MACHINE LEARNING ENGINEER NANODEGREE CAPSTONE PROPOSAL

OUSHNIK DEY

Domain Background

Machine learning techniques are widely used these days to target potential customers by companies across all fields. Technique such as clustering has proven to be extremely helpful in this area. Arvato Financial Solutions has paired up with Udacity for their nanodegree program to create a customer segmentation report for them and predict which individuals are most likely to become a customer of their company.

Problem Statement

Analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. I will have to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then I will have to apply this learning on a third dataset with demographics information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company.

Datasets and Inputs

The data that I will use has been provided by Udacity's partners at Bertelsmann Arvato Analytics, and represents a real-life data science task.

There are four data files associated with this project:

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).
4. **Udacity_MAILOUT_052018_TEST.csv**: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. I will be using the information from the first two files to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS"), then use this analysis to make predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company.

Solution Statement

In the first part of the project I will use unsupervised learning technique (k-means clustering) to describe the relationship between the demographics of the company's existing customers and the general population of Germany. This will help us to describe parts of the general population that are more likely to be part of the mail-order company's main customer base, and which parts of the general population are less so.

In the second part of the project I will be using supervised learning techniques to predict which individuals are most likely to become a customer of the company. As a part of this I will try out various classifiers like AdaBoost, Random Forest, Gradient Boosting. I will be using Grid Search to find out the best estimator.

Benchmark Model

Since this is a part of Kaggle competition, the benchmarking will be done based on the highest score (0.0819) for this competition.

Evaluation Metrics

The evaluation metric for this Kaggle competition is AUC for the ROC curve, relative to the detection of customers from the mail campaign.

Project Design

- Programming language: Python 3
- Libraries : Pandas, Numpy, Scikit-learn, Matplotlib, Seaborn
- Workflow:
    - Preprocessing step to assess missing data and re-encode categorical and mixed features.
    - Feature selection by applying feature scaling and then perform dimensionality reduction using PCA.
    - Apply k-means clustering on PCA transformed data to find out which parts of the general population are more likely to be part of the mail-order company's main customer base.
    - Use classifiers like AdaBoost, Random Forest, Gradient Boosting to find out which individuals are most likely to become a customer of the company. Apply grid search to come up with the best classifier and then tune its hyperparameters to improve the performance.
    - Test the model on Kaggle competition.