**Bibliotheca E-Book Subscription Recommendation**

**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) | **Checkpoint 1** | **75** |
| 2.1 Visualization using Python(20 marks)  2.2 Exploratory Data Analysis(30 marks)  2.3Visualization using Power-BI (25 marks)  2.4 - Model Building using ML algorithms (50 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:**

**Bibliotheca E-Book Subscription Recommendation**

**About:**

Bibliotheca is a platform that gives its users an exclusive chance to subscribe to E-books at discounted price. Users can also get a preview of the books by reading a sample of pages even before they buy a subscription for any specific book. Each purchase for the subscription gives users access to the relevant E-book for a fixed period.

**Challenges:**

We have information about all the books that a User has previewed/purchased a subscription for; on the platform in the past and the problem statement asks you to predict the 10 most probable books the user is going to buy in the coming weeks.

For this case study, you need to predict the right books that User is most likely to buy in the coming weeks and your overall task is to Maximize accuracy.

The training data will include the information about the Books, Users and their history. The books that users have previewed/subscribed to in the past has been made available as part of the visiting history dataset. [The visit data will not be available for the future period].

**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.

**Recommendations:**

* As a data analyst, what are the approaches you suggest to the platform to recommend the right books to subscribe or buy.

**Data Dictionary:**

* BookID: Row Number indicating transactions
* GENRE: Book Type- Fiction or Biography etc..
* BOOKTITLE: Title of the Book along with author name
* BOOKNAME: Name of the Book
* SERIES: Name of the Book Series
* AUTHOR: Author name
* USERRATINGS: Ratings by the user
* OTHERPRINTEDINFO: Any Other information like Total no. of pages, published year etc.
* DETAILS: Book Details
* SUMMARY: Summary of the Book
* AUTHORDESC: Information about the author.
* COVERPAGE: Book cover image link
* Popularity: Popularity value of the book.

**Book Purchase History:**

* UserID: ID of the User
* TIMESTAMP: Date and Time of Purchase
* SUBSTATE: Location of purchase
* WeekofYear: Week and year of purchase

**Book Visit History:**

* SESSIONID: unique number that a Web site's server assigns a specific user for the duration of that user's visit
* REFERRALID: Referral Code
* TIMESTAMP: Date and Time
* BOOKPREVIEWED: 1 if book is previewed
* SUBSCRIBED: 1 if user has subscribed else 0
* WeekofYear: Week and Year of subscription

**User Data:**

* SIGNUPDATE: Sign up date and time of User
* DELETED: Date and Time if deleted
* AGEGROUP: Age group of the User
* GENDER : Gender of the User
* STATE: State to which the user belongs

**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

1. Identify the missing values.
2. Extract users and ratings.
3. Extract books that have received ratings more than 3.5.
4. Print any one book title and description randomly
5. Obtain the author description for any one Book with User rating more than 4.5.
6. Number of unique books and unique users.
7. Most read books.

**Task 1.2 (SQL-Oracle)**

**Stage 1:**

1. Construct an ER-Diagram for the above-mentioned Requirement
2. Construct Tables as per the ER-Diagram.
3. Identify the relationships between tables and use appropriate standards for the same where applicable
4. Insert the appropriate data into the identified tables from the sample dataset provided.

**Stage 2:**

1. Display bookname and author details whose user ratings range between 3.5 to 4.5.
2. Display top 5 bookname based on their popularity.
3. Display the statistics(ie percentage) of the users who have been signed up with total count based on gender, and age group between 20-30,31-40;
4. Display the top 5 substates from where the books have been purchased the most.
5. Display the name of the top 3 authors whose books are purchased the maximum number of times.
6. Display the genre of the books which are commonly purchased and also the substate from where they are subscribed.
7. Display the details of the books which are frequently previewed.

**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 on PowerBI, Machine Learning)**

**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:

Visualization on some important parts like:

* Most rated books
* most popular books
* most popular authors
* Average rating distribution for all books
* states and substates with most number of books and
* Age group of the user with most read books based on Genre.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**TASK 2.2 (Exploratory Data Analysis)**

Perform the below data preparation/analysis tasks relevant to the problem statement

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

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

**Connect the data with Power BI desktop and perform Data Manipulation using Power Query Editor. Perform the below tasks in Power BI Desktop.**

Come up with appropriate results and visuals for the following: Please note that this is not an exhaustive list, you may add more,

* Top 5 Genre based on User Ratings.
* Identify the Best Books based on Popularity belonging to the following Genre categories: Fiction, Science Fiction, Mystery and Biography
* Top 10 highest rated books
* Top 10 most popular books
* Top 10 authors with the most number of books
* Top 5 Substates and states based on Book purchase.
* Most number of ratings
* Which User age group has most read books on Science Fiction?

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

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

You have to build recommender system/ model  where we have to recommend similar books to the reader based on his/her interest.

* Build appropriate predictive model/s on the data.
* Compare various recommendation models
* 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.

**Recommendations:**

* As a data analyst, what are the approaches do you suggest to the sales team to identify the customer interests on eBooks and recommend suitable eBooks to subscribe.

**NOTE:** Results and graphs must be backed with appropriate inferences and insights.

**CheckPoint 3**

**Task 3.1**

**What is Expected?**

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 Parquet format.
* Perform profiling of the data through PySpark and ensure that it is migrated correctly wherever 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 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**

**AWS**

1. Move the Datasets to AWS s3
2. Create Redshift Instance
3. Ensure you create required tables in Redshift
4. Create a data pipeline/copy command to move the data from storage to data warehouse(Redshift). You are allowed to use other copy commands as well to move the data from storage to data warehouse.
5. Connect the Redshift data to PowerBI
6. Perform the tasks mentioned in Task 2.3(Only 4-5 core reports)

**AZURE**

1. Move the Dataset to Azure Synapse Storage Gen1
2. Create a serverless SQL pool to query the data from Storage gen1
3. Create a Linked service to PowerBI
4. Ensure you have sufficient privileges on Synapse to access the serverless sql pool.
5. Perform various analytics on PowerBI
6. Create the dashboard in PowerBI like Task 2.3 only 4-5 core reports

**Task 3.3**

Deploy the Machine Learning Model created Task 2.4 in flask application.