MACHINE LEARNING ENGINEER NANODEGREE

CAPSTONE PROJECT PROPOSAL

Arvato-Bertelsmann Customer Acquisition

UDACITY

Ahmad Shapiro
ahmad.shapiro@alexu.edu.eg

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 ecommerce 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 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 a 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?"

All of the projects I've done , I've been following a great strategy that always helped me , "Divide and Conquer" , so according to this strategy we will divide our project into two phases.

Phase 1:

We will use the population data with an unsupervised learning approach to identify the population different clusters and then apply them to the customers data set to see if we can detect any sort of pattern.

Phase 2:

After learning about customer's clusters we will be switching into 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 Input

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**: 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.

Solution Statement

1. Data Wrangling :-

Data will be cleaned and columns with proportion of missing values more than 50% will be dropped , the remaining missing values will be imputed by iterative imputer of sklearn library using BayesianRegressor with a mean strategy for numerical variables and mode (most-frequent) strategy for Categorical Variables.

2. Scaling :-

Data will be scaled and saved for later use in the unsupervised learning phase.

3. Unsupervised Learning Phase :

3.1. **Principal Component Analysis** : we will compute the 95% variance explaining principal components to be used later in the clustering process.

3.2. **K-Means Clustering** : After the PCA step we will cluster the data into the best number of clusters based on some metrics like (elbow-method) , after that we will see if the cluster makes a good separation between the customers and population data set, if so. We will explain the most high clusters in the customer's data set and the features contributing to their main components (highest centroids)

4. Supervised Learning Phase :-

From a previous experience I noticed that using categorical data principal components to do a supervised learning task isn't efficient at all and it always harms the model (according to many previous tasks).

Categorical features will be one-hot encoded and then fed into the following models to determine which of them will be our selected models to tune later.

4.1. XGBoost Algorithm

4.2. Catboost Algorithm

4.3. StructuredDataClassifier of Auto-Keras library

Benchmark Models

We fitted a Logistic Regression Model on the data and it score an ROC AUC of 0.66 , so this will be our benchmark to evaluate the rest of models.

```
X_train,X_test,y_train,y_test=prepare_training(k_means=False,pca_=False)
```

```
log_reg=LogisticRegression(max_iter=10000)
log_reg.fit(X_train,y_train)
```

```
LogisticRegression(max_iter=10000)
```

```
y_pred=log_reg.predict_proba(X_test)[:,-1]
roc_auc_score(y_true=y_test,y_score=y_pred)
```

```
0.6596403822938968
```

So our benchmark is 0.66 AUC_ROC

Evaluation Metrics

Phase 1 (Unsupervised Learning) :-

1. PCA variance ratio for the dimensionality reduction .
2. Sum of squared error for the K-Means Clustering algorithm to determine the best k according to elbow method.

Phase 2 (Supervised Learning) :-

1. Area under ROC Curve :

   Before explaining this metric , we should first explain its components.

   As we are doing a binary classification task. We have four possibilities of any prediction.

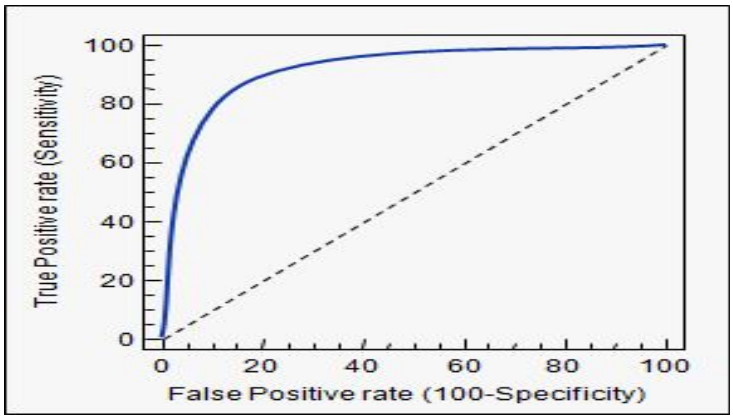
   1. Predicted : Positive , Real : Positive which is TRUE Positive.
   2. Predicted : Positive , Real : Negative which is False Positive.
   3. Predicted : Negative , Real : Negative which is TRUE Negative.
   4. Predicted : Negative , Real : Positive which is False Negative.

After Calculating the above numbers , we are left with two other metrics :-

1. Sensitivity : Which is true positive rate = true positives / true positives + false negatives.
2. Specificity : Which is true negatives rate = true negatives / true negatives + false positives.
3. False Positive Rate : 1 - Specificity

The ROC curve is simply a plot of Sensitivity against the false positive rate for different cutoffs variables, each point on the curve is a pair of both specificity and sensitivity

corresponding to a threshold , a good result will have 100% in both respictly which results in area of 1 under the curve, but the worse may have 0.5 area under the curve as illustrated below



Project Design

Pre-Phase :

1. **Data Wrangling :-**
    1.1. Encoding missing values with NaN and checking for outliers according to the encoding levels in the **DIAS Attributes — Values 2017.xlsx** sheet.
    1.2. Dropping Columns with percentage of missing values above 50%
    1.3. Imputing the rest of the missing values after dropping columns using sklearn's iterative imputation.
    1.4. One-hot encode string categorical columns and scale the whole data set to be ready for the next phase.
    1.5. Saving a clean version of the population , customers, train and test datasets.

Phase-1 (Unsupervised Learning) :

1. **Dimensionality Reduction (PCA) :**
    1.1. Calculating Principal Components explaining 95% of the variability in our data fitted on the population dataset.

1.2. Transforming the customers and population datasets to their principal components.

1.3. Scaling the datasets again to be ready for the next phase.

1.4. Saving the customers and population dataset to a clean PCA ".csv" version.

2. Clustering K-means :-

2.1. We will test multiple values of hyperparameter "K" to choose the best K that clusters the data according to the "elbow-method" , fitted on the population data.

2.2. Predicting clusters for both population and customers then comparing both with visualizations to know which clusters are more among customers.

2.3. After knowing customers clusters we will try to interpret them using visualization and highest value centroids among those clusters and feature's weights for their PCA components.

Phase-2 (Supervised Learning) :

1. **Model Selection :-**Testing the models we mentioned above and choosing the best two models to baby set (tuning their hyperparameters) them.

2. **Model Tuning :-**Tuning the hyper parameter using BayesearchCV from skopt library (the best based on previous experience)

3. **Prediction :-** Choosing the best model among the tuned models then predicting the test data set and submitting to kaggle .