**Integrated Capstone Project**

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

|  |  |  |
| --- | --- | --- |
| **Check Point Topics** | **Remarks** | **Max Marks** |
| 1.1 Data manipulation using Python ( 50 marks)  1.2 Analysis using SQL Queries (50 Marks) | **Checkpoint 1** | **100** |
| 2.1Visualization using Power-BI (50 marks)  2.2 Data Analysis using Big Data Tools(50 marks) | **Check Point 2** | **100** |
| 3.1 Data Analysis and ML Model Training and deployment on Cloud (100 marks) | **Checkpoint 3** | **100** |
| 4.1Final Presentation and Viva( 50 marks) | **Check point 4** | **50** |

**Domain:**

Financial services data set

**About:**

Finance is a field that is concerned with the allocation (investment) of assets and liabilities over space and time, often under conditions of risk or uncertainty. Finance can also be defined as the art of money management. Participants in the market aim to price assets based on their risk level, fundamental value, and their expected rate of return.

The dataset consists of customer information of L&T financial services. It is a finance dataset, which consists of customers demographics, loan disbursed, asset cost being purchased and the customers previous account and loan history. The dataset also consists of the state and branch id of L&T from where the loan was disbursed, the customer’s account history. It also contains the CNS score and score description provided by the Credit Bureaus of India.

**Challenges:**

It is a challenge for any financial services to target the right people for disbursing the loan. The credit team must analyse various details like CIBIL score, payment history (if available), credit history, geographical location, profession, income, age, education etc. of the customers. This will help in understanding whether the person is capable of paying back the loan amount. Which in turn reduces its NPAs and increases the profitability.

You need to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

**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 the credit team to identify factors that decide high probability of default? Recommend based on your analysis.

.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Loan\_Id | Unique Loan id |
| Disbursed\_Amount | Amount of Loan disbursed |
| Asset\_Cost | Cost of the Asset |
| LTV | Loan to Value of the asset |
| Branch\_Id | Branch where the loan was disbursed |
| City\_Code | Code representing City |
| State\_Code | Code representing State |
| Postal\_Code | Postal code of the area |
| Date\_of\_Birth | Date of birth of the customer |
| Employment\_Type | Employment Type of the customer (Salaried/Self Employed) |
| DisbursalDate | Date of disbursement |
| Region\_ID | Code for region of disbursement |
| MobileNo\_Avl\_Flag | If Mobile no. was shared by the customer, then flagged as 1 |
| Aadhar\_flag | If Aaddhar was shared by the customer then flagged as 1 |
| PAN\_flag | If PAN was shared by the customer then flagged as 1 |
| VoterID\_flag | If voter id was shared by the customer then flagged as 1 |
| Driving\_flag | If DL was shared by the customer then flagged as 1 |
| Passport\_flag | If passport was shared by the customer then flagged as 1 |
| PERFORM\_CNS.SCORE | Bureau Score |
| DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS | Number of delinquent accounts in the last six months |
| CREDIT.HISTORY.LENGTH | Credit history in terms of years |
| NO.OF\_INQUIRIES | Number of inquiries made by the customer |
| Loan\_Default (target variable) | Payment default in the first EMI on due date |

**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:
  + Analysis of percentage of default and check whether it is imbalanced data?
  + Replace missing values with appropriate techniques
  + Use the right treatment for outliers in the data
  + Analyse default variable with demographic related data
  + Determine and draw insights on association between default and other variables, both categorical and numerical.

**Task 1.2 (SQL-Oracle)**

**Stage 1:**

* + Construct and ER-Diagram for the above-mentioned Requirement
  + Construct Tables as 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:**

* Generate Reports of the customer who has approached for the loan and all kyc is submitted and disbursal date is not given, or loan is not disbursed.
* Generate Reports of the customer who has approached for the loan whose Bureau Score is less than 650 and min. enquires made is >3 and the cost of asset is between 50000 to 70000 and ltv is between the range of 50-60.
* Generate Reports of the customer who has approached for the loan whose Bureau Score is less than 650 and min. enquires made is >3 and the cost of asset is between 50000 to 70000 and ltv is between the range of 50-60.
* Generate Reports of the self-employed customer who has approached for the loan whose Bureau Score is less than 650 and min. enquires made is >3 and loan default are 0 and and if all kyc is submitted and cross checked if any dues to previous loans.
* Generate Reports of those customers where the loan is disbursed based on the Cibil score ranging between 670 -780 and also who are acquired the loan for more than 1 time and loan default is 1 or 0.

**Check point 2 (Visualization using Power-BI)**

**TASK 2.1(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.**

* What were the total enquiries done?
* What was the maximum asset cost?
* What is the average asset cost for each employment type?
* What is the average loan default for each driving flag?
* Display to Key Influencer Visual for the appropriate columns and indicate your inferences.
* Display loan default by employment type and indicate which employment type has the highest loan default.
* Display a decomposition tree for the data.

**TASK 2.2** **Data Analysis using Big Data Tools**

**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 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 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.

**Deliverables/Submission guidelines of Checkpoint 2**

1. You have to prepare a power point presentation with screenshots of outputs (10 -15 slides) for each check point.
2. Mention Problem Statement and your approach to the problems
3. Task 2.1
   * 1. PowerBI .pbix file to be submitted.
     2. Have all comments written properly in the .pbix file.
     3. The .pbix file should contain the Batch Name and the Group Number, Group member names at the top.

Task 2.2

* + Submit Jupyter code file in html format. The code file(html file for Task 2.2) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
    1. All comments/inferences/insights/reasons for doing a particular tasks etc should be written as a ‘markdown text’, but **NOT** using a comment lines with # or ‘’’.
    2. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).
    3. Name of the file must be in the form of:
    4. *BatchName\_\_GroupNumber\_FirstName\_SecondName.html*
* Put all Tasks 2.1 & 2.2 as zip file (Mentioning batch name, Group number and your name) and upload it on the LMS.

**CheckPoint 3**

**Task 3.1 - Data Analysis + ML Model Training and Deployment on Cloud**

**AWS**

1. Redshift to PowerBI Connectivity
2. Move the Datasets to AWS s3
3. Create Redshift Instance
4. Ensure you create required tables in Redshift
5. 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.
6. Connect the Redshift data to PowerBI
7. Perform the tasks mentioned in Task 2.1(Only 4-5 core reports)
8. Dynamodb to s3 bucket confgiure SNS notifications for any new records added in the Dynamodb
9. Transfer the AWS s3 data to AWS Quicksight perform the same analysis doe using powerBI(Any 5 core reports)
10. 50% of storage exceeds then cloud watch has to trigger the alaram
11. Write a Lambda function which logs in cloud trail about S3 file type and size.
12. Build appropriate ML model/s on the data using AWS Sagemaker , Identify the right metric to evaluate the performance of the model **and Deploy on AWS Sagemaker**

**AZURE**

1. Azure Synapse to PowerBI Connectivity
   1. Move the DataSet to Azure Synapse Storage Gen2
   2. Create a serverless SQL pool to query the data from Storage gen2
   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. Perform the tasks mentioned in Task 2.1(Only 4-5 core reports)
2. Enable Azure blob storage monitoring by adding sample data and upon processing if storage receives more than 20 bytes of data
3. Azure blob to azure data bricks using notebook options databricks to powerbi connectivity
4. Azure blob to Azure SQL copy option using datafactory and connect Azure SQL to Databricks
5. Write Azure functions to trigger to trigger when blob storage exceeds 20 bytes of data.
6. Build appropriate ML model/s on the data using Azure Machine Learning , Identify the right metric to evaluate the performance of the model **and Deploy on Azure Machine Learning**

**GCP**

1. BigQuery to PowerBI Connectivity
   1. Move the Datasets to Google Storage (Bucket)
   2. Create Bigquery Instance
   3. Ensure you create required tables in Bigquery
   4. Create a data pipeline/copy command to move the data from storage to data warehouse. You are allowed to use other copy commands as well to move the data from storage to data warehouse.
   5. Connect the BigQuery to PowerBI
2. Write Cloud Function by adding sample data in the cloud storage and upon processing if storage storage receives more than 20 bytes of data as inbound or outbound
3. Transfer the data from bucket to Looker and perform any 5 reports performed in the step 2.1
4. Configure GCP monitoring services when storage exceeds 20 bytes of data, notify using pub/sub.
5. Configure Google Big Query and enable monitoring services (Cloud Logging) for every record insertion or deletion.
6. Build appropriate ML model/s on the data using Google Big Query Models/Vertex AI , Identify the right metric to evaluate the performance of the model **and Deploy the model on GCP Machine Learning.**

**Deliverables/Submission guidelines of Checkpoint 3**

Task 3.1

Complete all the above tasks on your respective Cloud Platform allotted and for submission take screenshots of each task specified with step by step flow in a word document with proper caption mentioned along with your Batch/Group/Team member names convert as a PDF file and submit the PDF document on the LMS

**CheckPoint 4**

**Task 4**

Prepare crisp Final presentation including all three Checkpoint achievements and appear for Q&A session

**Deliverables/Submission guidelines of Checkpoint 4**

* You have to prepare a power point presentation with screenshots of outputs (10 -15 slides)
* Submit the ppt.

The above four Checkpoints completes UNext Capstone Project