Machine Learning Engineer Nanodegree

Capstone Project Proposal

Customer Segmentation – Arvato Financial Solutions

Ambresh Patil

January 5th 2020

Domain Background

Bertelsmann is a media company based in Germany founded in 1835. It is one of the world's largest media companies and also active in the service sector and education. Its principal divisions include the RTL Group, Penguin Random House, Gruner + Jahr, BMG, Arvato, the Bertelsmann Printing Group, the Bertelsmann Education Group and Bertelsmann Investments.

Arvato Financial Solutions offers Consulting Services and one of its clients which deals with organic products is looking to expand its client base in Germany. The Capstone Project will be developed in this domain. The client's strategy involves sending free product samples to increase the customer base out of potential customer segment. It is imperative for the client to send the product samples to the customer segment of interest to optimize the efforts and costs involved.

This Capstone Project aims at leveraging the datasets with Arvato and the client's customer dataset in anayzing the attributes and demographic feature to create customer segmentation. Machine Learning puts us at great advantage in recognizing hidden patterns to improve customer segmentation.

Problem Statement

How can the efficiency of customer acquisition be increased of a client who is into selling of organic product and intends to acquire new customer by sending sample products to customer segment of interest

My proposal consists of 3 sections

1. An unsupervised learning approach to identify potential customer segment by analyzing attributes of existing customer against the general population data.
2. Building a Supervised learning model based on the former analysis, which predicts if an individual will respond to the campaign or no. The dataset used for prediction will targets of client's mail order campaign.
3. The chosen model will be used to make prediction on campaign data as a part of Kaggle competition.

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:

1. Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns)
2. Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns)
3. Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
4. 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:

1. DIAS Information Levels — Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category
2. 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 Statements

Customer Segmentation using Unsupervised Learning Approach

- As a first step of feature engineering, I will be encoding any categorical features into numerical values and perform feature scaling to have continuous features in identical range. For dimensionality reduction, I will be using PCA and for prediction I will be using K-Means clustering.

Prediction using Supervised Learning approach.

- Once we have customer segments of interest, I will be finding out the best working model among the below models for customer conversion prediction on the given datasets
  1. Logistic Regression
  2. Grid Search
  3. Decision Tree Regression

Apart from the above approach, I will also keep my options open to check if any other approach might yield better results.

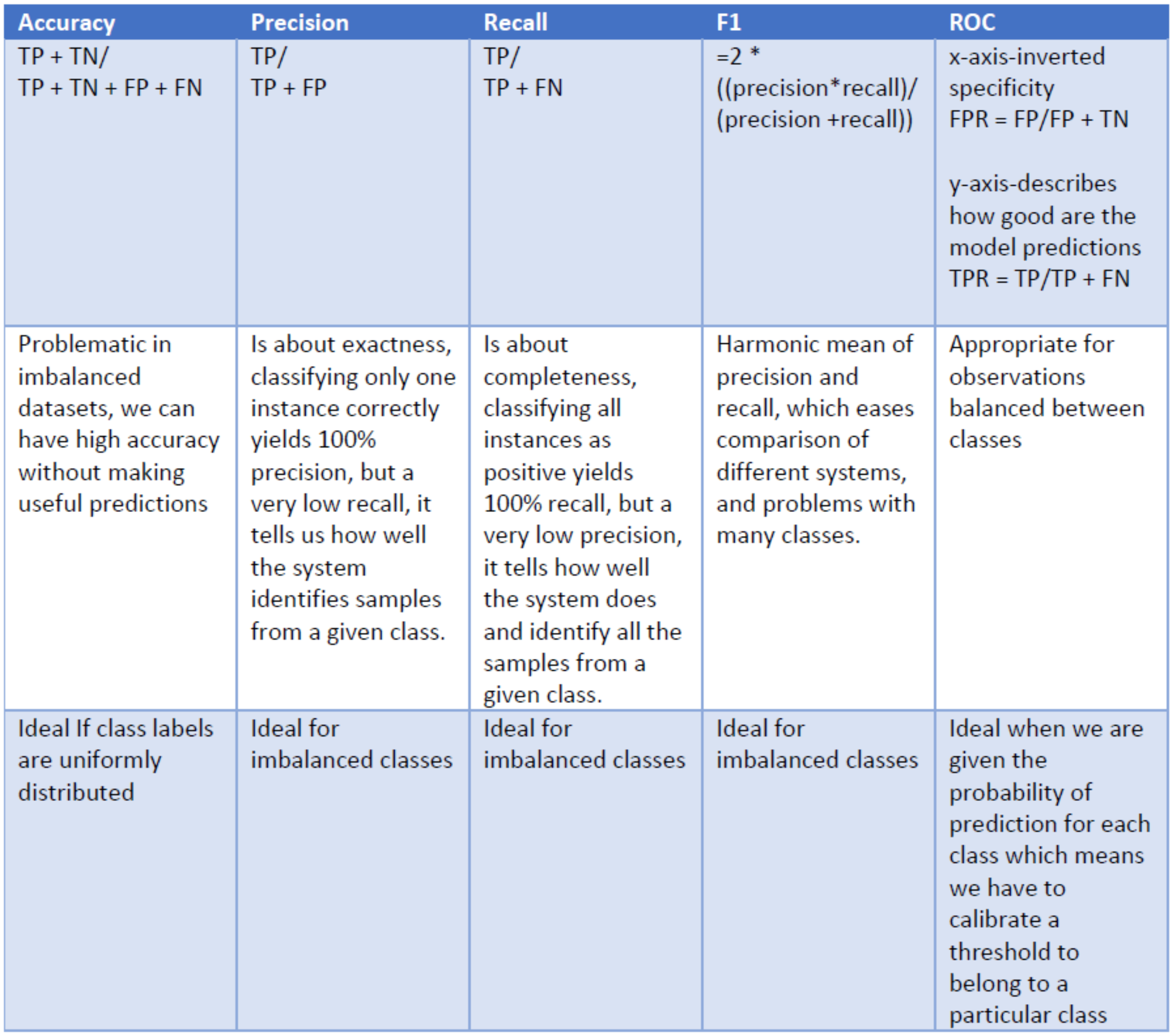
Benchmark Models

Researching on Kaggle for targeted marketing and customer conversion, Gradient Boosting Classifier looks like a good choice where performances are upto 80%.

Evaluation Metrics

- Explained Variance can be used as evaluation metrics during PCA implementation as it considers feature variance, helping us determine important features to be considered for dimensionality reduction.
- Precision, Recall, Accuracy, Mean Absolute error, Mean Squared Error can be used as evaluation metrics in supervised learning part for prediction.

Exploratory data analysis plays a vital role in determining evaluation metrics to be used finally. For instance in K-Means we cannot really have a single right evaluation of model's performance. For example the number of K is provided as a hyperparameter. For the unsupervised learning, the decision depends 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:** The data received is raw typically and requires cleanup for improper data entries and missing values. Missing values percentage will be found out for each feature, type of feature (binary, categorical, continuous) and outliers are identified. Missing data will be dropped or filled on a case by case approach.
2. **Data Visualization:** Helps us in detection of specific patterns like correlations between predictors and target variables, scales and ranges. Seaborn, Matplotlib and pandas will be used for preliminary summary statistics.
3. **Feature Engineering:** PCA will be implemented to find most relevant features and drop less important features for optimal model training. Confusion matrixes can help to further identify features that should be eliminated due to dependency/high intra-correlation.
4. **Model Selection:** Research and experiment with algorithms to get the best suited for this problem, namely KMeans for the unsupervised learning approach and Logistic Regression, Decision Tree Regression, Grid Search for the supervised learning approach to predict costumer acquisition through targeted marketing.
5. **Model Tuning:** After the selection of right model, to improve performance without overfitting we will adjust model parameters within a range and increase awareness for possible data leakage.
6. **Test and Predict:** using proposed metrics, explained in the table as an indicator of success in our predictions.