Machine Learning Engineer Nanodegree: Capstone Proposal

Arvato Solutions: Customer Targeting

Author: Mike Griffin

Date: 22nd May 2020

DOMAIN BACKGROUND

This is a Udacity capstone project representing a real-life data science problem using information provided by Betrelsmann Arvato Analytics.

Arvato is a global services company that develops and implements solutions for business customers in a diverse range of industries, with a focus on innovative data analysis and automation. For this project, Arvato provided services to a German mail-order sales company aiming to understand demographics and improve customer acquisition. This project will explore unsupervised clustering to identify customer groups as well as classification modelling to predict which individuals will respond to campaigns.

Ultimately this work was designed allow the client to improve client targeting for future campaigns. My analysis is for exploration and learning only.

PROBLEM STATEMENT

The overarching objective is to identify which individuals are most likely to respond to a marketing campaign and become customers of the client. This is a binary classification problem using individual-level attributes to estimate the likelihood of a customer response. Specifically, the project will aim to maximise the ROC score on a held-out test set on Kaggle.

DATASETS AND INPUTS

Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company, stored at the level of an individual.

- 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).

365 fields are common across the dataset with information on the fields and labels stored in two data dictionaries:

- DIAS Information Levels - Attributes 2017.xlsx: high-level descriptions of attributes
- DIAS Attributes - Values 2017.xlsx: detailed mappings of levels within fields

SOLUTION STATEMENT

The solution will have two components:

1) **Customer segmentation**– a report which will separate the population in to distinct clusters and explore the variation in attributes by clusters. This will require appropriate imputation, feature engineering, dimensionality reduction and clustering via k-means. This will allow comparison between the population and customer data to explore which clusters are over/under-represented in the customer set.

2) **Customer acquisition** - a set of models will be trained and compared on the binary classification of whether an individual will respond to a campaign. Segment information from (1) will be computed and fed in to the classification model. Model performance will be assessed and tuned using train/validation splits before measurement on a held-out test set on Kaggle.

BENCHMARK MODEL

A simple decision tree will be trained to determine baseline performance, with constraints on leaf size or depth to control overfitting. Baseline performance will be assessed using the ROC AUC score and confusion matrices. This will enable clear interpretation of important metrics, thresholds and interactions which may be used to engineer features with predictive value for the main modelling.

EVALUATION METRICS

For the first stage involving customer segmentation:

- Explained variance will be used to assess the effectiveness of dimensionality reduction.
- For clustering, the “elbow method” will be used to assess the optimal number of clusters using the inertia metric to minimise cluster distances.

For the second stage involving binary classification:

- In line with the Kaggle competition, the AUC ROC score will be the main measure to assess model performance, supplemented by analysis of confusion matrices.

PROJECT DESIGN

A standard data science pipeline will be adopted to cover the two steps:

Customer segmentation

1) **Exploratory data analysis** – value distributions and missingness will be explored to enable appropriate data processing.
2) **Data processing and feature engineering** – standard techniques will be used for imputation, scaling and one-hot encoding. Principal component analysis will be explored to reduce dimensionality and feature engineering may be used to capture feature interactions.
3) **Clustering analysis** – unsupervised clustering methods (k-means etc) will be used to explore data groupings. The elbow method will be used to determine the appropriate cluster number.

Customer acquisition

4) **Supervised modelling** – a benchmark decision tree model will be used to define baseline performance. A selection of models will be trained, with a focus on classical ML (logistics regression, random forests, gradient boosting) as this is a relatively small tabular dataset.
5) **Tuning and model selection** – parametrised models will be optimised and compared using clear data splits:
    - A held-out test set of 25% of the dataset to compare model
    - Cross validation within the remaining data for train/validation splits for hyperparameter tuning. ROC score will be the primary performance measure
6) **Kaggle submission** – a limited number of Kaggle submissions will be made to avoid overfitting the private test set, with benchmarking against other models via the leader board.
7) **Model interpretation** – measures of feature importance/partial dependence will be explored to understand model and implications for future targeting strategy.