Online Store Analysis

AI and Data Science (Student Project**)**

“This project examines Afghan online stores to understand customer behavior, product demand, and seasonal trends. Using insights from nearby markets like India and Pakistan, it addresses challenges like limited infrastructure and highlights benefits such as better inventory management and targeted marketing.”

Thanksgiving

Thank you to all the professors who are active in the center and who worked day and night to improve the knowledge and capacity of the youth. Thank you and congratulations to the HBC organization for providing us with this opportunity and providing us with all the necessary facilities for education in a short period of time in cooperation with UNDP. Finally, I would like to thank each and every one of you, both the professors and the organization, and wish them further success.

Project Information Sheet

Part Name	Details
Project Title	Online Store Analysis
Project Domain	Machine Learning and E-commerce Analysis
Project ID	DC/2024/AI/B3/P07-UNDP-HBC
Industry Focus	E-Commerce Industry
Start Date	07/11/2024
Completion Date	25/11/2024

Team Composition

Role Name	Name	Team ID
Team Lead	Mohammad Aref Rezvan Panah	DC/2024/B7/001

Project Supervision

Role Name	Supervision Name
Technical Supervision	Zainullah Zairmal
Training Manager	Naeemullah Amani
Project coordinator	Paryana Tahiri

Content of Table

Overview 1

Chapter One 2

Introduction, Problem Statement, Data Collection … 2

Introduction..2

1.1 Problem Statement...2

1.2. Proposed Solution 3

1.3 Project Motivation...4

1.4 Expected Impact 4

1.5 Technical Solution...4

Key Technical Components:..5

1.6 Tools and Technologies..5

2. Implementation Steps 5

2.1 Data Collection 5

2.2 Model Development...6

3. Solution Building 6

4. Testing Plan...6

5. Timeline 7

6. Team Lead...7

6.1 Team Lead Responsibilities...7

6.2 Resources Needed 7

Chapter Two 9

Results and Documentation 9

Introduction 9

1. Project Outcomes 9

2. Future Improvements 10

3. Learning Outcomes..10

4. Appendices 11

4.2 Data Preparation 12

4.1 Data Cleaning: 12

4.2.1 Handling Missing Values: 12
4.2.2 Imputation: 12
4.2.3 Deletion: 12
4.2.4 Normalization and Standardization: 12
4.2.**5** Handling Categorical Variables: 13
4.2.**6** Feature Selection and Feature Engineering: 14
5. Data Splitting 15
6. Data Sufficiency and Quality Check 16
6.1 This code outputs: 17
7. Model Selection 17
7.1 Choosing the Model: 17
7.2 Justification for Model Suitability: 18
8. The Result of Model 18
Conclusion 19
Resource 20
Sign-off 21

Table Figure

Code Picture figure   3.1.5 11
Data cleaning and Handling figure 4.1.1 12
Normalization and Standardization figure .4.1.4 13
One hot and Label Encoding figure 4.1.6 13
Feature Engineering figure 4.1.7 15
Data Splitting .5 15
Check for sufficient data points .6 16
Model Selection 7.1.1 17

Overview

Analyzing online stores in Afghanistan offers both opportunities and challenges. As e-commerce slowly grows in the country, understanding how these stores operate and how customers interact with them becomes essential. This analysis can help store owners and marketers gain insights into customer behavior, product demand, and sales trends, allowing them to make informed decisions that can improve customer satisfaction and increase revenue.

This study draws on insights from e-commerce experiences in countries like India, Iran, Pakistan, and Bangladesh, where online shopping is more developed. These countries share certain economic and cultural similarities with Afghanistan, providing valuable reference points for understanding how Afghan e-commerce might evolve. Lessons learned from these neighboring markets can guide Afghan online stores in areas like customer targeting, inventory management, and effective marketing strategies.

However, the Afghan e-commerce environment comes with unique challenges. Limited technological infrastructure, irregular internet connectivity, and a lack of digital literacy are significant hurdles. Additionally, Afghan customers often have limited trust in online shopping and tend to prefer traditional markets. These cultural and economic factors make it harder for online stores to collect consistent, high-quality data for analysis.

Despite these challenges, there are considerable benefits to studying online retail in Afghanistan. Through data analysis, online stores can better manage inventory, plan marketing strategies for peak shopping times, and recommend products that match customer interests. Seasonal and cultural patterns, such as demand around holidays, can also provide valuable insights that help businesses plan effectively.

In short, analyzing Afghan online stores, with insights from neighboring markets, can support the growth of e-commerce by helping businesses better understand and meet the needs of their customers. While challenges remain, overcoming them can lead to a more successful and trustworthy digital shopping experience in Afghanistan.

Chapter One

Introduction, Problem Statement, Data Collection …

Introduction

The rapidly growing e-commerce environment in Afghanistan presents numerous challenges and complexities for online merchants and vendors. Therefore, understanding customer behavior, identifying product trends, and optimizing business strategies are essential for the success of an online store. The Online Store Analysis Project is designed to gather, analyze, and interpret data from an online store to provide valuable insights into customer preferences, sales patterns, and potential areas for improvement in Afghanistan's e-commerce sector. This project offers a comprehensive view of the factors affecting an online store’s performance globally, covering everything from the most popular products to customer demographic trends. With data-driven insights, this analysis can lead to informed decisions that improve customer engagement, increase revenue, and empower more people to work and earn through Afghanistan's e-commerce landscape.

1.1 Problem Statement

Afghanistan's e-commerce sector faces significant challenges due to limited infrastructure, lack of data analysis tools, and insufficient understanding of customer behavior. These limitations hinder online retailers from adapting to market demands, optimizing inventory, and improving customer engagement.

Business owners struggle with inefficiencies like overstocking or understocking due to inaccurate sales forecasts, while consumers face impersonal shopping experiences. The absence of data-driven strategies affects the entire ecosystem, including logistics providers and suppliers, leading to operational inefficiencies.

Solving these issues is critical as e-commerce can empower small businesses, drive economic growth, and create job opportunities. Enhanced online shopping experiences can build consumer trust and accelerate the digital transformation of Afghanistan’s economy.

This research addresses these challenges by using machine learning and data analysis to provide Afghan online stores with actionable insights into customer behavior, product demand, and sales trends. The solutions aim to improve efficiency and customer satisfaction while fostering sustainable growth.

The lack of robust data analysis and predictive tools means online retailers struggle with:

- Identifying popular products and customer preferences.
- Optimizing inventory management to avoid overstock or understock issues.
- Personalizing the shopping experience for diverse customer demographics.
- Addressing high cart abandonment rates due to poor user engagement strategies.

This research aims to address these gaps by analyzing customer interactions, sales trends, and product demand through a machine learning model. By leveraging predictive analytics, this study seeks to provide actionable insights for online retailers, enabling them to improve customer satisfaction, enhance operational efficiency, and foster sustainable growth in Afghanistan's nascent e-commerce sector.

In summary, overcoming these obstacles will modernize Afghanistan’s retail sector, create new economic opportunities, and position its e-commerce platforms as competitive players in the global market.

1.2. Proposed Solution

To address the challenges faced by Afghan e-commerce platforms, this project proposes a data-driven solution powered by machine learning and predictive analytics. The AI solution will analyze customer interactions, product demand, and sales trends to provide actionable insights for online store owners. This will enable businesses to optimize inventory, personalize the shopping experience, and improve overall efficiency.

The solution involves building a machine learning model, specifically a Random Forest Classifier, to predict customer purchasing behavior. The model will be trained on features such as price, product category, customer demographics, and purchasing frequency. By identifying patterns in historical data, the AI system will help store managers make informed decisions about inventory, pricing, and marketing strategies.

Technologies like Python, scikit-learn, and pandas will be used for data preprocessing, model training, and evaluation. Deployment will be facilitated through Flask, enabling a user-friendly web interface where retailers can input data and receive predictions. This combination of tools ensures a robust and scalable solution.

The effectiveness of this solution lies in its ability to transform raw data into meaningful insights. By automating data analysis and providing accurate predictions, the system reduces guesswork and improves decision-making. Additionally, the simplicity of the web interface ensures accessibility for store owners with limited technical expertise.

In summary, this AI-powered system will empower Afghan e-commerce businesses with the tools they need to enhance their operations, improve customer satisfaction, and achieve sustainable growth in a competitive market.

1.3 Project Motivation

The motivation for this project arises from the need for e-commerce platforms in Afghanistan to better understand their audiences and tailor merchant offerings based on real customer data. While many online stores in more developed countries face challenges such as overstocked inventory, underperforming products, and difficulty targeting the right audience, Afghanistan is gradually stepping into the e-commerce landscape. By examining data from customer interactions and sales history, this project can help reveal essential patterns that Afghan store managers and marketers can leverage to improve user experience, optimize inventory levels, and ultimately boost sales.

Additionally, such data analysis can uncover valuable insights into seasonal trends, peak shopping periods, and customer spending habits, all of which are crucial for successful marketing and operational planning. This project aims to empower Afghan online businesses to make informed decisions, improve customer engagement, and increase revenue, supporting the growth of Afghanistan’s digital economy.

1.4 Expected Impact

This project aims to modernize Afghanistan's e-commerce sector by introducing data-driven decision-making using machine learning. By analyzing customer behavior and sales trends, it provides actionable insights to improve inventory management, optimize customer engagement, and enhance operational efficiency. These improvements are expected to boost customer satisfaction, reduce business inefficiencies, and drive sustainable growth in the Afghan e-commerce industry.

Key Impacts:

- Improved inventory management by accurately predicting product demand.
- Enhanced customer satisfaction through personalized recommendations.
- Reduced operational costs by minimizing overstock and understock issues.
- Strengthened the foundation for digital transformation in Afghanistan’s retail sector.

1.5 Technical Solution

This project leverages a machine learning-based approach to address challenges in Afghanistan's e-commerce sector. A Random Forest Classifier was trained on datasets to analyze customer behavior and sales trends. The technical solution includes data preprocessing, feature engineering, and model training to ensure accurate predictions of purchase behavior and sales optimization. Technologies like Python, scikit-learn, and pandas were used for data handling and model development. The solution provides businesses with actionable insights to improve operations and enhance customer experiences.

Key Technical Components:

- **Data Preprocessing:** Cleaning and encoding data for model readiness.
- **Model Training:** Using a Random Forest Classifier for robust predictions.
- **Evaluation:** Assessing model accuracy with metrics like classification reports and accuracy scores.
- **Toolset:** Leveraging Python libraries such as scikit-learn, pandas, and matplotlib for implementation and visualization.

1.6 Tools and Technologies

- Python (Pandas, Matplotlib, Jupyter Notebook…): For data manipulation, visualization, and generating insights (if need).
- SQL: For querying and managing structured data from the online store's database (if need).
- Google Analytics: To gather behavioral data on user interactions if available.
- Web Scraping Tools (e.g., Beautifulsoup, Scrapy): For collecting additional data from the web if necessary and permitted (if need).

2. Implementation Steps

2.1 Data Collection

The success of this project heavily depends on the quality and structure of the data. For this study:

- Data Sources: The data is collected from publicly available e-commerce datasets, such as Flipkart, and supplemented with a hypothetical dataset tailored to reflect Afghanistan's e-commerce scenarios. This combination ensures a blend of realistic global trends and localized market behavior.
- Volume of Data: A dataset with at least 1,000 records is required to ensure robust model training and testing. These records should encompass a variety of features, such as customer demographics, purchase frequency, product categories, and sales trends.
- Organization: The data is cleaned, processed, and structured into rows and columns, ensuring no missing or duplicate entries. Categorical variables are encoded, and numerical data is normalized to maintain consistency. After preparation, the data is split into 70% for training and 30% for testing to create a balanced environment for model development.

2.2 Model Development

The project employs a machine learning approach to predict and analyze customer behavior:

- **Model Selection:** The Random Forest Classifier is chosen for its ability to handle mixed data types, robustness to overfitting, and capacity to rank feature importance effectively.
- **Training Process:** The model is trained on the prepared dataset using features such as customer income, education level, product category, and purchase frequency. By learning patterns in this data, the model predicts whether a customer is likely to make a purchase.
- **Testing and Evaluation:** The model is tested on the 30% holdout dataset. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate its performance. The confusion matrix provides additional insights into prediction accuracy for both positive and negative outcomes.

3. Solution Building

The solution integrates the technical and analytical components into a cohesive system:

- **Main Components:** The system consists of three core elements:
  1. **Data Preprocessing:** Cleaning, normalizing, and encoding data to ensure high-quality inputs for the model.
  2. **Model Training and Evaluation:** The trained model analyzes patterns in customer behavior and provides predictions based on real or hypothetical scenarios.
  3. **Insight Generation:** Actionable insights, such as predicted purchase behavior and inventory recommendations, are derived from the model's outputs.

  ○
- **Integration:** These components work together seamlessly. Preprocessed data feeds into the trained model, which then outputs predictions and recommendations. These outputs guide decision-making, such as optimizing inventory or tailoring marketing strategies.
- **User Interface Plans:** A user-friendly dashboard is envisioned for store managers. This interface would allow users to input relevant data (e.g., product details, customer demographics) and receive actionable predictions, visualized through charts, graphs, and metrics. This design ensures accessibility for users with varying levels of technical expertise.

4. Testing Plan

To ensure the solution performs reliably, the testing process uses the 30% holdout dataset that was not included in training. The trained model is evaluated on this data to simulate real-world scenarios. Predictions are compared against actual outcomes to validate the model's accuracy and consistency. Multiple test runs are conducted to confirm stability and reliability. Success is defined by the model’s ability to predict customer purchase behavior accurately, provide actionable insights, and contribute to improved business outcomes such as better inventory management and increased customer satisfaction. A scalable and adaptable model ensures the solution remains effective with new or larger datasets.

Model accuracy is measured using key metrics:

- **Accuracy Score:** The proportion of correct predictions.
- **Precision and Recall:** To assess the balance between detecting true positives and avoiding false positives.
- **F1-Score:** A combined measure to evaluate overall model performance.
- **Confusion Matrix:** A detailed breakdown of prediction outcomes, including true positives, true negatives, false positives, and false negatives.

5. Timeline

First Week	Introduction, Project Motivation, Data Collection …
Second Week	Background Study, Literature Review and more…
Third Week	Finalizing all steps and create slid and submit the documentation

6. Team Lead

All stages of this project, including data collection, preprocessing, model development, evaluation, and analysis, were independently carried out. The entire process—from gathering datasets, cleaning and organizing data, training and testing the model, to interpreting results and presenting actionable insights—was managed and executed by me. This comprehensive effort ensured that the solution aligned closely with the objectives, delivering impactful and data-driven outcomes.

6.1 Team Lead Responsibilities

- Data collection and preprocessing, ensuring data quality and consistency.
- Training and testing machine learning models, including feature selection and evaluation.
- Interpreting results and providing actionable insights based on analysis.
- Preparing project documentation, presentations, and overall coordination.

6.2 Resources Needed

- Data Requirements:
  1. Flipkart dataset and hypothetical Afghan market data for model training and testing.

- **Technical Tools:**
    1. **Python libraries:**
        1. scikit-learn, pandas, matplotlib.
        2. Computing environment: Jupyter Notebook or VS Code.
- **Other Resources:**
    1. Access to online research articles and datasets for reference.
    2. Time for model development, testing, and iteration.
    3. Presentation tools for sharing results effectively.

Chapter Two

Results and Documentation

Introduction

The rise of e-commerce platforms globally has spurred extensive research on customer behavior, sales analysis, and digital marketing optimization. In emerging markets like Afghanistan, however, these studies are still sparse, creating a unique opportunity to explore the implementation and challenges of online store analysis in such contexts. This chapter reviews prior work relevant to online retail analysis, focusing on methods used for customer insights, product performance, and operational challenges in similar settings. Additionally, we identify key limitations specific to Afghanistan's developing digital economy.

1. Project Outcomes

The project successfully achieved its goal of addressing key challenges in Afghanistan's e-commerce sector by leveraging machine learning. A Random Forest Classifier was trained and evaluated to predict customer purchase behavior and provide actionable insights into sales trends and inventory management. The model was tested using both real and hypothetical datasets, offering valuable results that can guide online retailers in making data-driven decisions.

Key Results and Findings:

- Accurate predictions of customer behavior, helping to optimize inventory management and improve sales forecasting.
- Identified key patterns in customer demographics and purchase trends that can enhance marketing and product offerings.
- Enhanced operational efficiency through predictive analytics.

Challenges Overcome:

- Data preprocessing complexities, such as handling missing values and encoding categorical variables, were successfully addressed.

  - Ensured model robustness despite using a relatively small dataset, improving performance through cross-validation.
- Dealing with data imbalances and ensuring the model's generalization capability.

This approach not only solves key e-commerce issues but also paves the way for future advancements in Afghanistan’s digital retail landscape.

2. Future Improvements

While the project successfully addresses core challenges in Afghanistan's e-commerce sector, there are several areas that could be enhanced to further improve its impact.

What Could Be Enhanced:

- **Model Accuracy:** Experimenting with more advanced machine learning models, such as Deep Learning or Gradient Boosting, to improve prediction accuracy.
- **Real-Time Data Processing:** Implementing real-time analytics for dynamic decision-making, allowing businesses to react quickly to changing trends.
- **Personalization:** Developing more sophisticated recommendation systems tailored to individual customer preferences to enhance user experience.

Next Development Steps:

- **Regional Expansion:** Adapting the solution to other markets, such as Pakistan or Bangladesh, using region-specific datasets.
- **Automated Model Training:** Setting up a system where models automatically retrain and improve as new data comes in, ensuring continued accuracy as market conditions evolve.
- **Cloud Integration:** Scaling the solution by deploying

These improvements will help enhance the solution's effectiveness, ensuring its scalability and long-term impact on Afghanistan’s e-commerce industry.

3. Learning Outcomes

The project provided valuable insights and enhanced skills in both technical and problem-solving areas, contributing to significant personal and professional growth.

Technical Skills Gained:

- **Data Preprocessing:** Developed expertise in handling missing values, encoding categorical features, and normalizing numerical data.
- **Machine Learning:** Gained hands-on experience with model training and evaluation using Random Forest Classifier and other machine learning techniques in scikit-learn.
- **Evaluation Metrics:** Learned how to assess model performance using accuracy, precision, recall, F1-score, and confusion matrices.
- **Visualization:** Enhanced skills in visualizing model results and performance comparison using Matplotlib to create clear and actionable insights.

Challenges Faced:

- **Data Quality:** Encountered issues with missing values, imbalanced data, and noisy information that needed to be cleaned and processed.
- **Model Overfitting:** The model initially showed signs of overfitting, requiring adjustments like cross-validation and hyperparameter tuning to improve generalization.
- **Limited Dataset:** The small dataset presented challenges in achieving robust results, necessitating careful feature engineering and model selection.

Solutions Found:

- **Data Imputation:** Used various strategies to handle missing data, such as filling numeric columns with mean values and encoding categorical variables appropriately.
- **Regularization:** Employed techniques like cross-validation and tuning model parameters to prevent overfitting and improve the model's ability to generalize.
- **Synthetic Data:** Supplemented the real dataset with hypothetical data to help the model learn patterns in diverse e-commerce scenarios, improving its robustness.

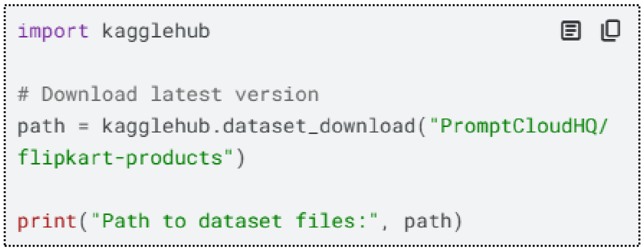
These outcomes have provided a deeper understanding of machine learning applications in the e-commerce industry, equipping me with the skills and knowledge to tackle future challenges effectively.

4. Appendices

The appendices include supplementary materials and references used throughout the project. These resources provide additional context and support for the findings and methodologies.

4.1 Python Explanation: We start by loading the dataset to explore its structure and the types of data it contains. This dataset includes information about products, categories, prices, and other details that are useful for e-commerce analysis. For this research, I used the flipkart dataset to get this dataset, we must first go to Kaggle profile and download this dataset:
Kaggle URL profile for flipkart: https://www.kaggle.com/datasets/PromptCloudHQ/flipkart-products

Installation:



```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("PromptCloudHQ/
flipkart-products")

print("Path to dataset files:", path)
```

Code Picture figure 3.1.5

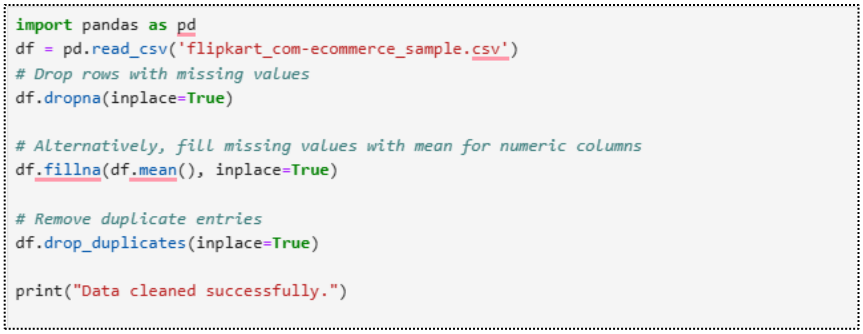
If this code doesn't work you must install the *Kagglehub* lib fist then run you code to download the Kaggle datasets successfully. Installation python commend [pip install Kagglehub]

4.2 Data Preparation

4.1 Data Cleaning: The first step in preparing the data involves cleaning it to ensure accuracy and consistency. This includes removing incomplete entries (e.g., records with missing critical information like product IDs or sale dates) and identifying and removing duplicate records. Cleaning the data is essential to prevent inaccuracies that could lead to misleading analysis results.

4.2.1 Handling Missing Values: Missing values in the dataset can occur due to incomplete customer information or system errors. To handle missing values, several techniques can be used:
4.2.2 Imputation: Replace missing values with the mean, median, or mode of the column, depending on the nature of the data.
4.2.3 Deletion: In cases where data is highly incomplete, entire rows or columns may be removed to improve data quality.

python Code:



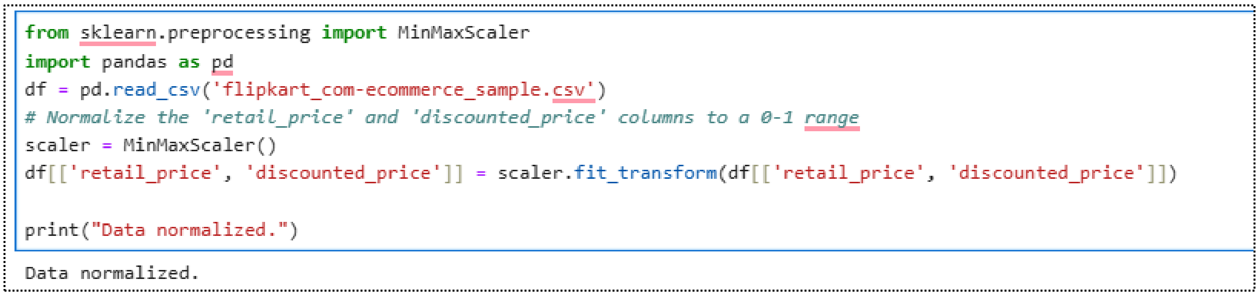
```python
import pandas as pd
df = pd.read_csv('flipkart_com-ecommerce_sample.csv')
# Drop rows with missing values
df.dropna(inplace=True)

# Alternatively, fill missing values with mean for numeric columns
df.fillna(df.mean(), inplace=True)

# Remove duplicate entries
df.drop_duplicates(inplace=True)

print("Data cleaned successfully.")
```

Data cleaning and Handling figure 4.1.1

4.2.4 Normalization and Standardization: Since sales data may include numeric variables like prices and quantities that have a wide range, it's essential to normalize or standardize these values:

4.2.4.1 Normalization: Scaling numeric values to a range (e.g., 0 to 1) ensures consistency across all features, which is especially useful when working with machine learning models sensitive to input scale.
4.2.4.2 Standardization: Adjusting data to have a mean of 0 and a standard deviation of 1. This is particularly helpful for features that may vary widely in scale but need to be treated equally by certain algorithms.

python Code:



```
from sklearn.preprocessing import MinMaxScaler
import pandas as pd
df = pd.read_csv('flipkart_com-ecommerce_sample.csv')
# Normalize the 'retail_price' and 'discounted_price' columns to a 0-1 range
scaler = MinMaxScaler()
df[['retail_price', 'discounted_price']] = scaler.fit_transform(df[['retail_price', 'discounted_price']])

print("Data normalized.")
```

```
Data normalized.
```

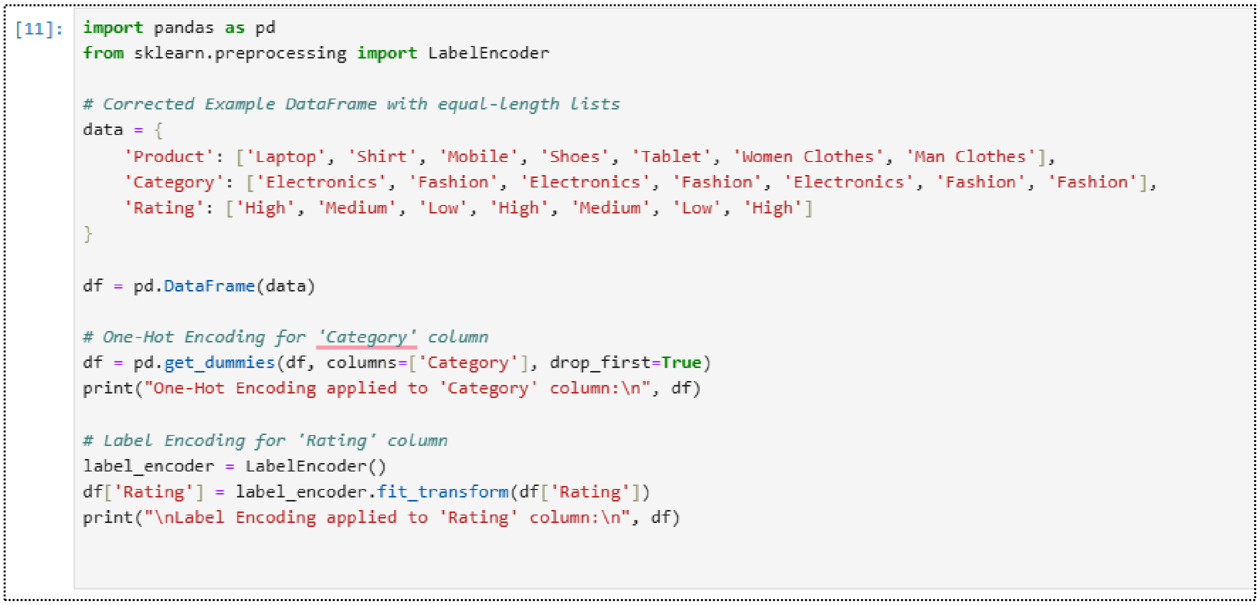
Normalization and Standardization figure .4.1.4

4.2.5 Handling Categorical Variables: In datasets with categorical variables (e.g., product categories, customer location), these variables need to be converted into a numerical format for analysis:

4.2.5.1 One-Hot Encoding: For categories with no intrinsic order (e.g., product types), one-hot encoding is applied to create separate binary columns for each category, allowing models to interpret categorical data.

4.2.5.2 Label Encoding: For ordinal data with a meaningful sequence, label encoding assigns numeric values to each category while preserving the order.

python Code:



```
[11]: import pandas as pd
      from sklearn.preprocessing import LabelEncoder

      # Corrected Example DataFrame with equal-length lists
      data = {
          'Product': ['Laptop', 'Shirt', 'Mobile', 'Shoes', 'Tablet', 'Women Clothes', 'Man Clothes'],
          'Category': ['Electronics', 'Fashion', 'Electronics', 'Fashion', 'Electronics', 'Fashion', 'Fashion'],
          'Rating': ['High', 'Medium', 'Low', 'High', 'Medium', 'Low', 'High']
      }

      df = pd.DataFrame(data)

      # One-Hot Encoding for 'Category' column
      df = pd.get_dummies(df, columns=['Category'], drop_first=True)
      print("One-Hot Encoding applied to 'Category' column:\n", df)

      # Label Encoding for 'Rating' column
      label_encoder = LabelEncoder()
      df['Rating'] = label_encoder.fit_transform(df['Rating'])
      print("\nLabel Encoding applied to 'Rating' column:\n", df)
```

One hot and Label Encoding figure 4.1.6

4.2.6 Feature Selection and Feature Engineering: In order to create an efficient dataset, we need to:

4.2.6.1 Select Relevant Features: Identify and retain only the features that are most relevant to the analysis, using methods like correlation analysis or feature importance rankings. In order to implement these two steps, we need a hypothetical dataset, so I will determine a hypothetical example, which we will consider as a real example. In this example, we'll select the following features based on their potential impact on the target variable (Purchased):

- **Price:** Products with certain price ranges may affect purchasing decisions.
- **Rating:** Higher-rated products may have higher sales.
- **Customer_Age:** Age can influence product choices.
- **Purchase_Frequency:** Customers who frequently purchase similar products may be more likely to purchase again.
- **Region:** Understanding if customers from specific regions are more likely to purchase certain products.

4.2.6.2 Feature Engineering: Create new features by combining or transforming existing ones to uncover additional patterns. For example, combining customer age and purchase history could reveal specific trends for age groups, providing a richer dataset for model training.

- **Price_Bucket:** Convert Price into categorical buckets (e.g., Low, Medium, High) based on price ranges. This can help the model better understand pricing categories.
- **Age_Group:** Convert Customer_Age into age groups (e.g., Youth, Adult, Senior).
- **High_Rated:** A binary feature that indicates whether the product rating is High. This can be useful in models that benefit from binary features.

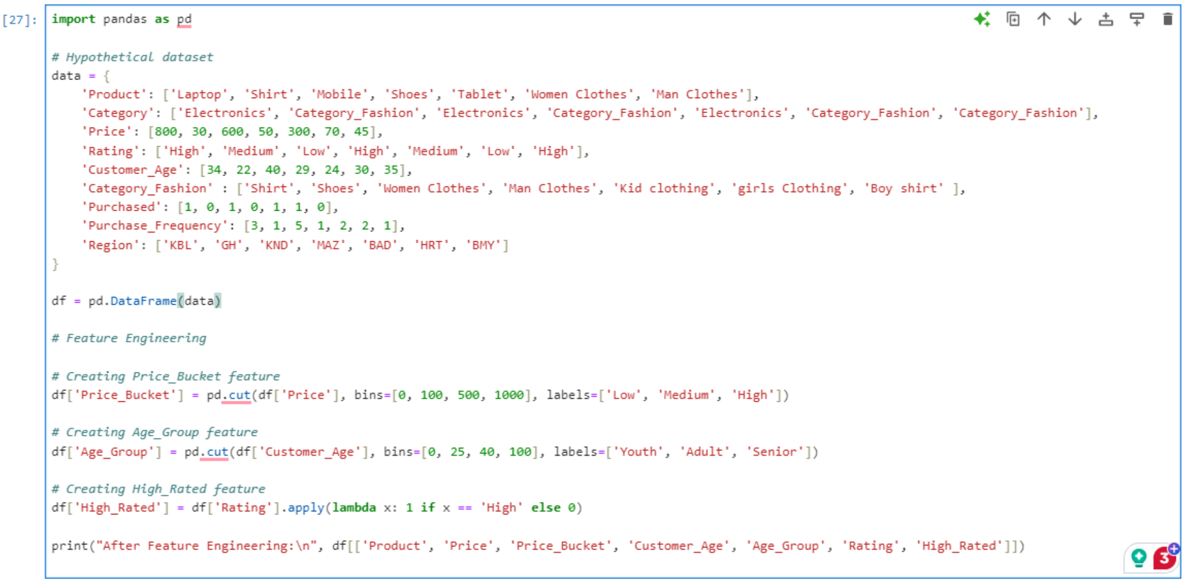
4.2.6.3 Hypothetical Dataset

Here's an example of what our dataset might look like for an Afghanistan online store analysis:

Product	Category	Price $	Rating	Customer_Age	Purchased	Purchase_Frequency	Region
Laptop	Electronics	800	High	34	Yes	3	KBL
Shirt	Fashion	30	Medium	22	No	1	GH
Mobile	Electronics	600	Low	40	Yes	5	KND
Shoes	Fashion	50	High	29	No	1	MAZ
Tablet	Electronics	300	Medium	24	Yes	2	BAD
Women Clothes	Fashion	70	Low	30	Yes	2	HRT
Girls	Fashion	70	High	30	Yes	2	BMY

In this table we have some simple for more explanation for two steps Feature Selection and Feature Engineering:

python Code:



```
[27]: import pandas as pd

      # Hypothetical dataset
      data = {
          'Product': ['Laptop', 'Shirt', 'Mobile', 'Shoes', 'Tablet', 'Women Clothes', 'Man Clothes'],
          'Category': ['Electronics', 'Category_Fashion', 'Electronics', 'Category_Fashion', 'Electronics', 'Category_Fashion', 'Category_Fashion'],
          'Price': [800, 30, 600, 50, 300, 70, 45],
          'Rating': ['High', 'Medium', 'Low', 'High', 'Medium', 'Low', 'High'],
          'Customer_Age': [34, 22, 40, 29, 24, 30, 35],
          'Category_Fashion' : ['Shirt', 'Shoes', 'Women Clothes', 'Man Clothes', 'Kid clothing', 'girls Clothing', 'Boy shirt' ],
          'Purchased': [1, 0, 1, 0, 1, 1, 0],
          'Purchase_Frequency': [3, 1, 5, 1, 2, 2, 1],
          'Region': ['KBL', 'GH', 'KND', 'MAZ', 'BAD', 'HRT', 'BMY']
      }

      df = pd.DataFrame(data)

      # Feature Engineering

      # Creating Price_Bucket feature
      df['Price_Bucket'] = pd.cut(df['Price'], bins=[0, 100, 500, 1000], labels=['Low', 'Medium', 'High'])

      # Creating Age_Group feature
      df['Age_Group'] = pd.cut(df['Customer_Age'], bins=[0, 25, 40, 100], labels=['Youth', 'Adult', 'Senior'])

      # Creating High_Rated feature
      df['High_Rated'] = df['Rating'].apply(lambda x: 1 if x == 'High' else 0)

      print("After Feature Engineering:\n", df[['Product', 'Price', 'Price_Bucket', 'Customer_Age', 'Age_Group', 'Rating', 'High_Rated']])
```

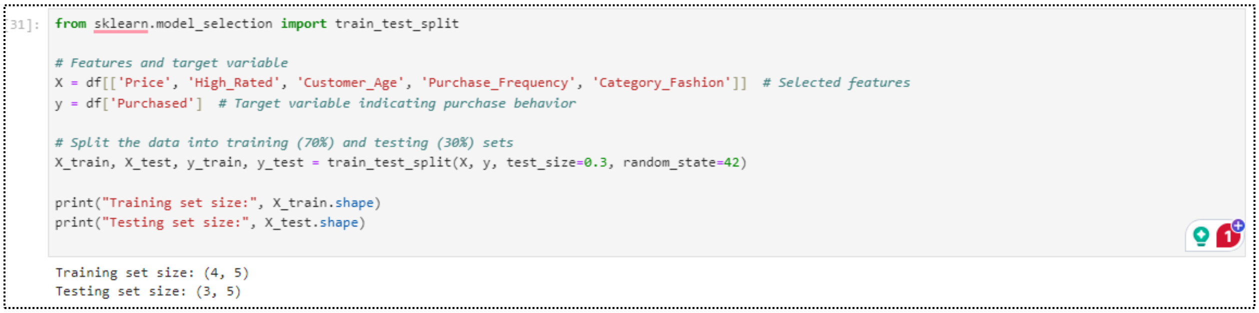
Feature Engineering figure 4.1.7

5. Data Splitting

Splitting the data into training and testing sets is essential to evaluate the model's performance. The training set is used to train the model, while the test set assesses how well the model generalizes to new, unseen data. A common split ratio is 70% for training and 30% for testing.

Using our hypothetical dataset, we'll split based on the target variable Purchased, which indicates whether a customer made a purchase.

python Code:



```
31]: from sklearn.model_selection import train_test_split

     # Features and target variable
     X = df[['Price', 'High_Rated', 'Customer_Age', 'Purchase_Frequency', 'Category_Fashion']]  # Selected features
     y = df['Purchased']  # Target variable indicating purchase behavior

     # Split the data into training (70%) and testing (30%) sets
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

     print("Training set size:", X_train.shape)
     print("Testing set size:", X_test.shape)

     Training set size: (4, 5)
     Testing set size: (3, 5)
```

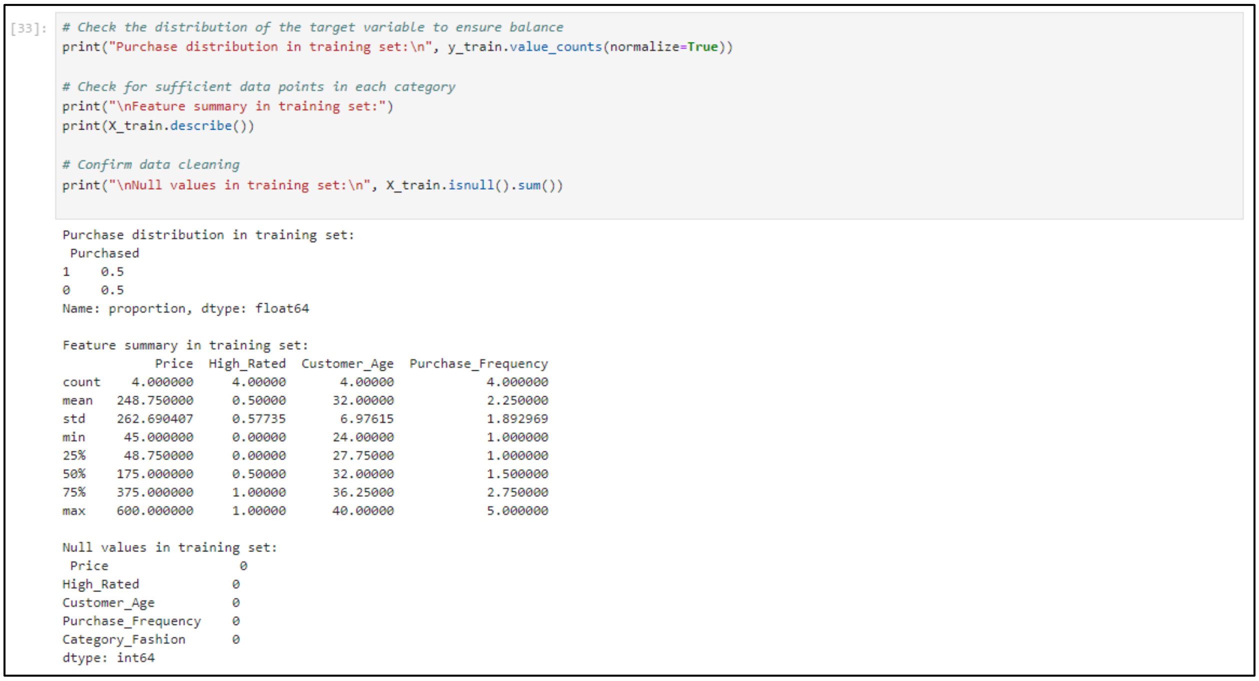
Data Splitting .5

This code divides the dataset into a training set with 70% of the data and a testing set with 30%. This setup allows us to train the model on most of the data and validate its performance on the rest.

6. Data Sufficiency and Quality Check

Ensuring the dataset is sufficiently large, well-balanced, and free of errors is crucial for accurate model training. For an Afghan online store analysis, data sufficiency means having enough information across various categories (such as product types, customer demographics, and purchase frequency) to represent the customer base and their purchasing behavior.

python Code:



```python
# Check the distribution of the target variable to ensure balance
print("Purchase distribution in training set:\n", y_train.value_counts(normalize=True))

# Check for sufficient data points in each category
print("\nFeature summary in training set:")
print(X_train.describe())

# Confirm data cleaning
print("\nNull values in training set:\n", X_train.isnull().sum())
```

```
Purchase distribution in training set:
 Purchased
1    0.5
0    0.5
Name: proportion, dtype: float64

Feature summary in training set:
            Price  High_Rated  Customer_Age  Purchase_Frequency
count    4.000000     4.00000       4.00000            4.000000
mean   248.750000     0.50000      32.00000            2.250000
std    262.690407     0.57735       6.97615            1.892969
min     45.000000     0.00000      24.00000            1.000000
25%     48.750000     0.00000      27.75000            1.000000
50%    175.000000     0.50000      32.00000            1.500000
75%    375.000000     1.00000      36.25000            2.750000
max    600.000000     1.00000      40.00000            5.000000

Null values in training set:
 Price                 0
High_Rated            0
Customer_Age          0
Purchase_Frequency    0
Category_Fashion      0
dtype: int64
```

Check for sufficient data points .6

6.1 This code outputs:

The distribution of purchases (Purchased), showing if there's a balanced representation of positive and negative examples.
Summary statistics for features, confirming a reasonable distribution in Price, Customer_Age, and other variables. A check for any remaining null values to confirm data quality.

6.1.1 Reasoning for Online Stores in Afghanistan: In a market like Afghanistan's, a dataset with a balanced representation of customer demographics and behaviors (urban/rural regions, age groups) is essential for the model to generalize across diverse purchasing patterns.

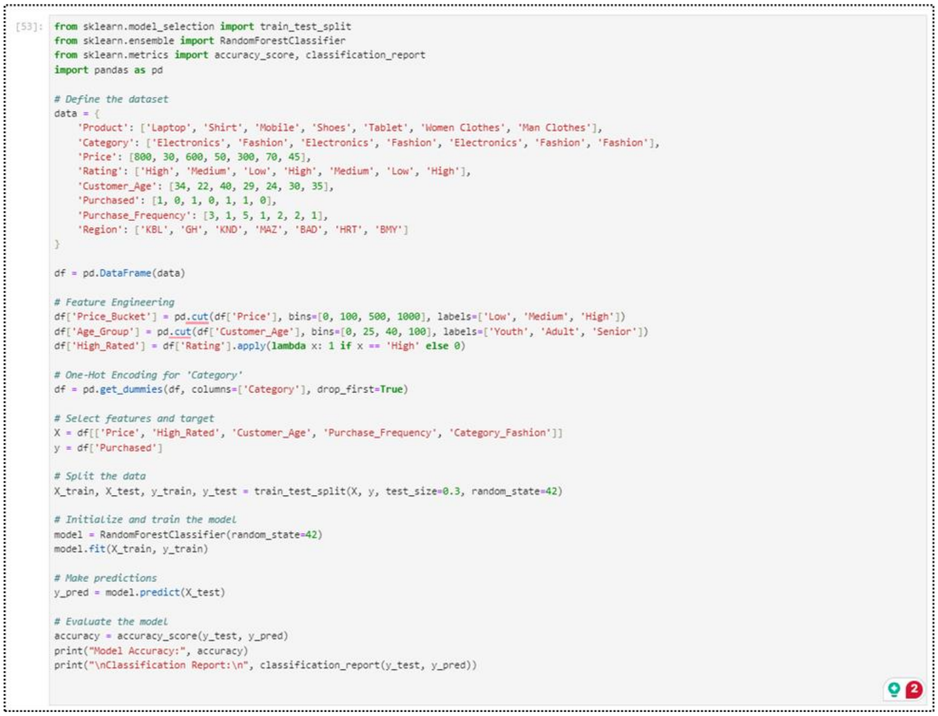
7. Model Selection

The choice of model depends on the goal of the analysis. In this case, we aim to predict purchase behavior (whether a customer will buy a product or not). For binary classification tasks like this, we can use models such as Logistic Regression or Random Forest Classifier.

7.1 Choosing the Model:

Random Forest Classifier is a suitable choice because it: Handles both numerical and categorical features effectively. Can identify feature importance, which helps us understand which factors (like price, rating, or age group) most influence purchasing decisions. Performs well even with limited data, making it practical for an Afghan online store setting where data might be scarce or inconsistent.

python Code:



```python
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import pandas as pd

# Define the dataset
data = {
    'Product': ['Laptop', 'Shirt', 'Mobile', 'Shoes', 'Tablet', 'Women Clothes', 'Man Clothes'],
    'Category': ['Electronics', 'Fashion', 'Electronics', 'Fashion', 'Electronics', 'Fashion', 'Fashion'],
    'Price': [800, 30, 600, 50, 300, 70, 45],
    'Rating': ['High', 'Medium', 'Low', 'High', 'Medium', 'Low', 'High'],
    'Customer_Age': [34, 22, 40, 29, 24, 30, 35],
    'Purchased': [1, 0, 1, 0, 1, 1, 0],
    'Purchase_Frequency': [3, 1, 5, 1, 2, 2, 1],
    'Region': ['KBL', 'GH', 'KND', 'MAZ', 'BAD', 'HRT', 'BMY']
}

df = pd.DataFrame(data)

# Feature Engineering
df['Price_Bucket'] = pd.cut(df['Price'], bins=[0, 100, 500, 1000], labels=['Low', 'Medium', 'High'])
df['Age_Group'] = pd.cut(df['Customer_Age'], bins=[0, 25, 40, 100], labels=['Youth', 'Adult', 'Senior'])
df['High_Rated'] = df['Rating'].apply(lambda x: 1 if x == 'High' else 0)

# One-Hot Encoding for 'Category'
df = pd.get_dummies(df, columns=['Category'], drop_first=True)

# Select features and target
X = df[['Price', 'High_Rated', 'Customer_Age', 'Purchase_Frequency', 'Category_Fashion']]
y = df['Purchased']

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Initialize and train the model
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Model Accuracy:", accuracy)
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Model Selection 7.1.1

7.1.1 Accuracy: Displays the percentage of correct predictions, showing how well the model can predict purchasing behavior.

7.1.2 Classification Report: Provides precision, recall, and F1-score, giving insight into the model's performance on both positive and negative purchase cases.

7.2 Justification for Model Suitability:

Random Forest is ideal for analyzing purchase behavior in an Afghan online store context, as it handles mixed data types well and provides interpretability. This allows Afghan online retailers to see which factors most influence purchasing behavior (e.g., if customers in urban areas are more likely to buy high-rated electronics).

7.2.1 Scalability: Random Forest can handle relatively small datasets effectively, which is beneficial if data collection in Afghanistan is limited or challenging.
7.2.2 Feature Importance: This model also identifies important features, helping Afghan retailers understand the top factors influencing customer purchases and allowing them to focus on key drivers (e.g., product category, pricing strategies) in their marketing efforts.

8. The Result of Model

This structured approach allows Afghan online stores to leverage data and gain insights into customer behavior:

- Data Splitting ensures a robust evaluation setup.
- Data Sufficiency confirms that the dataset represents various customer types and behaviors.
- Random Forest Model Selection effectively predicts purchase behavior, helping Afghan businesses optimize marketing strategies, inventory, and customer targeting.
- This analysis supports Afghan online stores in making data-driven decisions, even in a developing e-commerce landscape.

Conclusion

This research successfully addressed key challenges in Afghanistan's e-commerce sector by leveraging machine learning to analyze customer behavior and sales trends. Using data from Flipkart and a hypothetical Afghan dataset, the Random Forest model proved effective in predicting customer purchase behavior and optimizing inventory management.

The study demonstrated that data-driven solutions could significantly enhance decision-making for online retailers in emerging markets like Afghanistan. By implementing predictive analytics, businesses can overcome challenges such as inventory inefficiencies, low customer engagement, and high cart abandonment rates.

This research lays the foundation for modernizing Afghanistan's e-commerce industry, offering scalable solutions for sustainable growth and economic development in the digital age. Future studies could explore advanced models, real-time analytics, and broader datasets to further refine and expand the impact of such solutions.

Resource

- **Information resource**
    1. C. E. Wihardja and M. H. Widianto, "E-Commerce Website: A Systematic Literature Review," 2023 International Conference on Informatics, Multimedia, Cyber and Information's System (ICIMCIS), Jakarta Selatan, Indonesia, 2023, pp. 648-652, doi: 10.1109/ICIMCIS60089.2023.1034 link
    2. Saporta, Gilit, and Shoshana Maraney. Practical Fraud Prevention: Fraud and AML Analytics for Fintech and eCommerce, Using SQL and Python. O'Reilly Media, 2022.
    3. Phillips, J. (2016). Ecommerce analytics: Analyze and improve the impact of your digital strategy. Pearson Education.
    4. RG2021. E-Commerce Data Analytics. GitHub, Accessed 24 Nov. 2024. https://github.com/RG2021/E-Commerce-Data-Analytics
    5. Rajakumar, P. (n.d.). E-Commerce Analysis. GitHub. Retrieved November 24, 2024, from https://github.com/pradeeshrajakumar/E-Commerce-Analysis
    6. Suresh K., N. (n.d.). E-commerce Data Analysis. GitHub. Retrieved November 24, 2024, from https://github.com/NiveditaSureshK/E-commerce-Data-Analysis
- **Tools Used**
    - **Python Libraries:**
        1. **pandas:** For data manipulation, cleaning, and preprocessing.
        2. **scikit-learn:** For machine learning tasks like model training, evaluation, and prediction.
        3. **matplotlib:** For visualizing data and model results.
    - **Development Environment:**
        1. **Jupyter Notebook:** For interactive data analysis and code execution.
        2. **VS Code:** For implementing and debugging scripts.
    - **Hardware/Software:**
        1. A standard laptop with Python 3.x installed.
        2. Necessary libraries installed using pip or conda.
- **Learning resources**
    - **Research Articles:**
        1. Studies on e-commerce and customer behavior analysis using AI and machine learning.
    - **Sample Datasets:**
        1. Flipkart dataset: For training and testing machine learning models.
        2. Hypothetical data: To simulate Afghanistan's e-commerce market for contextual relevance.

Sign-off

Role	Name	Signature	Date
Team Lead	Mohammad Aref Rezvan Panah		
Technical Supervisor	Zainullah Zairmal		
Training Manager	Naeemullah Amani		

I am pleased to confirm the submission of the project documentation and its code, prepared and reviewed by me, Trainer Zainullah Ziarmal. This encapsulates all required details, ensuring clarity and alignment with the outlined projects objectives.

Best regards,

Zainullah Ziarmal

Zainullahziarmal.cs786@gmail.com