Machine Learning Engineer Nanodegree

Capstone Project

Mridul Gangwar
March 09, 2020

I. Definition

Project Overview

Finding new customers for a product is always a challenge. In an ideal world where you have a lot of funds available to acquire new customers, you can run promotions, discounts, online and offline campaigns to get new customers. However, when you have limited funds and staff hours available, you should narrow down your efforts to a target segment, which is more probable of buying your products.

"If I'm trying to expand sales, I have to find out who my existing customers are. What are their demographics? What do they look like?" Jerry Osteryoung, director of outreach for the Jim Moran Institute for Global Entrepreneurship at Florida State University. In this project, I will be creating a machine learning model to assist marketing team to target the right segment.

Problem Statement

In this project, I will be working on data provided by Arvato Bertelsmann. I will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population.

1) I will be using a clustering algorithm to create a segment of customers and identifying a similar group of people in the global population dataset. Since the characteristics of these two groups are the same, we can expect them to more probable to become customers.
2) The learnings from the previous comparison will be used to create a model to predict which individuals are most likely to convert in the marketing campaign.

Metrics

It's a classification problem, whether the person will respond to the mail or not. I am expecting the data to be unbalanced as very few people react to the mail campaigns; the best metrics to be used are AUC/ROC and f1 score metrics. I will be using AUC to evaluate the models.

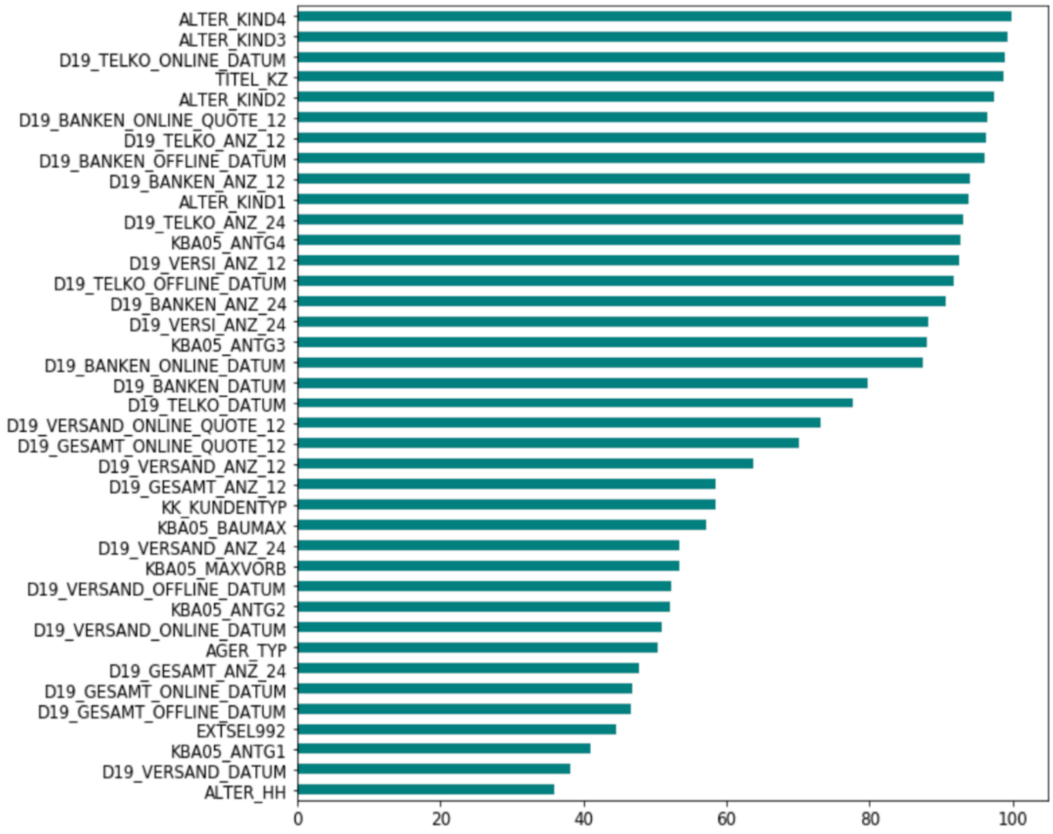
II. Analysis

Data Exploration:
Six datasets included in this project:

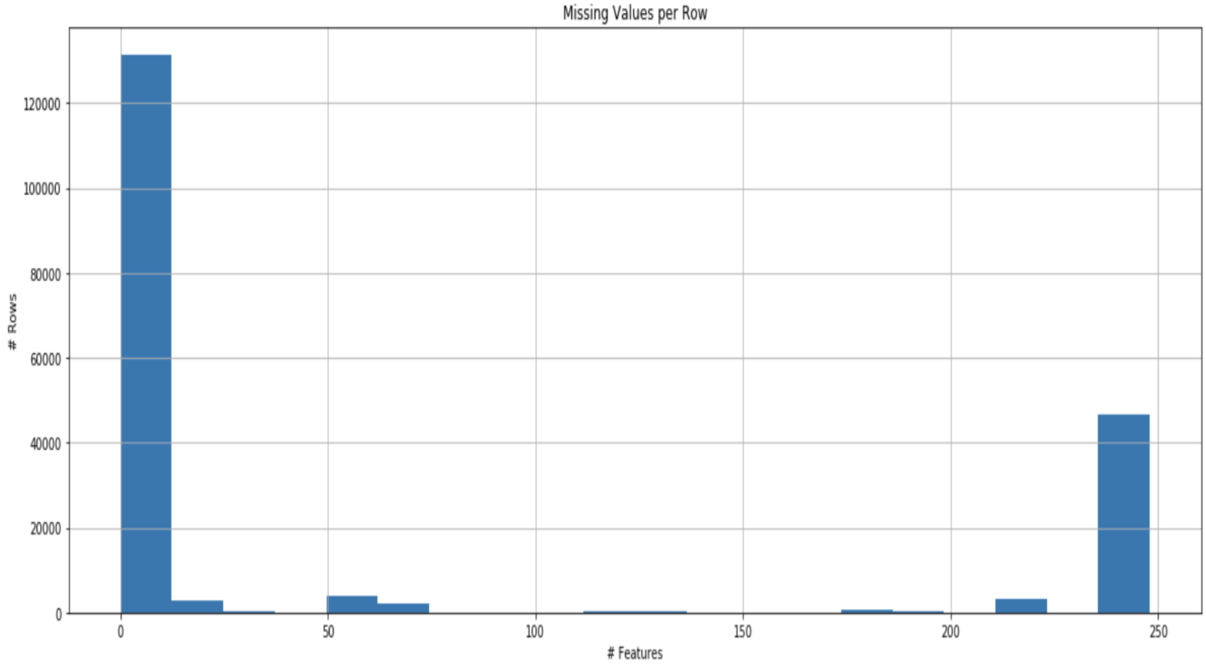
1) AZDIAS: demographics data for the general population of Germany
2) CUSTOMERS: demographics data for customers of a mail-order company
3) DIAS Attributes: detailed mapping of data values for each feature in alphabetical order
4) DIAS Information Level: top-level list of attributes and descriptions, organized by informational category
5) Udacity_MAILOUT_052018_TRAIN: This dataset contains all the demographic information of the targeted population and the response whether they responded to the mail or not. This dataset is used to train the machine learning model
6) Udacity_MAILOUT_052018_TEST.csv: This is the dataset used to generate prediction, which is submitted to the Kaggle competition

Steps performed in the data exploration:

1) I calculated the shape of the dataset. While working with azdias, I was getting timeout error or idle workspace error. That is why I decided to work with a sample of the azdias dataset. This sample is created by randomly selected 1/5 of the rows of the total azdias dataset
2) I calculated mean, minimum, different percentiles of all the numerical columns of both the azdias and customer datasets. Most of the columns have similar values and not deviated a lot.
3) After reading data, I realized the datatype in column 18 and 19 (i.e. 'CAMEO_DEUG_2015', 'CAMEO_INTL_2015') are of mixed type. From the DIAS Attributes dataset, I realized there are some garbage values in these two columns. Hence, I created a function to replace these garbage values with Nan.
4) Using DIAS Attributes dataset, I realized for most of the columns, some value represents null or unknown values. I identified those values for each column and then replaced those values with nan.
5) I visualized which columns have more than 40% of the rows as missing. I identified those columns and later deleted from further analysis. Below is the snippet for the customer's dataset:



6) Next, I visualized the distribution of the rows with the number of null values in them. Later, I deleted rows with more than 250 null values. Below is the snippet for the customer's dataset:



7) Also, by manual inspecting DIAS Attributes dataset, identified a few features which could be used to represent data in a better way. For example, reducing the number of categories in order to better represent the data.

Benchmark

I chose a threshold of 40% to remove columns with null values more than the threshold. I prefer this threshold as a small value will result in a loss of data, and a large value could result in some garbage columns.

III. Methodology

Data Preprocessing