AWS Machine Learning Engineer Nanodegree Capstone Proposal
Arvato Customer Acquisition Prediction Using Supervised Learning

Fady Morris Milad Ebeid

January 23, 2022

Contents

1 Domain Background

ARVATO is a company that provides financial services, IT services, and supply chain management services solutions to other businesses. It has a large base of global customers. It's solutions focus on automation and data analytics.[KGaa]

Arvato's customer base come from a wide variety of businesses, such as insurance companies, telecommunications, media education and e-commrece.

Arvato analytics is helping businesses to take important decisions ang gain insights from data. It uses data science and machine learning to meet business goals and gain customer satisfaction.

Arvato is owned by Bertelsmann [KGab], which is a media, services and education company that operates in about 50 countries around the world.

In this project, Arvato is helping a mail-order company that sells organic products in Germany to build a model to predict which individuals are most likely to convert into becoming customers for the company by analyzing marketing campain mailout data.

Customer retention and churn were addressed in the following academic research papers: [AF20] and [Zhu18]

2 Problem Statement

The problem can be stated as: " Given the existing marketing campain demographic data of customers who responded to marketing mails, how can we predict whether a new person will be a potential customer for the mail-order company?"

A supervised learning algorithm will be used to train a model that will help the company make such predictions and decide whether is a person is a potential candidate to be a customer for the company or not.

3 Datasets and Inputs

The dataset is a private dataset. It is used with permission from Arvato for use in the nanodegree project.

There are two files of the dataset that we are concerned with:

- `Udacity_MAILOUT_052018_TRAIN.csv`: Demographic data for people in Germany that were targeted by the marketing mailing campain. It contians data for 42,982 individuals.

  This training daataset has 367 columns, 366 of them are demographic features and 1 label column `'RESPONSE'`

- `Udacity_MAILOUT_052018_TEST.csv`: The testing dataset, It contains data for 42,833 individuals and it has it has 366 columns of the demographic features, this dataset has no label columns and will be tested using kaggle api for Udacity+Arvato: Identify Customer Segments competition.

There are also two metadata files that contain a data dictionary for the demographic features in the previous dataset files.

- `DIAS Information Levels - Attributes 2017.xlsx`: An excel sheet that contains a top-level organization of demographic features, their description and some notes.
- `DIAS Attributes - Values 2017.xlsx`: An excel sheet that contains demographic features sorted alphabetically, their description, their values, and meaning of each value.

A quick examination of the dataset showed that the dataset is highly skewed (imbalanced). only 1.24% of the individuals targeted by the marketing campaign would respond to it as shown in the following table:

RESPONSE	count	percentage
0	42430	98.76%
1	532	1.24%

4 Solution Statement

In this project I will use the provided data to predict whether an individual will be a customer for the mail-order company or not. The general solution steps are:

- The data will be pre-processed. The data will be explored for missing values, invalid values outside the range defined in `DIAS Attributes - Values 2017.xlsx`. Then, categorical features will be encoded as numerical features. Finally, The features will be z-scaled and normalized to have similar ranges of values and accelerate learning convergence.
- A supervised machine learning algorithm will be used to train a model on the training dataset in `Udacity_MAILOUT_052018_TRAIN.csv`, using the column `'RESPONSE'` as a label for training.

  Some of the candidate algorithms for model training are logistic regression, decision tree classifier, random forest and XGBoost.
- Hyperparameter tuning of model hyperparameters to obtain the best performance metric. I will use AWS Sage maker Hyperparameter Tuning Job for this task.
- The trained model will be tested using the test set from `Udacity_MAILOUT_052018_TEST.csv` and the results will be validated using kaggle api for the competition Udacity+Arvato: Identify Customer Segments.

For full details of the technologies and techniques that will be used for the solution refer to Section 7 - Project Design.

5 Benchmark Model

A benchmark model will be a simple Logistic Regression model trained using the default hyperparameters. Other proposed supervised learning model will be compared to this baseline model in terms of performance and training time.

6 Evaluation Metrics

Since this is a binary classification problem, and the dataset is imbalanced as shown in Section 3 - Datasets and Inputs, the proposed classification metrics are:

- F1 Score.
- Area under the receiver operating curve (AUROC).

Since the dataset is highly skewed, then accuracy score will not be a good choice for such a problem.

Some of the records from `Udacity_MAILOUT_052018_TRAIN.csv` will be used as a validation dataset and the performance metric will be compared against.

7 Project Design

In this project I will use Amazon Web Services and their Sagemaker Compute Instances for data cleaning, model training, hyperparameter tuning and generating predictions.

Amazon SageMaker is a fully managed cloud computing service that provides a machine learning engineer with the ability to prepare build, train, and make inferences for machine learning models quickly.

1. The dataset files in Section 3 - Datasets and Inputs will be uploaded to an Amazon S3 bucket.
2. Data pre-processing, cleaning, exploration and feature engineering: The data the data will be checked for missing values. Invalid values will be spotted and fixed using the provided metadata file `DIAS Attributes - Values 2017.xlsx`

   The pre-processing step will be completed inside an Amazon Sagemaker Notebook Instance
3. A baseline Scikit-learn logistic regression model will be trained on the dataset, the metric obtained will be recorded for future reference.
4. Candidate supervised learning models will be evaluated against the dataset and the best performing model will be selected.
5. Hyperparameter tuning will be done on the selected model using a Sagemaker Hyperparameter Tuning job. This step will use regression on the hyperparameter search space to find the best and optimal hyperparameters to train the model.
6. Model training using the best hyperparameters obtained in the previous steps. I will use one of Amazon Sagemaker built-in algorithm images for model training using appropriate compute instance.
7. Generating Predictions: The test dataset from `Udacity_MAILOUT_052018_TEST.csv` will be used by the final model trained in the previous step to generate predictions that can be evaluated using Kaggle API for the Udacity+Arvato: Identify Customer Segments competition.

References

[AF20] Atallah M. AL-Shatnwai and Mohammad Faris. "Predicting Customer Retention using XG-Boost and Balancing Methods". en. In: *International Journal of Advanced Computer Science and Applications* 11.7 (2020). ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2020.0110785. URL: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=7&Code=IJACSA&SerialNo=85 (visited on 01/23/2022).

[KGaa] Bertelsmann SE & Co KGaA (www.bertelsmann.com). *Arvato - Bertelsmann SE & Co. KGaA*. en. Publisher: Bertelsmann SE & Co. KGaA, Carl-Bertelsmann-Straße 270, D-33311 Gütersloh, www.bertelsmann.com. URL: https://www.bertelsmann.com/divisions/arvato/ (visited on 01/23/2022).

[KGab] Bertelsmann SE & Co KGaA (www.bertelsmann.com). *Company - Bertelsmann SE & Co. KGaA*. en. Publisher: Bertelsmann SE & Co. KGaA, Carl-Bertelsmann-Straße 270, D-33311 Gütersloh, www.bertelsmann.com. URL: https://www.bertelsmann.com/company/ (visited on 01/23/2022).

[Zhu18] Yayun Zhuang. "Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm". en. In: *Management Science and Engineering* 12.3 (Sept. 2018). Number: 3, pp. 51–56. ISSN: 1913-035X. DOI: 10.3968/10816. URL: http://flr-journal.org/index.php/mse/article/view/10816 (visited on 01/23/2022).