Clickstream Analysis & Prediction Web Application

Project Overview

This project aims to develop an intelligent and interactive web application to analyze e-commerce clickstream data. By leveraging machine learning techniques, the application predicts customer conversion, estimates potential revenue, and segments users based on their browsing behavior. This helps businesses optimize marketing strategies, increase sales, and improve customer satisfaction.

Business Goals:

- Predict whether a customer will complete a purchase (classification).
- Estimate potential revenue from a customer (regression).
- Segment customers into distinct groups for targeted marketing (clustering).
- Detect users likely to bounce or revisit pages for improved engagement.

Dataset Details

Source: UCI Machine Learning Repository - Clickstream Data

- **Train dataset:** `train_data.csv`
- **Test dataset:** `test_data.csv`

Key Variables:

Column	Description
year	Year of session (2008)
month	Month of session (4-8)
day	Day of month

order	Sequence of clicks in a session
country	Country of visitor
session_id	Unique session identifier
page1_main_category	Main product category
page2_clothing_model	Product model code
colour	Product colour
location	Position of product on page
model_photography	Photo type (en face/profile)
price	Product price in USD
price_2	Price above category average (target for classification)
page	Page number on website

Project Workflow

1. Data Preprocessing

- **Missing Value Handling:**
  - Numerical columns → median imputation
  - Categorical columns → mode imputation
- **Column Standardization:**
  - Convert all column names to lowercase and replace spaces/special characters with underscores.
- **Initial Check:**
  - Inspect categorical and numerical columns
  - Prepare for feature engineering

---

2. Feature Engineering

Performed on **raw data** before encoding or scaling:

Session Metrics:

- `session_length` → max order in session
- `num_clicks` → total clicks per session
- `time_per_category` → number of clicks per category

Clickstream Patterns:

- `first_page` → first page visited in session
- `last_page` → last page visited
- `unique_categories` → number of unique categories viewed

Behavioral Metrics:

- `is_bounce` → session length = 1
- `exit_page` → last page of session
- `is_revisit` → repeated visit to same product

---

3. Encoding & Scaling

- **Categorical Features:** Label Encoding (`first_page`, `last_page`, `exit_page`, etc.)
- **Numerical Features:** StandardScaler for session metrics and other numeric columns

---

4. Exploratory Data Analysis (EDA)

- Visualizations:
    - Bar charts, histograms, pairplots

    - Correlation heatmaps
    - Session duration & pageview distribution
- Goal: Understand feature distributions and relationships

---

5. Handling Imbalanced Classes

- Classification target (`price_2`) often imbalanced
- Techniques applied:
    - SMOTE (oversampling minority class)
    - Random undersampling (if needed)
    - Class weight adjustment in models

---

6. Model Building

Supervised Learning:

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

Unsupervised Learning:

- **Clustering:** K-Means, DBSCAN, Hierarchical Clustering

Pipeline:
`Data Preprocessing → Feature Scaling → Model Training → Hyperparameter Tuning → Evaluation`

---

7. Model Evaluation

Task	Metrics

Classification	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Regression	MAE, MSE, RMSE, R-squared
Clustering	Silhouette Score, Davies-Bouldin Index

8. Streamlit Web Application

Features:

- Upload CSV or input session values manually
- Real-time prediction of:
  - Customer conversion (classification)
  - Potential revenue (regression)
- Display customer segments (clustering)
- Visualizations:
  - Bar charts, pie charts, histograms
  - Session metrics dashboards

Deployment:

- Runs locally using:

```
streamlit run app.py
```

9. Results

- Classification model predicts conversion with high accuracy (~70% target).
- Regression model estimates revenue per session.
- Clustering segments customers for personalized marketing.
- Streamlit app provides a user-friendly interface for business users.

---

10. Tools & Technologies

- **Programming:** Python, Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn
- **Machine Learning:** Scikit-learn, XGBoost, TensorFlow (for Neural Networks)
- **Web App:** Streamlit
- **Version Control:** GitHub

---

11. Project Deliverables

1. **Source Code:** Preprocessing, feature engineering, modeling, deployment
2. **Streamlit App:** Interactive application for predictions & visualization
3. **Documentation:** This detailed report
4. **Presentation Deck:** Summarized findings and visualizations

---

References

- UCI Machine Learning Repository – Clickstream Data
- Scikit-learn Documentation
- Streamlit Documentation
- Imbalanced-learn (SMOTE) Documentation