



Virtual Internship (Data Science)

Data Intake Report

Group Name: Project Group 1

Members:

No	Name	Email	Country	College/company	Specialization
1	Preeti Verma	vermapreeti.dataanalyst@gmail.com	Canada	-	Data Science
2	Thanuja Modiboina	thanujayadav953@gmail.com	UK	-	Data Science
3	Abishek James	abishekjames1998@gmail.com	Ireland	-	Data Science

Name: Bank Marketing (Campaign)

Report date: 13-04-2023

Internship Batch: LISUM19

Data intake by:

Data intake reviewer: Data Glacier

Data storage location:

Problem Description :

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps to understand whether a particular customer will buy their product or not (based on the customer's past interaction with the bank or other Financial Institution). This is an application of the company's marketing data.

Business Understanding :

The goal is to build a Machine Learning model that helps in predicting the outcomes of each customer's marketing campaign and analyzing which features have an impact on the outcomes will help the company to understand how to make the campaign more effective. Additionally, categorizing the customer group that subscribed to the term deposit helps to determine who is more likely to purchase the product in the future, thereby developing more targeted marketing campaigns.

This can be accomplished by using an ML model that shortlists the customers whose possibility of purchasing the product is higher. So, marketing such as telemarketing, SMS or email marketing can concentrate only on those customers. It will save time and resources by doing this.

Project Lifecycle

Deadline (Date/week)	Plan and Deliverables
19 April 2023(Week 7)	<ul style="list-style-type: none">● Problem statement● Business understanding● Dataset collection
26 April 2023(Week 8)	<ul style="list-style-type: none">● Data understanding● Data analysis - finding null values, and outliers.● Data processing
2 May 2023(Week 9)	Data cleaning and transformation
9 May 2023(Week 10)	EDA and Model Recommendation
16 May 2023(Week 11)	EDA Presentation and Proposed Modeling Technique

23 May 2023(Week 12)	Model selection and Building the model
30 May 2023(Week 13)	Final project report and code submission

Tabular data details:

File 1: bank_additional_full.csv

Total number of observations	41189
Total number of files	2
Total number of features	21
Base format of the file	.CSV
Size of the data	5836800(5.56MB)

File 2: bank_additional.csv

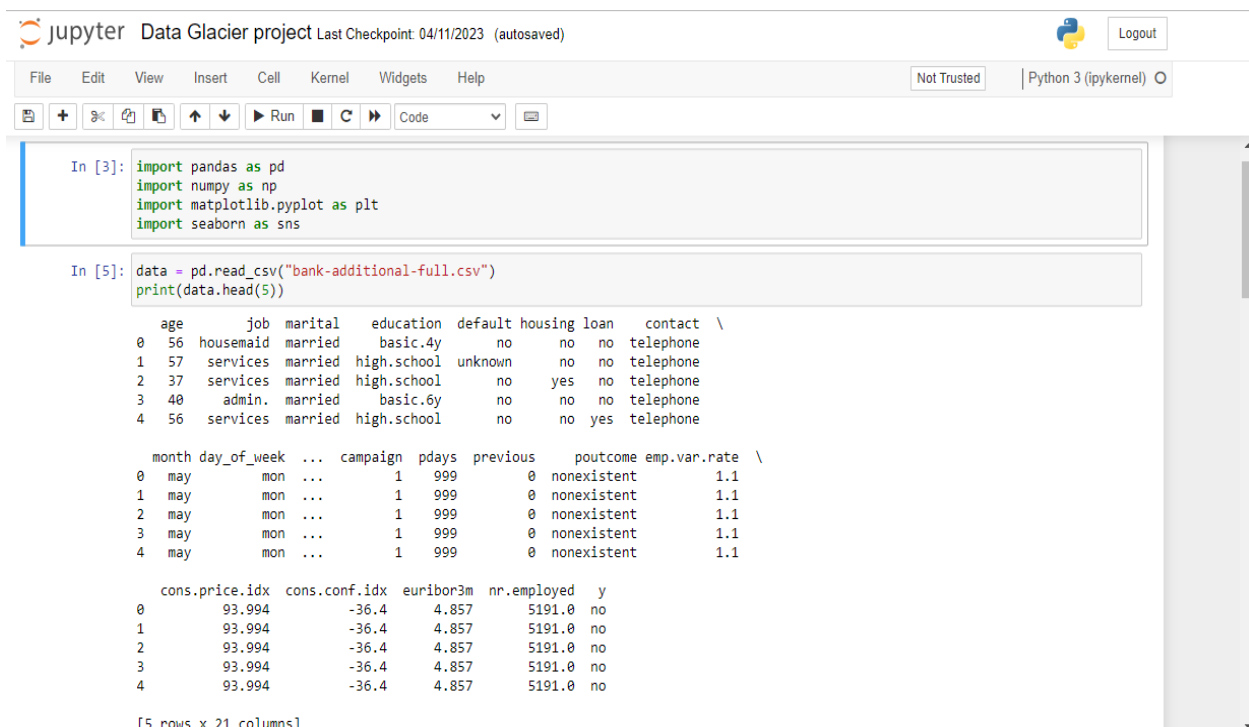
Total number of observations	4120
Total number of files	2
Total number of features	21

Exploratory Data Analysis

1. The data covers the period from May 2008 to November 2010.
2. There are 2 datasets, the second dataset is a sample of the first dataset.
3. There are 10 integers and 11 categorical variables.
4. The missing values in both datasets are presented by an "unknown" string. We changed it to NaN.
5. There are missing values in six variables namely, job, marital status, education, default, housing, and loan. This will be imputed using various methods.
6. There are 12 duplicates in the first dataset and no duplicates in the sample dataset, this will be dropped since they are minimal and will not affect our analysis

Assumptions

We assume the data provided is correct and up to date.



The screenshot shows a Jupyter Notebook titled "Data Glacier project" with a last checkpoint of "04/11/2023 (autosaved)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and code execution. The notebook contains two code cells. The first cell imports the necessary libraries: pandas, numpy, matplotlib.pyplot, and seaborn. The second cell reads a CSV file named "bank-additional-full.csv" and prints the first five rows of the resulting DataFrame. The output shows two tables of data. The first table has columns: age, job, marital, education, default, housing, loan, and contact. The second table has columns: month, day_of_week, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, and y. The output indicates that there are 5 rows and 21 columns in the dataset.

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

In [5]: data = pd.read_csv("bank-additional-full.csv")
print(data.head(5))
```

	age	job	marital	education	default	housing	loan	contact	
0	56	housemaid	married	basic.4y	no	no	no	telephone	
1	57	services	married	high.school	unknown	no	no	telephone	
2	37	services	married	high.school	no	yes	no	telephone	
3	40	admin.	married	basic.6y	no	no	no	telephone	
4	56	services	married	high.school	no	no	yes	telephone	

	month	day_of_week	...	campaign	pdays	previous	poutcome	emp.var.rate	
0	may	mon	...	1	999	0	nonexistent	1.1	
1	may	mon	...	1	999	0	nonexistent	1.1	
2	may	mon	...	1	999	0	nonexistent	1.1	
3	may	mon	...	1	999	0	nonexistent	1.1	
4	may	mon	...	1	999	0	nonexistent	1.1	

	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	93.994	-36.4	4.857	5191.0	no
1	93.994	-36.4	4.857	5191.0	no
2	93.994	-36.4	4.857	5191.0	no
3	93.994	-36.4	4.857	5191.0	no
4	93.994	-36.4	4.857	5191.0	no

[5 rows x 21 columns]

Jupyter Data Glacier project Last Checkpoint: 04/11/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

4 93.994 -36.4 4.857 5191.0 no

[5 rows x 21 columns]

```
In [6]: print(data.dtypes)
```

age	int64
job	object
marital	object
education	object
default	object
housing	object
loan	object
contact	object
month	object
day_of_week	object
duration	int64
campaign	int64
pdays	int64
previous	int64
poutcome	object
emp.var.rate	float64
cons.price.idx	float64
cons.conf.idx	float64
euribor3m	float64
nr.employed	float64
y	object
dtype:	object

```
In [7]: data.isna().sum()
```

Jupyter Data Glacier project Last Checkpoint: 04/11/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

```
In [7]: data.isna().sum()
```

```
Out[7]: age      0
job      0
marital   0
education 0
default   0
housing   0
loan      0
contact   0
month     0
day_of_week 0
duration  0
campaign  0
pdays    0
previous  0
poutcome  0
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
y          0
dtype: int64
```

```
In [9]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
#   Column              Non-Null Count  Dtype  
```

Home Page - x Data Glacier - x Data ingestio - x understanding - x ThanujaModi - x ThanujaModi - x how to give g - x Jodebu/uplo - x + - □ ×

localhost:8888/notebooks/Data%20Glacier%20project.ipynb

jupyter Data Glacier project Last Checkpoint: 04/11/2023 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

In [9]: data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype  
---  --
0   age                    41188 non-null  int64  
1   job                    41188 non-null  object  
2   marital                41188 non-null  object  
3   education              41188 non-null  object  
4   default                41188 non-null  object  
5   housing                41188 non-null  object  
6   loan                   41188 non-null  object  
7   contact                41188 non-null  object  
8   month                  41188 non-null  object  
9   day_of_week            41188 non-null  object  
10  duration                41188 non-null  int64  
11  campaign                41188 non-null  int64  
12  pdays                   41188 non-null  int64  
13  previous                41188 non-null  int64  
14  poutcome                41188 non-null  object  
15  emp.var.rate            41188 non-null  float64 
16  cons.price.idx           41188 non-null  float64 
17  cons.conf.idx            41188 non-null  float64 
18  euribor3m                41188 non-null  float64 
19  nr.employed              41188 non-null  float64 
20  y                        41188 non-null  object  
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```