Arvato Financial Solutions: Identify Customers from a Mailout

Machine Learning Engineer Nanodegree

Dec 5, 2021

Project Background

To fulfill Udacity Machine Learning nanodegree, I am required to choose a project provided by Udacity or a customized project at my own interest. My passion goes to data-driven market growth, in particular, using data science to analyze customer activity patterns in relation to products, and how to use such pattern to optimize market growth. Thus, I selected this project for further research.

Domain Background

Arvato Financial Solutions, a Bertelsmann subsidiary, provides their data and outline to be one of the Udacity listed projects. Arvato Financial Solutions provides professional financial services to renowned international brands as well as respected local businesses. Their services center around cash flow in all segments of the customer lifecycle: from identity, fraud and credit risk management, to payment and financing services and debt collection.

In this project, Arvato requires a report for customer segmentation, and solution to identify potential customers that will respond to their marketing campaign. The final product from this project will be a notebook clearly explained the data pattern, centered around customer segmentation, and an effective supervised learning model evaluated by ROC AUC on Kaggle. As for Udacity, an additional final report of the entire project will also be required.

Regardless of company size, the adoption of data science and machine learning for marketing has been rising in the industry. I am excited to analyze this real world data to find out actionable insights for business owners, and support their decisions with scientific findings. In particular, I will research and investigate on feature analysis, clustering algorithms, and classification algorithms, together with other essential data cleaning and model evaluation techniques.

Problem Statement

A clear insight of customer segmentation is at the center of interests for Arvato Financial Solutions. Moreover, Arvato desires to optimize their mail marketing campaign to send mails to selected individuals, ideally those who are likely to be customers. To be specific, this project intends to answer two questions:

i. How to identify parts of the general population to be ideal customers for Arvato mail-order business? Essentially, the dissimilarity between the general population and target customers shed light on the pivotal characteristics of segmentation. My task is to find out the dissimilarity among these two populations.

ii. A method to identify individual who is likely to be a customer for mail-order business. The solution will most likely point to classification algorithms.

Datasets and Inputs

The data have four main files and two supporting files:

Main Files

- *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).
- *Udacity_AZDIAS_052018.csv*: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

Supporting Files

The supporting file provides information about the columns depicted in the files, in Excel format.

- */DIAS Information Levels - Attributes 2017.xlsx* is a top-level list of attributes and descriptions, organized by informational category.
- */DIAS Attributes - Values 2017.xlsx* is a detailed mapping of data values for each feature in alphabetical order.

Solution Statement

For the first part of the project, I will apply K means clustering, DBSCAN(Optional) and HDBSCAN(Optional) clustering algorithms.

- *k means clustering*

K means is a popular clustering method. First, you choose k - the number of clusters. Then you randomly put k feature vectors, called centroids, to the feature space. We then compute the distance from each example x to each centroid c using metric, like the Euclidean distance. Then we assign the closed centroid to each example. For each centroid, we calculate the average feature vector of the examples labeled with it. The average feature vectors become the new locations of the centroids.

- *DBSCAN*

While k-means and similar algorithms are centroid-based, DBSCAN is density-based clustering algorithm. Instead of guessing how many clusters you need, by using DBSCAN, you define two hyperparameters: ϵ and n. Similar to the K means algorithm in iterating the data points assigned around a center, the distance is measured against ϵ, and the number of samples assigned in the cluster is measured against n. The advantage of DBSCAN is that it can build clusters that have arbitrary shape.

For the second part of project, I will apply classification methods as below:

- *Logistic Regression - Random Forest Classifier - KNN - SVM with different kernels - Boosting algorithms (catboost, adaboost, xgboost)*

Benchmark Model

For the supervised learning task, I will use Logistic Regression model as my benchmark model. This model raw outputs probability of a predicted class, indicating an ordinal ranking order. Besides this, the first ranked candidate in Kaggle Leaderboard scores 0.81063, serving as another benchmark for my model reference.

Evaluation Metrics

The evaluation metric for Kaggle competition is AUC score. AUC stands for "Area under the ROC Curve", where ROC curve is a graph plot with True Positive Rate (TPR) on Y axis and False Positive Rate (FPR) on X axis. This ROC curve variates from the change of classification threshold. Lowering the threshold classifies more item as positive, thus increasing true positives and false positives.

AUC provides an aggregate measure of the performance across all classification thresholds. It ranges from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; while a model has perfect predictions on all instances has AUC of 1.0.

Project Design

The project follows those three stages given by Arvato Financial Solutions. In short, the first two stages will involve data cleaning, data analysis with visualization, and then building model, evaluate and interpret model. The last stage will be submitting predicted results to Kaggle. After this step, I will write the final report.

Stage 1: Customer Segmentation Report

"This section will be similar to the corresponding project in Term 1 of the program, but the datasets now include more features that you can potentially use. You'll begin the project by using unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments."

This explanation is given by Udacity. The underlying requirement is to segment customer into meaningful cohorts. Each cohort with similar characteristics, while between cohorts there is at least one distinguished nature.

We will first preprocess the general population dataset and customer datasets respectively, by addressing below problems, followed by my proposed solutions:

1. Handle missing values: cleaning and/or filling missing values

2. Implement feature selection and transformation:

2.1 correlation analysis

2.2 feature transformation such as categorical value encoding, numerical value scaling

2.3 dimension reduction such as Principle Component Analysis

3. Build clustering models

4. Interpretation of the results, mainly comparing the dissimilarity of general population segmentation and Arvato customer segmentation

5. Conclusion

Stage 2. Supervised Learning Model

"You'll have access to a third dataset with attributes from targets of a mail order campaign. You'll use the previous analysis to build a machine learning model that predicts whether or not each individual will respond to the campaign."

The third dataset may or may not be the same as the first two datasets in stage 1. This new dataset will be used for training a supervised model, and later, evaluated against additional new dataset without target variable. The performance measurement here is AUC. We need to pay attention to the ordinal ranking of individuals to represent the likelihood of being a customer.

If data run into imbalanced class problem, I will address it by using two methods: in data level, using data augmentation tool, e.g. SMOTE; in algorithm level, adjust class weights in algorithm directly, e.g. XGBoost.

Stage 3. *Final Report*

After the project is completed, a project report (in PDF format only) is required, addressing the five major project development stages.

Reference

- Arvato Financial Solutions
- Kaggle
- Area Under the curve
- ROC Curve and AUC