Machine Learning Engineer Nanodegree

Capstone Proposal

Joe Abraham

September 26, 2020

Proposal

The project is done to fulfill the requirements of Machine learning nanodegree, I am choosing a project and dataset provided by Udacity and Arvato Financial Solutions.

Domain Background

About Arvato Financial Solutions

Arvato Financial Solutions provides professional financial services to renowned international brands as well as respected local businesses — allowing them to leave their credit management to a professional, so they can focus on what matters most for their business.

The services center around cash flow in all segments of the customer lifecycle: from identity, fraud and credit risk management, to payment and financing services and debt collection.

The Arvato Financial Solutions team is made up of proven and reliable experts in around 20 countries, including 7,000 IT, analytics, process and legal specialists, dedicated to revealing the advantages of predictive analytics, leading-edge platforms and big data.

All employees are aligned by a common goal: to make sure client's credit management runs effortlessly and efficiently, ultimately resulting in optimized financial performance.

For further information please visit finance.arvato.com

Source : https://www.linkedin.com/company/arvato-financial-solutions/about

Problem Statement

The Arvato Financial Solutions wants to find the potential customers, additionally they want to know more about the customers and their background. For this reason, analysis of demographics data for customers is done against demographics information for the general population.

1. How to identify parts of the general population to be ideal customers?
2. A method to identify individuals who are likely to be a customer ?

Datasets and Inputs

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).

Additional support files available :

1. DIAS Information Levels - Attributes 2017.xlsx is a top-level list of attributes and descriptions, organized by informational category.
2. DIAS Attributes - Values 2017.xlsx is a detailed mapping of data values for each feature in alphabetical order.

Due to privacy concerns - Datasets are protected and are not publicly available.

Solution Statement

Part 0 - Knowing the Data

Data preprocessing - Inspect the data and necessary data cleaning needs to be done.

Part 1 - Unsupervised learning

Clustering needs to be done with the demographic data provided in Azdias(general population) and customers. For that PCA and K-Means clustering will be the best method and clustering needs to be done with the best model.

Part 2 - Supervised Learning

Here the supervised learning algorithms have to be applied on Udacity_MAILOUT_052018_TRAIN.csv and the best performing model will be tuned for best results.

Benchmark Model

Supervised Learning, Logistic regression is used as a benchmark model.

Evaluation Metrics

The evaluation metric for this competition is AUC for the ROC curve, relative to the detection of customers from the mail campaign. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers).

The line plotted on these axes depicts the performance of an algorithm as we sweep across the entire output value range. We start by accepting no individuals as customers (thus giving a 0.0 TPR and FPR) then gradually increase the threshold for accepting customers until all individuals are accepted (thus giving a 1.0 TPR and FPR). The AUC, or area under the curve, summarizes the performance of the model. If a model does not discriminate between classes at all, its curve should be approximately a diagonal line from (0, 0) to (1, 1), earning a score of 0.5. A model that identifies most of the customers first, before starting to make errors, will see its curve start with a steep upward slope towards the upper-left corner before making a shallow slope towards the upper-right. The maximum score possible is 1.0, if all customers are perfectly captured by the model first. (It should be noted that this particular task is very difficult with a lot of noise, and so you should not expect extremely high scores!)

Source(https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation)

Project Design

The whole project is divided into 4 parts:

- Part 0 : Know the data, where you try to understand the data
- Part 1 : Customer Segmentation (Unsupervised Learning)
- Part 2 : Supervised Learning
- Part 3 : Kaggle competition

Part 0 : Know the data, where you try to understand the data

This section contains - Data preprocessing.

- Replacement of strings to float
- Replacement missing values to np.nan - data taken from the excel provided.
- Replacement na with mean
- LNR set as index
- Cleaning columns if it doesn't have enough data (if the dataset is used for clustering)

Part 1 : Customer Segmentation (Unsupervised Learning)

This section contains -

- PCA (Dimension Reduction)
- Scaling
- K-Means Clustering with different centeroids
- Model

Part 2 : Supervised Learning

- Training in different models and check of their learning curve
- Hyper parameter tuning on the best performing model.
- Scoring

Part 3 : Kaggle competition

With the best performing supervised model prediction is done to the dataset Udacity_MAILOUT_052018_TEST.csv and the results are to submitted on the kaggle website