**Predicting Bank Marketing Campaign Success Using Logistic Regression with Feature Selection and Cross Validation**

Ian Kinskey, Jack Rasmus-Vorrath, and Alice Karanja

MSDS 6372 Applied Statistics: Inference and Modeling

Section 403

August 18, 2017

**Bank Marketing Data Set**

**Problem Statement**

Direct Marketing is the practice of delivering promotional messages directly to current or prospective customers on an individual basis rather than using a mass medium. Predictive models are a great tool in analyses used to assess and increase the effectiveness of such marketing campaigns.Logistic regression remains one of the most popular techniques used to predict customer behavior. Millions of dollars are spent annually on marketing activities that utilize logistic regression models. Therefore, it is essential to build robust logistic models that have strong predictive ability for a successful direct marketing campaign. Using some combination of the 16 predictor variables in the data, how well can we predict the probability of a customer subscribing to a term deposit from the bank in question using a logistic regression model?

**Data Set Description**

This data was obtained from the UC Irvine Machine Learning Repository, and relates to the direct marketing campaigns of a Portuguese banking institution attempting to get its clients to subscribe to a term deposit. The marketing campaigns were conducted by making multiple phone calls to the clients. The client responses and predictor variable information are used to assess whether the subject will subscribe to the bank term deposit or not.

The dataset has 16 predictor variables (categorical or numeric) and a (binary) response variable consisting of either “yes” or “no” to the term deposit subscription. Some of the factors included in this study are demographic (age, job, education, marital status); others concern information on customer financial history (whether they have defaulted on a loan before) as well as their current financial status (whether they have a home loan or personal loan, and the average annual balance of their account). There is also information about the marketing campaign process (how long since the bank contacted the customer, whether it was by mobile phone or landline, and the duration of the conversation). All of this information is compiled and used to predict the response of the customer to the offer of a term deposit with this bank. In preparation for the analysis, we converted all factor predictor variables to dummy variables with the corresponding number of levels.

This data set can be obtained at the following URL: <https://archive.ics.uci.edu/ml/datasets/bank+marketing#>.

**Constraints and Limitations**

This marketing dataset *observes* information on current and prospective customers of the bank, suggesting that the data sampling procedure did not incorporate random selection, and that random allocation to factor levels was not possible. To this extent, the findings of this study cannot be generalized to other marketing strategies, financial institutions, or their customers. Inferences warrant serious qualifications, and causal determinations are not possible. It is possible there are unreported dependencies between observations. For example, there may be two members of the same household, and, if one subscribes for a term deposit, there may be a lower likelihood that the other would as well. Any dependencies such as these are not reported by the submitters of the data set, and we do not attempt to detect them in this paper.

**Exploratory Data Analysis (EDA)**

**Assumptions:**

1. Binary logistic regression requires that the dependent variable be binary. In this case the response variable “subscribed” is a factor of two levels: “yes” and “no”.
2. Linearity: logistic regression assumes linearity of independent variables and log odds.
3. Independence of errors: we assume independence of observations; two different outcomes from the same customer should not exist in the data, nor should the response of one customer affect another.
4. Sample size requirements: insofar as maximum likelihood estimates are statistically less powerful than those calculated by ordinary least squares, the large number of observations in the data (n = 45211) should be adequate.
5. Multicollinearity: the explanatory variables should not be too highly correlated with one another. This was checked using the scatterplot matrix and pearson correlation matrix (below).

|  |
| --- |
| **Fig. 1. Correlation Matrix Plot** |
|  |

No serious correlations between numeric variables are present in the scatterplot matrix.

**Dataset summaries**

|  |
| --- |
| **Fig. 2 Plot of Response Variable Subscribed** |
| **subscriber.frequency.png** |

As shown in Fig. 2, the classes of the response variable, ***subscribed***,are highly imbalanced, with 88% of the class in the “no” category, and only 12% of the class in the “yes” category. This imbalance may cause problems in generating a model that will accurately predict the minority class because it is such a rare event.

|  |
| --- |
| **Fig. 3. Summary of Numeric Variables** |
|  |

We see in Fig. 3 that some customers have a negative average annual account balance. Additionally, the average annual balance variable shows very significant skew, as demonstrated by the large difference between its mean and median.

|  |
| --- |
| **Fig. 4 Pearson Correlation Matrix Plot** |
|  |

Consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables, the heterogeneous correlation matrix above (Fig. 4) shows pairwise measures between all predictors as well as the response. The highest correlation is between “outcome of the previous marketing campaign“ (***poutcome***, with levels failure < other < success < unknown) with “number of days that passed by after the client was last contacted from a previous campaign “ (***pdays***) = 0.858, reflecting the period between successful customer contact, and indicating that following up on unknown outcomes of previous contact may contribute to marketing success. This relation is also reflected in the correlation between “outcome of the previous marketing campaign“ (***poutcome***) and “number of contacts performed before this campaign and for this client“ (***previous***). There is also some indication of positive correlation between ***poutcome*** and ***contact*** (with levels cellular < telephone < unknown), suggesting that telemarketing to customers at home is more effective than doing so by cellphone, as customers are presumably more occupied with other tasks when underway between locations. Customers without housing loans (***housing***, with levels no < yes ) also appear somewhat more receptive to the subscription campaign (r^2 = -.139). Lastly, though it is used primarily for benchmark purposes, the predictor duration (last contact duration, in seconds) has the highest correlation with the response (y), with r^2= 0.395. The longer one has the customer on the phone, the better the chances of marketing success.

|  |  |
| --- | --- |
| **Fig. 5. Contingency Table Heatmaps - 1** | |
| Subscribed and Contact Type | Subscribed and Credit Default |
|  |  |
| Subscribed and Has Housing Loan | Subscribed and Has Personal Loan |
|  |  |

Fig. 5 shows heatmaps of contingency tables for several explanatory variables and the response, ***subscribed***. The ***contact.type*** for both levels of ***subscribed*** is dominated by the cellular telephone numbers. Similarly, customers who have not had credit default (***credit.default***)dominate both levels of the ***subscribed*** response. This is not too surprising, as those who tend to be savers--a likely target demographic for this sort of campaign--are less likely to have defaulted on a debt. Current or potential customers are more evenly divided between the levels of the ***has.housing.loan*** predictor, but there is a much stronger tendency amongst observations towards having a personal loan (***has.personal.loan***).

|  |  |
| --- | --- |
| **Fig. 6. Contingency Table Heatmaps - 2** | |
| Subscribed and Marital Status | Subscribed and Previous Outcome |
|  |  |
| Subscribed and Highest Education | Subscribed and Job Type |
|  |  |

Fig. 6 shows additional contingency table heatmaps. The proportions of ***marital.status*** are quite different for the levels of ***subscribed***; those who did not subscribe are substantially more likely to be married than those who did. Overwhelmingly, the subscription status of individuals contacted in the previous marketing campaign (***prev.outcome***)was unknown. With respect to the highest education factor (***highest.educ***), the greatest concentration of observations in the sample is at the “Secondary” level. Looking at the response across different job types (***job.type***), ratios of subscribers to non-subscribers are higher amongst semi-employed customers (“retired”, “student”, “unemployed”) than amongst work-force regulars, with those in “management”, “administrative”, and “technical” positions more likely to subscribe than those in “blue-collar” jobs.

**Logistic regression analysis - Tentative Model**

Interpretation for the model is as described below:

Y = β0 + e β1X1 + e β2X2 + e β3X3 + e β4X4 ………….. βnXn

Y = e-2.536 + age\*e1.127^-4………..

**If β = 0 (eβ = 1), the odds of subscribing to the term deposit remain the same as x changes.**

**Β > 0 (eβ > 1), the odds of subscribing to the term deposit increase as x increases**

**Β < 0 (eβ < 1), the odds of subscribing to the term deposit decrease as x increases**

AIC = 21,648

|  |  |
| --- | --- |
| **Fig. 7 Tentative Model Results** | |
| **Tentative model estimates** |  |

After fitting the full model, comparison of the Null Deviance (**32631)** to the Residual Deviance (**21562**) indicated a significant decrease, meaning that the model with the predictors fits the data better than the model with only an intercept.

Interpretation of the difference in Deviance between models is as shown below:

1. Test the difference

Find the chi-square difference between null and residual by subtracting:

chidiff = Logit1$null.deviance – Logit1$deviance

**32631 - 21562 = 11069**

1. Find the difference in degrees of freedom between null and residual by subtracting:

dfdiff = Logit1$df.null - Logit1$df.residual

**45210 - 45168 = 42**

1. Determine significance of the difference:

**Pchisq** <**0.0001**

A chi-sq P Value of <0.0001 means that the error is significantly less than it would be using a model with no predictors. As shown above, most levels of nearly all predictors demonstrated statistical significance, indicating their relative usefulness at a high level.

One can also use an analysis of the deviance table to check for predictor contributions to the rich model. Here, terms are added sequentially, making relative contributions to model improvement more transparent.

|  |
| --- |
| **Fig. 8 Tentative Model - Analysis of Deviance Table** |
| Model: binomial, link: logit Response: y Terms added sequentially (first to last)   Df Deviance Resid. Df Resid. Dev Pr(>Chi)  NULL 45210 32631  age 1 28.2 45209 32603 1.074e-07 \*\*\* job 11 720.2 45198 31883 < 2.2e-16 \*\*\* marital 2 156.6 45196 31726 < 2.2e-16 \*\*\* education 3 103.8 45193 31622 < 2.2e-16 \*\*\* default 1 18.9 45192 31603 1.362e-05 \*\*\* balance 1 53.0 45191 31550 3.262e-13 \*\*\* housing 1 525.0 45190 31025 < 2.2e-16 \*\*\* loan 1 138.2 45189 30887 < 2.2e-16 \*\*\* contact 2 798.7 45187 30088 < 2.2e-16 \*\*\* day 1 54.9 45186 30033 1.270e-13 \*\*\* month 11 1367.2 45175 28666 < 2.2e-16 \*\*\* duration 1 5612.5 45174 23054 < 2.2e-16 \*\*\* campaign 1 139.4 45173 22914 < 2.2e-16 \*\*\* pdays 1 133.5 45172 22781 < 2.2e-16 \*\*\* previous 1 51.9 45171 22729 5.734e-13 \*\*\* poutcome 3 1166.7 45168 21562 < 2.2e-16 \*\*\* --- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

The table above shows incremental drops in deviance when adding predictors one at a time. Adding ***duration*** results in the most significant reduction to residual deviance. However, all predictors indicate significant p-values. For now, we conclude that all variables in the rich model are adding some predictive value and leave them in. Refinements to the model with recursive feature selection are subsequently investigated to improve prediction and reduce error.

**Model Selection**

|  |
| --- |
| **Fig. 9. Modeling Workflow** |
|  |

Figure 9 depicts an overview of the logistic regression modeling process. The first step in the modeling process is to partition the data into training and validation subsets. This partitioning enables the checking of the proposed logistic regression model against data it has never seen in order to mitigate the risk of overfitting. The data are partitioned by first randomly shuffling the rows of the full data set so as to eliminate any bias that may be introduced by the ordering of the data. Next, the data are divided, with 75% (33,908 rows) allocated to a training set, and 25% (11,303 rows) allocated to a validation set. The training set is then passed to the recursive feature elimination function, a feature selection process which utilizes k-fold cross validation.

|  |
| --- |
| **Fig. 10. Recursive Feature Elimination Algorithm** |
|  |

Fig. 10 depicts the recursive feature elimination algorithm, where *k* = 10. For each of the 10 folds, a full logistic regression model is fit using all *S* predictors, where *S* = 16. From this model, the predictors are ranked according to the absolute value of their t-statistics. After this ranking process, (1) a loop is entered: for *Si* in *S*, a logistic regression model is fit, the area under the receiver operating characteristic curve (AUC) is calculated, and the least important predictor is then dropped before repeating. These AUC statistics are compiled and averaged for each feature across the k repetitions. The *Si* having the highest mean AUC is then selected as the best logistic regression model. A final logistic regression model is then fitted using the *Si* predictors against the full training set.

In choosing an optimal cut-point for the class probabilities, our model results in a trade-off between the overall accuracy rate of predictions and the true positive rate (TPR) of predictions. This makes sense, for if we desired to have a perfect TPR, we would simply predict all new data as producing a positive subscriber outcome. As we expect the 88:12 no-to-yes ratio to remain fairly similar, this would of course result in a very large number of false positives, resulting in very poor accuracy. We arrived at a cutpoint of 0.104 by searching the cut-point space of 0 to 1 in 0.001 increments for the maximum geometric mean of accuracy and TPR. This cut-point results in an accuracy of 75% and a TPR of 84%. Fig. 11a depicts the accuracy-TPR trade-off of this model, which shows the receiver operating characteristic of the model, while Fig. 11b shows its area-under-curve statistic of 0.87.

|  |  |
| --- | --- |
| **Fig. 11a.** True Positive Rate vs. Accuracy | **Fig. 11b.**  Receiver Operating Characteristic |
| unbalanced.accuracyVtpr.png | unbalanced.roc.png |

Models trained on datasets with imbalanced response levels sometimes perform poorly in predicting the minority class. Methods for balancing the class proportions for modeling purposes include upsampling and downsampling. Upsampling involves generating additional row-wise data for the minority class by sampling from the minority class with replacement. Downsampling involves reducing the number of majority class rows by sampling with replacement from the majority class. The modeling process explained in the previous section is repeated by employing both upsampling and downsampling methods in an attempt to improve the true positive ratio.

|  |  |
| --- | --- |
| **Fig. 12. Accuracy vs. TPR and ROC Plots for Class Balancing Sampling Method Models** | |
| Upsample Model - Accuracy vs. TPR | Upsample Model - ROC |
| upsample.accuracyVtpr.png | upsample.roc.png |
| Downsample Model - Accuracy vs. TPR | Downsample Model - ROC |
| downsample.accuracyVtpr.png | downsample.roc.png |

Both alternative models produced ROC-AUC values only slightly better than those of the unbalanced set.

The final logistic regression model is used to predict the subscriber success response in the validation set, and these predictions are then used to evaluate the final model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fig. 13. Comparison of Accuracy, TPR, and FPR by Method** | | | | | |
| Model | Best Cut-Point | Accuracy at Best C.P. | TPR at Best C.P. | FPR at Best C.P. | Accuracy-TPR Geom. Mean at Best C.P. |
| Original (Unbalanced) | 0.10 | 0.75 | 0.84 | 0.26 | 0.80 |
| Upsampled | 0.46 | 0.78 | 0.83 | 0.22 | 0.81 |
| Downsampled | 0.45 | 0.77 | 0.84 | 0.24 | 0.81 |

Ultimately, neither of the two alternative models using upsampling and downsampling produced superior results which would justify their added complexity, and thus we determined that the original model with unbalanced response classes was the best one.

**Model Evaluation & Diagnostics**

|  |
| --- |
| **Fig. 14. Final Model Summary** |
|  |

Evaluation of the final model with 13 features made use of measures of variable importance, Wald’s test on individual predictors, the model likelihood ratio test, several (rank) discrimination indexes, bootstrapped validation of these indexes, as well as a resampling calibration curve.

|  |
| --- |
| **Fig. 15. Variance Inflation Factors** |
|  |

VIFs of the features in the final model are all < 4.0, indicating no serious multicollinearity issues.

**Variable Importance**

Individual predictor contributions to the strength of the model classification was observed using the caret::varImp() function, which lists model parameters in order of the absolute value of their t-statistics.

|  |
| --- |
| **Fig. 16. Final Model Variable Importance** |
|  |

As the original description of this dataset notes, the attribute of ***last.contact.duration*** highly affects the output target (if duration = 0, then y = ‘no’). Duration is not known before a call is performed, and after the end of the call the response is obviously known. For the purposes of realistic prediction, this feature is taken into account primarily for benchmark purposes.

Interestingly, the coefficient estimate of the ***has.housing.loan*** feature (-1.068e+00) reflected the negative correlation observed in EDA (-.139), indicating the influential factor of whether or not the customer has housing loans. Bearing in mind that only homeowners have housing loans, the strength of this factor appears to outweigh whatever influence is exerted by the mode of contact with the customer, by cell, home telephone, or otherwise (***contact.type***). The number of contacts before the current campaign with the customer (***contacts.num.prev***) appeared to matter more than the number during (***contacts.num***), which is consistent with the positive value of the coefficient (3.509e-03) for ***days.passed***-- the number of days since last contact with the client in a previous marketing campaign. Also of importance, customer ownership of a personal loan (***has.personal.loan***) reflected in its coefficient estimate (-7.102e-01) the same negative relationship with the response observed in the correlation matrix.

**Wald’s tests**

As a measure of the ratio of the square of the regression coefficient to the square of its standard error, the Wald test is useful for determining if removing a predictor will harm fit. Low p-values for the model coefficients demonstrate the usefulness of the corresponding predictors to the fitting procedure.

|  |  |  |
| --- | --- | --- |
| **Fig.17. Final Model Wald Test Results** | | |
| *Predictor* | *F-score* | *P-val* |
| last.contact.duration | F = 3087.066 | p= < 2.22e-16 |
| has.housing.loan | F = 593.3766 | p= < 2.22e-16 |
| contact.type | F = 437.2825 | p= < 2.22e-16 |
| contacts.num.prev | F = 145.9943 | p= < 2.22e-16 |
| contacts.num | F = 127.3638 | p= < 2.22e-16 |
| days.passed | F = 121.63 | p= < 2.22e-16 |
| has.personal.loan | F = 116.4509 | p= < 2.22e-16 |
| highest.educ | F = 58.06867 | p= 2.5971e-14 |
| prev.outcome | F = 42.43241 | p= 7.4191e-11 |
| last.contact.moy | F = 36.25536 | p= 1.7486e-09 |
| marital.status | F = 33.32687 | p= 7.8577e-09 |
| age | F = 15.5384 | p= 8.1008e-05 |
| avg.annual.balance | F = 14.85314 | p= 0.00011643 |

**Likelihood Ratio Test & Discrimination Indexes**

Output from the rms::lrm() function grants particular insight into the evaluative process with a number of readily available model performance measures:

|  |
| --- |
| **Fig. 18. Final Model Likelihood Ratio Test Result** |
|  |

At a high level, the likelihood ratio test statistically quantifies the usefulness of the predictors in comparison with reduced models that exclude them. The null hypothesis holds that the reduced model is true, and a low p-value (<.0001) for the overall model fit statistic compels one to reject the null hypothesis.

Max |deriv| is the maximum (over βs) of the absolute value of the first derivative of the log-likelihood function at the apparent maximum likelihood estimates. The value of 3e-08 indicates that convergence happened.

Although the R-square value is adversely affected by imbalanced data, the g-indexes and Brier-scores indicate relatively low levels of response variation, and a reasonably small sum of square differences between predicted and actual outcomes.

The latter is consistent with the Somers' Dxy rank correlation (Dxy = .741) between the predicted probabilities and the observed responses, as well as with the characteristically optimistic (C = .871) AUC or concordance-index (c-index), which has simple relationship with Somer’s Dxy: Dxy=2(c−0.5). With values between 0 and 1, Dxy suggests that the model's predictions are random when equaling 0; at Dxy=1, the model is perfectly discriminating.

To quantify the model’s optimism and validate these evaluative measures, bootstrapping with 1000 repetitions was applied:

|  |
| --- |
| **Fig. 19. Bootstrap Results** |
|  |

Comparisons were conducted with bias-corrected measures, indicated in the column above entitled “index.corrected”, which adjusts the index in accordance with the model’s “optimism”.

**Calibration Curve using Resampling**

As a final evaluative procedure, a calibration curve using resampling was constructed:

|  |
| --- |
| **Fig. 20. Calibration Curve** |
|  |

Inferences drawn from this model should be duly qualified with a view to the above plot, whose curve indicates some evidence of overfitting, underestimating low probabilities, and overestimating high ones.

**CONCLUSION**

|  |  |
| --- | --- |
| **Fig. 21. Accuracy vs. TPR and ROC Plots for Tentative Model** | |
| Accuracy vs. TPR | ROC |
| **tent.accuracyVtpr.png** | **tentativemodel.roc.png** |

Returning to the tentative model, we use it to predict against the validation data set and assess it in terms of TPR, accuracy, and ROC-AUC as shown in Fig. 21 & 22.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fig. 22. Comparison of Accuracy, TPR, and FPR by Method** | | | | | |
| Model | Best Cut-Point | Accuracy at Best C.P. | TPR at Best C.P. | FPR at Best C.P. | Accuracy-TPR Geom. Mean at Best C.P. |
| Final Model | 0.10 | 0.75 | 0.84 | 0.26 | 0.80 |
| Tentative Model | 0.12 | 0.78 | 0.81 | 0.22 | 0.80 |

The final model shows a performance profile very similar to the tentative model. While the TPR of the final model is slightly better than that of the tentative model, the accuracy and FPR are slightly worse, and the geometric mean of accuracy and TPR are effectively equivalent. What’s more, the ROC-AUC metrics are nearly indistinguishable, leading us to a conclusion that, from a predictive perspective, the final model is likely not superior to the tentative model.

In effect, both models appear to be approaching a performance wall in their predictive ability. True positives are found only at the expense of accurately predicting an event whose rarity introduces uncorrected model bias. The marginal success of the upsampling and downsampling procedures corroborates these results, which are also observable in the resampling calibration curve. In light of this finding, application of Firth’s bias-reduced penalized-likelihood logistic regression warrants further investigation. The statistical significance of nearly all levels of all predictors in the tentative and final models suggests insufficient features as another likely factor in predictive performance. Having verified the low variance inflation factor amongst final model predictors (< 4.0), there are presumably more factors at work in bank term deposit subscription than a first-order combination of the available features can account for; higher-order interactions and dummy variable encoding of individual factor levels also warrant closer study. Worthy of note, however, is the success of recursive feature elimination in identifying a cross-validated final model that performs on par with the tentative, full model despite having fewer predictors. Though the classifier inadequately discriminates the rare event of positive bank term deposit subscription, the parsimonious result of the final model does grant insight into the more important features as well as indirect knowledge of what factors into a negative response. Intuitively, individuals with personal and home loans are less likely to subscribe. Though customers are presumably more receptive to direct marketing in the relatively undistracted environment of the home, the proportion of direct marketing campaigns conducted by cell phone factors strongly into the coefficient estimate for contact type. While the covariance with the response of previous outcome and number of previous contacts contrasts with the correlation estimates of these features, suggesting the need for finer factor-level encoding, it also indicates the importance of targeting repeat customers and maintaining established points of contact. From this interpretive perspective, the positive coefficient estimate of days elapsed since last customer contact may be as much a reflection of previous success as it is one of customer non-responsiveness. Importantly, such inference is consistent with the notion that previous subscribers, as well as potential customers who have not yet removed themselves from the direct marketing contact list, are more likely to subscribe in the future. These findings point toward a predictive model for the likelihood of bank term deposit subscription that is less the function of demographic factors than it is one of observed patterns of customer behavior.

**REFERENCES**

[*http://www.statmethods.net/graphs/scatterplot.html*](http://www.statmethods.net/graphs/scatterplot.html)

[*https://stats.idre.ucla.edu/r/dae/logit-regression/*](https://stats.idre.ucla.edu/r/dae/logit-regression/)

[*http://www.researchmanuscripts.com/July2014/2.pdf*](http://www.researchmanuscripts.com/July2014/2.pdf)

[*www.stat.ufl.edu/~winner/sta6127/chapter15c.ppt*](http://www.stat.ufl.edu/~winner/sta6127/chapter15c.ppt)

[*www.statstools.com*](http://www.statstools.com)

[*https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/*](https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/)

**APPENDIX**

**Exploratory Data Analysis**

|  |
| --- |
| **#install.packages('dplyr')**  library(dplyr)  library(glmnet)  library(ROCR)  library(MASS)  **# set the working directory**  setwd("~/Documents/MSDS 6372 STATS TERM 2/PROJECTS/Project3")  **# load the data**  bank.dat <- read.csv("C:\\Users\\me\\Documents\\Documents\\MSDS 6372 STATS TERM 2\\PROJECTS\\Project3\\bank-full-ak.csv", sep=",", header = TRUE )  is.na(bank.dat) <- bank.dat=="" #look for missing values  sapply(bank.dat,function(x) sum(is.na(x))) #count of missing values per variable  sapply(bank.dat, function(x) length(unique(x))) #levels per variable  **# summary of data**  head(bank.dat)  str(bank.dat)  summary(bank.dat)  table(bank.dat$y)  prop.table(table(bank.dat$y))  **#Summary of numeric variables**  library(psych)  bank.num<- as.data.frame(subset(bank.dat, select = c(1,6,10,12,13,14,15,17)))  head(bank.num)  describeBy(bank.num, "y")  **# Scatterplot Matrices to check for multicollinearity**  install.packages("gclus")  library(gclus)  correl.r <- abs(cor(correl)) # get correlations  correl.col <- dmat.color(correl.r) # get colors  # reorder variables so those with highest correlation  # are closest to the diagonal  correl.o <- order.single(correl.r)  cpairs(correl, correl.o, panel.colors=correl.col, gap=.5,  main="Continuous Variables Ordered and Colored by Correlation")  **#logistic regression**  Logit1= glm(y ~ ., data= bank.dat, family= binomial (link = "logit"))  summary(Logit1)  anova(Logit1, test="Chisq") **# to analyze the table of deviance**  **#Heterogeneous Correlation Matrix**  library(sjPlot)  new\_bank.dat <- as.data.frame(lapply(bank.dat, as.integer))  sjp.corr(new\_bank.dat) |

**Model Selection**

|  |
| --- |
| library(caret)  library(dplyr)  library(ROCR)  library(MASS)  **# load the data**  bank.dat <- read.csv('../data\_sets/bank/bank-full.csv', sep=";")  **# summary of data**  head(bank.dat)  #str(bank.dat)  #summary(bank.dat)  ##---------------------------------------  **## BEGIN DATA PREPARATION**  ##---------------------------------------  **# manipulate data so we can run it through rfe/glm**  c1 <- as.double(bank.dat$age)  c2 <- as.factor(bank.dat$job)  c3 <- as.factor(bank.dat$marital)  c4 <- as.factor(bank.dat$education)  c5 <- as.factor(bank.dat$default)  c6 <- as.double(bank.dat$balance)  c7 <- as.factor(bank.dat$housing)  c8 <- as.factor(bank.dat$loan)  c9 <- as.factor(bank.dat$contact)  c10 <- as.factor(bank.dat$day)  c11 <- as.factor(bank.dat$month)  c12 <- as.double(bank.dat$duration)  c13 <- as.double(bank.dat$campaign)  c14 <- as.factor(bank.dat$pdays)  c15 <- as.double(bank.dat$previous)  c16 <- as.factor(bank.dat$poutcome)  c17 <- as.factor(as.character(bank.dat$y))  c18 <- as.factor(as.numeric(ifelse(c17 == "yes", 1, 0)))  bank.dat.2 <- data.frame(cbind(c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11, c12, c13, c14, c15, c16, c18))  colnames(bank.dat.2) <- c("age"  ,"job.type"  ,"marital.status"  ,"highest.educ"  ,"credit.default"  ,"avg.annual.balance"  ,"has.housing.loan"  ,"has.personal.loan"  ,"contact.type"  ,"last.contact.dom"  ,"last.contact.moy"  ,"last.contact.duration"  ,"contacts.num"  ,"days.passed"  ,"contacts.num.prev"  ,"prev.outcome"  ,"subscribed")  # after dataframe, 'subscribed' reverts to numeric, so set it to factor, table flip--> (╯°□°）╯︵ ┻━┻  bank.dat.2$subscribed <- as.factor(bank.dat.2$subscribed)  **# Housekeeping**  rm(c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11, c12, c13, c14, c15, c16, c17, c18) # de-clutter  rm(bank.dat) # get rid of bank.dat (free up memory)  ##---------------------------------------  **## END DATA PREPARATION**  ##---------------------------------------  ##---------------------------------------  ## **BEGIN TRAIN / TEST SPLI**T  ##---------------------------------------  **# Set seed**  set.seed(1984)  **# Shuffle row indices: rows**  rows <- sample(nrow(bank.dat.2))  **# Randomly order data**  bank.dat.2 <- bank.dat.2[rows, ]  **# train-test sizes**  split <- round(nrow(bank.dat.2) \* 0.75,0)  **# Create train**  bank.train <- bank.dat.2[1:split, ]  **# Create validation set**  bank.validate <- bank.dat.2[(split + 1):nrow(bank.dat.2), ]  **# housekeeping**  rm(bank.dat.2) # get rid of bank.dat.2 (free up memory)  rm(rows) # de-clutter  rm(split) # de-clutter  **##---------------------------------------**  **## END TRAIN / TEST SPLIT**  ##---------------------------------------  **##---------------------------------------**  **## BEGIN MODELING**  ##---------------------------------------  **# number of predictors**  number\_predictors <- dim(bank.train)[2]-1 # less 1 for the response var!  **# partition of training set into Xs and Y objects**  x <- dplyr::select(bank.train, -subscribed) # predictors  y <- as.factor(as.numeric(bank.train$subscribed)) # response  myLRFuncs <- lrFuncs  myLRFuncs$summary <- twoClassSummary  rfe.ctrl <- rfeControl(functions = myLRFuncs,  method = "cv",  number = 10,  verbose = TRUE  , returnResamp = "all")  train.ctrl <- trainControl(method="none",  classProbs=TRUE,  summaryFunction=twoClassSummary,  verbose=TRUE)  glm\_rfe\_ROC <- rfe(x, y,  sizes=c(1:number\_predictors),  rfeControl=rfe.ctrl,  # family="binomial",  family=binomial(link="logit"), # this is the default link, but making explicit for clarity  method="glm", # using the generalized linear models package  metric="ROC", # set selection metric to receiver operating characteristic  trControl=train.ctrl)  **# housekeeping**  rm(myLRFuncs) # de-cluttering the environment  rm(number\_predictors) # de-cluttering the environment  rm(rfe.ctrl) # de-cluttering the environment  rm(train.ctrl) # de-cluttering the environment  ##---------------------------------------  **## END MODELING**  ##---------------------------------------  ##---------------------------------------  #**# BEGIN MODEL METRICS**  ##---------------------------------------  **# summary of the feature selection/cv process**  glm\_rfe\_ROC  **# histograms of various metrics**  hist(glm\_rfe\_ROC$resample$ROC)  hist(glm\_rfe\_ROC$resample$Sens)  hist(glm\_rfe\_ROC$resample$Spec)  **# produce the performance profile across different subset sizes**  trellis.par.set(caretTheme())  plot(glm\_rfe\_ROC, type = c("g", "o"))  x.validate <- bank.validate[1:(ncol(bank.validate)-1)]  y.validate <- ifelse(as.numeric(bank.validate$subscribed) == 2, 1, 0)  y.pred <- predict(glm\_rfe\_ROC, x.validate)  y.pred$pred <- as.numeric(as.character(y.pred$pred))  y.pred$pred <- ifelse(y.pred$pred == 2, 1, 0)  y.actual <- as.numeric(ifelse(bank.validate$subscribed == 2, 1, 0))  **# ROC Curve**  validation.prediction <- prediction(y.pred$pred, y.validate)  roc.perf = performance(validation.prediction, measure = "tpr", x.measure = "fpr")  plot(roc.perf)  **# Confusion matrix (reveals how "confused" the model is)**  v\_actual <- as.character(ifelse(y.actual == 0, "no", "yes"))  p\_class <- as.character(ifelse(y.pred[2] > 0.5, "yes", "no")) # 50% cut point (not sure what model uses)  cm <- confusionMatrix(p\_class, v\_actual , positive="yes") # this is giving a confusion matrix with vars in wrong order  model.accuracy <- cm$overall[1] # Accuracy = classifications correct / all classifications  model.kappa <- cm$overall[2] # Kappa = (observed accuracy - expected accuracy)/(1 - expected accuracy)  model.sens <- cm$byClass[1] # Sensitivity = true positive / (true positive + false negative) [also called true positive rate]  model.spec <- cm$byClass[2] # Specificity = true neg / (actual 'no' + actual 'no')  model.ppv <- cm$byClass[3] # Pos Pred Value = (sensitivity \* prevalence)/((sensitivity\*prevalence) + ((1-specificity)\*(1-prevalence)))  model.npv <- cm$byClass[4] # Neg Pred Value = (specificity \* (1-prevalence))/(((1-sensitivity)\*prevalence) + ((specificity)\*(1-prevalence)))  model.prec <- cm$byClass[5] # Precision = true positive / (true positive + false positive)  model.f1 <- cm$byClass[7] # F1 = (1+beta^2)\*precision\*recall/((beta^2 \* precision)+recall) \*where beta = 1 for this function.  model.prev <- cm$byClass[8] # Prevalence = (true positive + false negative) / sum( all cells)  model.detr <- cm$byClass[9] # Detection Rate = true positive / sum(all cells)  model.detp <- cm$byClass[10] # Detection Prevalence = (true positive + false positive ) / sum( all cells)  model.bacc <- cm$byClass[11] # Balanced Accuracy = (sensitivity+specificity)/2  save(glm\_rfe\_ROC, file='../mdl\_obj/unbalanced.RData')  **# housekeeping**  rm(my\_RMSE)  rm(x)  rm(y)  rm(y.actual)  rm(y.validate)  rm(x.validate)  rm(bank.train)  rm(bank.validate)  rm(y.pred)  rm(cm)  rm(model.accuracy)  rm(model.kappa)  rm(model.sens)  rm(model.spec)  rm(model.ppv)  rm(model.npv)  rm(model.prec)  rm(model.f1)  rm(model.prev)  rm(model.detr)  rm(model.detp)  rm(model.bacc)  rm(p\_class)  rm(v\_actual)  ##---------------------------------------  **## END MODEL METRICS**  ##--------------------------------------- |

**Model Evaluation & Diagnostics**

|  |
| --- |
| **#Final Model Summary**  summary(glm\_rfe\_ROC$fit)  **#Variance Inflation Factor**  library(rms)  vif(glm\_rfe\_ROC$fit)  **#Variable Importance**  varImp(glm\_rfe\_ROC$fit)  **#Wald Test**  library(survey)  regTermTest(glm\_rfe\_ROC$fit, "last.contact.duration")  regTermTest(glm\_rfe\_ROC$fit, "has.housing.loan")  regTermTest(glm\_rfe\_ROC$fit, "contact.type")  regTermTest(glm\_rfe\_ROC$fit, "contacts.num.prev")  regTermTest(glm\_rfe\_ROC$fit, "days.passed")  regTermTest(glm\_rfe\_ROC$fit, "contacts.num")  regTermTest(glm\_rfe\_ROC$fit, "has.personal.loan")  regTermTest(glm\_rfe\_ROC$fit, "highest.educ")  regTermTest(glm\_rfe\_ROC$fit, "prev.outcome")  regTermTest(glm\_rfe\_ROC$fit, "last.contact.moy")  regTermTest(glm\_rfe\_ROC$fit, "marital.status")  regTermTest(glm\_rfe\_ROC$fit, "age")  regTermTest(glm\_rfe\_ROC$fit, "avg.annual.balance")  **#Likelihood Ratio Test & Discrimination Indexes**  library(rms)  mod\_1 <- lrm(subscribed ~ contact.type + contacts.num + prev.outcome + age + last.contact.duration + contacts.num.prev + has.personal.loan + last.contact.moy + avg.annual.balance + has.housing.loan + days.passed + highest.educ + marital.status, x=TRUE, y=TRUE, data = bank.validate)  print(mod\_1)    my.valid <- validate(mod\_1, method="boot", B=1000)  my.valid  **#Calibration Curve using Resampling**  my.calib <- calibrate(mod\_1, method="boot", B=1000)  par(bg="white", las=1)  plot(my.calib, las=1) |

**DATASET CITATION**

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

Available at: [pdf] http://hdl.handle.net/1822/14838

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

1. Title: Bank Marketing

2. Sources

Created by: Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012