UDACITY

Machine Learning Nanodegree Program

Capstone Proposal: Customer Segmentation Report for Arvato Financial Services

MADHAV

JUNE 10th, 2021

Domain Background

"Arvato is an international based service sector company providing it's various services through software and hardware technologies with a focus on innovations in automation and data/analytics. Globally renowned companies from various domains such as telecommunications providers , IT services, Net

Banking, e-commerce– rely on Arvato's portfolio of solutions. Arvato is fully owned by Bertelsmann. [1]"

Arvato is looking all the possible ways to support a client-base(mail-order company selling organic products) with available datasets to find the target audience for their campaign. To achieve this, I will use the available dataset to segment the customers based on their interest with attributes and demographic feature that are more suitable.

With the help of Data Analysis and Machine Learning concepts many underlying patterns can be discovered and also handle very high volumes of data efficiently.

Problem Statement

The Problem Statement for this project is pretty straight forward "How can an existing mail order company selling organic products can acquire more number of clients in order to expand their business?"

The proposed solution is divided into 2 subparts.

The first part would deal with the segmentation of the customers using unsupervised machine learning techniques based on the intersection of the data with current customers and population provided.

Secondly, a supervised model will be used to predict the outcome whether or not a customer is likely to be in the targeted audience or not.

Datasets and Inputs

The dataset is already provided by Bertelsmann Arvato Analytics, following are the files along with their description provided for this project:

- **Udacity_AZDIAS_052018.csv**: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- **Udacity_CUSTOMERS_052018.csv**: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).

- **Udacity_MAILOUT_052018_TRAIN.csv**: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- **Udacity_MAILOUT_052018_TEST.csv**: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Additionally, there were 2 more files for describing attributes:

- **DIAS Attributes - Values 2017.xlsx**: Explains values encoding.
- **DIAS Information Levels - Attributes 2017.xlsx**: Explains column names meanings.

MAILOUT_TRAIN and MAILOU_TEST are provided for the building training and testing a supervised model.

Solution Statement

There are several steps to be performed towards problem statement.

1. Data exploration and preprocessing

   Firstly, the given data must be explored such as its shape, features, null values, any mixed types etc. Followed by data cleaning such as dropping empty values based on a condition and dealing with mixed type formats and removing outliers if any. Then, the data is transformed by encoding categorical values into numeric types.

2. Segmentation

   Here, the dimensionality reduction technique known as Principle Component Analysis will be used to reduce the dimensions because original data contains 366 features not all of them play a crucial role. Afterwards elbow method will be used to find appropriate number of clusters to be used in the K-Means technique.

3. Predictive modeling
   In last section a predictive model using supervised machine learning models will be build such as XGBoost, ADABoost, RandomForests etc. At this stage it is pretty difficult to say which one would be suitable or which does not. It’s just an assumptions

Benchmark Model

The benchmark model would be XGBoost because of its feasibility and efficiency and more than that one cannot ignore the fact that it is the most preferred than other existing ones.

Evaluation Metrics

For the dimensionality reduction algorithm PCA, a plot between experienced variance and number of components can be taken into consideration to choose number of components.

For predictive based modelling, since we are going to use classification models metrics such as confusion matrix and AUC_ROC curves can be used for evaluation.

Project Design

The proposed architecture of the project should look as follows:

1. Data Preprocessing: This section is completely based on data exploration as explained above such as filling the missing values or dropping them, dealing with mixed types, removing the outliers.
2. Feature engineering: Determining the most relevant features with the help of the unsupervised techniques such as PCA and K-Means algorithms.
3. Supervised model implementation: Several above mentioned supervised learning algorithms will be used to predict the accurate outcome.
4. Model tuning: Several model performance will be closely evaluated through metrics and the one which gives good performance will be considered for the evaluation.
5. Evaluation and testing: Finally, the best tuned model will be used in the Kaggle competition.

References

[1] Available: https://www.bertelsmann.com/divisions/arvato/#st-1 – Brief description about the company[Accessed June 08-2021]

[2] https://towardsdatascience.com/how-customer-centric-marketers-use-machine-learning-387df1a33850 [Online].

Available: https://towardsdatascience.com/how-customer-centric-marketers-use-machine-learning-387df1a33850 [Accessed June 9th, 2021]