What Makes You Click *Buy Now*?:

A Look at the Effect of Demographic Factors on Receptiveness to Marketing

Eric Geisler

John Carroll University

May 09, 2024

Introduction

Data is growing at an unprecedented rate. Because of the rapid increase in the presence of the digital universe, there exists more data in the modern world than at any other point in history. Consequently, companies have more data on consumers than ever before. Simply put, more data means more opportunities to leverage data. Audience segmentation has become the norm in strategic marketing. Brands use demographic, socioeconomic, and lifestyle factors to target the consumers that are most likely to engage with their products and services. Thus, with a better understanding of what factors influence consumer behavior, companies can market in a more efficient manner.

To determine possible factors that have a large effect on this consumer behavior, a linear regression model was utilized to analyze the association between age, income, and number of children when predicting a positive response to a marketing campaign. When looking at the results of this test, income and number of children were determined to have significant predictive power for consumer responses to a marketing campaign.

All group members contributed equally to both the analysis and presentation.

Problem

Do certain demographic and financial factors relating to age, socioeconomic status, and family life help predict consumer responses to marketing campaigns? Specifically, do age, annual income, and number of children of an individual influence their response to a marketing campaign? This research will analyze these aforementioned variables and their relationship with corporate marketing. Before conducting any analysis, age, annual income, and number of children are hypothesized to significantly affect a consumer response to a marketing campaign.

Data

data is linked here: https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign

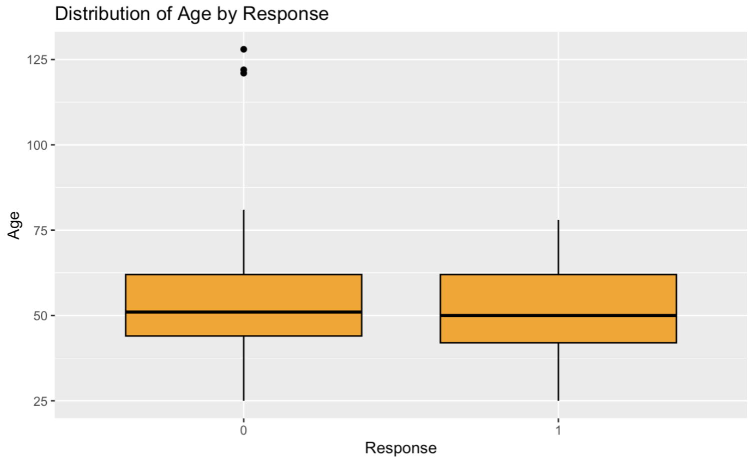
The data used for conducting this analysis came from a general survey of consumers and their behaviors. Information on 2240 consumers was gathered, with twenty-nine specific values for each individual. Specifically, the research will use age, annual income, and number of children in a household to predict consumer response to marketing campaigns. The original copy of the data contained the birth year of a consumer, denoted in a column named *Year_Birth*. Additionally, annual income was stored as a continuous variable in a column named *Income.* For the number of children, the original data contained columns *Kidhome* and *Teenhome*, which observed the number of young children and teenagers in the household of an individual. Lastly, consumer response to marketing campaigns was stored in a column named *Response*. Again, this is the response variable for this research. In *Response*, values of *0* corresponded to a negative response to marketing campaigns and values of *1* corresponded to a positive response to marketing campaigns.

Data Cleaning and Preprocessing

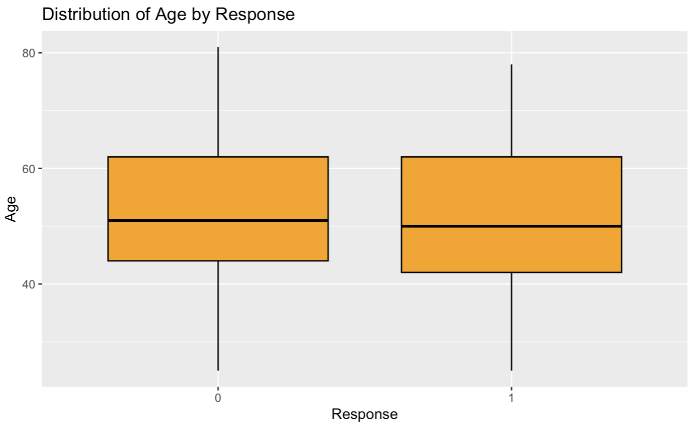
The process of data cleaning and feature engineering for the dataset used in this research was relatively simple. Firstly, all missing data in the dataset was removed. In order to get the age of each individual, each value in *Year_Birth* was subtracted from 2021 and stored in a new column *Age*. The number 2021 was utilized as the constant because the survey was conducted in the year 2021. Additionally, data transformation was necessary to get an all-encompassing measure for the number of children in a given household. The values of *Kidhome* and *Teenhome* were added and stored in a new column named *ChildrenHome*. The columns *Income* and *Response* were left unchanged.

Data Visualization

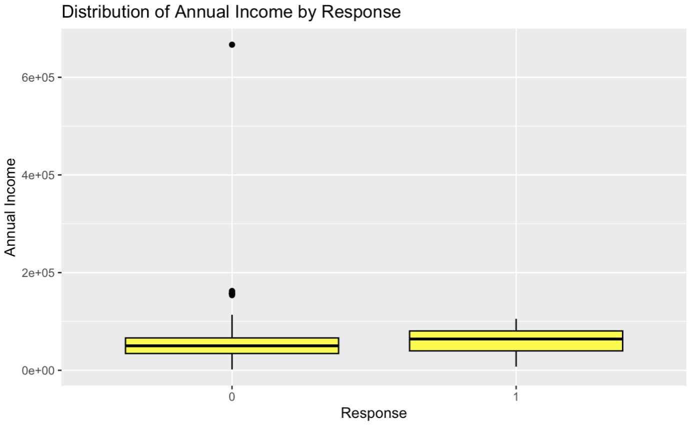
After data cleaning and feature engineering was complete, the distributions of each variable for negative and positive marketing responses were visualized in an attempt to glean insights. First, *Age* and *Response* were visualized in a box plot, seen below:



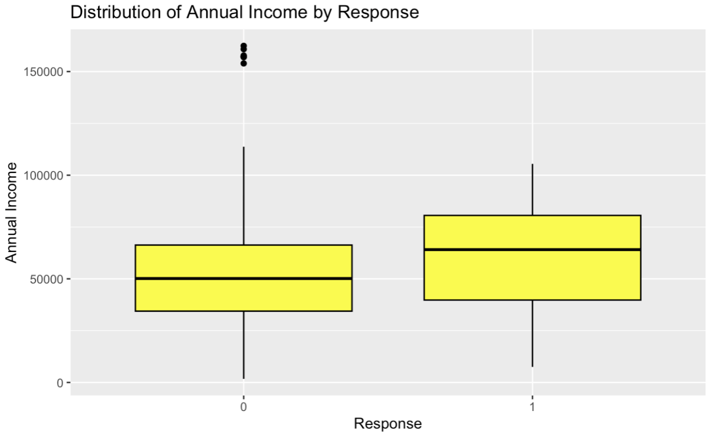
Before actually analyzing the visualization, the outliers were filtered out. Since these values were above 110, it was assumed that these were erroneous data points. The new visualization is shown below:



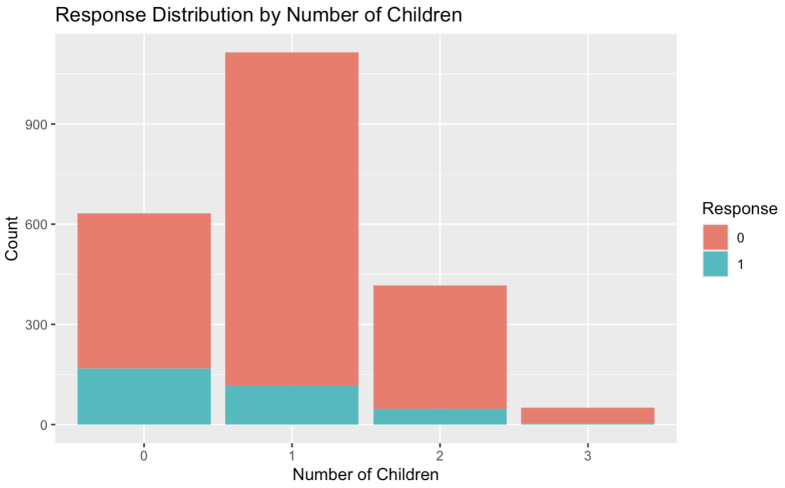
After removing outliers, patterns can be more easily observed. As the distributions of age for negative and positive marketing responses appear to be similar, it can be assumed that *Age* is not a significant predictor of *Response*. Moving on, the same visualization was created for *Income*, shown below:



Again, the removal of outliers was necessary before analyzing any trends. The new visualization is shown below:



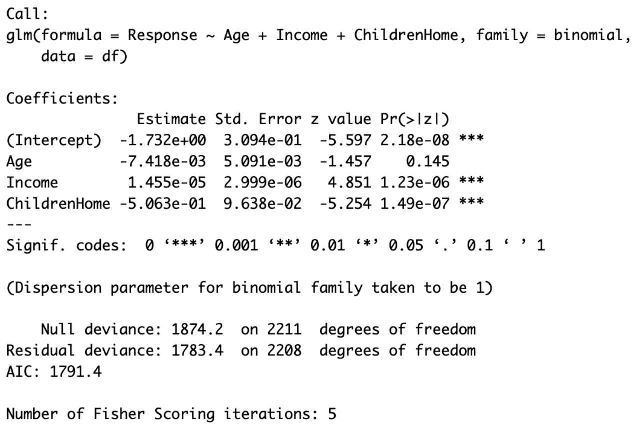
Evidently, the distribution of annual incomes for positive responses to marketing campaigns is higher than the distribution of annual incomes for negative responses to marketing campaigns. In other words, a higher annual income appears to be correlated with a higher probability of a positive response. From the visualization, *Income* seems to be a significant predictor of *Response*. Lastly, a bar graph was created to visualize distribution of *Response* for each number of children in a household in the dataset, seen below:



Based on the visualization, *ChildrenHome* also appears to be a significant predictor of *Response*. The proportion of positive responses appears to decrease for each additional child. Now that a cursory examination of all predictor variables has been conducted, a model can be created.

General Linear Model

The data is ready to be tested. A linear regression model will be utilized because the response variable is binary. For each variable, the null hypothesis being tested is that a given variable is not a significant predictor of the response to marketing campaigns. The results from the test are seen below:



```
Call:
glm(formula = Response ~ Age + Income + ChildrenHome, family = binomial,
    data = df)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.732e+00  3.094e-01  -5.597 2.18e-08 ***
Age          -7.418e-03  5.091e-03  -1.457    0.145
Income        1.455e-05  2.999e-06   4.851 1.23e-06 ***
ChildrenHome -5.063e-01  9.638e-02  -5.254 1.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1874.2  on 2211  degrees of freedom
Residual deviance: 1783.4  on 2208  degrees of freedom
AIC: 1791.4

Number of Fisher Scoring iterations: 5
```

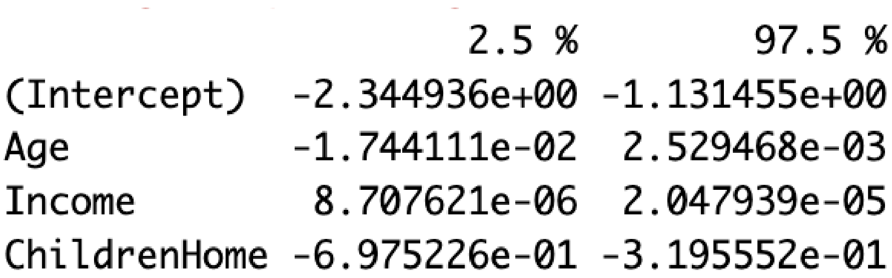
Examining the results of the linear regression model, there are two significant predictors of *Response*. The p-value for *Age* of .145 is above .05, and thus, at the .05 significance level, there is no evidence to suggest that age is a significant predictor of a response to marketing campaigns. The p-value for *Income* of .00000123 is below .05, and thus, at the .05 significance level, there is evidence to suggest that annual income is a significant predictor of a response to marketing campaigns. Lastly, the p-value for *ChildrenHome* of .000000149 is below .05, and thus, at the .05 significance level, there is evidence to suggest that the number of children in a household is a significant predictor of a response to marketing campaigns. The full model equation is as follows:

log(odds) = -1.732 + -0.007418*Age + 0.00001455*Income + -0.5063*ChildrenHome

In the absence of any predictor variables, there is a .177 chance of a positive response to marketing campaigns, calculated by exp(-1.732). For *Age*, we predict that the odds of an individual with one year higher of an age and the same income and number of children in the home are multiplied by exp(-0.007418), or .993. As the age of an individual increases, their odds of giving a positive response to marketing campaigns decreases (not significant). For *Income*, we predict that the odds of an individual with one dollar (very small unit) higher of an annual income and the same age and number of children in the home are multiplied by exp(0.00001455), or 1.00. As the annual income of an individual increases, their odds of giving a positive response to marketing campaigns increases. For *ChildrenHome*, we predict that the odds of an individual with one more child in their household and the same age and annual income are multiplied by exp(-0.5063), or .603. As the number of children in the household of an individual increases, their odds of giving a positive response to marketing campaigns decreases.

Confidence Interval

The confidence interval calculated for the model coefficients can be seen below:

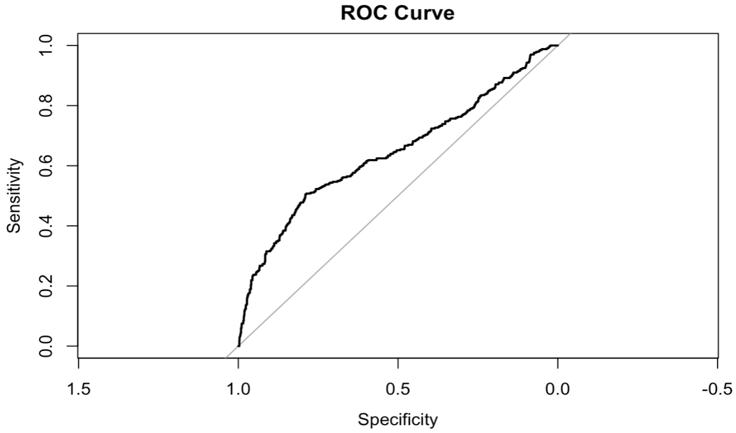


```
                    2.5 %        97.5 %
(Intercept)  -2.344936e+00 -1.131455e+00
Age          -1.744111e-02  2.529468e-03
Income        8.707621e-06  2.047939e-05
ChildrenHome -6.975226e-01 -3.195552e-01
```

The log(odds) for each value are shown in the results. For the following interpretations, this was undone by calculating exp(x) for each value. In the absence of other variables, the model predicts with 95% confidence that an individual's odds of providing a positive response to a marketing campaign are between .096 and .322. For *Age*, the model predicts with 95% confidence that the odds of an individual with one year higher of an age and the same income and number of children in the home are multiplied by a value in the range of .982 to 1.002. As the confidence interval contains 1, *Age* is not a statistically significant predictor of *Response*. For *Income*, the model predicts with 95% confidence that the odds of an individual with one dollar higher of an annual income and the same age and number of children in the home are multiplied by a value in the range of 1.000009 to 1.00002. Lastly, for *ChildrenHome*, the model predicts with 95% confidence that the odds of an individual with one more child in their household and the same age and annual income are multiplied by a value in the range of .498 to .726.

Model Accuracy

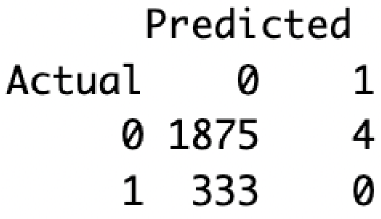
In order to evaluate the accuracy of the linear regression model with the given predictions, the ROC curve can be analyzed, shown below:



The area under the ROC curve is .6439. Thus, the model is better than random chance (.5) but clearly still not that accurate.

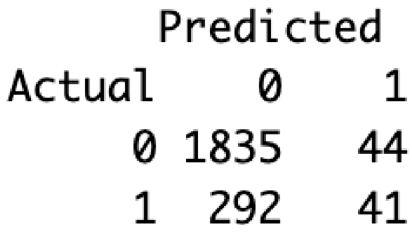
Predictions

In regards to model predictions, classification based on the model's predicted probabilities was first calculated with a threshold of 50%. Meaning, if a predicted probability was greater than 50%, the prediction was classified as a positive response. If it was not, then the prediction was classified as a negative response. The confusion matrix of the predictions is shown below:



```
         Predicted
Actual     0    1
     0  1875    4
     1   333    0
```

Problematically, the model did not make a single correct prediction of a positive response, and only predicted four positive responses in general. While the overall accuracy was 84.76%, the precision and recall were both 0%. Since the probability of a positive response was low in general, classification was redone with a new threshold of 30% for the predicted probabilities, resulting in the confusion matrix below:



```
         Predicted
Actual     0    1
     0  1835   44
     1   292   41
```

With the new 30% threshold for classification, the model had a better array of predictions. The accuracy actually improved to 84.81%, with a precision of 48.23% and a recall of 12.31%.

Conclusion

From this research, annual income and number of children in a given household have been determined to be significant predictors of a response to marketing campaigns. Accordingly, a corporate marketing strategy should focus on targeting an audience with high incomes and low numbers of children in the household. Marketing campaigns should not be generally focused towards a specific age group. However, there is more to discover. A model with the only significant predictors being annual income and number of children in a given household does not have great predictive ability for responses to marketing campaigns.

Future Work

Going forward, it would be beneficial to examine more variables that conceptually seem to be plausible predictors of responses to marketing campaigns. This can begin with other variables in the same dataset and later extend to more data. With the discovery of more significant predictors, the linear regression model can better predict these responses. Thus, the model could have more practical application to marketers looking to target their desired audience.

.