arvato BERTELSMANN

UDACITY

Customer Segment Identification

Capstone Proposal

Ken Chen

chken4869@hotmail.com

1. Domain Background

Arvato Financial Solutions, a Bertelsmann subsidiary. It provides solutions including ID & Fraud Management, Credit Risk Management, Payment & Financial Services and Debt Collection Services. [1]

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. By enabling companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention.[2]

Customers are different in many ways: Customers are different in many ways:[3]

- Needs
- Capabilities
- Business economics and strategies
- Willingness to engage in business with you
- Demographic characteristics

In this project, we will analyze demographics data for customers of a mail-order sales company in Germany, find custom segments and predict potential customs' response to market campaign. [4]

2. Problem Statement

1) Customer segmentation for the company's existing customers.

We will use unsupervised machine learning algorithms to analyze the market and segments, selecting the main characteristics that can best describe the company's consumer.

2) Predict people's response to the marketing campaign of our client.

This is a binary classification problem with highly imbalanced class, with only about 1.23% of our customers in the data responds positively to the campaign.

3. Datasets and Inputs

The datasets are provided by Arvato Financial Solutions.

Data	Usage	Description	Size
DIAS Information Levels - Attributes 2017.xlsx	Metadata	Top-level list of attributes and descriptions, organized by informational category	
DIAS Attributes - Values 2017.xlsx	Metadata	Detailed mapping of data values for each feature in alphabetical order.	
Udacity_AZDIAS_052018.csv	Custom Segmentation	Demographics data for the general population of Germany	891211 persons x 366 features
Udacity_CUSTOMERS_052018.csv	Custom Segmentation	Demographics data for customers of a mail-order company	191652 persons x 369 features
Udacity_MAILOUT_052018_TRAIN.csv	Predict Model	Demographics data for individuals who were targets of a marketing campaign	42982 persons x 367 columns
Udacity_MAILOUT_052018_TEST.csv	Predict Model	Demographics data for individuals who were targets of a marketing campaign	42833 persons x 366 columns

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether each recipient became a customer of the company.

4. Solution Statement

The solution comprises following steps:

Step 1: Exploratory Data Analysis and Data Cleaning

In this step we will explore the datasets by visualizing. We will clean the datasets according to their description, handling missing values, separating categorical values and create basic features for the next step.

Step 2: Feature Engineering, Custom Segmentation with Unsupervised Learning Models

In this step we will create useful features from the cleaned data and build unsupervised learning models such as K-means for the clustering problem.

Step 3: Feature Engineering, Detect Potential Customers with Supervised Learning Models

In this step, we will use previous analysis and build supervised models such as KNN and Random Forrest to predicts whether or not each individual will respond to the campaign.

Step 4: Prediction, Kaggle Competition

In this step, we will use the previous model to generate a submission file for the related Kaggle Competition.[5]

5. Benchmark Model

1) Custom Segmentation

The benchmark model is simple K-means model, which we can just include basic features.

2) Detect Potential Customers Problem

The benchmark model is simple Logistic Regression because it's a linear classifier and easy to train.

6. Evaluation Metrics

1) Custom Segmentation

Squared Error.

2) Detect Potential Customers Problem

Precison, Recall, F1 Score, AUC Metric.

7. Outline of Project Design

Step 1: Data Processing

The data will be loaded and applied some descriptive statistics. We will handle the conditions of missing values, incorrect values etc. We can handle in the following ways:

i. Drop any incomplete rows of data.

ii. Re-code the numbers for text descriptions so the model can read them.

iii. Normalize the data using methods from Scikit-Learn.

Step 2: Feature Engineering

We will use some feature engineering techniques to extract useful features for future use. Due to the huge number of features, we will get explained variance of features in the dataset and determine the necessary features using dimensionality reduction techniques such as PCA.

Step 3: Model Training

We will use the features to build appropriate models for the two problem.

Firstly, we will build unsupervised learning model for customer segments identification. We will use K-means as our base model and then try models such as Affinity propagation, Mean-shift, Hierarchical clustering, DBSCAN and OPTICS for improvements.

Then, we will build supervise learning model to predict a person's willingness to become a customer. We will use Logistic Regression as our base model and then improve its performance using Decision Tree, Random Forests and other classification models.

In this part, we will split data into train/validation/test sets and compare among different models. And we will

Step 4: Parameter Tuning

For some complicated models, we should do some parameter tuning job to improve the performance. A hyper parameter tuning algorithm like Grid Search will be used to determine the best set of hyper parameters.

Step 5: Model Combine

In this step, we can combine previous good models and generate predictions.

Step 6: Final Predictions

In the end, we can upload our prediction to the Kaggle competition.

References

[1]. Bertelsmann, Arvato. About.2020.

https://www.bertelsmann.com/divisions/arvato/

[2]. Custom Segmentation Definition.

https://searchcustomerexperience.techtarget.com/definition/customer-segmentation

[3]. Reasons for Customer Segmentation

https://ocw.mit.edu/courses/sloan-school-of-management/15-902-strategic-management-i-fall-2006/lecture-notes/custseg.pdf

[4]. Udacity, Bertelsmann/Arvato Project Overview. 2019.

https://classroom.udacity.com/nanodegrees/nd009t/parts/2f120d8a-e90a-4bc0-9f4e-

43c71c504879/modules/2c37ba18-d9dc-4a94-abb9-066216ccace1/lessons/4f0118c0-20fc-482a-81d6-b27507355985/concepts/8400bad9-69b4-4455-826c-177d90752f00

[5]. Kaggle.com. (2020). Udacity+Arvato: Identify Customer Segments | Kaggle.

https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard .