### Research with solid fill7.3 Hands-on Case Study:

Build Ensemble Technique : Decision Tree & Gradient Booster Model

*Download Source Data Set from GitHub link* : <https://github.com/Smartbrain2024/Mastering_AI_2.git>

**Chapters/Chp\_07/7.3/Hands\_on/online\_shoppers\_intention.csv**

**About Dataset**

**Context**

**Data Set Information:**

The dataset consists of feature vectors belonging to 12,330 sessions.  
The dataset was formed so that each session  
would belong to a different user in a 1-year period to avoid  
any tendency to a specific campaign, special day, user  
profile, or period.

**Content**

The dataset consists of 10 numerical and 8 categorical attributes.  
The ‘Revenue’ attribute can be used as the class label.

“Administrative,” “Administrative Duration,” “Informational,” “Informational Duration,” “Product Related” and “Product Related Duration” represent the number of several types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g., moving from one page to another. The “Bounce Rate,” “Exit Rate” and “Page Value” features represent the metrics measured by “Google Analytics” for each page in the e-commerce site. The value of “Bounce Rate” feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session. The value of “Exit Rate” feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The “Page Value” feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The “Special Day” feature indicates the closeness of the site visiting time to a specific special day (e.g., Mother’s Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

Objective : Build AI Model with highest accuracy for Revenue using Ensemble technique – Decision Tree ,Gradient Booster Algo in KNIME workflow .

Step 1 EDA: Convert Number into String

A diagram of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step 2: Univariate and Bi variate EDA analysis :

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Revenue CountsA screenshot of a computer

Description automatically generated

Visitor Type analysis

A screenshot of a graph

Description automatically generated

Conditional Box Plot : Revenue Vs Page value

A screenshot of a graph

Description automatically generated

Cross Tab Analysis: Revenue Vs Month

A screenshot of a computer

Description automatically generated

Linear correlation Analysis:

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Data Preprocessing:

A yellow square with black text and black text

Description automatically generated

Step 3: Data Preprocessing

A yellow square with black text and black text

Description automatically generated

Missing Values Component:

A screenshot of a computer

Description automatically generated

One to Many :

A screenshot of a computer

Description automatically generated

Step 4: Column FilterA screenshot of a computer

Description automatically generated

Step 5 : Partitioning Train and Test Data

A screenshot of a computer

Description automatically generated

Decision Tree Modelling:

A diagram of a tree

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Train Scorer:

A screenshot of a computer

Description automatically generated

Test Scorer

A screenshot of a computer

Description automatically generated

Gradient Booster:

A diagram of a tree

Description automatically generated

A screenshot of a computer

Description automatically generated

Train Scorer:

A screenshot of a computer

Description automatically generated

Test Scorer:

**Conclusion:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algo Model | Data Sample | Accuracy % | Cohen’s Kappa % | Error % |
| Decision Tree | Train | 92.09 | 0.67 | 7.9 |
| Decision Tree | Test | 89.8 | 0.59 | 10.03 |
| Gradient Booster | Train | **93.57** | **0.73** | **6.43** |
| Gradient Booster | Test | 89.97 | 0.59 | 10.03 |

Choosing Gradient Booster Model with Highest Accuracy is the best choice and further modifying Hyper parameters will gain more accuracy in determining higher Revenue.