6372: Project 2

Simerpreet Reddy, Rinku Lichti, Megan Ball

# Introduction

Summary

# Data Description

Short description of our data

# Exploratory Data Analysis

Walkthrough of key findings in EDA

PCA discussion

Graphs

# Objective 1:

## Restatement of Problem

Restatement of Problem and the overall approach to solve

Model Selection

**Modeling strategy**:

1. Upon doing EDA and creating a few derived variables, we started with creating a simple logistic regression model using all the variables and then step by step removing statistically insignificant variables as well as variable with high multicollinearity based on both VIFs and scatter plots. We kept poutcome and prevly\_cntctd in the simple model because even though these had comparatively higher VIFs, these are practically independent columns. After every column removal, we reran the model and repeated the exercise till we found the model with statistically significant as well moderate VIFs variables.
2. We then ran feature selection methods– Step and LASSO, doing the same exercise as above.
3. Using ROC curves to find a good cut off range, we predicted the response variable with few cut off values in the range 0.1 and 0.5. We then compared the Accuracy, Sensitivity and Specificity of all the different models – simple, step, lasso on various cut offs. Considering the dataset and the goal, we chose better specificity over better accuracy and sensitivity.

**Model**

For simple interpretable model, the following model gave us the best confusion matrix statistics at cut off=0.15.

Variables considered significant for the model to predict the odds of a client subscribing to a term deposit as part of the marketing campaign - education,default,month,duration,campaign,poutcome,cons\_price\_idx,euribor3m and Age\_Grp.

Log(odds of a client subscribing to a term deposit) = -31.2267 + (-0.3241)\*education\_ basic.4y + (-0.1680)\*education\_ basic.6y + (-0.3459)\*education\_ basic.9y + (-0.2058)\*education\_ high.school + (1.1845)\*education\_ illiterate + (1)education\_university.degree

+ (2.4913)\*default\_no + (2.1216)\* default\_unknown + (1)\* default\_yes

+ (1.1850)\* month\_mar (-0.6265)\* month\_apr + (-1.1457)\* month\_may + (-0.00265)\* month\_jun + (0.1381)\*month\_jul + (0.2490)\* month\_aug + (0.3729)\* month\_oct + (-0.2051)\* month\_nov + (0.0619)\* month\_dec + (1)\* month\_sep

+ (-0.7414)\* poutcome\_failure + (-0.3325)\* poutcome\_nonexistent + (1)\* poutcome\_success

+ (0.1205)\* Age\_Grp\_17\_31 + (-0.1167)\* Age\_Grp\_32\_37 + (-0.1933)\* Age\_Grp\_38\_47 + (-0.0187)\* Age\_Grp\_47\_55 + (1)\* Age\_Grp\_>55 + (0.00473)\* duration + (-0.0534)\* Campaign + (0.3002)\* cons\_price\_idx + (-0.7190)\* euribor3m

## Checking Assumptions

Lack of fit test:

* We performed Hosmer-Lemeshow Goodness-of-Fit Test to test if the data fit the logistic regression model. With chi-sq=333.62 at 8 degrees of freedom and p-value <0.0001. TEST FAILED, probably because of lasrge data set.
* We performed Global test for beta=0, with Likelihood ratio, Score and Wald test all agreeing at p-value<0.001, we reject the null hypothesis and conclude to say that the model is valid.
* Confusion matrix for accuracy, sensitivity and specificity:

**Influential point analysis** (Cook’s D and Leverage):

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 mode, 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. Please see Appendix sectionXYZ for details on outliers, model output and plots after removing the outliers from the training data set.

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

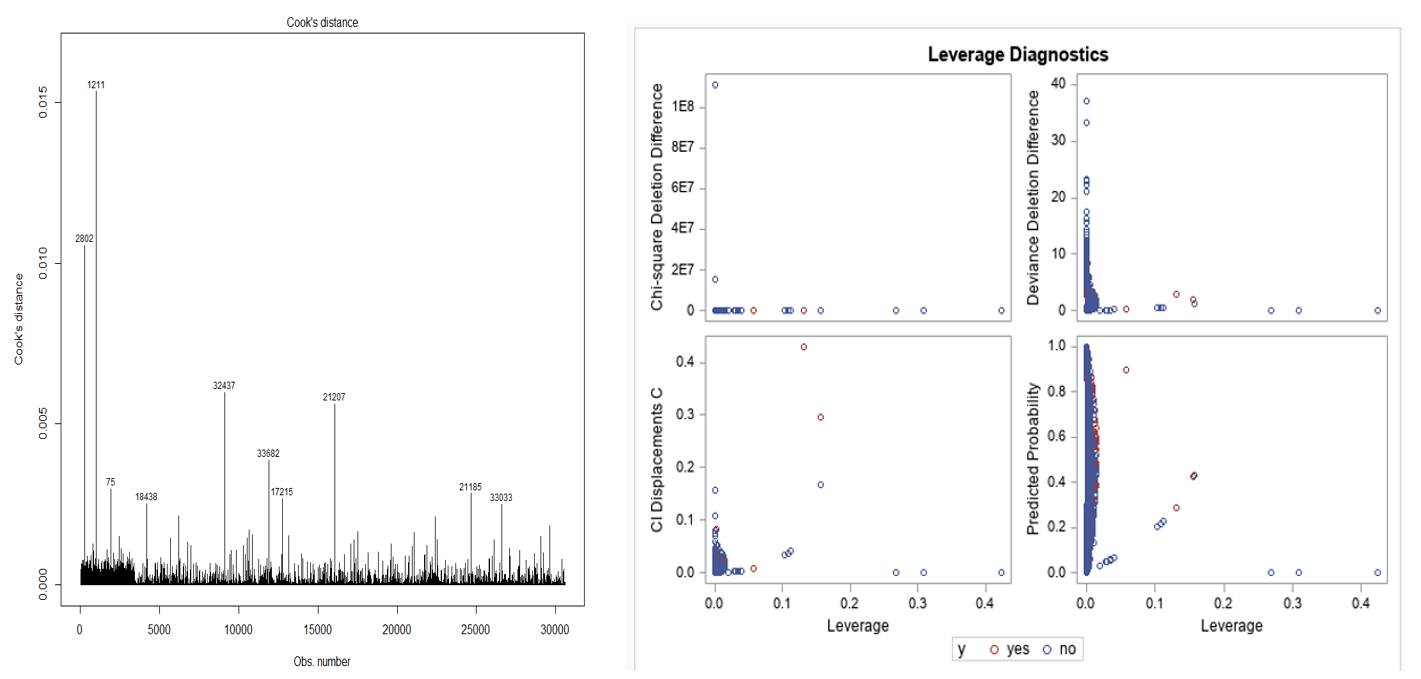
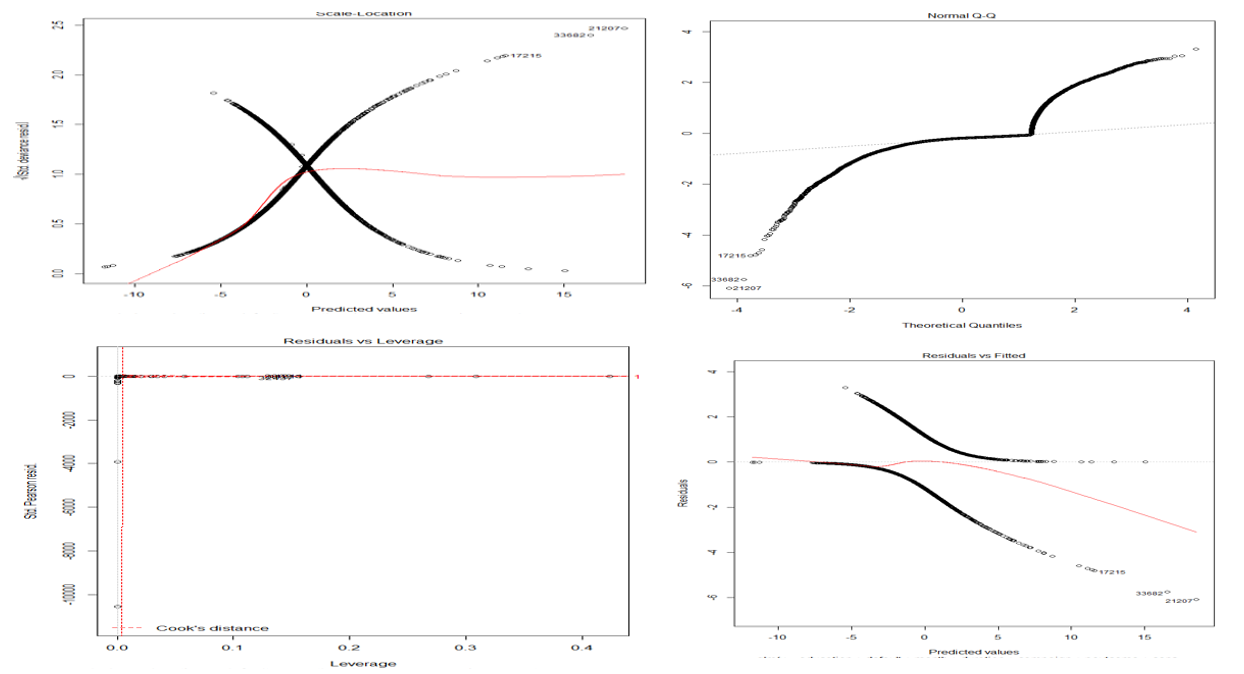


Fig: Residual Plots from the model with full training data set.



**Model Interpretation**:

Education:

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

Holding all other variables constant, the odds of a client (saying yes to a term deposit) with a basic.6y education are 0.863 times lower than those of a client with a university.degree. The 95% confidence interval of these odds is between 0.687 and 1.083.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) with a basic.9y education are 0.722 times lower than those of a client with a university.degree. The 95% confidence interval of these odds is between 0.619 and 0.843.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) with a basic. high.school education are 0.831 times lower than those of a client with a university.degree. The 95% confidence interval of these odds is between 0.736 and 0.938.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who is an illiterate are 3.337 times higher than those of a client with a university.degree. The 95% confidence interval of these odds is between 0.700 and 15.903.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) with a professional.courseeducation are 0.905 times lower than those of a client with a university.degree. The 95% confidence interval of these odds is between 0.782 and 1.048.

Default:

Default no

Default unknown

Month:

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Mar are 3.360 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 2.425 and 4.654.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Apr are 0.549 times lower than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.418 and 0.721.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of May are 0.327 times lower than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.252 and 0.424.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Jun are 1.024 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.777 and 1.350.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Jul are 1.179 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.890 and 1.563.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Aug are 1.318 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.997 and 1.742.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Oct are 1.491 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 1.091 and 2.040.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Nov are 0.837 times lower than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.626 and 1.118.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) who was last contacted in the month of Dec are 1.093 times higher than those of a client last contacted in September. The 95% confidence interval of these odds is between 0.668 and 1.789.

Poutcome:

Holding all other variables constant, the odds of a client (saying yes to a term deposit) whose last campaign outcome was a failure are 0.163 times lower than those whose last campaign outcome was a success. The 95% confidence interval of these odds is between 0.134 and 0.197.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) whose last campaign outcome is nonexistent are 0.245 times lower than those whose last campaign outcome was a success. The 95% confidence interval of these odds is between 0.207 and 0.290.

Age\_Grp:

Holding all other variables constant, the odds of a client (saying yes to a term deposit) in the age group 17-31 are 0.916 times lower than those who are >55yrs in age. The 95% confidence interval of these odds is between 0.775 and 1.082.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) in the age group 32-37are 0.723 times lower than those who are >55yrs in age. The 95% confidence interval of these odds is between 0.610 and 0.856.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) in the age group 38-47 are 0.669 times lower than those who are >55yrs in age. The 95% confidence interval of these odds is between 0.562 and 0.796.

Holding all other variables constant, the odds of a client (saying yes to a term deposit) in the age group 47-55 are 0.797 times lower than those who are >55yrs in age. The 95% confidence interval of these odds is between 0.664 and 0.957.

Duration: The odds ratio is 1.005 which means that for every single unit increase in the call duration, the odds a client subscribing to a term deposit increase by 0.5%. The 95% confident interval for the odds is between 1.005 and 1.005.

Campaign: The odds ratio is 0.948 which means that for every single unit increase in the number of contacts, the odds a client subscribing to a term deposit decreases by 0.52%. The 95% confident interval for the odds is between 0.923 and 0.973.

cons\_price\_idx: The odds ratio is 1.350 which means that for every single unit increase in the cons\_price\_idx, the odds a client subscribing to a term deposit increase by 35%. The 95% confident interval for the odds is between 1.229 and 1.483.

euribor3m: The odds ratio is 0.487 which means that for every single unit increase in euribor3m, the odds a client subscribing to a term deposit decreases by 0.513%. The 95% confident interval for the odds is between 0.468 and 0.507.

|  |  |
| --- | --- |
|  |  |

**Any or all**: LASSO, RIDGE, ELASTIC NET,

Stepwise, Forward, Backward

Manual / Intuition

## Checking Assumptions

Lack of fit test

Influential point analysis (Cook’s D and Leverage)

**Optional** Residual Plots

## Model Interpretation

Interpretation **Required**

Confidence Intervals **Required**

# Objective 2:

Make sure it is clear how many models were created to compete against the one in Objective 1. Make note of any tuning parameters that were used and how you came up with them (knn and random forest logistics) **Required**

* Brief summary of analysis and tuning parameters for complex model, LDA/QDA, and random forest

## Metrics & Model Comparison

Table of accuracy, sensitivity, specificity for each model

ROC curve for all our models

# Conclusion & Final Recommendations

# Appendix

Figures referred to in report

R markdown/code