MACHINE LEARNING ENGINEER NANODEGREE

Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Beatriz Sousa

November 19th 2019

Domain Background

Bertelsmann found its origins as a publishing house in 1835 (Schuler, 2010), and through steady growth and development made its way to the software and hardware distribution market in the 80's (Computerwoche, 1983). By 1999 the company received its current name Arvato Bertelsmann (Neuer Name, neue Ziele, 1999) and over the next decade fully entered the domain of high-tech, information technology, and e-commerce services (Paperlein, 2012).

Arvato offers financial solutions in the form of diverse segments, from payment processing to risk management activities. It is in this domain that that this capstone project will be developed. Arvato is looking to use its available datasets to support a client (mail-order company selling organic products) in identifying the best data founded way to acquire new client base. To achieve this goal I will explore Arvato's existing datasets to identify attributes and demographic features that can help segment customers of interest for this particular client.

Customer centric marketing is a growing field that benefits greatly from accurate segmentation, with the help of machine learning hidden patterns can be found in volumes that could easily be missed without computational help, requiring very little maintenance or human intervention, leading to an improved experience from customer seekers and customers alike.

Problem Statement

The problem statement for this project is “How can a client – mail order company selling organic products – acquire new clients in a more efficient way?”.

The solution I propose for this problem is divided in 3 subproblems.

I will use an unsupervised learning approach to identify customer segments of value based on demographics data of existing customers versus general population data, and will follow-up on the discovered customer segments with a supervised learning approach using a dataset with demographics information for the target customers for the advertising campaign and predict which individuals would be more likely to convert to company customers.

Datasets and Inputs

All the datasets were provided by Arvato in the context of the Udacity Machine Learning Engineer Nanodegree, on the subject of Customer Acquisition / Targeted Advertising prediction models.

There are 4 datasets to be explored in 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).

And 2 metadata files associated with these datasets:

- 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

Which can help mapping the attributes to its particular type or missing value encoding.

Solution Statement

For a two stepped problem I propose a two stepped solution.

Since the first portion of the solution requires the usage of unsupervised learning methods I will make sure to select and encode any non-numerical features, followed by feature scaling to guarantee that the natural scale of the features does not affect their overall weight on the principal components, I will use PCA for dimensionality reduction and as a part of the data pre-processing for the prediction step I will implement KMeans as a form of partitioned based clustering (efficient and good performer for medium to large datasets which is our case).

Once the data is pre-processed and the customer segments are identified I will approach the supervised learning component of this project by testing which models, out of the considered habitual options for customer conversion prediction work best for these particular datasets, namely:

- Logistic Regression
- DecisionTreeRegressor (with an ensemble of RandomForestRegressor and GradientBoostingCLassifier)
- GridSearchCV

Since at this point this is just a proposal, and for now there is no way to predict how good of a fit this approach can be, I shall keep an open mind to tried different approaches that can reveal themselves to be more suited.

Benchmark Models

For this problem it is suggested to use Gradient Boosting Classifier based on consulted data sets of historical relevance on Kaggle relating to customer conversion and targeted marketing response (performances nearing 80%).

Evaluation Metrics

For the first part of the problem using unsupervised learning, explained variance ratio can be used when we are implementing PCA, as it accounts for the description of feature variance, allowing for the determination of more important features that stand out with more explained variance.

For the prediction portion of the project (supervised learning) precision can be used as a metric, as does accuracy and recall. Regression based models also benefit from using Mean Absolute Error and Mean Squared Error.

The final decision on which evaluation metrics to use highly depend on the information obtained through exploratory data analysis. In the case of KMeans (unsupervised learning) there is no ground truth to evaluate a model's performance, there is no single right answer to evaluation since for instance the number of k clusters is a hyperparameter input. And for the unsupervised learning portion of the problem solution the decision will depend on data balance of classes.

Accuracy	Precision	Recall	F1	ROC

TP + TN/
TP + TN + FP + FN
	TP/
	TP + FP
		TP/
		TP + FN
			=2 * ((precision*recall)/(precision +recall))	x-axis-inverted specificity
				FPR = FP/FP + TN
				
				y-axis-describes how good are the model predictions
				TPR = TP/TP + FN

Problematic in imbalanced datasets, we can have high accuracy without making useful predictions	Is about exactness, classifying only one instance correctly yields 100% precision, but a very low recall, it tells us how well the system identifies samples from a given class.	Is about completeness, classifying all instances as positive yields 100% recall, but a very low precision, it tells how well the system does and identify all the samples from a given class.	Harmonic mean of precision and recall, which eases comparison of different systems, and problems with many classes.	Appropriate for observations balanced between classes
Ideal If class labels are uniformly distributed	Ideal for imbalanced classes	Ideal for imbalanced classes	Ideal for imbalanced classes	Ideal when we are given the probability of prediction for each class which means we have to calibrate a threshold to belong to a particular class

Project Design

1. **Data Cleanup:** most of the data that is received raw requires an extensive step of cleanup for improper data entries and missing values. For each feature I will examine the percentage of missing values, identify outliers and the type of feature (binary, categorical, continuous, etc). Missing data will filled or dropped on a case by case approach.
2. **Data Visualization:** Allows for a birds-eye view of the data and early detection of particular patterns, namely, correlations between predictors and target variables, ranges and scales. For this we can take advantage of the matplotlib library and seaborn as well as pandas for preliminary summary statistics.

3. **Feature Engineering:** Implement PCA, find most relevant features, eliminate features of low importance for optimal model training further in the project. Confusion matrixes can help to further identify features that should be eliminated due to dependency/high intra-correlation.
4. **Model Selection:** Experiment with the before-mentioned algorithms to find the ideally suited for this problem, namely KMeans for the unsupervised learning portion and Logistic Regression, DecisionTreeRegressor (with an ensemble of RandomForestRegressor and GradientBoostingCLassifier), GridSearchCV for the supervised learning portion in which we are to predict costumer acquisition through targeted campaigns.
5. **Model Tuning:** Once we find the model that best suits our data, adjust model parameters within a range that allows for increased performance without overfitting, increase awareness for possible data leakage.
6. **Test and Predict:** use the previously proposed metrics, explained in the table present in the section for evaluation metrics as an indicator of success in our predictions.

Works Cited

Computerwoche. (1983, October 21). Bertelsmann vertreibt Rechner von TI. *Computerwoche*.

Neuer Name, neue Ziele. (1999, June 9). *Darmstädter Echo*.

Paperlein, J. (2012, April 5). E-Commerce ist weltweit ein Thema. *Horizont*, p. 14.

Schuler, T. (2010, June 18). Erst Drucker, dann Verleger. *Berliner Zeitung*, p. 30.