**Integrated Capstone Project**

**This Case Study has three check points defined in it.**

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python ( 25 marks)  1.2 Statistical Analysis using Python (25 Marks) | **Check point 1** | **25** |
| 2.1 Visualization using Python(20 marks)  2.2 Exploratory Data Analysis(40 marks)  2.3 - Model Building using ML algorithms | **Check Point 2** | **50** |
| 3.1 Deployment of ML model using Flask (30 marks) | **Check point 3** | **24** |

**Domain:**

Retail data set

**About:**

Black Friday is an informal name for the Friday following Thanksgiving Day in the United States, which is celebrated on the fourth Thursday of November. The day after Thanksgiving has been regarded as the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until more recent decades.

Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving. Black Friday is not an official holiday, but California and some other states observe "The Day After Thanksgiving" as a holiday for state government employees, sometimes in lieu of another federal holiday, such as Columbus Day. Many non-retail employees and schools have both Thanksgiving and the following Friday off, which, along with the following regular weekend, makes it a four-day weekend, thereby increasing the number of potential shoppers.

Black Friday has routinely been the busiest shopping day of the year in the United States since 2005, although news reports, which at that time were inaccurate, have described it as the busiest shopping day of the year for a much longer period. Similar stories resurface year upon year currently, portraying hysteria and shortage of stock, creating a state of positive feedback.

**Challenges:**

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month.

**What is Expected?**

Being a data analyst, you must come up with a first step document that lists output of your exploratory analysis, any issues or problems you may see with data that need follow up, and some basic descriptive analysis that you think highlights important outcomes/findings from the data. Based on your findings, the next level of analysis will be charted out.

.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **User\_ID** | User ID |
| **Product\_ID** | Product ID |
| **Gender** | Sex of User |
| **Age** | Age in bins |
| **Occupation** | Occupation (Masked) |
| **City\_Category** | Category of the City (A,B,C) |
| **Stay\_In\_Current\_City\_Years** | Number of years stay in current city |
| **Marital\_Status** | Marital Status |
| **Product\_Category\_1** | Product Category (Masked) |
| **Product\_Category\_2** | Product may belong to other category also (Masked) |
| **Product\_Category\_3** | Product may belong to other category also (Masked) |
| **Purchase** | Purchase Amount (Target Variable) |

**Check Point 1**

**Task 1.1(Data Manipulation using Python)**

Here are some indicative types of analysis you can perform. Please note that this is not an exhaustive list, you may add more

* Come up with appropriate results for the following:
* Maximum, spend in different categories of products
* Based on above data which set of customers can be offered personalised discount vouchers
* And in which category the voucher should be offered
* Or it should be on the total amount.

**Task 1.3 (Statistical Analysis using Python)**

* + Descriptive statistics for both numerical and categorical and draw few insights from them.
  + Perform relevant hypothesis testing (t, chi-Square, Anova tests)

**Check point 2 (Visualization using Python, EDA, Model building using ML Algorithms)**

* **TASK 2.1 (Visualization using Python)**
* Here are some indicative types of visualization you can perform. Please note that this is not an exhaustive list, you may add more
  + Come up with appropriate results and visuals for the following:
  + Maximum, spend in different categories of products
  + Based on above data which set of customers can be offered personalised discount vouchers
  + And in which category the voucher should be offered
  + Or it should be on the total amount.

Prepare the data by handling missing values, outlier analysis, data transformation and normalization.

**TASK 2.2 (Exploratory Data Analysis)**

Data Preparation/Analysis tasks including (but not limited to) the following.

* Univariate, Bi- Variate Analysis and Multi- Variate Analysis
* Missing values identification and treatment
* Outlier analysis and treatment
* Data scaling using min-max and/or Z-score normalisation
* Data transformation
* Feature Engineering

**Task 2.4(Model building using ML algorithms)**

**Predictive Analysis:**

* Build appropriate predictive model/s on the data.
* Compare various predictive models with appropriate regularization and/or hyper-parameter tuning.
* Evaluate the performance of the model.
* Identify the right metric to evaluate the performance of the model.
* Identify issues and concerns on the given data and suggest the best technique/s to overcome the issues.

**Task 3.3 -Deployment of Models using Django/Flask**

Deploy the Machine Learning Model created in Task 2.4 using the Flask application.