Machine Learning Engineer Nanodegree

Capstone Proposal

Jaweria Shabbir
November 12, 2020

Domain Background

Arvato-Bertelsmann financial solutions company is headquartered in Germany; it provides customer support to various companies in different domain. Arvato-Bertelsmann uses latest technology to help their customers in the best way possible. It focuses on using cutting edge technology combined with predictive analytics to aid businesses make important decisions based on large datasets and analysis.

Customer Segmentation has become a very popular practice with today's market. This is a very common targeted problem. It helps sellers, identify their customers and perform targeted advertisement for them. This usually helps save money, and gives greater return on the investment. This application can be altered and used in many other applications aside from selling products.

Customer Segmentation divides populations based on different interests and characteristics. It predicts future customers based on features and characteristics. Arvato-Bertelsmann financial solutions uses given data from business to help define customer segmentation for their business and make the best choices based on those predictions. This is crucial for business growth, and effective marketing strategy. Predictive analytics help build long term relationship between business and customers, the continuous relations is significant to a business success *(James M. Curran, Sajeev Varki, Deborah E. Rosen (2010)).*

Arvato-Bertelsmann financial solutions company works with a mail-order company to sell their organic products. The data provided by Arvato for this project is based on the customers of mail-order company and general population Germany.

Problem Statement

Based on the data given, what demographics in Germany are most likely to become customers for a mail-order company that sells organic products? Identify the population of people that are more likely to become future customers for the mail-order company among the general population?

Datasets and Inputs

The data has been provided by Arvato-Bertelsmann Financial Services. The data provided is based on general polulation of Germany and the customers of a mail-order sales company. The data provided can predict the future customers based on thorough analysis and predictive model. The first two files [Azdias and Customers] will be used to see similarities and differences between customers vs. population at large. The analysis from those to files will be used to make prediction on the other two files [MAILOUT] and predict which individuals are more likely to become customers. The 'customers' data includes three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP') which provide further information about the customers. The MAILOUT_TRAIN file contains an additional column "RESPONSE", that indicates whether or not the individual became a customer of the mail-order sales company.

The four separate datasets are listed below with brief descriptions:

- Udacity_AZDIAS_052018.csv: contains demographics data for the general population of Germany. Each row represents an individual and the columns represent features [891,211 rows and 366 features]
- Udacity_CUSTOMERS_052018.csv: contains demographics data for the existing customers of mail-order Company [191,652 rows (individuals) with 369 columns (features)].
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

There are two other excel sheets which provide additional attributes to the datasets.

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

Solution Statement

- Use the first two datasets [Udacity_AZDIAS_052018.csv and Udacity_CUSTOMERS_052018.csv] to compare and identify the customers from the general population.
    - The two datasets will be explored for its shape, attributes, NAN, non-numerical and missing values.
    - Non-numeric data will be replaced using one-hot-encoding, some missing values will be replaced with mean of the data. If there are large numbers (ex. more than 30%) of NAN or missing values in a column, the column will be dropped.

    - The two excel sheet files will be used as a reference to map and understand the datasets in details
    - Unsupervised learning algorithm, such as clustering will be applied along with PCA (to choose significant features among the several features present), and scaling to segment general population in clusters.
    - The section of population that are more likely to be customers from the general population will be identified using unsupervised learning algorithm (clustering).
- The second set of dataset will be used to predict what set of population will successfully become customers for mail-order sales company
    - Clean the data if needed (non-numerical values, NAN, missing data etc.), using similar methods as used for previous datasets
    - Build a prediction model using supervised learning algorithm
    - Some of the supervised learning technique used will be: logistic regression, support vector machines, K-nearest neighbor, and decision trees.
    - The model will be tested to determine the ideal algorithm for this task
    - The model will indicate the set of population that will be successfully acquired by the mail-order sales company.
    - This procedure might be repeated depending on the results. The goal is to achieve appropriate scores on accuracy, precision, and recall; roughly greater than 60% on at least two of the given scales. The runtime estimates will also be evaluated for each model.
- Alas the model will be submitted to the Kaggle completion and tested there.

Benchmark Model

I will be using the default setting of three models [logistic regression, Naïve Bayes, and SVM] without any model tuning as a benchmark model. Precision, recall, and runtime estimates for each model will be compared to evaluate the predicted solutions.

Evaluation Metrics

Precision, recall, F1 score and runtime estimates will be used as evaluation metric that quantifies the performance of both the benchmark model and the solution model. Since the data is imbalanced, confusion matrix will be ideal for evaluation. Precision and recall will be determined using the confusion matrix; score of greater than 60% will be accepted.

Project Design

Analysis and preprocessing of data:

As mentioned above, the data provided by Arvato-Bertelsmann Financial Solutions is not clean and requires some preprocessing before being used. Analyze the data for any NAN/missing values, and remove noise. Fill missing data with mean of the column as appropriate. Convert all the data into numerical values so algorithms can be performed with ease. Also keep in mind the data is imbalance while choosing different algorithms.

Unsupervised learning:

Scale the data and then perform unsupervised learning such as K clustering. Apply PCA to narrow down the features since there are lots of features available in the datasets. Scale the data as well if needed.

Supervised Learning:

Use supervised learning modeling on the labeled data. Experiment with several supervised learning models such as logistic regression, Naives Bayes classifier, decision trees and SVM. Use evaluation metrics such as precision, recall, runtime estimate and benchmark model to determine how well your model performs.

Hyperparameter Tuning:

Choose a model that performs the best and tune it further using hyperparameter tuning algorithm such as Grid Search.

Conclusion:

Answer the problem statement, and identify the appropriate section of the population that is most likely to be acquired by the mail-order sales company.

References:

- Arvato Financial Solutions. 2020. *Arvato Financial Solutions*. [online] Available at: <https://finance.arvato.com/en-us/> [Accessed 11 November 2020].

- James M. Curran, Sajeev Varki, Deborah E. Rosen. (2010) Loyalty and Its Antecedents: Are the Relationships Static?. *Journal of Relationship Marketing* 9:4, pages 179-199. https://www.tandfonline.com/doi/citedby/10.1080/15332660902991197?scroll=top&needAccess=true

- Talabis, Mark Ryan M., et al. “Analytics Defined.” *Information Security Analytics*, Syngress, 5 Dec. 2014, www.sciencedirect.com/science/article/pii/B9780128002070000010.