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.

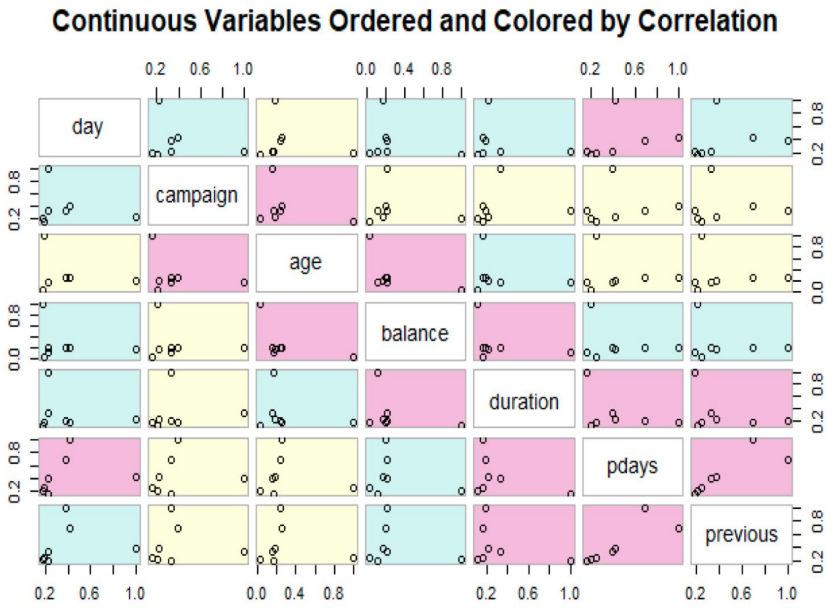
Exploratory Data Analysis (EDA)

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).

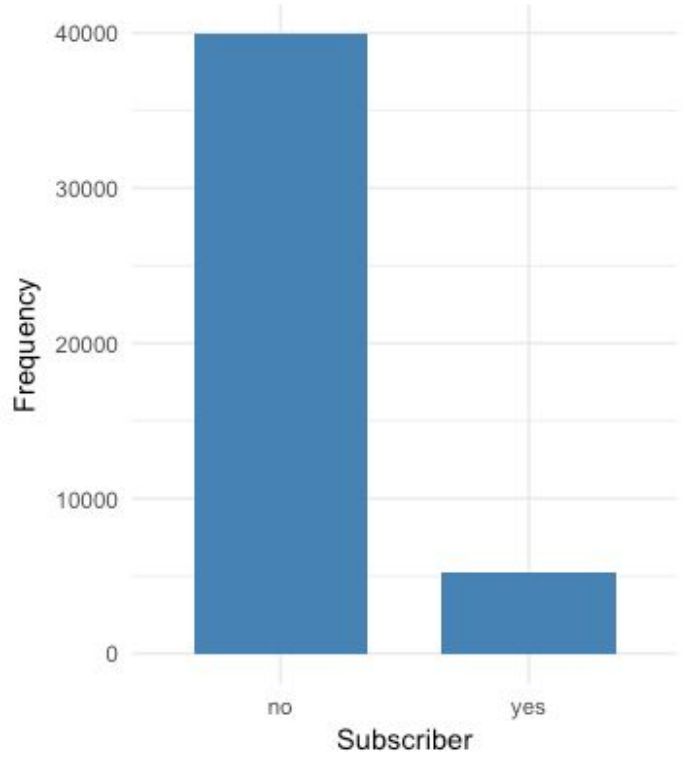
Fig. 1. Correlation Matrix Plot



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

Dataset summaries

Fig. 2 Plot of Response Variable Subscribed



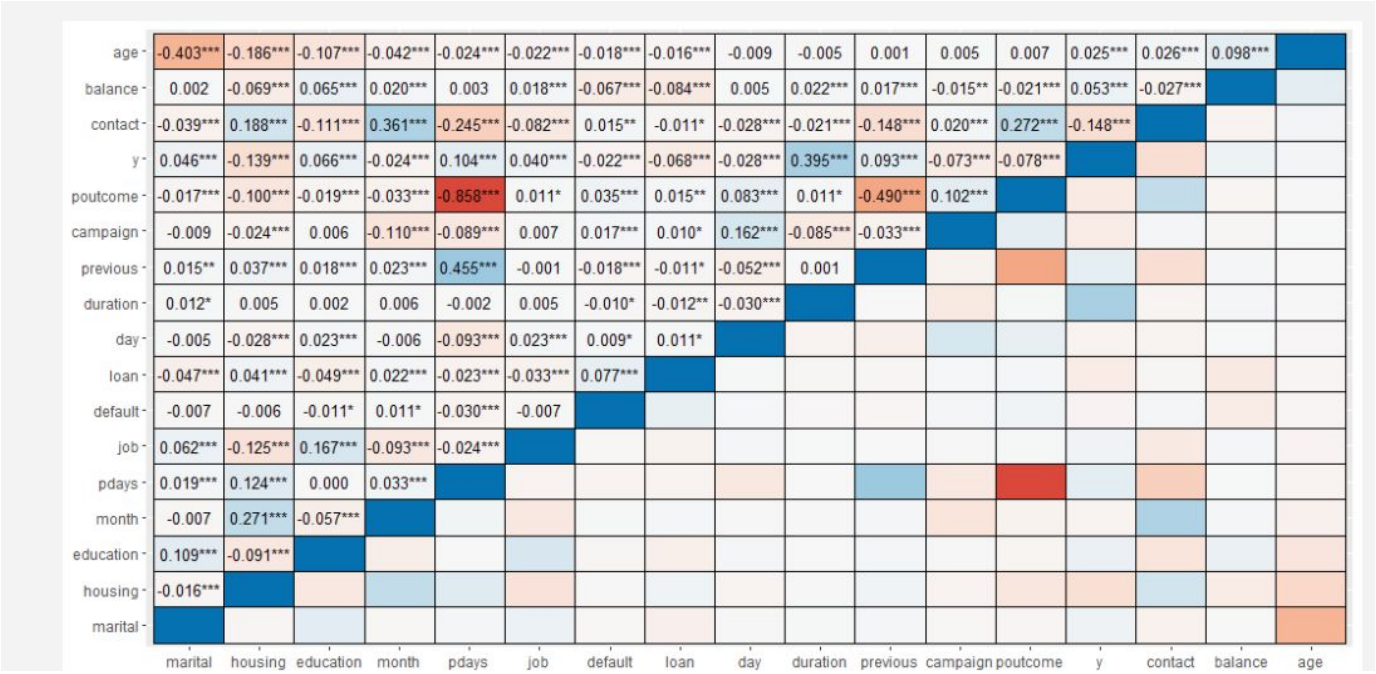
As shown in Fig. 2, the classes of the response variable, ***subscribed***, are highly imbalanced, with 88% of the class in the "no" category, and only 12% of the class in the "yes" category. This imbalance may cause problems in generating a model that will accurately predict the minority class because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

Fig. 3. Summary of Numeric Variables

```
 Descriptive statistics by group
group: no
          vars     n    mean      sd median trimmed    mad   min    max  range  skew kurtosis    se
age          1 39922   40.84   10.17     39   40.25  10.38    18     95     77  0.59     0.05  0.05
balance      2 39922 1303.71 2974.20    417  719.55 618.24 -8019 102127 110146  8.31   140.50 14.89
day          3 39922   15.89    8.29     16   15.78  10.38     1     31     30  0.08    -1.06  0.04
duration     4 39922  221.18  207.38    164  186.61 121.57     0   4918   4918  3.59    28.95  1.04
campaign     5 39922    2.85    3.21      2    2.18   1.48     1     63     62  4.80    37.30  0.02
pdays        6 39922   36.42   96.76     -1    8.02   0.00    -1    871    872  2.70     7.03  0.48
previous     7 39922    0.50    2.26      0    0.06   0.00     0    275    275 49.09  5522.73  0.01
y*           8 39922    1.00    0.00      1    1.00   0.00     1      1      0   NaN      NaN  0.00
------------------------------------------------------------------------------------------------------------
group: yes
          vars    n    mean      sd median trimmed    mad   min   max range skew kurtosis    se
age          1 5289   41.67   13.50     38   40.36  11.86    18    95    77 0.87     0.28  0.19
balance      2 5289 1804.27 3501.10    733 1149.40 985.93 -3058 81204 84262 8.44   133.87 48.14
day          3 5289   15.16    8.50     15   14.95  10.38     1    31    30 0.17    -1.05  0.12
duration     4 5289  537.29  392.53    426  481.42 318.76     8  3881  3873 1.70     4.96  5.40
campaign     5 5289    2.14    1.92      2    1.75   1.48     1    32    31 4.37    35.09  0.03
pdays        6 5289   68.70  118.82     -1   41.67   0.00    -1   854   855 2.14     5.50  1.63
previous     7 5289    1.17    2.55      0    0.60   0.00     0    58    58 6.61    97.97  0.04
y*           8 5289    2.00    0.00      2    2.00   0.00     2     2     0  NaN      NaN  0.00
>
```

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.

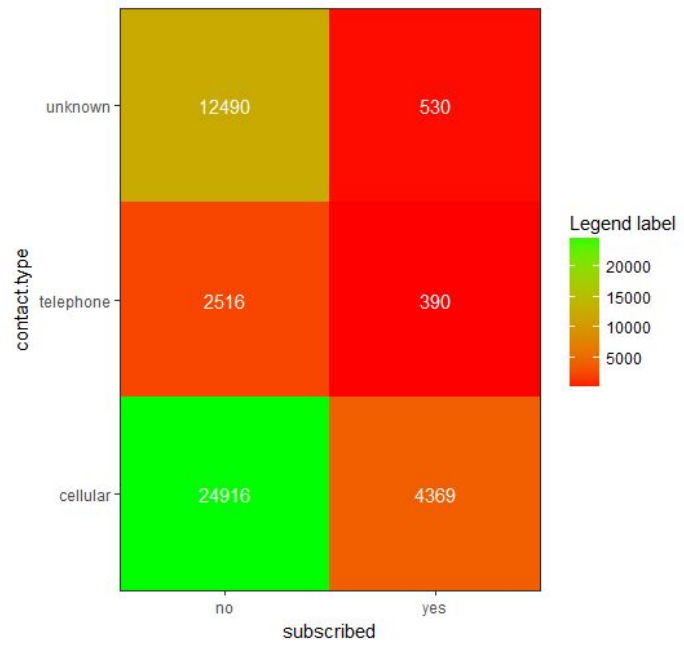
Fig. 4 Pearson Correlation Matrix Plot



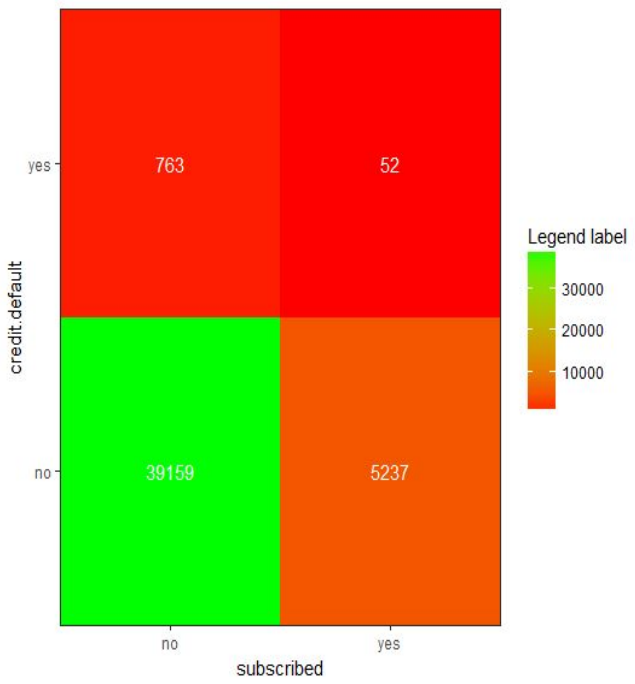
Consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables, the heterogeneous correlation matrix above (Fig. 4) shows pairwise measures between all predictors as well 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

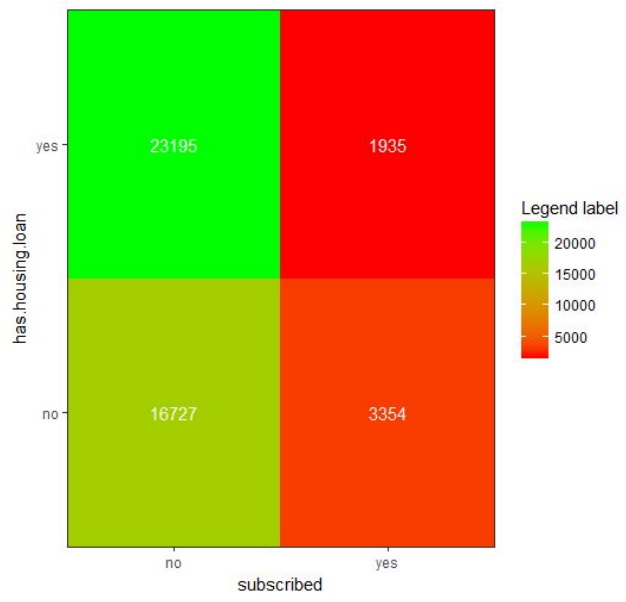
Subscribed and Contact Type



Subscribed and Credit Default



Subscribed and Has Housing Loan



Subscribed and Has Personal Loan

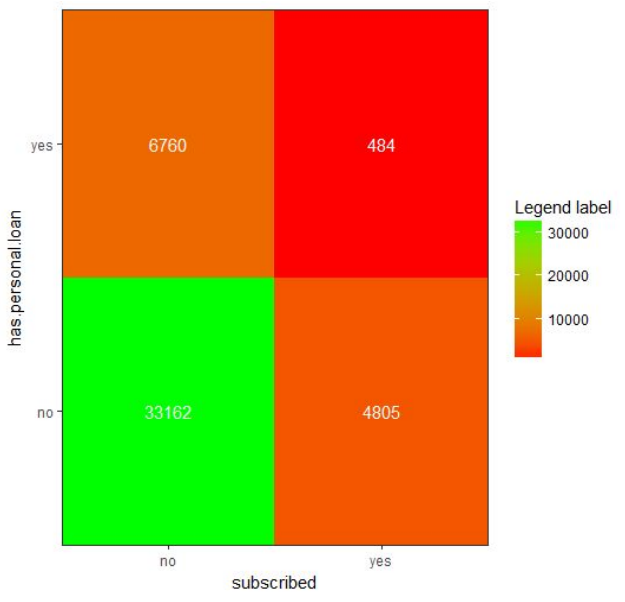
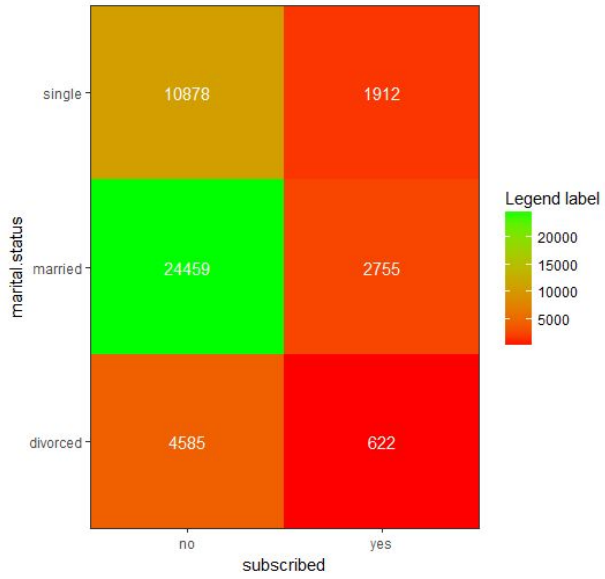


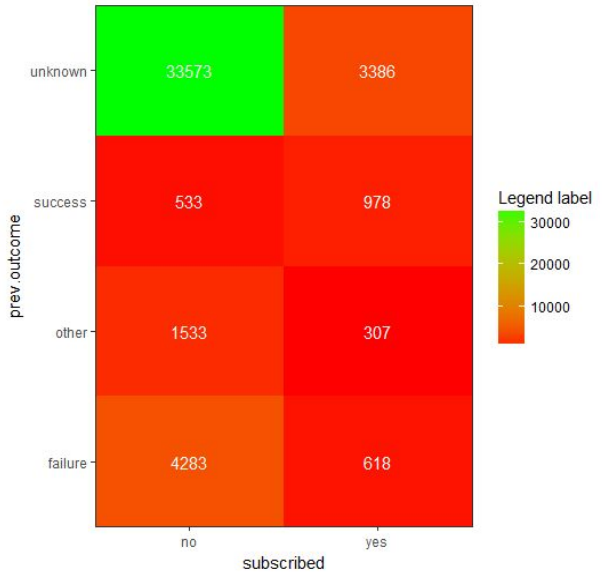
Fig. 5 shows heatmaps of contingency tables for several explanatory variables and the response, ***subscribed***. The ***contact.type*** for both levels of ***subscribed*** is dominated by the cellular telephone numbers. Similarly, customers who have not had credit default (***credit.default***) dominate 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

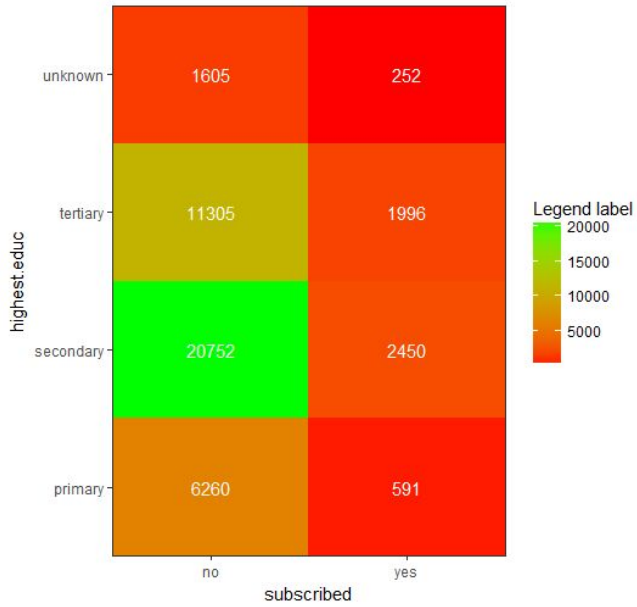
Subscribed and Marital Status



Subscribed and Previous Outcome



Subscribed and Highest Education



Subscribed and Job Type

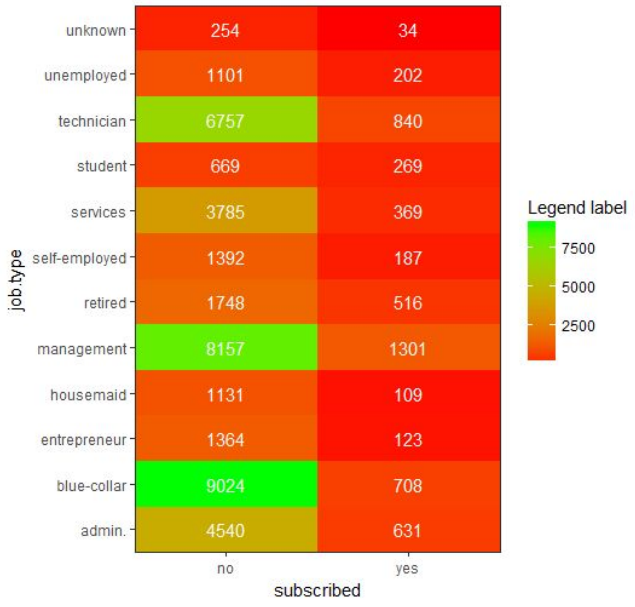


Fig. 6 shows additional contingency table heatmaps. The proportions of ***marital.status*** are quite different for the levels of ***subscribed***; those who did not subscribe are substantially more likely to be married than those who did. Overwhelmingly, the subscription status of individuals contacted 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:

$Y = \beta 0 + e^{\beta 1}X^1 + e^{\beta 2}X^2 + e^{\beta 3}X^3 + e^{\beta 4}X^4$ …….. βnXn

$Y = e^{-2.536} + age*e^{1.127\text{^}-4}$……..

If β = 0 (e^{β} = 1), the odds of subscribing to the term deposit remain the same as x changes.

B > 0 (e^{β} > 1), the odds of subscribing to the term deposit increase as x increases

B < 0 (e^{β} < 1), the odds of subscribing to the term deposit decrease as x increases

AIC = 21,648

Fig. 7 Tentative Model Results

Tentative model estimates

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7286  -0.3744  -0.2530  -0.1502   3.4288

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.536e+00  1.837e-01 -13.803  < 2e-16 ***
age                1.127e-04  2.205e-03   0.051 0.959233
jobblue-collar    -3.099e-01  7.267e-02  -4.264 2.01e-05 ***
jobentrepreneur   -3.571e-01  1.256e-01  -2.844 0.004455 **
jobhousemaid      -5.040e-01  1.365e-01  -3.693 0.000221 ***
jobmanagement     -1.653e-01  7.329e-02  -2.255 0.024130 *
jobretired         2.524e-01  9.722e-02   2.596 0.009436 **
jobself-employed  -2.983e-01  1.120e-01  -2.664 0.007726 **
jobservices       -2.238e-01  8.406e-02  -2.662 0.007763 **
jobstudent         3.821e-01  1.090e-01   3.505 0.000457 ***
jobtechnician     -1.760e-01  6.893e-02  -2.554 0.010664 *
jobunemployed     -1.767e-01  1.116e-01  -1.583 0.113456
jobunknown        -3.133e-01  2.335e-01  -1.342 0.179656
maritalmarried    -1.795e-01  5.891e-02  -3.046 0.002318 **
maritalsingle      9.250e-02  6.726e-02   1.375 0.169066
educationsecondary 1.835e-01  6.479e-02   2.833 0.004618 **
educationtertiary  3.789e-01  7.532e-02   5.031 4.88e-07 ***
educationunknown   2.505e-01  1.039e-01   2.411 0.015915 *
defaultyes        -1.668e-02  1.628e-01  -0.102 0.918407
balance            1.283e-05  5.148e-06   2.493 0.012651 *
housingyes        -6.754e-01  4.387e-02 -15.395  < 2e-16 ***
loanyes           -4.254e-01  5.999e-02  -7.091 1.33e-12 ***
contacttelephone  -1.634e-01  7.519e-02  -2.173 0.029784 *
contactunknown    -1.623e+00  7.317e-02 -22.184  < 2e-16 ***
day                9.969e-03  2.497e-03   3.993 6.53e-05 ***
monthaug          -6.939e-01  7.847e-02  -8.842  < 2e-16 ***
monthdec           6.911e-01  1.767e-01   3.912 9.17e-05 ***
monthfeb          -1.473e-01  8.941e-02  -1.648 0.099427 .
monthjan          -1.262e+00  1.217e-01 -10.367  < 2e-16 ***
monthjul          -8.308e-01  7.740e-02 -10.733  < 2e-16 ***
monthjun           4.536e-01  9.367e-02   4.843 1.28e-06 ***
monthmar           1.590e+00  1.199e-01  13.265  < 2e-16 ***
monthmay          -3.991e-01  7.229e-02  -5.521 3.36e-08 ***
monthnov          -8.734e-01  8.441e-02 -10.347  < 2e-16 ***
monthoct           8.814e-01  1.080e-01   8.159 3.37e-16 ***
monthsep           8.741e-01  1.195e-01   7.314 2.58e-13 ***
duration           4.194e-03  6.453e-05  64.986  < 2e-16 ***
campaign          -9.078e-02  1.014e-02  -8.955  < 2e-16 ***
pdays             -1.027e-04  3.061e-04  -0.335 0.737268
previous           1.015e-02  6.503e-03   1.561 0.118476
poutcomeother      2.035e-01  8.986e-02   2.265 0.023543 *
poutcomesuccess    2.291e+00  8.235e-02  27.821  < 2e-16 ***
poutcomeunknown   -9.179e-02  9.347e-02  -0.982 0.326093
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32631  on 45210  degrees of freedom
Residual deviance: 21562  on 45168  degrees of freedom
AIC: 21648

Number of Fisher Scoring iterations: 6
```

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

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

dfdiff = Logit1$df.null - Logit1$df.residual
45210 - 45168 = 42

3. 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

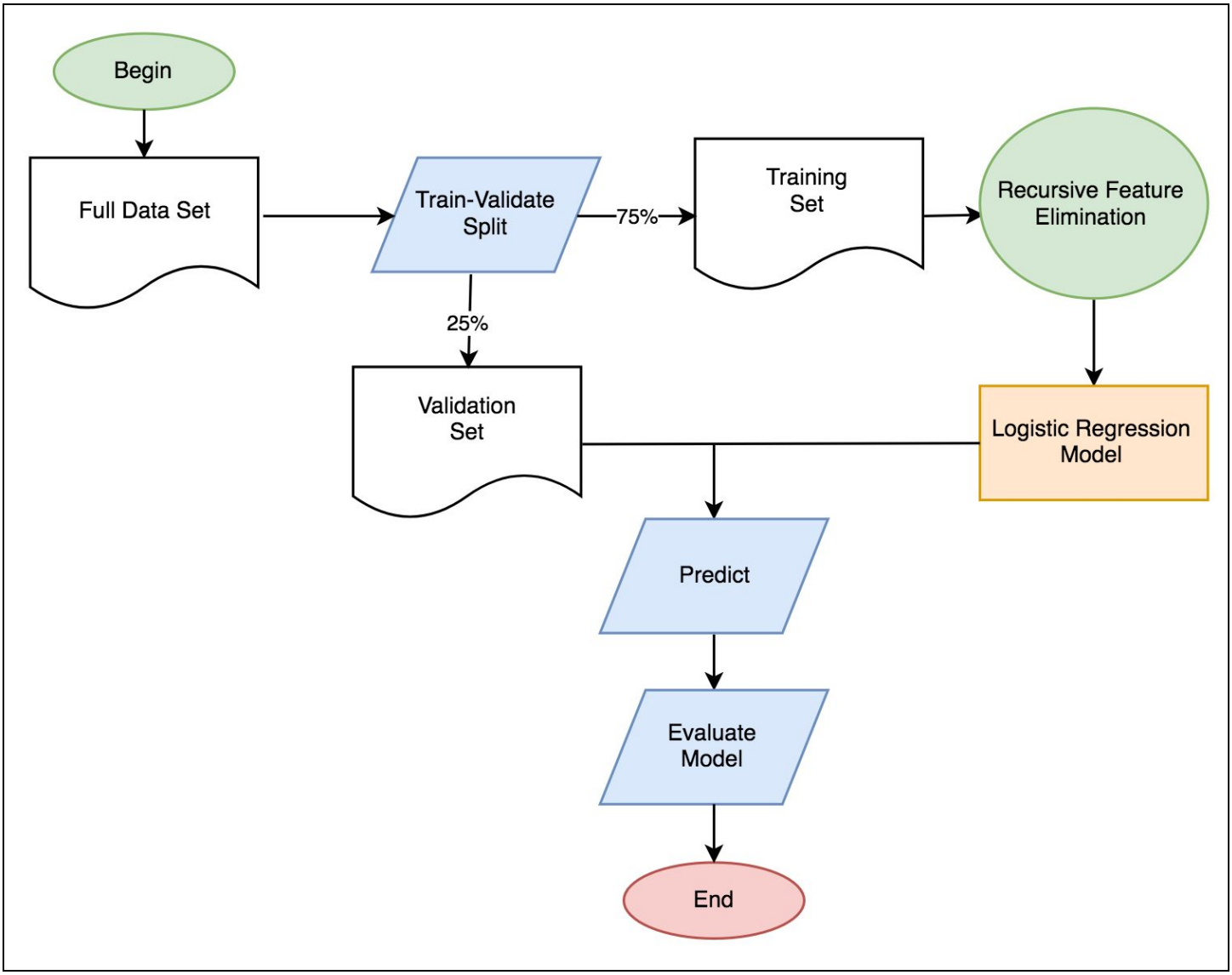


Figure 9 depicts an overview of the logistic regression modeling process. The first step in the modeling process is to partition the data into training and validation subsets. This partitioning enables the checking of the proposed logistic regression model against data it has never seen in order to mitigate 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

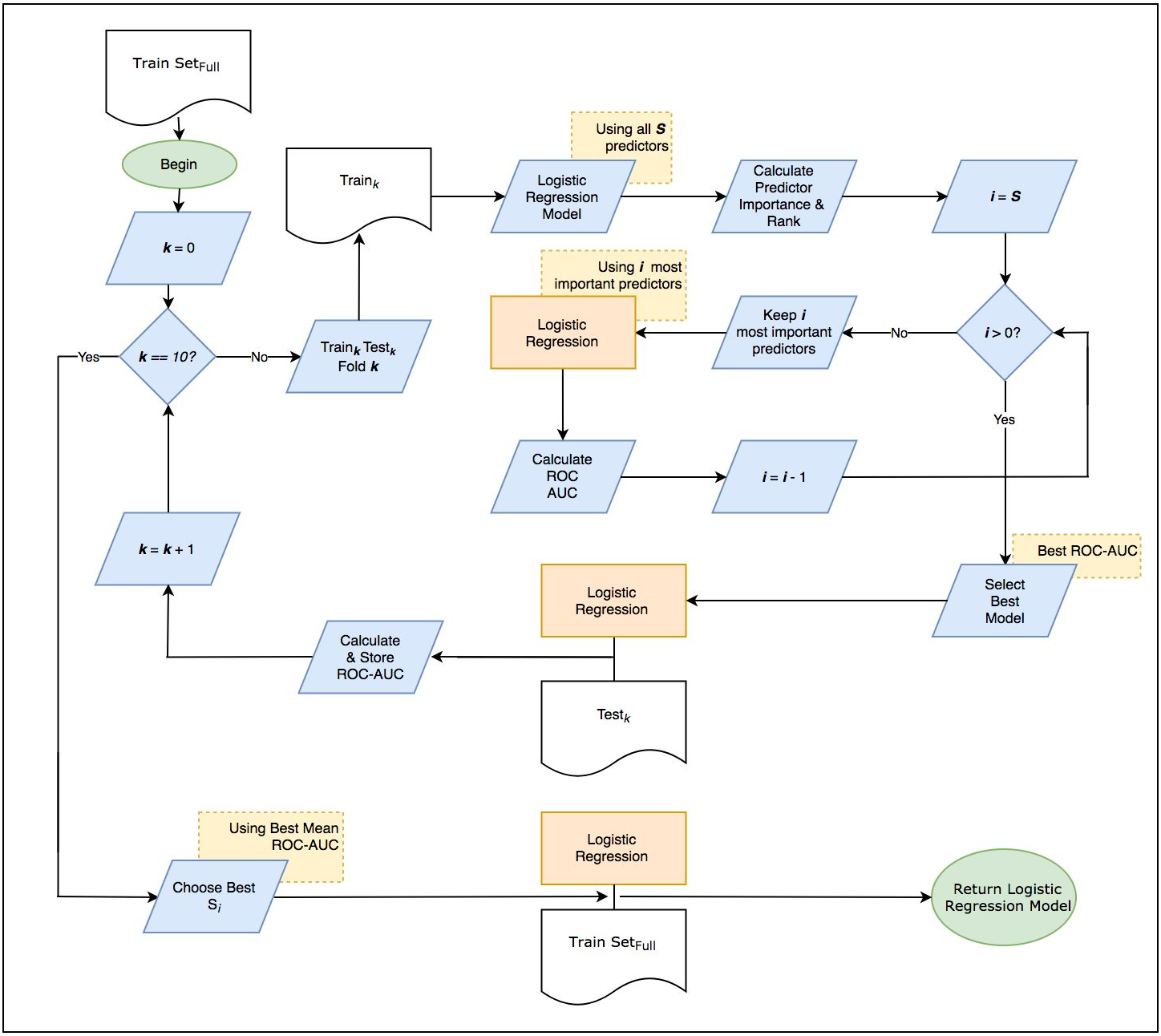
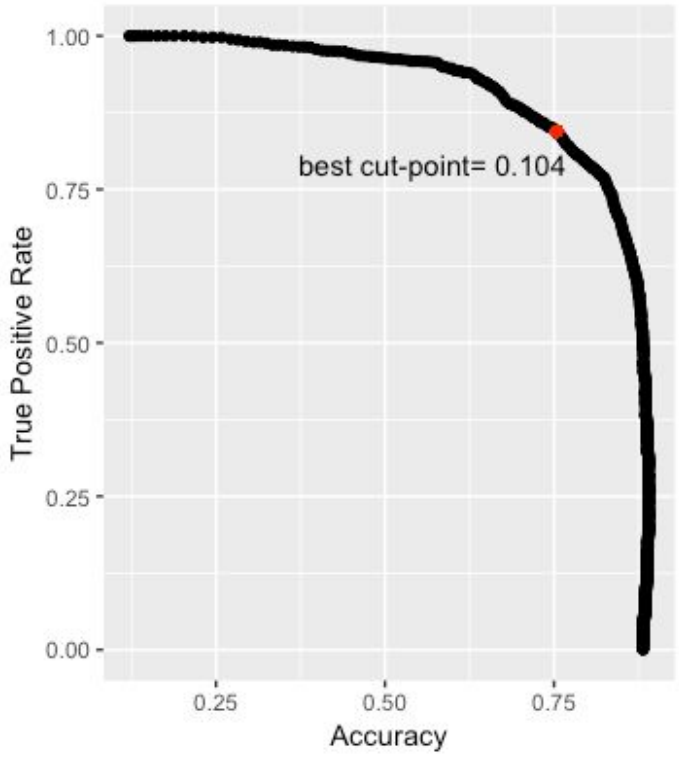
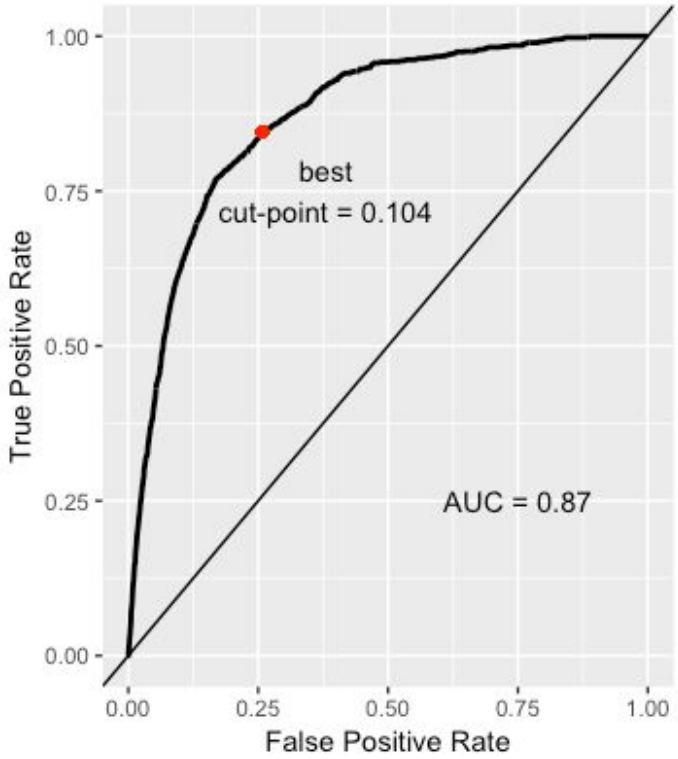


Fig. 10 depicts the recursive feature elimination algorithm, where k = 10. For each of the 10 folds, a full logistic regression model is fit using all S predictors, where S = 16. From this model, the predictors are ranked according to the absolute value of their t-statistics. After this ranking process, (1) a loop is entered: for S_i 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 S_i having the highest mean AUC is then selected as the best logistic regression model. A final logistic regression model is then fitted using the S_i predictors against the full training set.

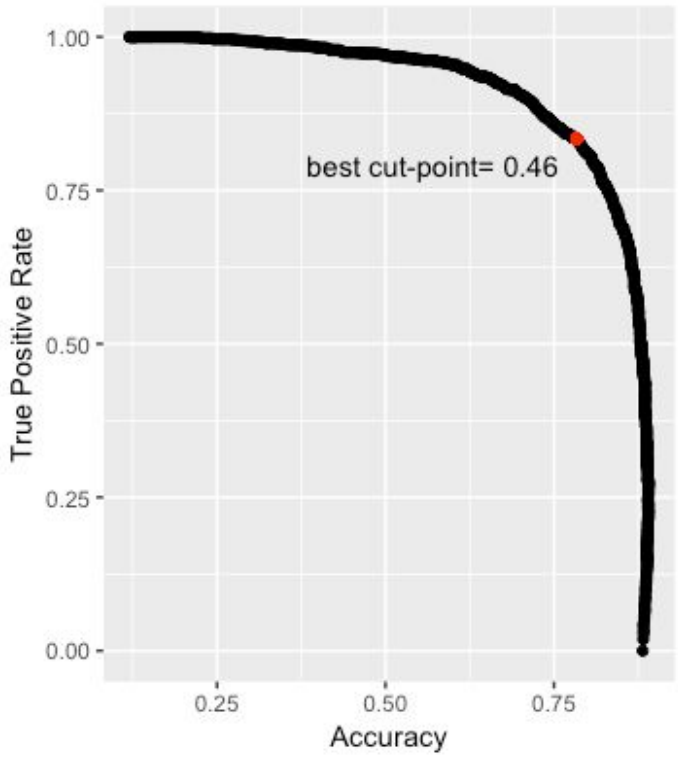
In choosing an optimal cut-point for the class probabilities, our model results in a trade-off between the overall accuracy rate of predictions and the true positive rate (TPR) of predictions. This makes sense, for if we desired to have a perfect TPR, we would simply predict all new data as producing 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

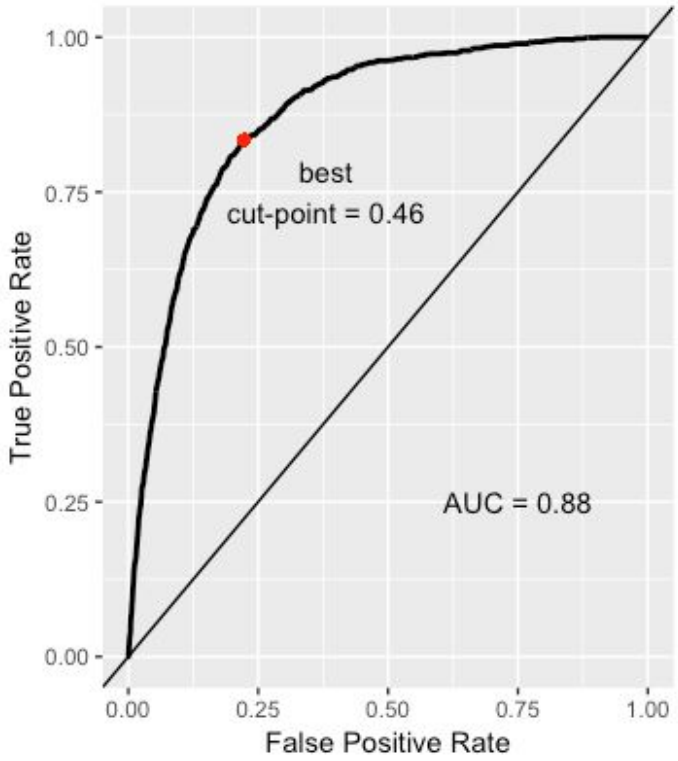
model, which shows the receiver operating characteristic of the model, while Fig. 11b shows its area-under-curve statistic of 0.87.

Fig. 11a. True Positive Rate vs. Accuracy	**Fig. 11b.** Receiver Operating Characteristic
	

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.

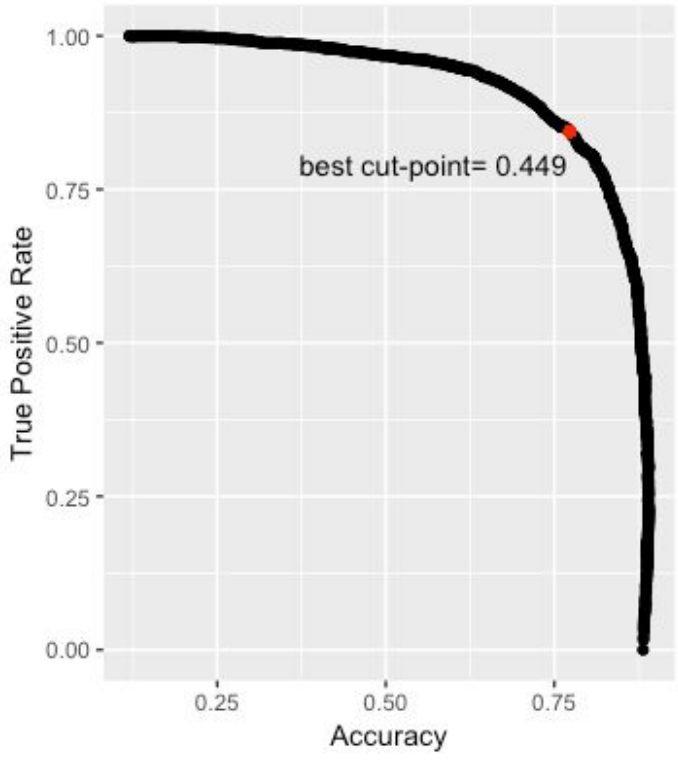
Fig. 12. Accuracy vs. TPR and ROC Plots for Class Balancing Sampling Method Models	
Upsample Model - Accuracy vs. TPR	Upsample Model - ROC

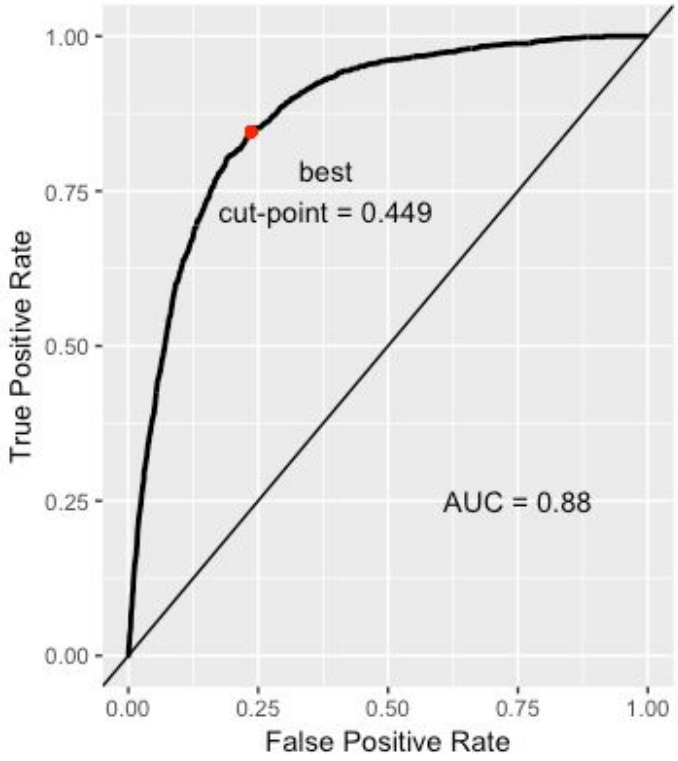




Downsample Model - Accuracy vs. TPR

Downsample Model - ROC





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

```
> summary(glm_rfe_ROC$fit)

Call:
glm(formula = Class ~ ., family = "binomial", data = tmp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5762  -0.4484  -0.2874  -0.1638   3.6014

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.280e+00  2.377e-01  -9.591  < 2e-16 ***
last.contact.duration  3.913e-03  7.043e-05  55.561  < 2e-16 ***
has.housing.loan    -1.068e+00  4.385e-02 -24.359  < 2e-16 ***
contact.type        -6.589e-01  3.151e-02 -20.911  < 2e-16 ***
contacts.num.prev    1.108e-01  9.171e-03  12.083  < 2e-16 ***
days.passed          3.509e-03  3.181e-04  11.029  < 2e-16 ***
contacts.num        -1.305e-01  1.156e-02 -11.286  < 2e-16 ***
has.personal.loan   -7.102e-01  6.582e-02 -10.791  < 2e-16 ***
highest.educ         2.019e-01  2.650e-02   7.620 2.53e-14 ***
prev.outcome         2.232e-01  3.427e-02   6.514 7.32e-11 ***
last.contact.moy     3.843e-02  6.383e-03   6.021 1.73e-09 ***
marital.status       2.065e-01  3.577e-02   5.773 7.79e-09 ***
age                  7.623e-03  1.934e-03   3.942 8.08e-05 ***
avg.annual.balance   1.959e-05  5.083e-06   3.854 0.000116 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24405  on 33907  degrees of freedom
Residual deviance: 18067  on 33894  degrees of freedom
AIC: 18095

Number of Fisher Scoring iterations: 6
```

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

```
> vif(glm_rfe_ROC$fit)
last.contact.duration      has.housing.loan          contact.type     contacts.num.prev
             1.078638              1.188268              1.143578              1.463514
          days.passed          contacts.num     has.personal.loan          highest.educ
             3.570160              1.034276              1.015503              1.041952
         prev.outcome      last.contact.moy        marital.status                   age
             3.805457              1.118619              1.272479              1.342984
   avg.annual.balance
             1.026240
```

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

```
> varImp(glm_rfe_ROC$fit)
                        Overall
last.contact.duration 55.561368
has.housing.loan      24.359323
contact.type          20.911301
contacts.num.prev     12.082811
days.passed           11.028599
contacts.num          11.285558
has.personal.loan     10.791241
highest.educ           7.620280
prev.outcome           6.514016
last.contact.moy       6.021243
marital.status         5.772943
age                    3.941878
avg.annual.balance     3.853978
```

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

```
Logistic Regression Model

 lrm(formula = 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, data = bank.validate,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs          11303    LR chi2    2130.34    R2        0.332    C         0.871
  1            9964    d.f.            13    g         1.632    Dxy       0.741
  2            1339    Pr(> chi2) <0.0001    gr        5.113    gamma     0.744
 max |deriv| 3e-08                          gp        0.140    tau-a     0.155
                                            Brier     0.080
```

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

```
> my.valid
          index.orig training    test optimism index.corrected    n
Dxy           0.7416   0.7436  0.7407   0.0029          0.7386 1000
R2            0.3323   0.3354  0.3293   0.0060          0.3262 1000
Intercept     0.0000   0.0000 -0.0217   0.0217         -0.0217 1000
Slope         1.0000   1.0000  0.9851   0.0149          0.9851 1000
Emax          0.0000   0.0000  0.0073   0.0073          0.0073 1000
D             0.1884   0.1903  0.1866   0.0037          0.1847 1000
U            -0.0002  -0.0002  0.0001  -0.0003          0.0001 1000
Q             0.1886   0.1904  0.1864   0.0040          0.1846 1000
B             0.0803   0.0799  0.0804  -0.0005          0.0808 1000
g             1.6319   1.6495  1.6242   0.0253          1.6066 1000
gp            0.1395   0.1401  0.1390   0.0011          0.1384 1000
```

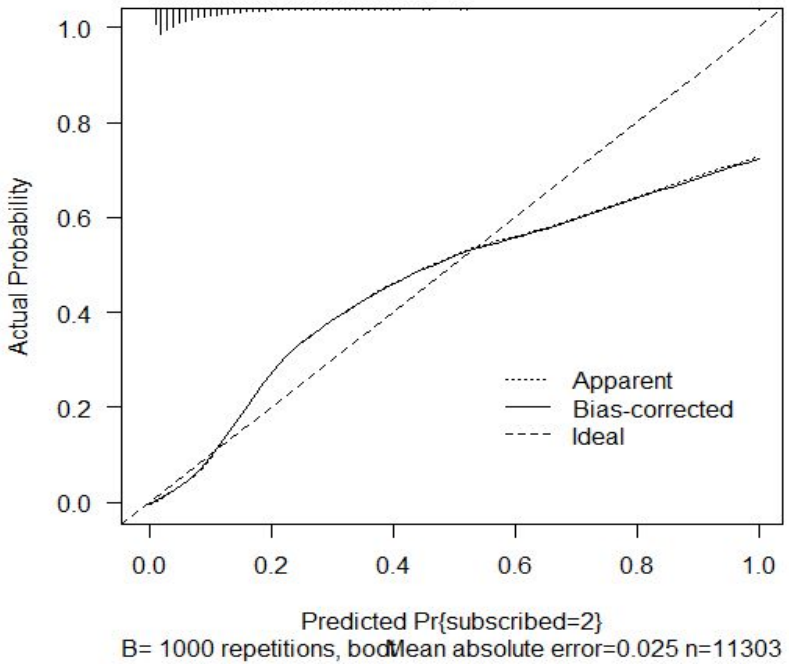
Comparisons were conducted with bias-corrected measures, indicated in the column above entitled "index.corrected", which adjusts the index in accordance with the model's "optimism".

Calibration Curve using Resampling

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

Fig. 20. Calibration Curve

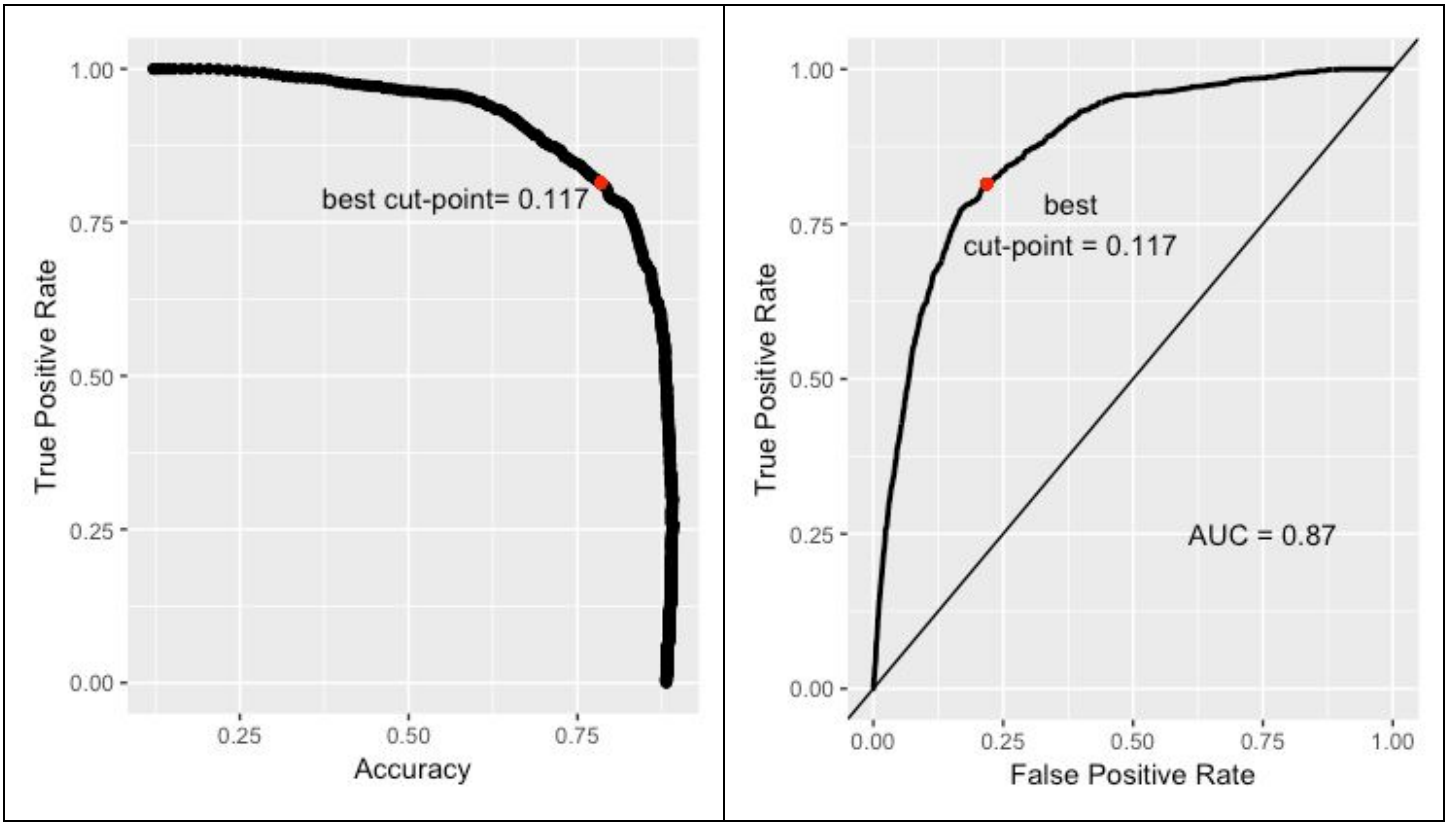
```
> my.calib <- calibrate(mod_1, method="boot", B=1000)

> par(bg="white", las=1)
> plot(my.calib, las=1)

n=11303   Mean absolute error=0.025   Mean squared error=0.00178
0.9 Quantile of absolute error=0.074
```



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



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
https://stats.idre.ucla.edu/r/dae/logit-regression/
http://www.researchmanuscripts.com/July2014/2.pdf
www.stat.ufl.edu/~winner/sta6127/chapter15c.ppt
www.statstools.com
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-->
(ノ°□°)ノ ┻━┻
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 SPLIT
##-------------------------------------

# 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