Machine Learning Nanodegree Program

Capstone Proposal: Customer Segmentation Report for Arvato Financial Services

Kanarovska Yuliia

September 26th, 2020

Domain Background

"Arvato is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include SCM solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation and data/analytics. Globally renowned companies from a wide variety of industries – from telecommunications providers and energy providers to banks and insurance companies, e-commerce, IT and Internet providers – rely on Arvato's portfolio of solutions. Arvato is wholly owned by Bertelsmann. [1]"

Arvato efficiently helps its customers with digital transformation, valuable insights, and analysis of the data as well as making better business decisions. "Customer-centric marketing is an approach to marketing that prioritizes customers' needs and interests in all decisions related to advertising, selling, and promoting products and services. [2]" Understanding correlations and customers' behavior from given data is key to successful customer-centric marketing.

Data analysis techniques and Machine Learning helps to uncover hidden patterns and effectively manipulate large volumes of data with minimum human intervention.

Problem Statement

The formulation of the problem is "Under provided demographic data of the German population and of current customers of the mail-order company, determine the description of the targeted customers' groups for the given company and verify which new individuals might be acquired as new customers and on which basis."

The proposed solution is divided into 2 subsections. The first step would be to use unsupervised machine learning techniques to segmentize the customers based on the intersection of the data for the current customers and the general population. Secondly, a supervised model will be used on the discovered before data.

Datasets and Inputs

All the data is provided by Bertelsmann Arvato Analytics and there are given four files for 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).

Additionally, there were 2 more files for describing attributes:

- **DIAS Attributes - Values 2017.xlsx**: Explains values encoding
- **DIAS Information Levels - Attributes 2017.xlsx**: Explains column names meanings

Each row in the demographic data files represents and describes a person as well as his or her environment, such as their household, building, and neighborhood. The general structure of the AZDIAS and CUSTOMERS data files is similar. MAILOUT...TEST and MAILOUT…TRAIN are provided for the development and testing of the supervised model.

Solution Statement

There are several steps to e performed towards the solution.

1. Data exploration and preprocessing
   The first few steps would be to explore data, its shape, and structure, to understand the data values. Afterward, some cleaning should be performed, empty and undefined values should be dropped, mixed types should be fixed. Later, the sequence of transformations that could be performed such as categorical values should be encoded and data, in general, will be scaled.
2. Features selection (segmentation)
   At this moment the work will be devoted to the appropriate features' selection. By now 366 features are representing each person, not all of them are required and important for modeling. Therefore, there will be used dimensionality reduction technique like Principal Component Analysis (PCA). Then, Elbow Curve will be used to determine the optimal number of clusters for the K-Means technique, so that, finally, the K-Means algorithm might be used.
3. Predictive modeling
   The last step will be devoted to building a supervised model.
   Proposed algorithms are:

- K-nearest neighbors
- Decision trees (C5.0/CART)
- XGBoost

However, at this point, it is hard to state without hesitation which model will be more suitable and give a better outcome.

Benchmark Model

The benchmark model for this project would be XGBoots since it is proved to be quite flexible and efficient.

Evaluation Metrics

Two different parts of the project should undergo different evaluations. For the dimensionality reduction algorithm PCA it is better to look at a data variance to decide how many top components to include.

For predictive modeling exists different approaches to evaluation. While regression models benefit from Root Mean Squared Error (RMSE) evaluation metric, for decision trees should be considered something else: it is better to implement the AUC-ROC curve and/or confusion matrix.

Project Design

The proposed architecture of the project should look as follows:

1. Data cleaning and visualization: this section is devoted to the exploration of the data for missing and/or improper values, identification of the outliers. Based on the revealed information, missing data should be dropped or filled if it is possible.
2. Features engineering: determining the most relevant features with the help of the unsupervised learning algorithms: PCA and K-Means algorithms. Afterward, inappropriate features should be eliminated.
3. Supervised model implementation: after a solution with feature engineering is settled, several above-mentioned supervised models for predictive analysis will be used.
4. Model tuning: after primary evaluation of the different algorithms' performance, further work should be preceded with the outstanding one. Therefore, particularly one should undergo hyperparameter tuning for improving the performance.
5. Evaluation and testing: finally, the best and tuned model should be used for predictions and Kaggle competition.

References

[1] Arvato-Bertelsmann, "Arvato", Bertelsmann. [Online].

Available: https://www.bertelsmann.com/divisions/arvato/#st-1 [Accessed Sept. 28, 2020]

[2] HelpScout, “How to Build a Winning Customer-Centric Marketing Strategy”, Sarah Chambers. [Online]. Available: https://www.helpscout.com/blog/customer-centric-marketing/ [Accessed Sept. 28, 2020]