Proposal for Udacity Machine Learning Nanodegree

Customer segmentation and Mailout's response classification

Korntewin Boonchuay
30 Mar 2021

Domain Background

Arvato is a data-driven organization and an internationally services company who provide various solutions to the business customers, including Supply Chain Management, financial services and IT services which continuously developed using the power of data and analytics [1].

Interestingly, there are around 200,000 existing Arvato's customers and 360 demographic features for each individual. With this much data, it is almost impossible to manually explore and identify customer's insight. Thus, data mining and machine learning techniques will be an integral part in finding customer's insight. Additionally, with the mindset of data-driven organization, the results from data analytics will lead to business action and making impact to Arvato business.

Market segmentation technique is suitable for the current Arvato business both in terms of increasing existing customers satisfaction [2] and effectively approaching potential customers. Thus, in this project, various machine learning techniques will be implemented to cluster and identify the potential customers. The underlying latent and insight from market segmentation will also be used to predict mailout responses for each individual and improve the mailout strategies for the Arvato marketing team.

Problem Statement

To cluster the population and identify the potential customers, the data of the general population is needed. In this project, AZ DIAS information database [3] is provided as a general population data with the strict terms of use.

The project will be divided into 2 main processes. First, the unsupervised machine learning technique will be implemented to cluster the population using general population data/Arvato's existing customers data and identify the demographic pattern of the potential customers. Finally, the underlying latent and information from the previous process will be used to effectively classify the population who are likely to respond to the Arvato's mailout using supervised machine learning.

Datasets and Inputs

There are 4 given files containing demographic data for Arvato's existing customer and general population in Germany provided by Arvato Bertelmann as follow:

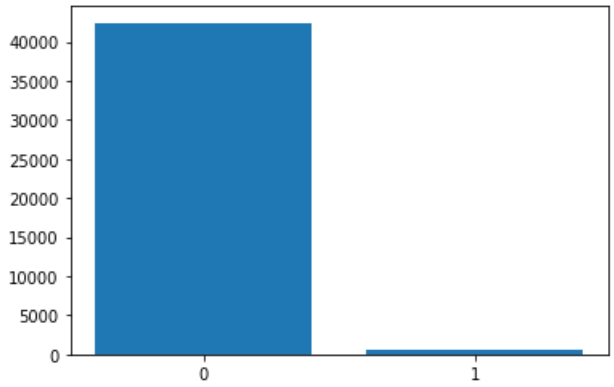
1. Udacity_CUSTOMERS_052018.csv
    a. This file contains demographic data for Arvato's existing customers of 191,652 individuals with 369 features.
2. Udacity_AZDIAS_052018.csv
    a. This file contains demographic data for the general population in Germany of 891,221 individuals with 366 features.
3. Udacity_MAILOUT_052018_TEST.csv
    a. This file contains demographic data for the population who were targeted by a marketing campaign of 42,833 individuals with 366 features.
4. Udacity_MAILOUT_052018_TRAIN.csv
    a. This file contains demographic data for the population who were targeted by a marketing campaign of 42,982 individuals with 367 features.

Furthermore, there are 2 data dictionary or metadata files describing the information, all possible values and encoding null value of each features as follow:

1. DIAS Attributes - Values 2017.xlsx
    a. This file contains all possible values and how null value is encoded for each feature.
2. DIAS Information Levels - Attributes 2017.xlsx
    a. This file contains the description of each feature.

These individuals data will be explored, feature engineer and used as an input for customer segmentation and mailout response prediction.

But first, the balance of the label in mailout datasets should be observed. Since it is the most important thing in choosing metrics to evaluate the classification model. The number of label for "0" (no response) and "1" (does response) can be illustrated as below:



There is only 1.2% of the label “1” which indicates the heavy imbalance for classification model. Thus, the metrics for evaluating should be precision, recall, F1 score or Area Under Curve of ROC.

Solution Statement

There are 2 main goals for this project, customer segmentation and mailout response prediction. The solution to solve these goals can be described as follow:

1. Customer segmentation
    a. The machine learning pipeline will be implemented to cluster the individuals. The pipeline components are consist of
        i. Transformer to scale the feature using standardization
        ii. Principal component analysis estimator to reduce the dimensionality of the input features
        iii. Minibatch KMeans to cluster the data
    b. The clustering result will be used for marketing campaigns, since we can assign the individual to the cluster and know before approaching them that this individual is a potential customer for Arvato or not.
2. Mailout response prediction
    a. The BayesSearchCV in scikit-optimize will be wrapped to the machine learning pipeline in this step. There are 2 advantages for this method, first is, the hyper parameters can be tuned easily. Second is, data leakage can be prevented since the whole pipeline will be fit only on the training folds and evaluate on testing folds. Furthermore, the customer segmentation results from the previous step will be appended to the original data as an additional feature to exploit the latent information from unsupervised model.
    b. The pipeline components are consist of
        i. Transformer to scale the feature using standardization
        ii. Estimators to predict the mailout response, the candidate models are as follow:
            1. Logistic regression as the representative of the basic model
            2. Random forest classifier as the representative of the bagging model
            3. Adaboost classifier as the representative of the boosting model
            4. Xgboost classifier as the representative of the state of the art model
    c. The prediction result will be used to filter out some of the individuals to approach using mailout so that the marketing team can focus their time on potential individuals to increase efficiency.

Benchmark Model

Since there are no clear formulas to cluster the customer segmentation and to predict the mailout response, the benchmark model for customer segmentation and mailout response will be “random guessing”. Meaning that, the models have to be better than uniform randomly assigned population to the cluster centers on customer segmentation task or uniform randomly predicted responses on mailout response prediction task. If it is not better, then the random guessing should be implemented to solve these tasks instead.

Evaluation Metrics

1. Customer segmentation task
    a. The distribution of the density in each cluster for the general populations and existing customers is not uniform. This can be measured by comparing the distribution of the cluster density between unsupervised model and uniform random model using chi square test with confidence level of 95%
2. Mailout response prediction task
    a. Area under the curve of ROC curve. It has to be greater than 0.5 which is the random guessing result.

Project Design

The workflows for approaching the solutions of this project are designed as follow:

1. Manually identify the null value in JSON file
    a. From preliminary observing the “DIAS Information Levels - Attributes 2017.xlsx”, the null value for each feature in the demographic data is encoded differently. If it is not treated correctly, it will lead to confusion in data exploration and definitely pose a problem in the machine learning pipeline.
2. Metadata exploration
    a. In this step, the description of each feature will be skimmed and filter out irrelevant or biased features (sex for example). Also, the level of measurement will be identified in this step and save it in JSON format.
3. Data exploration
    a. The individuals data will be explored, such as data distribution, data missing percentage or identifying quasi constant features. This information will be used in feature engineering steps to filter out unnecessary features or individuals to boost the performance of machine learning models.
4. Feature engineering
    a. The feature will be filtered out using many criterias, such as correlation on the other feature, missing value percentage, quasi constant feature or the number of cardinality. The selected features will be saved and can be used in both customer segmentation and mailout response prediction.

5. Data cleaning
    a. In this step, for reusable purpose in customer segmentation and mailout response prediction, the data cleaning pipeline will be created using sklearn pipeline and self-custom transformer. The components of the pipeline are as follow:
        i. Transformer for selecting the specific features from full dataframe
        ii. Transformer for encoding null value on each feature
        iii. Transformer to impute missing value using median imputer
        iv. Transformer to encode the nominal level features using one hot encoder
6. Customer segmentation
    a. The machine learning pipeline will be implemented to cluster the individuals. The pipeline components are consist of
        i. Transformer to scale the feature using standardization
        ii. Principal component analysis estimator to reduce the dimensionality of the input features
        iii. Minibatch KMeans to cluster the data
    b. The results in this step will be helpful in mailout response prediction because we can know beforehand the potential customer from unsupervised machine learning model.
7. Mailout response prediction
    a. The Cross Validation Model selection in sklearn will be wrapped to the machine learning pipeline in this step. There are 2 advantages for this method, first is, the hyper parameters can be tuned easily. Second is, data leakage can be prevented since the whole pipeline will be fit only on the training folds and evaluate on testing folds. Furthermore, the customer segmentation results from the previous step will be appended to the original data as an additional feature.
    b. The pipeline components are consist of
        i. Transformer to scale the feature using standardization
        ii. Estimators to predict the mailout response, the candidate models are as follow:
            1. Logistic regression as the representative of the basic model
            2. Random forest classifier as the representative of the bagging model
            3. Adaboost classifier as the representative of the boosting model

References

[1] "Arvato - Bertelsmann SE & Co. KGaA," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/divisions/arvato/#st-1. [Accessed 15 03 2021].

[2] P. Premkanth, "Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC," Global Journal of Management and Business Research, pp. 33-40, 2012.

[3] "Customers Insight," AZ Direct, [Online]. Available: https://www.az-direct.com/site/en/products/customer-insights/. [Accessed 15 03 2021].