MACHINE LEARNING ENGINEER NANODEGREE

Capstone Proposal

Customer Segmentation – Arvato Financial Services

Santosh Kannan

1. Domain Background

Arvato is a services company that develops and provides Supply Chain Management (SCM) solutions, Financial Services and IT services to clients all over the world. Key highlights include development of innovative solutions with focus on automation and data analytics. List of clients for Arvato range from insurance companies, e-commerce, energy providers to internet providers [1]. Arvato is a venture of the Bertelsmann group, which is giant in the media and education industry.

The business model for Arvato revolves around helping clients with data analytics to assist with key business decisions. This includes uncovering hidden patterns, identifying customer behaviour from the raw data. This strategy for customer centric clients utilises various aspects of Data Science and Machine Learning.

In this project, Arvato is assisting a Mail-Order company in Germany, which sells organic products, to understand its customers. The ask is to identify the possible future customers in a deterministic manner. On the basis of existing customer data across Germany, I propose a Machine Learning model which can identify probable customers.

2. Problem Statement

On the basis of existing demographics data for customers of a mail-order company in Germany, analyse and identify individuals who are most likely to convert into customers for the company.

The proposed solution will utilise an unsupervised model for the purpose of identifying customer segments in the dataset. This will be followed up by a supervised model which uses demographic information of target customers for the advertising campaign and predicts the individuals who are likely to convert into customers.

3. Datasets and Inputs:

The datasets are provided by Arvato for Udacity Machine learning Nanodegree program's project. The dataset consists of demographic data of customers in Germany. This comprises of 4 datasets and 2 metadata files associated with them.

Data File	Description
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).
DIAS Information Levels - Attributes 2017.xlsx	Top-level list of attributes and descriptions, organized by informational category
DIAS Attributes - Values 2017.xlsx	Detailed mapping of data values for each feature in alphabetical order.

4. Proposed Solution

This problem statement can be tackled in phases:

Phase 1: Analysing the demographic data with unsupervised learning algorithms. This will help us identify classes in the data and features which are critical for customers of the company.The primary motive is to classify the different segments of customers in the provided dataset and reconcile it against the general population dataset.

- **Step 1**: As in any data science problem, the first step is to carry out data exploration. This includes identification of outliers in the data and resolving them. This is followed by normalizing the data.
- **Step 2**: Since the dataset contains a high number of features, identifying the minimum number of features required to represent the dataset is essential. This task can be achieved using techniques such as Principal Component Analysis.
- **Step 3**: Finally, the customers need to be classified using an unsupervised learning algorithm. For this specific problem, we'll use the K-Means clustering algorithm.

Phase 2: Applying a supervised learning model to predict whether the individual is likely to become a customer. This is achieved as below

- **Step 1**: Data pre-processing steps are carried out and the data is split into train and validation sets
- **Step 2**: A model will be trained and evaluated against the validation dataset
- **Step 3**: The trained model will be used on the provided test data to make predictions

Various supervised learning models can be used for predicting the future customers; however, we'll explore the below algorithms:

- Logistic Regression – Simple Binary Classification algorithm
- Decision Tree Classifier – Rule based algorithm for classification
- XG Boost Classifier – Another Decision tree algorithm

5. Benchmark Model

The benchmark model for this case is the simple Logistic Regression model due to its ease at training and efficiency. Although this model would not be the most accurate, it still would perform well, which can be used as a comparison point for other algorithms.

6. Evaluation Metrics

As mentioned earlier, the project is tackled in two phases:

Phase 1: Customer classification with unsupervised algorithms

This phase deals with reducing the feature space of the dataset with Principal Component Analysis technique. This can be evaluated using the explained variance of each feature, meaning, the minimum number of features explaining a high variance in the dataset can be chosen. For K-Means algorithms it is essential to choose a good value for K. Thus, evaluating the squared error will be a guideline for choosing the appropriate value for K.

Phase 2: Customer identification with supervised algorithms

After the model is trained and evaluated using the validation set, the model is applied on the test set. A couple of metrics can be used in this case to evaluate the model:

- Accuracy
- Confusion Matric, Recall, Precision

These metrics will help us identify how the model is able to generalize to a new dataset. We can thus make changes to the model if required. Using the confusion matrix, we can gauge the inclination of the model using the False Positives and False Negatives.

7. Project Design

The roadmap for model development can be summarized as below:

- **Step 1: Data Profiling**
  The data will be checked for any outliers such as missing values, incorrect values etc. The incorrect values will be validated and fixed according to the directions provided. The missing values will be quantified, and a decision will be made on extent of data to be discarded for further processing.

- **Step 2: Feature Extraction**
  This step involves understanding the variance in the data and identification of features which explain maximum features in the dataset. Popular feature extraction technique called as ‘Principal Component Analysis’ will be used to achieve this task.

- **Step 3: Training the Model**
  Once the features are extracted, it is essential to identify the customer segments in the provided dataset. Since the dataset is unlabelled, we use an unsupervised algorithm to uncover any patterns in the data. Post this step, we’ll use different supervised models on the training set to make predictions. The evaluation metrics mentioned in the previous section would be used to evaluate the best performing model amongst them.

- **Step 4: Tuning the Model**
  Once the best performing model is identified, it is essential to tune the model and improve its performance. Different hyperparameter tuning methods can be utilised to achieve this task

- **Step 5: Final Predictions**
  Finally, the best model will be used to make predictions. These predictions will be submitted on the Kaggle competition page.

References

[1] Arvato–Bertelsmann,“Arvato”,Bertelsmann,[Online], “https://www.bertelsmann.com/divisions/arvato/#st-1”