Machine Learning Engineer Nanodegree

Capstone Proposal

Domain Background

Arvato offers financial solutions offers a wide range of services from Id & Fraud Management to payment financial services. The company is looking to use its available dataset to analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population and predict which individuals are most likely to convert into becoming customers for the company. To achieve this goal I will explore Arvato's existing datasets to identify attributes and demographic features that can help segment customers of interest for this particular client.

Problem Statement

The problem statement for this project is "How can company identify which individuals are most likely to convert into becoming customers for the company?"
I will use an unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments, and then follow-up on the discovered customer segments with a supervised learning approach using a dataset with demographics information for the target customers for the advertising campaign and predict which individuals would be more likely to convert to company customers

Datasets and Inputs

There are four data files associated with 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).

. And 2 metadata files associated with these datasets:

- DIAS Information Levels — Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category

- DIAS Attributes — Values 2017.xlsx: a detailed mapping of data values for each feature in alphabetical order

Which can help mapping the attributes to its particular type or missing value encoding.

Solution Statement

Since there are two sub-domain problems which are unsupervised learning and supervised learning methods, the first step will include pre-process the data before feeding to unsupervised learning method. At first glance through the dataset, there are 4 type of values which are ordinary, numeric, categorical, and mixed types. So I will encode non-numerical values followed by feature scaling to guarantee that the natural scale of the features does not affect their overall weight on the principal components. PCA would be used to reduce dimension before applying KMeans as a form of partitioned based clustering (efficient and good performer for medium to large datasets which is our case).

Once we finish creating customer segmentation, supervised learning method will be used to identify attributes and demographic features to predict which individuals would be more likely to convert to company customers. There are number of models that we should test out:

- Logistic regression
- DecisionTreeRegressor

Since this is just a proposal, these models are not fixed but open to another approches that get high efficiency.

Benchmark Model

For this problem, it is suggested to use Gradient Boosting Classifier based on consulted data sets of historical relevance on Kaggle relating to customer conversion and targeted marketing response (performances nearing 80%).

Evaluation Metrics

For the first part of the problem using unsupervised learning, explained variance ratio can be used when we are implementing PCA, as it accounts for the description of feature variance, allowing for the determination of more important features that stand out with more explained variance.
For the prediction portion of the project (supervised learning) precision can be used as a metric, as does accuracy and recall. Regression based models also benefit from using Mean Absolute Error and Mean Squared Error.

Project Design

1. **Get to know data:** Examine how may rows, columns, features, what kind of values in each dataset

2. **Data preprocessing:** Filling unknown values, missing values, or drop rows, columns that missing the most

3. Feature Engineering: Implement PCA, find most relevant features, eliminate features of low importance for optimal model training further in the project.

4. Model Selection: Experiment with the before-mentioned algorithms to find the ideally suited for this problem, namely KMeans for the unsupervised learning portion and Logistic Regression, DecisionTreeRegressor

5. Model Tuning: Once we find the model that best suits our data, adjust model parameters within a range that allows for increased performance without overfitting, increase awareness for possible data leakage.

6. Test and Predict: use the previously proposed metrics, explained in the table present in the section for evaluation metrics as an indicator of success in our predictions.