A REPORT

ON

PREDICTING ONLINE SHOPPERS PURCHASING INTENTIONS

BY

NIKHIL DHANARAJ	2021A7PS0427H
PRITAM REDDY NANDIKONDA	2021A7PS1502H
SRIJA GUDA	2022A7PS2003H

CS F415

DATA MINING

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

Apr 2024

Contents

Abstract

This research project aims to predict online shoppers' purchasing intentions using machine learning techniques. The primary objective of this project is to find newer state-of-the-art classification models which will be able to classify sessions into two categories: those ending with a shopping transaction (positive class) and those that do not (negative class). This task presents a significant challenge due to the inherent imbalance within the data. Typically, the number of sessions that don't result in a purchase (negative class) vastly outnumbers (84.5%) those that do (positive class), creating a skewed dataset. We address this imbalance through techniques like upsampling to ensure a more balanced training process. We compared seven classification models, including classical models such as Naive Bayes, Decision Forests, Support Vector Machines and Logistic Regression and newer models such as Deep Neural Decision Forest, K Nearest Neighbours and XGBoost algorithms.Through our experiments, we learned that XGBoost, with 150 boosting rounds, and Decision Forest, with 20 trees, outperformed the other models for our task.

Keywords: Online shopper intentions, Classification, Comparing Models.

1. Introduction

The main goal of this project is to discover state-of-the-art classification models capable of effectively categorising sessions into two groups: those concluding with a shopping transaction (considered as the positive class) and those that do not (considered as the negative class). The aim is to identify advanced models that outperform conventional classification techniques.

Forecasting the likelihood of online purchases is crucial for driving revenue and optimising marketing efforts. Precise classification empowers businesses to customise website experiences and marketing strategies, enhancing conversion rates. Evaluating a website's impact on revenue aids in strategic decisions, like enhancing profitability or discontinuing underperforming sites.

Online shopper behaviour is difficult to predict for a variety of reasons. First, real-world e-commerce data is imbalanced, with far more sessions not leading to purchases. This can trick traditional models into favouring the majority class and neglecting the crucial minority (purchases). Second, online shopper behaviour is influenced by complex, non-linear interactions between various factors. Simple models might struggle to capture these nuances, leading to inaccurate predictions. Finally, real-world features often depend on each other. For instance, visiting the shopping cart may be correlated with the amount of time spent on product pages. These relationships can be overlooked by naive models that assume feature independence, which would reduce their performance.

Predicting online shopper behaviour is a tedious challenge.The field of predicting online shopper behaviour is constantly evolving as new techniques and algorithms emerge. While advancements are made, there is no perfect solution for every problem. Consumer preferences and online shopping habits are constantly changing, requiring adaptable models. Recognizing these complexities, our research focuses on evaluating various classification models to identify those that perform best with our specific dataset and address the inherent class imbalance.

The key component of our approach is to compare classical models and state-of-the-art models on our dataset to classify online shoppers with a purchase (positive class) and shoppers without a purchase (negative class). By using advanced techniques, these newer models offer improved accuracy and efficiency in distinguishing between the classes. Through rigorous testing and evaluation, our results showcase the superior performance of these modern approaches compared to traditional classification models using a range of metrics including accuracy, recall, F1 score and precision. This result not only highlights the evolution of classification models but also highlights the potential for enhanced decision-making and targeted strategies - tailor website experiences for our dataset and it classifies online shoppers' intentions of purchase.

2. Related Work

In the paper Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data by Md Rayhan Kabir, Faisal Bin Ashraf, and Rasif Ajwad published in 2019, they have analysed different classification algorithm such as Decision Tree, Random Forest, Naive Bayes, SVM to predict whether a customer, visiting the webpages of an online shop, will end up with a purchase or not. The results from the paper show that Random Forest is most suited to predict the customer's purchase intention. Moreover, if we choose to do gradient boosting using this algorithm, it can predict with the highest accuracy, which is 90.34%. The current gradient boosting approach achieves acceptable accuracy on this diverse yet sparse dataset. However, there's room for improvement. Exploring XGBoost, a more powerful algorithm, could potentially lead to better performance. Additionally, deep learning techniques might be suitable for uncovering complex patterns in the data. Finally, utilising a larger dataset could further enhance the model's ability to generalise and potentially increase accuracy.

In the paper Predicting Online Purchasing Intention by Akash Deoraj (2021), the algorithms that were used are Random Forest, Extra Trees , Artificial Neural Network and Logistics Regression for which the accuracies are 89.6%, 89.1%, 88.64% and 87.1% respectively, after taking an average of the accuracy before feature selection, after feature selection and after hyperparameter optimization. The research paper primarily focused on comparing the accuracy, macro average and weighted average of the various algorithms before and after feature selection. The paper concluded with achieving an accuracy of 89.9% using the Random Forest model. This paper can be highly improved upon by testing the dataset on more models- rather than the classical models, using more state-of-the-art models to improve the accuracy.

3. Approach/Methodology

The core objective of this project is to identify state-of-the-art classification models that can accurately sort sessions into two categories: those that end with a shopping transaction (classified as the positive class) and those that do not (classified as the negative class). The goal is to find newer models that surpass traditional classification methods in performance and effectiveness.

The dataset that we are using is from UC Irvine Machine Learning repository. There are a total of 12,330 sessions. In these sessions, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest 15.5 % (1908) were positive class samples ending with shopping. There are 17 attributes in the dataset we are using of which 8 are categorical and 10 are numerical. In order to avoid biases of special day, user profile or period, the dataset was formed such that each session would belong to a user in a period of 1 year.

The dataset includes various features related to user behaviour on an e-commerce website. These features cover the types of pages visited, the time spent on each type of page, and metrics such as bounce rate, exit rate, and page value provided by Google Analytics. Additionally, it includes information about special days, operating system, browser, region, traffic type, visitor type, weekend visits, and the month of the year.

3.1 Classification Algorithms:

Random Forest: Random Forest is a bagging method where you build multiple decision trees and combine their predictions through a voting mechanism to make a final decision. This approach makes it more reliable than a single decision tree because errors in individual trees tend to cancel out, resulting in a more accurate outcome.

Naive Bayes: The naive Bayes algorithm is based on Bayes' theorem. Here the assumption is that the features are independent from each other. In this case, the prior probability of features is constant. It determines the posterior probability of each class using the likelihood of the sample's features considering the given class and the prior probability of each class.

Logistic Regression: It estimates the probability of occurrence of the dependent variable for given independent variables. It uses a logistic function to predict binary outcomes by employing a linear combination of input variables. Thus, it shows an S-shaped curve to map any real values between 0 and 1. The coefficients of this combination are used to represent the maximum likelihood estimation.

SVM: The approximation of the support vector machine algorithm is to find a line separating the data in two different classes. At first, data are projected in a high dimensional feature space and the data are mapped by the kernel. The kernel can be linear or nonlinear. Then the algorithm creates a hyperplane to separate data from one class to another. For each test case, the data descriptors are mapped to the same feature spaces and predict the class of the data by using the hyperplane.

K-Nearest Neighbors (KNN) : It operates on the principle of proximity, where a data point is classified based on the class most common among its k nearest neighbours in the feature space. KNN does not involve explicit training; instead, it memorises the entire training dataset and uses it for making predictions during inference.

Deep Neural Decision Forests (DNDF): This is a hybrid model that combines the strengths of deep neural networks (DNNs) and decision forests. DNDF consists of multiple neural networks, with each network individually making a decision. The model learns both the feature representations and decision boundaries simultaneously, allowing it to effectively handle high-dimensional and non-linear data.

XGBoost: This operates by sequentially building an ensemble of decision trees, correcting errors at each iteration to optimise overall performance. With built-in regularisation techniques, XGBoost effectively controls overfitting. Its cross-validation support ensures robust model tuning. XGBoost's ability to provide insights into feature importance makes it a valuable tool for classification.

4. Experiments

4.1 Dataset:

Initially, we checked whether there were duplicate values and dropped those values. This led to the number of total sessions changing from 12,330 to 12,205. After this, we applied label encoding to the attributes ‘Month’ (which consisted of the months of the year) and ‘VisitorType’ (which consisted of New_Visitor and Returning_Visitor). Since the number of positive samples were 84.5% (10,422) and negative samples were 15.5 % (1908), we applied up sampling. This led to the number of true samples and false samples to be 8218 each. We also showed the correlation matrix for all the attributes to which ones are dependent on one another. Following this, we applied normalisation to some of the features in our dataset with large maximum values. These features included Administrative_Duration, Informational_Duration, ProductRelated_Duration, PageValues, Administrative, Informational, Region, TrafficType and ProductRelated.

4.2 Evaluation Methods:

To evaluate the proposed methodology for predicting online shoppers' intentions, we employ several evaluation metrics to assess the performance of the classification models. The following methods are utilised:

Accuracy: Accuracy measures the proportion of correctly classified instances out of the total instances. It provides an overall assessment of the model's predictive performance but may not be suitable for imbalanced datasets.

Precision: Precision calculates the proportion of true positive predictions among all positive predictions made by the model. It indicates the model's ability to avoid false positives.

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances. It assesses the model's ability to capture all positive instances.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is particularly useful when dealing with imbalanced datasets.

These evaluation metrics collectively provide insights into the classification model's performance in distinguishing between sessions that result in a shopping transaction and those that do not. By considering multiple metrics, we gain a comprehensive understanding of the models' strengths and weaknesses in predicting online shoppers' intentions.

In the models we compared, we used binary cross entropy loss function or mean squared error loss function for training.

4.3 Experimental Setup:

In our experimental setup, we carefully selected hyperparameters and configurations for each classification model to ensure fair comparisons and optimal performance. Here are the key aspects of our experimental setup:

Hyperparameters Tuning: For models such as Random Forest, XGBoost, and Deep Neural Decision Forest, we tuned hyperparameters such as the number of trees, boosting rounds, and epochs, respectively. This involved conducting grid searches or random searches over a predefined range of values to identify the best combination of hyperparameters.

Feature Engineering: Prior to model training, we performed feature engineering to preprocess and transform the input data. This included tasks such as handling missing values, encoding categorical variables, and scaling numerical features to ensure consistency and compatibility across models.

Overall, our experimental setup was designed to evaluate the performance of various classification models and identify the most effective approaches for predicting online shoppers' intentions. By carefully controlling experimental variables we aimed to ensure the validity and reliability of our findings.

5. Results and Discussion

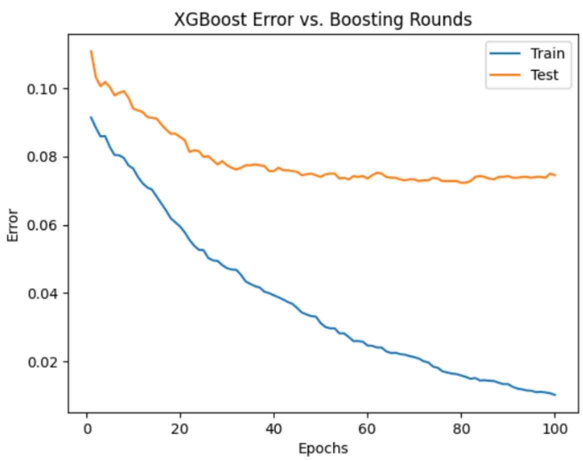
Here are the classification performance measures averaged on 10 different random states for the 7 classification models that we tested:

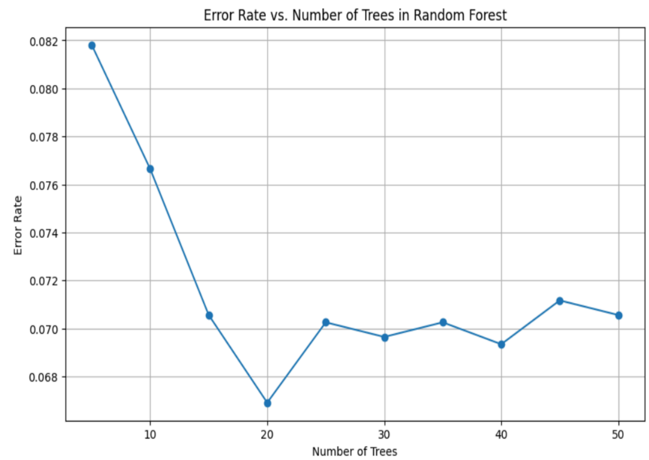
Model	Accuracy	Precision	Recall	F1 Score
Random Forest	88.14	88.69	88.14	88.09
Naive Bayes	71.53	74.85	71.53	70.56
Vanilla Logistic Regression	84.21	84.47	84.22	84.18
Support Vector Machines	84.45	85.02	84.45	84.38
K Nearest Neighbours	84.52	84.52	84.52	84.52
Deep Neural Decision Forest	87.77	89.12	85.72	87.39
XGBoost	92.84	92.86	92.84	92.85

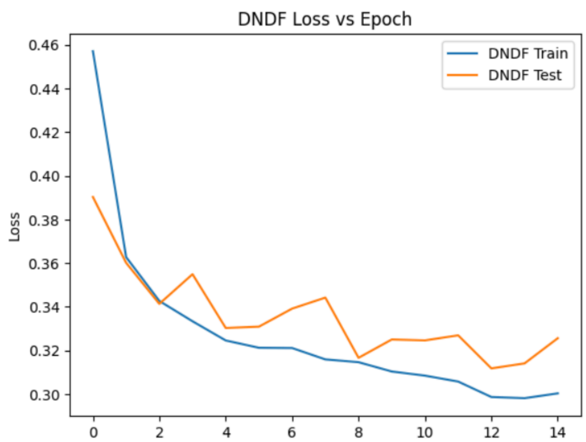
As we can see from the above table the worst performing model is the Naive Bayes model. This is probably because of the faulty assumption of the model that all its features are independent when it is not as seen in the correlation matrix. The three best performing models were

XGBoost, Random Forest and Deep Neural Decision Forest.

Below are the graphs of the 3 best models against some of their key hyperparameters:







The above graphs show that XGBoost performs better as the number of boosting rounds increases. Random Forest performs better for 20 trees than the rest. This may be due to overfitting. We also notice that the Deep Neural Decision Forest performs better when the number of epochs increases until it reaches 13 epochs. Below, we compare the three models after tuning them for best performance.

Model	Accuracy
XGBoost with 100 boosting rounds	93.50
Random Forest with 20 trees	93.31

Deep Neural Decision Forest trained on 13 epochs	88.32

From the above, we can infer that XGBoost and Random Forest classification models are the best models for predicting online shoppers' intentions.

We then tested our 2 best models on a Titanic Dataset which is a small dataset that classifies whether a person who bought tickets for a ride on the Titanic survived or not.Our results from the experimentation are shown below

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	81.11	89.36	65.62	75.67
Random Forest	79.02	78.85	79.02	78.79

The poorer results on this new dataset can be attributed to the fact that our new dataset contains less than 600 data points. Even with these lesser amounts of data we were able to obtain a fairly decent classification measure with our XGBoost and Random Forest models.

6. Conclusion

In this project, we used the Online Shoppers Intention dataset to find the best classification model that can be used to predict the intentions of online shoppers. We compared seven classification models, including classical models such as Naive Bayes, Decision Forests, Support Vector Machines and Logistic Regression and newer models such as Deep Neural Decision Forest, K Nearest Neighbours and XGBoost algorithms.

Through our experiments, we learned that XGBoost, with 150 boosting rounds, and Decision Forest, with 20 trees, outperformed the other models for our task. We can attribute this to the fact that XGBoost can handle non-linear relationships between features very well and that our Random Forest model with 20 trees can leverage the diversity of its trees and generalise the predictions.

Though the other models performed relatively well, our Naive Bayes model performed poorly on our task. This is primarily due to the incorrect assumption of the Naive Bayes model that our features are independent of each other, but our dataset, as it is a real-world dataset, exhibits a degree of correlation between the features.

We also experimented with our models on another dataset, the Titanic dataset, where we predicted whether a person who bought a ticket to the Titanic survived. This is a small dataset, and our two best models still gave us decent classification metrics on these models. This experiment further verifies that our models are comparable to current SOTA models.

From the above, we can conclude that ensemble methods and tree-based algorithms are best suited for predicting online shoppers' intentions.

7. References

[1] M. R. Kabir, F. B. Ashraf and R. Ajwad, "Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data," *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038521.

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785