Data Glacier - Week 8

Team Members:

- Ray Ng
  - rayng2018@gmail.com
  - Canada
  - University of British Columbia
  - Specialization: Data Science
- Rita Uzoka
  - rita.uzoka@yahoo.com
  - United Kingdom
  - Sheffield Hallam University
  - Specialization: Data Science
- Fatemeh Bagheri
  - f.bagheri13@gmail.com
  - France
  - Université Jean Monnet St Etienne - Université de Lyon
  - Your Specialization: Data Science

Problem: Customer Segmentation - XYZ bank wants to roll out Christmas offers to their customers. But the bank does not want to roll out the same offer to all customers, instead they want to roll out personalized offers to particular sets of customers. If they manually start understanding the category of customer then this will not be efficient and also they will not be able to uncover the hidden pattern in the data (pattern which groups certain kinds of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 groups as this will be inefficient for their campaign.

Data understanding

Data storage location:
https://drive.google.com/drive/folders/1bfCpJIKmp6IHxiLPWvOS2nU1dc24pViB

Tabular data details:

Total number of observations	1000000
Total number of files	1
Total number of features	48
Base format of the file	.csv
Size of the data	154 MB

The variables are continuous data while the categorical variables have been converted to continuous of 1s and 0s. Hence the data is a continuous data.

Data Issues and Solutions:

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the

recurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions. These are parts of data that should be considered as problems:

50	ZARAGOZA	1		0

NA value

1085568				NA	

- **Data completeness :various columns, such as the ind_empleado(Employee index), pais_residencia(Customer's Country residence), sexo(Customer's sex) etc have empty values in certain records**
    - Mitigation: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.
    - Recommendation: Make input for such fields required, especially easily obtainable ones such as customer sex
- **Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)**
    - Mitigation: Make data transformations to ensure consistent data types in a column. Examples: Convert selected records in characters to numeric; Remove non-numeric characters from the string.
    - Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at the later stage.
- **Inconsistent values for the same attribute (e.g. -999999 as a value for seniority in months)**
    - Mitigation: The seniority in months cant have a negative number,Use regular numbers to replace extended values into positive numbers or zero to ensure consistency across seniority in months.
    - Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. Make sure only positive integers (0,1,2,...) are valid.

Outliers are often detected through graphical means, though you can also do so by a variety of statistical methods using your favorite tool. (Excel and R will be referenced heavily here, though SAS, Python, etc., all work).

5 ways to deal with outliers in data

1. Set up a filter in your testing tool
2. Remove or change outliers during post-test analysis
3. Replace outliers with appropriate values where context is known
4. Consider the underlying distribution of each variable
5. Consider the value of mild outliers - What is the effect of these outliers on the analysis?

GitHub Repo Link:
https://github.com/faba13/VC.git