AVRATO CUSTOMER SEGMENTATION AND CAMPAIGN CONVERSION

Md Ahsanur Rashid

10th July, 2020

Direct marketing or direct response marketing was first coined in 1958 [1]. Response channel has evolved over time from reply cards, forms and 800-numbers to websites and emails.

Optimization is the key driving force behind every operation today. Advertisement campaigns are no different. There used to be a time when mass advertisement campaigns were all the rage. But with changing times, optimized, targeted advertisement campaigns are more popular.

Banks have a lot of information on their clients and it is easy to use that information to learn key characteristics about clients and be able to target persons with a higher chance of winning over with a direct marketing campaign.

Problem Statement

- We need to better understand the customers of Avrato. This can be done by using unsupervised learning and grouping customers into different buckets based on their attributes.
- We need to create a model to predict whether a person with a given set of attributes will be a good direct marketing campaing customer or not. This can be done by using supervised learning.

The steps that I envision are:

- Download and process customer data
- Use unsupervised learning to bucket customers based on attributes and clean
- Use supervised learning on a different set of data and create a model for predicting good customers for mail-in campaign
- Test model accuracy

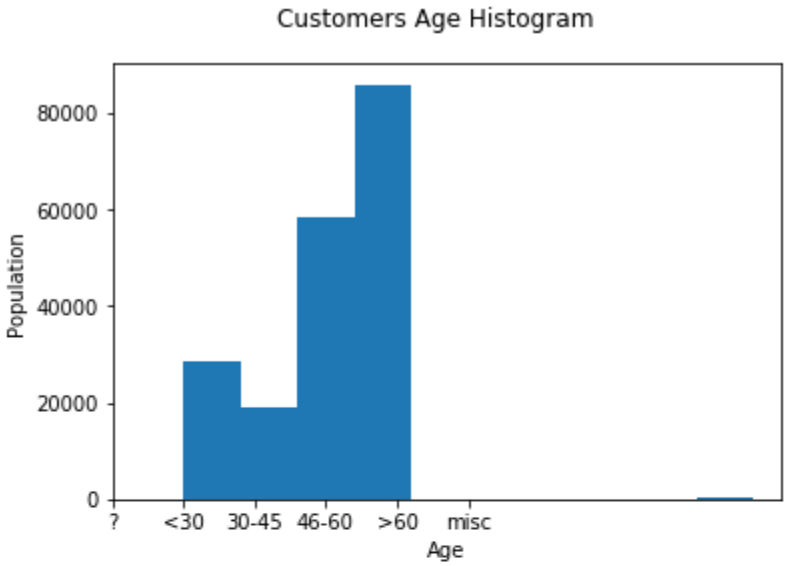
Datasets and Inputs

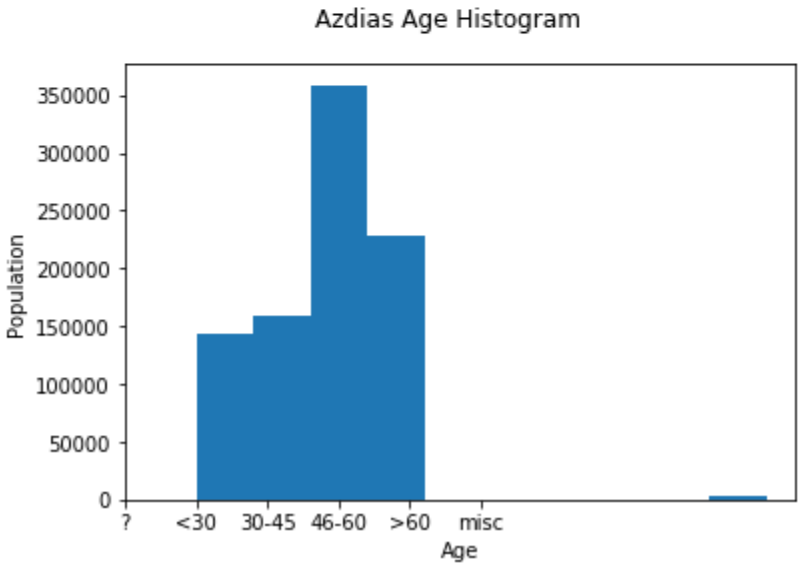
Dataset and all inputs have been provided by Avrato.

There are four total data files provided:

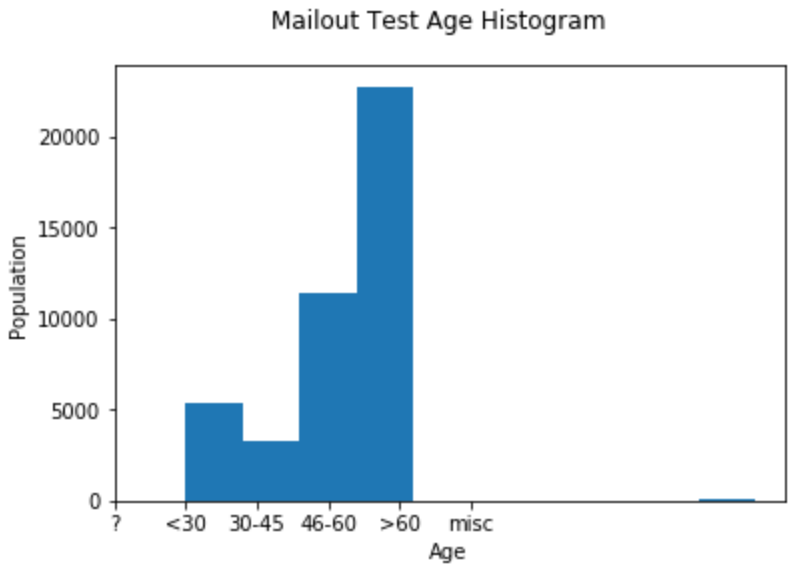
- AZDIAS
    - Data of general population in Germany
    - 891221 rows and 366 columns
- CUSTOMERS
    - Data of customers of the company
    - 191652 rows and 369 columns
- MAILOUT_TRAIN
    - Data for individuals who were targets of direct marketing campaign
    - 42962 rows and 367 columns
- MAILOUT_TEST
    - Data for individuals who were targets of direct marketing campaign
    - 42833 rows and 366 columns

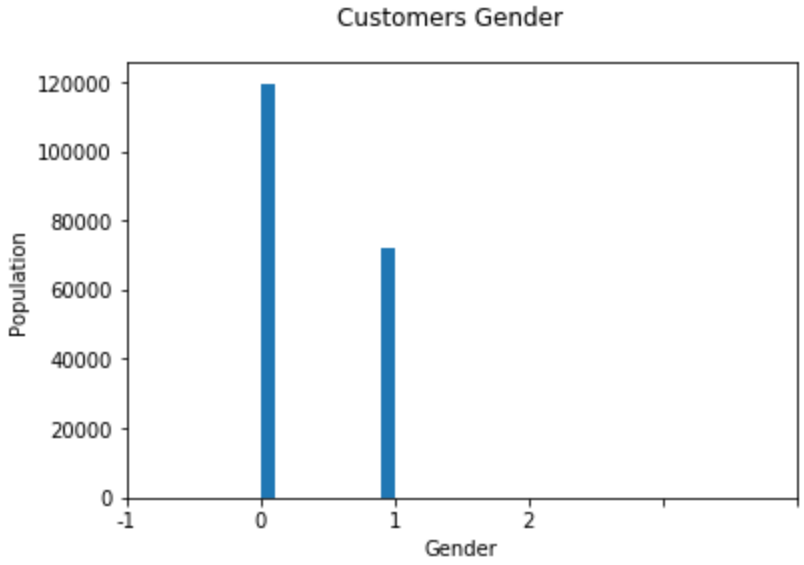
Age and gender distribution among the four data set visualized as the following:

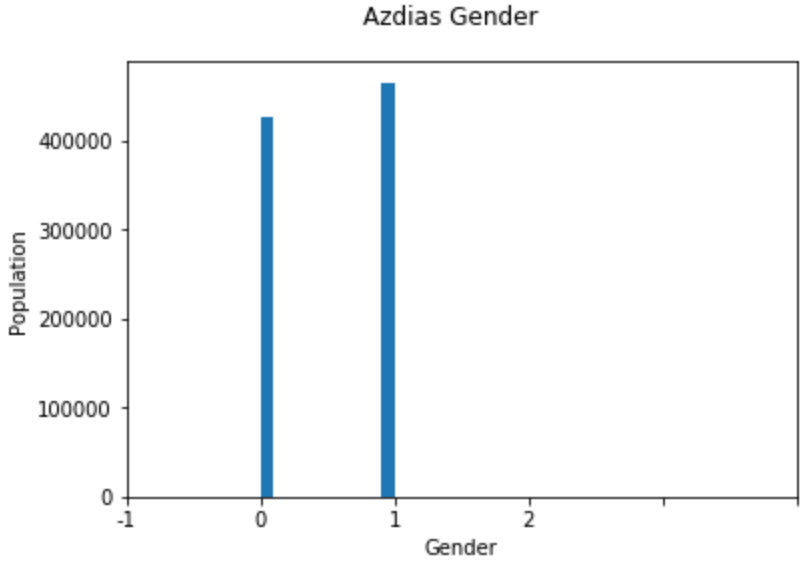


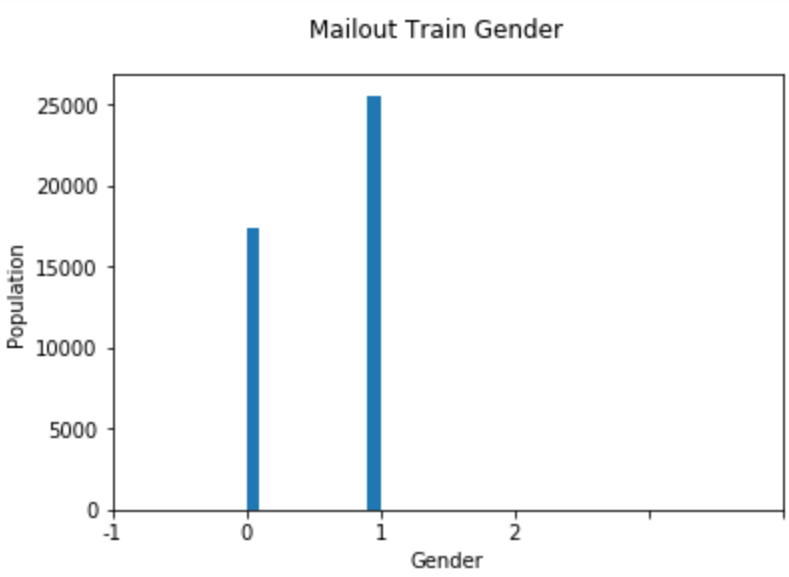


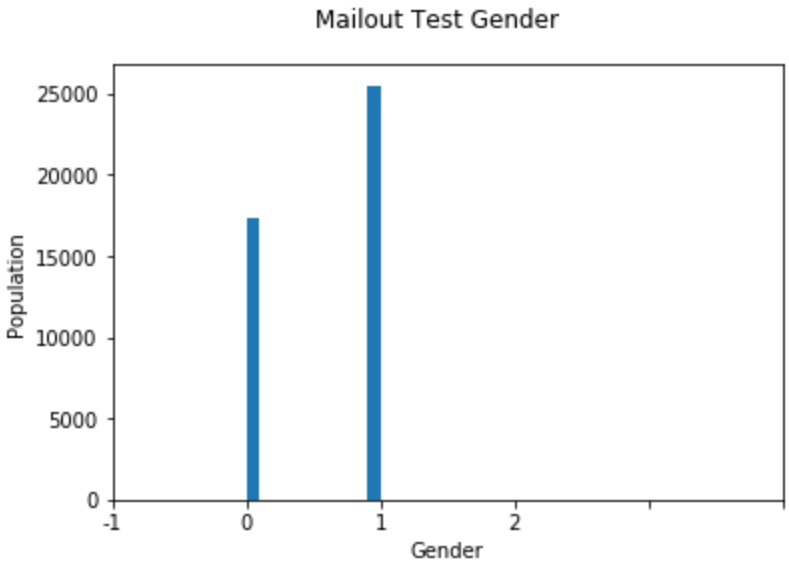

Mailout Train Age Histogram
Population
20000
15000
10000
5000
0
?
<30
30-45
46-60
>60
misc
Age


Mailout Test Age Histogram
Population
20000
15000
10000
5000
0
?
<30
30-45
46-60
>60
misc
Age


Customers Gender
120000
100000
80000
60000
40000
20000
0
Population
-1
0
1
2
Gender


Azdias Gender
400000
300000
200000
100000
0
Population
-1
0
1
2
Gender





The data quality is not very good, quite a few values are missing and if we clean up by brute force, then we only have a handful of data to work with. As seen in the gender visualization above, most data we have is either for males or unknown/missing.

Algorithms and Techniques

I am envisioning using the following algorithmic approaches:

- Logistic regression
- Convolutional Neural Network
- XG Boost
- KNN
- SVM

Benchmark Model

On Kaggle, the highest accuracy achieved so far for this ML problem is 81.063%. I will try to beat that accuracy by choosing the right model and tuning the hyperparameters.

Evaluation Metrics

Accuracy is a common evaluation metric for binary classifiers. We can evaluate the model/project by using the following equation, given equal weights to all prediction

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Predictions}$$

We take into consideration both true positives and true negatives here, since both have a positive impact on the success of Avrato's direct marketing campaign's success.

Other evaluation metrics include:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Since we only want to opt-in customers who will more likely respond to direct marketing campaigns, False Negative is preferred over False Positive. As such, increasing Precision is prioritized over increasing Recall.

References

[1] *Robert D. McFadden (14 January 2019). "Lester Wunderman, Father of Direct Marketing, Dies at 98". The New York Times.*