Bertelsmann/Arvato Project

Domain Background: The data and outline of this project was provided by Arvato Financial Solutions, a Bertelsmann subsidiary. The background of this project is in 'Marketing Analysis' or using predictive analysis using exploratory data analysis and machine learning techniques.

Problem Statement: The problem we are trying to solve here is to extract the customer segment from the data that will have a population who will be potential enough to convert as a customer.

Datasets and Inputs:

1. Udacity_AZDIAS_052018.csv - Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
2. Udacity_CUSTOMERS_052018.csv - Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
3. Udacity_MAILOUT_052018_TRAIN.csv - Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
   This training set is specially curated for the Supervised learning process.
4. Udacity_MAILOUT_052018_TEST.csv - The dataset we need to run our predictions on. It consists of the demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Solution Statement:
Through exploratory data analysis we can understand the patterns in the data and get a summary of the features, about it's sparsity, completeness.
By applying pre-processing techniques to the data we can select the features that has more relevant data, replace the NULL values and trim the datasets in order to make it compatible for the pipeline.
Apply feature scaling to bring uniformity to data
Apply unsupervised learning techniques like K-Means and supervised techniques like Logistic Regression or Random Forest to classify the potential customer to non potential customers.
With this prediction algorithm we can apply it on future data and help the company to optimize their spend for maximum conversion.

Benchmark Model:

The benchmark models to be used are :

- Predictive analysis methods used via unsupervised and supervised learning.
- Preprocessing of data with Imputers and Scaling techniques
- Feature selections
- Evaluation metrics

- Proper analysis charts for the comparison of contribution of general data into actually converted data

Evaluation Metrics:
Either use ROC metrics or Precision - Recall metrics depending on the sparsity of the data

Project Design:
- Use matplotlib for exploratory data analysis
- Use Imputer, StandardScaler from sklearn.preprocessing to make uniformity in the data
- Use unsupervised learning techniques like K-Means to cluster the data
- Use supervised learning techniques like Logistic Regression or Random Forest algorithm to classify each customer in the training set and evaluate using evaluation metrics
- Use evaluation metrics that focus on the accuracy and precision of the model to choose the model for evaluation
- Choose the best model and save it
- Use the saved model to predict on the test data
- Create CSV data for Kaggle submission

Proposal Review Link:

Link to Udacity review for the proposal