Udacity - Machine Learning Engineer Nanodegree Program

Capstone Proposal
Kadichari Victória Coelho Farias Almeida
April 17, 2020

Create a Customer Segmentation Report for Arvato Financial Solutions

Domain background

This project aims at a real-life data science task that was provided by partners like Bertelsmann Arvato Analytics.
Analyzes of the demographic data of customers of a sales company, where they are located in Germany and comparing with demographic information of the general population, will be carried out.
Unsupervised learning techniques will be used to segment customers, identifying groups of the population that best describe the company's main customer base. In a third set of data, supervised learning techniques will be applied to predict which people are most likely to order by mail who may become your customers.

Problem statement

This project aims to predict which people are most likely to become a customer of a mail order company in Germany.
For that it is necessary:
1. Pre-processing of data: Analyzing the dataset, it is necessary to convert fields that have no value to NaN. Remove empty rows and columns.

2. Customer Segmentation Report: use unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments.

3. Supervised Learning Model: You will have access to a third set of data with attributes of the destinations of a direct mail campaign. Where you will use the previous analysis to create a machine learning model that predicts whether each individual will respond to the campaign or not.

4. Kaggle competition: After choosing a model, it will be used to make predictions in the campaign data as part of a Kaggle competition. Sorting individuals by the likelihood of becoming customers and you will see how their modeling skills compare to others.

Datasets and Inputs

There are four data files associated with this project that were made available on the Udacity platform:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Each line in the demographic files represents an individual, but includes external information, such as your family, for example.
Using information from the AZDIAS and CUSTOMERS datasets, it is possible to verify the similarities and differences between customers and the general population. And then use the analysis to predict which recipients are most likely to become a customer for the mail order company, using the two MAILOUT files. These will be used for supervised machine learning.

Solution statement

To solve this project, it is necessary to divide by tasks:

1. Pre-processing and data analysis:
   - Clear columns and rows without values
   - Fields that are empty, convert to a numeric value
   - Analyze NaN values
   - Replace NaNs with -1
   - Normalize the data

2. Dimensionality reduction with Principal Component Analysis (PCA):
   - It is necessary to reduce the dimensionality of the data.

3. Client segmentation with K-means cluster:
   - Use K-means as this is a very efficient tool for dividing dimension points into clusters.

4. Binary classification:
   There are several algorithms that are quite efficient for binary classification, however it is necessary to evaluate the data set to determine which model to use, for this case I have the suggestion LogisticRegression, RandomForestRegressor and KNN.
Finally, Python Flask and Docker will also be used to create an endpoint to access the model and make predictions.

Benchmark model

The reference model will be given by logistic regression and then its performance will be compared with RandomForestRegressor and KNN.

Evaluation metrics

To evaluate the model, the accuracy value will be observed and we will also use the confusion matrix.

Project design

1. Pre-processing and data analysis:

- Clear columns and rows without values
- Fields that are empty, convert to a numeric value
- Analyze NaN values
- Replace NaNs with -1
- Normalize the data

2. Dimensionality reduction with Principal Component Analysis (PCA):
   - It is necessary to reduce the dimensionality of the data.

3. Client segmentation with K-means cluster:
   - Use K-means as this is a very efficient tool for dividing dimension points into clusters.

4. Model selection:
   - After training different types of models with standard hyperparameters, choose the most efficient.

5. Create a REST API to make predictions:
   - Python Flask and Docker were used to create an endpoint to access the model and make predictions.