Predicting Client Banking Using Portuguese Banking Institution Campaign Data

University of North Carolina at Chapel Hill

BIOS 735 - Spring 2020

East Lee, Tianyou Luo, Saifa Pirani, & Josh Rutsohn

Introduction

Commonly used in the banking industry, predictive analytics have become a popular method to increase transparency and understanding of risk exposure, to improve the ability to target products and services for customers, to enhance product pricing, track performance of products and services, and to identify "high-potential" prospects and customers (Spenser & Sohail, 2012). The goal of predictive analytics is to use machine learning, data mining, and artificial intelligence techniques to understand the unknown underlying function of the data and often to derive decision support systems (DSS) that support managerial decision. DSS are information technology systems that use business data to support decision making for banks. They enable banks to rethink their marketing strategy and target key customers to optimize profits (Moro, Cortez, & Rita, 2014).

Predictive analytics have become a vital asset for banks everywhere. For example, in order to study the discounts its private bankers were offering to customers, a US bank performed predictive analytics. Results showed that while the bankers claimed that they were only offering valuable discounts, the discounts were actually unnecessary and costing the bank essential revenue. When the bank implemented changes using DSS to identify essential discounts, revenue rose by 8 percent in just a few months (Garg, Grande, Macias-Lizaso Miranda, Sporleder, & Windhagen, 2017). Another bank in Europe, applied machine-learning algorithms to predict which currently active customers are likely to reduce their business with the bank in an attempt to counter a shrinking customer base. Predictive analytics resulted in DSS that were used in a marketing campaign to target key customers. By implementing these changes, the bank reduced churn by up to 15 percent (Gard et al., 2017).

These are only a few cases of the potential and importance of predictive analytics for banks. The applications of predictive analytics are endless. Particularly, predictive analytics can help banks identify the success of a telemarketing program. One such telemarketing program was implemented in a Portuegese bank between May 2008 and November 2010. The bank collected information on customer demographics, telemarketing attributes, and economic influence variables (Moro et al., 2014). In the primary study, Moro et al, (2014) used these data to model telemarketing success for selling bank long-term deposits.They analyzed over a 150 set of features using a forward stepwise regression to derive a final model with 22 features. They used the final model and applied it to four predictive analytics techniques: logistic regression, decision trees, neural network, and support vector machine. When they compared the four models using the area of the receiver operating characteristic curve (AUC) and the area of the LIFT cumulative curve (ALIFT) as metrics, they found that the neural network model presented the best results.

While the primary study produced meaningful results, the study had many limitations. For example, they neglected to address the issue of imbalanced data. In predictive analytics there are two key assumptions: a) the goal of the machine learning algorithm is to maximize accuracy and b) the classifier will operate on data drawn from the same distribution as the training data. With imbalanced data, these key assumptions lead to unsatisfactory classifiers (Provost, 2000). Further, they only used the semi-automatic stepwise forward selection process to derive their final model. It is commonly known that the stepwise forward selection process has many limitations and when compared to automatic techniques such as LASSO and least angle regression (LAR) stepwise forward selection greatly underperforms (Efron, Hastie, & Tibshirani, 2004; Folm, 2018; Hastie & Friedman, 2001). The authors also failed to cross-validate their

results. It is vital to cross-validate a model to assess predictive error and ensure reproducibility of results (Folm, 2018; Krstajic, Buturovic, Leahy, & Thomas, 2014).

The primary objectives of the current study are to address the limitations of the primary study and to use predictive analytics to identify a data-driven model that can predict whether a client will open a term bank account (success) or not (failure). First, we address the issue of imbalanced and missing data by applying an optimal method of imputation. Next, we employ an automatic feature selection process by using the LASSO regularization method and our own feature selection algorithm.Finally, we apply logistic regression and the support vector machine technique to build and to cross-validate our classification model.

Method

Participants and Procedures

The data ($N = 41{,}188$) come from a direct marketing campaign conducted by a Portuguese banking institution. This campaign aimed at encouraging clients to open a term deposit--a financial deposit that has a set maturity date. These marketing campaigns were completed over the phone, and sometimes multiple calls were made in order to ascertain a yes or no response to a term deposit. Each record consists of a binary outcome--success or failure in convincing the client of opening a term deposit--as well as telemarketing attributes, client demographics, and economic influence variables (see Table 1 & 2). These economic influence variables, such as the unemployment rate consumer confidence index, etc., were originally collected from external data courtesy of the Portuguese Republic statistical website. These data were collected between May 2008 and November 2010.

Data Imputation

Close examination of the data revealed that the data contained highly imbalanced and missing completely at random data (see Table 1). To address this issue we chose to examine and compare two methods of imputation: MICE and missForrest. Using a simulation study, we aimed to identify the best algorithm for the sample data. Performance of both methods were judged based on a low overall proportion of falsely classified entries (PFC; see Table 3). A detailed description of each algorithm is provided before the simulation study is discussed below

MICE. Let $Y = (Y_1, ..., Y_p)$ be all variables and Y_{-i} contain all the variables of Y, except the ith component. $Y_i|Y_{-i}$ follows some conditional distribution with parameters θ. We should specify this conditional distribution and it will be our model. Then posterior distribution of theta is derived from the conditional distribution we set.

$$\begin{aligned} \boldsymbol{Y} &= (Y_1, \ldots, Y_p)^T \\ \boldsymbol{Y}_{-i} &= (Y_1, \ldots, Y_{i-1}, , Y_{i+1}, \ldots, Y_p)^T \\ Y_i|\boldsymbol{Y}_{-i} &\sim f(Y_i|\boldsymbol{Y}_{-i}, \boldsymbol{\theta}) \end{aligned}$$

When current state is t, $\theta_1^{(t+1)}$, the parameters of conditional distribution of Y_1 is drawn from the posterior distribution, then $Y_1^{(t+1)}$ of the next state follows the conditional distribution with the updated $\theta_1^{(t+1)}$. Now we got updated $Y_1^{(t+1)}$, then $\theta_2^{(t+1)}$ at the stage $t+1$, follows the posterior distribution given updated $Y_1^{(t+1)}$. The remaining variables and parameters follow the same logic.

$$
\begin{aligned}
\boldsymbol{\theta_1}^{(t+1)} &\sim P(\boldsymbol{\theta_1}|y_1^{(t)}, y_2^{(t)}, \dots y_p^{(t)}) \\
Y_1^{(t+1)}|\cdot &\sim f(y_1|\boldsymbol{\theta_1}^{(t+1)}, y_2^{(t)}, \dots, y_p^{(t)}), \\
\boldsymbol{\theta_2}^{(t+1)} &\sim P(\boldsymbol{\theta_2}|y_1^{(t+1)}, y_2^{(t)}, \dots y_p^{(t)}) \\
Y_2^{(t+1)}|\cdot &\sim f(y_2|\boldsymbol{\theta_2}^{(t+1)}, y_1^{(t+1)}, y_3^{(t)}, \dots, y_p^{(t)}), \\
&\vdots \\
\boldsymbol{\theta}_p{}^{(t+1)} &\sim P(\boldsymbol{\theta_2}|y_1^{(t+1)}, y_2^{(t+1)}, \dots y_p^{(t)}) \\
Y_p^{(t+1)}|\cdot &\sim f(y_p|\boldsymbol{\theta_3}^{(t+1)}, y_1^{(t+1)}, y_2^{(t+1)} \dots, y_{p-1}^{(t+1)}),
\end{aligned}
$$

The MICE algorithm starts by specifying univariate conditional distributions, and filling in missing values with samples from the observed values of the same variable. Next, the algorithm draws parameters of conditional distribution from the posterior distribution and updates missing values using the obtained parameters. This procedure is repeated for each *p* variable M times. Steps for the MICE algorithm are described below:

1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \dots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \dots, M$.
4. Repeat for $j = 1, \dots, p$.
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \dots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \dots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

missForest. One shortcoming of MICE is that a prior specification of a conditional distribution is necessary to estimate and draw parameters. Therefore, it is based on a parametric model. Like MICE, missForest imputes missing values on a variable-by-variable basis, but the main difference between MICE and missForest is that missForest is based on a random forest instead of conditional distributions. Therefore, estimating parameters from a conditional distribution needed in missForest. Steps for missForest are described below:

1. Make initial guess for missing values;
2. $\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
3. **while** not γ **do**
4. $\quad \mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ store previously imputed matrix;
5. $\quad$ **for** s in $\mathbf{k}$ **do**
6. $\quad\quad$ Fit a random forest: $\mathbf{y}_{\text{obs}}^{(s)} \sim \mathbf{x}_{\text{obs}}^{(s)}$;
7. $\quad\quad$ Predict $\mathbf{y}_{\text{mis}}^{(s)}$ using $\mathbf{x}_{\text{mis}}^{(s)}$;
8. $\quad\quad \mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ update imputed matrix, using predicted $\mathbf{y}_{\text{mis}}^{(s)}$;
9. $\quad$ **end for**
10. $\quad$ update γ.
11. **end while**
12. **return** the imputed matrix $\mathbf{X}^{\text{imp}}$

Simulation Study. For the simulation study, we first generated simulation data whose missing pattern was close to the original data as much. Next, we extracted rows without any missing values and let ‘complete’ the extracted dataset. Second, NA's were artificially produced. Since we wanted the generated simulation data to be similar to the original data as much as possible, the proportion of NA in simulation data was the proportion of the original data set. For example, In the original data, the categorical variable job has missing values at a rate of 0.8%.

Therefore, NA's are produced in the `complete` data at a rate of 0.8%. The data simulation process was completed five times.

MICE and missForest were applied to the data and the PFC's were extracted from the results. The PFC indicates how many imputed values are different from the true values (1-Accuracy). We can calculate PFC because all variables with missing were categorical and we know the true values

$$PFC = \frac{\Sigma_{j=1}^{p}\Sigma_{i=1}^{n} I(Y_{imp} \neq Y_{true})}{\#NA}$$

The PFC's are given in Table 3. We can see that missForest shows uniformly smaller PFC's compared to MICE. In addition, the running time of MICE took 11 times longer than missForest.

Though both MICE and missForest can effectively handle missing not at random (MNAR) data, it is known that their performance is affected by missing pattern of data. Their results become more accurate when the data are missing at random (MAR). Since MAR is an assumption that is impossible to verify statistically, further investigation is needed. If the MAR assumption is true, when a model that has missing values in a dependent variable is estimated, the residuals should be independent of whether the value of the dependent variable is an imputed value or the original one. Based on these ideas, we designed the following diagnosing step: We fitted a multinomial logistic regression model using the data imputed by missForest. The dependent variable was each variable that had missing values before imputation. Classification

errors for each model were examined based on the original data and the imputed data. The results are given in Table 4.

The multinomial logistic regression models on each variable have generally larger classification error in the original data samples. The results indicate that the MAR assumption might not hold in this data. However, missForest is known for relatively better performance than other imputation methods under MNAR setting as well. Therefore, the data set imputed by missForest will be used for the current study.

Data Analytic Plan

Feature Selection. In order to build a model to predict if a client opens a term bank account (success) or not (failure), we first applied logistic regression analyses to the data using the LASSO regularization method with telemarketing attributes, client demographics, and economic influence variables collected in the campaign as potential predictors. The analyses were employed using the package *glmnet* in R (Friedman, Hastie, & Tibshirani, 2010; R Core Team, 2014) and with teampkg4 (see Appendix A). The area of the receiver operating characteristic curve (AUC) and five-fold cross validation procedure were used to select the tuning parameter lambda and determine the final model.

Predictive Analytics. After selecting the final model from the feature selection phase, logistic regression and the support vector machine technique were employed to build the predictive model. Support Vector Machine (SVM) is a machine learning method for classification and regression. The technique finds the “best” margin (distance between the line and the support vectors) in the kernel space that separates the classes, thereby reducing the risk

of error. Compared to other classification techniques, SVM works well with high dimensional data (Meyer, Leisch, & Hornik, 2003). However, with large data sets, the computational burden for the SVM technique increases, thereby it takes a long time to compute (Moro et al., 2014). Thus, both logistic regression and SVM analyses were conducted using 10,000 randomly chosen subjects to compare the results, and the rest of the data will be naturally used as a test set. Kappa statistics were used to select tuning parameters for SVM. Results calculated on the full dataset using logistic regression and five-fold cross validation are reported.

For the logistic regression results, an alpha level of .05 was chosen to interpret the significance of predictors. For SVM, we used radial basis kernels, because surprisingly, linear kernels took even much longer running time and experienced convergence issues. The accuracy, Kappa, specificity, and sensitivity of both models were examined.

Results

Feature selection results are presented in Figures 1 & 2. Figure 1 compares the parameter estimates of the logistic regression model obtained from *glm* and from logistic() in *teampkg4*. Parameter estimates were identical for both methods, which shows that our self-written function has performed optimization correctly and converged successfully. Additionally, Figure 1 displays the time in seconds it took for the code to execute. Clearly, results display that compared to *glm*, logistic() from *teampkg4* took a longer time to execute. The relationship between the AUC and the lambda value obtained from the LASSO regularization method is showcased in Figure 2. Here, the AUC monotonically decreases as the value of lambda increases. Thus, we elected to retain all variables in the final model.

The cross-validation results for the logistic regression and the support vector machine model are presented in Table 5. For the logistic regression model, accuracy was .90 and Kappa was 0.30. Interestingly, while the sensitivity, the ability of the model to correctly identify positive cases (true successes) was 0.98, the specificity, the ability of the model to correctly identify true negative cases (true failures) was 0.24. Similarly, for the SVM model, accuracy was .90, Kappa was 0..28, sensitivity was 0.99, and specificity was 0.21.

The logistic regression model results on the full dataset are presented in Table 6. Note, the model contained many categorical variables, so it was difficult to interpret the results. The results in Table 6 convey parameter estimates relative to clients with an administrative job, who are divorced, have a basic education (4 year), have a housing loan, have a personal loan, were contacted via mobile, and the previous marketing campaign was a failure for that client. Since the model results are compared to a unique subpopulation of clients, the results for the model reported should be interpreted with caution. For example, for one of the predictors, with the subpopulation described above reference, while holding all other variables constant, blue-collar workers are significantly less likely by 22% than administrative workers to open a term bank.

Discussion

The current study aimed to address limitations of the primary study and to use predictive analytics to create a model to predict if a client will open a term bank account. The data were first cleaned and imputed using the missForest imputation technique. A subset of 10,000 random subjects was selected to account for the large computational time the SVM technique requires to analyze large data sets. LASSO regularization was used to aid with feature selection. Once the final model was selected, the data were reviewed using logistic regression and SVM.

Cross-validation results using k - fold cross-validation, where k was set to 5 for both models were reported. Finally, the logistic regression parameter estimates were interpreted.

During the model selection phase, both *glmnet* and logistic() from *teampkg4* were used to aid feature selection. While parameter estimates from both were identical, logistic() from *teampkg4* took longer to execute. The LASSO regularization process indicated that there was no need to drop any features, thus we retained all variables as predictors.

When logistic regression and SVM were applied to the data, both models had a high accuracy rate and a low Kappa metric. This may be because Kappa is known as an inadequate metric in imbalanced data (Delgado & Tibau, 2019).This phenomenon is known as the *Kappa Paradox*. The *Kappa Paradox* is often a product of the prevalence paradox- when probability of chance agreement is high (high accuracy) low values of Kappa are produced and/or of the bias paradox -- high Kappa values occur because of imbalanced marginal distribution. Byrt, Bishop, & Carlin (1993) suggest that a correction for bias and prevalence should be applied to Kappa before interpretation. However, a limitation of this study was that the bias correction was not applied. Thus, we elected to only interpret the accuracy statistic.

The high sensitivity and a low specificity rates, for both models, indicate that the predictive model will incur high false-positive cases (Akobeng, 2007; Stephanie, 2014). Since both models have nearly 99% sensitivity, we can conclude that the predictive models are credible and valuable for telemarketing campaign managers to accurately identify clients that would open a term bank account. However, the low specificity rate of between 21-24% is mildly concerning. While there will be no high revenue loss if clients who are misclassified as “successes” are targeted during a campaign, a large amount of false-positives can lead to a loss of time and

man-power that could be spent targeting clients properly classified as "successes." Overall, while the classification models are not perfect in terms of accuracy, specificity, and sensitivity, we still conclude that the models are credible and highly useful predictive models.

Parameter estimates for the logistic regression model were interpreted at an alpha of .05. Blue-collar workers, retired workers, service workers, students, clients contacted by phone, clients not contacted on the previous campaign, clients who opened a bank term account, the employment variation rate, the consumer price index, and the consumer confidence index were all significant predictors of predicting "success" in the model reported below. As aforementioned, these results should be considered with interpretation and implication of categorical variables in mind.

Although the current study improves upon the primary study, the current study still contains many limitations. Due to the high computational time burden, we were only able to analyze 10,000 randomly chosen subjects. We intend to use our package to reduce the computational burden by writing the code in *Rcpp* so that we can use the full data. Further, all available metrics to assess classification models and all available classification models (e.g., neural network) were not used in the current study.

Overall, the implications of the current study are numerous. The logistic regression and SVM models lay the groundwork to explore further outcomes. By selecting different categorical variables as reference variables, results can be uniquely tailored to target clients as a marketing strategy. Our next step will be to automate these results in decision support systems. We intend for these results to provide key information for managerial decisions, subsequently leading to optimization of the telemarketing campaigns and an increase in profits.

References

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatrica, 96*(3), 338-341.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423-429.

Delgado, R., & Tibau, X. A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one, 14*(9), 1 -26.

Flom, P. (2018, December 11). Stopping stepwise: Why stepwise selection is bad and what you should use instead.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software, 33*(1), 1–22.

Garg, A., Grande, D., Macias-Lizaso Miranda, G., Sporleder, C., & Windhagen, E. (2017, April). Analytics in banking: Time to realize the value. Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value

Hastie, R. T. & Friedman, J. (2001), *The elements of statistical learning*, Springer-Verlag, New York.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics, 6*(1), 1-15.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing, 55*(1-2), 169-186.

Moro, S., Cortez, P. & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems. 62*, 22-31

Provost, F. (2000, July). Machine learning from imbalanced data sets 101. In Proceedings of the AAAI'2000 workshop on imbalanced data sets (Vol. 68, pp. 1-3). AAAI Press.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Spenser, P. & Sohail, O. (2012). Banking Analytics: a three minute ride. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-us-ba-bankinganalytics3minguide.pdf

Stephanie. (2014, May 19). Sensitivity vs Specificity and Predictive Value. Retrieved from https://www.statisticshowto.com/sensitivity-vs-specificity-statistics/

Tables

Table 1. Characteristics of the Data ($N = 41188$)

Variable	N (%)	N missing
Job		330 (1%)
Administrative	10422 (25%)	
Blue Collar	9254 (23%)	
Entrepeneur	1456 (4%)	
Housemaid	1060 (3%)	
Management	2924 (7%)	
Retired	1720 (4%)	
Self-Employed	1421 (4%)	
Services	3969 (10%)	
Student	875 (2%)	
Technician	6743 (16%)	
Unemployed	1014 (3%)	
Marital Status		80 (< 1%)
Divorced	4612 (11%)	
Married	24928 (61%)	
Single	11568 (28%)	
Education Status		1731 (4%)
Basic Education (4 yr)	4176 (10%)	
Basic Education (6 yr)	2292 (6%)	
Basic Education (9 yr)	6045 (15%)	
High School	9515 (23%)	
Illiterate	18 (< 1%)	
Professional Degree	5243 (13%)	
University Degree	12168 (30%)	

Table 1 Cont.

Defaulted Credit		8597 (21%)
	Yes	3 (< 1%)
	No	32588 (79%)
Housing Loan		990 (2%)
	Yes	21576 (52%)
	No	18622 (45%)
Personal Loan		6248 (15%)
	Yes	6248 (15%)
	No	33950 (82%)
Contact Method		0 (0%)
	Mobile	26144 (64%)
	Landline	15044 (36%)
Previous Marketing Campaign Success		0 (0%)
	Failure	4252 (10%)
	Non-existent	35563 (86%)
	Success	1373 (3%)
Previous Contact		0 (0%)
	No	39673 (96%)
	Yes	1515 (4%)

Table 2. *Variable Mean and Standard Deviation*

Variable	Mean (SD)
Age	40 (10.4)
Duration of Last Contact (seconds)	258.3 (259.3)
Number of Contacts During Campaign	2.6 (2.7)
Number of Contacts from Previous Campaign	0.17 (0.49)
Employment Variation Rate (quarterly)	0.08 (1.57)
Consumer Price Index (Monthly)	93.6 (0.6)
Consumer Confidence Index (Monthly)	-40.5 (4.6)
Euribor 3 month rate	3.6 (1.7)
Number of Employees (quarterly)	5167 (5191)

Table 3 - *PFC for MICE and missForest*

Simulation	missForest	MICE
1	0.1328	0.1577
2	0.1323	0.1655
3	0.131	0.1628
4	0.1288	0.1609
5	0.1325	0.1587

Table 4. *Classification Error for Original Samples and Imputed*

Variable	Original	Imputed
Job	0.48	0.55
Marital	0.32	0.23
Education	0.47	0.23
Default	0.00	0.00
Housing	0.44	0.24
Loan	0.16	0.06

Table 5. *Sensitivity, Specificity, and Kappa Statistics*

Logistic Regression			SVM		
	Reference			Reference	
Predictor	0	1	Predictor	0	1
0	27241	2678	0	27281	2755
1	445	824	1	405	747
Accuracy	0.900		Accuracy	0.899	
Kappa	0.304		Kappa	0.281	
Sensitivity	0.984		Sensitivity	0.985	
Specificity	0.235		Specificity	0.213	

Note : SVM = Support Vector Machine, 0 = Failure to open term account, 1 = Success to open term account

Table 6. *Logistic Regression Model*

Variable	β	*p*
(Intercept)	-107.60	0.00
age	0.00	0.65
Blue-Collar	**-0.22**	**0.00**
Entrepeneur	-0.12	0.25
Housemaid	-0.11	0.39
Management	-0.12	0.12
Retired	**0.32**	**0.00**
Self-Employed	-0.07	0.47
Services	**-0.17**	**0.02**
Student	**0.30**	**0.00**
Technician	0.00	0.97
Unemployed	-0.02	0.89
Married	0.03	0.64
Single	0.10	0.12
Basic Edu (6 year)	0.08	0.40
Basic Edu (9 year)	-0.07	0.40
High School	0.02	0.77
Illiterate	0.99	0.12
Prof De, urse	0.04	0.66
Universi ee	0.15	0.05
Housing Loan (No)	-0.02	0.59
Perspnal Loan (Yes)	-0.04	0.42
Contact (phone)	**-0.91**	**< 2e-16**
Previous Campaign	**-0.04**	**0.00**
Days	**0.00**	**0.00**
Previous Contanct	-0.07	0.23
Previous Marketing None	**0.49**	**0.00**
Previous Marketing Success	**0.75**	**0.00**
Employment Variation Rate	**-0.73**	**< 2e-16**
Consumer Price Index	**1.19**	**< 2e-16**
Consumer Confidence Index	**0.04**	**0.00**
Euribor 3 Month	0.01	0.90
Number of Employees	0.00	0.65

Note: Bolded values indicate significant predeictors at $p < .05$

Figures

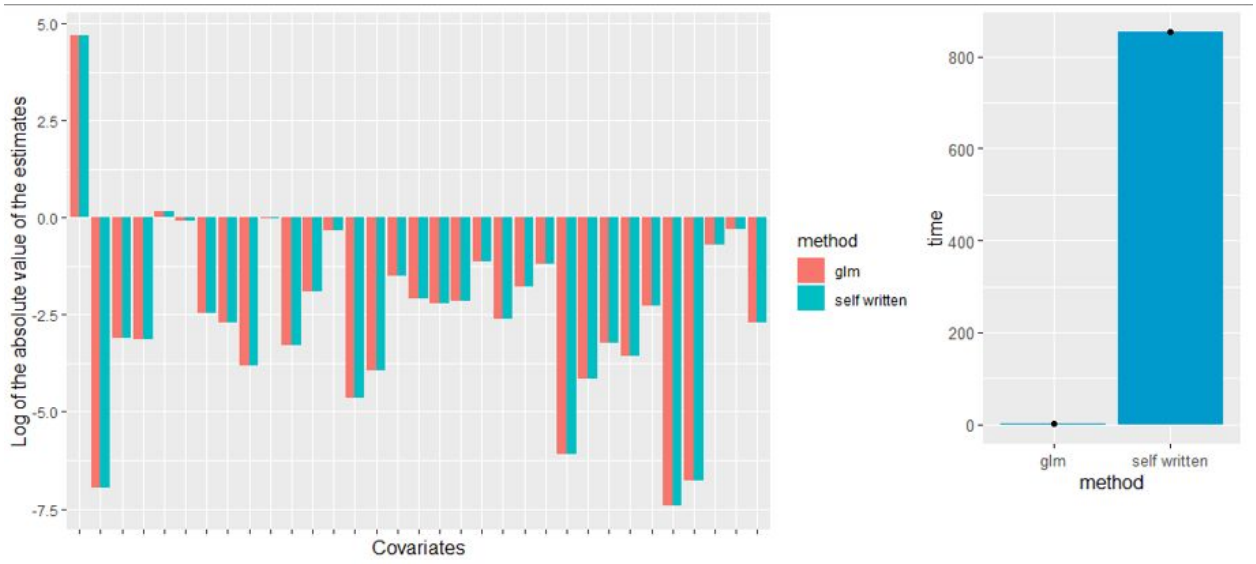

Figure 1. GLM and team4pkg Feature Selection Performance

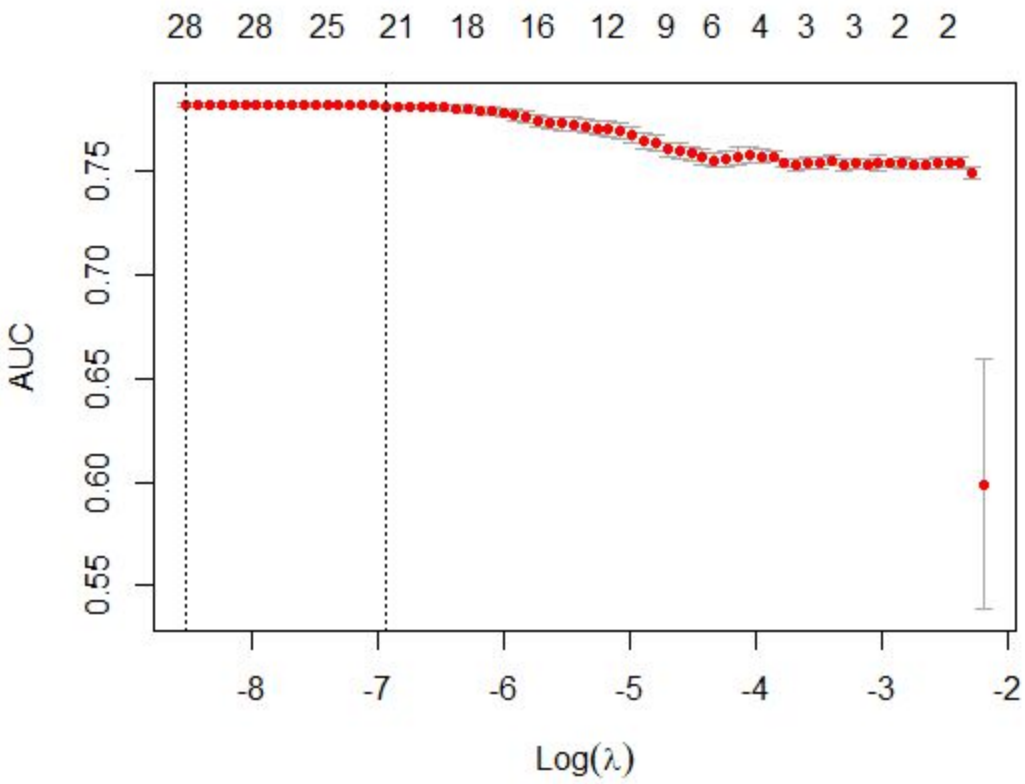

Figure 2. LASSO Regularization Results

Appendix A

The logistic regression optimization (Module 2) and prediction (Module 3) functions are found in the R package **team4pkg**. Four functions can be found associated with these tasks: deriv1.logistic(), deriv2.logistic(), logistic(), and predict.logistic(). A final function, to.factor(), is used simply to convert a list of features from numeric to factor in order to perform imputation and feature selection. The functions predict.logistic(), deriv1.logistic(), and deriv2.logistic() are used to produce the MLE μ, first derivative, and second derivative for the function logistic(). The logistic() function then uses the Newton-Raphson algorithm to maximize the estimate of π. This function uses two stopping criteria that may be specified: maximum iterations and tolerance. The function will also suppress the iterations given option 'verbose = 0'. These functions can be explored further using the R documentation in the "man" folder.

Several tests can be run to determine whether the package is working as intended. Including standard multiplication and addition tests, the package tests for conformable dimensions for matrix multiplication and matrix solving.

A separate folder was produced in the package, "projectcode", that provides commented R code on how to replicate the results used within this paper and its associated presentation. These files are not necessary to run the logistic regression optimization or prediction code, but rather it is only provided in order to show how the results for the imputation and prediction methods were derived.

The logistic regression optimization code in **team4pkg** does not have any dependencies and can be run entirely in base R. The separate folder for replicating the project code has several dependencies that can be viewed as imports in the DESCRIPTION file. The only external

requirements to replicate the results are the data files bank-additional-full.csv and imputedData.csv, both found on the GitHub repository Team4. These data were not included in the package due to their large sizes. The compiled version **team4pkg** can also be found in the Team 4 repository titled team4pkg_1.0.0.tar.gz.