Machine Learning Engineer Nanodegree

Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Naga Sai Pranay Modukuru

9th April 2020

Domain Background

Arvato is a services company that provides financial services, Information Technology (IT) services and Supply Chain Management (SCM) solutions for business customers on a global scale. It develops and implements innovative solutions with a focus on automation and data analytics. Arvato's customers come from a wide range of industries such as insurance companies, e-commerce, energy providers, IT and Internet providers [1]. Also, Arvato is wholly owned by Bertelsmann, which is a media, services and education company [2].

Arvato is helping its customers get valuable insights from data in order to make business decisions. Customer centric marketing is one of the growing fields. Identifying hidden patterns and customer behaviour from the data is providing valuable insights for the companies operating in customer centric marketing. Data Science and Machine Learning are immensely used now a days to fulfil business goals and to satisfy customers.

In this project, Arvato is helping a Mail-order company, which sells organic products in Germany, to understand its customers segments in order to identify next probable customers. The existing customer data and the demographic data of population in Germany are to be studied to understand different customer segments, and then building a system to make predictions on whether a person will be a customer or not based on the demographic data.

Problem Statement

The problem statement can be formulated as, "Given the demographic data of a person, how can a mail order company acquire new customers in an efficient way".

Proposed solution:

First, the demographic data of the general population and the customers will be studied with the help of unsupervised learning algorithms. The goal in this step is to identify segments in general population and segments in the existing customers, and then discovering what demographic features correspond to a person being a customer for the mail-order company.

Second, a supervised learning algorithm will be used to make predictions on whether a person is a probable customer or not, based on the demographic data.

Dataset and Inputs

There are four data files associated with 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).

Additionally, 2 metadata files have been provided to give attribute information:

- 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

All the files associated with the project have been provided by Arvato in the context of Machine Learning Nanodegree Program for analysis and customer segmentation purposes. The four csv files are the demographic data files, in which each row represents demographics of a single person. Each row also includes additional information about their household, building and neighbourhood in addition to their demographics. Customers data has three additional columns indicating their specifics with regard to the mail order company. The Train and Test data have been provided to evaluate supervised learning algorithms.

Solution Statement

In the first part of the project, the task is to identify any customer segments present in the provided dataset and Match these segments with the segments of population present in the general population dataset.

- In the first step, the dataset will be explored to examine if there are any missing values or mis recorded values in the data and fix them. Also, any categorical features need to be re encoded into numerical features with the help of Label encoders. Finally, the data will be scaled in order to ensure no single feature will have higher weights the later steps.
- The second step is to identify the minimum number of features that would be sufficient to explain the dataset. Since there are 366 features that represent a single person and not all the features will be important in forming the segments. A dimensionality reduction technique like Principal Component Analysis (PCA) can be used here to identify minimum number of features which explain the variation in the dataset.
- The third step is to segment the general population and the customers into different segments based on the selected features with the help of unsupervised learning algorithm. K-means clustering is a good choice for this step as this algorithm tries to assign each data point to a cluster based on the distance from a cluster centre.

In the second part of the project, the task is to predict whether the mail order company can acquire a customer.

- In the first step, the data is pre-processed (the first two steps of the first part will be done again on train and test data).
- In the second step a supervised learning algorithm will be trained and evaluated on the pre-processed training data.
- In the last step, the trained model will be used to make predictions on the test data provided.
- Proposed algorithms for supervised learning.
  - Logistic Regression – A simple binary classification algorithm
  - Decision Tree Classifier – A tree-based algorithm which uses rule-based approach for classification.
  - Random Forest Classifier and XG Boost Classifier can also be used since they are derived from the decision trees algorithm

A grid search algorithm can also be used in order to select the best hyperparameters for the selected algorithm.

Benchmark Model

A benchmark model in this case would be a Logistic Regression model since it is easy to train and test within less amount of time. The performance of this model will be considered as a baseline for further steps, where different algorithms can be used to compare the performance with this benchmark to decide whether to proceed with an algorithm or not.

Evaluation Metrics

The project is divided into two parts

Customer Segmentation using unsupervised learning algorithms

This part of the project uses a dimensionality reduction technique PCA to reduce the number of dimensions. The explained variance ratio of each feature could be the reference in selecting the number of dimensions for the later steps. The minimum number of dimensions explaining as much variation as possible in the dataset can be chosen in this step.

Also, in case of segmenting the customers into different clusters, an unsupervised learning algorithm like K-Means Clustering is proposed. Also, in this case the number of clusters will be a hyper parameter and it will be selected based on the squared error i.e. the distance between all the clusters.

Customer Acquisition using supervised learning algorithms

In the second part of the project, the task is to predict whether or not the mail-order company should approach a customer. Here the given training data will be split into train and evaluation sets, the model will be trained on the training split and will be evaluated on the evaluation split. In this step evaluation metrics for classification can be used.

The evaluation metrics for classification include:

- Accuracy
- Confusion Matrix – F1 score, Recall, Precision
- Area Under the Receiver Operating Curve (AUROC)

The decision on these metrics will be based on the problem at hand. If the target labels are highly imbalanced then accuracy would be a bad choice to evaluate the model. Then having a look at the confusion matrix will give a better idea about the predictions. Also, we can tune the models to target for maximum precision or recall or F1 score based on the problem statement i.e. whether we can afford to have false positives or true negatives.

Here, as per the leader board on the Kaggle AUROC has been the metric on which the predictions on the test set are evaluated. The AUROC gives an idea about overall performance of the model, where the curve is created by plotting True positive rate and False positive rate under different threshold settings. A good performing model will have an AUROC of 1. So higher the AUROC better the performance of the model.

An appropriate evaluation metric will be chosen after analysing the data and observing the balance between different classes in the provided data.

Project Design

A brief explanation of the proposed steps of the project

1. Data Cleaning and Visualisation: The data needs to be checked for any missing values and any mis recorded values. All the mis recorded values will be verified and will be fixed based on the information provided in the metadata files. An analysis on how many missing values are there per feature will be performed to decided on which features to neglect. A visualization analysis will be don on the data to understand any noticeable patterns in the data.
2. Feature Engineering: Understanding explained variance of features in the dataset and determining the required number of features that can amount for maximum variance in the dataset using a dimensionality reduction technique like PCA. Determining correlations between features will also help in identifying redundant features.
3. Modelling: First step is to identify the customer segments using unsupervised learning algorithms. A K-Means Clustering algorithm will be used to segment the data into desired number of clusters. In the second step, different supervised algorithms will be trained and evaluated in the context of predicting whether a person will be our next customer or not. Algorithms like Logistic Regression, Decision Tree, Random Forests and Gradient Boosted Trees will be used to make predictions and will be evaluated. The previously proposed evaluation metrics will be used to determine the best model in this step.
4. Model Tuning: After evaluating different algorithm's performance on the evaluation data. The algorithm which has a good score will be selected and tuned to improve the performance. A hyper parameter tuning algorithm like Grid Search will be used to determine the best set of hyper parameters.
5. Predictions on Test data: Finally, the best model will be used to make predictions on the test data and the predictions will be submitted on the Kaggle competition page.

References

[1] Arvato-Bertelsmann, "Arvato," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/divisions/arvato/#st-1. [Accessed April 2020].

[2] Bertelsmann, "Company," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/company/. [Accessed April 2020].