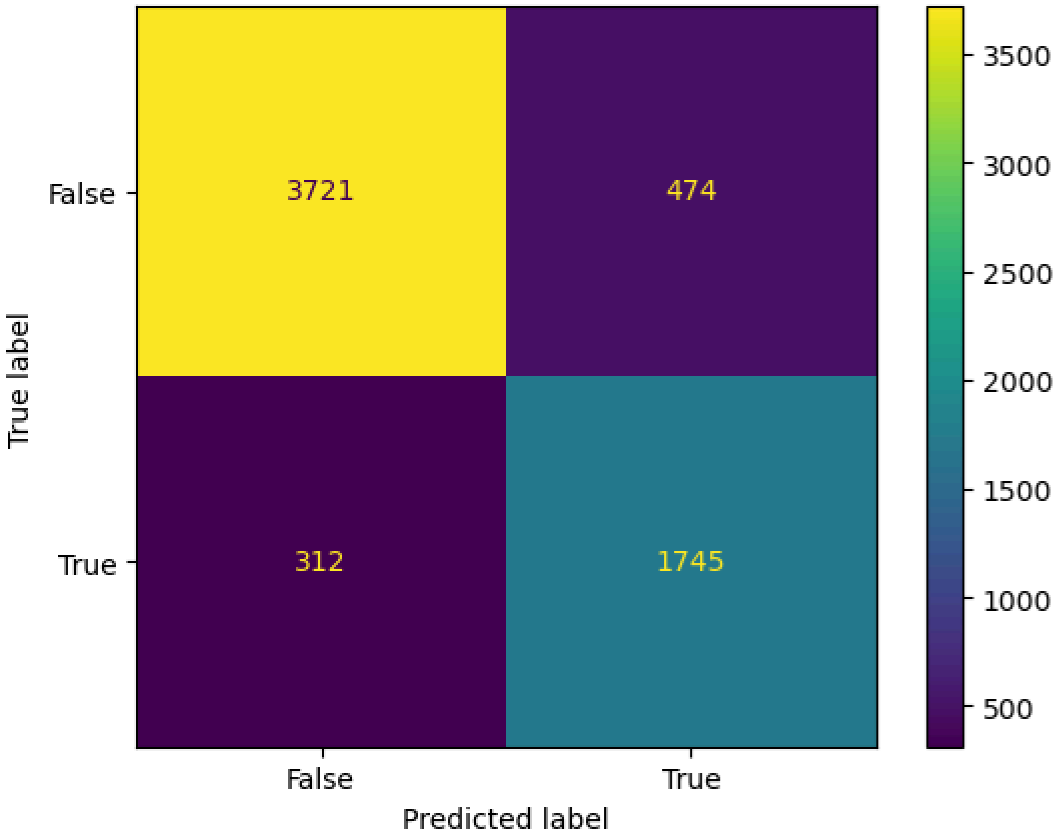
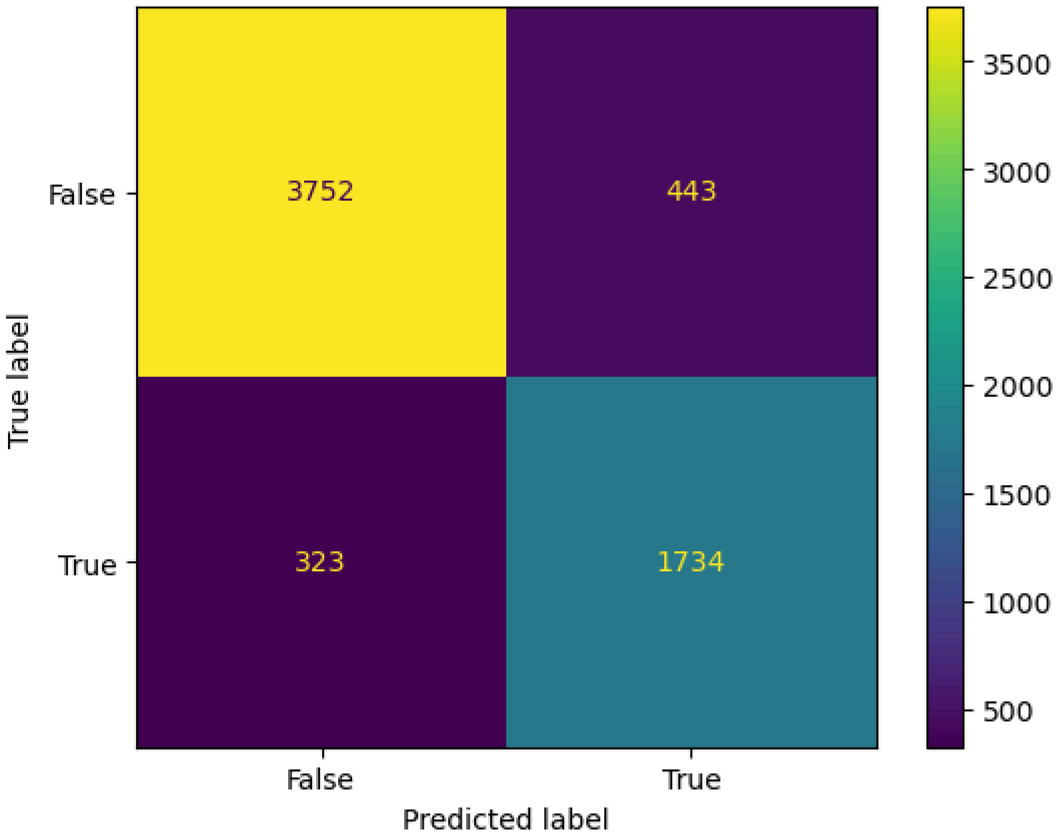


Figure 3. *Confusion matrix for the performance of the Bagged Trees on the test set*



The K Nearest Neighbours model performs well for how simple it is and is able to hold its own in terms of accuracy. Since it does not use any algorithm such as boosting and bagging it loses out in the metrics of F1 and Precision. Though this model is capable and highly interpretable it cannot overcome the problem of imbalance that persists through the target variable.

The SVM classifier with a radial Kernel seems to have performed poorly with a very low precision score and relatively low F1 and accuracy too. Alternative Kernels did not fare any better.

Discussion and Conclusion

This paper demonstrates that the problem of binary classification on an imbalanced dataset can be solved using Tree Based methods, particularly the Gradient Boosted Trees classifier and the Bagging Trees Classifier with the Random Forest Classifier as the estimator. Notably, while Random Forests alone yielded subpar F1 scores and lower precision compared to Boosted and Bagged Trees, the simple KNN algorithm exhibited promising performance, albeit hindered by the imbalanced nature of the data. Similarly, the parametric method of Logistic Regression showed commendable precision but was significantly impacted by the data's imbalance, warranting caution in its application.

A critical aspect of this process was feature selection, as all models exhibited notably poorer performance when utilising alternative subsets of predictors. The adoption of the mRMR process for feature selection led to a substantial increase in model performance. Interestingly, the most influential predictor, "PageValues," provided by Google Analytics, reflects the average value of a web page visited by a user before completing an e-commerce transaction. Despite the anticipation that predictors such as "Product Related" and "Product Related Production" would exhibit the highest correlation with the output, they instead demonstrate a strong correlation with the PageValue variable, which is prioritised first by the algorithm.

A key limitation of the project is that due to the sole use of machine learning models, the choice of categorical variables in the predictors is very limited.

Moving forward, the next steps in addressing this problem involve exploring the utility of Deep Learning models, such as neural networks, on this dataset. These models hold potential for leveraging more categorical variables, as they can handle the increased dimensionality resulting from one-hot encoding.

To conclude, the purchasing intention of a user can be predicted with the use of ensemble tree based methods.

Moreover, the ability to predict whether a user will end their session presents opportunities for the implementation of integrated systems that deliver tailored content to individual customers, thereby enhancing user retention and, ideally, facilitating purchase decisions.

References

Sakar, C. O., Polat, S., Katırcıoğlu, M. A., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing & Applications*, *31*(10), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

Ding, A. W., Li, S., & Chatterjee, P. (2015). Learning user Real-Time intent for optimal dynamic Web page transformation. *Information Systems Research*, *26*(2), 339–359. https://doi.org/10.1287/isre.2015.0568