Sebastian Koenig

Bertelsmann/Arvato Customer Segmentation & Data Driven Marketing

December 17th 2019

Domain Background

This project combines unsupervised and supervised machine learning on structured data to enable data driven marketing.

Problem Statement

This project will aim to identify the most likely customer acquisition prospects for a German mail order company by analyzing what portion of the general population most resembles the existing customer base. In a second step we then hope to predict individual customer responses to the marketing campaign.

Datasets and Inputs

The data for this project has been provided by Bertelsmann/Arvato in the form of csv files, which we will need to clean and preprocess. This includes data gained through market research about the general population and the existing customer base, as well as labeled data of customer responses to the marketing campaign and an unlabeled test set.

Solution Statement

K-means clustering appears to be the most logical approach for population segmentation since we are looking for a subgroup of the population that is most similar to the existing customer base. XGBoost has performed well on the supervised learning portion of this project in the past. We will compare it to Google's AutoML performance and if neither provides satisfying results be constructing a custom model.

Benchmark Model

The unsupervised model should show us a population segmentation where the vast majority of the existing customer base belongs to a single cluster of the general population.

We will compare the performance of our supervised models by submitting them to the existing kaggle competition and see how they measure up. As of this writing the top performing models on the leaderboard come in above 0.8 on the AUR metric. The score to beat is 0.80819.

Evaluation Metrics

As mentioned above the evaluation metric for the Kaggle competition is AUR, and will thus be the metric we employ. Accuracy would not deliver good results since the input data is highly asymmetric.

Project Design

This project will have four stages. First the data for the unsupervised learning tasks will have to be cleaned and then segmented using k-means clustering. We may have to iterate over several versions with different numbers of clusters and selected features.

We will then train the first two supervised models. First Google AutoML, to see how well the lowest effort method performs, then XGBoost using AWS SageMaker. If neither of these models archive an AUR of 0.8 or above we will search for a custom architecture and perhaps hand engineer features for a third model.