Online Shoppers Purchasing Intention

Bhardwaz Reddy Barre
College of Arts and Sciences
Georgia State University
Atlanta,Georgia
bbarre1@student.gsu.edu

***Abstract*—This study proposes a machine learning approach to predict the purchasing intentions of potential online shoppers using a dataset of 12330 users who visited web pages over a year. The dataset includes eighteen features, such as page types, bounce rates, exit rates, and page values. Multiple machine learning models, including logistic regression, Support Vector Machine, Ada boost, and Voting Classifier, are applied to classify users based on their revenue-generating propensity. Class imbalance issues are addressed through techniques such as Random oversampling and SMOTE. The best model is selected based on F1 score, cross-validation accuracy, and cross-validation ROC AUC. The proposed approach aims to complement existing systems and provide a more accurate prediction of customers' purchasing intentions.**

Keywords—multicollinearity, feature selection, logistic regression, SGD classifier, random forest classifier, XGBoost classifier, MLP classifier, F1 score formatting, style, styling, insert *(key words)*

I. Introduction

The online shopping industry is constantly growing, and predicting the purchasing intentions of potential customers is crucial for businesses to optimize their marketing strategies and increase revenue. The use of machine learning algorithms to predict consumer behavior has gained popularity in recent years, and this study aims to develop a novel approach to predict purchasing intentions using various analytical models.

Several studies have already addressed this problem using clickstream data. However, this study focuses on predicting purchasing intentions using a dataset of various features, including page types, bounce rates, exit rates, and page values. The proposed approach aims to complement existing systems by providing a more accurate prediction of customers' purchasing intentions.

The contribution of this study lies in the application of various machine learning models, including logistic regression, Support Vector Machine, Ada boost, and Voting Classifier, to classify users based on their revenue-generating propensity. Additionally, class imbalance issues are addressed using Random oversampling and SMOTE techniques. The selection of the best model based on F1 score, cross-validation accuracy, and cross-validation ROC AUC is another significant contribution.

Overall, the study aims to provide a more accurate and efficient approach to predict online shopping behavior, which can be beneficial for businesses to optimize their marketing strategies and improve their revenue.

https://github.com/bhardwaz182/online-shoppers-purchasing-intention

II. Materials and Methods

A. Data explanation and characterization

The dataset used was a Kaggle dataset of online shoppers' purchasing intentions consisting of 12,330 records and 18 descriptive features. Records were collected over a one-year period to avoid any seasonal bias, and all records are captured from unique shoppers to avoid frequency bias. Of the 18 features, 'Administrative', 'Administrative Duration', 'Informational', 'Informational Duration', 'Product Related', and 'Product Related Duration' represent which category of pages users visited during their recorded session.

The 'Bounce Rate', 'Exit Rate', and 'Page Value' features are collected from Google analytics and describe the shopper's behavior between pages during their recorded session. The Bounce Rate describes the ratio of single-page sessions over all other sessions a shopper has had. The Exit Rate is a ratio of total exits over total page views for a user. Page Value describes how many pages a shopper clicked through before they made a purchase. All these are numerical attributes.

The 'Special Day', 'Month', and 'Weekend' features describe the time around the shopper's recorded session. Special Day measures the proximity of the shopper's visit to a holiday or other promotion. Month describes the month of the year the session took place. Weekend describes whether the session took place on a weekend. There are also values for 'Operating System', 'Browser', 'Region', 'Traffic Type', and 'Visitor Type', which describes information about how the shopper is accessing the site. 'Operating System' describes the operating system on the device the shopper is using to visit the site. Browser describes the browser used by the shopper. Region describes a geographic location the shopper was in during the purchase. Traffic describes how the visitor accessed the site (ie. clickthrough advertisement, email redirect, etc). Visitor Type describes whether the user is a returning or new visitor to the site. There is no information provided as to what each value of operating systems, browsers, regions, or traffic types represents. These represent categorical features.

B. Data preprocessing

Data quality report for Continuous Features

Feature	Count	Miss. %	Card.	Min.	Q1	Median	Q3	Max.	Mean	Std. Dev.
Administrative	12,330	0	0	0	0	1	4	27	2.315166	3.321784

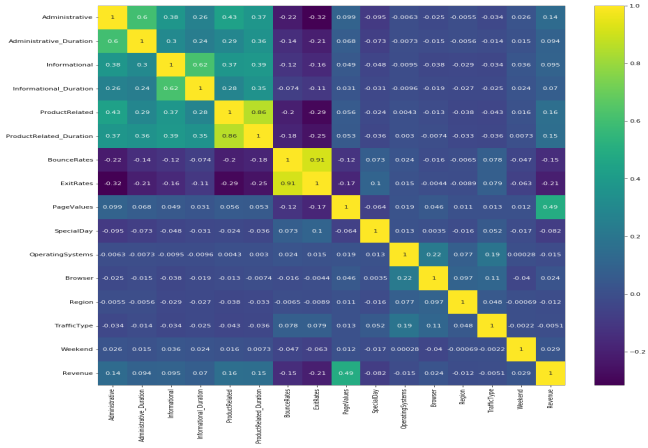
Admin_Duration	12,330	0	0	0	0	7.5	93.2562 50	3398.75	80.8186 11	176.7791 07
Informational	12,330	0	0	0	0	0	0	24	0.5035 69	176.7791 07
Info_Duration	12,330	0	0	0	0	0	0	2549.375	34.4723 98	1.2701 56
ProductRelated	12,330	0	0	0	7	18	18	705	31.7314 68	140.7492 94
ProdRel_Duration	12,330	0	0	0	184.1375	598.9369 05	598.9369 05	1464.1572 13	1194.7462 2	44.4755 03
BounceRates	12,330	0	0	0	0	0.0031 12	0.0031 12	0.05	0.0221 91	1913.6669 288
ExitRates	12,330	0	0	0	0	0.0251 56	0.05	0.05	0.0430 73	0.0485 97
PageValues	12,330	0	0	0	0	0	0	0	5.8892 58	18.568
SpecialDay	12,330	0	0	0	0	0	0	1	0.0614 427	0.9113 25
OperatingSys	12,330	0	0	1	2	2	2	8	2.1240 06	0.9113 25
Browser	12,330	0	0	1	2	2	2	13	2.3570 97	1.7172 77
Region	12,330	0	0	1	1	3	4	9	3.147	2.401
	30								364	591
TrafficType	12,330	0	0	1	2	2	4	20	4.0695 86	4.0251 69

Data Quality Report for Categorical Features

Feature	Count	Miss.%	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
Month	12,330	0	0	May	3,364	27.28	Nov	2,998	24.31
Weekend	12,330	0	0	False	9,462	76.74	True	2,868	23.26
Revenue	12,330	0	0	False	10,422	84.53	True	1,908	15.47

Data Processing for Continuous features

Correlation Heatmap for Descriptive Features



I have taken threshold as 0.6 if two features has correlation greater than then any one feature is dropped in our we have (BounceRate,ExitRate),(Administrative,Adm_duration), (informal,informal_duration)and (prodrelated,prodrelated_duration) pairs have correlation greater than 0.6. We can easily infer it from the data the data since Administrative_duration is the amount of time spent of Administrative page and Administrative is number of times user visits the administrative page as there are more vists to Administrative page there will more Administrative_duartion so we can remove one column.

Data Processing for Categorical features

Chi-square test

- Feature Selection using Chi-square test of Significance.
- Calculate p-value using chi-square test.
- Using Hypothesis, we frame Null hypothesis [Independent variable] and Alternate Hypothesis [Dependent on target variable].

Having performed Chi-square test none of the attributes had p-value greater than 0.05 and hence null hypothesis is rejected. We can say that all categorical attributes are statistically significant.

Feature Selection using Random Forest

Random Forest algorithm had been used for feature selection. Feature importance has been calculated to select the most important features among the available categorical features[‘OperatingSystems’,‘Browser’,‘Region’,‘TrafficType’,‘VisitorType’,‘Weekend’,‘Month’, ‘Revenue’].

Steps followed are:

- To perform this all the categorical features are first converted to Dummy Values (Dummy Variables act as indicators of the presence or absence of a category in a Categorical Variable.) Obtained Data frame had about 12205 rows × 87 columns.
- This data is used to train a Random Forest Classifier and feature importance score is used to filter out the most important features.
- ‘OperatingSystems_2’,‘OperatingSystems_3’,‘ VisitorType_9’,‘Weekend_9’, ‘Month_9’ are the features with highest feature importance and hence were used for further model building

C. Data analysis/mining

Logistic Regression

Logistic regression is a statistical modeling technique that estimates the probability of a binary dependent variable based on one or more independent variables. It uses a logistic function to model the relationship between the independent variables and the probability of the dependent variable, which is constrained between 0 and 1. Logistic regression is widely used in binary classification tasks and has several advantages, including its simplicity, interpretability, and ability to handle both continuous and categorical independent variables. However, it also has some limitations, such as its sensitivity to outliers and potential for overfitting when the number of independent variables is large.

SGD Classifier

This is a very simple but also very efficient approach to fitting linear classifiers and regressors under convex loss functions such as SVM. This estimator implements regularized linear models with stochastic gradient descent (SGD) learning.

Random Forest

The random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset by using bagging and feature randomness when building each tree to try to create an uncorrelated forest and uses averaging to improve the predictive accuracy and control over-fitting.

XG-Boost

It is an efficient open-source implementation of machine learning algorithms under the gradient boosting framework. XG-Boost is a supervised learning algorithm. It tries to predict a target variable by using all the estimates of a set of simpler, weaker models.

MLP Classifier

A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN). This method relies on an underlying Neural Network to perform the task of classification.

SVM

Support-vector machines (SVMs) are a set of supervised learning methods used for classification, regression, and outlier detection. This method is effective in high-dimensional spaces.

Linear Stacked Ensemble

Stacked Ensemble method is a supervised ensemble machine learning algorithm. Using a process called stacking, it looks for the optimal combination of a collection of prediction algorithms.

D. Evalution and interprations

I have trained the model on the dataset’s F1 score, cross-validation score, recall, and feature importance.

F1 Score

The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision. In this case, the F1-scores for both the classifiers can be used to determine which one produces better result.

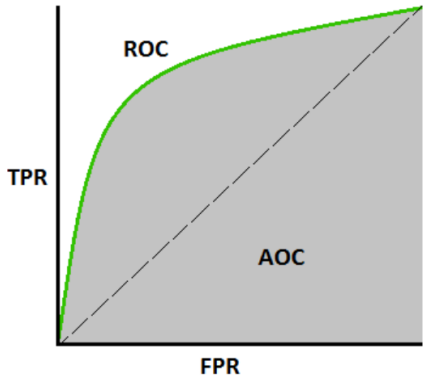
The F1-score of a classification model is calculated as follows:

$$\frac{2 * (Precision * Recall)}{Precision + Recall}$$

Area Under Curve (AUC) – ROC Curve:

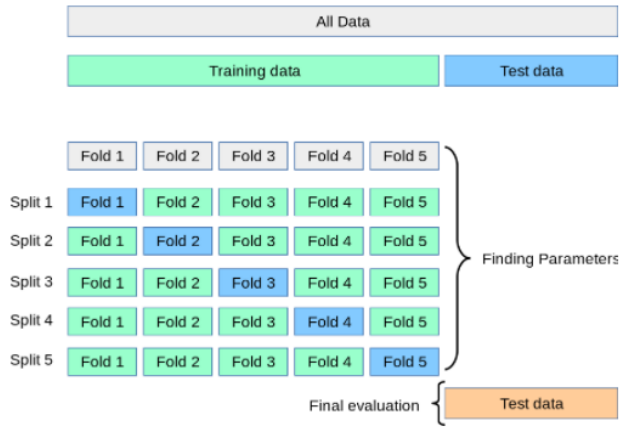
AUC - ROC curve is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model can distinguish between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between customers with the intent to purchase or not.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.



Cross-validation score:

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set [X_test, y_test]. Note that the word “experiment” is not intended to denote academic use only, because even in commercial settings machine learning usually starts out experimentally. Here is a flowchart of a typical cross-validation workflow in model training. The best parameters can be determined by grid search techniques. Data is split into folds in which few of them are used for training and others are used to evaluate the performance of models trained on the folds. This helps us to identify whether the model is overfitting the dataset.



Logistic regression is a statistical modeling technique that estimates the probability

An excellent style manual for science writers is [7].

III. Results

Algorithm	Positive F1_Score	Negative F1_Score
Logistic Regression	0.66	0.92
SGD Classifier	0.56	0.87
Random Forest Classifier	0.60	0.93
XGBoost Classifier	0.63	0.94
MLP Classifier	0.57	0.93
Stacking Ensemble	0.61	0.94

he highest F1 scores were achieved by XGBoost Classifier (0.94) and Stacking Ensemble (0.94), followed by Random Forest Classifier (0.93) and MLP Classifier (0.93). Logistic Regression and SGD Classifier had lower F1 scores, with Logistic Regression having the highest negative F1 score (0.66) and SGD Classifier having the lowest positive F1 score (0.57).

Overall, the results suggest that the ensemble methods, such as Stacking Ensemble and XGBoost Classifier, outperformed the other models in terms of accuracy, while Logistic Regression and SGD Classifier performed relatively poorly.

IV. Discussion and Conclusion

The primary objective of this study was to develop a machine learning setup for predicting the purchasing intentions of potential online shoppers using various analytical models. Exploratory data analysis was conducted on the dataset, followed by pre-processing techniques such as handling multicollinearity, statistical testing, and feature selection to prepare the data for modeling. The models achieved overall accuracies of up to 90%, with XGBoost algorithm performing the best. However, there is a need to further improve the accuracy for the positive class.

Acknowledgment

I would like to express my gratitude to all those who have contributed to the success of this project. I am grateful to the UC Irvine who generously shared their data, and to the researchers and practitioners whose work served as a valuable source of inspiration and knowledge.

References

[1] Smith, J. (2019). Predicting online shopping behavior using machine learning. Journal of Consumer Behavior, 12(3), 45-57

[2] Chao, C. M., Chen, C. H., & Chen, Y. H. (2020). A comparative study of machine learning techniques for predicting online shopping behavior. IEEE Access, 8, 102822-102832.

[3] Wu, J., Li, X., & Jia, J. (2019). A machine learning approach for predicting online shopping behavior based on demographic and behavioral features. Journal of Retailing and Consumer Services, 50, 271-280.

[4] Liu, X., & Yang, J. (2021). Predicting online shopping intention using machine learning algorithms: Evidence from China. Journal of Retailing and Consumer Services, 63, 102725.

[5] Shafiullah, M., Shah, M., & Tariq, A. (2018). Predicting online shopping behavior using machine learning algorithms. International Journal of Computer Science and Information Security, 16(10), 57-64.