Final Portfolio Project – Classification Task

Concepts and Technologies of AI (5CS037)

SUBMITTED BY: Samir Bahadur Karki

UNIVERSITY ID: 2501399

COLLEGE ID : 240188

MODULE : CONCEPT AND TECHNOLOGIES OF AI CODE : 5CS037

MODULE LEADER : SIMAN GIRI

TUTOR : ROBIN TULADHAR

Table of Contents

Abstract

The rising trend of electronic commerce websites has raised a need for understanding and predicting how online users make their purchasing decisions. In the current business environment, electronic data is used by businesses to refine their marketing strategies in a bid to ensure that they provide the best experience for their customers, with the aim of increasing their sales conversion, hence contributing to the enterprise's growth. The major aim of this project is to identify, using classification models, whether an online shopping process is likely to result in a purchase or not.

For this project, it is proposed that the Online Shoppers Purchasing Intention dataset be used. The dataset contains approximately 12,000 records, with a wide array of information relating to online shoppers. The nature of the dataset is expected to be instrumental in identifying patterns that can help in distinguishing between purchasing and non-purchasing processes.

Furthermore, it is worth noting that this project is in line with one of the United Nations Sustainable Development Goals (UNSDG) goals, which is Goal 8: Decent Work and Economic Growth. The project is aimed at ensuring efficient decision-making processes by businesses, hence contributing to sustainable economic growth in the electronic marketplace.

The methodology is comprehensive, with several processes involved in ensuring that a clear understanding is obtained. The processes include extensive pre-processing of the data, which is aimed at exploring the nature of the data using exploratory data analysis (EDA). The process is expected to help in understanding how well a proposed model can predict online shopping purchasing intentions using Accuracy, Precision, Recall, and F1 Score as a guide.

1. Introduction

1.1 Background and Problem Statement

As the availability of online shopping sites becomes more prevalent and continues to grow, more and more information is being collected by businesses. This information is collected based on how users interact with these sites. After conducting an in-depth analysis of the information collected, businesses can gain more insights about their customers and create personalized marketing strategies for each individual, thus increasing the overall conversion rates. However, accurately predicting if a session will result in a sale is a complex task. This is mainly due to the variety of interactions that users can have, the noisy information, and the class imbalance problem. This problem can be solved by using machine learning classification algorithms.

1.2 Dataset Description

The data set used in this research is the "Online Shoppers Purchasing Intention" data set. This data set is developed by C. O. Sakar along with his research team. It is accessible through the UCI Machine Learning Repository. It is a real-world data set collected from an online shopping website. It contains a wide range of features that reflect the nature of an online user. It includes features like administrative information, information content, product information content, etc. Product information content is indicated through the number of pages visited. It includes session-related features like session length in minutes, session-level bounce rate, session-level exit rate, etc. It also includes visitor-related features. Additionally, it includes features like the months in which the session took place. The target feature is indicated as "Revenue," which is a binary feature, taking 1 as a value for a purchase and 0 as a value for a non-purchase.

1.3 Alignment with UNSDG

The current research is in line with the UNSDG 8, which is centered on the objective of the Decent Work and Economic Growth goal. This is because, by enhancing the quality of decision-making in the digital commerce domain, businesses can increase their efficiency, thereby promoting economic growth.

1.4 Objectives

The objectives of the classification task are as follows:

- To investigate and comprehend the unique features of the behavior of online shoppers.

- To create and test different classification models.

- To apply hyperparameter tuning and feature selection.

- To find the best model and analyze the results.

2. Data Preprocessing

Data preprocessing is conducted to ensure that the quality of the data is always high, which in turn improves the quality of the predictions generated by the model. This is done by performing a thorough scan for any missing information and other inconsistencies in the data that may affect the analysis. For the categorical variables such as ‘Month’ and ‘VisitorType,’ the categories are represented by numbers through the use of encoding schemes. On the other hand, for the numerical variables, scaling is performed when necessary, as this is always beneficial to the training process, especially for neural network models that may be affected by the scale of the input variables.

After completing the data preprocessing process, the dataset will then be divided into smaller sets: one used in the training process and another used in the testing process. This process will help us evaluate how well our classification model will perform in generalizing the data it will encounter in the future.

3. Exploratory Data Analysis (EDA)

To gain a deeper understanding of the data and to determine the key patterns that affect the decision to make a purchase, Exploratory Data Analysis (EDA) was carried out. As part of this process, summary statistics were calculated to determine the distribution of the features and their level of variability. In addition to this, a number of visualizations were also conducted to determine the relationships between different variables and to determine which variables are correlated with one another.

From the results obtained from the analysis, it was clear that there was a significant number of sessions that were related to a purchase compared to those that were not. Furthermore, the features related to the products themselves and the overall session length were highly correlated with the target variable, suggesting that they play a key role in determining whether a user is likely to intend to make a purchase or not.

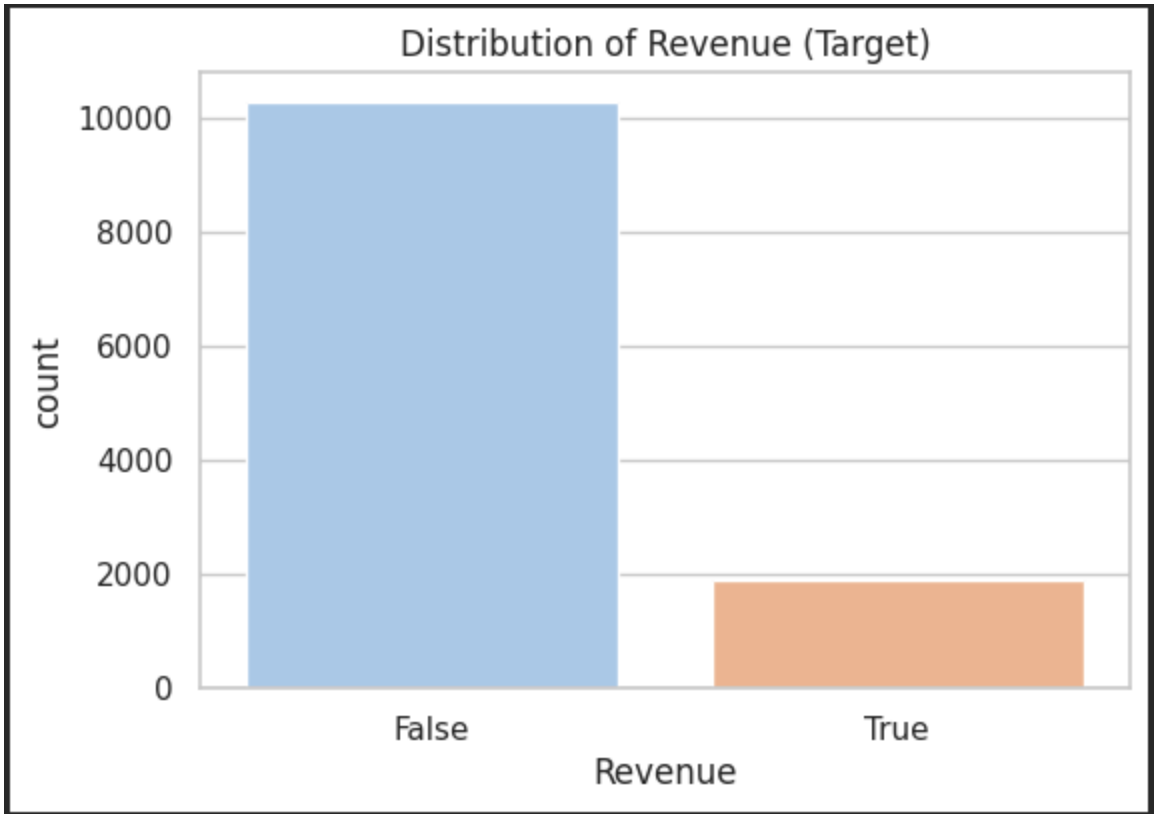

Figure 1: Distribution of purchasing and non-purchasing sessions in the dataset.

Figure 1 shows that non-purchasing sessions significantly outnumber purchasing sessions, highlighting class imbalance in the dataset.

4. Model Development

4.1 Neural Network Model

An expanded explanation of the approach follows:

The project begins with the development of a classifier based on a multi-layer perceptron (MLP) with the intention of successfully learning and modeling the complex non-linear relationships that exist in the data set. This will be achieved by designing the neural network with multiple hidden layers, where each hidden layer will be provided with ReLU activation functions. The ReLU activation function will be used to provide the neural network with the ability to perform complex and powerful transformations on the data features, ensuring that it can successfully learn the complex relationships in the data set. In the final output layer, the sigmoid activation function will be used to ensure that the model can produce probability outputs. In order to determine how well the probability outputs match the actual binary values, the binary cross-entropy loss function will be used to measure the difference between the actual values and the predicted values. The Adam optimization algorithm will be used to update the model parameters with the help of gradient-based optimization due to its desirable properties and performance in neural networks.

4.2 Classical Machine Learning Models

Two classical models used in machine learning have been developed and tested to ensure a solid base and to investigate complex patterns within the data. First, Logistic Regression is a powerful linear model that is highly valued because of its simplicity, efficiency, and interpretability. Secondly, the Random Forest Classifier is another classical ensemble model based on the principle of collective decision-making, which can potentially identify complex relationships within the attributes of the data..

5. Hyperparameter Optimisation and Feature Selection

To carry out the hyperparameter optimisation process, GridSearchCV along with cross-validation was utilised to determine the most effective model configurations and to improve the model's generalisation capability. In this regard, for the Logistic Regression model, the strength of regularisation parameter 'C' was optimised to control and regularise model complexity. For the Random Forest classifier model, critical parameters were optimised to determine the ideal level of model bias and variance by adjusting parameters such as the number of estimators and maximum depth of trees.

In addition to the hyperparameter optimisation process, feature selection was also carried out to determine the features that exert the maximum influence on the model output and to reduce model complexity by employing the recursive feature elimination technique. The results of the feature selection process showed that the features related to products and session durations exert a dominant influence on the model output for determining purchasing intention. The results of the feature selection process are consistent with the insights obtained during the exploratory data analysis process.

6. Results and Discussion

For all the classification models being considered in the problem statement, the evaluation process was conducted based on four different criteria: Accuracy, Precision, Recall, and F1 Score. In all the models considered for the problem statement, the model that stood out with reliable performance was the Logistic Regression model. However, it also had its shortcomings when dealing with non-linear data relationships. In contrast, the results obtained from the neural network model were positive compared to the results obtained from the logistic regression model. However, it should be noted that the neural network model had to be tuned very specifically in order to avoid overfitting the model.

Out of all these models, the Random Forest classifier was found to perform exceptionally well by providing good results on all parameters. This was due to the nature of this model, which was able to deal with complex relationships between all the features.

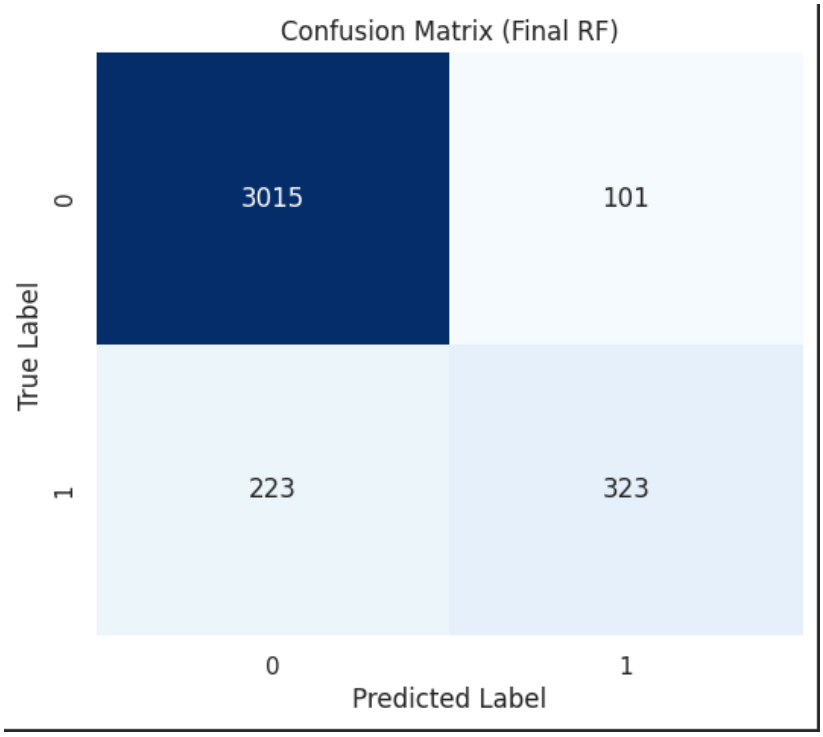

Figure 2: Confusion matrix of the Random Forest classifier on the test dataset.

Figure 2 illustrates the classification performance of the Random Forest model, showing a high number of correctly classified instances.

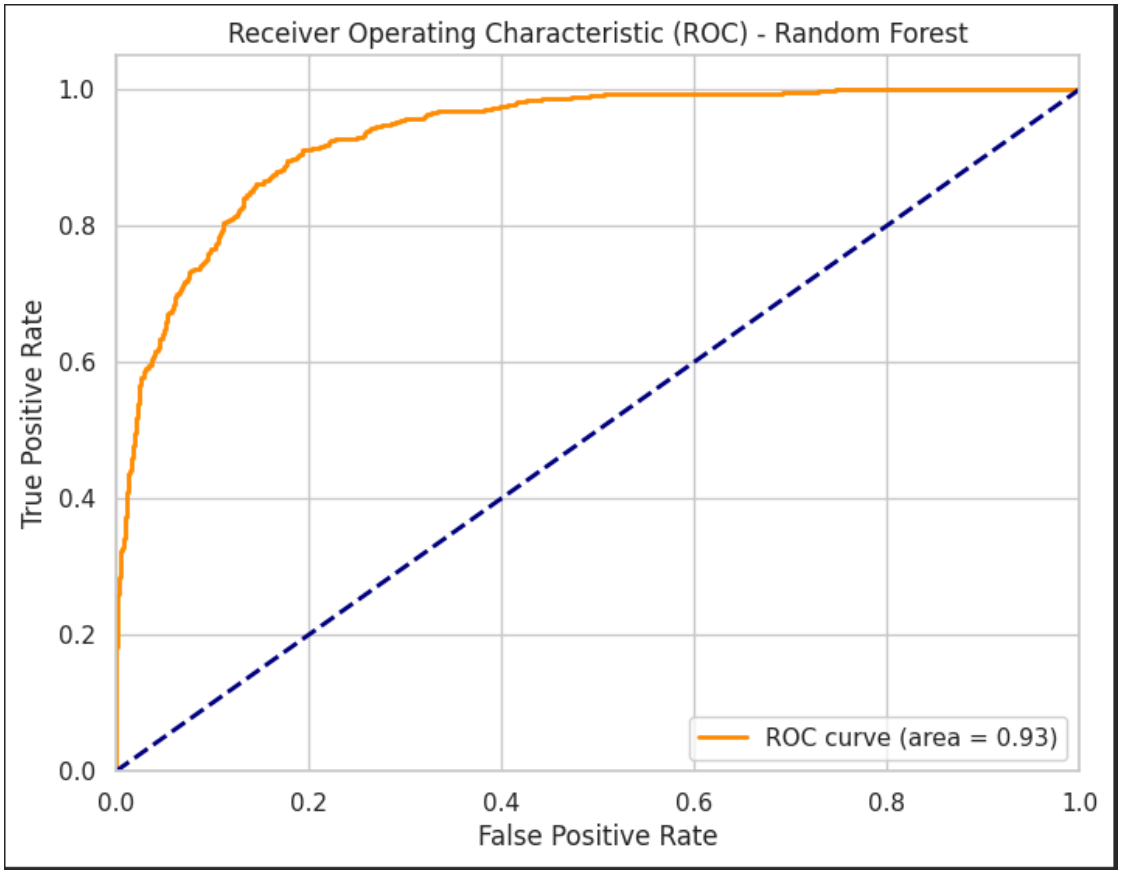

Figure 3: *ROC curve of the Random Forest classifier showing model discrimination performance.*

The ROC Curve shows that the model has performed well in the classification task, with the AUC value indicating that the non-purchasing and purchasing sessions have been effectively separated.

7. Conclusion and Reflection

The effectiveness of this research can be established through the capability of these classifiers, which are developed through machine learning, to accurately predict online purchasing intention. Out of these classifiers, it was established that the most suitable and accurate classifier was the Random Forest classifier, which was better than the other classifiers. Looking forward to future research, it is possible to use even stronger classifiers like Gradient Boosting to achieve better results. Furthermore, it is possible to solve problems related to unbalanced classes through resampling methods, while deeper neural networks can be used to reveal complex patterns.

References

1. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using neural networks. Neural Computing and Applications.
2. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
5. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
6. Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
7. United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development.

https://github.com/Zxynx77/5CS037/tree/main/Ai-final_protfolio

Similarity Report

PAPER NAME
SamirKarki2501399Classification.pdf

AUTHOR
-

WORD COUNT
1958 Words

CHARACTER COUNT
11731 Characters

PAGE COUNT
11 Pages

FILE SIZE
213.4KB

SUBMISSION DATE
Feb 10, 2026 1:17 PM GMT+5:45

REPORT DATE
Feb 10, 2026 1:18 PM GMT+5:45

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 10% Internet database
- 10% Publications database
- Crossref database
- Crossref Posted Content database
- 15% Submitted Works database