Machine Learning Nanodegree Program

Capstone Proposal:

Identify Customer Segments for Arvato Financial Services

Seyedmeysam Hadigheh

July 15th, 2021

Domain Background

“Arvato is an internationally active services company that develops and implements innovative solutionsfor 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 problem we will be working on in this project is the following:

"How can the German mail-order company acquire new customers more efficiently, given the access to German demographics data?"

Essentially, given the demographics data of a single person, what can we do to predict, with sufficiently high/ significant accuracy, whether this person will be a new customer to the mail-order company? Out of all of these people with their associated demographics information (third dataset), can we predict with confidence how many of them could be future customers with high probabilities of becoming customers?

The problem can be quantified in the following terms: number of current/established customerclusters (customer segmentation unsupervised problem) and probability of being a new customer to the company (supervised problem).

Machine Learning techniques can be employed in the two main subsections of the project:

- Using unsupervised learning methods on the data of established customers and thegeneral population's demographics data, we can create customer segments.

- Using supervised learning methods on a third dataset, we can train a model to predict the probability of a person becoming a new customer (above a certain threshold, the model will assign the person to be a highly probable new customer), and use this modelfor future predictions.

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; 891211 persons (rows) x 366 features (columns).
- **Udacity_CUSTOMERS_052018.csv**: Demographics data for customers of a mail-ordercompany; 191 652 persons (rows) x 369 features (columns).
- **Udacity_MAILOUT_052018_TRAIN.csv**: Demographics data for individuals who weretargets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- **Udacity_MAILOUT_052018_TEST.csv**: Demographics data for individuals who were targetsof 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

Ultimately, Arvato Financial Solutions' goal is to enable their client company to gain insight from their current established customer base in order to better target the German population atlarge, by predicting in advance and with sufficient accuracy who would become a future customer.

To make that possible, after initial data exploration and cleaning, we will first employ unsupervised learning techniques to identify customer segments (*customer segmentation*): theseinclude applying **PCA** (Principal Component Analysis) for Dimensionality Reduction, and an algorithm such as **K-Means Clustering** to obtain the meaningful 'clusters' of customers.

Then, we will make use of supervised learning techniques for the second part of the project, which consists of predicting future potential clients from the German Population dataset, basedpartially on insight gained by customer segments. For this task, we will try different supervised

Benchmark Model

For the final step of this project, where we will apply supervised machine learning techniques to our binary classification problem (new customer 1 – not a new customer 0), an appropriate benchmark model to compare our model's performance could be a Logistic Regression Model.Thus, our benchmark model will be a standard Logistic Regression model with outcomes 1 = new customer, 0 = not a new customer.

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 beeliminated.
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. In Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Arvato#cite_note-3
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]
3. Udacity+Arvato: Identify Customer Segments. In Kaggle. Retrieved from: https:// www.kaggle.com/c/udacity-arvato-identify-customers
4. Classification: ROC Curve and AUC. In Google Developers. Retrieved from: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
5. PMLB: a large benchmark suite for machine learning evaluation and comparison. Retrieved from : https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0154-4