Customer Conversion Prediction ML Project - Evangelos Protopapadakis

This project aims to understand the customer characteristics of a mail order company to improve their targeted marketing campaigns. We leveraged datasets supplied by Arvato Analytics, to understand the key characteristics of the customer base and built a supervised model to predict the likelihood of a customer conversion. The final model was submitted in a Kaggle competition and evaluated based on the Receiver Operating Characteristics Area Under Curve score. The model has a score of 0.8878 ranking 1st among 438 participants.

Project Overview

This project was one of the proposed capstone projects of the Udacity Data Scientist Nanodegree. This project was of personal interest to me due to the widespread application of customer segmentation and the fact that it combined both unsupervised and supervised machine learning methods.

The original project brief is below:

"In this project, we will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. You'll use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, you'll apply what you've learned on a third dataset with demographics information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company. The data that you will use has been provided by our partners at Bertelsmann Arvato Analytics, and represents a real-life data science task."

Data

4 datasets were provided by Bertelsmann Arvato Analytics through Udacity for this project.

1. Demographics data for the general population of Germany: 891 211 persons (rows) x 366 features (columns)
2. Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns)
3. Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns)
4. Demographics data for individuals who were targets of a marketing campaign; 42 833 persons x 366 (columns)

In addition, two supporting documents were provided that described the encoding and meaning of features across datasets:

1. Dias Information Levels Attributes - Contains descriptive information for each feature
2. DIAS Attribute Values - Contains detailed information about the meaning of the different value levels for each feature in the 4 datasets

Problem Statement

The problem statement is composed of two parts:

1. Can we identify a demographic in our current customers with respect to the general German to inform better marketing products?
2. Can we reduce the number of people targeted and still retain a high proportion of our target demographic?

Metric

The latter part of the problem statement will be evaluated in a Kaggle competition using a dataset with witheld labels. Area Under Curve (AUC) will be the performance metric used to evaluate the model.

The AUC performance metric is appropriate for a highly imbalanced labeled dataset where the number of positive responses is significantly smaller than the number of negative responses. This metric is able to evaluate the quality of our model when, despite being a good predictor, we would still expect a fairly small conversion rate from our target demographic.

Project Outline

The project is broken into three major sections:

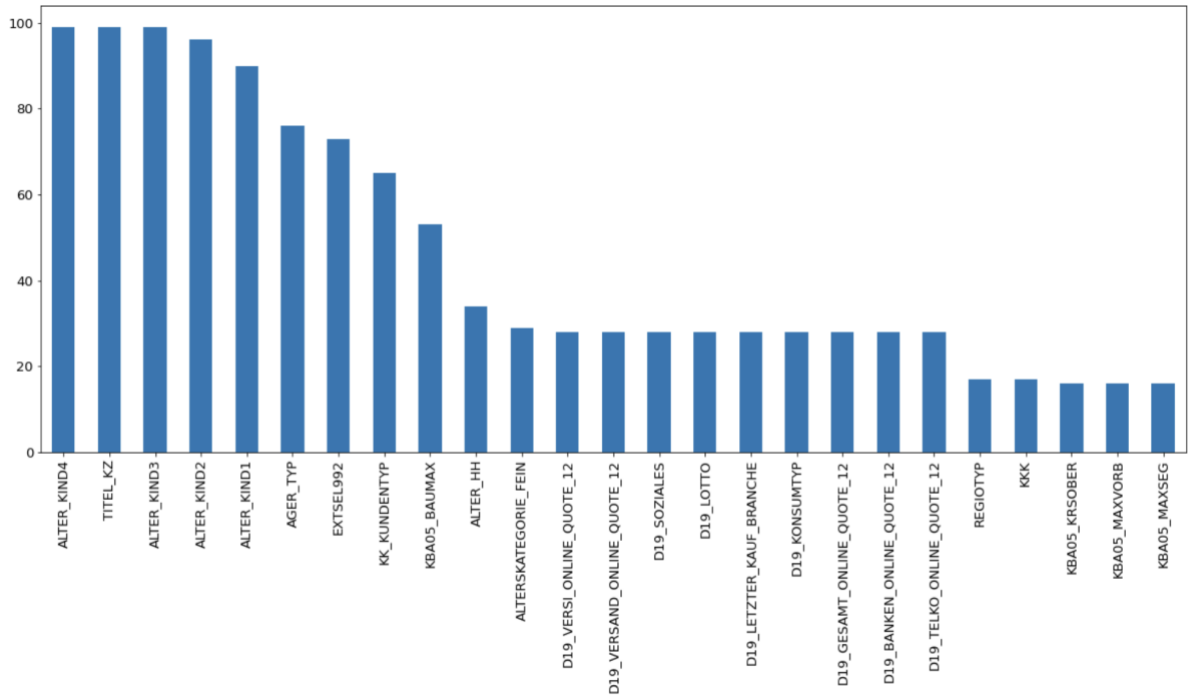
1. Data Understanding and Data Cleaning
2. Unsupervised model to segment customers and identify key characteristics of customers
3. Supervised model that can predict the likelihood of an individual converting to a customer

Data Understanding & Data Cleaning

At first we need to understand the features. The main questions we need to answer are:

- Do we have any sparse features?
- Are numerical features ranked?
- How to treat categorical features? Are categorical features ordinal?
- How to treat missing values?

Many of the numerical features encoded an ‘unknown’ as -1 or 0. So the first step was to convert everything to NaN. There were 8 features that had more than 50% missing values that we thought are sparse enough so we decided to drop them.



% Missing Values

There are various categorical data in the dataset that we need to transform before proceeding with further analysis.

- **CAMEO_DEU_2015** : The 'CAMEO_DEU_2015' column which specifies the family classification of the person has XX and X values. Because there is no specification what those columns mean, we would replace them with NaN values. Then we will one-hot encode the column with binary variables given that the order of the current values is not meaningful.

- **CAMEO_DEUG_2015**: This column had numerical values but in string format. We decided to convert them to integers. There is no description provided by Arvato but we decided to keep it as we don't want to have any information loss.

- **CAMEO_INTL_2015:** The 'CAMEO_INTL_2015' feature is a ranked numerical feature that combines two categories Wealth (Wealthy, Prosperous, etc) and Life Stage (Pre-Family, Young Family, etc). We decided to split this feature into two features. To do so, we observed that the first value specifies the Wealth and the second specifies the Life Stage.

- **D19_LETZTER_KAUF_BRANCHE:** We one-hot encoded this categorical variable as well.

- **OST_WEST_KZ:** Defines East/West so we converted it to binary.

For the rest of the missing data, we decided to impute them using the mode of the population data. So, we calculated the most frequent value by column and we filled any missing values for all datasets provided.

Unsupervised Model for Customer Segmentation

We want to understand how customer demographic is distinct from a general population using an unsupervised model. We will leverage K-means clustering to defined distinct clusters that represent the customer population.

In order to build this unsupervised model, we need to reduce the dimensionality of the original dataset. The higher the number of dimensions, the less able the KMeans model is able to cluster features. Known as the 'curse of dimensionality'. In low dimensional datasets, the closest points tend to be much closer than average, but two points are only close if they're close in every dimension. So when there are lots of dimensions, it's more likely that the closest points aren't on average much closer than average.

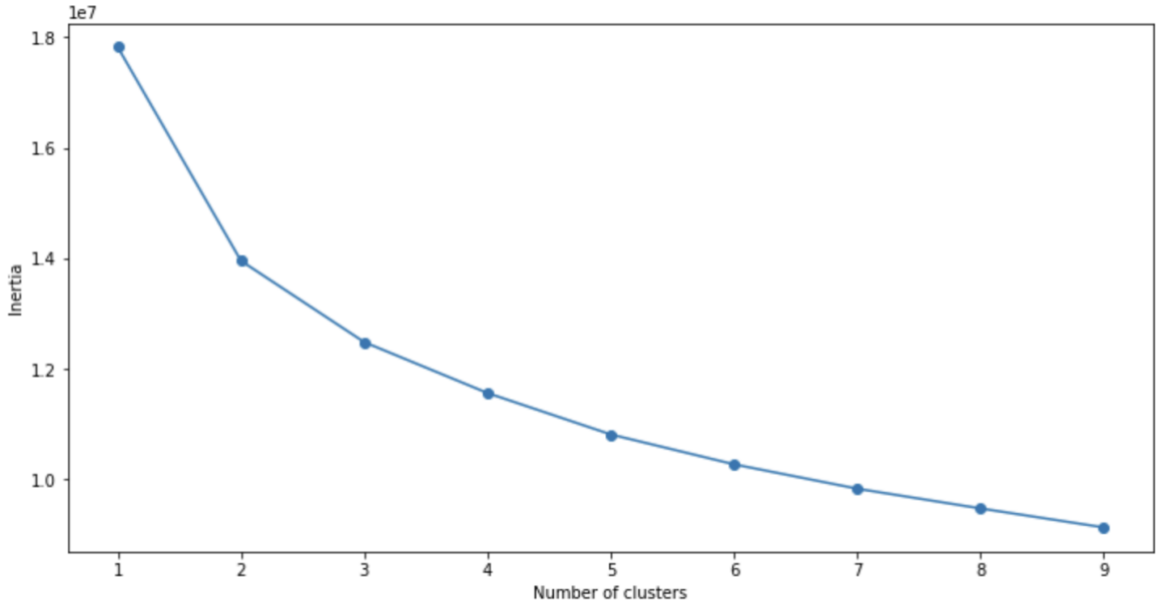
Dimensionality Reduction

To reduce the dimensions of the dataset, we concatenated the population data with the customer data and we labeled the the general population having a RESPONSE==0 and the customers having a RESPONSE==1. In other words, if this were a mail order campaign, all current customers would have a conversion of 1 and all non- customers would have a conversion of 0. Then this dataset was fed to an xgbooster classifier and a random forest models. The xgbooster classifier achieved a higher AUC score in the test data so we decided to move on with this model. In any case, both models had a significant overlap in terms of deciding which features are important.

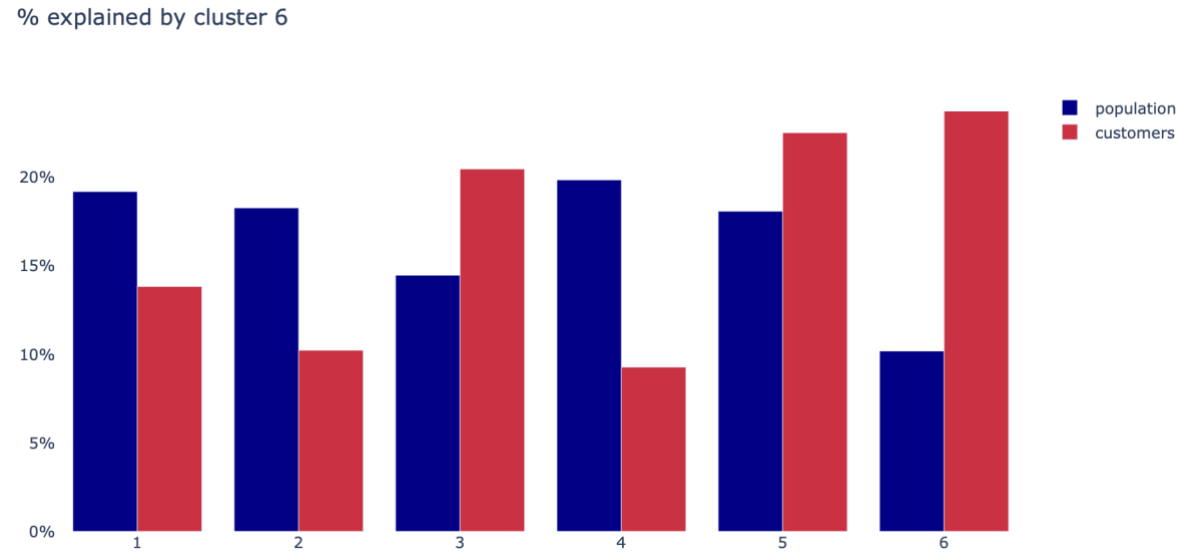
Unsupervised Learning with KMeans

After we reduced the dimensionality of the dataset we fitted a KMeans model to segregate the population and customer characteristics.

To find an optimal number of clusters I used the inertia criterion and Elbow method. The inertia decreases as a function of the number of clusters and plotting the inertia against number of clusters shows the rate of change of inertia. The optimal cluster can be found at an inflexion point where the rate of change sharply decreases.



The Elbow method result can be seen above. There wasn't a clear inflexion point but the rate of change seemed to decrease most between the value of 4 and 6. We decided to move on with cluster 6 as it provided the most differentiation among population and customer data:



<u>Customer Characteristics:</u>

Most of the variation of the customers dataset was explained by clusters 3, 5 and 6. In particular:

Cluster 3:
Customers have a higher net income compared to the general population (HH_EINKOMMEN_SCORE) . The average person is defined as dreamily (SEMIO_VERT). In terms of financial situation, it is defined as "be prepared" but less of a money saver compared to the general population (FINANZ_VORSORGER, FINANZ_SPARER). Those

customers are also defined as family type (LP_FAMILIE_FEIN) with more than one person in the household familiar with the products of the company (ANZ_PERSONEN).

Cluster 5:
In this cluster, there are more female than male customers compared to the general population (ANREDE_KZ) and the average person is defined as traditionally minded (SEMIO_TRADV). This cluster has also higher income and are less of money savers compared compared to the average population. In terms of buying habits they are classified as CJT Type 6 (Customer-Journey-Typology relating to the preferred information and buying channels of consumers).

Cluster 6:
In this cluster, the customers are defined as money savers (FINANZ_SPARER) with low estimated net income (HH_EINKOMMEN_SCORE) and younger in age (ALTERSKATEGORIE_GROB). In terms of buying habits they are classified as CJT Type 1 (Customer-Journey-Typology relating to the preferred information and buying channels of consumers) and have high exposure to KOMBIALTER (classification based on Arvato analytics).

Supervised Model

We want to build a supervised model based on the mailout datasets that can predict the likelihood of an individual converting to a customer. Then we will test our prediction in a Kaggle competition. We tested two different classifiers Random Forrest and XGBoost. After submitted both results to Kaggle the Random Forest classifier gave superior predicton currently ranking as #1 out of 438 submissions.