Capstone Project

Problem Statement

The dataset is from a bank, data related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be (or not) subscribed. The data and attribute description are in the folder.

Dataset Description

The dataset has the following attributes:
1 - unique sequence id
2 - age (numeric)
3 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
4 - marital_status : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
5 - education (categorical: "unknown","secondary","primary","tertiary")
6 - default: has credit in default? (binary: "yes","no")
7 - balance: average yearly balance, in euros (numeric)
8 - housing: has housing loan? (binary: "yes","no")
9 - loan: has personal loan? (binary: "yes","no")
10 - contact: contact communication type (categorical: "unknown","telephone","cellular")
11 - day: last contact day of the month (numeric)
12 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
13 - duration: last contact duration, in seconds (numeric)
14 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
15 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
16 - previous: number of contacts performed before this campaign and for this client (numeric)
17 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
18 - opened_new_td_act_no_yes - has the client subscribed to a term deposit? (binary: "yes","no")

Business Requirements:

The idea is we have set of information about a person whether it’s their age the job they have the marital status whether they have loan with the bank or not etc. we're trying to decide here is whether this person open a new deposit account with this bank. So, the theory here is that if I can better target the right customers with the right offer at right time they'll convert over to actually open a deposit account with the bank so that I don't have to reach out whole population and waste a lot of money with ads or mailer or whatever have you but target only the perspective customers.

Technical Requirements:

Create a solution using different Azure services learnt in the sessions to read the data from below mentioned datasets, performed the transformations mentioned below and write the data in ADLS in delta format. Once the data is written, create views using serverless synapse to read the data from ADLS location.

Datasets: Portuguese banking institution

Prerequisites: All the services discussed in the sessions needs to be configured in your Azure account.

The Activity:

1. Create a GitHub repo(source), add all the documents related to the capstone project.
2. Place the raw data files in the Bronze layer in your ADLS storage.
3. Create a pipeline in ADF with notebook activities to trigger the ADB notebook, you will be creating.
4. In the pipeline there should be a notebook activities created.

    a. The notebook activity will read the data from Bronze / Silver L1 layer. Below mentioned transformations needs to be performed on this data read and they need to be written to Gold layer. Data transformation logic has been mentioned below.

    The data written in Gold, must have data columns from both the files.

    b. Once the processing in ADB notebooks is complete, use web activity and Logic apps to trigger an email to yourself with a success message.

    c. Once this transformed data is written onto Gold layer, create views using serverless synapse and Read the data from synapse view.

Transformation Logic:

The below mentioned transformation needs to be done as a part of the second notebook activity you will be creating in the ADF pipeline.

As stated above, you will be **reading the Joined data from Silver L1 layer**, **Transforming the data** and **writing it to Gold layer**.

1. Define the schema for the dataset and use the schema to read the file bank_data.csv
2. Using the above schema read the data and the data frame.

3. Verify the schema
4. Check the datatypes
5. Cache the dataframe
6. Verify the first few records
7. Verify the total number of rows and columns
8. Verify the summary statistics
9. Find the maximum and minimum values in each column
10. Find if there are any negative balances in the columns
11. Replace the negative balances with zero.
12. Define a table/view on the spark dataframe created to run sql queries on the dataframe.
13. Verify the target distribution.
14. Find the pairwise frequency between target and loan columns
15. Find the term deposit opted for different job categories. (Plot a visualization for the same).
16. Find the term deposit not-opted for different job categories. (Plot a visualization for the same).
17. Find the term deposit not-opted for different education categories. (Plot a visualization for the same).
18. Find the term deposit not-opted for different marital status category. (Plot a visualization for the same).
19. Plot a heat map for the dataframe
20. Remove null values in the dataframe
21. Split the data into training and testing to make it ready for Spark ML pipeline.

General Instructions:

1. The project must be done by an individual.
2. Queries regarding the project need to be posted on chat / discuss with the SME.
3. Design the project as per the problem statement given below.
4. The project evaluation is for 100 marks.

Submission:

- Detailed presentation need to prepared by taking a screen shot of all the steps mentioned above with your name/id that is present in ADB and ADF on top left corner in a document.
- Also attach the notebook code files (Download the DBC archive file from ADB) in the respective folders you will be creating for submission.
- ipynb, ppt, screenshot, dataset, any relevant document
- System will accept only ZIP file submissions i.e., in .zip format (Max size- 100 MB).
- Review the .zip file before uploading it.
- Please ensure that your submission is complete in all aspects.
- Multiple submission is not accepted.
- We strongly recommend you submit at least 60 minutes before your deadline.
- There will be no extension so please make sure to submit before the deadline.
- Result of capstone project will be shared the business team.

Project Start Date	23-June-2023
Project End Date	30-June-2023
Project Submission Date	30-June-2023
Naming Convention for the file	<empid_firstname_Capstone_Project> F11035_Kiran_Capstone_Project.zip

Assessment Criteria

Participants will be graded on Approach, Solution and Presentation (25%,50%,25%)

S.No	Criteria		Marks
1	Approach (25)	Design of the solution	12.5
2		Domain, Azure Services and Technical Understanding	12.5
3	Solution (50)	Best programming practices Completeness Readability	15
4		Data Ingestion & Pipeline	15
5		Data exploration	20
6	Presentation (25)	Domain Business understanding	7.5
7		Completeness of presentation	7.5
8		Visualization Approach	5
9		Future Work	5

All the Best!!!