Group Name: Team Coltenback

Name: Richard Coltenback

Email: rdcoltenback@gmail.com

Country: United States of America

College: Drew University

Specialization: Data Analyst

GitHub Repository Link: https://github.com/RColtenback/Cross-Selling-Data-Analysis.git

Problem Description:

XYZ credit union in Latin America is performing very well in selling the banking products (e.g.: credit card, deposit account, retirement account, safe deposit box etc.). However, their existing customers are not buying more than one product which means bank is not performing well in cross-selling (Bank is not able to sell their other offerings to existing customer). XYZ Credit Union decided to approach Team Coltenback to solve their problem.

Data Understanding:

Train.csv

Total number of observations	13647309
Total number of files	1
Total number of features	48
Base format of the file	.csv
Size of the data	2.29 GB

The data we will be looking at has to deal with banking and different products that go along with it. Some examples of the variables are loans, pensions, mortgages, etc. Fortunately, due to the data collected being based on each customer, we also have access to personal traits like their ages, sex, country of residence, etc. This dataset has a mix of different types of variables. Some variables like fecha_dato and fecha_alta are dates. Some variables like indext and indfall are Boolean or dichotomous as they are true/false or yes/no. Some variables like ind_empleado are string or nominal as they have letters for each customer. Finally, some variables like ncodpers and age are integer or scale as they have numbers for each customer.

There are some issues with the dataset. First, not all of the data is filled in. Some of the values of the variables are either missing or set as NA. This isn't too big of an issue for me as I'm planning to analyze the data in ways that don't limit me by losing variables due to their being missing data or NA. I do believe visuals might need some work around so they still depict what is wanting to be shown. Since the dataset is so big, I won't be able to use a friendly program like Tableau, which is easy to work around, so I'll have to be clever. A second issue that could arise is the large number of data being available to us causing slow loading times and trouble processing commands. I could use modin and dask instead of regular pandas but there may be more that needs to be done to optimize the exploration of the dataset. Issues like outliers and skew won't be an issue for me as I will use means and medians in places that need it. I believe keeping the outliers is good practice so people who have the data explore all of it and know everything about it, painting the clearest picture possible. However, should I need to explore the dataset without outliers, I could also set minimum and maximum values on each variable before running analysis.