**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:**

Marketing for Financial Services

**About:**

DB Bank is a large public sector bank which has branches across the cities. It provides various services like savings accounts, current account, term deposits, personal loans, home loans etc. to customers. Whenever the bank conducts marketing on its new schemes, it will keep track of data related to customers’ personal, social and economic details. Also, it maintains the detailing on efforts made to achieve success in the campaign.

Recently, the bank has conducted a campaign to market their term-deposit scheme. Campaigns was conducted based mostly on direct phone calls, soliciting the bank's customers to place a term deposit. After all the marking efforts, if the client had agreed to place a deposit, then the campaign is success, otherwise not (Target variable marked 'yes', or 'no').

**Challenges:**

It is a challenge for bank officials to target right people for a successful campaign. The marketing team must analyse various details like profession, income, age, education etc. of the customers. Also, the possibility of any existing loans, credit history etc. will give economic status of the customer and hence helps in understanding whether the person is capable of depositing money into term-deposit scheme of the bank.

The first step of this analysis is to assess what data is available and perform some exploratory and descriptive analytics to identify interesting and useful patterns, trends, and insights.

And the next step is to build a predictive model on the give data.

**What is Expected?**

Being a data analyst, you must come up 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.

Also, you need to build appropriate predictive model for classifying successful and unsuccessful campaigns. You can perform comparative study of several predictive models with various approaches and give your inferences accordingly.

.

**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 with 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
* euribor3m: euribor 3 month rate - daily indicator
* nr.employed: number of employees - quarterly indicator

**Target variable:**

* responded - has the client subscribed 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-economical 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 has 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

**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)**

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:
  + 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-economical analysis of the customers
  + Demographic analysis of the marketing campaign using the master files.

**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 2.3(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.**

* 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 2.4(Model building using ML algorithms)**

**Model Building:**

* Build appropriate ML model/s on the data.
* Compare various ML 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.

**CheckPoint 3**

**Task 3.1 - 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 PySpark routines to load data Spark DataFrames and save it into Hive tables in an optimized format on a Hadoop cluster, de-normalizing the data if required.

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

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

● 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**

**AWS**

1. Move the Datasets to AWS s3
2. 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.
3. In AWS load the dataset to Athena load from storage
4. Configure Data Lake and Athena for your data
5. Ensure you have required privileges in Data Lake to access your table.
6. Connect the Athena data to PowerBI
7. Perform the tasks mentioned in Task 2.3

**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. Perform the tasks mentioned in Task 2.3

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

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