E-Commerce Customer Purchase Prediction Project

Project Overview

This project challenges students to predict customer purchasing behaviour using an e-commerce dataset. Students will determine whether a customer will make a purchase (binary classification) based on their browsing behaviour, demographics, and session characteristics. This follows the same structured approach as your Adult Income dataset assignment but with a retail/e-commerce focus.

Dataset Selection

Online Shoppers Purchasing Intention Dataset (UCI Machine Learning Repository)

- **Size:** 12,330 records with 18 features
- **Target:** Revenue (binary: purchase/no purchase)
- **Features:** Administrative pages, Informational pages, Product-related pages, bounce rates, exit rates, page values, special day indicator, month, operating system, browser, region, traffic type, visitor type, weekend indicator
- **Publicly Available:** https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Project Structure

Day 1: Data Loading & Preprocessing (20 marks)

1.1 Library Imports (3 marks)

1.2 Dataset Loading (4 marks)

- Load the Online Shoppers dataset using pandas
- Display first 5 rows and basic dataset information

1.3 Data Exploration (5 marks)

- Display dataset info() and describe()
- Check for missing values and data types
- Show unique values for categorical columns

1.4 Data Preprocessing (8 marks)

- Handle missing values (if any)
- Encode categorical variables (Month, VisitorType, Weekend) using LabelEncoder
- Normalize numerical features using StandardScaler
- Create binary target variable (Revenue: True=1, False=0)

Day 2: Exploratory Data Analysis (20 marks)

2.1 Distribution Analysis (5 marks)

- Create histograms for key numerical features (Administrative, Informational, ProductRelated)
- Show distribution of target variable (Revenue) using bar chart

2.2 Correlation Analysis (5 marks)

- Generate correlation heatmap for numerical features
- Identify strongest correlations with target variable

2.3 Outlier Detection (5 marks)

- Use boxplots to identify outliers in BounceRates, ExitRates, and PageValues
- Document outlier handling strategy

2.4 Categorical Analysis (5 marks)

- Show purchase rate by Month using grouped bar charts
- Analyze purchase behavior by VisitorType (New vs Returning)
- Weekend vs Weekday purchase patterns

Day 3: Feature Engineering (15 marks)

3.1 Feature Selection (5 marks)

- Remove highly correlated features (correlation > 0.8)
- Drop irrelevant features based on domain knowledge

3.2 Feature Creation (5 marks)

- Create new feature: `TotalPages` = Administrative + Informational + ProductRelated

- Create `HighEngagement` binary feature based on PageValues > median
- Create `SessionQuality` categorical feature based on BounceRates and ExitRates

3.3 Dimensionality Reduction (5 marks)

- Apply PCA to reduce dimensions to 10 components
- Create scatter plot of first two PCA components colored by Revenue
- Show explained variance ratio

Day 4: Classification Models (25 marks)

4.1 Model Training and Evaluation

Train and evaluate four models using 80/20 train-test split:

4.1.1 Logistic Regression (6 marks)

- Train model and generate predictions
- Calculate accuracy, precision, recall, F1-score
- Display confusion matrix

4.1.2 Decision Tree (6 marks)

- Train with max_depth=10 to prevent overfitting
- Calculate performance metrics
- Display confusion matrix

4.1.3 Random Forest (6 marks)

- Train with n_estimators=100, random_state=42
- Calculate performance metrics
- Show feature importance plot

4.1.4 Support Vector Machine (7 marks)

- Train SVM with RBF kernel
- Calculate performance metrics
- Display confusion matrix
- Compare training time with other models

Day 5: Model Evaluation & Prediction (20 marks)

5.1 ROC Curve Analysis (8 marks)

- Plot ROC curves for all four models on same graph
- Calculate AUC scores for each model
- Identify best performing model based on AUC

5.2 Model Comparison (7 marks)

- Create comparison table with all performance metrics
- Discuss which model performs best and why
- Analyze trade-offs between precision and recall

5.3 Purchase Prediction (5 marks)

Predict purchase likelihood for this customer profile:

```
sample_customer = {
    'Administrative': 5,
    'Administrative_Duration': 150.0,
    'Informational': 2,
    'Informational_Duration': 45.0,
    'ProductRelated': 15,
    'ProductRelated_Duration': 800.0,
    'BounceRates': 0.02,
    'ExitRates': 0.05,
    'PageValues': 25.0,
    'SpecialDay': 0.0,
    'Month': 'Nov',
    'OperatingSystems': 2,
    'Browser': 2,
    'Region': 1,
    'TrafficType': 2,
    'VisitorType': 'Returning_Visitor',
    'Weekend': False
}
```

Deliverables

1. **Jupyter Notebook (.ipynb)** with complete analysis and well-commented code (your comments not AI)
2. **Number the exact question that you are answering in your code together with all the details as part of your comments**

3. **Also submit a PDF version of your code, after you have run it to demonstrate that it is working.**
4. **File Naming:** StudentNumber_EcommercePrediction_ProjectName
5. **MAKE SURE THAT YOU PUT YOUR STUDENT NUMBER AND FULL NAME AT THE BEGINNING OF YOUR CODE.**
6. **You will be heavily penalised if there are signs of copy and paste from AI tools or there are signs of code reuse from downloaded sources.**

Evaluation Criteria (100 marks total)

- **Data Processing & Preprocessing (20 marks):** Proper data loading, cleaning, encoding, and normalisation
- **Exploratory Data Analysis (20 marks):** Comprehensive visualisations and statistical analysis.
- **Feature Engineering (15 marks):** Creative feature creation and appropriate dimensionality reduction
- **Model Implementation (25 marks):** Correct implementation of all four classification algorithms with proper evaluation
- **Analysis & Interpretation (20 marks):** ROC analysis, model comparison, and business insights

Key Learning Outcomes

Students will demonstrate proficiency in:

- E-commerce data analysis and customer behaviour understanding
- Binary classification problem solving
- Feature engineering for behavioural data
- Model comparison and selection criteria
- Business application of machine learning results

Technical Requirements

- **Python 3.7+** with standard data science libraries
- **Jupyter Notebook** environment
- **Dataset:** Download from UCI ML Repository
- **Evaluation Focus:** Classification accuracy, precision, recall, and AUC metrics