Arvato Marketing Data Project Proposal

Project Overview

In this project the overall goal is to anaylyze a marketing data set with various demographic information (using unsupervised learning techniques) then develop an operationalized endpoint that is capable of determining if a particular user should receive marketing content or not based on if they are predicted to respond or not.

Domain Background

The general domain of this project is that of marketing and marketing segmentation. With a finite number of resources determining who to market to and in what way in the most efficient way possible. I decided to take on this project because my current position is that of one that markets to large audiences and I would like to explore new techniques in developing more effective and efficient segments.

Problem Statement

There are two main "problems" were trying to solve in this project.

1. Analyze demographic information and segment users into various groups, determine which features in the data are most relevant.
2. Produce a model that will determine if we should market to a person based on previous results

Datasets and Inputs

There are 4 files that are provided for this project that will be used to analyze the population set and create a predictive model.

population_demo.csv

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

customer_demo.csv

Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
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.
This file includes three extra columns from the population_demo.csv, ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file.

train.csv

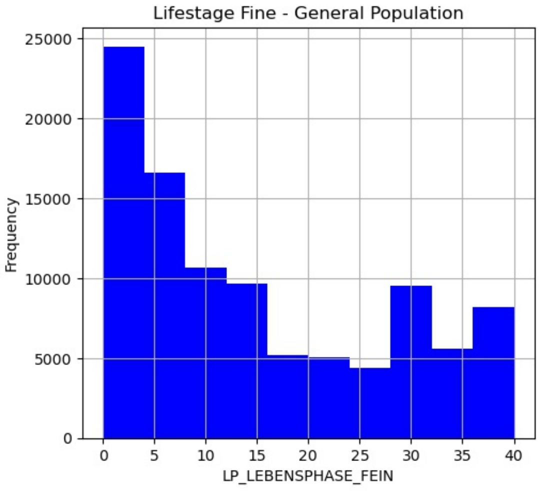
The train file contains instances where the company actually mailed out advertorials. Many of the same demographic columns as the other files but includes a "RESPONSE" column that indicates which of these customers actually purchased.

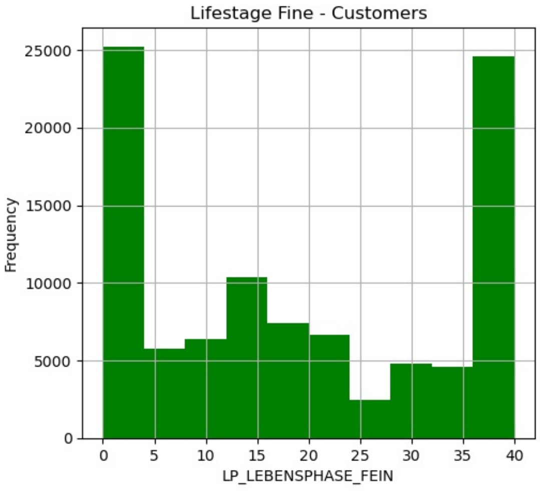
test.csv

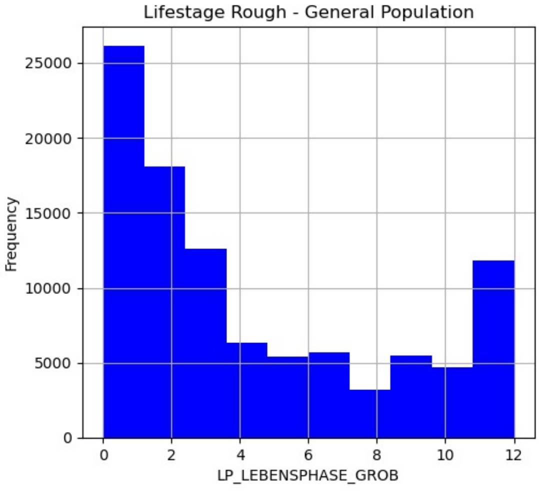
This file is very similar to the train except the "RESPONSE" column is left out, we will use our model to fill in these responses and submit to [this kaggle competition (https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview)](https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview).

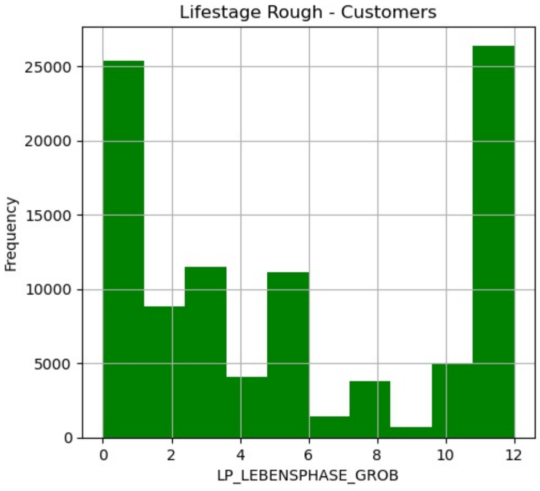
Initial Exploration

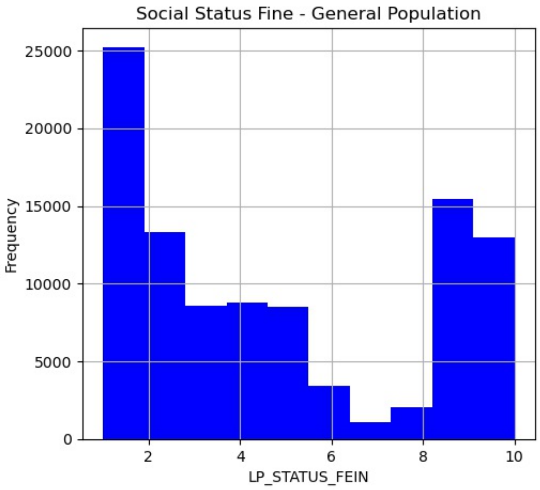
There are many features in this data set, one of the ones I thought might be representative is LP_LEBENSPHASE_FEIN (lifestage fine), LP_LEBENSPHASE_GROB (lifestage rough), LP_STATUS_FEIN (lifestyle fine) & ANREDE_KZ (sex). Below are several histograms between the observed customer base and the general population.

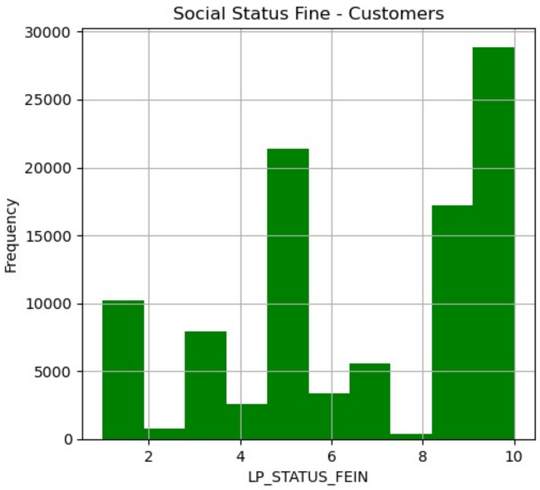

Lifestage Fine - General Population

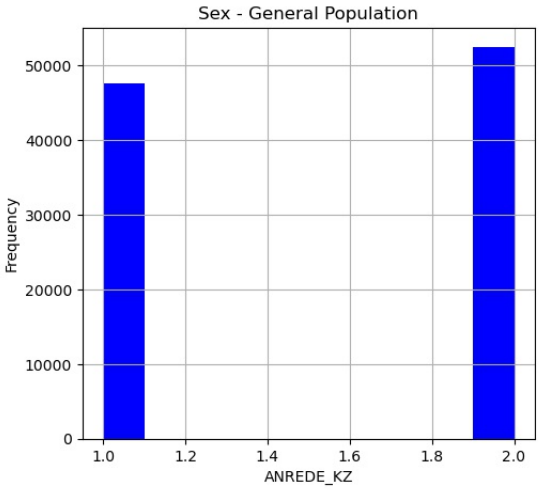

Lifestage Fine - Customers

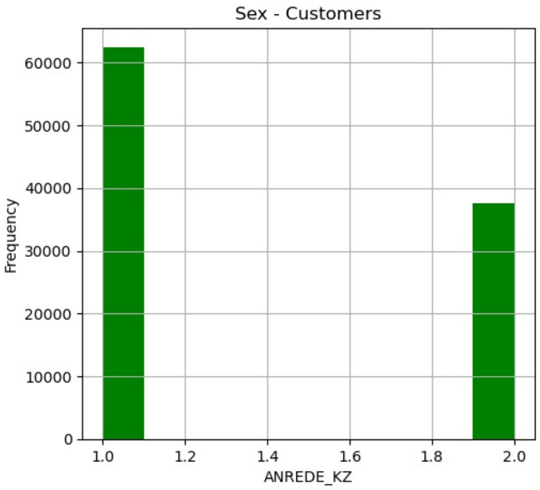

Lifestage Rough - General Population


Lifestage Rough - Customers









When looking at these charts for the first three lower numbers generally mean less affluent and higher numbers mean more affluent and for the last chart 1 represents men and 2 represents women.
So comparing the two distributions we can see our customers are more affluent and more regularly men compared to the rest of the population.

Solution Statement

Evaluation Metrics

The primary evaluation metric will be to properly classify if the users RESPONSE column. When selecting the evaluation metric to plug into auto gluon based on my readings here (https://www.kaggle.com/code/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model/notebook) log loss (the parameter label log_loss for auto gluon (https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.html)) because we are uncertain about the number of samples that will be buyer or not and this algorithm will reduce prediction error (accuracy requires a balanced and binary classification set).

Benchmark Model

The benchmark model will be an AutoGluon model trained on the data before any data cleaning or feature selection has been done, this will be the first submission to the Kaggle competition.

Solution Steps

As there are two problems stated in the problem statement I see this project being broken into two different solutions.

1. For the analysis portion the data will need to be columned over, determining which features in the data are relevant and which are not. This will involve some visualizations of the data along with some unsupervised learning analysis to determine which features are most correlated to the RESPONSE column. At first glance I feel sklearn.feature_selection.SelectPercentile (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html#sklearn.feature_ may be a good method for feature selection. Once some of these determinations are made (relevant features) a workflow will need to be created within data wrangler to fill in gaps in the data / clean the data.
2. Once we have a cleaned the train file I will most likely split this file by some proportion to create a validation and "local" test (separate from the test file submitted to Kaggle) file. I will then use a AutoGluon Classifier process to fine tune and train the model for the optimal results. From there I will use this model to fill out Kaggle submission and submit to Kaggle. The end goal is to have a classification accuracy (if the user will be a buyer or not) greater than the original benchmark Kaggle Solution.
3. EXTRA CREDIT: I would like to deploy this model in a way where a user can make a REST request to an endpoint (that the body of the request contains certain features) and the endpoint will respond back with if the customer should be marketed to our not.

Citations

- AutoGluon Documentation (https://auto.gluon.ai/stable/api/autogluon.tabular.TabularPredictor.html)
- VIPUL GANDHI - How to Choose Right Metric for Evaluating ML Model (https://www.kaggle.com/code/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model/notebook)