Identify Potential Deposit Subscribers of The Portuguese Retail Bank Market

Contents

1 Introduction **2**

1.1 Background . . . 2

1.2 Statistical Objective . . . 2

2 Exploratory Exploration **2**

2.1 Data Processing . . . 2

2.2 Exploratory Analysis . . . 2

3 Binary Logistic Regression Model **4**

3.1 General Model Form and Variable Selection . . . 4

3.2 Model Assumptions . . . 5

3.3 Model Validation . . . 5

3.4 Model Diagnostics . . . 5

4 Random Forests Model **6**

4.1 General Model Form and Variable Selection . . . 6

4.2 Model Assumptions . . . 6

4.3 Model Validation . . . 7

5 Support Vector Machine (SVM) Model **7**

5.1 General Model Form and Variable Selection . . . 7

5.2 Model Assumptions . . . 7

5.3 Model Validation . . . 7

6 Model Comparison **8**

6.1 Single Model Performance . . . 8

6.2 Composite Model Performance . . . 9

7 Conclusions, Limitations, and Future Research **11**

8 References **12**

9 Appendix **13**

Team ID: 4

Kenneth Broadhead

Koral Buch

Min Kim

Nanhao Chen

1 Introduction

1.1 Background

A Portuguese retail bank initiated a telemarketing campaign from 2008 to 2013 aiming to maximize the subscription of new clients to a long-term deposit. This campaign used a direct method of marketing through cellphone or telephone. A subset of the data, of the years 2008 to 2010, was uploaded in February 2012 to the UC Irvine Machine Learning Repository and publicly available for research purposes [1].

1.2 Statistical Objective

The report focuses on the construction of a model for predicting whether or not a banking client will subscribe to a long-term deposit. In order to investigate the potential for constructing a predictive model, some summary statistics and visual representations are included and explained. To formally assess the potential to predict the success or failure of telemarketing for subscription to a long-term deposit, predictive models such as logistic regression, random forest, and support vector machine (SVM) are utilized [2]. Corresponding model diagnostics and comparisons of the models' performance are discussed.

Finally, we note that there are many ways the quality of a predictive model may be assessed. One can use overall accuracy and similar metrics like the AUC (see section 3.1); or one can focus on more specific aspects of a model's predictive performance. In this case, the risk associated with type I and type II errors are not equal. If a client is contacted who isn't going to sign on to a deposit, the bank simply wastes a few minutes of their time; but if someone who is likely to sign on is missed, this potentially costs the bank money. Thus, in addition to the standard overall accuracy measures, we will also make special note of the type II error rate for each model.

2 Exploratory Exploration

2.1 Data Processing

Since our ultimate goal is constructing predictive models, we utilize the full data set, with all 41,188 observations, available at the repository. This provides us with as many potentially relevant predictors as possible, as well as enough data to split into training and validation sets. Due to the large size of the data set, we split the data (randomly) 50/50 into training and validation sets [3]. We split the data into these sets after processing and exploratory analysis. Initial exploration of the data shows there are several categorical predictors that have 'unknown' as a level. Since this lack of knowledge could be potentially useful to bank telemarketers, we treat these missing values as factors.

We removed the "duration" variable, since the duration of a call between a bank and client isn't known in advance, and would be unhelpful in building a predictive model. Additionally, we removed the variables concerning personal loans ("loan") and employment variate rate ("emp.var.rate"), for they cause extreme collinearity among the predictors. Finally, we removed the variable concerning defaulting on credit ("default"), for it is extremely unbalanced (only 3 'yes' values), resulting in instability in our logistic regression and random forest models.

2.2 Exploratory Analysis

In this section, we briefly investigate some of the features of the data set. We first note the extreme imbalance in the response variable in this data set. While this does not pose a problem for logistic regression, it is potentially problematic for random forestsand SVM models. Consequently, we take care to account for this imbalance in model construction (see below for details).

Below we provide a few summary plots of interest (Figures 1 and 2). The frequency bar charts show some interesting behavior. The job plot shows that students and retired individuals are more likely than others to subscribe to a long-term deposit. The plot of marital status shows that each class of individuals is roughly equally likely to subscribe to a long-term deposit, with single and unknown individuals appearing to have a very slightly increased chance of subscribing.

In the conditional probability plots, the relationships between the response and the predictors age, the consumer price index (CPI), and the consumer confidence index (CCI) are examined. A clear pattern is shown in the plot involving age: older clients have a higher probability of signing on to a long-term deposit. No clear patterns are seen in the CCI and CPI plots, but there are pockets of increased probability of subscription, possibly indicating more complex behavior that could be exploited for predictive purposes in conjunction with other predictors.

The few plots provided here do not show a comprehensive summary of the data set, but they suggest the potential for successful predictive model construction is high, in line with the primary focus of this report. More advanced predictive methodologies are outlined below. We utilize all predictors, excluding those mentioned in section 2.1, to construct these models.

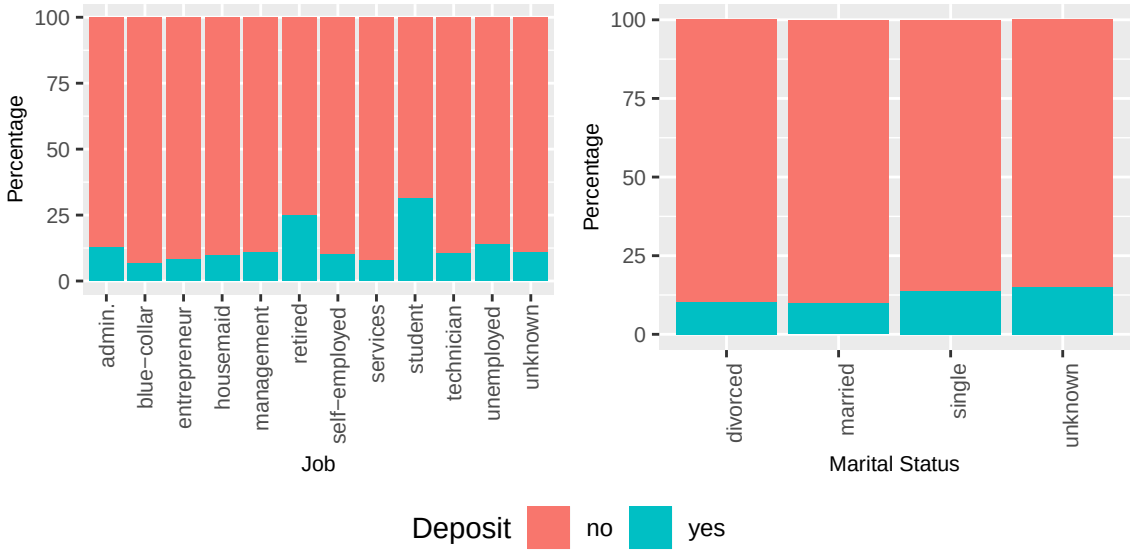

Figure 1: Categorical Stacked Bar Plots

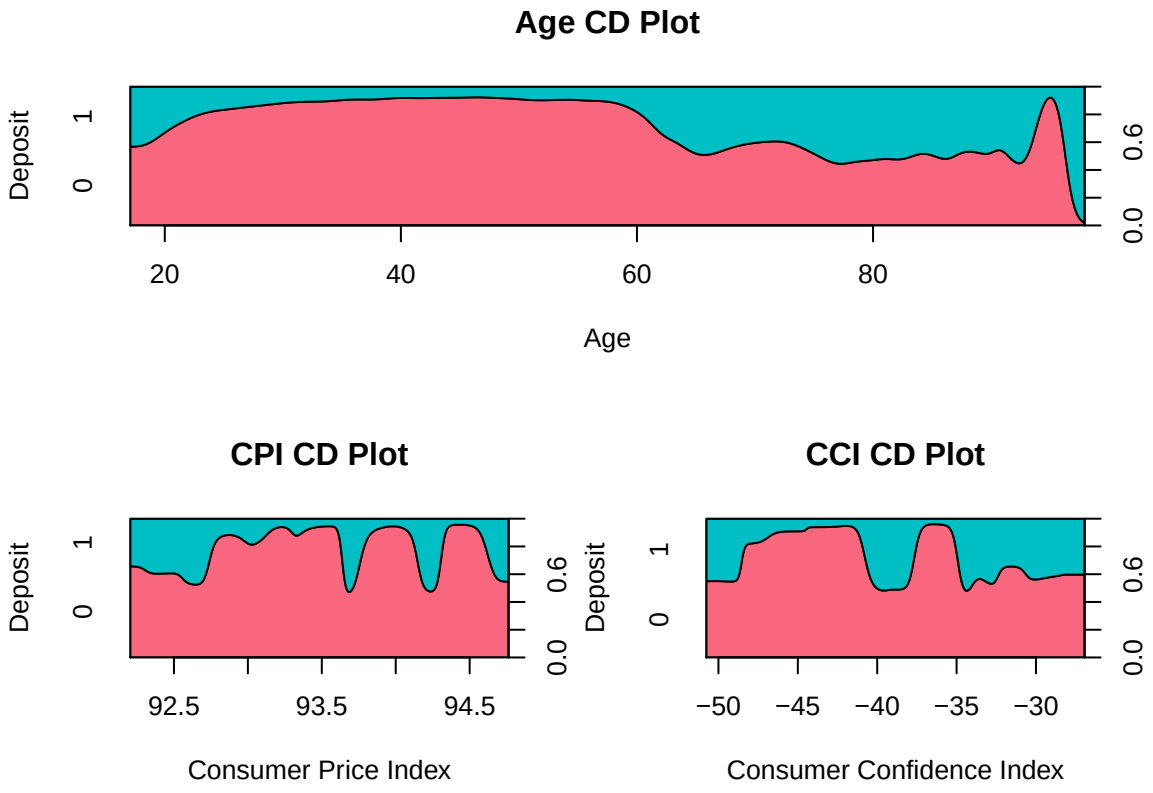

Figure 2: Conditional Density Plots

3 Binary Logistic Regression Model

3.1 General Model Form and Variable Selection

The binary logistic regression model is:

$ln(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$

Where:

Y is the binary response variable, $Y = 1$ means the client subscribed for a deposit, and $Y = 0$ means the client did not subscribe for a deposit;

p is the probability that $Y = 1$;

b_0 is the interception at y-axis;

$x_1, ...x_n$ are the predictor variables;

$b_1, ...b_n$ are the regression coefficients of $x_1, ...x_n$, respectively.

In order to improve the prediction capabilities of the model, we fit several different models until we maximized the performance, as measured by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, and the Matthews Correlation Coefficient (MCC). First, we fit a linear additive model with all 16 variables. Then, we added reasonable interaction and quadratic terms. While an exhaustive search for important second order effects was not feasible, we found that the addition of the following reasonable terms

gave a model with the best prediction performance: Quadratic terms for age ("age"), the number of days that passed before a client was last contacted ("pdays"), the consumer confidence index ("cons.conf.idx"), and the Euro three-month Interbank Offered Rate ("euribor3m"). Additionally, an interaction term between age and marital status ("age" · "marital") and an interaction term between education and the consumer price index ("education" · "cons.price.idx") was found to improve model performance.

3.2 Model Assumptions

The assumptions for a logistic regression model are:

- Assumption of Appropriate Outcome Structure - For the binary logistic regression, the type of the dependent variable (outcome) should be binary. In case of the dataset we analyze, we build a binary logistic regression model since we are interested in a predictive model for binary response variables (Subscription: Yes, No).
- Assumption of Independent Observations - Logistic regression requires all observations to be independent of each other.
- Assumption of Absence of Multicollinearity - Logistic regression requires the independent variables to be not highly correlated with each other.
- Assumption of linearity of Independent variables and Log Odds - Logistic regression requires that the independent variables are linearly related to the log odds.

3.3 Model Validation

To validate our logistic regression model's predictive capabilities, we first fit the logistic regression model to the training data set. We then use this fitted model to make predictions based on the validation data set. Below we provide summaries of the model's performance in and out of sample performance. Table 1 and 2 show confusion matrices for the model's performance in the training data set (table 1) and the validation set (table 2). Note the strikingly similar performance. Furthermore, Figure 6 in the Appendix shows ROC curves for the model's performance in the training and validation data sets. Note the remarkable similarity of the two curves. The AUC for each curve is 0.7962 for the training data set, and 0.7907 for the validation set. The MCC for the training set is 0.372, while the MCC for the validation set is similar, at 0.332. The similar performances of the fitted logistic regression model in these two data sets suggests that these measures for performance accurately characterize the predictive performance of the logistic regression model.

Table 1: Confusion Matrix For Training Set

Target	Prediction	
	No	Yes
No	17983 (87.3%)	1734 (8.4%)
Yes	289 (1.4%)	588 (2.9%)

Table 2: Confusion Matrix For Validaiton Set

Target	Prediction	
	No	Yes
No	17944 (87.1%)	1788 (8.7%)
Yes	332 (1.6%)	530 (2.6%)

3.4 Model Diagnostics

We note that our first assumption (Appropriate Outcome Structure) is trivially satisfied, for we have binary response data. Furthermore, observations are independent of one another, as clients were contacted individually of one another. We addressed potential problems of collinearity in chapter 3.1 above. Finally, to investigate the linear relationship of independent variables and Log Odds assumption, we examine a plot of the residuals against a plot of the linear predictor. If the overall model is correct, a Lowess smooth of the plot should approximate a horizonal with zero intercept. Figure 3 below shows such a plot (for the logistic

model fit on all available data), with the Lowess smooth roughly approximating a horizontal line with zero intercept. Thus, the independent variables appear roughly linearly related to the log odds. This suggests the overall appropriateness of the fitted logistic regression model.

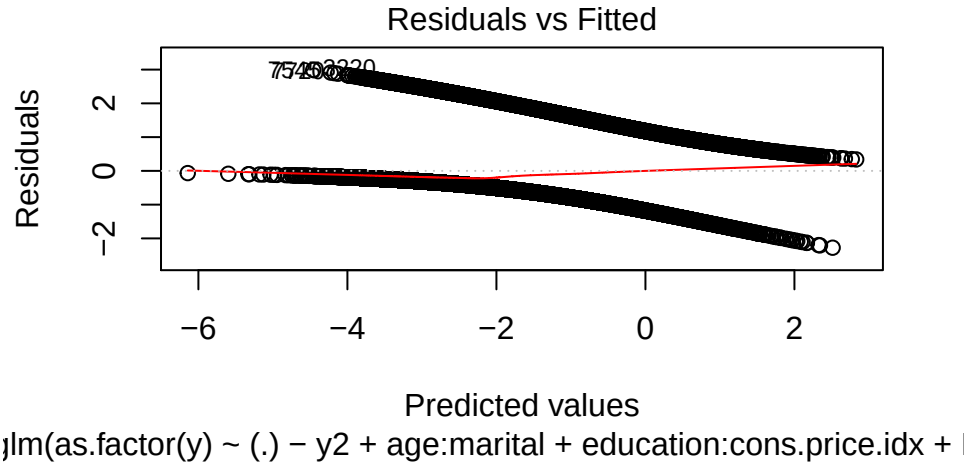

Figure 3: Residuals vs Fitted Plot For Logistic Regression Model

4 Random Forests Model

4.1 General Model Form and Variable Selection

To obtain a different predictive model, a random forests model was trained with the same training data. The random forests model is constructed by an ensemble of classification or regression decision trees. The model uses the random feature selection in the tree induction process and makes the prediction by cumulating the predictions of the branches. In general, the random forests model is fast to calculate, comparing to the other complex machine learning algorithms, and is as good as the best supervised learning algorithms [4]. At the same time, the random feature selection in the random forests model makes this model less possible to overfit the data. Although the depth of the random forests method results in the difficulty of the data interpretation, this method can give us a good model with relatively low cost. Herein, the random forests model is constructed based on the same variable options as the logistic regression model used above in order to compare their performance.

4.2 Model Assumptions

The random forests method usually requires a balanced dataset, because an unbalanced dataset makes this method biased in the same direction as the majority class. Since the response variable in our dataset is unbalanced, the class weight is used to adjust for this. Herein, the 'NO' and 'YES' classes are weighted proportional to how frequently the other class appears in the data set.

4.3 Model Validation

Table 3: Confusion Matrix For Training Set

Target	Prediction	
	No	Yes
No	17937 (87.1%)	0 (0%)
Yes	335 (1.6%)	2322 (11.3%)

Table 4: Confusion Matrix For Validaiton Set

Target	Prediction	
	No	Yes
No	17325 (84.1%)	1561 (7.6%)
Yes	951 (4.6%)	757 (3.7%)

The confusion matrix of the random forests model is seen in Table 3 and 4. The random forests model has a good performance on the training dataset with the AUC value as high as 0.998. The AUC value of the validation set is about 0.74, noticeably smaller than that of the training set (see Figure 7 in the Appendix). The MCC values for the training set and validation set are 0.92 and 0.31 respectively. This difference between the training set in the random forests Model may be caused by the unbalance of the dataset even though the re-weighted parameters have been considered when setup the model. It is also possible that the random forests model is simply overfitting the training data.

According to the importance variable analysis (see Figure 9 in the Appendix), the way to contact the clients ('contact' variable) plays an important role (11%) in the random forests Model. Based on the whole dataset, the people contacted by cellular have more chance to subscribe a term deposit than those contacted by telephone. Besides, the 'euribor3m' (11%), describing the Euribor 3-month rate, and the 'age' of the clients (13%) are another two important variables in the dataset. It is reasonable that the investment behavior is associated with the loan interest rate as well as the age of the clients. The low loan interest rate (or even negative) will encourage the clients to do the investment, and at the time, older people probably tend to have more money to purchase the financial product. There are other important variables, such as the number of contacts to the client ('campaign', 8%), and the 'education' (10%).

5 Support Vector Machine (SVM) Model

5.1 General Model Form and Variable Selection

In order to compare the prediction power of other models, another possible model, the Support Vector Machine (SVM) model, was built using the training dataset and its predictive power was tested using the validation dataset. The model consists of the same selection of variables as the logistic regression and the random forest model. The SVM is a classification method that creates a line that separates data points into classes. For the dataset we analyzed, the use of SVM is to predict the numbers of long-term deposit subscription (YES) by having a line that separates data points "Yes" and "No" into 2 classes.

5.2 Model Assumptions

The SVM is not robust to the imbalanced dataset, because an imbalance in numbers of two classes will lead to a bias in classification of two classes. Since the response variables are completely imbalanced with outnumbered 'NO's compared to 'YES's, the 'NO' and 'YES' classes are weighted proportional to how frequently the other class appears in the data set.

5.3 Model Validation

The confusion matrices for the SVM's training and validation performance is seen in Table 5 and 4. The AUC values for training and validation data are 0.8380 and 0.7644 respectively, with ROC curves shown in Figure 8 in the appendix. The slight difference could indicate potential overfitting; however, it is likely not too extreme due to the agreement seen between confusion matrices. The MCC values for the training

set and validation set are 0.1122 and 0.1126 respectively, further suggesting there is no extreme overfitting. The similar performances in training and validation data sets suggest that the metrics discussed provide an accurate measure of the predictive performance for the SVM Model. We thus proceed with model comparisons and discussion in the next section.

Table 5: Confusion Matrix For Training Set

Target	Prediction	
	No	Yes
No	18216 (88.5%)	2269 (11%)
Yes	56 (0.3%)	53 (0.3%)

Table 6: Confusion Matrix For Validaiton Set

Target	Prediction	
	No	Yes
No	18216 (88.5%)	2264 (11%)
Yes	60 (0.3%)	54 (0.3%)

6 Model Comparison

6.1 Single Model Performance

While the logistic regression and SVM models performed similarly in and out of sample, random forests appeared to suffer from overfitting. Thus, to make meaningful comparisons between models, we will only compare the performances of each model in the validation set.

The logistic regression model out performed both random forests and SVM in measures of overall accuracy (MCC and AUC). Thus, for this performance metric, it appears to be the superior model of those presented here. However, the random forests model had a much lower type II error rate compared to both SVM and logistic regression models, and correctly identified more clients who were likely to sign up for a long-term loan. Given the differences in risk between the two error rates (see section 1.2), type II error rate might be a better measure of a model's performance. Thus, random forests would be the superior model of those presented here. While the SVM model did not perform as well as the other two models in either of these metrics, we note briefly that it did have a much lower type I error rate than the either of the other two models (which could be important for some applications).

The differences in performance between these two models are likely due to the difference in model structure: each model makes different assumptions about the form of the relationship between the variables used to make a prediction, or uses the variables in a different way. For example, the logistic regression model assumes that the log odds of the dependent variable is linearly related to the independent variables, while the random forests and SVM do not make this assumption.

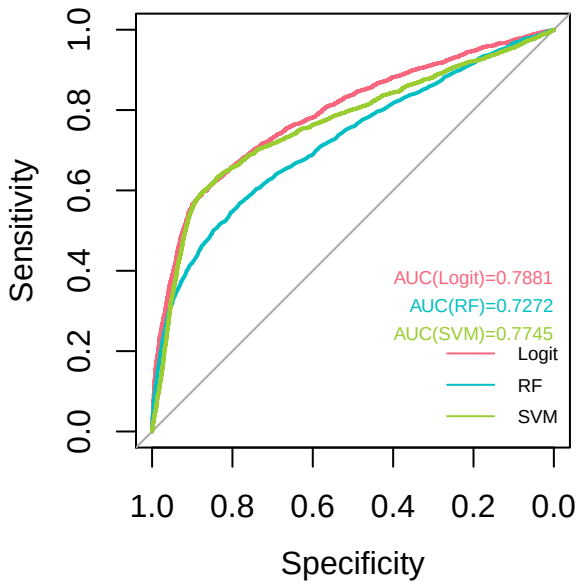

Figure 4: ROC Plot for Logistic Regression Model and Random Forest Model

Table 7: False Negative Rate

Logit	RF	SVM
1788 (8.7%)	1561 (7.6%)	2264 (11%)

6.2 Composite Model Performance

In this section we consider the possibility of building a composite classification model. We do this by using a convex combination of the predicted probabilities from a combination of models to form a new predicted probability of a client subscribing to a long-term loan. The combination of forecasts has been utilized before in forecasting time series [5]. As mentioned above, each model makes use of predictor variables in different ways, and thus provides the possibility of combining forecasts to produce better forecasts than could be obtained using each model individually. We consider two examples of a composite model here, one using the results of the logistic regression and random forests models and one using the results of all three models. We note that we have not employed any rigorous optimization routine, for our aim in considering composite models is to show the greater flexibility that they provide a modeler, and not to maximize some specified criterion.

Our first model is obtained weighting the predictions of the logistic regression model and random forest model by 0.2 and 0.8 respectively. This model's performance is a mix of the logistic regression and random forests models. A confusion matrix for this model can be seen in table 8. The AUC for this model is 0.7438 as seen in Figure 5, indicating that it has a slightly higher overall accuracy than the logistic regression model. Additionally, it has a much lower type II error rate than the logistic regression model (similar to the random forests model).

A slightly different composite model, with weights of 0.1, 0.8, and 0.1 for the logistic regression, random forests, and SVM model respectively shows only subtle improvements: an AUC of 0.7453 (Figure 5) and a confusion matrix (table 9) revealing a slightly lower type II error rate. While the SVM did not have a promising performance on its own, and its contribution to this composite model is very small, for some

applications even subtle improvements may be important and beneficial. It is then noteworthy that such improvements can be made by incorporating a model whose individual performance was not very promising.

Table 8: Confusion Matrix For Logit and RF

Target	Prediction	
	No	Yes
No	17633 (85.6%)	1632 (7.9%)
Yes	643 (3.1%)	686 (3.3%)

Table 9: Confusion Matrix For Logit, RG and SVM

Target	Prediction	
	No	Yes
No	17617 (85.5%)	1626 (7.9%)
Yes	659 (3.2%)	692 (3.4%)

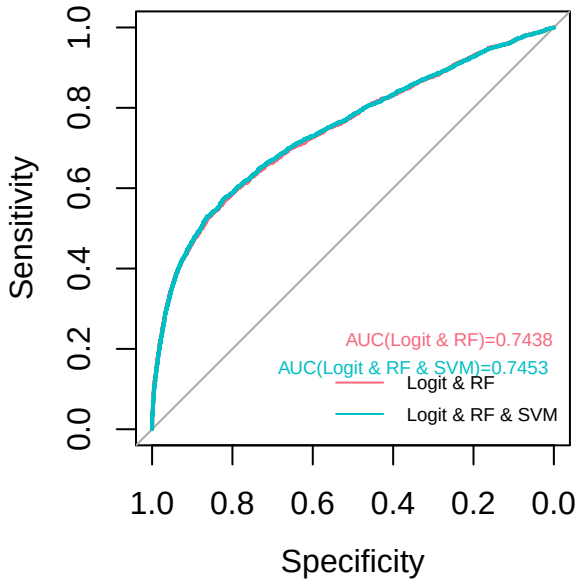

Figure 5: ROC Plot for Composite Models

We note that as there are many ways the quality of a predictive model may be assessed, choosing the weights for a general composite model is not a rigorous optimization problem. The overall measure of a model's performance is up to the modeler and their goals. However, once a specified performance metric(s) has been selected, a rigorous optimization problem can be formulated. The weights of each individual model in a composite model determine how much of that model's characteristics are seen in the final model's performance. For example, in our first composite model, the largest improvement is seen in Type II error, for greater weight was given to the random forests model, while only small improvements were seen in overall accuracy, for less weight was given to the logistic regression model. Thus, once a modeler has a clear view of what is required, they can work on finding an optimal composite model using this knowledge as a guiding principle.

Finally, we note that, as a composite model pools the performance of the individual models that make it up, an effective way of constructing them would be to construct individual models whose performance are optimized for one particular performance metric, and then combining them to form a final model. Some models may be constructed for overall accuracy, while others specialize in reducing a given error type, while more models may be constructed specifically to accurately classify/predict specific subpopulations. A final composite model may then be constructed that performs incredibly well by combining the best features of these specifically designed models.

7 Conclusions, Limitations, and Future Research

In conclusion, while the logistic regression model had better overall accuracy performance, the random forests model has lower type II error, thus it seems likely that the random forests model may be the preferred model for this application. However, in many statistical modeling applications, the true measure of a model's performance is mediated by the clients and problems at hand. We thus experimented with extending the flexibility of the models considered by considering combining their predictions using composite models. These composite models offer great flexibility, and allow a modeler to more effectively address answers to problems across various fields, and more accurately respond to the needs of clients. The treatment of such composite models here in the classification setting is far from complete, and we urge further research into this topic.

8 References

1. Center for Machine Learning and Intelligent Systems: UCI Machine Learning Repository, Bank Marketing Data Set. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

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

3. Kutner, M.H. 2005. Applied Linear Statistical Models, 5th.

4. Yiu, T. 2019. Understand Random Forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

5. BATES, J.M, and GRANGER C.W.J, 1969. The Combination of Forecasts. OR, Vol. 20, No. 4 (pp. 451-468).

9 Appendix

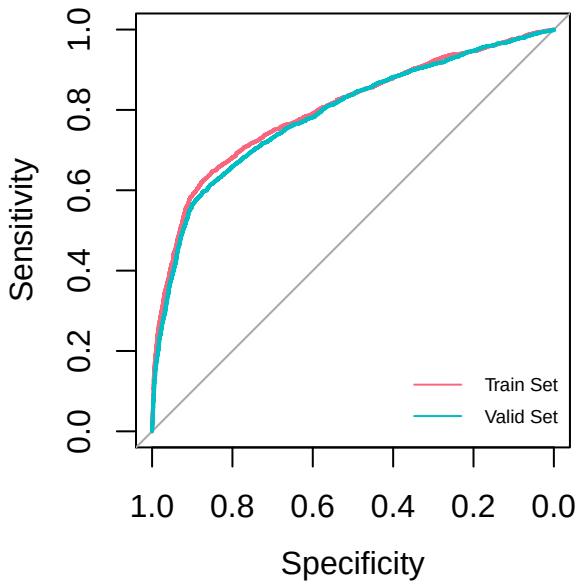

Figure 6: ROC Plot For Training and Validation Datasets of Logistic Regression Model

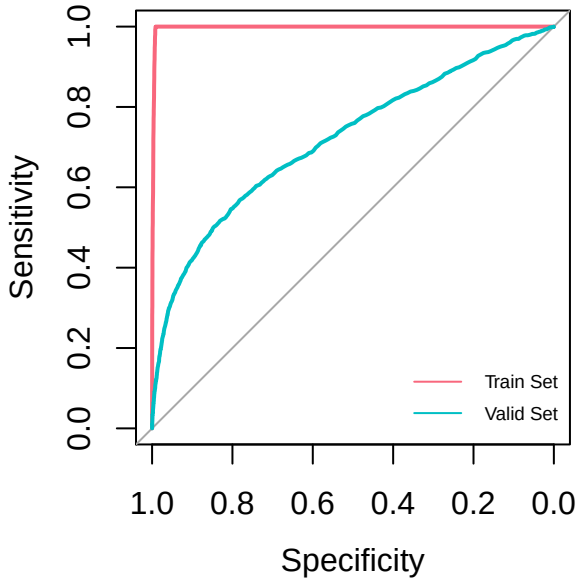

Figure 7: ROC Plot For Training and Validation Datasets of Random Forest Model

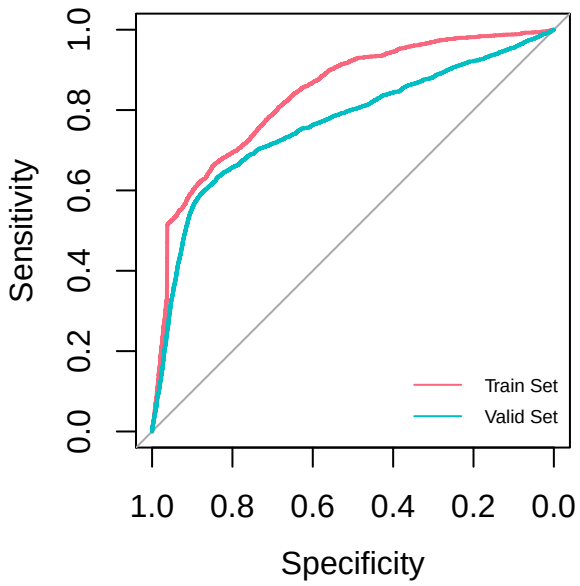

Figure 8: ROC Plot For Training and Validation Datasets of SVM Model

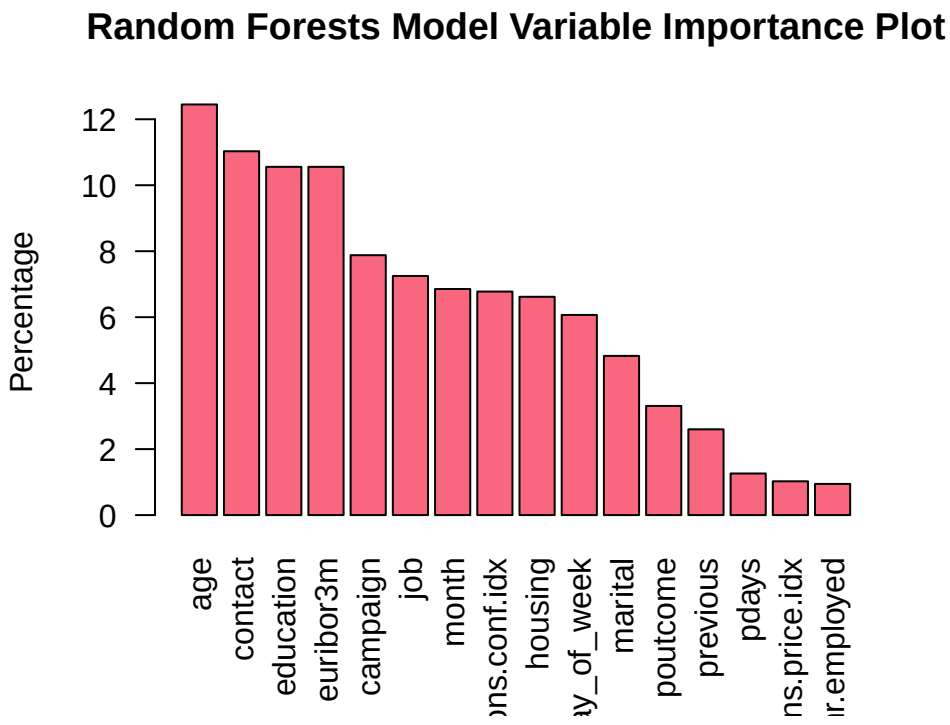

Figure 9: Plot of Random Forests Model Variable Importance