Machine Learning Engineer Nanodegree

Capstone Proposal: Targeted Advertising (Arvato)

Shuo Cheng Jan 11th, 2020

Proposal

Domain Background

Arvato Bertelsmann is an International media company. One of the firm's key focuses is to provide market analysis for its clients to target and acquire new customers more accurately and efficiently. [1] Levering creative data science providing robust conviction and timing, a company could find the right potential target customers in the most cost/time efficient manner. The research and models applied in the study would include Exploratory Data Analysis[2], unsupervised learning (PCA & K-means) and supervised learning (Logistic Regression, Ada Boost and Gradient Boosting Classifiers).

Problem Statement[3]

This project is to aid a mail-order sales company to acquire new German clients for their mail-out campaign based on analyzing and comparing the attributes of general population and potential target Germen population. The end goal is to identify and predict a group of target audience of the campaign that could bring the highest return for the company. This study would point out a general direction for the company to move forward with higher return on investment.

Datasets and Inputs[4]

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).

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.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether each recipient became a customer of the company.

Solution Statement

We will utilize 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 your 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.

Benchmark Model

The benchmark for this project would be Logistic Regression model. Ada Boost and Gradient Boosting Models have more capability to distinguish the binary classes than Logistic Regression. Beating this benchmark model could imply that the project has adequately solve the problem.

Evaluation Metrics [5,6]

For the classification problem, the evaluation metric is AUC - Area under the ROC (Receiver Operating Characteristic) curve. The axes of plane divides data into 4 segments:

- FP: False Positive; The population mistakenly grouped as potential customers
- TP: True Positive; The target customers predicted that are actual customers
- FN: False Negative; The real customers missed by the model prediction
- TN: True Negative; The population that is not potential clients

ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve gradually increases the threshold of Positive Rate, from 0 to ALL. The AUC represents validity of the model, ranging from 0.5 to 1.

Project Design

1. Customer Segmentation Report - Supervised Learning Model [2]
   We begin the project by using unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments.

i. Data loading and exploration
ii. Data cleaning and pre-processing
    - EDA -exploratory data analysis
        - Drop any incomplete rows of data
        - Create a new DataFrame, and drop unnecessary columns
        - Re-code the numbers for text descriptions so the model can read them
        - Normalize the data using MinMaxScaler to transform the numerical values so that they all fall between 0 & 1
iii. Dimensionality reduction with PCA

    PCA attempts to reduce the number of features within a dataset while retaining the "principal components", which are defined as weighted, linear combinations of existing features that are designed to be linearly independent and account for the largest possible variability in the data
iv. Use k-means, to segment customers using their PCA attributes
v. Pass in the training data and assign predicted cluster labels "RESPONSE"
vi. Exploring the resultant clusters & visualization for research purpose

2. Supervised Learning Model [8]

With a third dataset including attributes from targets of a mail order campaign, we use the previous analysis to build a machine learning model that predicts whether each individual will respond to the campaign.

i. Loading and exploring the data
    - Calculate the percentage of fraudulent data and deal with the imbalance
ii. Splitting the data into train/test sets
    - Shuffle the transaction data, randomly
    - Split it into two & get train/test features and labels
iii. Define and train a other models (Ada and Gradient Boosting Models) that performs binary classification. This model should be designed to:
    - Accept several input features (the number of anonymized features).
    - Create some Linear, hidden layers of a desired size
    - Return a single output value that indicates the class score
    - The returned output value should be a sigmoid-activated class score; a value between 0-1 that can be rounded to get a predicted, class label.
iv. Create an estimator making improvements on the model
v. Evaluating and comparing model test performance

3. Consolidating prediction data

After evaluating the test performance on which segment of population are the target consumers, we output the results in a csv file with two columns "LNR" & "RESPONSE" enter the Kaggle competition and assist the firm pin-pointing potential clients.

References

1. Bertelsmann, Arvato. *About.*2020. URL: https://www.bertelsmann.com/divisions/arvato/
2. Tukey, John W. (1977). Exploratory Data Analysis. Pearson. ISBN 978-0201076165.
3. Udacity, Bertelsmann/Arvato Project Overview. 2019. URL: https://classroom.udacity.com/nanodegrees/nd009t/parts/2f120d8a-e90a-4bc0-9f4e-43c71c504879/modules/2c37ba18-d9dc-4a94-abb9-066216ccace1/lessons/4f0118c0-20fc-482a-81d6-b27507355985/concepts/8400bad9-69b4-4455-826c-177d90752f00
4. Udacity, Bertelsmann/Arvato Project Workspace. 2019. URL: https://classroom.udacity.com/nanodegrees/nd009t/parts/2f120d8a-e90a-4bc0-9f4e-43c71c504879/modules/2c37ba18-d9dc-4a94-abb9-066216ccace1/lessons/4f0118c0-20fc-482a-81d6-b27507355985/concepts/e9553619-113b-4565-a34e-a9ef450659de
5. China Science Communication, 2019. URL: https://baike.baidu.com/item/AUC/19282953?fr=aladdin#3
6. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
7. Udacity, Solution: Simple Neural Network, 2019. URL: https://classroom.udacity.com/nanodegrees/nd009t/parts/670d5990-4694-47a7-887d-c84710e15a45/modules/6712751e-f328-4956-8881-bfe0c7e6cd7b/lessons/c461a06f-883b-4a29-bc43-edc1ffba821b/concepts/ed1a320d-db94-4609-aaba-f02b1f473466
8. Udacity, Notebook: Fraud Detection, Exercise, 2019. URL: https://classroom.udacity.com/nanodegrees/nd009t/parts/670d5990-4694-47a7-887d-c84710e15a45/modules/6712751e-f328-4956-8881-bfe0c7e6cd7b/lessons/cf391ace-c3c9-476b-b357-83ef349eb800/concepts/c30267d5-ac17-47de-a6e1-38fd07709a0d