Machine Learning Engineer Nanodegree

Capstone Project Proposal

Bertlesmen-Arvato Customer Segementation Project

J. P. Bedran

Domain Background

Bartlesmen – Arvato Customer segmentation analysis.

Arvato financial solutions is engaged in a project to bring more customers to a Mail-Order Company in Germany.
Arvato is company focused in financial solutions utilizing technology to deliver excellent results to their clients. In this scenario, we will be utilizing Machine Learning models to achieve the goals specified by the client.

Problem Statement

As stated by Timo Reis, our main goal with this project is to acquire more customers efficiently for our client utilizing our technical expertise and the datasets provided by Arvato and German govt.

Datasets and Inputs

Arvato provides the datasets and inputs that are going to be used in this project. They are as follows:

- 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).

Solution Statement

By utilizing unsupervised models in the dataset such as k-means, I will segment population and compare it to the customers. Indicating similarities between the groups, making it easier to identify individuals with greater probability of becoming our client's customers.

Benchmark Models

Utilizing GBC on consulted data sets of Kaggle relating targeted marketing

Evaluation Metrics

Will be utilizing accuracy (TP + TN/ TP + TN + FP + FN) if data is not imbalanced, uniform. If data is proven to have imbalance, will utilize recall (TP/ TP + FN) and precision (TP/ TP + FP) two have a more thorough view of the result the model rendered

Project Design

1. Data Cleanup: Examining missing values for features replacing or dropping them, accordingly. Examine, analyze and understand data in general and prepare for step 2.

2. Data Visualization: Utilizing matplotib in conjunction with pandas, makes able to thoroughly analyze and makes sense of this amount of data, better understanding and adjusting approaches to solve problem in the best way.

3. Feature Engineering: Apply PCA to data, choose most relevant features that cause greatest variance to model. Utilize most relevant features in next step.

4. Model Selection: Create model starting with KMeans for unsupervised learning, trying different models for optimum results. For supervised learning, initially try GBC, being pssible to go through GSCV, mainly focusing on optimal results.

5. Model Tuning: Adjust model parameters of chosen model to increase performance.

6. Test and Predict: Utilize skleran.metrics library to apply accuracy, recall and/or precision tests to data.