Week 8 Delivereables

Members: Uthej Goud
Email: goudtigulla@gmail.com
Country: Ireland
College: Griffith College Dublin
Specialization: Data Science
Submitted To: Data Glacier
Date: 24/07/2021
Batch Name: LISUM01

Problem Description:

XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also they will not be able to uncover the hidden pattern in the data (pattern which group certain kind 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 group as this will be inefficient for their campaign.

Data Understanding:

- The provided dataset has 1000000 observations and 47 columns.
- This is a unsupervised machine learning problem, since there is no classification or target variable, the data is to be segmented into groups to provide targeted advertisements.
- **Type of data**
  - The dataset consists of mainly ordinal columns and some integer columns
  - The dataset also consists of 2 datetime columns
  - Ordinal and datetime variables are not that useful for unsupervised learning methods such as clustering, while they are great for association.
  - In the case of this problem we have to find clusters rather than find patterns in the data i.e association, so these columns will be dealt with in Data Cleaning.

Problems with Data:

- The columns ult_fec_cli_1t and conyuemp have 99% null values.
- The renta coulmn represents the gross income of the family, and has about 17% of null values
- There are null values in other columns but they amount to around 1% of the dataset

- Few columns have wrong dtypes . have been changed manually, such as datetime columns have been assigned correct dtypes and certain float items have been assigned int dtypes on a case to case basis.

Data Cleaning Approaches:

- The columns ult_fec_cli_1t and conyuemp have 99% null values so the best action would be to drop these two columns as there are not many insights that can be gained for this problem, and if these columns are used it can create a skewed model.
- The null values in renta column can be replaced by the average of column.
- There are null values in other columns but they amount to around 1% of the dataset so I have decided to remove them from the dataset, as they are overlapping null values and replacing them is not possible since they are incomplete observations.
- All columns have been assigned correct dtypes.
- Outilers from columns such as antiguedad have been removed.

Github Repo link: https://github.com/uthej12/Customer_Segmentation