**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 Analysis using SQL Queries (25 Marks)  1.3 Statistical Analysis using Python (25 Marks) | **Check point 1** | **75** |
| 2.1 Visualization using Python(20 marks)  2.2 Exploratory Data Analysis(40 marks)  2.3Visualization using Power-BI (25 marks)  2.4 - Model Building using ML algorithms (40 marks) | **Check Point 2** | **125** |
| 3.1 Data Analysis using Big Data Tools(35 marks)  3.2 Data Analysis on Cloud (35 marks)  3.3 Deployment of ML model using Flask (30 marks) | **Check point 3** | **100** |

**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)**

Samyukta, Chandana

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 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 – (replace), outlier analysis, data transformation and normalization.

**Task 1.2 (SQL-Oracle)**

**Reshma, Suresh, Monika**

**Stage 1: -**

* Construct and ER-Diagram for the above-mentioned Requirement
* Construct Tables has per the ER-Diagram.
* Identify the relationships between tables and use appropriate standards for the same where applicable
* Insert the appropriate data into the identified tables from the sample dataset provided.

**Stage 2. Write SQL queries for below**

* 1)TOP 50 best selling products
* 2) TOP best selling product categories
* 3) count\_sales by gender
* 4) favourite products by gender(Female)
* 5) favourite products by gender(male)
* 6) sales\_volume by occupation
* 7)sales\_volume by age
* 8)sales\_volume by city category
* 9)sales\_volume by stay in city years
* 10)sales\_volume by Marital\_Status
* 11)purchase amount by age and gender
* 12)purchase amount by age, gender and marital\_status

**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, Visualization using Power-BI, Model building using ML Algorithms)**

**TASK 2.1 (Visualization using Python)**

**Monika, Reshma**

* 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)**

**Samyukta, Chandana, Suresh**

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 ?Should we do it with gender, marital status too
* Data transformation -> Converted qualitative to quantitative data, what else comes under this?0-17, 51-55?
* Feature Engineering

**TASK 2.3(Visualization using Power-BI)**

**Data Visualization in Power BI**

* Which Age group is purchasing the highest products from the store? What is the purchase amount of the age group between 0-17?
* Which gender is purchasing the highest products from the store?
* Display the total purchase happening in the store
* Visualize top 5 occupations by purchase
* Does marital status have any impact on the purchase amount?
* Display the maxim purchase that happened from the store
* Which product category was sold in maximum from the store.
* Display the top 5 product categories sold in maximum.
* Built a decomposition for purchase amount with appropriate variables affecting the purchase amount.
* Visualize a Key influencer visual for the purchase amount to explain it by product category. Identify the impact of product category on purchase amount

**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.

Regression, RFs, DT, Recomm, Time-Series, ANNs, Clustering

Clustering – Recomm

Regression – TimeSeries

ANNs – Shallow, Deep, Hypertuning, activation func

**CheckPoint 3**

**Task 3.1 - Data Analysis using Big Data Tools**

**What is Expected?**

**24th**

Big Data technologies like HDFS, Hive and PySpark need to be used as the historical data increases in size. As part of this task the following activities need to be done.

● Develop a PySpark application to load data Spark DataFrames and save it into Hive tables on a Hadoop cluster in an optimized format.

● Perform profiling of the data through PySpark and ensure that it is migrated correctly whereever the source is an RDBMS

● Write PySpark routines to cleanse the data, prepare the data to handle missing values, and the data transformations identified in task 1.1 again making sure that the data is written into Hive tables in an efficient format

● If the predictive model identified in task 2.4 available in Spark MLlib then develop a PySpark application to implement and evaluate the ML model identified with appropriate metrics\

● Ensure that the best practices are followed and the design & code use the features of Spark and take advantage thereof.

**Task 3.2 - Data Analysis on Cloud**

**26th -**

**AWS**

* Move the Datasets to AWS s3
* Create datapipeline to move the data from storage to datawarehouse(Redshift). You are allowed to use other copy command as well to move the data from storage to datawarehouse.
* In AWS load the dataset to Athena load from storage
* Configure Data Lake and Athena for your data
* Ensure you have required privileges in Data Lake to access your table.
* Connect the Athena data to PowerBI
* Perform the tasks mentioned in Task 2.3

**AZURE**

* Move the DataSet to Azure Synapse Storage Gen1
* Create a serverless SQL pool to query the data from Storage gen1
* Create a Linked service to PowerBI
* Ensure you have sufficient privileges on Synapse to access the serverless sql pool.
* Perform various analytics on PowerBI
* Perform the tasks mentioned in Task 2.3

1. Source – SQL; Target – Synapse using DataFactory
2. Try Queries like sales(products), purchase

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

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