Machine Learning Engineer Nanodegree

Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Rubén Cruz García

21st November 2020

Domain background

Arvato Financial Solutions is a German company that offers financial services, Information Technology services and Supply Chain Management solutions for business customers on a global scale. Their services also include the case we will tackle here: client profiles and acquisition.

More specifically, the business matter we will be focusing on here is the following:
a client mail-order company seeks our help/services to acquire new clients more efficiently. To accomplish our project, we will employ customer segmentation: this refers to the practice of dividing a customer base into groups of individuals, depending on well-defined specific features, such as age, gender, family, interests or spending habits.

Dividing the current customer base of the company into smaller meaningful groups will enable us to gain insight into their different types of customers, which will then permit the company to target the German population at large in an informed way: for example, marketing teams would highly benefit from such grouping information, as they would be able to determine which promotional campaign would most appeal to which demographic group before even launching these campaigns. Adding to customer segmentation, we wish to develop a supervised ML model capable to predict whether a person will be a new customer.

Problem statement

The main problem could be formulated as: “Given the demographic data of a person, how can a mail order company acquire new customers in an efficient way”.

Machine Learning techniques can be employed in the two main sections of the project:

- Unsupervised learning methods on the data of established customers and the German general population’s demographics data, we can create customer segments.

- Supervised learning methods on a third dataset, we can train a model to classify
If a person will become a new customer, and use this model for future predictions.

Datasets and inputs

The project makes use of four datasets:

- **Udacity_AZDIAS_052018.csv**: Demographics data for the general population of Germany; 891 211 people (rows) x 366 features (columns).
- **Udacity_CUSTOMERS_052018.csv**: Demographics data for customers of a mail order company; 191 652 people (rows) x 369 features (columns).
- **Udacity_MAILOUT_052018_TRAIN.csv**: Demographics data for individuals who were targets of a marketing campaign; 42 982 people (rows) x 367 (columns). This is the only dataset with the target variable, which is highly unbalanced (people hiring the services are the minority class).
- **Udacity_MAILOUT_052018_TEST.csv**: Demographics data for individuals who were targets of a marketing campaign; 42 833 people (rows) x 366 (columns).

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 the population present in the general German population dataset.

1. The dataset will be explored to examine if there are any missing values fix them. Also, any categorical features need to be re encoded into numerical features. Finally, the missing values will be imputed and the data will be scaled.
2. We will 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) will be used to identify minimum number of features which explain the variation in the dataset.
3. Segment the general population and the customers into different groups based on the selected features with the help of unsupervised learning algorithm, such as K-means clustering.

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

1. Data (train and test) is pre-processed following the same steps applied to the population datasets.
2. A supervised learning algorithm will be trained and evaluated on the processed training data. Proposed algorithms for supervised learning: Random Forest, Ada Boosting and Extra Trees.
3. The trained model will be used to make predictions on the test data provided.

A grid search algorithm should be used to select the best hyper-parameters for the proposed algorithms.

Benchmark model

The benchmark model will be a Logistic Regression algorithm since it is easy to train and test in a short time.

Evaluation metrics

In the unsupervised learning part, in order to choose the ideal number of clusters for the K-means algorithm, we will use the “elbow method”: we select the value of k at the “elbow” i.e. the point after which the distortion/inertia start decreasing in a linear fashion.

Regarding the supervised learning models, where we predict the outcome of a given customer, we will use the Area Under Curve (AUC) metric, since the target labels are highly imbalanced and the accuracy would be a bad choice to evaluate the model. Also, the AUC is the metric chosen to evaluate the final predictions in the Kaggle submissions section.

Project design

A brief explanation of the proposed steps of the project:

1. **Data Cleaning and Visualization**: The data needs processed to identify and clean missing values. An analysis on how many missing values are there per feature will be performed to decide on which features to neglect.
2. **Feature Engineering**: Determining the required number of features that can amount for maximum variance in the dataset using a dimensionality reduction technique like PCA.
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, Random Forest, Ada Boosting and Extra Trees will be used to make predictions. A hyper parameter tuning algorithm like Grid Search will be used to determine the best set of hyper parameters. The previously proposed evaluation metrics will be used to determine the best model in this step.
4. **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.