Project Title	**Customer Conversion Analysis for Online Shopping Using Clickstream Data**
Skills take away From This Project	**Data Preprocessing and Cleaning** **Exploratory Data Analysis (EDA)** **Feature Engineering** **Supervised and Unsupervised Machine Learning Techniques** **Classification, Regression, and Clustering Models** **Model Evaluation and Hyperparameter Tuning** **Pipeline Development for Data Processing and Modeling** **Streamlit Application Development** **Deployment of Interactive Machine Learning Models**
Domain	**E-commerce and Retail Analytics**

Problem Statement:

Imagine you are a data scientist working at a leading **e-commerce giant like Amazon, Walmart, or eBay**. Your goal is to develop an intelligent and user-friendly **Streamlit web application** that leverages **clickstream data** to enhance customer engagement and drive sales.

The application should:

1. **Classification Problem:** Predict whether a customer will complete a purchase (1) or not (2) based on their browsing behavior.

2. **Regression Problem:** Estimate the potential revenue a customer is likely to generate, helping the business forecast revenue and optimize marketing strategies.
3. **Clustering Problem:** Segment customers into distinct groups based on their online behavior patterns, enabling targeted marketing campaigns and personalized product recommendations.

By building this application, you aim to empower the business with **data-driven insights** to increase conversions, boost revenue, and enhance customer satisfaction.

Business Use Cases:

1. **Customer Conversion Prediction:** Enhance marketing efficiency by targeting potential buyers.
2. **Revenue Forecasting:** Optimize pricing strategies by predicting user spending behavior.
3. **Customer Segmentation:** Group users into clusters for better personalization.
4. **Churn Reduction:** Detect users likely to abandon carts and enable proactive re-engagement.
5. **Improved Product Recommendations:** Suggest relevant products based on browsing patterns.

Approach:

1. Data Preprocessing:

- **Dataset Details:**
  - **Train.csv**: Used to train machine learning models.
  - **Test.csv**: Used to validate model performance and simulate real-world scenarios.
- **Handling Missing Values:**
  - Replace missing values using mean/median for numerical data and mode for categorical data.
- **Feature Encoding:**
  - Convert categorical features into numerical using **One-Hot Encoding** or **Label Encoding**.
- **Scaling and Normalization:**

    - Apply **MinMaxScaler** or **StandardScaler** for numerical features to improve model performance.

---

2. Exploratory Data Analysis (EDA):

- **Visualizations:**
    - Use bar charts, histograms, and pair plots to understand distributions and relationships.
- **Session Analysis:**
    - Analyze session duration, page views, and bounce rates.
- **Correlation Analysis:**
    - Identify relationships between features using correlation heatmaps.
- **Time-based Analysis:**
    - Extract features like hour of the day, day of the week, and browsing duration.

---

3. Feature Engineering:

- **Session Metrics:**
    - Calculate session length, number of clicks, and time spent per product category.
- **Clickstream Patterns:**
    - Track click sequences to identify browsing paths.
- **Behavioral Metrics:**
    - Bounce rates, exit rates, and revisit patterns.

---

4. Balancing Techniques (For Classification Models):

- **Identify Imbalance:**
    1. Analyze the distribution of target labels (converted vs. not converted).
- **Techniques for Balancing:**
    1. **Oversampling:** Use **SMOTE (Synthetic Minority Oversampling Technique)** to create synthetic samples.
    2. **Undersampling:** Randomly remove majority class samples to balance the dataset.
    3. **Class Weight Adjustment:** Assign higher weights to the minority class during model training.

5. Model Building:

Supervised Learning Models:

- **Classification:** Logistic Regression, Decision Trees, Random Forest, XGBoost, and Neural Networks.
- **Regression:** Linear Regression, Ridge, Lasso, Gradient Boosting Regressors.

Unsupervised Learning Models:

- **Clustering:** K-means, DBSCAN, and Hierarchical Clustering.

Pipeline Development:

- Use **Scikit-learn Pipelines** to automate:
  - Data preprocessing → Feature scaling → Model training → Hyperparameter tuning → Evaluation.

6. Model Evaluation:

- **Classification Metrics:** Accuracy, Precision, Recall, F1-Score, ROC-AUC Curve.
- **Regression Metrics:** MAE, MSE, RMSE, and R-squared.
- **Clustering Metrics:** Silhouette Score, Davies-Bouldin Index, and Within-Cluster Sum of Squares.

7. Streamlit Application Development:

- **Interactive Web Application:**
  - Build a **Streamlit** interface that allows users to upload CSV files or input values manually.
- **Key Features:**
  - Real-time predictions for conversion (classification).
  - Revenue estimation (regression).
  - Display customer segments (clustering visualization).
  - Show visualizations like bar charts, pie charts, and histograms.

Results:

- Predict customer conversion with high accuracy and precision.
- Estimate potential revenue from users based on browsing behavior.
- Generate meaningful customer segments for targeted marketing strategies.
- Deploy an easy-to-use **Streamlit** application for end-users.

Project Evaluation Metrics:

Classification:

- Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Regression:

- RMSE, MAE, and R-squared.

Clustering:

- Silhouette Score and Davies-Bouldin Index.

Technical Tags:

- Python, Pandas, NumPy, Matplotlib, Seaborn
- Machine Learning: Scikit-learn, XGBoost, TensorFlow
- Pipelines, Data Preprocessing, Feature Engineering
- Streamlit for Web Applications
- Model Deployment

Dataset:

- **Source:** UCI Machine Learning Repository - Clickstream Data
- **Train Dataset:** `train_data` for training machine learning models.
- **Test Dataset:** `test_data` for evaluating model performance.

Dataset Explanation:

Data description

Variables:

1. YEAR (2008)

==

2. MONTH -> from April (4) to August (8)

==

3. DAY -> day number of the month

==

4. ORDER -> sequence of clicks during one session

==

5. COUNTRY -> variable indicating the country of origin of the IP address with the

following categories:

1-Australia
2-Austria
3-Belgium
4-British Virgin Islands
5-Cayman Islands
6-Christmas Island
7-Croatia
8-Cyprus
9-Czech Republic

10-Denmark
11-Estonia
12-unidentified
13-Faroe Islands
14-Finland
15-France
16-Germany
17-Greece
18-Hungary
19-Iceland
20-India
21-Ireland
22-Italy
23-Latvia
24-Lithuania
25-Luxembourg
26-Mexico
27-Netherlands
28-Norway
29-Poland
30-Portugal
31-Romania
32-Russia
33-San Marino
34-Slovakia
35-Slovenia
36-Spain
37-Sweden
38-Switzerland
39-Ukraine
40-United Arab Emirates
41-United Kingdom
42-USA
43-biz (*.biz) 44-com (*.com)
45-int (*.int) 46-net (*.net)
47-org (*.org)

==

6. SESSION ID -> variable indicating session id (short record)

==

7. PAGE 1 (MAIN CATEGORY) -> concerns the main product category:

1-trousers
2-skirts
3-blouses
4-sale

==

8. PAGE 2 (CLOTHING MODEL) -> contains information about the code for each product

(217 products)

==

9. COLOUR -> colour of product

1-beige
2-black
3-blue
4-brown
5-burgundy
6-gray
7-green
8-navy blue
9-of many colors
10-olive
11-pink
12-red
13-violet
14-white

==

10. LOCATION -> photo location on the page, the screen has been divided into six parts:

1-top left
2-top in the middle
3-top right
4-bottom left
5-bottom in the middle
6-bottom right

===

11. MODEL PHOTOGRAPHY -> variable with two categories:

1-en face
2-profile

===

12. PRICE -> price in US dollars

===

13. PRICE 2 -> variable informing whether the price of a particular product is higher than

the average price for the entire product category

1-yes
2-no

===

14. PAGE -> page number within the e-store website (from 1 to 5)

+++

Project Deliverables:

- **Source Code:** Scripts for preprocessing, modeling, and deployment.
- **Streamlit Application:** Interactive tool for predictions and insights.
- **Documentation:** Explanation of methodology, approaches, and results.
- **Presentation Deck:** Summarized findings and visualizations.

Project Guidelines:

- Use **GitHub** for version control.
- Follow PEP8 coding standards.
- Test each module using unit tests.
- Maintain detailed comments and logs in the code.

References:

Project Live Evaluation	Project Live Evaluation
EDA Guide	Exploratory Data Analysis (EDA) Guide
Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
ML FLOW Tutorial 1	ML FLOW 1
ML FLOW Tutorial 2	ML Flow Documentation ML FLOW 2
Project Orientation (English)	Recording Link
Project Orientation (Tamil)	Recording Link
Pipeline Sklearn	Documentation

Pipeline Tutorial English	Recording Link
Pipeline Tutorial Tamil	Recording Link

Timeline:

1 week

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.
Note: Book the slot at least before 12:00 Pm on the same day

Timing: Monday-Saturday (4:00PM to 5:00PM)

Booking link :https://forms.gle/XC553oSbMJ2Gcfug9

For DE/BADM project/class topic doubt slot clarification session:

Booking link : https://forms.gle/NtkQ4UV9cBV7Ac3C8

Session timing:

For DE: 04:00 pm to 5:00 pm every saturday
For BADM 05:00 to 07:00 pm every saturday

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.
Note: This form will Open only on Saturday (after 2 PM) and Sunday on Every Week

Timing:

For BADM and DE
Monday-Saturday (11:30AM to 1:00PM)

For DS and AIML
Monday-Saturday (05:30PM to 07:00PM)

Booking link : https://forms.gle/1m2Gsro41fLtZurRA

Created By:	Verified By:	Approved By:
Aravinth Meganathan	Shadiya P P	Nehlath Harmain