Capstone Project Proposal

Create a Customer Segmentation Report for Arvato Financial Services

Domain Background

Correctly predicting the respond to campaigns is important for companies so that they target the most appropriate groups of people, hence saving on cost. It also helps them acquire new customers more efficiently as instead of reaching out to everyone, they would just reach out to people identified as most likely to respond to the marketing campaign. If a company were to market their product without a strategy, they might not be able to reach potential customers as easily.

Arvato is a services company that develops and implements innovative Supply Chain Management (SCM), financial and IT solutions for business customers globally ("Arvato - Bertelsmann SE & Co. KGaA", 2020). In this project, a client of Arvato Financial Services, i.e. a German mail-order sales company is looking into targeted marketing on the population to acquire new customers. Attributes of past clients of the company will be analysed and matched to the attributes of the general population in Germany to identify new clients for the company.

Problem Statement

The problem statement of this project is “How can a mail-order company acquire new clients in an efficient manner?”

We can break this down further into, what are the attributes of general population that customers correspond to? And how likely are customers to respond to marketing campaigns?

Datasets and Inputs

4 datasets are provided by Arvato Financial Services 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).

The first 2 datasets will be used in an unsupervised learning task to divide the population and customers into clusters. The final 2 datasets will be used for training and testing in a supervised learning task to predict which customers are likely to respond to a marketing campaign.

Solution Statement

The project is broken down into 3 parts. First, doing data exploration. Next, I will use unsupervised learning techniques to divide the population and customers into clusters, hence, identifying the cluster of the population that customers are in. Lastly, through supervised learning, I will predict customers that are most likely to respond to the marketing campaign held by the company.

Benchmark Model

The benchmark model will be a decision tree classifier. They are simple to implement and can be interpreted easily.

Evaluation Metrics

Since the classes are highly imbalanced, with most people not likely to respond to the campaign, using accuracy will not accurate determine how good our model is. The Area under the ROC curve (AU ROC) is used to evaluate how good the model is. The ROC curve is a plot of the True Positive Rate against the False Positive Rate, which can be thought of as the proportion of correct predictions for the positive class against the proportion of errors for the negative class. Ideally, it should be in the top left of the plot. The AU ROC provides a score between 0 and 1 to quantify how good the model is.

Project Design

Data Visualisation and Cleaning

I will explore the missingness of data in the dataset. I will also do a check between the columns in the population and customer dataframe.

Feature Engineering

I will then do some feature engineering to extract more useful information from the dataset. Categorical columns will also be one-hot encoded. An imputer will be used to fill in any missing values left.

PCA and Clustering

The data is then scaled before doing principal component analysis on it to extract the important components in the data. I will then fit a K Means clustering algorithm to cluster the population and customers into groups. By comparing the proportion of population and customer in each cluster, we can determine which cluster has much more or less customers compared to the general population. Through an inverse transformation of the principal component analysis, we can then determine which factors are important in determining the attributes of a customer.

Training, Modelling and Tuning

I will then apply the same cleaning, feature engineering and imputation transformations on the training and testing dataset and train supervised learning models to predict how likely an individual is likely to respond to a marketing campaign. Different classifiers such as a random forest and XG Boost model can be used for this purpose and the hyperparameters of the models can be tuned for the best effect.

Customer Prediction

The tuned model will be used to make predictions on the test data and submitted to the Kaggle competition.

References

Arvato - Bertelsmann SE & Co. KGaA. Bertelsmann.com. (2020). Retrieved 24 September 2020, from https://www.bertelsmann.com/divisions/arvato/#st-1.