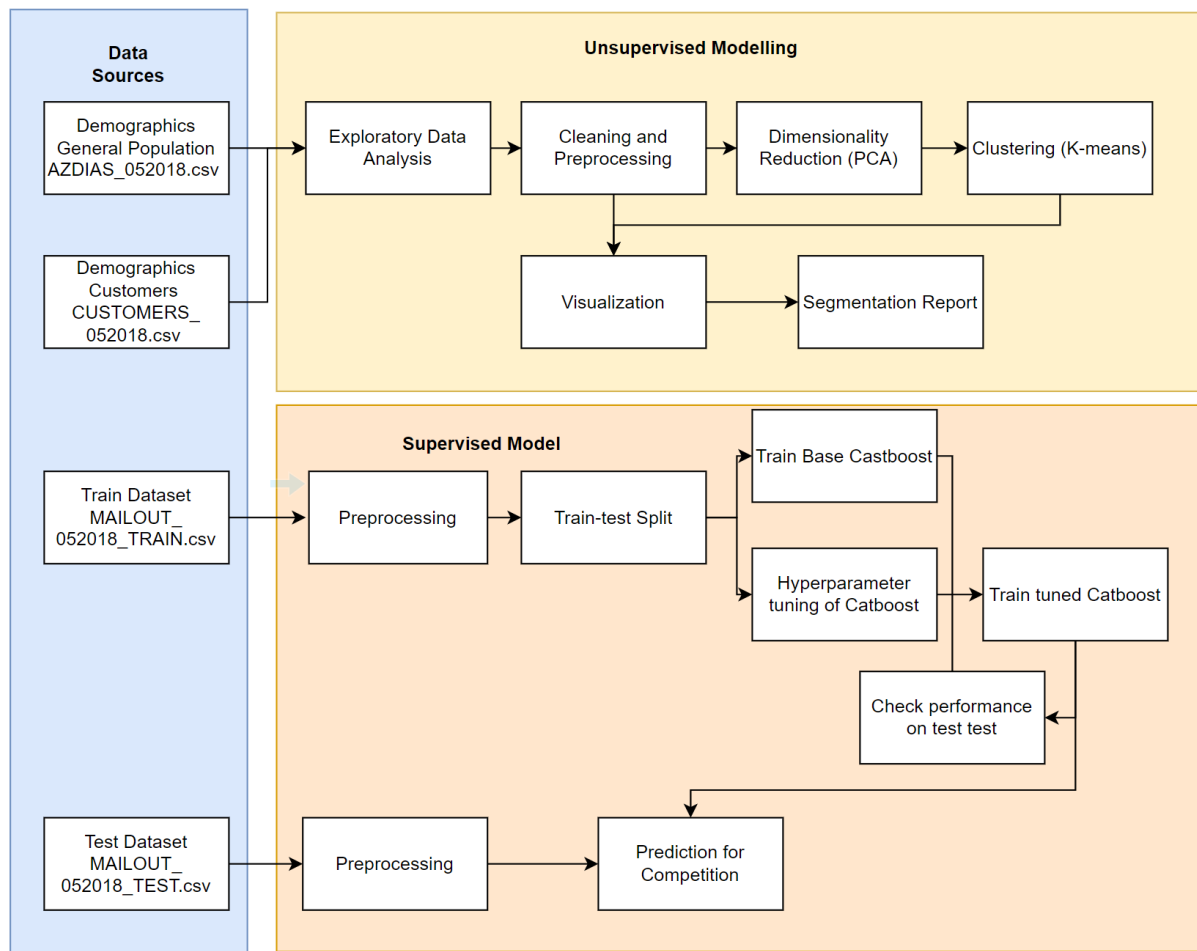


1. Project Overview

This project addresses a customer segmentation problem for a German financial institution. The first part deals with the analysis of the customer base and compares its attributes in relation to those of the general population. The second part deals with the construction of a predictive model to identify clients with a greater propensity to use the institution's services.

The following diagram shows the high-level workflow used to address both challenges.



2. Problem Statement and Metrics

The challenge is to increase the understanding of the customer base and generate a good characterization for commercial campaigns. In order to develop this I used data from the institution and data from a sample of the general population are used.

The second challenge is predicting from a prospect base whether they will respond to a marketing campaign.

For this, a machine learning model is built using the insights from the first part. The AUC, which provides a general measure of the quality of the model, will be used as the metric for the predictive model.

3. Methodology

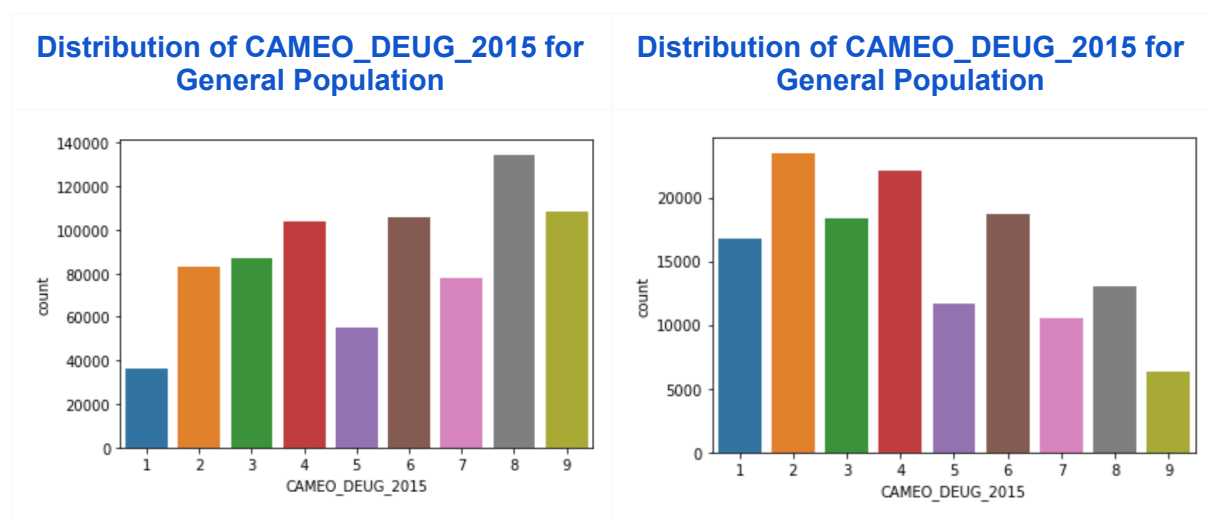
In this section I present the methodology used to analyze, prepare and model the data for the business case.

3.1. Exploration

The first step is to carry out an exploratory analysis of the data to understand what variables are available, their level of data quality and possible preliminary insights.

Specifically, sociodemographic attributes such as age, socioeconomic level, among others, are analyzed.

The socioeconomic dimension is reflected in the CAMEO_DEUG_2015 variable. Here I compare this variable for the general population sample and the customers base. In this variable the lower the number is associated to higher economic classes.



The distribution for the general population is right oriented and show a minor percentage of people belonging to the higher class. This distribution is different from the customer base,

this shows a left oriented distribution. The customers have mainly a higher socio-economic class.

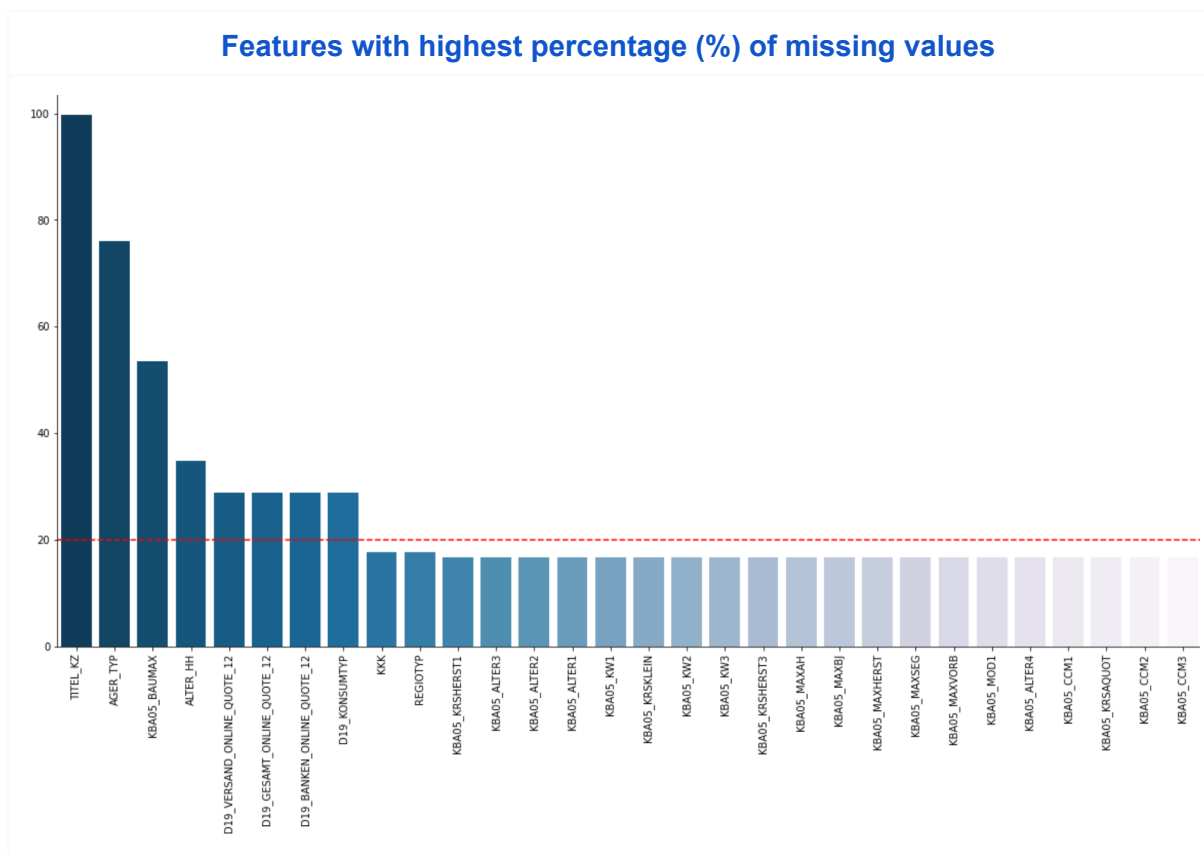
Is these distributions are different this variable could be a good candidate for propensity models.

3.2. Data Cleaning

The second step of the methodology is to clean the data, this process was carried out in several stages:

1. Delete columns that are not described in the data dictionary
2. Replace encoded missing values
3. Identify the presence of other strange or extreme values
4. Remove columns with high percentage of missing values

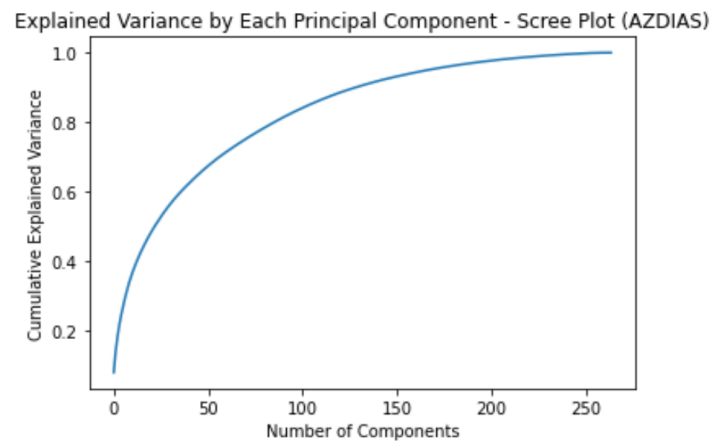
En la figura se presentan las columnas con el mayor porcentaje de valores faltantes



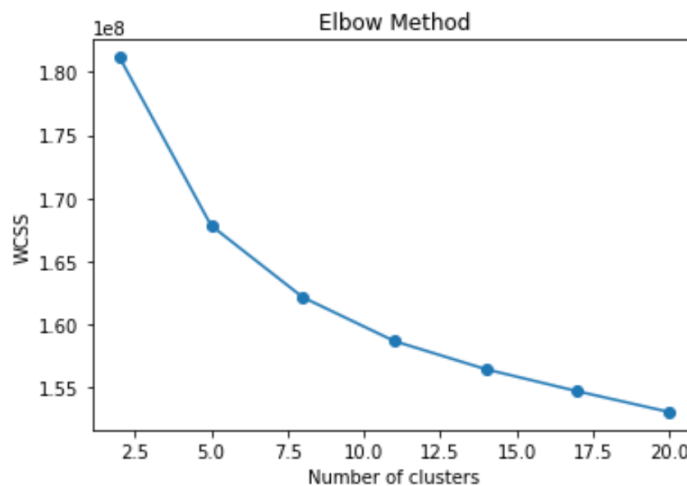
5. Delete rows with high percentage of missing values
6. Imputation of missing values
7. Create cleanup function that consolidates previous steps

3.3. Unsupervised Model for Segmentation

For the unsupervised model, a dimensionality reduction using Principal Component Analysis (PCA) was used. The graph shows the percentage of the variable explained by the components. The first 150 components explain about 90% of the variability in the data set.



For determining the number of clusters I used the Elbow method. Here I used k of 10 as a proper number of customer segments.



I present the clusters in the section 4.

3.4. Supervised Model for Campaign Sending

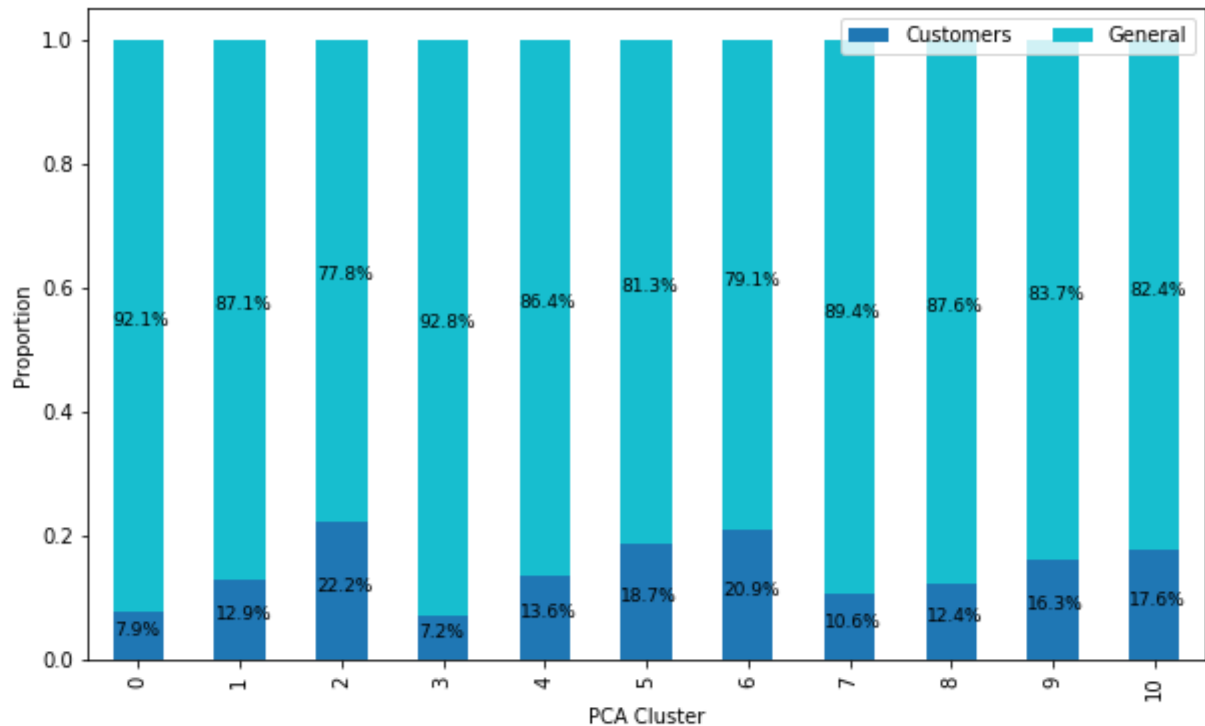
I trained a machine learning model for the classification task. The model used is Catboost due to his flexibility, his capacity of handling categorical features with a high number of different values. It also works very good with the standard scikit-learn functions like GridSearchCV and the scoring functions. The results of this model are presented in section 5.

4. Results: Customer Segmentation

The graph shows the percentage of clients among the different clusters obtained through K-means.

The clusters with the highest rate of clients, and which therefore would be the most attractive, are **2, 5 and 6**. They all have more than 18% penetration.

Therefore, to prioritize commercial campaigns, these clusters should be taken into consideration.



5. Results: Model

For this stage, multiple models were tested, such as Logistic Regression, Gradient Boosting Model and Catboost. The best performance was observed in the latter so the detail is presented around that model.

Base Model:

For the Catboost model, a base model was trained with the default hyper parameters:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(train_1, y, test_size=0.2, random_state=0)

# Base Catboost Model

catboost_base = CatBoostClassifier(random_seed = 278, custom_loss='AUC', silent = True)
catboost_base.fit(X_train,y_train)
```

In the test set the base model scored AUC 0.753 and f1-score of 0.0.

Hyper-parameter Tuning:

To optimize the parameters of the Catboost Model, the GridSearchCV function was used in conjunction with a grid. The detail of the grid is presented below:

```
# Hiperparameter tuning for Catboost

clf = CatBoostClassifier(random_seed = 278, custom_loss='AUC', silent = True, scale_pos_weight=30)


grid = {'learning_rate': [0.03, 0.02],
        'depth': [4,7],
        'l2_leaf_reg': [1, 9, 13]}

Grid_CBC = GridSearchCV(estimator=clf, param_grid = grid, cv = 3, n_jobs=-1, scoring = "roc_auc")
Grid_CBC.fit(X_train, y_train)
```

The optimized model scores AUC of 0.787 and f1-score of 0.08 on the test set. Both metrics improved in relation to the base model.

After evaluating the performance on the test set I retrained the model with all the available data and the best hyper-parameters for predicting the score for the Kaggle set.

I uplodaded the prediction to Kaggle for evaluating the score and I got AUC of 0.722 (image)

262	▲ 2	Encripteduser		0.72209	17
-----	-----	---------------	---	---------	----

6. Conclusion

This project was notoriously complex due to the extensive treatment of the data required, since it was necessary to carefully review the data dictionaries, clean the data and apply advanced modeling techniques.

One of the main skills that I worked on was working meticulously since it was very easy to get lost among hundreds of variables, this is because I had not worked with such large datasets before.

Regarding the models, it was interesting to develop both a supervised and a supervised model. An important aspect of improvement for these models is that variables that were not in the data dictionary could be tested. I eliminated these since I was not clear about their meaning, but they could certainly be an aid in the performance of the models.