Bertelsmann/Arvato Capstone Project Proposal

Yusong Zhou

September 29, 2020

1 Domain Background

Customer segmentation involves identifying certain groups of people with a usually huge demographic dataset. The size of the dataset, especially the number of features contained, makes machine learning models suitable for this job, as other simple statistical models may not do well with such a high dimension of features.

Bertelsmann is a media, services and education company in Germany. It's one of the largest media conglomerates in the world and have divisions in music, printing and investments as well. Arvato, the service provider division of Bertelsmann, is an internationally active services company that develops and implements innovative solutions for business customers from around the world. The datasets and objective of this project are provided by them.

This project will use both unsupervised and supervised machine learning models to first classify a population sample into several groups and then try to make predictions of whether a certain customer will respond to the company, based on another sample with the same set of features.

2 Problem Statement

Bertelsmann, a mail-order sales company in Germany, wants to identify potential customers from a general population. It provides four main datasets for analyzing the relation of the potentiality of becoming its customer to all features.

The unsupervised learning problem involves using two of the datasets, one for the general population and the other for some former customers of the Germany company. A clustering model will be applied to find certain groups in the general population that's most similar to the most of the former customers.

The supervised learning problem involves using the third dataset with a response column indicating whether this person replied the mail to predict the reaction of people from the last dataset, where the response column is empty. A regression model will be applied to generate a possibility for the empty response column.

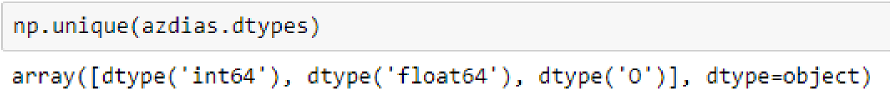
3 Datasets and Inputs

There are six datasets in total, four of them have been mentioned in the last section:

- *Udacity_AZDIAS_052018.csv*: contains 891211 people × 366 features, denoting general population. Most of the features are numerical types and a few are string;
- *Udacity_CUSTOMERS_052018.csv*: contains 191652 people × 369 features, denoting customers of the company. It has three additional features compared to azdias dataset;
- *Udacity_MAILOUT_052018_TRAIN.csv*: contains 42982 people × 367 features, denoting target people of marketing. It has one additional feature compared to azdias dataset, which is individual's response;

- *Udacity_ MAILOUT_ 052018_ TEST.csv*: contains 42833 people × 366 features, denoting test set of marketing without the response record of the last file.

Figure 1 shows the data types of azdias dataset, including *int*, *float* and *object*. The last one represents a column with string values, in Pandas DataFrame.

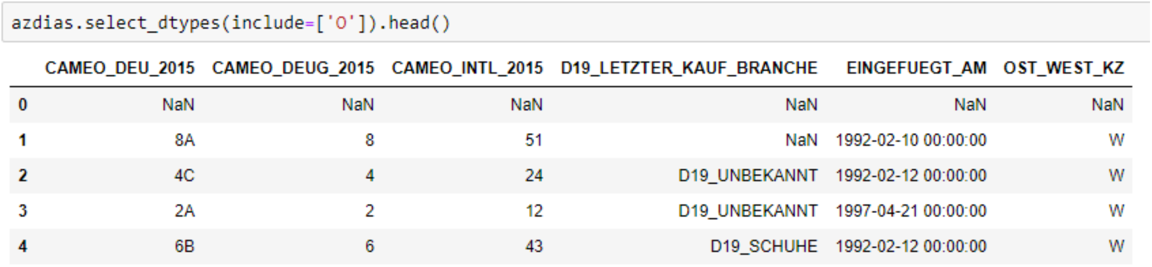


```
np.unique(azdias.dtypes)
```

```
array([dtype('int64'), dtype('float64'), dtype('O')], dtype=object)
```

Figure 1: Data Types of Dataset Azdias

Figure 2 gives an overview of these string columns. Some of them, like the second and third column, are actually numerical types that shall be converted. Others are mostly categorical, which can also be converted to numerical types. The fifth column contains time.



```
azdias.select_dtypes(include=['O']).head()
```

	CAMEO_DEU_2015	CAMEO_DEUG_2015	CAMEO_INTL_2015	D19_LETZTER_KAUF_BRANCHE	EINGEFUEGT_AM	OST_WEST_KZ
0	NaN	NaN	NaN	NaN	NaN	NaN
1	8A	8	51	NaN	1992-02-10 00:00:00	W
2	4C	4	24	D19_UNBEKANNT	1992-02-12 00:00:00	W
3	2A	2	12	D19_UNBEKANNT	1997-04-21 00:00:00	W
4	6B	6	43	D19_SCHUHE	1992-02-12 00:00:00	W

Figure 2: Columns of Type "Object"

Another two files *DIAS Information Levels - Attributes 2017.xlsx* and *DIAS Attributes - Values 2017.xlsx* contain information about the features, including description, meaning of data values and feature type (categorical or not).

4 Solution Statement

For unsupervised learning part, Principle Component Analysis and K-Means Clustering will be applied to find the group of general population that's most similar to the customers sample. The number of components will be determined by an explained variance of at least 95% and the number of clusters will be determined by the curve of the distances of clusters relative to the number of clusters.

For supervise learning part, several frequently used models will be applied to compare preliminary result, including Logistic Regression, Random Forest, XGBoost and other possible choices as well, to determine the model to use. Then hyperparameter tuning will be applied to achieve a high score at Kaggle's assessment of result.

5 Benchmark Model

Logistic Regression will be the benchmark model here as it's basically the simplest probability regression model and the final model should have a better performance than the simple one.

As Kaggle provides a rank of results of others' submitted models, the result obtained here will also be compared to those scores to evaluate the relative performance of the final model. Due to the existence of Bayes Error Rate [1], there might be an upper limit of model performance, and it's useful to make a comparison with others' best performance.

6 Evaluation Metrics

The Kaggle competition uses Area Under Curve as the score of a model. AUC will be applied here to be compared to Kaggle's ranked scores. Another reason to use AUC is that there's class imbalance in the dataset, as records with responses are a lot less than records without responses. Using classification accuracy would not be a good idea.

7 Project Design

- **Step 1: Data Cleaning**

First, raw data files have several issues that need to be addressed, including mixed type, missing values represented as numerical values and records with too many missing values. As all four datasets have 366 common features, a unified data-cleaning procedure need to be established and applied to all four datasets for further analysis.

- **Step 2: Unsupervised Learning**

PCA will be applied to the first two datasets to find the minimum subsets of components that obtain an explained variance above given threshold. With the number of components determined, two datasets will be transformed with the specific PCA model and perform K-Means Clustering, the number of clusters will be determined by certain pattern of the SSE curve.

- **Step 3: Supervised Learning**

A set of models will be trained and compared with the training set split into a smaller training set and a cross-validation set. The final model is determined using AUC on the training set. The preliminary score will be recorded as optimizing goal for hyperparameter tuning. The highest score will be compared to Kaggle's ranked scores and final conclusion will be made.

Reference

[1] Bertelsmann SE & Co.KGaA. Retrieved Sep 30, 2020, from https://www.bertelsmann.com/#st-1
[2] Bayes error rate. Retrieved Sep 30, 2020, from https://en.wikipedia.org/wiki/Bayes_error_rate
[3] Classification: ROC Curve and AUC. Retrieved Sep 30, 2020, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[1] The lowest possible error rate for any classifier of a random outcome