Customer Segmentation

Name: XYZ Bank

Location: Zimbabwe

Team: DGI

Date: 15-05-2021

Agenda

- Problem Description
- Business Understanding
- Approach
- Data Overview
- EDA Summary
- Proposed Model Technique
- Model Evaluation
- Recommendations

Problem Description

- The Bank XYZ wants to roll out personalized Christmas offers for certain customers instead of rolling out the same offers for all customers. As an alternative of trying to manually decide which customer is which category. The bank seeks an efficient approach that enables them to uncover hidden patterns in their customer data and categorize customers into a 5 unique groups.

Business Understanding

- Customer segmentation is the process of categorizing the customers into various groups according to their characteristics or behaviors.
- This will help the companies effectively match their products to the exact customers groups.

Data Glacier
Your Deep Learning Partner

Approach

For this analysis we will look at these factors:

1. Find out which customer belongs to which group (Group: 1, 2, 3, 4 and 5).

2. Find any hidden patterns in customer behavior that would help the bank match their products to the exact customers groups.

Data Overview

- 1 data set (cust_seg.csv)
- The full data set consists of 48 variables and 1,000,000 observations.

Assumptions:

- The data was skewed so we dropped the outliers.
- Some variables were dropped because they were considered useless i.e. some customers didn't have a province code.
- Variables are arranged in order of importance.

Manipulations:

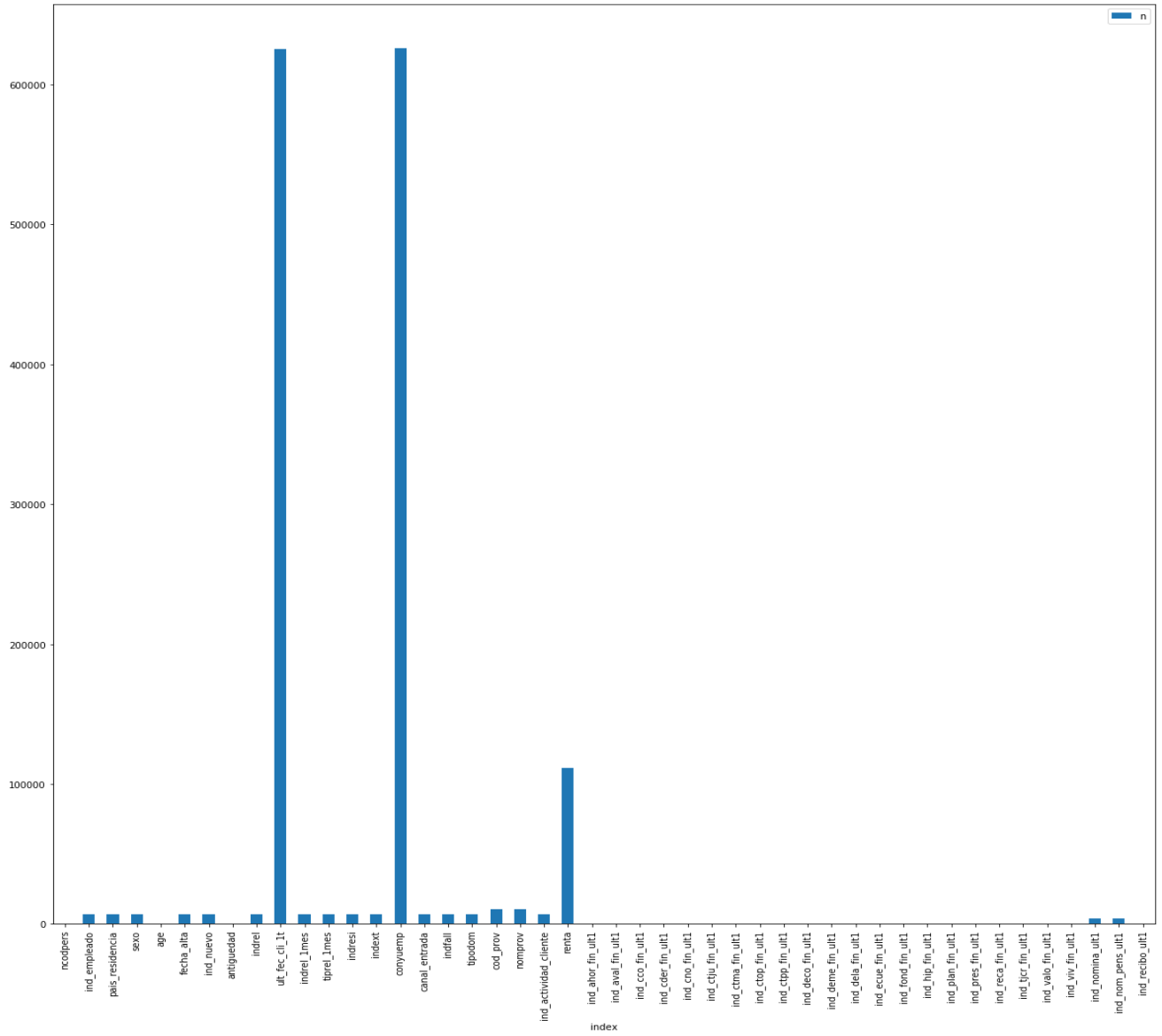
- The fecha_alta variable which is the date which the customer became the first holder of a contact at the bank was changed by subtracting the customer's date with the max date 2015-01-28 to get the number of years since first contract. This created a new column called fecha_alta_year.
- The 24 variables that start with the letters ind_ and end with the letter _ult1 which describe the different accounts or products the customers use was summed up to one variable which counts the number of products each customer makes use of in the bank. This created a new variable named number_of_accounts.

Exploratory Data Analysis

- Duplicate values: approximately 37.3% of the data was duplicates
- Approximately 40% of the data was discarded for the analysis after data cleaning and transformation

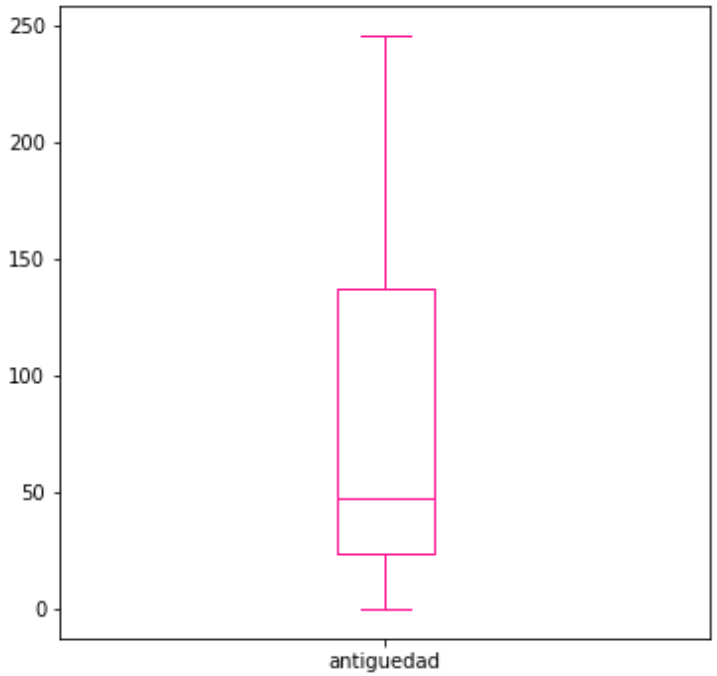
EDA: Missing Values

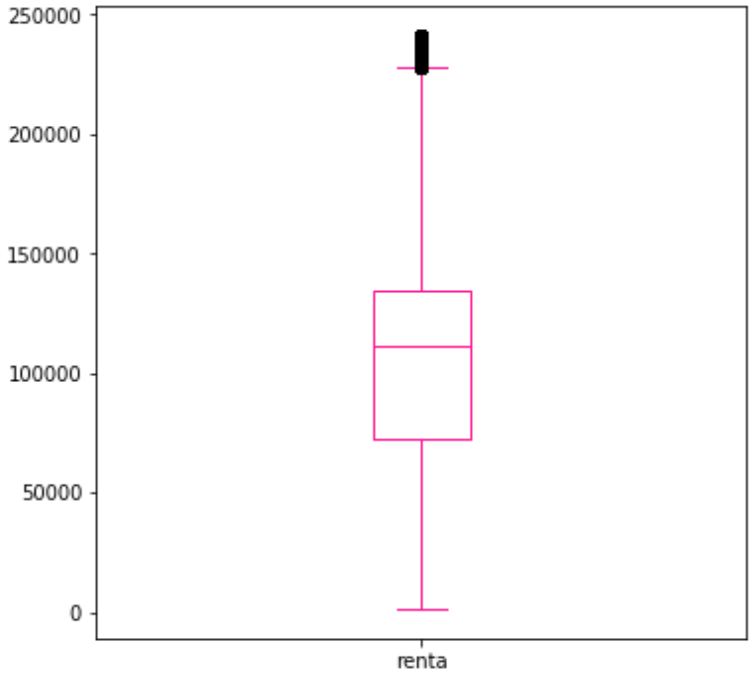
- Variables with more than 80% missing values were removed.
- The other missing values were replace using different machine learning methods.

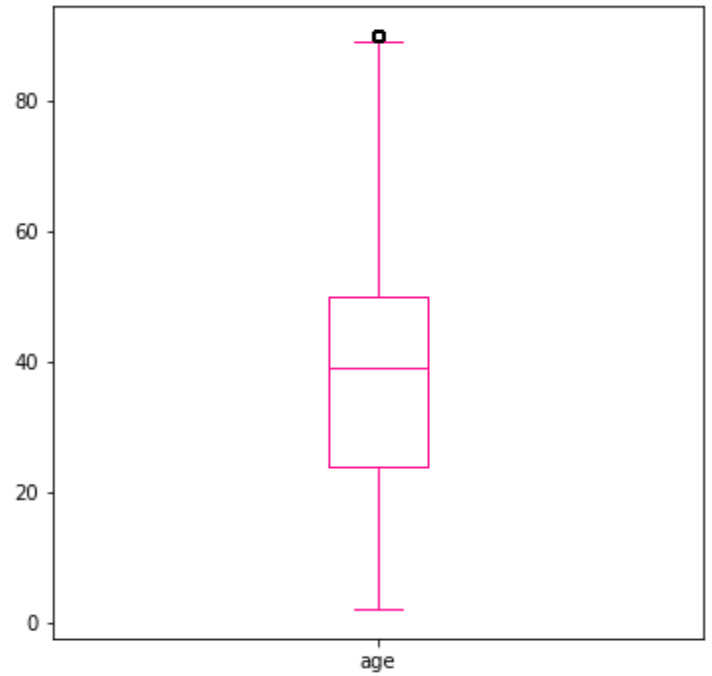


EDA: Outliers

- The renta variable and age variable contained some outliers. Thus, the outliers were removed from this analysis to obtained accurate results.

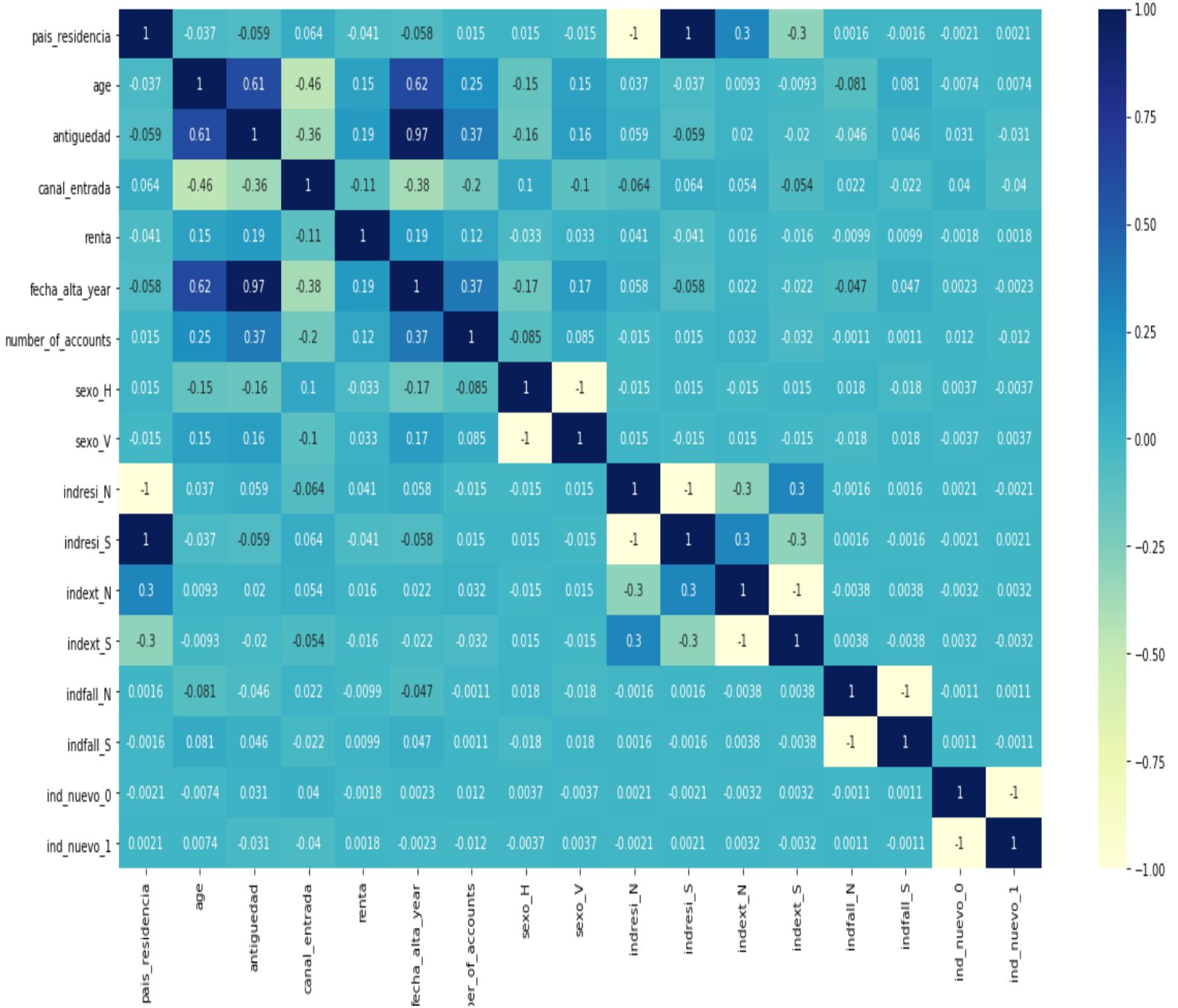






EDA: Correlation

- Most variables in the data set were correlated. For example, pais_residencia and residence index (indresi), indresi depends on pais_residencia.

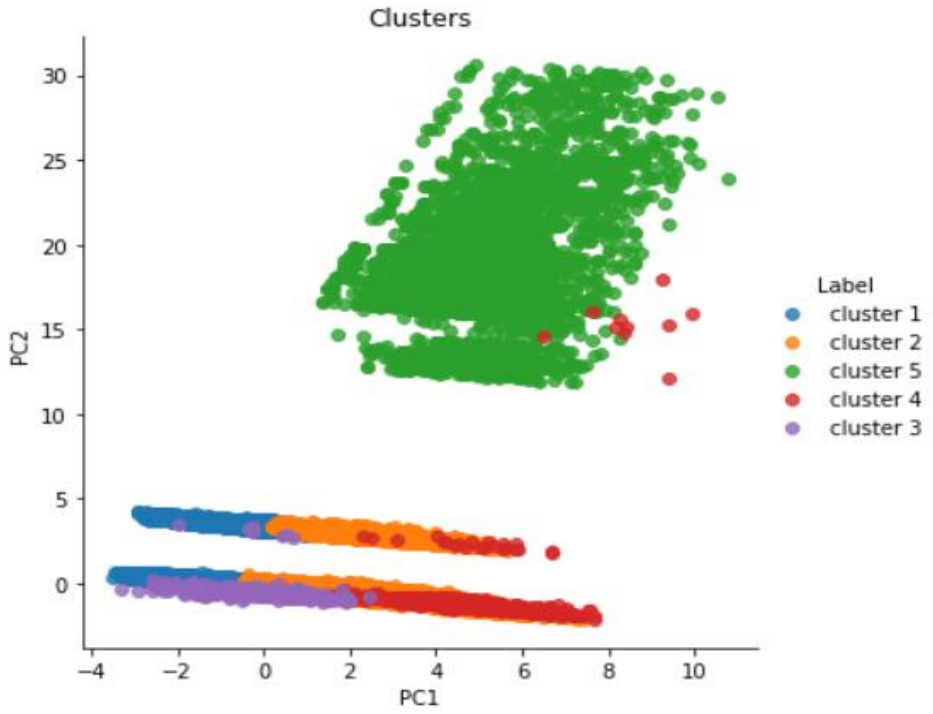


Proposed Model Technique

Data Glacier
Your Deep Learning Partner

KMeans

- KMeans is a partitioning method non-hierarchical clustering algorithm , it is less complex and easy to implement when compared to others clustering algorithm.
- It is an iterative process where you are trying to minimize the distance of the data point to the average data point in the cluster.
- One of the major application of KMeans clustering is segmentation of customers to get a better understanding of them which in turn could be used to roll out the personalized offers from the bank.

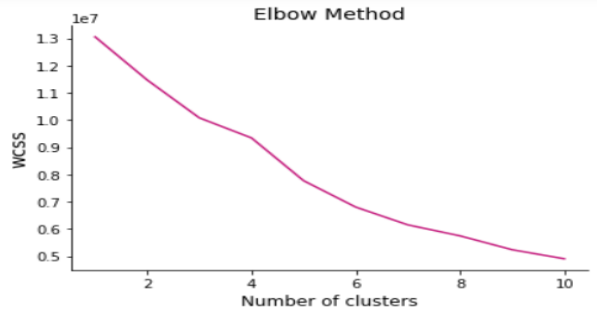


Model Evaluation

Data Glacier
Your Deep Learning Partner

a. Elbow Method

- The elbow method was utilized to evaluate the amount of clusters that fit with our dataset.
- 6 clusters were the optimal number.




```
from kneed import KneeLocator

k1 = KneeLocator(range(1, 11), wcss, curve = "convex", direction = "decreasing")
k1.elbow
```

```
6
```

b. Calinski-Harabasz Index

- Also known as the Variance Ratio Criterion - can be used to evaluate the model.
- A higher Calinski-Harabasz score relates to a model with better defined and well separated clusters.

```
from sklearn import metrics
from sklearn.metrics import pairwise_distances

metrics.calinski_harabasz_score(principal_components, clusters)
```

```
444179.50001355895
```

c. Davies-Bouldin Index

- This index signifies the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.
- This high score implies that the model does not have better separation between the clusters.

```
from sklearn.metrics import davies_bouldin_score

davies_bouldin_score(principal_components, clusters)
```

```
0.926732825121233
```

Recommendations

1. Data: A better description of the data variables would be appreciated for the next projects.

2. Model: The customer limited the groups to 5. That was a limitation as we found that the optimal clusters were 6.

3. Clusters: Clusters 3, 4 and 5 are all customer type former primary and customer relation type is active and inactive. This gives room for reactivations.

Thank You