Capstone Project Proposal

Description

The idea behind this project is to perform customer segmentation based on historical data provided by Arvato Financial Solutions.

We need to analyze a “CUSTOMERS” dataset and figure out how customers are similar to or differ from the general population at large (“AZDIAS” dataset). Then, using information from that analysis, we need to make predictions on users who were the target of a marketing campaign (“MAILOUT” dataset).

Key points of the proposal

1. The project’s domain background — the field of research where the project is derived

This project is based on analysing customers behaviour. We are trying to spot similarities and differences between customers population and the general population. The idea is to spot socio-economical traits that could give us some insights about what characteristics our customers have. This way, we can mine data from a bigger population and know in advance which persons could be more receptive to our products and services, hence targeting them with our marketing campaigns.

2. A problem statement — a problem being investigated for which a solution will be defined

With this project we want to answer: Who will buy our products?

We want to do an efficient use of resources to target people with bigger probabilities to become customers.

3. The datasets and inputs — data or inputs being used for the problem

The datasets to be used are those provided by Arvato Financial Solutions.

- **Udacity_AZDIAS_052018.csv** -> This dataset contains information about general population in Germany
- **Udacity_CUSTOMERS_052018.csv** -> This dataset could be considered as a subset of the previous

dataset, but here it only includes information about those german people that are also customers of this company.
- **Udacity_MAILOUT_052018_TEST.csv** -> This dataset contains users that were targeted by a marketing campaign. It includes socio-demographic variables and different affinities. It also contains whether or not the campaign worked, e.g. the target value to predict. This is the training partition.
- **Udacity_MAILOUT_052018_TRAIN.csv** -> This dataset contains users that were targeted by a marketing campaign. It includes socio-demographic variables and different affinities. It also contains whether or not the campaign worked, e.g. the target value to predict. This is the test partition.

4. A solution statement — the solution proposed for the problem given

The solution comprises two steps:

- **Step 1: Exploratory Data Analysis with Unsupervised Learning**
  In this step we will manually explore all datasets to spot interesting patterns. We will rely on a heavy use of data visualization, especially for comparing different populations. Also, some unsupervised techniques will be used to group people into clusters.

- **Step 2: Modelling and Prediction**
  For this step different Machine Learning models are going to be used: from most simple ones to more complex. Also, we will use model interpretability techniques to extract business insights by revealing which are the features more useful to distinguish potential customers from those that are not.

- **Step 3: Kaggle Competition**

- In this step we will generate a submission file for the related Kaggle Competition. We have been giving a very important tip:
  "The exact values of the "RESPONSE" column do not matter as much: only that the higher values try to capture as many of the actual customers as possible, early in the ROC curve sweep"
  hence, we should focus on having a high recall. We may tune the cost function to give more weight to customer class versus non-customer class when doing the classification.

5. A benchmark model — some simple or historical model or result to compare the defined solution to

We will start with a simple clustering using K-means with the numerical features only. As this is an unsupervised algorithm it is very difficult to numerically compare with a more powerful approach, hence I will compare the result of this baseline and the following clusterings qualitatively using 2D projections. The next model will use both categorical and numerical values. Categorical values will be encoded using embeddings obtained from a neural network executing a classification task.

For both cases, a dimensionality reduction technique is likely to be used, whether PCA or t-SNE.

For the second part of the project, a logistic regression will be used with only numerical features. This will be our baseline. Then we will use a Random Forest using all relevant features, encoding categorical values using numerical encoding. The last model will be a Neural Network with embeddings for encoding categorical values. This part could be easily measured using standard metrics that will be detailed in Section 6.

6. A set of evaluation metrics — functional representations for how the solution can be measured

As this is potentially an unbalanced problem (in general it's expected to have much more non-customers than customers in a big general population) we will rely on confusion matrixes and F1 score. However, as it is required by the Kaggle Competition requirement of the project AUC metric will be provided too.

7. An outline of the project design — how the solution will be developed and results obtained

The first step will be to use *pandas_profiling* to do a quick exploratory analysis. It will show us where to look deeper. Then, after a manual study of the data, we will assess if the previous hypothesis were correct or if we should reformulate the approach to tackle this project.

We will proceed following agile principles: iterative and incremental. Thus the use of baselines and incremental increasing the difficulty of the technics used.