

Team Detail: Overview

Group Name: Data Glacier Team

Member:

- Bogdan-Remus Pintilie
 - Email: bogdanremuspintilie@gmail.com
 - Country: Romania
 - Status: College Student of Warwick University
 - Intern Path: Data Analyst
- Raka Prasetya Nugraha
 - Email: careermagic780@gmail.com
 - Country: Indonesia
 - Status: Professional Data Analyst
 - Intern Path: Data Analyst
- Parwinder Singh
 - Email: parsingh048@gmail.com
 - Country: U.S.
 - Status: Recent Graduate
 - Intern Path: Data Analyst

Problem Description:

XYZ Credit Union, a successful bank in Latin America, is facing a challenge with cross-selling their banking products. While they have been successful in selling individual products such as credit cards, deposit accounts, retirement accounts, and safe deposit boxes, their existing customers are not purchasing more than one product. This indicates a lack of success in cross-selling their other offerings to their current customer base.

Business Understanding:

XYZ Credit Union offers a range of products and services including credit cards, deposit accounts, retirement accounts, safe deposit boxes, and more. However, they are struggling to effectively sell multiple products to their existing customers. The bank is seeking our assistance in analyzing their product offerings to determine whether they should focus on up-selling or cross-selling. The goal is to identify which products should be prioritized for up-selling (encouraging customers to upgrade or purchase higher-tier versions of their existing product) or cross-selling (offering additional products to existing customers).

Project Timeline:

Week 7:

- Understand the problem description and business understanding
- Create the project timeline
- Prepare the project report and push it to the repository

Week 8:

- Deepen the business understanding and explore the data
- Analyze the chosen approach based on the available data to solve the problem
- Prepare the project report and push it to the repository

Week 9:

- Implement ETL (Extract, Transform, Load) processes for data cleaning and transformation
- Perform data filtering (handling outliers and basic calculations) and clustering for data categorization
- Hold a progress meeting to discuss the results and update the project report

Week 10:

- Determine the relevant data and parameters needed for creating visualizations
- Ensure the data obtained is sufficient to support or solve the problem
- Update and enhance the project report

Week 11:

- Develop visualizations to present the findings
- Conduct a progress meeting to review the results and gather feedback
- Update the project report

Week 12:

- Finalize the visual and data dashboards
- Prepare presentations of the results
- Update the project report

Week 13:

- Review and update the project report
- Perform final checks and refinements
- Complete the project presentations and deliver the final report

Data Intake Report

Name: Data Analyst: Cross selling recommendation - Group Project

Report date:

Internship Batch: LISUM21

Version: 1.0

Data intake by: Bogdan-Remus Pintilie

Data intake reviewer: Raka Prasetya Nugraha, Parwinder Singh, Kevin Wang

Data storage location:

<https://drive.google.com/file/d/1OCOZGjoL14tgKZtCl4Q9cifuL8xwIAo6/view?usp=sharing>

<https://drive.google.com/file/d/1wCwF2D6lzxh50rBFWXwW330ByiHhVK0w/view?usp=sharing>

Tabular data details:

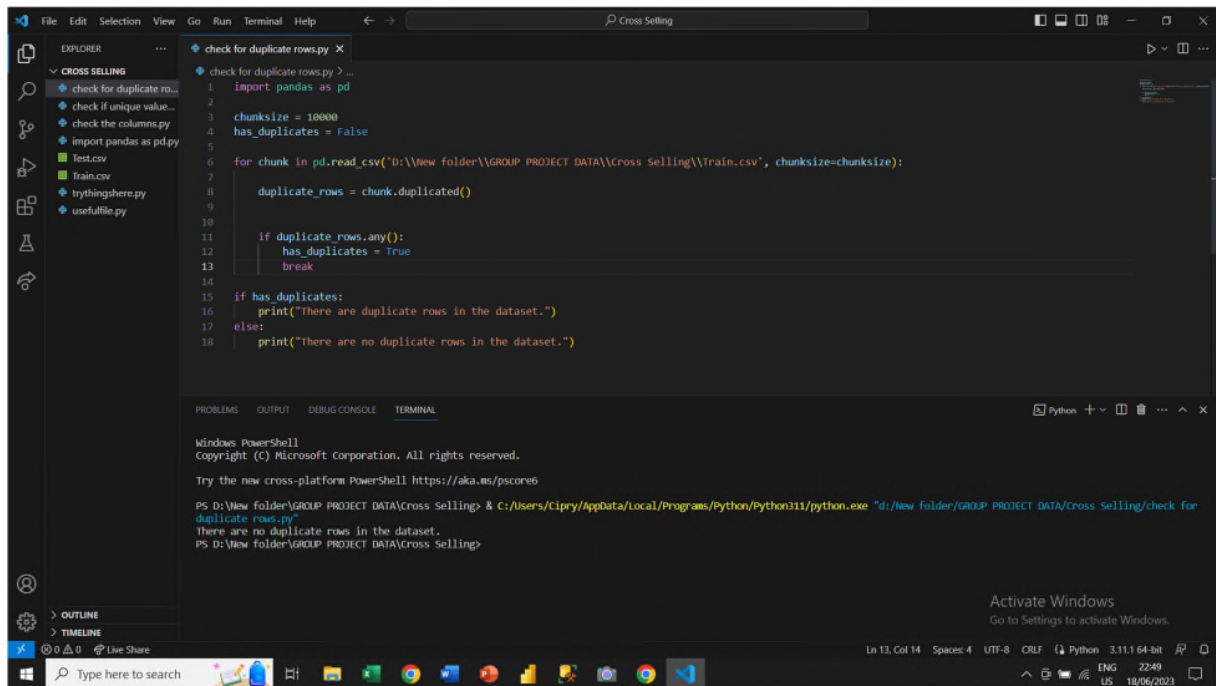
Total number of observations	929615
Total number of files	1 (Test.csv)
Total number of features	24
Base format of the file	.csv
Size of the data	105.2 MB

Total number of observations	13647309
Total number of files	1 (Train.csv)
Total number of features	48
Base format of the file	.csv
Size of the data	2.14 GB

Proposed Approach:

1. Independence: The assumption of independence suggests that the data points being analyzed are not influenced by each other. Independence allows for unbiased analysis and statistical tests that rely on the assumption of independence.
2. Validity: The assumption of validity implies that the data represents what it claims to represent. Valid data is relevant and appropriate for the analysis objectives, and it accurately captures the intended information or measurements.
3. Integrity: The assumption of data integrity suggests that the data has not been tampered with or altered inappropriately. Data integrity ensures that the data retains its original state and remains reliable throughout the analysis process.
4. The dataset does not contain any duplicate rows and to find this we used the code provided in the first picture below.
5. Duplicate identification in the key field: The second picture contains the code that shows the key field (ncodpers) contains duplicates.

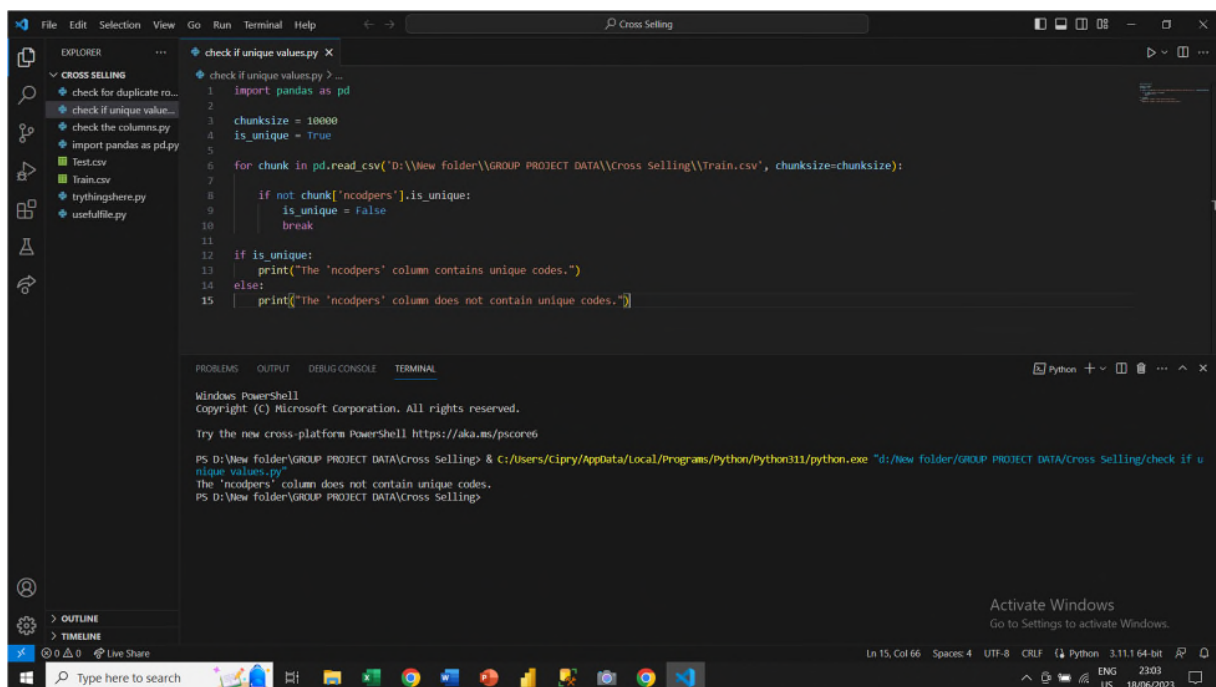
6. We improve data relevance by omitting the following fields since they do not provide valuable insights and do not help address the problem statement: conyuemp



The screenshot shows the Visual Studio Code interface with a file explorer on the left showing a project named 'CROSS SELLING'. The main editor displays a Python script named 'check for duplicate rows.py'. The script imports pandas and iterates through chunks of a CSV file to check for duplicate rows. The terminal at the bottom shows the command to run the script and its output, which states that there are no duplicate rows in the dataset.

```
1 import pandas as pd
2
3 chunksize = 10000
4 has_duplicates = False
5
6 for chunk in pd.read_csv('D:\\New folder\\GROUP PROJECT DATA\\Cross Selling\\Train.csv', chunksize=chunksize):
7
8     duplicate_rows = chunk.duplicated()
9
10
11     if duplicate_rows.any():
12         has_duplicates = True
13         break
14
15 if has_duplicates:
16     print("There are duplicate rows in the dataset.")
17 else:
18     print("There are no duplicate rows in the dataset.")
```

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Try the new cross-platform PowerShell <https://aka.ms/pscore6>
PS D:\New folder\GROUP PROJECT DATA\Cross Selling> C:\Users\Cipry\AppData\Local\Programs\Python\Python311\python.exe "d:\New folder\GROUP PROJECT DATA\Cross Selling\check for duplicate rows.py"
There are no duplicate rows in the dataset.
PS D:\New folder\GROUP PROJECT DATA\Cross Selling>



The screenshot shows the Visual Studio Code interface with a file explorer on the left showing a project named 'CROSS SELLING'. The main editor displays a Python script named 'check if unique values.py'. The script imports pandas and iterates through chunks of a CSV file to check for unique values in the 'ncodpers' column. The terminal at the bottom shows the command to run the script and its output, which states that the 'ncodpers' column does not contain unique codes.

```
1 import pandas as pd
2
3 chunksize = 10000
4 is_unique = True
5
6 for chunk in pd.read_csv('D:\\New folder\\GROUP PROJECT DATA\\Cross Selling\\Train.csv', chunksize=chunksize):
7
8     if not chunk['ncodpers'].is_unique:
9         is_unique = False
10         break
11
12 if is_unique:
13     print("The 'ncodpers' column contains unique codes.")
14 else:
15     print("The 'ncodpers' column does not contain unique codes.")
```

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.
Try the new cross-platform PowerShell <https://aka.ms/pscore6>
PS D:\New folder\GROUP PROJECT DATA\Cross Selling> C:\Users\Cipry\AppData\Local\Programs\Python\Python311\python.exe "d:\New folder\GROUP PROJECT DATA\Cross Selling\check if unique values.py"
The 'ncodpers' column does not contain unique codes.
PS D:\New folder\GROUP PROJECT DATA\Cross Selling>

Data Descriptions

Column Name	Description
fecha_dato	The table is partitioned for this column
ncodpers	Customer code
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia	Customer's Country residence
sexo	Customer's sex
age	Age
fecha_alta	The date in which the customer became as the first holder of a contract in the bank
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
antiguedad	Customer seniority (in months)
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)
indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential primary), 4(former co-owner)
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
conyuemp	Spouse index. 1 if the customer is spouse of an employee
canal_entrada	channel used by the customer to join
indfall	Deceased index. N/S
tipodom	Addres type. 1, primary address
cod_prov	Province code (customer's address)
nomprov	Province name
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household
segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated

Column Name	Description
ind_ahor_fin_ult1	Saving Account
ind_aval_fin_ult1	Guarantees
ind_cco_fin_ult1	Current Accounts
ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	Medium-term deposits
ind_dela_fin_ult1	Long-term deposits
ind_ecue_fin_ult1	e-account
ind_fond_fin_ult1	Funds
ind_hip_fin_ult1	Mortgage
ind_plan_fin_ult1	Pensions
ind_pres_fin_ult1	Loans
ind_reca_fin_ult1	Taxes
ind_tjcr_fin_ult1	Credit Card
ind_valo_fin_ult1	Securities
ind_viv_fin_ult1	Home Account
ind_nomina_ult1	Payroll
ind_nom_pens_ult1	Pensions
ind_recibo_ult1	Direct Debit

The product:

Credit card, deposit account, retirement account, and safe deposit boxes