VAUTECH SOLUTIONS — AI INTERNSHIP REPORT

Task 1: Real-World Dataset Sourcing & Problem Framing

Project Title:
Customer Purchase Prediction using Machine Learning

Intern Name:
Lokesh Girish Bharambe

Intern ID:
VT26AI001

Domain:
Artificial Intelligence & Data Science

Mentor:
Vishal Ramkumar Rajbhar

Company:
Vautech Solutions IT Solutions

Topic Name

Real-World Dataset Sourcing & Problem Framing for Customer Purchase Prediction

Topic Description

This task focuses on identifying a real-world business problem and framing it as an Artificial Intelligence problem using a suitable dataset. The e-commerce industry was selected, where the objective is to predict whether a customer will make a purchase based on demographic and behavioral data. A publicly available dataset was sourced and analyzed using Python and pandas to understand its structure, features, and data quality. The problem was defined as a binary classification task, and suitable business impact and success metrics were identified.

Abstract

E-commerce platforms collect large volumes of customer data through browsing behavior, purchase history, discount usage, and loyalty program participation. Despite high website traffic, many customers leave without completing a purchase, resulting in low conversion rates and inefficient marketing strategies.

This project aims to frame this real-world business challenge as an Artificial Intelligence problem. A publicly available e-commerce customer dataset containing 1,500 records was explored using Python and pandas. The dataset includes features such as age, income, number of previous purchases, time spent on the website, loyalty program membership, and discounts availed. The target variable represents whether a customer made a purchase.

The outcome of this task is a clear understanding of the business problem, dataset structure, AI problem definition, and evaluation metrics, forming a strong foundation for future machine learning model development.

1. Introduction

Artificial Intelligence (AI) is transforming industries such as finance, healthcare, education, and e-commerce. In e-commerce, companies track customer activities like browsing duration, purchase frequency, and interaction with discounts. However, many users visit websites without making purchases.

This results in low conversion rates and wasted marketing efforts. Predicting customer purchase behavior helps businesses focus on high-potential customers and optimize marketing strategies. This internship task focuses on converting this real-world challenge into a structured AI problem.

2. Problem Statement

2.1 Business Problem

E-commerce platforms experience high traffic but low sales conversions. Marketing campaigns are often sent to all customers, leading to high costs and low effectiveness.

Business Objective:
To identify customers who are most likely to make a purchase.

2.2 AI Problem Definition

The problem is framed as a binary classification problem, where:

- 1 → Customer will make a purchase
- 0 → Customer will not make a purchase

Predictions are based on demographic and behavioral features such as income, purchase history, time spent on the website, loyalty program membership, and discounts used.

3. Dataset Description

3.1 Data Source

The dataset was sourced from publicly available platforms such as Kaggle and open repositories, commonly used for learning and AI experimentation.

3.2 Dataset Structure

- Rows: 1500
- Columns: 9
- Missing Values: 0

The dataset is clean and suitable for machine learning.

3.3 Feature Description

Feature	Description
Age	Customer’s age
Gender	Encoded gender
AnnualIncome	Yearly income
NumberOfPurchases	Past purchases
ProductCategory	Product category
TimeSpentOnWebsite	Time spent (minutes)
LoyaltyProgram	Membership status
DiscountsAvailed	Discounts used
PurchaseStatus	Target variable

4. Methodology

1. Selected e-commerce as the industry use case
2. Identified a real-world business problem
3. Sourced an open dataset
4. Loaded and explored the dataset using Python
5. Checked structure and missing values
6. Identified the target variable
7. Defined AI problem type
8. Identified success metrics

5. Implementation (Code & Output)

5.1 Tools Used

- Python
- pandas
- CSV Dataset

5.2 Python Code Used

```
import pandas as pd
# Load the dataset
df = pd.read_csv("C:\\Users\\lokes\\Downloads\\customer_purchase_data.csv")
# Display first few rows
df.head()
# Display dataset information
df.info()
# Display statistical summary
df.describe()
```

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

Output

```
PS C:\Users\lokes> python -u "c:\Users\lokes\OneDrive\Desktop
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 1500 non-null   int64
 1   Gender              1500 non-null   int64
 2   AnnualIncome        1500 non-null   float64
 3   NumberOfPurchases   1500 non-null   int64
 4   ProductCategory     1500 non-null   int64
 5   TimeSpentOnWebsite  1500 non-null   float64
 6   LoyaltyProgram      1500 non-null   int64
 7   DiscountsAvailed    1500 non-null   int64
 8   PurchaseStatus      1500 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 105.6 KB
Missing values in each column:
 Age                  0
Gender               0
AnnualIncome         0
NumberOfPurchases    0
ProductCategory      0
TimeSpentOnWebsite   0
LoyaltyProgram       0
DiscountsAvailed     0
PurchaseStatus       0
dtype: int64
```

5.3 Output Explanation

After executing the code:

- The dataset contains 1500 rows and 9 columns
- No missing values were found
- Data types include integer and float values
- Dataset is clean and machine-learning ready
- Target variable identified as PurchaseStatus

This confirms that the dataset is suitable for further modeling without additional preprocessing.

6. Business Impact

- Improved marketing targeting
- Increased conversion rates
- Reduced marketing costs
- Better customer personalization
- Data-driven decision making

7. Success Metrics

- Accuracy – Overall correctness
- Precision – Correct buyer predictions
- Recall – Detection of actual buyers
- F1-Score – Balance between precision and recall

8. Conclusion

This task demonstrates how a real-world business problem can be framed into an AI problem using proper dataset sourcing and analysis. The project builds a strong foundation in AI thinking, data understanding, and business impact analysis. Future tasks will include model training, evaluation, and deployment.

End of Report