Predicting Whether an Online Session Will Result in a Purchase

Isaac U Umukoro
Postgraduate Diploma in Science in Data Analytics
(PGDDA)
National College of Ireland
(NCIRL)
Dublin, Ireland
x23215640@student.ncirl.ie

***Abstract*— Forecasting the purchasing intentions of online shoppers is a crucial undertaking in the field of e-commerce, to improve customer satisfaction and boost sales conversion rates. This study makes use of the Online Shoppers Purchasing Intention Dataset to develop a model that can predict whether an online shopping session would result in a purchase. To do that, we utilize, different machine learning and deep learning models, including logistic regression, decision trees, K-nearest Neighbors, and others. The application of the Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance dataset. We then apply Hyperparameter tuning and cross-validation to the two best-performing models to optimise their performance, ensure robustness, and avoid overfitting. According to the results, the most effective approach is the Extra Trees model (Extremely Randomized Trees), an Ensemble Learning method based on decision trees that select the best split among a random subset of features to attain better accuracy and robustness in projecting whether an online shopping session will produce a purchase. This model shows promise for practical use in e-commerce systems and beats others.**

Keywords—Machine learning, oversampling, hyperparameter tuning, ensemble method, cross-validation.

I. Introduction

In recent times, purchase intentions have become more important in e-commerce because they affect how people behave when they are shopping. The experiment for this study looks at how people shop online. The dataset used in this study comes from the UCI Machine Learning Repository, more specifically the Online Shoppers Purchasing Intention dataset. The dataset Features encompass the duration of user engagement with various pages, the types of incoming traffic, and additional attributes that provide insights into the probability of whether a user's online shopping session would result in a purchase. There are various machine-learning techniques applied to the provided dataset, which include, Logistic Regression, Decision Trees, K-Nearest Neighbors, Naive Bayes, Linear Discriminant Analysis, Support Vector Machine, Random Forest, Gradient Boosting Machine, XGBoost, LightGBM, AdaBoost, CatBoost, Extra Trees, feed-forward neural network, Hybrid model (Feed-forward neural network combined with ensemble methods & Meta Learner), and Ensemble methods [Random Forest, Gradient Boosting, XG Boost] Stacking model.

Although each of these models possesses distinct advantages, the objective of this analysis was to determine the most effective model for forecasting online shopping purchases. Consequently, the top-performing models from the previous stage, specifically Extra Trees and the Combination of Ensemble Methods, underwent optimization using hyperparameter tuning. The Extra Trees was determined to be the superior model, particularly when cross-validation was used to assess its performance.

II. Literature Review

In a preliminary review of the relevant literature, several papers were identified that contribute to predicting whether an online session would result in a purchase using the online shoppers purchasing intention dataset. These papers provide valuable insights into the methodologies, algorithms, and techniques employed in this research. Below is a brief review of the selected papers:

- Sakar et al. (2019) carried out a study on real-time prediction of online shoppers' purchasing intentions using multilayer perceptron (MLP) and Long Short-Term Memory (LSTM) recurrent neural networks. Their work demonstrated the effectiveness of deep learning techniques in capturing the temporal dynamics of user behaviour on e-commerce platforms [1].
- Mobasher et al. (2002) researched two different clustering techniques based on transactions of users and pageviews to find useful aggregate profiles that recommendation systems can use to achieve effective personalization at the early stages of user visits in an online store [2].
- Moe (2003) work was focused on differentiating between online shoppers. behaviour such as buying, searching, or browsing. This was done by analyzing in-store navigational clickstream data. This work is vital for predicting purchasing intentions to help in understanding the varying behaviours exhibited by users during their online sessions [3].
- Suchacka, Skolimowska-Kulig, and Potempa (2015) examined the grouping of e-customer sessions using support vector machines (SVM) and k-nearest neighbours (KNN) techniques. Their research showed that these traditional machine-learning methods can predict buying intentions and efficiently classify user sessions [4].
- Suchacka and Chodak (2017) used association rules in another study to evaluate purchase likelihood on internet retailers. Their research showed how rule-based approaches might offer insightful analysis of the elements affecting purchasing decisions [5].
- Budnikas (2015) explored computerised recommendations for e-transaction finalization through machine learning. His studies underlined the need to use machine learning techniques to generate recommendations in real time and increase e-commerce conversion rates [6].

- Clifton (2012) laid a basis for understanding and evaluating web traffic data to project user behaviour and purchase intentions emphasizing advanced web metrics with Google Analytics [7].
- Peng et al. (2005) and Kotsiantis et al. (2007) reviewed many feature selection and supervised machine learning methods respectively. consumers [8] [9].
- Baati et al. (2017, 2019) developed new classifiers and modified Naïve Bayes classifiers for numerical and categorical data. Their work Demonstrates the flexibility of these classifiers in many spheres, including e-commerce [10].
- Langley and Sage (1994) addressed the induction of selective Bayesian classifiers. Their work is relevant in the framework of probabilistic prediction models for e-commerce [11].
- Both Quinlan (2014) and Breiman (2001) offered a fundamental understanding of random forests and decision trees, respectively, which are extensively applied in online purchase intention prediction [12] [13].
- Tian et al. (2011) and Ding et al. (2015) studies were based on the use of support vector machine ensembles and real-time intent learning for optimal dynamic web transformation, respectively [14] [15].
- Chawla et al. (2002) were responsible for the introduction of the synthetic minority over-sampling technique (SMOTE), which is often used to address class imbalance issues in the dataset, and this includes the Online Shoppers Purchasing Intention dataset [16].

III. Dataset Description

The online shopper's intention dataset is loaded in Jupyter Notebook and then explorative data analysis is carried out on the dataset to help us understand the structure of the data. The dataset contains 18 features We then checked for missing data and found none. Furthermore, we carried out univariate and Bivariate analyses of the dataset. Lastly, a correlation heatmap is computed to help identify the features that are highly correlated in the dataset.

IV. Data Processing

There are several methods involved in the processing of the dataset, This includes the following:

A. Handling Duplicate Data

This step involves the identification and handling of duplicate rows in the dataset to avoid redundant records that can cause bias in the analysis. 125 duplicate rows were removed from the dataset.

B. Detecting and Handling of Outliers

Outliers were detected in the dataset using the interquartile range and then Capping and flooring were applied to handle the outliers. This method helps to reduce the influence of extreme values. Additionally, we applied log transformation to normalised skewness in the dataset.

C. Normalisation and Encoding of Categorical Features

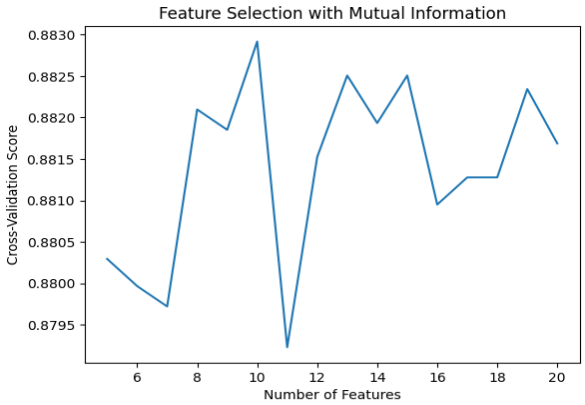
In this step, continuous numerical features are normalized using Standardscaler. One-hot encoding was applied to encode categorical features such that they fit a numerical form for model development. This stage also entails separating the data into target variables (y) and features (X).

D. Feature Selection

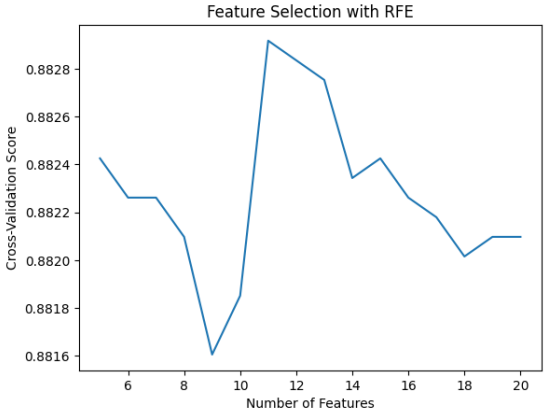
Three different feature selections were applied to the dataset. They include Mutual Information, Recursive Feature Elimination, and SelectKBest [8].

- Mutual Information: This feature selection identifies 16 optimal features which include: num PageValues, num ExitRates, num ProductRelated_Duration, num ProductRelated, num BounceRates, num Administrative Duration, num Administrative, cat Month Nov, cat TrafficType 7, cat Region 4, cat VisitorType_Returning Visitor, num Informational Duration, cat Browser 11, cat Traffic 9, num informational and cat Region 6

.

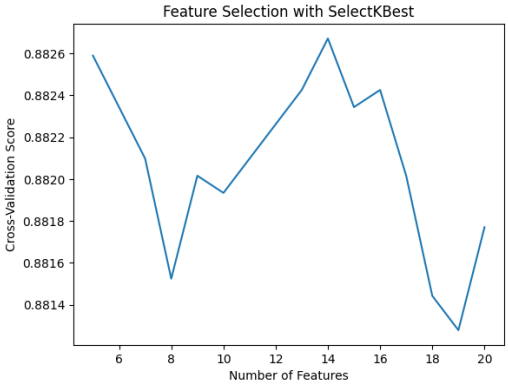


- Recursive Feature Elimination: This feature selection identifies 11 optimal features which include: num exit rates, num page values, cat month Dec, cat Month Feb, cat Month March, cat month May, cat month Nov, cat browser 12, cat browser 3, cat traffic type 13, cat traffic type 8.



- SelectBest: This feature selection identifies 14 optimal features which include: num administrative duration, num product-related duration, num bounce rates, num_exit rates, num_ page values, num_special day, num_ administrative, num_informational, num _product related, cat_month_May, cat_month _Nov,

cat_traffic type_2, cat_traffict ype, cat visitor type_returning_visitor.



We then combined the top features from the three feature selections above to create robust feature sets for our predictive analytics. Due to the imbalanced nature of our dataset (with labels distributed 84.5 % and 15.5 %), the Synthetic Minority Over-sampling Technique (SMOTE) is applied to address the class imbalance [16]. Lastly, the data is split into training and testing sets to prepare it for model training.

E. Applicable Techniques

In this research, various machine-learning techniques were applied to the dataset to build models and then select the best-performing model. These include:

- Logistic Regression: This is a statistical model that uses a logistic function to model a binary dependent variable. It uses different input features to predict whether an online shopping session would result in a purchase or not [9].
- Decision Trees: These are non-parametric supervised learning methods used for classification. They can predict the value of a target variable by learning simple decision rules inferred from the data features [12].
- K-Nearest Neighbors (KNN): This is an instance-based learning algorithm. It classifies a sample based on the majority class among its k-nearest neighbours in the feature space [17].
- Naive Bayes: They are probabilistic classifiers based on Bayes' theorem. They assume independence between every pair of features.
- Linear Discriminant Analysis (LDA): LDA searches for the linear combination of features that best separates two or more classes of events.
- Support Vector Machine (SVM): This is a supervised learning model that makes use of classification algorithms for two-group classification problems [4].
- Random Forest: This is an ensemble method that involves the construction of multiple decision trees and outputs the mode of the classes of the individual trees [13].
- Gradient Boosting Machine (GBM): GBM is an ensemble method that builds a model in a stage-wise fashion and generalizes it by allowing optimization of an arbitrary differentiable loss function [18].
- XGBoost: This is an optimized extreme gradient boosting. It is highly efficient, flexible, and portable [19].
- LightGBM: This is a gradient-boosting framework that uses tree-based learning algorithms.
- AdaBoost: It is also known as Adaptive Boosting.it is a method of ensemble learning that combines several weak classifiers to form a powerful classifier.
- CatBoost: This is an open-source library that excels in performance and is specifically designed for gradient boosting on decision trees. It is specifically designed to handle categorical features without the need for preprocessing [20].
- Extra Trees: Extremely Randomized Trees, or extra trees is an ensemble learning technique that creates many decision trees and produces the class that is the mode of the classes of the individual trees [21].
- Hybrid Model (FNN + Ensemble Learning + Meta-Learning): This method integrates a Feedforward Neural Network (FNN) with several ensemble learning models using a meta-model (Logistic Regression) to integrate the forecasts from these models. This method uses the advantages of several models to raise predictive performance [1].
- Ensemble Methods (Random Forest, Gradient Boosting, and XGBoost) using Stacking: Stacking is a method of ensemble learning whereby a meta-model is trained to aggregatively merge the predictions of several base models. Using Random Forest, Gradient Boosting, and XGBoost as base models, their predictions were aggregated using a meta-model to increase the accuracy of predicting online shoppers' purchasing intentions [13] [18] [19].

F. Results and Model Evaluation

It is important to evaluate the predictive models to determine their generalizability and effectiveness to unseen data. We utilized different performance metrics for the implementation of the various machine-learning techniques mentioned above. These include:

- Accuracy: it is simply the ratio of correctly classified observations out of the total observations.
- Precision and Recall: Precision, also known as positive predictive value, is the proportion of accurately predicted positive observations to the total number of predicted positive observations. High precision is characterized by a low false positive rate. Recall, also referred to as sensitivity, is the proportion of accurately predicted positive observations to the total number of actual positive observations.
- F1 Score: The F1 Score is a metric that calculates the weighted average of Precision and Recall. Thus, this score considers both incorrect positive and incorrect negative results. It demonstrates a classifier's robust performance is an effective method.
- ROC AUC Score: The receiver operating characteristics (ROC) curve measures the model's discriminatory power in distinguishing between

different classes. The area under this curve quantifies the model's ability to accurately classify instances.

TABLE 1

SUMMARY OF THE PERFORMANCES OF VARIOUS MODELS

Model	Accuracy	Precision	Recall	F1 Score	Roc AUC Score
Logistic Regression	0.82	0.85	0.80	0.82	0.82
Decision Trees	0.89	0.89	0.90	0.89	0.89
K-Nearest Neighbors	0.88	0.83	0.97	0.89	0.88
Naïve Bayes	0.68	0.62	0.92	0.74	0.67
LDA	0.78	0.78	0.78	0.78	0.78
SVM	0.86	0.88	0.85	0.86	0.86
Random Forest	0.93	0.92	0.95	0.93	0.93
Gradient Boosting Machine	0.90	0.90	0.90	0.90	0.90
XGBoost	0.93	0.93	0.94	0.93	0.93
LightGBM	0.93	0.92	0.93	0.93	0.93
AdaBoost	0.88	0.89	0.88	0.88	0.88
CatBoost	0.93	0.93	0.94	0.94	0.93
Extra Trees	0.95	0.93	0.97	0.95	0.95
Deep Learning FNN	0.93	0.90	0.96	0.93	0.93
Ensemble methods (Stacking)	0.94	0.94	0.94	0.94	0.94
Hybrid model	0.94	0.92	0.96	0.94	0.94

After careful analysis of the above results, the Extra Trees model exhibits superior performance in terms of accuracy, recall, F1 score, and ROC AUC score. It was established as the most well-balanced model as it relates to both precision and recall.

Accuracy:0.95, Precision: 0.93, Recall: 0.97, F1 Score: 0.95, ROC AUC Score: 0.95.

The next high-performing model is the combination of Ensemble methods. This model has very good performance in terms of accuracy, precision, recall, F1 score, and ROC AUC score. It has better precision than the Extra Trees model, and although it has a slightly lower performance compared to the Extra Trees model, it still offers a strong predictive capability. This combination utilizes a variety of methods to create strong and precise forecasts.

Accuracy: 0.94, Precision: 0.94, Recall: 0.94, F1 Score: 0.94, ROC AUC Score: 0.94.

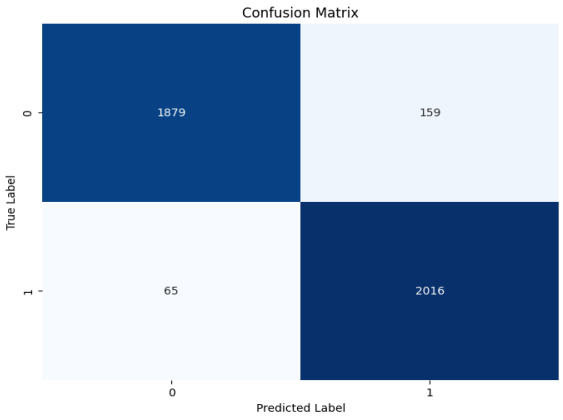
The two aforementioned models were selected for further optimization using the Radomnized search cv hyperparameter tuning. The randomized search evaluated 100 different parameter combinations over 5 folds (resulting in 500 total fits) and identified the optimal parameters that enhanced the model's predictive capabilities. Also, we applied stratified K-fold cross-validation to the models because of the imbalanced nature of the dataset, to ensure their robustness and to prevent it from overfitting. The results after the application of the above process show that the Ensemble method outperformed the Extra Tress.

TABLE II

SUMMARY OF FINAL MODEL PERFORMANCE

Model	Mean CV Accuracy	Accuracy	Precision	Recall	F1 score	ROC-AUC
Extra trees	0.91	0.95	0.93	0.97	0.95	0.95
Ensemble methods	0.93	0.94	0.94	0.94	0.94	0.94

The Extra Trees model has the highest accuracy, recall and an F1 score; hence, it is more suitable for this predictive analytics task. This model is very good at capturing most of the actual purchases, which can be very useful in trying to maximise potential revenue by identifying most purchasing sessions. Although it is slightly lower in precision when compared to the Combination of Ensemble Methods, the improvement in recall and F1 score makes it much more effective, overall, in this context. The results demonstrate that the extra trees model exhibits superior accuracy in correctly identifying both true positive and true negative cases, while also delivering a more equitable performance across a range of evaluation criteria. Below is a heatmap of the confusion matrix of the ensemble methods.



V. CONCLUSION

After an extensive analysis of various machine-learning models to predict online shoppers' purchasing intentions, the extra trees emerged as the optimal solution. This approach demonstrated exceptional performance across multiple metrics, solidifying its position as the best model for our predictive task.

Practically speaking, the confusion matrix revealed that with just 224 misclassifications, 1879 were true negatives out of 4119 sessions and 2016 were true positives. With this great degree of accuracy—0.95—the model is quite successful in accurately spotting sessions that lead to purchases, so reducing the false positives. Also, a very high recall of 0.97, shows that almost all real purchase sessions can be found by the model, and this guarantees that possible buyers are not disregarded. The harmonic mean of precision and recall (FI Score) was 0.95, suggesting a balanced performance in terms of both minimizing false positives and true positive capture. With a ROC AUC score of 0.95, the model's outstanding discriminating capacity—which helps to separate purchasing from non-purchasing sessions—is even more highlighted. The selection of Extra Trees as our optimal predictive model does provide significant value and strategic benefits for business. Its high accuracy and precision help the business to improve its targeted marketing efforts while focusing on users who are most likely to make a purchase. This helps to lower marketing costs, increase the conversion rates significantly, and ensure that resources are used efficiently to yield a higher return on investment.

Secondly, the model presents a very high recall, which indicates that almost all the potential buyers are correctly classified. This constant identification helps in the formulation of digital marketing strategies such as recommended products or services as well as, promotional offers to the customers thus enhancing their experience. Personalized marketing does more than just increase sales rate at any given time; it also enhances consumers' loyalty by making them purchase more often and spend more time on the brand.

Tactically, based on the analytical findings generated from the model's output, it is possible to address various business issues. The Knowledge of the pertinent purchasing influences ensures that the business modifies its product portfolio, adapts different strategies for its price, and enhances the techniques of its service delivery. Such an approach is useful in making sure that the business operations reflect the current market as well as consumer preferences implying the facet of constant advancement.

VI. References

[1] . O. C. Sakar, M. A. Katircioglu, S. O. Polat and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Computing and Applications 31(12),* p. 6893–6908, 10 2019.

[2] B. Mobasher, H. Dai, T. Luo and M. Nakagawa , "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," *Data Mining and Knowledge Discovery,* p. 61–82, Janauary 2002.

[3] W. W. Moe, "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream,," *Journal of Consumer Psychology,* vol. 13, no. 1-2, pp. 29-39, 2003.

[4] G. Suchacka, M. Skolimowska-Kulig, and A. Potempa, "Classification of e-customer sessions based on support vector machine.," *ECMS 15,* p. 594–600, 2015.

[5] G. Suchacka and G. Chodak, , "Using association rules to assess purchase probability in online stores.," *Information Systems and e-Business Management* , p. 751–780, 2017.

[6] "Computerised Recommendations on E-Transaction Finalisation by Means of Machine Learning," *Statistics in Transition. New Series,* vol. 16, pp. 309-322, 2015.

[7] B. Clifton, " Advanced Web Metrics with Google Analytics," *Wiley,* p. 608, 2012.

[8] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226 - 1238, 2005.

[9] S. Kotsiantis, I. Zaharakis and P. E. Pintela, "Supervised Machine Learning: A Review of Classification Techniques.," *Artificial Intelligence Review,* no. 3, pp. 159-190, 2007.

[10] K. Baati, T. Hamdani, A. Alimi and A. Abraham, "A new possibilistic classifier for mixed categorical and numerical data based on a bi-module possibilistic estimation and the generalized minimum-based algorithm," *Journal of Intelligent & Fuzzy Systems,* vol. 36, no. 4, p. 3513–3523, 2019.

[11] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Conference on Uncertainty in Artificial Intelligence*, 2013.

[12] J. R. Quinlan, C4.5: Programs for Machine Learning, Amsterdam: Elsevier, 2014.

[13] L. Breiman, "Random forests.," *Machine Learning,* vol. 45, no. 1, p. 5–32, 2001.

[14] J. Tian, G. Hong and W. Liu, "Imbalanced classification using support vector machine ensemble," *Neural Computing and Applications,* vol. 20, no. 2, pp. 203-209, 2011.

[15] A. W. Ding, S. Li and P. Chatterjee, " Learning user real-time intent for optimal dynamic web page transformation," *Information Systems Research,* vol. 26, no. 2, pp. 339-359, 2015.

[16] N. V. Chawla, K. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, no. 1, pp. 321-357, 2002.

[17] G. Suchacka, M. Skolimowska and A. Potempa, "A k-Nearest Neighbors Method for Classifying User Sessions in E-Commerce Scenario," *Journal of Telecommunications and Information Technology,* vol. 3(64), pp. 64-69, 2015.

[18] "Greedy function approximation: A gradient boosting machine.," *Annals of Statistics,* vol. 5, no. 29, pp. 1189-1232, 2001.

[19] T. Chen and . C. Guestrin, "XGBoost: A Scalable Tree Boosting System.," in *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , New York,, 2016.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, p. 6638–6648.

[21] P. Geurts, Ernst, D and L. Wehenkel, "Extremely randomized trees.," *Machine Learning* , vol. 63, no. 1, p. 3–42, 2006.