Capstone Project: Beterlsmann/Arvato Predicting Campaign Response

AWS Machine Learning Engineer Nanodegree
Jean Carlos da Cruz

Table of Contents

Domain Background

> Beterlsmann/Arvato is a company that provides financial, IT and supply chain management services. It is a global company focused on automation and analytics. Many global renowned companies from a wide variety of industries, from telecommunications providers, energy, banks, insurance, e-commerce, IT and others make use of the Arvato's portfolio of solutions. The company is helping businessses to retrive insights and make better data drive decisions by using data science and machine learning. In this project, we aim to help a mail-order company that sells organic products in Germany to build a solution that can predict which users are more likely to convert into becoming customwers for the company. The existing customer data and the demographic data of population in Germany are to be studied to understand different customer segments, and then building a system to make predictions on whether a person will be a customer or not based on the demographic data.

Problem Statement

> The problem statement can be describe as: "Given the demographic data of a person, how a mail order company can convert users into new customers?"
> In order to answer that question we will first indentify segments in general population and segments inside the existing customers, after that discover what demographic features correspond to a person being a customer for the company. Secondly, we will model a supervised learning algorithm to predict on whether a person is a probable customer or not, based on the demographic data.

Datasets and Inputs

> 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).
>
> Additionally, 2 metadata files have been provided to give 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.
>
> Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. Use the information from the first two files to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS"), then use your analysis to make predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company.
>
> The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed; it is against that withheld column that your final predictions will be assessed in the Kaggle competition.

Solution Statement

> The project will be divided into two parts, the first one we will identify any customer segments present in the provided dataset and match these segments with the segments of population present in the general population dataset, below a better description of each step in the first part: - Dataset will be explored to examine if there are any missing values or missing records in order to fix them in case it exists. Any categorical features will be encoded into numerical features using label encoders. After that, the data will be scaled in order to ensure any feature will have higher weights the later steps; - The second stop will be to identify the minimum number of features that would be sufficient to explain the dataset. Since there are more than 300 features that represent a single person and not all the features will be relevant to describe the segments. So, a dimensionality reduction thecnique like PCA (principal component Analysis) can be used here to identify minimum number of features which explain the variation in the dataset. The third and last step will be to create segments the general population and the customers based on the selected features using the K-Means Clustering, a unsupervised learning algorithm.
>
> The second part of the project will be the task to predict the likelyhood of a user to convert into a customer by doing each step as described below: - To begin we will pre-process the data once again, like we did in the first step of first part; - Secondly, we will train a supervised learning model and evaluated on the pre-processed training data, in order to get the best model possible we will test the following models: Logistic Regression, Decision Tree, Random Forest and XGBoost; - To close this this part, we will make predictions on the test data using the trained model.

Benchmark Model

> To use as baseline model as benchmark we will select a Logistic Regression model using the default hyperparameters since it is easy to train, test and explain the results.

Evaluation Metrics

> For the customer segmentation part we will use dimensionality reduction technique (PCA) to reduce the number of dimemsions, the selection will be based on the explained variance ratio of each feature, the mininum number of dimensions explaining as much variation as possible, the better. To create segments we will use K-Means, in this case the number of clusters will be a hyperparameter and it will be selected based on the squared error i.e. the distance between all the clusters. For the model prediction part the evaluation metrics will be the standard ones when it comes to classification models, such as: F1 Score and AUROC, because the dataset is highly skewed, the accuracy score will not be a good chose to evaluate the models.

Project Design

> In this project I will use AWS and their SageMaker instances for data cleaning, model training, hyperparameter tuning and generating predictions. Sagemaker is a fully managed cloud computing service that provides a machine learning engineer the ability to prepare, build, traind and make inferences for machine learning projects. The steps in the project will described below: - The files in Section 3 (Datasets and Inputs) will be uploaded to S3 Buckets; - Data wrangling, EDA and feature engineering will be done using Sagemaker notebooks instances aiming fixing possible missing data and mis recorded values. The misrecorded values will be fixed based on the information provided in the metadata files; - The modeling will be also done in Sagemaker, using K-Means Clustering to segment data into clusters and thein we will test severeal classification models in other to better predict the probability of convertion from users to customers; - The best model will be selected to be tuned to improve its performance. A hyper parameter tuning algorithm like Grid Search will be used to determine the best set of hyper parameters; - Finally, the best model (tuned) wil be used to make predictions on the test data and the predictions can be evaluated using Kaggle API.

References

[1] Arvato-Bertelsmann Web-site. Available at: https://www.bertelsmann.com/divisions/arvato/#st-1 [Acessed April 2022];