Udacity Machine Learning Engineer Nanodegree Capstone Project Proposal

Amiri McCain

August 19, 2019

Customer Segmentation Report for Arvato Financial Solutions

Project Overview

I have selected the Udacity specific Arvato Project. Completion of this capstone project includes three parts:

1. Create a customer segmentation report for Arvato Financial Solutions.
2. Create a supervised learning model and verify this prediction model with training data.
3. Submit the prediction model to the "Udacity+Arvato: Identify Customer Segments" Kaggle competition for testing.

Domain Background

Arvato Bertelsmann Financial Solutions is an IT service management company headquartered in Baden-Baden, Germany that specializes in optimizing the financial backbone of businesses and provides professional B2B (business-to-business) financial services. Arvato provides expertise in automation and data analytics that provide a comprehensive overview of an entire business and the business's client's journey.[1]

Problem Statement

Arvato Financial Solutions wants their client to be able to analyze attributes to identify payment behavior, predict payment behavior, recommend credit scores and calculate credit scores of their clients. This will allow Arvato's client to more efficiently target the right type of clients for their organic mail order catalog. This appears to be a clustering and classification problem.

The list below indicates how the results of the customer segmentation analysis report will be used and the questions Arvato's client desires to answer:

- How can the client, a mail order company acquire new clients more efficiently for its organic mail order catalog?
- The mission of the engagement with the client is to make decisions based on data instead of gut feel.
- Arvato's client will be using resultant datapoints to improve decisions and show rational behind decisions.
- Arvato wants their client to use and request reports more often.

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).

Solution Statement

Arvato Financial Solutions wants their client to be able to analyze attributes to identify payment behavior, predict payment behavior, recommend credit scores, and calculate credit scores of their clients. This will allow Arvato's client to more efficiently target the right type of clients for their organic mail order catalog.

Unsupervised learning techniques will be used to perform customer segmentation. This will allow parts of the population of Germany that best describe the core customer base of the company to be targeted.

Benchmark Model

The benchmark model will be the random forest model. A small portion of the training data will be used to benchmark the selected model. If this model proves deficient, a different, more appropriate model will be researched, tested and deployed. In addition, other custom algorithms may be tested against the random forest model. [3]

Evaluation Metrics

The prediction model will predict which people are most likely to respond to the advertising that they will receive through the mail. In addition to evaluating this model with training data, this prediction model will also be submitted to the "Udacity+Arvato: Identify Customer Segments" Kaggle competition.[2] The evaluation metric that Kaggle uses for this competition is the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve.[5] The AUC ROC is immune to class imbalance and this dataset will be highly imbalanced.

Project Design

The fist step will be to download all of the different data files and examine their shape, size, structure and data values along with examining the additional files that were provided "DIAS Attributes - Values 2017.xlsx" and "DIAS Information Levels - Attributes 2017.xlsx." These Excel files provide additional information, such as descriptions of

attributes and detailed mapping of data values for each feature and in alphabetical order, about the columns in the primary data files. After a thorough examination, I will begin the process of cleaning the data. Each data file provided will need to be cleaned. I will need to determine which features to keep and which ones to drop. Due to the numerous features in the data, I may need to use Principal Component Analysis (PCA) to reduce the number of features down to only what is pertinent.[4]

Next, I will need to implement an unsupervised learning technique, such as k-means clustering[6], to describe relationship between existing clients and the general population of Germany.

After successfully implementing unsupervised learning techniques for the Customer Segmentation Report, I will need to implement and deploy a successful supervised learning model and verify the model using training data.

Finally, I will submit my supervised learning model (prediction model) to the Kaggle competition for evaluation and scoring on the Leaderboard. For this model, I will likely use XGBoost[7] since it is fast, accurate and works well with large datasets.

References

1. Arvato Financial Solutions: https://finance.arvato.com/en-us/
2. Kaggle competition: https://www.kaggle.com/c/udacity-arvato-identify-customers
3. Random Forest: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
4. Principal Component Analysis (PCA): https://en.wikipedia.org/wiki/Principal_component_analysis
5. Receiver Operating Characteristic (ROC): https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve
6. k-means clustering: https://en.wikipedia.org/wiki/K-means_clustering
7. XGBoost: https://xgboost.readthedocs.io/en/latest/