James Griffin

Stat 536

Targeted Marketing Report

**Introduction**

Part of the reason The Super Bowl has the highest viewing of any event in the United States every year is for the commercials. Families who don’t know anything about football or the two teams gather in front of their tv’s to see what clever marketing and advertising strategies will be shown each year. Commercials like the original, “Jake from State Farm” can become cultural references. Targeting advertising and marketing messages at the correct audience and demographic is important for company growth. Figuring out who to target and how to target them is important to make sure the strategy has a high conversion rate. This data set seeks to help understand what makes certain strategies more effective, what demographics should be targeted, with the simple binary outcome variable if the person signed up for the credit card.

Our data set contains data on the demographics of each person in our data set, their age, marital status, type of job, amount of education, how they were contacted and how many times, things of that nature. Other than age, these data are almost strictly categorical. This coupled with the fact that our outcome variable had relatively very few successful outcomes were the tricky part of this analysis that we will discuss below.

A picture containing chart

Description automatically generated

**Methodology**

Given the binary outcome of the data, we will have to approach this problem from a logistic regression standpoint. There assumptions we need to satisfy, are the following:

* Outcome variable must be binary with 1 being the desired outcome
* No perfect multicollinearity or cereal auto correlation

When creating our feature set from the categorical data that we had. We wanted to limit the number of features we included. For example, with our education field, we had several levels: university.degree, high.school, basic.9y, professional.course, basic.4y, basic.6y, unknown and illiterate.

We grouped the basic four, six and nine years of education groups together and included the illiterate populations. This not only cut down on the number of features, but it grouped people in similar levels of human capital. These people on average only have the skills to be day laborers of some sort. We kept the high school, university and professional courses separate, and left out the unknown population, so as to preserve our assumption of no perfect multi collinearity. As part of this process, we created crosstab charts and for many of the variables[[1]](#footnote-1) to helps our intuition as we decided how to group features.

We followed this methodology with the rest of our categories, always leaving out our unknown observations, and grouping similar or low support groups together. The list of all features used can be found in the appendix.

Chart

Description automatically generated

We wanted to see how our assumption of multicollinearity and seral correlation was holding up. In large part, we can see that the correlation between our variables isn’t very significant.

**Models and Assumptions**

Mathematically, the logistic function can be written as

And we define our x to be equal to

Where are the variables used in the regression and is a 1if true and a 0 else for all categorical variables. This x value is then plugged into the logit function, which for any value x, will give us a number between (0,1). This y value will be the probability of opening an account. If the probability higher than 0.5, then the default classification is to give it a 1, else the prediction receives a 0. This is a hyper parameter that can be changed, however for this analysis we did not.

As part of this analysis, we needed to find which characteristics of consumers were more likely to take out a new credit card, this will define the betas used in the x equation above. However, before we proceeded with our variable selection, we needed to look at the disparity of support in our outcome variable

Because we have roughly 90% of our data who did not become a customer, if we had simply gave a model this data, it could pick 0, almost every time, and be a very accurate model. So, we determined it would be necessary to oversample the data to create evenly weighted classes, and we did this through the SMOTE algorithm.

SMOTE or Synthetic Minority Oversampling technique works by creating synthetic samples from the minority class, and then randomly choosing one of the k nearest neighbors. It then randomly tweaks that observation to create a new, observation that is unique. This is different from traditional bootstrapping, because while we do sample the minority data, and as we synthetically create more observations, it is with replacement, we don’t end up with any duplicates in the oversample set. We also only over sampled our training data, because by not over sampling the test set, none of the information in the test set was used in training. This will preserve the integrity of the test

When deciding on the proportion of oversampled data we wanted to create when compared to the non-subscription data. At first, we thought that creating perfectly balanced data sets would be best. This would give us an unbiased estimator; however, we do want a slightly biased estimator. Thew fact that most people don’t become customers means that in reality the probability is much less than 50%. We first created observations such that we would have 1/3rd as many observations for the subscription class, and compared it with balanced data. Comparisons of both of these approaches will be discussed below.

Having created more balanced data, we used the Recursive Feature Elimination or RFE to help select our most important features. RFE is based on repeatedly building a model to choose the best and worst features, then placing those in bins, and going again recursively. This is effective because essentially the model considers smaller and smaller feature sets, finding which features are able to contribute most to the model.

Luckily enough, for both the unbalanced and balanced data set, we got the same list of features.

['basic\_edu', 'high\_school', 'college', 'pro\_course', 'single', 'married', 'job\_self\_emp',

'has\_housing', 'no\_housing', 'contact\_call', 'contact\_summer', 'contact\_spring', 'contact\_fall',

'previously\_contacted']

These because the beta coefficients in our x equation above, which was used in the logistic regression as previously discussed.

**Results**

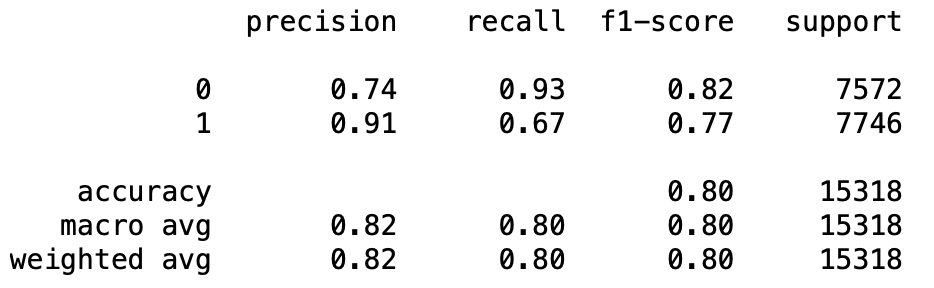
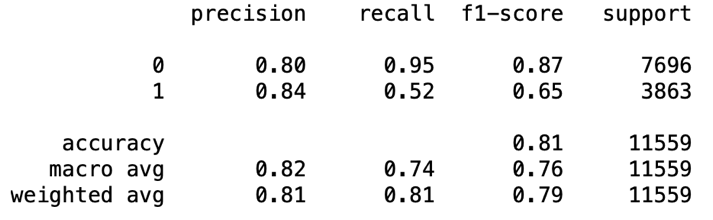
Shown below is the result for the balanced data retrogression. When we compared to the 1/3rd subscription 2/3rd non subscription regression[[2]](#footnote-2), we noticed that the pseudo r-squared was about 3% lower for this regression. We also noticed that we ran our predictions and found the accuracy, our accuracy was .80 vs .83 with the unbalanced data.

Shape

Description automatically generated with medium confidence

One might think that this would mean we should use the unbalanced data set, however when we investigate further and look at the types of errors we were getting, we get the following comparisons.

Unbalanced Balanced



We can see that the precision for our subscribing is substantially higher. We remember that precision is defined as: , while recall is [[3]](#footnote-3). Because we care about if someone actually subscribes, more than a few percentage points on overall accuracy or a few points on the pseudo r-squared, it is better if we use balanced data to train with, so that the model is unbiased in prediction.

When looking at what strategies are most effective, in particular if it is better for a company to call or use social media, we could only use one of these variables in the regression because each person was either one or the other. When we used contract\_call that coefficient was negative and equal to -1.65, when we used contact\_social\_media, the coefficient was positive and equal to 0.828. So, calling people appears to make them less likely to subscribe than social media. Lastly, we wanted to see if the repeated contacts had a positive effect. When looking at this, we saw that the vast majority of people had not been contacted. There were also people who had been contacted that day, so we couldn’t fill those null values with zeros. Instead, we turned this into a category as well, if someone had been previously contacted yes or no. This also had a very positive coefficient that was statistically significant.

In the appendix there are plots that show our histogram of predictions. This was also reassuring given the objectives of the analysis, because we are very sure when someone signs up, we are very confident in that. Our no subscribe predictions were more spread out which was san interesting take away.

**Appendix**

**Chart

Description automatically generated with medium confidence**Chart, histogram

Description automatically generated

**Chart

Description automatically generated**Chart, bar chart

Description automatically generated

**Chart, bar chart

Description automatically generated**Chart

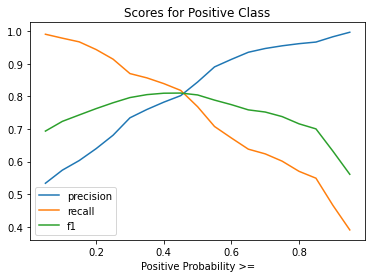
Description automatically generated

Regression of unbalanced populations, 1/3 vs 2/3

**Shape

Description automatically generated with medium confidence**

**Chart, histogram

Description automatically generated**

**A picture containing line chart

Description automatically generated**

1. See Appendix [↑](#footnote-ref-1)
2. See Appendix [↑](#footnote-ref-2)
3. What fraction of all positives do we actually predict? [↑](#footnote-ref-3)