**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 + ML Model Training and Deployment on Cloud (100 Marks) | **Checkpoint 3** | **100** |
| 4.1Final Presentation and Viva( 50 marks) | **Check point 4** | **50** |

**Domain:**

Unleashing the Potential of Strategic Marketing for Financial Services

**About:**

HBCF Bank is a prominent public sector bank operating numerous branches across various cities. It offers a wide range of financial services to its customers, including savings accounts, current accounts, term deposits, personal loans, home loans, and more. As part of its marketing strategies, the bank recently organized a campaign to promote its term-deposit scheme. Throughout the campaign, extensive data was gathered concerning customers' personal, social, and economic information. This data encompasses details on the marketing efforts undertaken by the bank to achieve success in the campaign.

The primary mode of marketing employed by the bank was direct phone calls, with bank representatives contacting existing customers and urging them to consider placing a term deposit. The outcome of the campaign was determined by the customer's decision to agree or decline the offer, and the corresponding target variable was categorized as 'yes' for success and 'no' for failure. By meticulously monitoring and analyzing this data, HBCF Bank aims to gain valuable insights into the effectiveness of its marketing strategies and the factors influencing customer decisions, enabling them to refine their future campaigns and better serve their clientele.

**Challenges:**

In today's competitive banking landscape, running successful marketing campaigns is an arduous task that requires strategic precision. For bank officials, identifying the right audience and tailoring campaigns to their specific needs is crucial. To accomplish this, the marketing team must delve into a wealth of customer data, including profession, income, age, education, existing loans, credit history, and more. Such comprehensive analysis helps determine a customer's economic status and their suitability for participating in term-deposit schemes or other financial offerings. Leveraging advanced data analytics techniques, banks can unlock valuable insights, discover patterns, and predict customer behaviour. This article explores the pivotal role of data analytics in targeted banking campaigns and the steps involved in harnessing its potential.

Conclusion: In an increasingly data-driven world, banks are realizing the immense potential of data analytics in targeted marketing campaigns. By leveraging exploratory and descriptive analytics techniques, banks gain valuable insights into customer demographics and economic indicators. Building predictive models empowers marketing teams to make data-driven decisions, ensuring their efforts are focused on the right audience. The integration of data analytics in banking campaigns leads to improved customer targeting, enhanced campaign success rates, and ultimately, greater customer satisfaction. As banks continue to embrace the power of data, the future of targeted banking campaigns looks promising, with personalized offerings that cater to individual needs and preferences.

**What is Expected?**

As a data analyst tasked with optimizing targeted banking campaigns, the first step is to conduct an exploratory analysis of the available data. This process involves extracting meaningful insights, identifying potential issues or problems in the dataset, and performing basic descriptive analysis to highlight important outcomes. Furthermore, building an appropriate predictive model for classifying campaign success will be essential. This article outlines the key steps involved in the exploratory analysis, highlights potential data issues, presents key descriptive findings, and discusses the subsequent predictive modelling phase, including a comparative study of various approaches.

Exploratory Analysis: During the exploratory analysis, several outputs and observations are derived from the dataset. These outputs serve as a foundation for further analysis and decision-making. Some of the key outputs include:

1. Data Distribution: Examining the distribution of key variables such as profession, income, age, education, loan status, and credit history provides insights into the customer base and potential segmentation opportunities.
2. Correlation Analysis: Conducting correlation analysis helps identify relationships between variables. It allows for the identification of influential factors that may impact campaign success, such as income level and loan status.
3. Missing Data Assessment: Identifying missing values and determining their impact on the analysis is crucial. Understanding the patterns of missing data and devising appropriate strategies to handle them ensures the reliability of subsequent modelling efforts.

Data Issues and Follow-Up: While conducting the exploratory analysis, potential issues or problems with the data may arise. These issues need to be addressed or followed up to ensure accurate and reliable predictive modelling. Some common issues include:

1. Data Inconsistencies: Identifying inconsistencies in data entry, such as incorrect or missing values, is crucial. These inconsistencies can affect the accuracy and reliability of the predictive model.
2. Outliers Detection: Outliers can significantly impact the analysis and model performance. Identifying and addressing outliers is important to avoid skewed predictions and biased outcomes.
3. Class Imbalance: Imbalanced classes, where the number of successful campaigns significantly differs from unsuccessful ones, can impact model performance. Techniques such as oversampling or under sampling may be required to address this issue.

Descriptive Analysis and Key Findings: The descriptive analysis provides a snapshot of the dataset and highlights important outcomes. Some essential descriptive findings from the data exploration may include:

1. Customer Segmentation: Identifying distinct customer segments based on demographic and economic indicators allows for personalized campaign targeting and tailored offerings.
2. Customer Behaviour Patterns: Analyzing historical campaign data provides insights into customer responses and behaviours, enabling banks to refine their marketing strategies and increase the likelihood of success.
3. Predictive Factors: Determining the key predictive factors, such as income levels, loan status, and credit history, helps build a foundation for developing accurate predictive models.

Building Predictive Models and Comparative Study: Once the exploratory analysis is complete, the next step is to build a predictive model to classify successful and unsuccessful campaigns. Several approaches, such as logistic regression, decision trees, random forests, and neural networks, can be employed. A comparative study of these models, assessing their performance metrics such as accuracy, precision, recall, and F1-score, will provide insights into the most effective approach for this specific banking campaign dataset.

Conclusion: By embarking on an exploratory analysis of the available data, data analysts can unveil key insights and potential issues, setting the stage for subsequent predictive modelling. Descriptive analysis highlights important findings that guide the campaign optimization process. The comparative study of predictive models aids in selecting the most appropriate approach to accurately classify successful and unsuccessful campaigns. Armed with these insights and modelling outcomes, banks can refine their marketing strategies, increase campaign success rates, and ultimately enhance customer satisfaction and loyalty.

**Data Dictionary:**

**Bank client data:**

* Customer id : Unique customer id
* CustAge: Age of the customer.
* Profession: type of job
* Marital: marital status
* Schooling: Educational qualification
* Default: has credit in default?
* Housing: has housing loan?
* Loan: has personal loan?
* State\_Code: Code representing unique state name
* Region\_Code: Code representing unique Region name
* City\_Code: Code representing City of the customer
* Postal\_Code: Postal code of the area to which the customer belongs to.

**Data related to the last contact of the current campaign:**

* Contact: contact communication type
* Month: last contact month of year
* Day\_of\_week: last contact day of the week
* Campaign: number of contacts performed during this campaign and for this client (includes last contact)
* Pdays: number of days that passed by after the client was last contacted from a Previous campaign (999 means client was not previously contacted)
* Previous: number of contacts performed before this campaign and for this client
* Poutcome: outcome of the previous marketing campaign

**Data related to social and economic context attributes**

* emp.var.rate: employment variation rate - quarterly indicator
* cons.price.idx: consumer price index - monthly indicator
* cons.conf.idx: consumer confidence index - monthly indicator
* euribor 3m: euribor 3 month rate - daily indicator
* nr.employed: number of employees - quarterly indicator

**Target variable:**

* responded - has the client subscribed to a term deposit?

**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 turnout of marketing campaign
  + Right mode to contact the customers (telephone or mobile)
  + Analysis on attempts made to turn a person into successful depositor
  + Personal data analysis on marital status, existing loans, education, profession etc. and its impact on the campaign’s success.
  + Socio-economic analysis of the customers
  + Demographic analysis of the marketing campaign using the master files.

**Task 1.2 (SQL-Oracle)**

**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:**

* Generate the list of customer per age and profession and display them according to ascending order of the age.
* Generate the list of customer who have no housing loan and personal loan.
* Generate the list of customer who has credit and has been contacted more than 2 times during the campaign.
* Generate the list of customer who has credit and has been contacted more than 2 times during the campaign and

outcome of the campaign is successful.

* Generate the list of customer who has credit and and has been contacted more than 2 times during the campaign and outcome of the campaign is failure and who have been contacted in the month of June and also display the day contacted.
* Generate report giving state-wise breakup of number of customers
* Generate report giving city-wise distribution of customers

**Deliverables/Submission guidelines of Checkpoint 1**

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. You need to submit all the code files - Task 1.1
4. The code file(html file for Task 1.1) should contain the Batch Name and the group name, group members (One of the group member) at the top (in Jupyter Notebook).
5. 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 ‘’’.
6. Submit the code file as HTML file format (you have an option in Jupyter Notebook to save the file as HTML).

Name of the file must be in the form of:

*BatchName\_FirstName\_SecondName.html*

1. Task 1.2 SQL code to be copied in the word doc
2. The presentation file should have the Batch name, group name, Project name, Group members, their responsibilities
3. Upload all the deliverables in the UNext LMS

**Check point 2 (Visualization using Power-BI , Data Analysis using Big Data Tools)**

**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.**

* Perform Monthly Response Analysis throughout the year for the campaign. Identify the highest Response Month.
* Right mode to contact the customers on all the days of the week (cellular or telephonic). Is there any day of the week where customers responded to telephonic call than to cellular calls?
* How customers age is affecting the marketing campaign success. Suggest which age group needs to be focused to improve the campaign success.
* Impact of Education and Job on Campaign Success.
* Impact of contact to the customer for a positive response. Identify the maximum number of times a customer needs to be contacted for a positive response.
* Customers of what qualification are responding positively for the campaign. Whether the previous outcomes are affecting the ongoing campaign.
* Is it a good strategy to contact a married customer having a house loan for the campaign?
* Which day of the week does the marketing campaign yields a high success rate?
* Is it a good strategy to contact a customer having a loan expecting a positive response for the campaign?

**Recommendations:**

* As a data analyst, what are the approaches do you suggest the marketing team to identify ideal target group to make the campaign successful? Recommend based on your analysis.

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

**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.

● 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

● 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 Datawarehouse(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. DynamoDB to s3 bucket configure SNS notifications for any new records added in the DynamoDB
8. Transfer the AWS s3 data to AWS Quick sight perform the same analysis doe using powerBI(Any 5 core reports)
9. 10 mb of storage of S3 exceeds then cloud watch has to trigger the alarm
10. Write a Lambda function which logs in cloud trail about S3 file type and size.
11. 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
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. Big Query 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.
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