A Predictive Model for Marketing Campaign Responses

Martize T. Smith

Introduction

Marketing is one of the biggest reasons why a company succeeds or fails. With rapidly changing markets, business environments, competition and other forces continuing to grow a business is becoming more and more challenging. The companies that rise to the challenge and strategically execute in their marketing are the ones that survive and continue to grow while others go out of business. In a report from MarketSplash, they confirmed that 82% of businesses use content marketing as a technique to grow sales and brand themselves. In a different report from WordStream, 63% of businesses have increased their marketing budgets. A global authority in marketing, Hubspot, released a marketing report highlighting data-informed marketing strategies have become vital and companies using data will greatly benefit from it. Understanding how critical marketing is to a business and how data helps guide marketing activities that yield results, in this this project I examine a dataset of business customers to discover what factors contribute the most to the customer's response to the business marketing campaigns then construct a logistic regression predictive model to predict future campaign responses.

Audience

This predictive model is not a one size fits all but serves as a guide for companies who want to have their data science team explore how other data scientists are sourcing and analyzing to create insights from data to enhance greater marketing effectiveness in campaign responses. Benefactors of this project is not limited to the marketing department or data and marketing professionals but any person or group searching for viewpoints on deriving impactful interpretations that are then transformed into actionable and measurable activities for achieving marketing success though a higher response from marketing efforts. Businesses are actively looking for more competitive advantages and even a supreme unfair advantage so they can accomplish market dominance. Regardless of the desired end goal observing this project will help act as one of many resources that contributes to helping businesses find a way to get greater outcomes in their marketing goals.

Data Source

The dataset that was downloaded and used in this project is from Kaggle.com. Kaggle is at the time of this project the world's largest data science community with massive amounts of resources. Kaggle exist for the purpose of strengthening and establishing a greater understanding as well as application of the data science industry. This dataset from Kaggle is marketing data from a company. The dataset provides around 2205 customer's information and demographics along with other data like but not limited to the company's marketing campaigns responses and marketing channels. This marketing data is great and will serve as a way to construct a predictive model for future campaign responses. Overall, this data source fits the purposeful pursuit of using data science techniques within the marketing framework to discover and leverage optimum ways to achieve greater marketing results altogether.

Data

All data is from one company and it contains the following on the company's 2205 customers and other subjects:

- Income
- Kids at home
- Teenagers at home
- Number of Days Since Last Purchased
- Amount Spent on Wines (Last 2 Years)
- Amount Spent on Fruits (Last 2 Years)
- Amount Spent on Meat Products (Last 2 Years)
- Amount Spent on Fish Products (Last 2 Years)
- Amount Spent on Sweet Products (Last 2 Years)
- Amount Spent on Gold Products (Last 2 Years)
- Number of Deals Purchased

- Number of Website Purchases
- Number of Catalog Purchases
- Number of In-Store Purchases
- Number of Website Visits a Month
- Marketing Campaign 1
- Marketing Campaign 2
- Marketing Campaign 3
- Marketing Campaign 4
- Marketing Campaign 5
- If Customer Accepted Offer in Last Campaign
- Age
- Date of Customer's Enrollemt with the company
- Divorced
- Married
- Single
- Widow
- College Graduate
- High School Graduate
- Master's Graduate
- PhD Graduate
- Amount Spent in Total (Last 2 Years)
- Amount Spent on Regular Products
- Marketing Campaign Overall

This dataset was generally cleaned but I still used basic cleaning checks such as checking for missing values, improper data types, duplicate values and misspelled words. I didn't need to use everything in this dataset nor all the columns. There is more than one campaign so I focused on the first one so this makes campaign one my dependent variable, which will be represented as "Y". I choose five independent variables for this project and they are represented as "X". The five independent variables all have descent correlations to customers responding to campaign one.

Method

This Kaggle dataset is marketing information from a company. The ultimate goal here is to discover what factors contribute the most to the customer's response to the business first marketing campaign then construct a logistic regression predictive model to predict future campaign responses. I applied the cross industry standard process for data mining (CRISP-DM) to express an overview of the methodology present in this project, which is as follow: (1) investigate and select independent and dependent variables (2) Visualize the data and analyze (3) Identify high correlations among variables (4) Build logistic regression model to predict marketing campaign response (5) Evaluate the Model. As I focused on the company's first marketing campaign, this is my dependent variable. Additionally, I choose five independent variables, which are the Amount Spent in Total (Last 2 Years), Amount Spent on Regular Products (Last 2 Years), Amount Spent on Wines (Last 2 Years), income and Number of Catalog Purchases because they had the highest correlation to the first marketing campaign relative to other contributing factors for selecting these independent variables. After selecting the variables and conducting data cleaning and exploration the next phase was model creation. In data science and mathematics there are countless ways to construct predictive models. Here are a few basic common predictive techniques:

Classification Model:

Classifications are supervised machine learning techniques and the model attempts to predict the accurate identification of provided inputted data. During classification, the model is completely trained using the training data. Next, the model is assessed on test data before being utilized to perform prediction on fresh invisible data.

Clustering Model:

A clustering model divides data into separate categories based on similar characteristics. Next, the model utilizes the data from each group to distinguish enormous outcomes for each cluster. Furthermore, this model functions by implementing two common types of clustering. One type is hard clustering, which categorizes data by deciding whether each point entirely fits to a certain cluster. The second type is soft clustering, which pinpoints a probability to each data point verses dividing them into specific clusters.

Decision Tree:

A decision tree is an algorithm that diagrams various sources of data into a tree-like framework to illustrate the possible scenarios of output of diverse decisions. The decision tree displays diverse decisions into branches and then expresses possible scenarios underneath each decision. Businesses and organizations generally apply decision trees to govern the most critical variables in a specific dataset.

I decided to implement logistic regression in this project. In data science, the Logistic Regression is applied as a classification technique during machine learning. It incorporates a logistic function to model the dependent variable.

Data Wrangling & Cleaning

Since I had used a Kaggle dataset a good deal of data cleaning had already taken place. I confirmed the cleaning and may have done additional cleaning, which is listed below.

Checked for Duplicates

It can be common for data from all sizes to be likely to have duplicated values. Duplicates can sometimes come from human mistakes when human is entering the data or completing a form while making a mistake in the progress. Mistakes can also come from computers as data is being pulled from one source then put somewhere else.

Checked for Wrong Data Types

It is important to make sure data types are correct so when moving along the data science process that exploration and analysis is properly conducted. Generally, many numbers are the common data type that will have to be converted as data cleaning takes place. Frequently numbers are entered or identified as text but for data processing to take place these numbers have to appear as numerals.

Checked for Missing Values

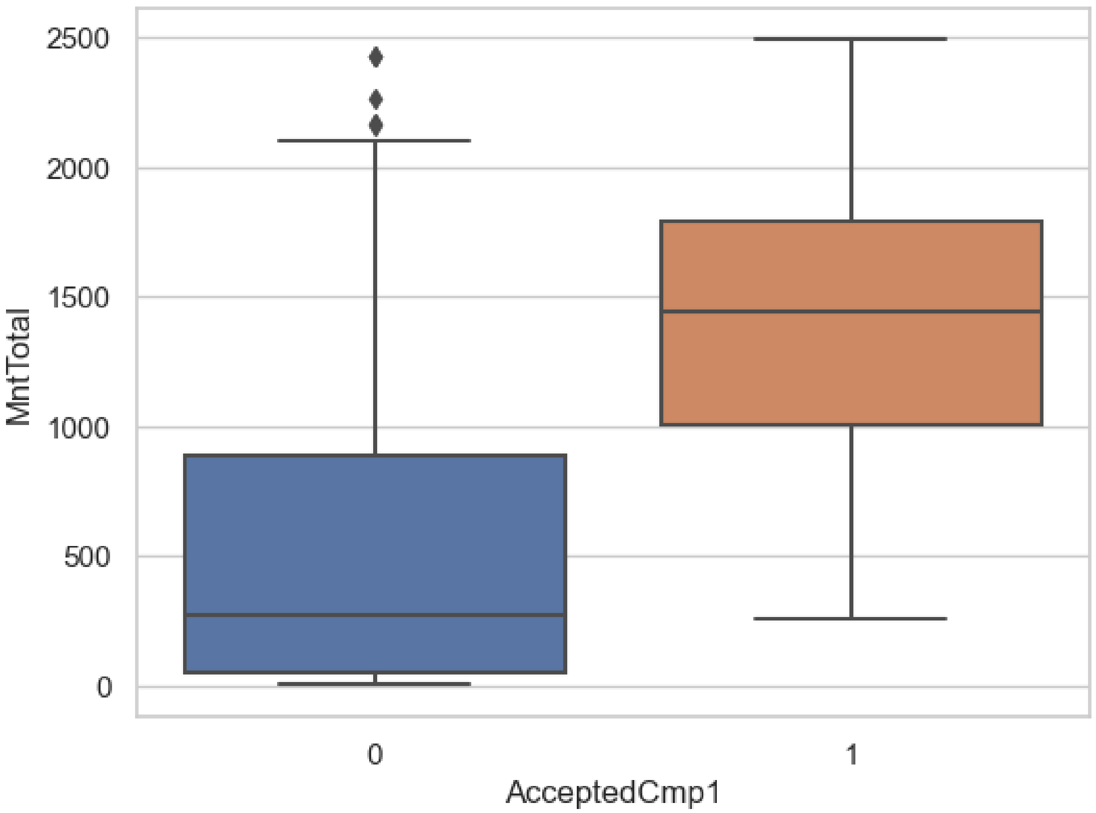
When observing datasets missing values are another common obstacle seen and must be dealt with. Two ways to handle missing values is to deleted them or replace the missing values. The right decision is based on what the overall project or purpose of the data science application is.

Checked for Irrelevant Data

Irrelevant data is sometimes in the dataset. Irrelevant data can give inaccurate or incomplete analysis. When examining data, the data professionals must know what is relevant and what is not before advancing to the data cleaning phase. An example is if you are investigating the age range of your customers, you would not need to add their email addresses.

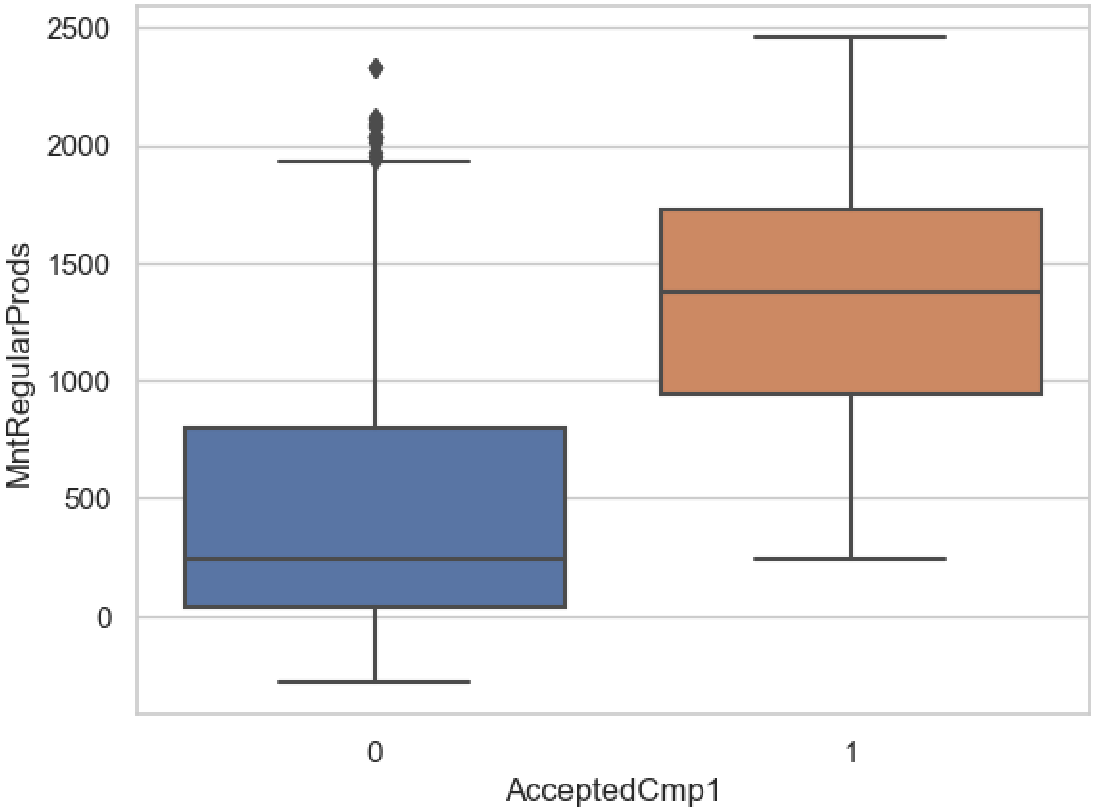
Exploratory Data Analysis (EDA)

During the EDA phase I used box plots for numerical correlation discovery and combination bar and line visuals because my independent variables were numerical and my dependent variable is categorical. Once I finished conducting the box plots visuals the data shows that in marketing campaign one more customers bought products after responding compared to those who bought from not responding. This is seen in figure one. In figure two, the same occurred, more customers bought products after responding compared to those who bought from not responding. In both figure one and two the number zero indicates customers did not respond to the marketing campaign while number one indicates customers who did respond. Regardless of whether the customers responded or not we see the buying activity from both as they eventually bought at some point.



[1]

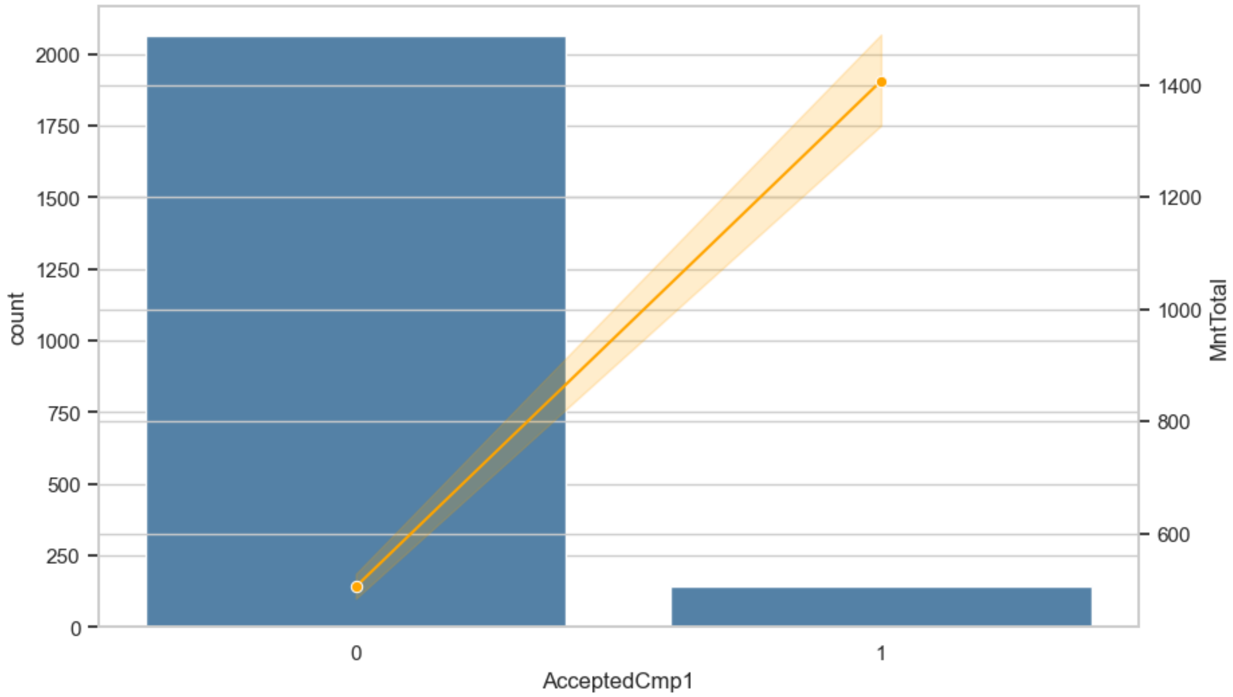
[1] Box-Plot of the number of customers who responded and those who didn't respond to the first marketing campaign compared to the average total amounts spent in total on all products in the last 2 years



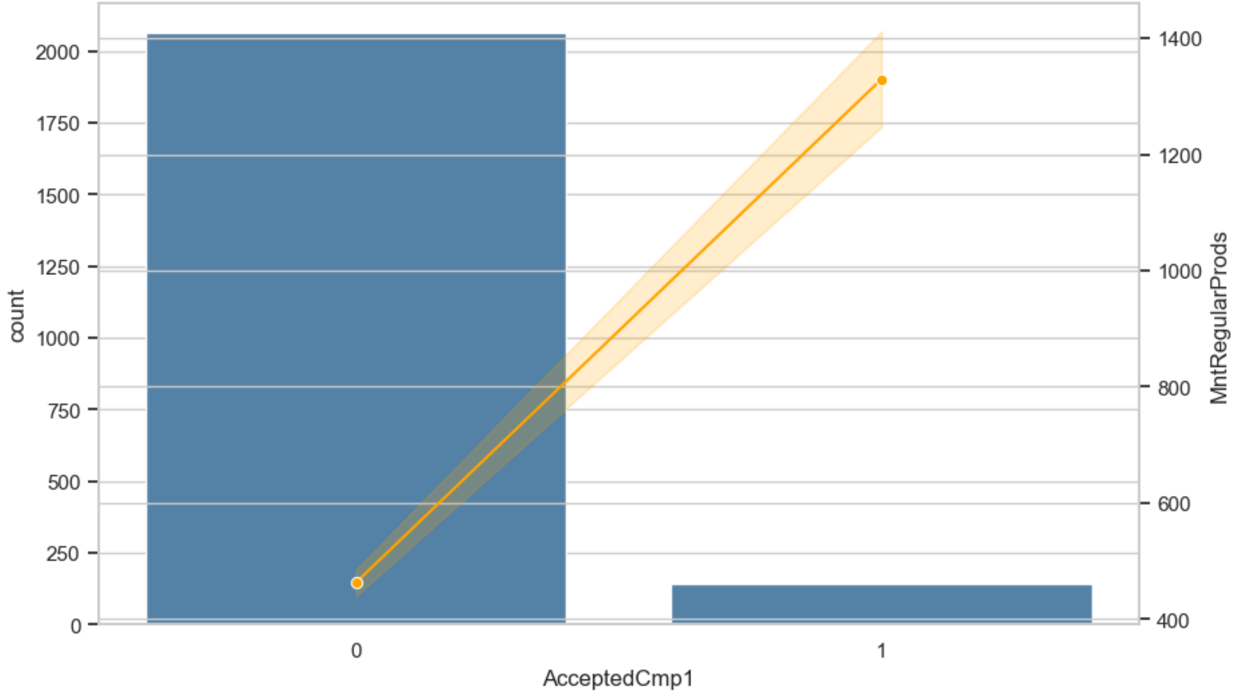
[2]

Below, in figure three, we have a combination bar and line visual that shows the trend or correlation of the number of customers who responded and those who didn't respond to the first marketing campaign compared to the total amounts spent in total on products in the last 2 years. The trend is showing that on average customers who responded to the campaign spend $1400 while customers who didn't respond spent around $150 from the last two years.

[2] Box-Plot of the number of customers who responded and those who didn't respond to the first marketing campaign compared to the average amounts spent on regular products in the last 2 years



3



4

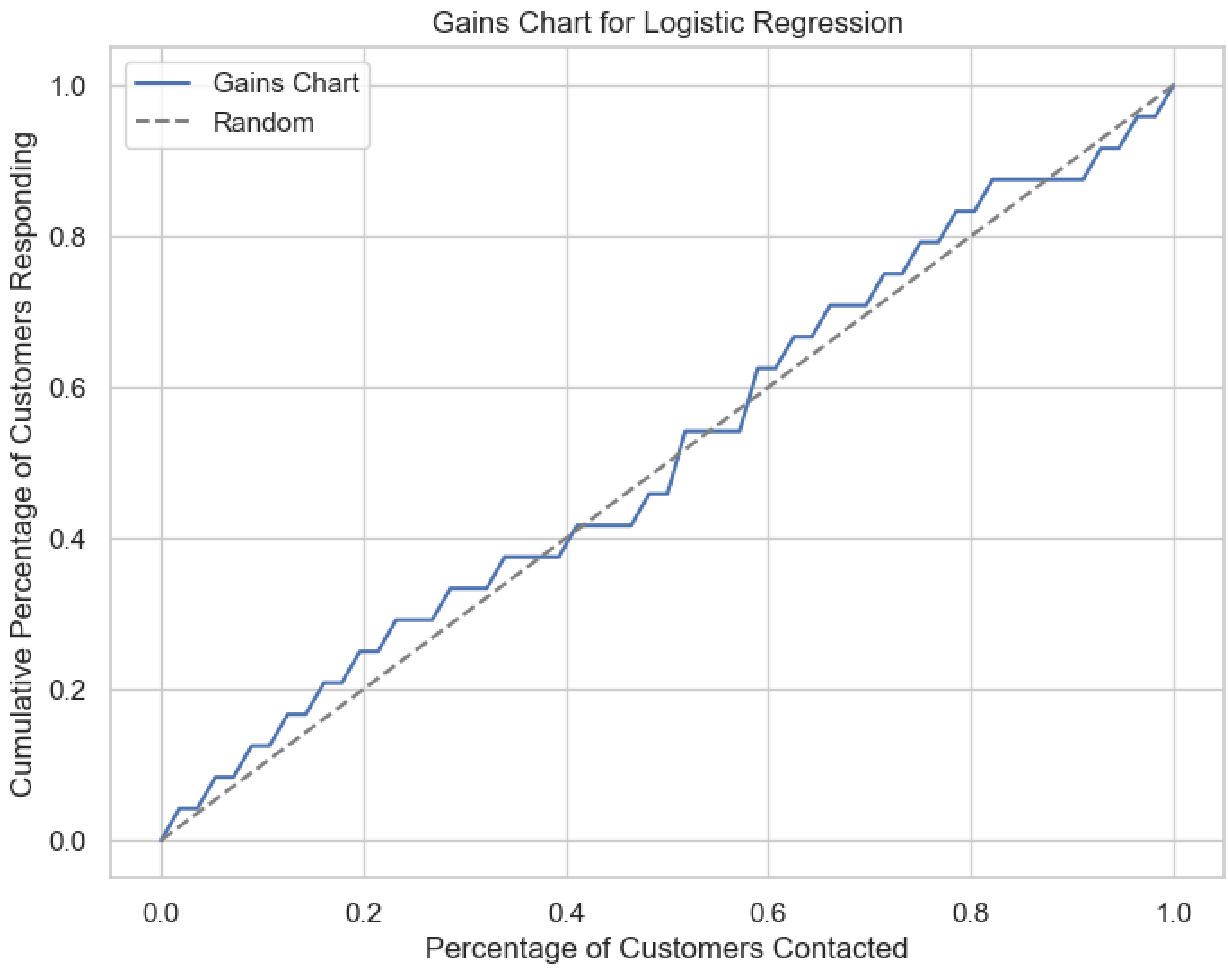
[3] Combination-bar graph for the number of customers who responded and those who didn't respond to the first marketing campaign compared to the total amounts spent in total on all products in the last 2 years

[4] Combination-bar graph for the number of customers who responded and those who didn't respond to the first marketing campaign compared to the amount spent on regular products in the last 2 years

Above, in figure four, we see a combination bar and line visual that shows the trend or correlation of the number of customers who responded and those who didn't respond to the first marketing campaign compared to the amount spent on regular products in the last 2 years. The trend is showing that on average customers who responded to the campaign spend $1300 while customers who didn't respond spent just under $150 from the last two years. In this overall report, I only plotted box-plots involving campaign one compared to total amounts spent in total on products (Figure 3) and amounts spent on regular products (Figure 4) in the last 2 years. I only showed these two to provide a snapshot and overview of my analysis of what I did when comparing marketing campaign one to all of my independent variables.

Model & Evaluation

In the modeling phase, I choose to use a logistic regression model as I did the train test split on the selected data chosen. For evaluation, I use a gains chart as well as the accuracy formula. Logistic regression approximates the chance of an event taking place using given independent and dependent variables. A gains chart is a model evaluation technique that measure how much improvement someone can assume to do with the predictive model comparing without a model. Accuracy is the fraction of predictions the model got correct. Formula: Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

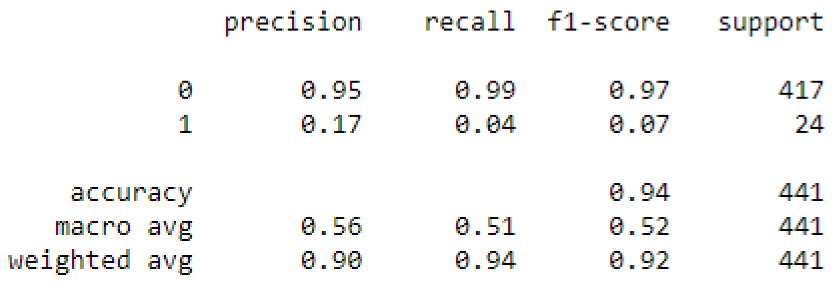


[5]

The gain chart in figure five, the dashed line expresses “no gain”, meaning, what companies would expect to achieve by marketing to customers at random. The nearer the cumulative gains line is to the top-left corner of the chart, the bigger the gain; the higher the amount of the customer’s responding that are reached for the lower amount of customers contacted. As far as accuracy, the accuracy scored 94%, good range is from 70% to 90%, but this doesn’t mean that the model is supreme.

Predictions

[5] Gains chart of the Logistic Regression Model



```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       417
           1       0.17      0.04      0.07        24

    accuracy                           0.94       441
   macro avg       0.56      0.51      0.52       441
weighted avg       0.90      0.94      0.92       441
```

[6]

Figure six is the classification report from the logistic model I created. This report shows the precision, recall, F1, and support scores for the model. This report helps show how the model will perform. Also, from classification report we can better gauge how I should improve the model for future usage.

Future Improvements

In the future, to improve this model I would likely add more data from current and new incoming customers to help give greater insights. Another improvement would be to use more advanced evaluation techniques to uncover more areas that could be targeted to enhance model. Lastly, another improvement is using different algorithms to broaden the scope of the predictive model’s performance then compare all to find the optimal model to act as the model of choice.

[6] Classification Report of Logistic Regression Model