Machine Learning Engineer Nanodegree

Capstone Proposal

Ming Zhong July 7th, 2019

Proposal

(approx. 2-3 pages)

Domain Background

(approx. 1-2 paragraphs)

In traditional markets, customer clustering / segmentation is one of the most significant methods used in studies of marketing. This study classifies existing customer cluster/segmentation methods into methodology-oriented and application-oriented approaches. Most methodology driven studies used mathematical methodologies; e.g statistics, neural net, generic algorithm (GA) and Fuzzy set to identify the optimized segmented homogenous group.

In recent years, it has been recognized that the partitioned clustering technique is well suited for clustering a large dataset due to their relatively low computational requirements. Behavioral clustering and segmentation help derive strategic marketing initiatives by using the variables that determine customer shareholder value. By conducting demographic clustering and segmentation within the behavioral segments, we can define tactical marketing campaigns and select the appropriate marketing channel and advertising for the tactical campaign. It is then possible to target those customers most likely to exhibit the desired behavior by creating predictive models.

A general literature review can be found in the paper **CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE** by Dr. Sankar Rajagopal from Tata Consultancy Services.

My main motivation to work on this project is to explore the possibility in applying ML in marketing areas. Besides, I have two internship experience in E-commerce companies,

that make me wonder how to combine ML and business together and create true value instead of doing simple class projects.

Problem Statement

(approx. 1 paragraph)

This competition is connected to one of Udacity's capstone project options for the Data Science Nanodegree program, in connection with Arvato Financial Solutions, a Bertelsmann subsidiary.

In the project, a mail-order sales company in Germany is interested in identifying segments of the general population to target with their marketing in order to grow. Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company in order to build a model of the customer base of the company. The target dataset contains demographics information for targets of a mailout marketing campaign. The objective is to identify which individuals are most likely to respond to the campaign and become customers of the mail-order company.

As part of the project, half of the mailout data has been provided with included response column. For the competition, the remaining half of the mailout data has had its response column withheld; the competition will be scored based on the predictions on that half of the data.

Datasets and Inputs

(approx. 2-3 paragraphs)

The data for this project is provided by Udacity partners at Bertelsmann Arvato Analytics, and represents a real-life data science task. It includes general population dataset, customer segment data set, dataset of mail-out campaign with response and test dataset that needs to make predictions.

There are four datasets, all of which have identical demographics features (only part of them are different)

Two dataset for customer segmentation analysis:

- *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)

Because these two datasets are the demographics characteristics of general population and company's customers. It is worthy to explore and compare the clustering analysis between these two groups. And match the general population with our customers. Then, we are able to find out the targeted population that share similar behaviors as our current customers. After cleaning these two datasets in similar fashion, I would like to perform clustering analysis on both datasets with same number of clusters. And then these two cluster distributions were then compared to see where the strongest customer base for the company is.

Two dataset for customer conversion prediction:

- *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). In addition to the above data, there are two additional meta-data:
- *DIAS Information Levels—Attributes 2017.xlsx*: a top-level list of attributes and descriptions, organized by informational category
- *DIAS Attributes—Values 2017.xlsx*: a detailed mapping of data values for each feature in alphabetical order

After customers segmentation, we also want to build a model that can predict whether a potential customer will convert or not in our mail-out list. However, only train set is retained. The test set is withheld for the purpose of Kaggle competition. In order to train and validate our classifier, we need to split the training set into training subset and validation subset. And because of the highly imbalanced nature of our dataset, we need to apply cross-validation(probably with 10 subsets) to train and validate our model.

After training and picking out the best kind of model, we may want to use the same train set to do parameters tuning to further improve our model performance. In the meantime, some classifying method may output the feature importance data that can allow us to communicate our result in business language easily.

In the last step, I prefer to use decision tree to repeat the process again for the sake of better understand and visualize classification results.

Solution Statement

(approx. 1 paragraph)

There are 4 steps to finish the project:

- Data pre-processing: clean and re-encode data.

  Missing values by columns and rows will be analyzed, data will be divided by types followed by subsequent transformations.

- Segmentation:create clusterings of customer and general population, and then identify the difference.

  Use principal component analysis (PCA) technique for dimensionality reduction. Then, elbow and other methods will be used to identify the best number of clusters for clustering algorithm. Finally, apply clustering to make segmentation of population and customers and determine description of target cluster for the company.

- Prediction: use the demographic features to predict whether or not a person became a customer after a mail-out campaign.

  Build machine learning model using response of marketing campaign and use model to predict which individuals are most likely to convert into becoming customers for the company.

  I will use several machine learning classifiers and choose the best using analysis of learning curve(ROC-AUC). Then, I will parametrize the model and make predictions.

- Kaggle competition:The results of this part need to be submitted for Kaggle competition

Benchmark Model

(approximately 1-2 paragraphs)

The benchmark model will be a binary classifier logistic model. Because logistic model is widely used in actual world and easy to apply. And the criteria to measure the performance is ROC-AUC because of the imbalanced nature of the dataset.

Evaluation Metrics

(approx. 1-2 paragraphs)

Because of the data imbalance, we cannot only use recall and accuracy as metrics to measure the model performance. We should consider TP and FP at the same time. Therefore, using ROC-AUC is much more suitable.

Project Design

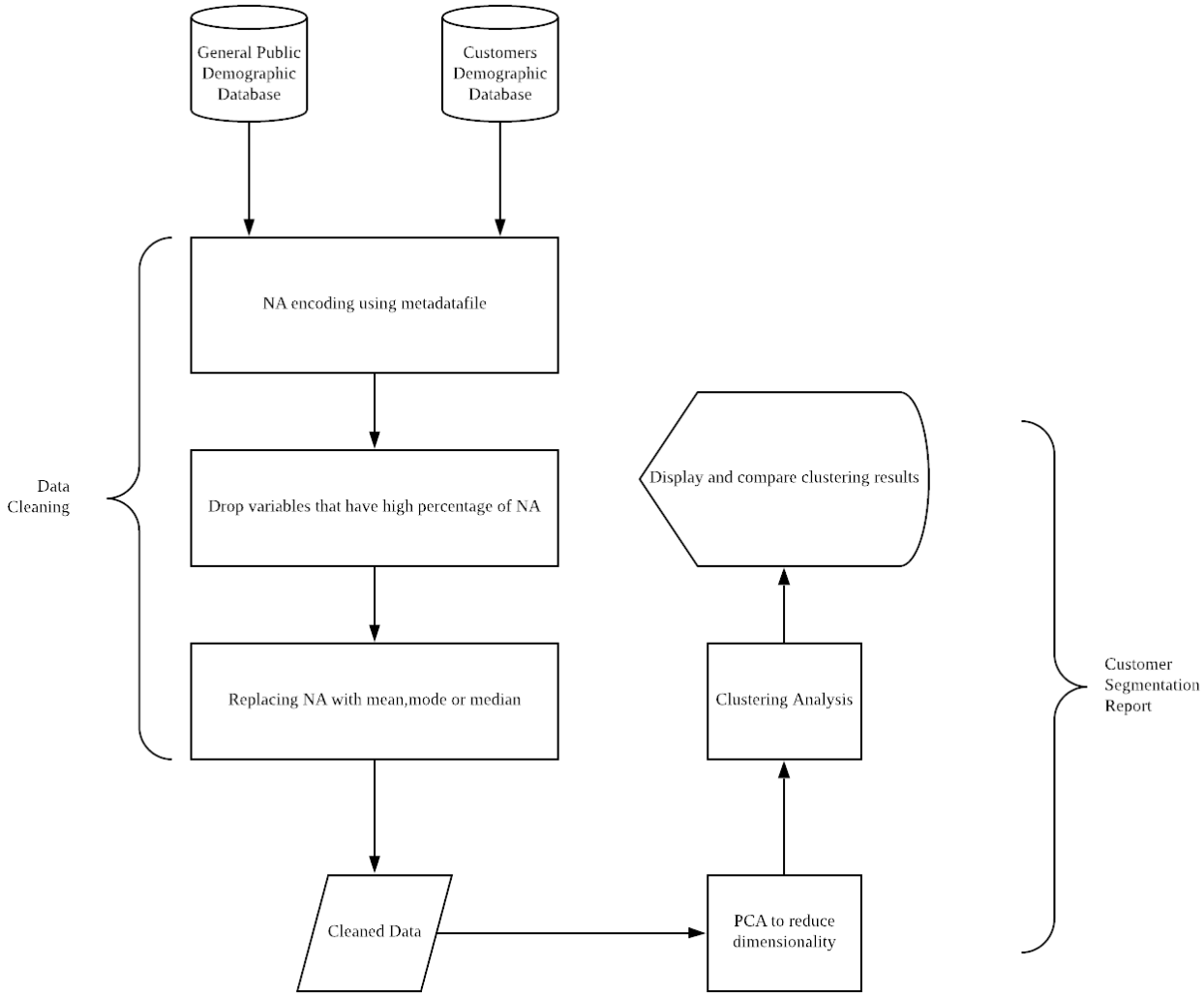
(approx. 1 page)

- Data Preprocessing

1. Encode NA values using the metadata excel file
2. Drop variables that have too many NA Values(30%)
3. Using PCA to reduce dimensionality and collinearity to generate our input file

- Customer Segmentation

1. Run Clustering Analysis on the dataset after PCA
2. Compare and match the clustering results from two datasets(general population and customers demographical datasets)
3. Explain the components and how they define customers' characteristics



- Predict Customers Conversion

1. Clean dataset in a similar fashion as the Customer Segmentation part.
2. Run classification algorithm to predict the conversion of target customers
3. Compare different models and select the best model
4. Generate submission file using test set and submit it to Kaggle platform

5. Use decision tree to repeat the processes for better understanding and visualization to business people