Member Names: Blaise Papa

Country: Kenya

Institution: Strathmore University

Specialization: Data-Science

Problem Description:

Christmas is around the corner and a certain bank would like to capitalize on the holiday. It would like to send out offers to its customers to increase customer interaction with the bank. The bank does understand it has different customer segments based on their habits. They hence would like to roll out specialized offers for different customers. They however are short time and resource cap they would like to automate the process of understanding how and which offers should be sent to which customers.

Business Understanding

From the problem description, this requires a customer segmentation approach. Customer segmentation involves analyzing customer behavior based on certain metric features, from this analysis we categorize customers into groups of similar behavior. This approach is handy since not all customers have the same needs and patterns, they however have similar actions to a particular customer group.

To achieve this, we build a classifier algorithm based on the data features collected by the bank.

Type of data:

The data collected for this project is customer data from the bank, the data encompasses various features about the customers, the features are mainly numeric and categorical.

Problems:

There exist columns with null values as much as 99% of the records this has to be addressed if we need to get better results from the model.

We also do have a column containing outliers, some machine models do not fare well with the presence of outliers and hence this should be addressed

The data columns also are in Spanish and hence not easily readable when developing the model

Approaches:

We intend to drop columns with null values above the 50 % threshold. Columns that exceed this do not give quality information and imputing the null values might make our data incorrect.

We also drop rows that have outlier samples to improve model performance. The presence of outliers affects model performance especially for linear and non-tree models that are majorly affected.

To bridge the language barrier we shall use google translator API to translate column names, hence understanding them better.

Github Link:

https://github.com/Blvisse/CustomerSegmentation/tree/data-analysis