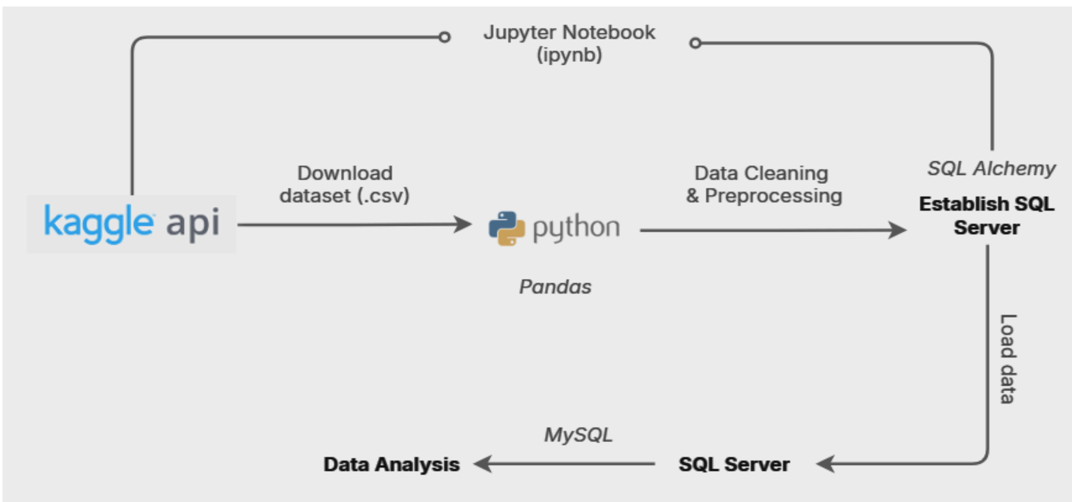
SQL Bank Target Marketing Analysis

Overview

This project involves a comprehensive analysis of marketing campaigns held by a bank aimed at increasing the number of term deposit subscribers. The goal is to identify the key factors influencing customer decisions, including their financial behaviors and demographic insights. The analysis provides a detailed account of the project's objectives, business problems, findings, and conclusions.



Objectives

1. Evaluate the outcomes of a bank's marketing campaign to acquire term deposit customers.
2. Identify key factors influencing customers' decisions to subscribe to deposits.
3. Provide recommendations to improve future targeted campaigns based on the analysis.

Dataset

The dataset for this project is sourced from the Kaggle dataset: [Bank Target Marketing]

Tech Stack

- **Kaggle API**: To download the dataset.
- **Jupyter Notebook**: IDE for data exploration.
- **Python** (Pandas): For initial data cleaning and manipulation.
- **SQLAlchemy**: To establish connection with the local SQL database server.
- **SQL Server**: For detailed query analysis.

Python Data Gathering

Importing Libraries

```
import os
import pandas as pd
import numpy as np
```

Data Collection & Extraction

```
# download dataset using kaggle api
# !pip install kaggle
import kaggle
!kaggle datasets download seanangelonathanael/bank-target-marketing -f "Bank Target Marketing Dataset.csv"
# The csv is packed into a zipped file.

# Extract data from zip file
import zipfile
zip_ = zipfile.ZipFile('data.csv.zip')
zip_.extractall()
zip_.close()
```

Data Cleaning

```
# read data
df = pd.read_csv('data.csv')
df.head()
df.info()
```

Check for duplicate values

```
df.duplicated().sum()   # 11162

# remove duplicates
print(f'Size of the dataset initially: {df.shape}')

df.drop_duplicates(inplace=True)
```

Check for null values

```
df.isna().sum()   # 0
```

Rename columns for easy analysis

```
# rename df columns:

new_cols = {
'default': 'credit_default',
'contact': 'contact_type',
'day': 'contact_day',
'month': 'contact_month',
'campaign': 'contact_count'
}

df = df.rename(columns=new_cols)
```

Establish a SQL Connection

Load data in SQL server

```
# !pip install pymysql
# !pip install mysqlclient

import sqlalchemy as sal
from sqlalchemy import create_engine

hostname = 'localhost'
username = '****'
password = '****'
port = 3306
database = 'my_sql'

engine = create_engine('mysql+pymysql://' + username + ':' + password + '@' +
hostname + ':' + str(port) + '/' + database)
conn=engine.connect()
```

Connect the dataset to SQL server

```
df.to_sql('bank_df', method=None, schema='my_sql', con=conn, index=False,
if_exists='append')

conn.commit()
```

SQL Data Analysis

Data Overview

- **Total records:** 45,211
- **Features:**
    - **Numerical:** (8 features)
      id, age, balance, contact_day, duration, contact_count, pdays, previous
    - **Categorical:** (10 features)
      job, marital, education, credit_default, housing, loan, contact_type, contact_month, poutcome, deposit
- **Target variable:** deposit (yes/no)
- **Null Values:** No missing values across all columns.

Schema

```sql
DROP TABLE IF EXISTS bank_df;

CREATE TABLE bank_df (
       age INT,
       job VARCHAR(20),
       marital VARCHAR(20),
       education VARCHAR(20),
       credit_default VARCHAR(10),
       balance FLOAT,
       housing VARCHAR(10),
       loan VARCHAR(10),
       contact_type VARCHAR(20),
       contact_day INT,
       contact_month VARCHAR(10),
       duration INT,
       contact_count INT,
       pdays INT,
       previous INT,
       poutcome VARCHAR(20),
       deposit VARCHAR(10)
);
```

Descriptive Statistics

1. What is the age distribution of customers?

```
SELECT MIN(age), MAX(age), AVG(age) FROM bank_df;
```

The dataset includes customers aged 18 to 95, with an average age of 41.

2. What is the average yearly balance of customers in the dataset?

```
SELECT MIN(balance), MAX(balance), AVG(balance) FROM bank_df;
```

- The average yearly balance of customers ranges from -$8k to $100k, with the negative balance suggesting that some may have financial issues due to outstanding debts.
- Approximately 83% of customers maintain a positive balance, while 16% have a zero or negative balance.

3. What impact does the frequency of prior contact (pdays) have on customer engagement in the campaign?

On average, customers were contacted 2 to 3 times during the campaign, with a significant 81.7% having no prior contact (pdays = -1).

```
SELECT pdays, count(*) count from bank_df
GROUP BY pdays
ORDER BY count DESC;
```

pdays	count
-1	36954
182	167
92	147
91	126
183	126
181	117

Categorical Data Analysis

1. Job Distribution: There are 12 distinct customer job categories, with the top 5 being 'blue collar,' 'management,' 'technician,' 'admin,' and 'services.'

```
SELECT job, COUNT(*) count
FROM bank_df
GROUP BY job
ORDER BY count DESC;
```

2. Marital Status: Among the customers, 60% are married, 28% are single, and 11% are divorced.

```
SELECT marital,
       COUNT(*) AS count,
       ROUND((COUNT(*) / (SELECT COUNT(*) FROM bank_df))*100,2) percentage
FROM bank_df
GROUP BY marital
ORDER BY count DESC;
```

job	count
blue-collar	9732
management	9458
technician	7597
admin.	5171

marital	count	percentage
married	27214	60.19
single	12790	28.29
divorced	5207	11.52

3. Education: Secondary education is the most prevalent educational level among customers, accounting for approximately 51.32% of individuals.

4. Communication: Over two-thirds of customers (64%) are contacted via *cellular*, while only 6% are contacted by *telephone*. Additionally, around 28% of customers have an *unknown* contact type.

5. Monthly Trends: *May* records the highest number of last contacts compared to other months.

These insights will be used later to understand customer spending behavior for further analysis.

Data Transformation

1. Adding a Primary Key column

Before we begin, we currently do not have a primary key column, so let's create one first:

```
ALTER TABLE bank_df
ADD client_id INT NOT NULL AUTO_INCREMENT PRIMARY KEY;
```

2. Combining columns

Columns such as *'housing'* and *'loan'* can be combined into a single column to make the data more efficient:

```
ALTER TABLE bank_df
ADD loan_type VARCHAR(20);

UPDATE bank_df
SET loan_type =
  CASE
    WHEN housing = 'yes' THEN 'housing-loan'
    WHEN loan = 'yes' THEN 'personal-loan'
    ELSE 'no loan'
  END;
```

The 'housing' and 'loan' columns can be dropped now.

3. Splitting columns

From the existing set of columns, new categories were created to classify customers:

CATEGORY BY AGE

```
ALTER TABLE bank_df
ADD age_group VARCHAR(20);

UPDATE bank_df
SET age_group =
      CASE
    WHEN age >= 18 AND age <= 25 THEN 'Young Adult'
    WHEN age > 25 AND age <= 34 THEN 'Early Career'
    WHEN age > 34 AND age <= 54 THEN 'Mid Career'
    WHEN age > 54 AND age <= 64 THEN 'Pre-Retirement'
    WHEN age > 64 THEN 'Retirement'
  END;
```

CATEGORY BY QUARTER
Next, a quarter column was created from the month contacted *to identify seasonal trends*.

```
ALTER TABLE bank_df
ADD quarter VARCHAR(20);
```

```
UPDATE bank_df
SET quarter =
  CASE
    WHEN contact_month IN ('jan', 'feb', 'mar') THEN 'Q1'
    WHEN contact_month IN ('apr', 'may', 'jun') THEN 'Q2'
    WHEN contact_month IN ('jul', 'aug', 'sep') THEN 'Q3'
    WHEN contact_month IN ('oct', 'nov', 'dec') THEN 'Q4'
  END;
```

We are now ready for exploratory data analysis.

Exploratory Data Analysis

In this part, we will investigate whether the campaign met its objectives and identify correlations between variables to pinpoint areas for improvement in future campaigns.

Key Variables to Investigate

- **credit_default**: Whether the customer has credit in default. (*yes/no*)
- **poutcome**: Result of the previous marketing campaign. (*success/failure*)
- **deposit**: Whether the customer subscribed to a term deposit. (*yes/no*)

Distribution of Variables:

1. Customer demographics (*age, income level*)
2. Product features (*interest rate, deposit term*)
3. Campaign channel (*email, phone call, social media*)

1. How many customers subscribed to a term deposit after the campaign?

```
SELECT deposit, COUNT(*) AS total_subscribed
FROM bank_df
GROUP BY deposit
```

It appears that only a small fraction responded positively (*yes*), indicating their limited interest in the new product.

deposit	total_subscribed
yes	5289
no	39922

2. What percentage of customers subscribed to a term deposit after the marketing campaign?

Only 11.7% of the 45,211 customers subscribed.

```
SELECT COUNT(*) AS total_subscribed,
  (COUNT(*) /
      (SELECT COUNT(*) FROM bank_df)
  )*100 AS percent_subscribed
FROM bank_df
WHERE deposit = 'yes';
```

Additionally, 83% of customers involved in the campaign have a balance above 0:

```
SELECT SUM(CASE WHEN balance > 0 THEN 1 ELSE 0 END) positive_bal,
      SUM(CASE WHEN balance <= 0 THEN 1 ELSE 0 END) negative_bal
FROM bank_df;
```

total_subscribed	percent_subscribed
5289	11.6985

positive_bal	negative_bal
37931	7280

3. What is the distribution of customers who subscribed to the term deposit based on their balance?

```
SELECT
  SUM(CASE WHEN deposit = 'yes' AND balance > 0 THEN 1 ELSE 0 END) AS
positive_subscribed,
  SUM(CASE WHEN deposit = 'no' AND balance > 0 THEN 1 ELSE 0 END) AS
positive_not_subscribed,
  SUM(CASE WHEN deposit = 'yes' AND balance <= 0 THEN 1 ELSE 0 END) AS
negative_subscribed,
  SUM(CASE WHEN deposit = 'no' AND balance <= 0 THEN 1 ELSE 0 END) AS
negative_not_subscribed
FROM bank_df;
```

positive_subscribed	positive_not_subscribed	negative_subscribed	negative_not_subscribed
4787	33144	502	6778

While only 1.1% of subscribers had a negative balance, 10.5% subscribed with a positive balance.

Banks targeting customers with low balance:

```
SELECT
  COUNT(contact_count) AS times_called,
  balance
FROM bank_df
WHERE deposit = 'yes'
GROUP BY balance
ORDER BY times_called DESC;
```

times_called	balance
292	0
17	1
16	2
15	5
12	3

These customers are not suitable for future campaigns, as they may not be reliable for long-term deposits. Campaigns should instead focus on customers with positive bank balances.

4. What are the demographic statistics of customers who subscribed to the term deposit despite having a negative balance?

Management and blue-collar workers are more likely to subscribe to a term deposit despite a negative balance:

```
SELECT job, COUNT(*) AS count
FROM bank_df
WHERE balance <= 0
  AND deposit = 'yes'
GROUP BY job
ORDER BY count DESC;
```

Most subscribers are in the **mid to early career group (ages 25 to 54)** and are likely seeking the best investment options for their future:

```
SELECT age_group, COUNT(*) count
FROM bank_df
WHERE balance <= 0
AND deposit = 'yes'
GROUP BY age_group
ORDER BY count DESC;
```

A negative balance may indicate *financial liabilities, family expenses, or unsettled loans*.

Thus, these customers are unlikely to subscribe to a term deposit and will not be considered for further analysis.

age_group	count
Mid Career	258
Early Career	152
Pre-Retirement	55
Retirement	19
Young Adult	18

5. Is there any impact of loan status on the subscription rates for term deposits?

- Entrepreneurs, self-employed, students, and housemaids are the weak targets for campaigns due to their lack of fixed incomes.
- The majority (61%) of subscribers have no loans on them.
- Customers with personal or housing loans are less likely to subscribe to a term deposit plan.

loan_type	count	percent_subscribed
personal-loan	177	3.70
housing-loan	1664	34.76
no loan	2946	61.54

```
SELECT
  loan_type,
  COUNT(*) AS count,
  ROUND(COUNT(*) /
    (SELECT COUNT(*)
    FROM bank_df
    WHERE deposit = 'yes'
      AND balance > 0) * 100, 2) AS percent_subscribed
FROM bank_df
WHERE balance > 0
  AND deposit = 'yes'
GROUP BY loan_type
ORDER BY count;
```

6. Are there any seasonal trends for customer subscriptions?

- **Q2 (Second Quarter)** sees the most contacts (over 34%).
- **May** has the highest customer interactions.

Further analysis reveals that over 66% of subscribers were contacted for the first time in this campaign. (*pdays = -1*)

pdays	count
-1	534
87	9
91	5
171	1

Is there a correlation between the frequency of customer contacts and their subscription rate?

```
WITH cte AS (
  SELECT
    contact_month,
    COUNT(*) AS contact_count,
    SUM(CASE WHEN deposit = 'yes' THEN 1 ELSE 0 END) AS subscribed,
    SUM(CASE WHEN deposit = 'no' THEN 1 ELSE 0 END) AS not_subscribed
  FROM bank_df
  GROUP BY contact_month
)
SELECT
  contact_month,
  contact_count,
  subscribed,
  not_subscribed,
  (subscribed / contact_count) AS call_to_subscriber_ratio
FROM cte
ORDER BY call_to_subscriber_ratio DESC;
```

contact_month	contact_count	subscribed	not_subscribed	call_to_subscriber_ratio
mar	477	248	229	0.5199
dec	214	100	114	0.4673
sep	579	269	310	0.4646
oct	738	323	415	0.4377
apr	2932	577	2355	0.1968
feb	2649	441	2208	0.1665
aug	6247	688	5559	0.1101
jun	5341	546	4795	0.1022
nov	3970	403	3567	0.1015
jan	1403	142	1261	0.1012
jul	6895	627	6268	0.0909
may	13766	925	12841	0.0672

- *March, December, September, and October* show a surprisingly **high call-to-subscriber ratio** of 50%, despite lower customer interactions.

This insight suggests that banks should increase outreach efforts at the end and beginning of the year to boost subscriber growth.

7. Are there significant variations in the subscription ratio among customers?

```
WITH cte AS (
  SELECT
    job,
    SUM(CASE WHEN deposit = 'yes' THEN 1 ELSE 0 END) AS subscribed,
    SUM(CASE WHEN deposit = 'no' THEN 1 ELSE 0 END) AS not_subscribed
  FROM bank_df
  GROUP BY job
)
SELECT
  job,
  subscribed,
  not_subscribed,
  (subscribed / not_subscribed)*100 AS ratio
FROM cte
ORDER BY ratio DESC;
```

- *Students and retired professionals* are more likely to subscribe to the term deposits.
- The *unemployed* rank third but are not potential targets due to income instability.

job	subscribed	not_subscribed	ratio
student	269	669	40.2093
retired	516	1748	29.5195
unemployed	202	1101	18.3470
management	1301	8157	15.9495
admin.	631	4540	13.8987
self-employed	187	1392	13.4339

Banks can focus on **students and retirees** as a part of their future marketing campaigns to increase the success rates.

Are these customers (students and retirees) actually a reliable choice for term deposits?

```
SELECT job, credit_default, COUNT(*) AS count
FROM bank_df
WHERE deposit = 'yes'
  AND job IN ('student', 'retired')
GROUP BY job, credit_default
ORDER BY job, count DESC;
```

job	credit_default	count
retired	no	515
retired	yes	1
student	no	269

contact_day	contact_count	subscribed	not_subscribed	call_to_subscriber_ratio
1	322	90	232	0.2795
10	524	121	403	0.2309
30	1566	271	1295	0.1731
22	905	154	751	0.1702
3	1079	178	901	0.1650
4	1445	230	1215	0.1592

- It seems students and retired professionals do not have any credit defaults, making them *strong candidates* for future campaigns.
- Having no credit default indicates that these customers are reliable and can consider long-term investments, which benefits banks.
- These customers are more likely to subscribe to term deposits **at the beginning of the month**, **but contact rates are low**.
- Banks should put more efforts on targeted marketing campaigns in the first weeks to attract more customers.

8. Are there any returning subscribers from previous campaigns?

poutcome	contact_count	subscribed	not_subscribed	call_to_subscriber_ratio
success	1511	978	533	0.6473
other	1840	307	1533	0.1668
failure	4901	618	4283	0.1261
unknown	36959	3386	33573	0.0916

Customers who showed interest in past campaigns are more likely to subscribe to term deposits in the new campaign, making them strong targets for future campaigns.

Is there a relationship between the average yearly balance and the loan status of subscribers?

```
SELECT loan_type, MIN(balance), MAX(balance)
FROM bank_df
WHERE deposit='yes'
GROUP BY loan_type
ORDER BY MAX(balance) DESC
```

As expected, there's a **positive correlation** between customers' outstanding loans and their average yearly balance. Customers without active loans maintain a higher average yearly balance.

loan_type	MIN(balance)	MAX(balance)
no loan	-546	81204
housing-loan	-3058	45248
personal-loan	-799	22125

9. How do past customer interactions affect the campaign outcomes?

```
WITH cte AS (
  SELECT DISTINCT poutcome, pdays
  FROM bank_df
)
SELECT poutcome,
  AVG(pdays) AS avg_pdays
FROM cte
```

```
GROUP BY poutcome
ORDER BY AVG(pdays) DESC;
```

Frequent contact shows higher success rates, while those who were contacted a long time ago had lower outcomes. Banks should **prioritize follow-ups with older and active customers** to maintain high turnover rates.

poutcome	avg_pdays
success	230.5647
unknown	232.8333
other	249.9896
failure	283.3737

10. What are the conversion rates across different customer segments by age group and quarter?

```
SELECT
    quarter,
    age_group,
    ROUND(AVG(CASE WHEN deposit = 'yes' THEN 1 ELSE 0 END) * 100, 2) AS
success_conversion_rate,
    ROUND(AVG(CASE WHEN deposit = 'yes' AND pdays = -1 THEN 1 ELSE 0 END) *
100, 2) AS new_customer_conversion_rate,
    ROUND(AVG(CASE WHEN deposit = 'yes' AND poutcome LIKE '%failure%' THEN 1
ELSE 0 END) * 100, 2) AS potential_customer_conversion_rate
FROM bank_df
GROUP BY age_group, quarter
ORDER BY success_conversion_rate DESC;
```

quarter	age_group	success_conversion_rate	new_customer_conversion_rate	potential_customer_conversion_rate
Q1	Young Adult	55.56	33.33	4.04
Q4	Young Adult	49.37	22.78	7.59
Q1	Retirement	48.13	27.81	4.81
Q4	Retirement	44.20	21.55	6.08
Q2	Retirement	42.51	26.35	2.40
Q3	Retirement	36.36	21.82	3.64
Q3	Young Adult	29.65	21.14	1.26
Q1	Pre-Retirement	22.44	11.99	2.71

- Marketing campaigns were **most effective in Q1 for young adults** (18-25), with a 55.5% *customer-turned-subscriber* rate.
- The **lowest rate** (7%) was for **mid-career customers in Q2**.

Stakeholders can use this data to optimize targeted marketing for specific age groups at the appropriate time of the year.

Conclusion & Key Insights

- **Q1 had the highest conversion rate (55.5%)** for young adults (18-25), making it the most effective time to target this age group.
- **Q2 had the lowest conversion rate (7%)** for mid-career individuals (35-55), which suggests room for improvement in strategies targeting this demographic.
- Campaigns in **March, September, October, and December** were the most successful, with high conversion rates despite lower interactions. Banks should increase their outreach during these months for high turnover rates.
- Customers in **management and blue-collar jobs**, even with negative bank balances, should be targeted for future campaigns.

Recommendations

- Prioritize follow-ups with older customers and those interested in past campaigns for higher subscription rates.
- Focus on engaging with customers with positive balances and no active loans, as well as those in management and blue-collar jobs.
- Increase outreach efforts in March, September, October, and December for better results.
- Target young adults (18-25) during Q1 for higher conversion rates.
- Improve existing strategies for mid-career individuals (35-55) to boost their engagement in Q2.