Arvato Financial Services

Capstone Proposal

Domain Background

Arvato is a company that provides Information Technology and financial services for business customers globally. Arvato solutions focus on automations and data analytics. The company is owned by Bertelsmann an educational media company. There customers range from large companies such as Internet providers or even e-commerce. Arvato services can bring data insights to company's that need to make valuable business decisions. Taking data from demographics and previous customers to understand the different segments of customers and create a system that predicts whether someone would be a potential customer based on demographics.

Problem statement

The problem statement derives from the question, "how can a mail order company bring in customers in an effective way, given the demographics of a someone". Without segmenting the customers in an effective way mail order company cannot spend unreasonable amounts of money on campaigns. The demographic data of the population needs to be studied with unsupervised learning algorithms. The solution here is to identify segments in the population and existing customers that correspond to a person being a potential customer. We will also need a supervised learning algorithm that will predict based on demographic data if that person will likely be a customer or not.

Dataset and Inputs

These are the files 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).

Two metadata files were also provided:

- DIAS Attributes – Values 2017.xlsx (level list of attributes and descriptions)
- DIAS Information Levels – Attributes Values 2017.xlsx (mapping of each data value)

These files have been provided by Arvato to Udacity for the customer segmentation and analysis for Machine learning nanodegree. The demographic files are the four csv files, each role represents a person. Each role includes information such as building, household and demographics. The customers data has three columns that relates them regarding the mail order company. The train and test data evaluate supervised learning algorithms.

Solution Statement

First, we must identify the customer segments in the provided dataset and match the segments to the current population in the general population dataset. The dataset needs to be analyzed to study if there are any missing data or errored values so we can correct them. With the help of label encoders, we can also encode any categorical features. Then the data will be measured to ensure that no single feature will have higher weights. Next, we need to identify the number of features that would suffice to explain the dataset. A PCA will be used to reduce the minimum number of features since 366 features would be too much for a single person. Then with the assistance of an unsupervised learning algorithm we would segment the general population and customers into different segments based on selected features.

Benchmark Model

Before classifying the model, I plan on testing to see for accuracy to see which model would be more reasonable. A logic regression could also be a great model since it is easy to train and test within less time. The performance of the model would be the baseline for where different algorithms can compare to this benchmark to choose whether to proceed with the model.

Evaluation metrics

The project will be divided into to parts, the customer segmentation using unsupervised learning algorithms and customer acquisition using unsupervised learning algorithms. The customer segmentation using unsupervised learning algorithm's part will use the PCA technique to reduce the number of dimensions. The variance ratio of each feature would be the would be reference in selecting the number of dimensions. The number of dimensions would explain the number of variances in the data set. The customer acquisition using unsupervised learning algorithm's part would predict whether the mail-order company should contact the customer. In this part the training data set will be split into train and evaluation sets. The model will be trained on the training split and evaluated on the evaluation split. The classification for the evaluation metrics would include: Accuracy, Matrix and Area under the receiver operating curve.

Project Design

1. Data cleaning: The dataset needs to be analyzed to study if there are any missing data or errored values so we can correct them. All the errored values will be examined and corrected based on the information of the

metadata files. A visualization will also be done to analyses any common patterns in the data.

2. Feature engineering: Understanding the variance of features in the data then determining the required number of features that can amount for maximum variance using dimensionality reduction techniques such as PCA.
3. Modeling: Identifying the customers using unsupervised learning algorithms. K-means clustering algorithm will be used to segment the data into the needed number of clusters. Then the supervised learning algorithm will be trained and evaluated whether someone could potentially be our next customer or not.
4. Model Tuning: After examining the different algorithms and performances the algorithm that has a good score will be chosen and tuned for improvements.
5. Prediction for the Test data: Then the best model will be chosen to for predictions on the test data and be submitted to the Kaggle competition.

References: Arvato – Bertelsmann, “Arvato,” – Bertelsmann [Online] https://www.bertelsmann.com/divisions/arvato/#st-1

Bertelsmann, “Company,” Bertelsmann [Online] https://www.bertelsmann.com/company/