



# Ecommerce purchase Intention prediction

## **Submitted by:**

Harsanbruno Maria Joseph Esuraj

Student id -20065575

B9DA110 ADVANCED DATA AND NETWORKING MINING (B9DA110\_2425\_TMD3)

Professor - Oleksandr Bezrukavyi

## 1. Introduction

Customer behavior in the current competitive e-commerce world is an essential element when it comes to boosting online purchasing and supporting better user experience. Data mining offers the possibility of integrating online shopping information so that business firms can analyze volumes of data and make purchases through the identification of the purchase trends and the ability to make precise forecasts regarding customer behavior in the future.

This project is aimed at predicting whether an online shopper makes a purchase in a browsing session or not. Through the aid of the data on Online Shoppers Intention, we will be trying to draw in an insightful piece as well as creating a predictive model which will be useful in allowing the businesses to target potential buyers adequately.

This work aims at the following considerations:

- Examine behavior trends in visitors utilizing attributes of sessions i.e. visits on particular page or the duration of them, and the bounce rates.
- Construct and test a classification model to determine the intention to purchase.
- Telephone communication Update to join and set up a unitary command.

The project shows that data mining may be applied as the strategic tool to increase conversion rates and consumer interest to online retailing platforms with the help of data mining, data preprocessing, exploratory data analysis (EDA) and machine learning algorithms.

---

## 2. Dataset Overview

This project will use the Online Shoppers Intention dataset that comprises data on the sessions level gathered by an online web shop. It consists of 12 330 records and 18 features, which explains how the users behave when visiting the site online.

### 2.1 Source of data

- Data set name: Online shoppers intention
- Source: UCI Machine Learning Repository
- Type: CSV file
- Records Total: 12,330
- Number of Features: 18 (17 independent features, 1 target feature)

### 2.2 Description of Features

The features are varied characteristics of the activity of users:

- **Administrative / Administrative Duration:** How many and total time of administrative related pages.
- **Informational / Informational Duration:** The count of pages and the time on informational pages.

- **Product Related / Product Related Duration:** The amount and the amount of time spent on the pages relating to the products.
  - **Bounce Rates:** This is the percentage of visitors that leave immediately after viewing one page.
  - **Exit Rates:** Probability of leaving the site out of a definite page.
  - **Page Values:** The average of page values regarding conversions.
  - **SpecialDay:** The physical closeness of the session to a special shopping day (i.e., Valentine Day).
  - **Month:** Month the session occurred in.
  - **Operating System, Browser, Region, Traffic Type:** Information about the technical and geographic session.
  - **Visitor Type:** The difference between a returning or a new customer.
  - **Weekend:** Did the session take place at the weekend.
  - **Revenue (Target Variable):** Boolean (indicates whether the session has led to a purchase (True/False)).
- 

## 2.3 Properties of Data

One of the important aspects of data is the characterization of the data.  
Is composed of numeric and dichotomous values.

- Target variable (Revenue) is skewed in that most of the sessions did not result in purchase.
  - A seasonal analysis of trends is possible, because the features are personified (Month, Special Day).
  - The data is broad in terms of illustrating customer behaviour and is hence suitable in the classification and prediction data mining exercise.
- 

## 3. Methodology

This section describes what was done in the analysis of the data and the construction of predictive models related to determining whether a shopping activity online will lead to a purchase.

---

### 3.1 Preprocessing Data

#### 1. Lifting the Dataset

The dataset (online\_shoppers\_intention.csv) was read with the help of Pandas.

- Simple inspection was done by using `.head()`, `.info()` and `.describe()` to familiarize with its structure.

#### 2. Processing of Missing values

- The data were inspected on null values; there were none that needed imputation.
3. **Coding Categorical Variables**  
Features such as Month and VisitorType were subjected to Label Encoding.
  - Booleans attributes (Weekend and Revenue) were modified to binary integers (0/1).
  4. **Feature Scaling**  
Standard variables like Administrative Duration, ProductRelated Duration, BounceRates, etc., were transformed into standard variables as StandardScaler is done to improve the model performance.
  5. **Train-Test Split**  
The training-test split of the data was done using train\_test\_split and 80 percent was used as training data and 20 percent testing.
- 

### 3.2 Exploratory Data Analysis (EDA) Revenue Analysis

The analysis on the target variable Revenue that determines whether there was a purchase (True) or not (False) in a session started.

- False: 84.53 percent of the session never converted to a sales.
- True: Just 15.47 percent of the sessions had resulted in purchase.

This huge disparity reveals the fact that most online shoppers visit and to a greater extent add items to the cart, but do not proceed to purchase. This might be as a result of comparison shopping or even indecisiveness when it comes to pricing or distractions at the check out.

As it is a very imbalanced dataset, classification models should be chosen attentively, and the accuracy might be considered only along with the precision, recall, and the F1- score to be accurate and reliable and guarantee the adequate prediction of the minority True class.

---

### 3.3 Models Used

To be able to predict whether a customer session yields a purchase (Revenue) six different machine learning models were used. They were both selected in order to compare different types of algorithmic solutions ranging across simpler interpretable solutions, to more complex ensemble and deep learning solutions.

1. **Logistic Regression**
  - A starting point statistical framework that is most suitable in binary classification.
  - It was selected because it is easily interpretable and allows a rapid determination of a performance benchmark.
2. **Decision Tree**
  - Visual, intuitive and non-linear pattern-capturing.
  - It can be used to provide an insight into what features tend to produce the most impact when making a purchase.

### 3. **Random Forest**

- A multi-layered model of decision trees which minimizes overfitting.
- Gives feature scores of significance, and can give better predictive accuracy than a single tree.

### 4. **Single-Layer Perceptron (SLP)**

- The simple feed-forward neural network.
- Chosen to examine the way the complex relationship can be handled between the session features and the purchase behavior by the deep learning.

### 5. **Support Vector Machine (SVM)**

- Works well in high-dimensional space and can deal with non-linear boundaries by use of kernel functions.
- Located to test how it will distinguish between purchasing and non-purchasing sessions.

### 6. **XGBoost**

- An algorithm is an optimized combination of a gradient boost algorithm, which works fastest.
- It is well-known to have high accuracy and work in non-balanced datasets.

The attempt to compare the different approaches helped this study further compare the different methods to find the best method in modeling which customers are likely to purchase.

## **4. Evaluation of the models and the results.**

Here we show the result of testing the performance of 6 machine learning models in predicting whether a customer of an online shop will make a purchase within a browsing period. The comparison is based on various performance measures that include precision, recall, F1-score, accuracy, ROC-AUC as well as confusion matrix. Since the dataset is highly imbalanced, it is paramount to use measures that would show the quality of how models would identify the minority class (purchases).

---

### **4.1 Test of Logistic Regression**

#### **Classification Report:**

## Logistic Regression Evaluation

### Classification Report:

	precision	recall	f1-score	support
0	0.89	0.98	0.93	2084
1	0.76	0.36	0.49	382
accuracy			0.88	2466
macro avg	0.83	0.67	0.71	2466
weighted avg	0.87	0.88	0.86	2466

### Confusion Matrix:

```
[[2042  42]
 [ 246 136]]
```

ROC-AUC Score: 0.8652

### Confusion Matrix:

Confusion Matrix = (204242246136) Confusion Matrix = (204224642136)

ROC-AUC: 0.8652

The results of the Logistic Regression indicate that the model works well to predict non-purchaser (class 0) but fails to predict those who can become buyers (class 1) based on the poor recall of the latter. At the level of 0.8652 AUC, which represents the moderate discriminative power, it is an adequate baseline model.

---

## 4.2 Evaluation in Decision Tree

### Classification Report:

#### Decision Tree Evaluation

### Classification Report:

	precision	recall	f1-score	support
0	0.92	0.91	0.91	2084
1	0.53	0.55	0.54	382
accuracy			0.85	2466
macro avg	0.72	0.73	0.72	2466
weighted avg	0.86	0.85	0.85	2466

### Confusion Matrix:

```
[[1894  190]
 [ 172  210]]
```

ROC-AUC Score: 0.7293

### Confusion Matrix:

Confusion Matrix = (1894172190210) Confusion Matrix = (1894172190210)\begin{pmatrix} 1894 & 172 \\ 190 & 210 \end{pmatrix}(1894190172210)

**ROC-AUC Score: 0.7293**

Decision Tree works quite well in predicting non-purchasers (class 0) however model could not perform well in predicting purchases (class 1) as indicated by its low precision and recall of class 1. The AUC score is 0.7293 which depicts a moderate performance to classify the class.

---


### 4.3 Random Forest Score

#### Classification Report:

Random Forest Evaluation

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.96	0.94	2084
1	0.73	0.56	0.63	382
accuracy			0.90	2466
macro avg	0.83	0.76	0.79	2466
weighted avg	0.89	0.90	0.89	2466

 Confusion Matrix:

```
[[2007  77]
 [ 169 213]]
```

ROC-AUC Score: 0.9182

#### Confusion Matrix:

Confusion Matrix = (20077169213) Confusion Matrix = (200777169213)\begin{pmatrix} 2007 & 77 \\ 169 & 213 \end{pmatrix}(200716977213) Confusion Matrix = (200716921377213)

**ROC-AUC: 0.9182**

Random Forest has good results in predicting both values of 0 (no purchase) and 1 (purchase) with an ROC-AUC of 0.9182. It is effective in separating the bought customers and the ones that do not, but the purchasing recall can be better.

---

### 4.4 Classification of MLP

#### Classification Report:

## MLP Classifier Evaluation

### Classification Report:

	precision	recall	f1-score	support
0	0.92	0.95	0.93	2084
1	0.65	0.52	0.58	382
accuracy			0.88	2466
macro avg	0.78	0.74	0.75	2466
weighted avg	0.87	0.88	0.88	2466

### Confusion Matrix:

```
[[1974  110]
 [ 182  200]]
```

ROC-AUC Score: 0.8868

### Confusion Matrix:

The Confusion Matrix is = (1974182110200)

**ROC-AUC Score:** 0.8868

- MLP Classifier achieves a high precision and recall in the prediction of class 0 (no purchase) but moderate recall in class 1 (purchase). ROC-AUC (0.8868) score is signifying that there is a good discriminating capacity, but improvement can be made concerning purchase prediction.

---

## 4.5 Performance of SVM Classifier

### Classification Report:

#### SVM Classifier Evaluation

### Classification Report:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	2084
1	0.71	0.43	0.53	382
accuracy			0.88	2466
macro avg	0.80	0.70	0.73	2466
weighted avg	0.87	0.88	0.87	2466

### Confusion Matrix:

```
[[2016   68]
 [ 218  164]]
```

ROC-AUC Score: 0.8520

---



## 4.6 Analysis of XGBoost Classifier

### XGBoost Classifier Evaluation

#### Classification Report:

	precision	recall	f1-score	support
0	0.92	0.95	0.94	2084
1	0.67	0.56	0.61	382
accuracy			0.89	2466
macro avg	0.80	0.76	0.77	2466
weighted avg	0.88	0.89	0.89	2466

#### Confusion Matrix:

```
[[1978 106]
 [ 167 215]]
```

ROC-AUC Score: 0.9161

#### Confusion Matrix:

Confusion Matrix = (1978167106215) Confusion Matrix = (1978167106215)

**Score:** 0.9161

XGBoost is performing well, in all measured dimensions, especially when differentiating non-purchasers (0) with high recall (0.95) and precision (0.92). It also does well as far as class 1 (purchase) is concerned but recall (0.56) remains moderate. A total score of 0.9161 in the ROC-AUC score gives a very good discriminatory power.

## 4.8 Comparison of Models

Here, we are comparing the performance of the six models, namely, Logistic Regression, Decision Tree, Random Forest, MLP, SVM, and XGBoost. Using the main indicators based on Accuracy, Precision, Recall, F1-score, and AUC-ROC, we will conclude which of the models is better suited to predicting whether an online shopper will make a purchase on the site during a browsing session.

#### Model Performances:

	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	Logistic Regression	0.8832	0.7640	0.3560	0.4857	0.8652
1	Decision Tree	0.8520	0.5220	0.5288	0.5254	0.7200
2	Random Forest	0.9006	0.7246	0.5785	0.6434	0.9158
3	MLP	0.8852	0.6533	0.5524	0.5986	0.8730
4	SVM	0.8840	0.7069	0.4293	0.5342	0.8520
5	XGBoost	0.8893	0.6698	0.5628	0.6117	0.9161

#### Analysis:

##### 1. Accuracy:

- Best: Random Forest (0.9006)

- Worst: Decision Tree (0.8520)

Accuracy indicates the frequency of the model getting predictions right. Random Forest is ahead in this case with XGBoost and MLP showing to be quite useful, as well. The accuracy of Decision Tree is the least probably as a result of overfitting.

## 2. Precision:

- Best: Logistic Regression (0.7640)
- Worst: Decision Tree (0.5220)

Logistic Regression has the highest precision, indicating that when it predicts a purchase, it's more likely to be correct. This matters in reducing false positives. Decision Tree fares badly on precision.

## 3. Recall:

- Best: Random Forest (0.5785)
- Worst: Logistic Regression (0.3560)

Random Forest has the highest accuracy in identifying genuine purchases (class 1), where Logistic Regression does not fare well, which means that it fails to identify many genuine buyers (false negatives).

## 4. F1-score:

- Best: Random Forest (0.6434)
- Worst: Logistic Regression (0.4857)

Random Forest achieves the best precision and recall balance followed by XGBoost. Logistic Regression presents lower F1-score which signifies that precision and recall are on the opposite sides of class 1 (purchases).

## 5. ROC-AUC:

- Best: XGBoost (0.9161)
- Worst: Decision Tree (0.7200)

XGBoost is the most outstanding in the differentiation between class 0 (no purchase) and class 1 (purchase). This ROC-AUC score is very high at 0.9161. Decision Tree is the model with the worst AUC, pointing out that the model is not able to differentiate between the two classes.

---

## The Best Model: Random Forest

Judging by the above comparison, the most optimal model is Random Forest due to the following reasons:

### 1. Balanced Performance:

Random Forest ranks high according to all the metrics with the best result being the accuracy and a good F1-score (0.6434), which demonstrates a good balance between precision and recall.

2. **High ROC-AUC:**

Random Forest has a good ROC-AUC score of 0.9158 implying that it is an effective splitter of purchasers and non-purchasers which is necessary in this e-commerce application.

3. **Distribution of Class Imbalance:**

The model is quite good with the unbalanced set of data, with the minority class of 1 (purchase). Random Forest model is more effective than the majority of models to work with this imbalance, and it is superior in terms of recall of class 0 (non-purchase) and rather good in terms of recall of class 1 (purchase).

4. **Interpretability:**

Random Forest also gives the feature importance that can be quite useful as a sort of explanation of the factors that affect the probability of making a purchase. This not only makes it a high-performance model but it is also an interpretable one.

---

## Summary of Coding Part

A number of machine learning algorithms were applied and tested in this project to perform a prediction task where the attribute was determined by whether an online shopper would make a purchase in a browsing session or not. Models that were selected included Logistic Regression, Decision Tree, Random Forest, MLP (Multi-Layer Perceptron), SVM (Support Vector Machine), and XGBoost. The different models were assessed in terms of their different performance measures such as accuracy, precision, recall, F1, and ROC-AUC.

The major results of the assessment are:

1. The best-performing model was Random Forest that scored the highest accuracy (0.9006) and ROC-AUC score (0.9158). It demonstrated an adequate performance with good precision, recall, and F1-score, especially in the case of class 0 (non-purchase) with reasonable performance regarding class 1 (purchase) as well. This renders Random Forest as the most trustworthy model with respect to this set of data.
2. XGBoost stood out as a competitor, as it has achieved the largest ROC-AUC score (0.9161) and can rank potential buyers and non-buyers well. It performed, however, a little worse than Random Forest with respect to the precision and recall of the minority class, class 1 (purchase).
3. Although a good baseline model, Logistic Regression performed poorer in the prediction of class 1 (purchase), as it recalls 36 percent and has 49 percent F1-score, meaning that many potential buyers are not found.
4. Decision Tree showed reflectivity to Logistic Regression with lesser proportion of precision and recall of class 1 (purchase), making it ineffective in predicting purchases.
5. The MLP and SVM showed decent performance in class 0, but did not perform well in class 1 with MLP having moderate performance in terms of recall (0.52) and SVM worse recall in class 1 (purchase).

Conclusively, the Random Forest makes the most appropriate model in this task since it has shown high reliability, accuracy, and AUC results, which makes it fit as baseline model to predict online

purchase of a shopper. Also, XGBoost can be perceived as a robust model, in particular, when additional tuning is conducted.

The following procedures include the use of RapidMiner in the second part of the task, which is aimed at investigating the same data and testing the models' performance with an alternative application.

---

## **5. Contributions**

The project involved my complete implementation of the project and assessment on the models. It included data preprocessing and model building as well as evaluation of performance. I made most of the running with data exploration, data preprocessing, and building of several models such as Logistic Regression, Decision Tree, Random Forest, and SVM. The application of XGBoost and MLP was assisted by my teammate.

### **Fields I worked in:**

- Data preprocessing
- Model Implementation
- Explorative Data Analysis (EDA)
- Model Evaluation
- Also did the ppt
- Helped in rapid miner in the aisttudio