Identifying customer segment using unsupervised learning

CAMILLE PAPILLON-HOGUE

Introduction

Goal : Apply unsupervised Learning techniques to identify segments of the population that form the core customer base for a mail-order sales company in Germany.

Why ? These segments can then be used to direct marketing campaigns towards audiences that will have the **highest** expected rate of return

This project was delivered during the Udacity Nanodegree class: Introduction to Machine Learning with TensorFlow.

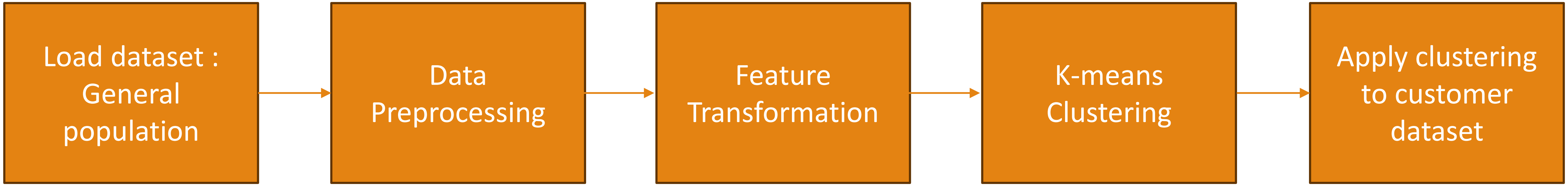
Project Link : https://github.com/camillepapillon/Unsupervised_ML_Identifying_Customer_Segment

Context

3 main datasets:

1. General population dataset = Demographics data for the general population of Germany; 891211 persons (rows) x 85 features (columns)
2. Customers population dataset = Demographics data for customers of a mail-order company; 191652 persons (rows) x 85 features (columns)
3. Data Dictionary : Detailed Information file about the features in the provided datasets

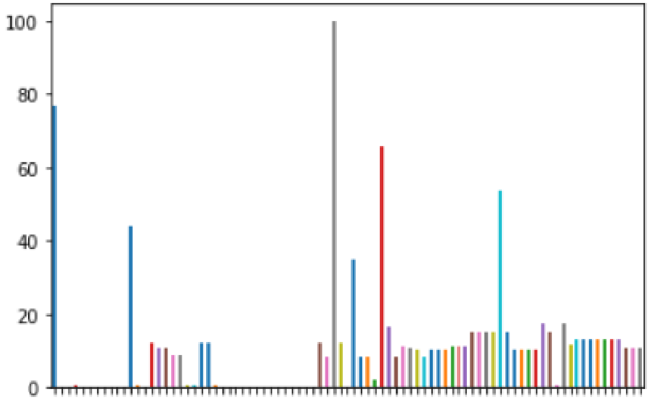
Methodology


Load dataset : General population
Data Preprocessing
Feature Transformation
K-means Clustering
Apply clustering to customer dataset

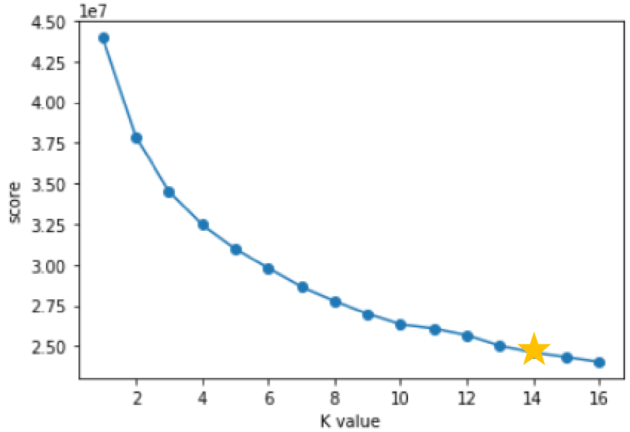
Key Findings & Insights

- Dropped 6 columns with more than 20% missing data
- Re-encoded binary features and dropped the 14 multilevel features
- Transformed missing values in rows with Imputer function => average missing values on all rows = 1,37 %
- Applied feature scaling with StandardScaler() on general population dataset
- Performed dimensionality reduction (PCA) and kept **30** principal components
- Applied Kmeans clustering to general population dataset and kept **14 clusters**

Percentage of missing data per column in general population dataset



Model score per count of clusters (K value) in Kmeans clustering



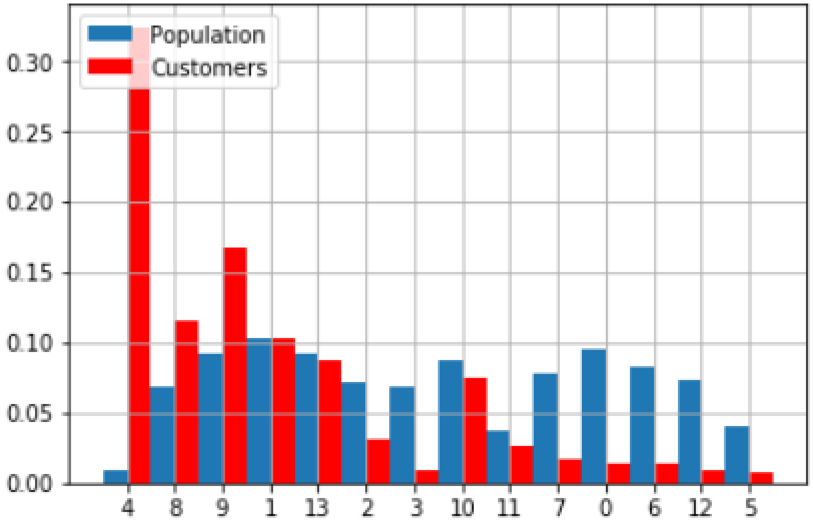
Key Findings & Insights

- Customer population is **overrepresented** by cluster 4 and **underrepresented** by cluster 0 compared to the general population

Principal Component analysis of cluster 4 and 0 :

- **Cluster 4** = male between 36-60 years old, lives alone or with someone, from prosperous household, high income, more likely to be homeowner.
- **Cluster 0** = women less than 30 years old, average income, financially minimalist, less likely to live in 6-10 family home.

Percentage of the population in each of the 14 clusters


Population
Customers