proposal.md

Capstone Proposal

Machine Learning Engineer Nanodegree

Ricard Trinchet Arnejo

2020/05/24

1. Domain Background

Arvato Bertelsmann is a german services company that operates worldwide. In this project, we analize data from a group of the company's customers, located in Germany. The main objective is to find some characteristics of the population of customers and then compare those characteristics with those of the general german population.

Further, we will analize labeled data from a targeted email campaign, which tells us which customers responded to that campaign. We can use that data to train a Supervised Machine Learning model that predicts if a given individual will respond to a new campaign or not.

The problem is really interesting, as its solution will certainly give added value to the company. Also, it is a problem that many companies would face, so it is likely that one would have to face a similar problem in a Data Scientist job.

2. Problem Statement

As mentioned in the last section, the problem consist of 2 main parts.

First, we have data of a group of german customers of the company, as well as data of a group of german population. The aim is to use Unsupervised Learning techniques to extract information from both groups and be able to compare them, as well as potentially identify clusters of general population that are similar to the company's customers and thus would be potential targets.

Second, there is also available data related to a previous email marketing campaign. The aim is to use a Supervised Learning technique to be able to understand which individuals responded to the campaign and also be able to predict which new individuals will eventually respond to a new campaign.

3. Datasets and Inputs

Arvato gives us four different datasets which we can use.

- Azdias data: dataset containing demographic information of general german population. It has over 800K rows and 366 features.
- Customers data: dataset with company customers information. It has about 190K rows and 369 features (3 extra variables for customers).
- Train data: Data with information about the 42K targets of the email campaign and the response of each individual.
- Test data: Data with individuals that can be potential targets of a new email campaign. It has about 42K rows.

We will use the first two datasets for the first problem and the two last for the second problem.

4. Solution Statement

We will work on one solution for each problem:

- For the first problem, we will use an Unsupervised Learning method (likely, a clustering method) to detect groups in the population. Then, we will describe each of the groups and compare them with the general population data.

- For the second problem, we will train a Supervised Learning classifier that allows us to classify the indivuduals from the test data in two groups: those who accept the offer and those who don't.

5. Benchmark Model

For the first problem, a benchmark model would be the k-means algorithm, which is a widely used clustering technique.

For the second, we could take a look at the associated Kaggle competition's Leaderboard. There, we see that most of the participants are scoring AUC (Area Under Curve) better than 0.7. That would be a good score to have at minimum and then try to improve it as much as possible.

6. Evaluation Metrics

For the first problem, it is harder to define a evaluation metric, as the success depends a lot on the data. Nevertheless, a metric like the Silhouette could be helpful.

For the second, we will use the AUC, as it is the one used by the Kaggle competition associated.

7. Project Design

The detailed steps for the workflow related to the project are as follow:

- Load the data and carefully look at some of its characteristics. As a result of this, we would obtain a function that would perform data cleaning on the data.

Some things to consider for this step are: % of missings, factors that need replacement, variables that can be dropped, etc. Once the cleaning is implemented, we can proceed with the following:

- Preprocessing. That is, prepare the data to be accepted by a model. This would include scaling the numerical columns, impute missing values and encoding the categorical columns. We can also try a dimensionality reduction technique like PCA, to get a reduced number of variables. Alternatively, we will train a RandomForest and see which features are more important for that model. Those are good candidates to be a subset of features.
- Modeling. For the first problem, a clustering technique like k-means will be used. If the results are not satisfactory, we can explore other clustering methods like DBSCAN. For the second problem, we will try a classifier like RandomForest. If the results could be improved, we will use other technique like GradientBoosting methods (lightgbm).

Also, we will need to keep an eye on some of the following questions.

- Is the data imbalanced? If so, we will try with Oversampling techniques.
- Could we resample the Azdias and Customers data and still obtain good results?
- Could we train our models for the second problem using Azdias and customers datasets?