Capstone Project Final Report

Predicting the ultimate customer in Portuguese bank marketing

Kimberly Kaufman

April 17, 2019

### Abstract

This project aims to predict which types of customers can be expected to subscribe to a term deposit product at a bank, and therefore should be targeted in current and future telemarketing campaigns. By analyzing 2.5 years’ of direct marketing campaign data from an unnamed Portuguese banking institution (via the UCI Machine Learning Repository), we can predict which subset of customers can be expected to purchase this product. This information provides valuable marketing insights that the banking industry can use to better target their potential customers.

### Analysis Approach

The dependent variable we are trying to predict is listed as “y”, which is categorical, binary, and consisting of two possible values: “yes” and “no.” The former indicates that the customer did subscribe to the term deposit product, the latter indicates the customer did not subscribe to the term deposit product. A subset of the other 20 available variables will be used in a logistic regression model in order to predict which customers can be expected to subscribe to a term deposit, represented by a “yes” observation of the y variable.

### Data

This project uses the Bank Marketing Data Set from the UCI Machine Learning Repository, available [here](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) via the UCI archive. The data was originally obtained and condensed for a published study in the June 2014 issue of Decision Support Systems1 that analyzed the success of telemarketing calls for selling bank long-term deposits. The data was collected from 2008 to 2013, therefore including the effects of the financial crisis, and was then reduced from 150 to 22 features.

All banking data is available as four CSV files, the largest of which will be used for this analysis: bank-additional-full.csv. The full data set consists of 41,188 observations of 21 variables. Each observation represents a unique client.

The attributes of the training dataset include:

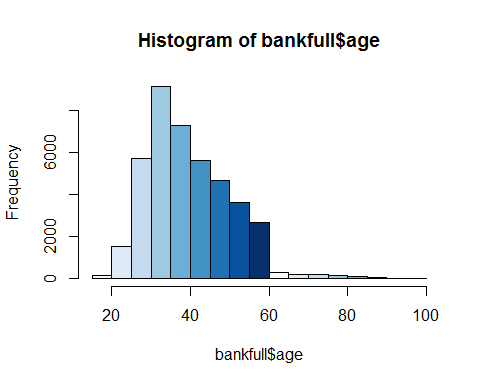
* 7 customer demographic attributes
  + Age, job, marital, education, default, housing, loan (categorical)
* 8 marketing campaign attributes
  + Contact, month, day of week, poutcome (categorical)
  + Duration, campaign, pdays, previous (continuous; numeric)
* 5 social and economic context attributes
  + Emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed (continuous; numeric)
  + Note: these abbreviations represent employment variation rate, consumer price index, consumer confidence index, Euribor 3 month rate, and number of employees, respectively.
* 1 output variable / target
  + Y (categorical; binary)

##### **Full Data Structure**

## 'data.frame': 41188 obs. of 21 variables:  
## $ age : int 56 57 37 40 56 45 59 41 24 25 ...  
## $ job : chr "housemaid" "services" "services" "admin." ...  
## $ marital : chr "married" "married" "married" "married" ...  
## $ education : chr "basic.4y" "high.school" "high.school" "basic.6y" ...  
## $ default : chr "no" "unknown" "no" "no" ...  
## $ housing : chr "no" "no" "yes" "no" ...  
## $ loan : chr "no" "no" "no" "no" ...  
## $ contact : chr "telephone" "telephone" "telephone" "telephone" ...  
## $ month : chr "may" "may" "may" "may" ...  
## $ day\_of\_week : chr "mon" "mon" "mon" "mon" ...  
## $ duration : int 261 149 226 151 307 198 139 217 380 50 ...  
## $ campaign : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ pdays : int 999 999 999 999 999 999 999 999 999 999 ...  
## $ previous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ poutcome : chr "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...  
## $ emp.var.rate : num 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...  
## $ cons.price.idx: num 94 94 94 94 94 ...  
## $ cons.conf.idx : num -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...  
## $ euribor3m : num 4.86 4.86 4.86 4.86 4.86 ...  
## $ nr.employed : num 5191 5191 5191 5191 5191 ...  
## $ y : chr "no" "no" "no" "no" ...

### Data Wrangling

This data set is relatively clean due to its use in a previous analysis, so minimal data wrangling was required. The data was checked for missing and NA values, and none were found. There were, however, a handful of outliers in our numeric variables. As an example, let’s take a look at a frequency histogram of the “age” variable, which represents the age in years of the client.



We can see from looking at this visualization that there is a long tail of outliers to the right of the majority of our distribution. In order to get a closer look at where the outliers begin, we can observe the range of ages by percentile.

##### **Age Variable by Percentile**

## 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75%   
## 26 28 30 31 32 33 34 35 37 38 40 41 43 45 47   
## 80% 85% 90% 95% 100%   
## 49 52 55 58 98

Since we can see that the large outliers indicated in the histogram occur after the 95th percentile, we can consider using a capping/flooring approach to outliers for this exercise. After performing this type of exploratory analysis on the rest of the numeric variables, caps were set at the 5th and 95th percentiles, and all outliers were imputed with those caps.

Due to the relatively even mix of continous and categorical variables, one-hot encoding was considered for categoricals in order to provide more numerous and accurate modeling possibilities. However, most of our categoricals had a large number of possible values, resulting in 29 extra dummy variables and 38 total variables. As this amount of variables proved to be unmanageable, an alternate approach was required instead. Let’s take a look at the “job” variable as an example, which represents the client’s profession.

##### **Job Variable by Unique Values**

##   
## admin. blue-collar entrepreneur housemaid management   
## 10422 9254 1456 1060 2924   
## retired self-employed services student technician   
## 1720 1421 3969 875 6743   
## unemployed unknown   
## 1014 330

We can see that there are 12 unique values to this variable, though it is unlikely we will need to utilize all of them in the model. Working from the smallest number of observations up, we can group the less significant values such as unknown, student, unemployed, housemaid, self-employed, entrepreneur, retired, and management into one consolidated group labeled “other.” This reduces our variable to 5 unique variables, which is then converted to a factor with 5 levels. This procedure was performed on all categorical variables with more than a handful of unique values.

After these steps, the data was also checked for variables with zero variance, near zero variance, and high correlation. A quick check for NZV revealed one variable with zero variance. Our numerical variables alone were checked for highly correlated variables, which returned a total of three variables. All four of these NZV and correlated variables were removed in order to prevent them from slowing down or skewing our model.

##### **Variables with Zero and Near Zero Variance**

## freqRatio percentUnique zeroVar nzv  
## age 1.059140 0.080120423 FALSE FALSE  
## job 1.036269 0.012139458 FALSE FALSE  
## marital 2.154910 0.009711566 FALSE FALSE  
## education 1.278823 0.012139458 FALSE FALSE  
## default 3.790625 0.007283675 FALSE FALSE  
## housing 1.158630 0.007283675 FALSE FALSE  
## loan 5.433739 0.007283675 FALSE FALSE  
## contact 1.737836 0.004855783 FALSE FALSE  
## month 1.573780 0.012139458 FALSE FALSE  
## day\_of\_week 1.012802 0.012139458 FALSE FALSE  
## duration 1.015534 1.743226182 FALSE FALSE  
## campaign 1.669063 0.016995241 FALSE FALSE  
## pdays 0.000000 0.002427892 TRUE TRUE  
## previous 6.322311 0.004855783 FALSE FALSE  
## poutcome 8.363829 0.007283675 FALSE FALSE  
## emp.var.rate 1.767639 0.019423133 FALSE FALSE  
## cons.price.idx 1.161257 0.046129941 FALSE FALSE  
## cons.conf.idx 1.161257 0.043702049 FALSE FALSE  
## euribor3m 1.097589 0.529280373 FALSE FALSE  
## nr.employed 1.902273 0.019423133 FALSE FALSE  
## y 7.876724 0.004855783 FALSE FALSE

##### **Highly Correlated Variables**

## [1] "euribor3m" "emp.var.rate" "nr.employed"

### Predictive Model Creation

Because the observations were all representative of unique clients and independent of each other, the data set was easily partitioned into training and test sets using a 90%/10% split, respectively.

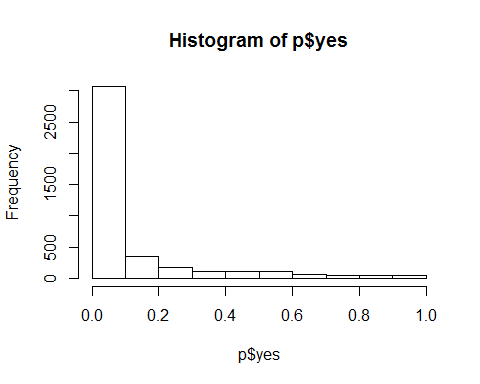
A generalized linear model, the simplest form of logistic regression available in the Caret package, was used in training. It was fitted on a “ROC” metric using a cross-validation setup with k = 5 as our number of folds. The model was then used to make predictions on the test set, providing outputs in the form of probabilities. See below for the respective model coefficients.

##### **Summary of Model Coefficients**

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2313 -0.3460 -0.2128 -0.1359 3.1781   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.569e+01 5.080e+00 8.993 < 2e-16 \*\*\*  
## age 3.039e-03 2.471e-03 1.230 0.218809   
## `jobblue-collar` -4.432e-01 7.761e-02 -5.711 1.12e-08 \*\*\*  
## jobother 8.830e-02 5.576e-02 1.584 0.113280   
## jobservices -3.545e-01 8.737e-02 -4.058 4.96e-05 \*\*\*  
## jobtechnician -1.748e-01 7.120e-02 -2.455 0.014101 \*   
## maritalmarried -2.524e-02 6.819e-02 -0.370 0.711253   
## maritalsingle 2.887e-01 7.674e-02 3.762 0.000169 \*\*\*  
## maritalunknown -3.940e-02 3.988e-01 -0.099 0.921291   
## educationhigh.school 3.242e-02 8.066e-02 0.402 0.687757   
## educationother 2.877e-01 7.707e-02 3.733 0.000189 \*\*\*  
## educationprofessional.course 1.443e-01 9.233e-02 1.563 0.118128   
## educationuniversity.degree 1.021e-01 7.882e-02 1.295 0.195360   
## defaultunknown -6.479e-01 6.518e-02 -9.941 < 2e-16 \*\*\*  
## defaultyes -8.644e+00 1.377e+02 -0.063 0.949947   
## housingunknown 5.122e-02 1.379e-01 0.372 0.710256   
## housingyes 2.725e-02 4.156e-02 0.656 0.512048   
## loanunknown NA NA NA NA   
## loanyes -7.159e-02 5.752e-02 -1.245 0.213277   
## contacttelephone -9.733e-01 7.968e-02 -12.215 < 2e-16 \*\*\*  
## monthjul 4.724e-01 8.642e-02 5.467 4.59e-08 \*\*\*  
## monthjun 1.514e+00 9.986e-02 15.166 < 2e-16 \*\*\*  
## monthmay 2.304e-01 8.827e-02 2.610 0.009054 \*\*   
## monthother 1.075e+00 7.410e-02 14.504 < 2e-16 \*\*\*  
## day\_of\_weekmon -9.556e-02 6.651e-02 -1.437 0.150773   
## day\_of\_weekthu -4.359e-02 6.445e-02 -0.676 0.498836   
## day\_of\_weektue 4.623e-02 6.553e-02 0.706 0.480442   
## day\_of\_weekwed -1.720e-02 6.581e-02 -0.261 0.793775   
## duration 6.437e-03 9.612e-05 66.966 < 2e-16 \*\*\*  
## campaign -7.157e-02 1.440e-02 -4.971 6.66e-07 \*\*\*  
## previous 2.341e-01 6.306e-02 3.712 0.000206 \*\*\*  
## poutcomenonexistent NA NA NA NA   
## poutcomesuccess 2.471e+00 9.058e-02 27.283 < 2e-16 \*\*\*  
## cons.price.idx -5.021e-01 5.470e-02 -9.180 < 2e-16 \*\*\*  
## cons.conf.idx 8.538e-02 5.871e-03 14.543 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26099 on 37069 degrees of freedom  
## Residual deviance: 16756 on 37037 degrees of freedom  
## AIC: 16822  
##   
## Number of Fisher Scoring iterations: 10

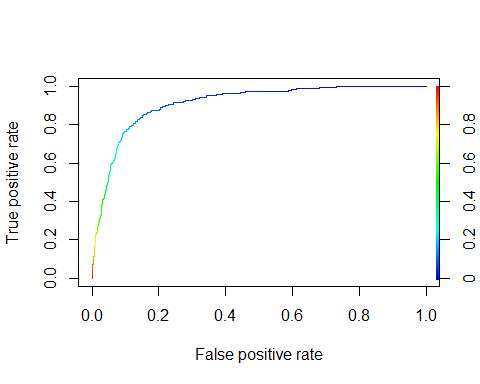
### Model Evaluation

The results of our predictions can be seen below. An initial look at the test prediction probabilities (p) provided a reasonable looking frequency histogram, with a majority of the observations at a very low probability (indicative of a “no” in the y variable, reflective of what we saw in the training data).



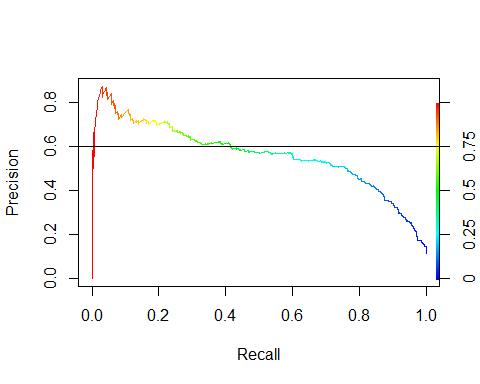
Evaluation of model performance was performed in a couple of different ways: ROC and a precision/recall curve. These approaches were exploratory methods to visualize the different ways to determine threshold value for our predictions.

The first approach used was an ROC curve.



We can see from this curve that the optimal balance of true positives and false positive occurs where the blue line begins to darken, at about 0.2. From a quick check of the AUC, we see that the area under the curve is about .9114, which is much higher than 0.5. So, this model adequately differentiates between “yes” and “no.”"

The second approach involved plotting a curve of precision vs recall.



This curve shows an optimal balance of precision and recall at 0.6 precision, 0.4 recall, the green area of the curve, giving us a threshold value of 0.5, very different from the 0.2 yieled by the ROC approach.

In this case, we are solving a marketing problem in trying to determine which types of customers to target. Since this problem is more concerned with marketing to the most inclusive and correct subset of customers, it is focused on determining the optimal number of true positives. For this reason, the ROC curve is the best approach for thresholding since it best allows us to prioritize this measure.

### Conclusion and Recommendations

In order to predict the target customers for future Portuguese bank telemarketing campaigns, a generalized linear model with a threshold value of 0.2 is sufficient to predict customers who will subscribe to a term deposit option with an accuracy rate of 83%. See below for model results against the test set.

##### **Confusion Matrix**

|  |  |  |
| --- | --- | --- |
| y | False | True |
| no | 347 | 3307 |
| yes | 352 | 112 |

We can see from this confusion matrix that we have 112 true positives, 352 false positives, 3307 true negatives, and 347 false negatives. That gives us a sensitivity rate of 24.4%, a specificity rate of 90.4%, and a total accuracy rate of 83.0%.

Based on the accuracy of the model created and tested in this study, implementation of a standard logistic regression for predicting potential customers to subscribe to a term deposit option might be recommended. The model created in this study generates a prediction that is 83% accurate. However, the model has a much higher specificity than sensitivity rate, meaning it is more successful at detecting true negatives than true positives. Given these statistics, the best use of this model would be to rule out all predicted “no” values from the y value, and focus future marketing campaigns on the remaining subset of clients.

Ideas for further research on this topic may include comparing additional and more complex forms of logistic regression models available in the Caret package.

### References

1. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014