6372: Project 2

Simerpreet Reddy, Rinku Lichti, Megan Ball

# Introduction

The goal of this analysis was to develop an optimal classification model to predict the success rates of a Portuguese bank telemarketing campaign. As part of this, a total of four models were developed for use in comparison for best overall combined accuracy, sensitivity, and specificity:

* Simple, interpretable linear regression model
* Complex linear regression model
* Linear discriminant analysis (LDA) model
* Random Forest model

The subsequent discussion will review the analysis steps, model metrics, and ultimate best model for classification.

# Data Description

The dataset used in our analysis initially consisted of 41,188 observations with a total of 21 different variables. The data was collected as part to determine success of getting customers to open accounts as part of a telemarketing campaign.

# Exploratory Data Analysis

Walkthrough of key findings in EDA

* Use graphs relevant to our interactions
* Use graphs relevant to simple model chosen vars

# 

# Objective 1

## Restatement of Problem

Our first goal was to develop a simple and interpretable model to predict and explain the relationship between our variables and the probability of customers opening an account. To do this, we chose logistic regression as our desired model type and performed variable selection and cross validation to select the best interpretable model.

Modeling Strategy

After doing EDA and creating a few derived variables, we started with creating a simple model using all the variables and then step by step removing statistically insignificant variables as well as variables with high multicollinearity based on both VIFs and scatter plots. After every column removal, we re-ran the model and repeated the exercise till we found the model with statistically significant as well moderate (generally with values less than 10) VIF variables.

We then ran both stepwise selection and LASSO as competing feature selection methods, and further limiting our variables based on VIF. Using ROC curves to find a good cut off range, we predicted the response variable with multiple cut off values in the range 0.1 and 0.5 to determine optimum value. We then compared the accuracy, sensitivity and specificity of all the different models – simple, step, and lasso – on various cut offs. Since the primary goal is to predict when a customer will subscribe to a term deposit (positive response variable = ‘yes’) we chose a cut-off value based on balancing our overall accuracy and the sensitivity.

Model

For the simple interpretable model, the following model gave us the best combined accuracy, sensitivity, and specificity at a cut off probability value of 0.15.

## Assumptions

* Data points are independent.

Lack of fit test:

* We performed a global test for the coefficients (beta) being equal to zero with Likelihood ratio, Score and Wald test all agreeing at p-value <0.001 that we reject the null hypothesis and conclude that the logistic regression model is valid.
* Cross validation was done in the form of a test/train split which demonstrated good metrics on our test data set.

|  |  |
| --- | --- |
| Metric | Result |
| Accuracy | 86.6% |
| Sensitivity | 83.7% |
| Specificity | 87.2% |

Influential points analysis

There were several points that stood out when we looked at Leverage, Cook’s D, Standard residuals, and deviance plots. We removed those and reran the code, even though the plots looked better, there were no significant changes in the model or the variable coefficients. So, we decided to keep those data points in the training data set.

## Model Interpretation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Odds Ratio | 95% CI: Lower Limit | 95% CI: Upper Limit | P-value |
| Intercept | 1.486 x 10-21 | 5.619 x 10-26 | 3.934 x 10-17 | < 0.001 |
| Job: blue-collar | 0.707 | 0.610 | 0.820 | < 0.001 |
| Job: entrepreneur | 0.871 | 0.600 | 1.151 | 0.332 |
| Job: housemaid | 0.818 | 0.589 | 1.136 | 0.230 |
| Job: management | 0.950 | 0.784 | 1.150 | 0.600 |
| Job: retired | 1.085 | 0.843 | 1.397 | 0.528 |
| Job: self-employed | 0.893 | 0.690 | 1.155 | 0.388 |
| Job: services | 0.741 | 0.615 | 0.892 | 0.002 |
| Job: student | 1.192 | 0.917 | 1.550 | 0.190 |
| Job: technician | 0.994 | 0.863 | 1.146 | 0.938 |
| Job: unemployed | 0.884 | 0.655 | 1.195 | 0.423 |
| Default: unknown | 0.722 | 0.620 | 0.841 | < 0.001 |
| Contact: telephone | 0.691 | 0.588 | 0.811 | < 0.001 |
| Month: August | 1.264 | 1.000 | 1.600 | 0.050 |
| Month: December | 1.052 | 0.643 | 1.720 | 0.841 |
| Month: July | 1.396 | 1.123 | 1.736 | 0.003 |
| Month: June | 1.485 | 1.202 | 1.835 | < 0.001 |
| Month: March | 5.741 | 4.332 | 7.610 | < 0.001 |
| Month: May | 5.304 | 0.446 | 0.631 | < 0.001 |
| Month: November | 1.072 | 0.854 | 1.346 | 0.545 |
| Month: October | 1.514 | 1.135 | 2.021 | 0.005 |
| Month: September | 0.992 | 0.725 | 1.359 | 0.961 |
| Duration | 1.002 | 1.002 | 1.003 | < 0.001 |
| Campaign | 0.956 | 0.931 | 0.981 | < 0.001 |
| Previous number of contacts | 0.762 | 0.694 | 0.836 | < 0.001 |
| Cons\_price\_idx | 1.692 | 1.513 | 1.893 | < 0.001 |
| Cons\_conf\_idx | 1.060 | 1.047 | 1.072 | < 0.001 |
| Euribor3m | 0.491 | 0.470 | 0.513 | < 0.001 |
| Age Group (32-37) | 0.829 | 0.727 | 0.946 | 0.005 |
| Age Group (38-47) | 0.743 | 0.646 | 0.855 | < 0.001 |
| Age Group (47 – 55) | 0.859 | 0.734 | 1.004 | 0.057 |
| Age Group (55+) | 0.930 | 0.753 | 1.149 | 0.502 |
| Previously contacted: Yes | 6.190 | 5.048 | 7.590 | < 0.001 |
| Duration Group: 5-10min | 2.983 | 2.600 | 3.424 | < 0.001 |
| Duration Group: 10+ min | 7.200 | 5.563 | 9.319 | < 0.001 |

Education:

Holding all other variables constant, the odds of a client (saying yes to a term deposit) with a basic.4y education are estimated to be 0.738 times lower than those of a client with a university.degree. The estimated 95% confidence interval of these odds is between 0.616 and 0.884.

# Objective 2

For competing models, we chose Complex Model, LDA/QDA, and Random Forest.

#### Complex Model

A complex logistic regression model was built off of the variables deemed most significant from our simple regression model. Based on EDA, as mentioned above, the factors that appeared to have some sort of interaction included default with duration, contact with duration, default with month, and month with euribor3m.

#### Linear Discriminant Analysis

In addition to a complex logistic model, a both linear and quadratic discriminant analysis were run on the numeric variables for comparison.

#### Random Forest

To tune Random Forest, I chose mtry and the split function. Mtry was an obvious choice as the default Random Forest hyperparameter, which tells Random Forest how many variables to consider when performing each split. Each time Random Forest splits, it choose mtry variables at random, then calculates the optimal split considering a split function to select a variable for the split. For the split function, I chose giri, extratrees, and hellinger, all of which are suitable for classification, but I weren’t sure which one would perform best given our data.

Optimizing for ROC, the winning parameters are an mtry of 5 predictors considered at each split, and the Hellinger split rule. It's interesting that Hellinger won. I found some papers suggesting Hellinger handles imbalanced data well because it is insensitive to skew.

Reference: <https://www3.nd.edu/~nchawla/papers/DMKD11.pdf>

Reference: <https://medium.com/@evgeni.dubov/classifying-imbalanced-data-using-hellinger-distance-f6a4330d6f9a>

To better understand how this Random Forest model is working, let’s look at the minimal depth for each predictor. The more important variables will be the ones used to split higher up the tree.

Chart, bar chart

Description automatically generated

Figure

From this I can see the most influential variable in the model by far is duration, and a derivative of duration, duration\_group10+ min is 3rd. I worry that duration of last call may be a strong “predictor” only because people who convert will of course have long duration of last call because they already decided to move forward and their last call probably is long because they are finalizing the deal. If so, then this model may be not so useful for predicting those who are likely to convert in the future. Predicting on duration feels like predicting rain based mostly on feeling raindrops starting to fall on my head.

Maybe duration should be replaced with the duration of the 2nd to last call? Or perhaps average duration of all calls? It may be worth following up with the company to refine the available variables.

Another way to interpret the model is to look at the influence of interactions:

Chart, bar chart, waterfall chart

Description automatically generated

Figure

Consumer Confidence most frequently interacted with other variables, like age, duration, and campaign, and in fact the min depth interaction is cons\_conf\_idx:duration.

## Metrics & Model Comparison

The best model performance was for the random forest model with a cut-off of 0.135 for probability.

Table

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity |
| Simple |  |  |  |
| Complex |  |  |  |
| LDA |  |  |  |
| Random Forest | 0.86 | 0.92 | 0.85 |

Chart

Description automatically generated

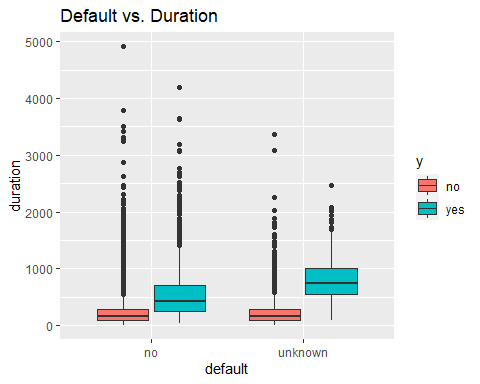
Figure . ROC curves for the top four models.

# Conclusion & Final Recommendations

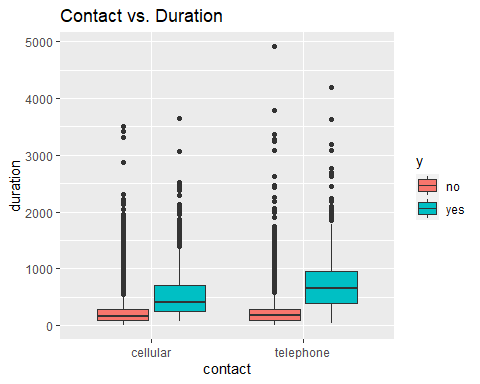
While these models performed well given the data, especially Random Forest, we recommend revisiting the problem with duration. It appears to have the most influence yet logically seems like any model based on it would only be useful to “predict” conversion after people have already converted. We would encourage rethinking about which data to include in the training set based on that data’s availability about people at the time accurate predictions would be most valuable to the business.

|  |
| --- |
| **Exploring an alternative model without duration…**  We didn’t have time to fully explore this, but as an experiment, I built a Random Forest model that excluded duration and duration\_group entirely, and here are the variables with min depth:  Chart, bar chart  Description automatically generated  If the purpose of the model is to predict which people to contact in the first place based on who is more likely to convert, then this model may be more effective since it seems to be driven mainly by predictors known in advance of any contact.  As for interactions:  Chart, bar chart  Description automatically generated  Consumer confidence index is again the most influential interaction, this time with campaign.  More work would need to be done to clear away other variables that would be unknown for people who have not yet been contacted, or perhaps were only contacted in the distant past, or for other campaigns. Then we would need to redo this analysis, but unfortunately we are out of time. Still, I wanted to call out this finding and possible next steps. |

# Appendix



Figure



Figure

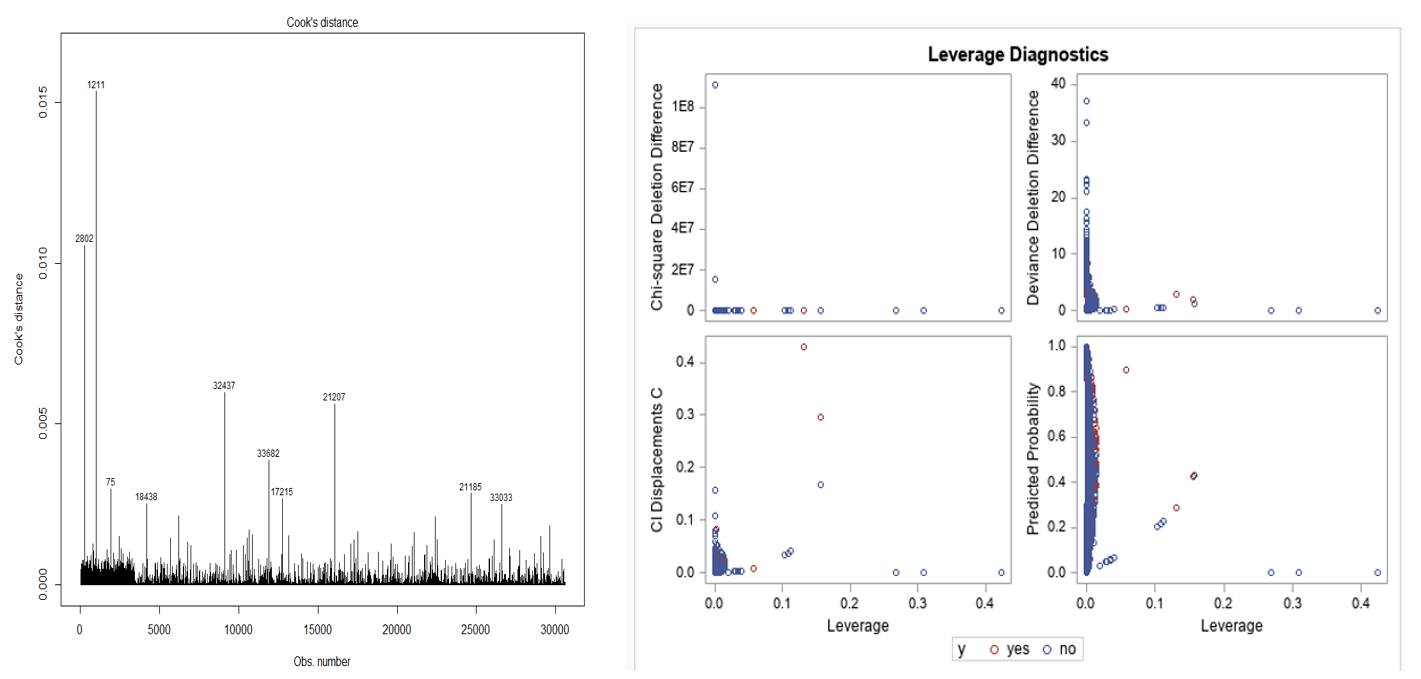


Figure 8 Cook’s D and Leverage plot from the model with full training data set:

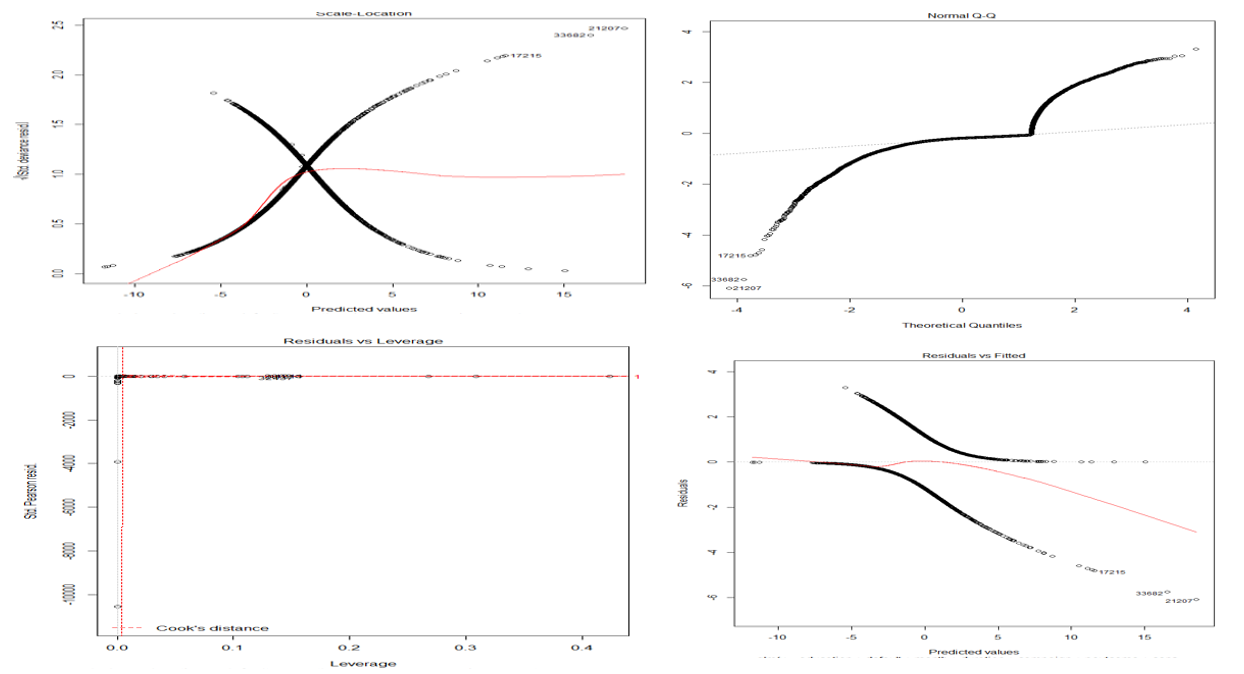


Figure 9 Residual Plots from the model with full training data set.

Chart, line chart

Description automatically generated

Figure 4

Link to Rmarkdown knit file (full notes/comments with graphics): <https://github.com/megball/6372-project-2/blob/main/Rmarkdown/Project2.html>

## R Code