Capstone Project

Customer Segmentation Report for Arvato Financial Services

Domain Background:

Arvato is a global services company headquartered in Germany. Its services include customer support, information technology, logistics, and finance. Arvato's financial solutions include payment processing, factoring, and debt collection services. Financial services have been one of the most profitable Arvato businesses and are also a strategic growth area for Bertelsmann. In this project we will be using available datasets on customers of Arvato to target new client base with the help of machine learning techniques.

Problem Statement:

Which individual's company should target who are most likely to convert into becoming customers of the company?

We will solve this problem in sub-parts:

1) Unsupervised learning techniques to perform customer segmentation
2) Supervised learning approach to target customers for the marketing campaign

Datasets and Inputs:

The data that we will use has been provided by Arvato Analytics and represents a real-life data science task.

There are 4 datasets and 2 metadata files used in this project:

Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany

Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company

Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign

Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign

DIAS Information Levels — Attributes 2017.xlsx: A top-level list of attributes and descriptions

DIAS Attributes — Values 2017.xlsx: A detailed mapping of data values for each feature

Solution Statement:

To develop unsupervised model, first we need to convert all non-numerical features to numeric features if we want to use those in the model then need to do feature scaling so that it does not impact overall weight of principal component. We will use PCA to reduce the dimension of the data. Then will use KMeans for clustering.

Once clustering is done, will use the supervised learning to predict which individuals we should target to increase our customer base. At this point of time I am not sure which model I will be using to solve this problem. But first I will start with the simple one Logistic regression then I will explore other complicated models if I do not get desired result.

Benchmark Model:

In Kaggle usually people have achieved the 70-80% performance relating to customer conversion and targeted marketing response.

Evaluation Metrics:

For unsupervised learning -> Explained variance ratio (PCA)

For supervised learning -> Precision, Recall and Accuracy can be used as a metric. For Regression based models we can use Mean Absolute Error and Mean Squared Error.

The final decision will be based on the data balance of classes and its distribution across multiple attributes.

Project Design:

Product will be distributed in the following parts:

Data Clean-up: First we need to clean-up the data to remove any missing data and improper data entries. Missing data will be handled on a case by case basis.

EDA: After data clean-up will do the exploratory data analysis to figure out which attributes should be used in the model. Its extremely useful to look at the relationship between different variables to get the better sense of the data.

Feature Engineering: After feature selection we will convert all non-numerical attributes to numerical. Need to normalize the values so that it can be used in the model.

Modelling: We need to try different models to see which works for our data. It will be tried-and-tested methodology to come up with the best model.

Testing and Prediction: will use the evaluation metrics (as mentioned above) to predict the success of our modelling.