Udacity Machine Learning Engineer Nano Degree

Capstone Proposal Customer Segmentation – Arvato Financial Solutions

Edward Amankwah

October 13, 2020

Domain Background

Bertelsmann is a leading publishing house since 1835 (Wikipedia, 2020), which over the years has expanded from media, services and education (Bertelsmann, 2020) into software and hardware distribution especially in the 1980s (Computerwoche, 1983). The company obtained its current name, Arvato Bertelsmann in 1999 and migrated more into high tech, information technology, and e-commerce services in 2012 (Wikipedia, 2020). Some of the services they offer include supply chain, financial solutions, e-commerce, insurance companies, energy providers, IT consulting, Information technology and Internet providers (Arvato-Bertelsmann, 2020).

Arvato is using artificial intelligence to derive data driven insights for their customers to implement business decisions. Customer services such as financial solutions, portfolio risk management, payment processing and customer targeting for marketing campaigns benefit immensely from artificial intelligence derived solutions. It is in this field that this project will be developed, as machine learning and data science practices are immensely used to attain business goals and fulfil customer satisfaction. Since customer centric marketing is a key component for modern companies to thrive, this project will focus on using available data from Arvato to help a Mail-order company selling organic products in Germany, to better understand its customer purchasing decisions. The available customers will be segmented to help identify probable future customers that should be targeted for marketing campaigns. The project will explore the existing customer and demographic data of population in Germany is crucial to understand the behavior of the different customer segments, and then build a trained model to predict who will be a potential future customer based on the demographic data.

Problem Statement

The problem statement for this capstone project is “How can a mail-order client selling organic products efficiently attract new customers, given the demographic information of a person?” My approach to this challenge includes firstly, I will compare the customer segments derived from both the demographic data of the current customers and the general population data by performing an unsupervised machine learning modelling. Secondly, I will conduct a supervised learning modelling based on the resultant customer segments using a dataset with demographic features that corresponds to a client being a customer for marketing campaign. Lastly, I will predict probable individuals that are good candidates to be converted to the company customers.

Dataset and Input

The following datasets were provided by Arvato for the sole purpose of the Udacity Machine Engineer Nanodegree, capstone project topic on customer segmentation and targeted marketing campaign.

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).

Moreover, two metadata files have been provided to give 2017 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.

The four demographic comma separated values (CSV) data files contain rows that represent the demographics of a single individual. Additionally, each row contains extra information about their household, building and neighborhood. The Customers data includes three extra columns containing specific mail order company for each individual. Also, train and test data have been provided for the evaluation of the supervised learning models.

Solution Statement

I will provide two solutions to solve the two major problems; unsupervised learning for customer segmentations and supervised learning for binary classification presented above.

The following steps will be taken in the unsupervised modeling task to match the segments in the provided dataset with the segments in the provided general population dataset:

- The data will be cleaned and preprocessed to examine the statistics of missing values and miss-recorded values so that they can be fixed. Since machines work with numbers, categorical features will be encoded into numerical features by using label encoders. The data will be scaled to ensure every feature is equally accounted for in terms of their weights contributions to the principal components.
- Each individual in the dataset is described by 366 features, which presents a challenge in terms of model dimensionality as well as speed during the modelling training. As part of the preprocessing steps, a linear principal component analysis (PCA) method will therefore be used to identify the minimum number of features that can explain the most variations in the data set.
- Based on the selected features, the general population and the current customers will be segmented by using the unsupervised K-Means learning algorithm. The elbow method will be used to determine the number of cluster centers to which each data point will be assigned based on the distance from a cluster center.

The second part of the project involved using a supervised algorithm to predict which potential customers can the mail order company convert.

- The first two steps in the first part will be repeated to preprocess the train and test datasets.

- Next, a number of supervised learning algorithms will be trained and evaluated on the pre-processed training dataset.
- Balancing the response data techniques as well as grid search and automated hyperparameter techniques will be used to select the best hyperparameters for the best-chosen algorithm.
- Lastly, the best trained model will be used to make predictions on the provided test data.
- The following algorithms may be used for the supervised learning.
    - Logistic Regression
    - Decision Tree Classifier
    - Support vector machines
    - Adaboost
    - Random Forest Classifier and
    - XGBoost Classifier

Benchmark Model

I will propose a Logistic Regression model with default parameters and without tuning as my baseline model because it can be implemented easily, trained and tested with less amount of computer resources and time. Other models will be compared to this baseline to determine if they are better or worse.

Evaluation Metrics

1) **Customer Segmentation using unsupervised learning algorithm**
   The PCA technique will be used to reduce the number of dimensions. The explained variance ratio will be used to determine the important features with the most variance that can be retained for the modelling exercise. In deciding the number of clusters to be used for the proposed K-Means algorithm, the distance metric between data points and nearest cluster center will be based on squared errors.
2) **Customer conversion using supervised learning algorithms**
   The supervised algorithms will predict whether the mail-order company should target a particular customer or not for marketing campaigns. The provided data set will be split into train and evaluation datasets for model training and evaluation, respectively. The evaluation metrics for the binary classification will include Accuracy, classification report and Area under the receiver operating curve (AUROC).

   The best evaluation metric to be used will depend on the balance between the binary response variable class and the business problem at hand, after all exploratory data analysis have been completed.

Project Design

1. **Data Cleaning**: real world datasets are raw and require extensive cleaning process which includes verifying and fixing missing data values. An analysis of the missing dataset will

result in dropping or fixing datasets based on the information provided in the meta dataset. Outliers in the feature datasets will be identified and non-numeric features will be pre-processed into numerical values.

2. **Data Visualization**: the feature dataset will be further exploited to determine, through visualizations and summary statistics, patterns, correlations and data ranges between the independent feature predictors and the dependent target variable. Standard libraries such as matplotlib, seaborn and pandas will be used throughout the analysis.
3. **Feature Engineering**: to reduce the number of dimensions in the feature sets, PCA techniques will be implemented to capture the relevant features that can explain the greater variability in the feature sets. The number of features to be used in the modelling will eventually be reduced to speed up the training process and all redundant features will be dropped.
4. **Model Selection**: to perform the customer segmentation step, a K-Means clustering algorithm will be implemented to obtain the desire number of clusters. This unsupervised learning step will lead to the application of a suitable supervised learning algorithm to predict the potential customers that can be targeted for marketing campaigns. The algorithms that were mentioned in the solution statement section will be implemented to make predictions and evaluated based on the metrics itemized in the evaluation metrics section.
5. **Model Optimization**: before predicting on the test dataset, the model that best fit the training and evaluation datasets will be tuned by adjusting the hyperparameters within a suitable range that reduces training overfitting and increase the model performance.
6. **Model Prediction on Test data**: the last step will include making predictions on the test data and submitting the best model for scoring on the Kaggle competition website.

References

Bertelsmann (2020). Arvato-Bertelsmann. Retrieved from https://www.bertelsmann.com/divisions/arvato/#st-1

Bertelsmann (2020). Bertelsmann Company. Retrieved from https://www.bertelsmann.com/company/

Computerwoche (1983). Bertelsmann vertreibt Rechner von TI. Retrieved from https://www.computerwoche.de/a/bertelsmann-vertreibt-rechner-von-ti,1180755

Wikipedia (2020). Arvato. Retrieved from https://en.wikipedia.org/wiki/Arvato