SS 3860 Final Report

by

DARIO MELCONIAN

SOHAM GH

SUN CHUWEN

LIU ZEZHEN

Contents

[Introduction 2](#_Toc69276954)

[Dataset Source and Description 2](#_Toc69276955)

[Methods Used 3](#_Toc69276956)

[Results 4](#_Toc69276957)

[Exploratory data analysis 4](#_Toc69276958)

[Main data analysis 5](#_Toc69276959)

[Significant Coefficients and Interpretations 7](#_Toc69276960)

[Model Performance 7](#_Toc69276961)

[Goodness of fit and model diagnostics 8](#_Toc69276962)

[Conclusion 9](#_Toc69276963)

[Contributions 10](#_Toc69276964)

[References 10](#_Toc69276965)

[Appendix 11](#_Toc69276966)

[Section 1: setup 11](#_Toc69276967)

[Section 2: Data cleaning and setting categorical variables as factors 11](#_Toc69276968)

[Section 3: descriptive statistics 12](#_Toc69276969)

[Section 4: plotting variables of interest 13](#_Toc69276970)

[Section 5: Violin plots to understand the distributions. 13](#_Toc69276971)

[Section 6: null values 14](#_Toc69276972)

[Section 7: First round of variable selection and initial model fit 14](#_Toc69276973)

[Section 8: model Selection and performance 15](#_Toc69276974)

[Section 9: interpreting the coefficients. 16](#_Toc69276975)

[Section 10 : Model diagnostic plots and Hosmer-Lemeshow statistic 16](#_Toc69276976)

# Introduction

In today’s digitally fueled world, the concept of marketing has become the heart of a business corporation.  Furthermore, in marketing, it is crucial that the obtaining of consumer information through promotion and telemarketing is given a great deal of importance.  In the case of bank institutions, marketing often consists of drawing clients to specific investment opportunities or offering long-term deposits contracts.

The purpose of the analysis of this data is to discover the factors that positively and negatively affect the probability of any given client subscribing to the ‘Bank Term Deposit’ offered by the bank institution. Furthermore, the goal of this analysis from a business perspective is to understand what factors impact the decision of a customer, i.e., to subscribe to the bank institution’s term deposit or not.  By understanding what factors impact this decision a bank can fine-tune its marketing effort to target the demographics that correlate with higher probabilities of subscription. This saves time and money for a business, which is a recipe for the success of any business.

# Dataset Source and Description

The dataset being investigated in this report was taken from the UCI Machine Learning Repository. It has about 40,000 rows and involves the direct marketing campaigns of a banking institution located in the country of Portugal, from 2008-2013. These marketing campaigns were implemented and performed by phone call.  In some cases, more than one method of contact was required to reach the client to assess if the product of a ‘Bank Term Deposit’ was to be subscribed or not (‘yes’ or ‘no’) by the client.

In this section we divide the variables in the dataset into subgroups depending on the context behind them.

|  |  |
| --- | --- |
| Variable | Description |
| age | (*numerical*) Age of the client |
| job | (*categorial*) Type of career job, such as ‘blue collar’, ‘technician’, ‘entrepreneur’, ‘admin’, ‘unemployed’, etc. |
| marital | (*categorical*) Marital status |
| education | (*categorical*) Education level; such as ‘illiterate’, ‘basic’, ‘university. Degree’, etc. |
| default | (*categorical*) Has credit in default; ‘yes’ or ‘no’ |
| housing | (*categorical*) Has a housing loan; ‘yes’ or ‘no’ |
| loan | (*categorical*) Has a personal loan; ‘yes’ or ‘no’ |

There are also some variables that contain information regarding the bank’s past and present interactions with customers.

|  |  |
| --- | --- |
| Variable | Description |
| contact | (*categorical*) Contact communication type, such as ‘cellular’ and ‘telephone’ |
| month | (*categorical*) Last month of contact with client, such as ‘jan’, ‘feb’, etc |
| day\_of\_the\_week | (*categorical*) Last day of contact with client, such as ‘mon’, ‘tue’, etc |
| duration | (*numeric*) Last contact with client duration, in seconds.  If duration = 0, then y = ‘no’.  Before a call is performed, duration = unknown, and after the call, y is known |
| campaign | (*numeric*) Number of contacts performed during this campaign for the specific client.  This means the number of clients called, and the number of times called for each client |
| pdays | (*numeric*) Number of days that passed after the client was last contacted from a previous campaign.  (where 999 means the client was not previously contacted) |
| previous | (*numeric*) Number of contacts performed before this campaign and for the client |
| poutcome | (*categorical*) Outcome of the previous marketing campaign, such as; ‘failure’, ‘success’, and ‘nonexistent’ |

This last set of variables represent the social and economic factors which might contribute to the success of a telemarketing strategy.

|  |  |
| --- | --- |
| Variable | Description |
| emp.var.rate | (*numeric*) Employment variation rate using a quarterly indicator. |
| cons.price.idx | (*numeric*) Consumer price index using a monthly indicator |
| cons.conf.idx | (*numeric*) Consumer confidence index using a monthly indicator |
| euribor3m | (*numeric*) Euribor 3-month rate using a daily indicator |
| nr.employed | (*numeric*) Number of employees using a quarterly indicator |

(Source: UCI ML Repository)

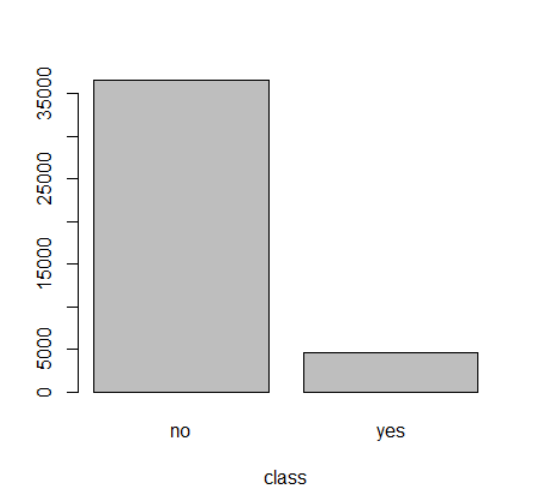
# Methods Used

For our exploratory analysis, we made detailed plots to check for the correlations between predictors and the response variable. This was done separately for the categorical variables and for the numerical variables. Next, the range of each variable was checked to ensure that none of the variables have very extreme values. Finally, a table was made to report the number of null values in each variable. Using the information from the exploratory analysis, a preliminary variable selection was performed on our model.

The response variable is binary, and it can be assumed for now that they are independent Bernoulli trails and thus a logistic regression model was used. Another round of variable selection using the AIC was performed as well as check of the feature significances using chi- square tests. Following this, the model coefficients were interpreted and possible explanations for the results were provided. Goodness-of-fit for the final model was checked, and plots were presented to visualize the residuals and the spread of predicted probabilities to explain the results. Finally, some discussion on the report was written up along with possible improvements to our methodology.

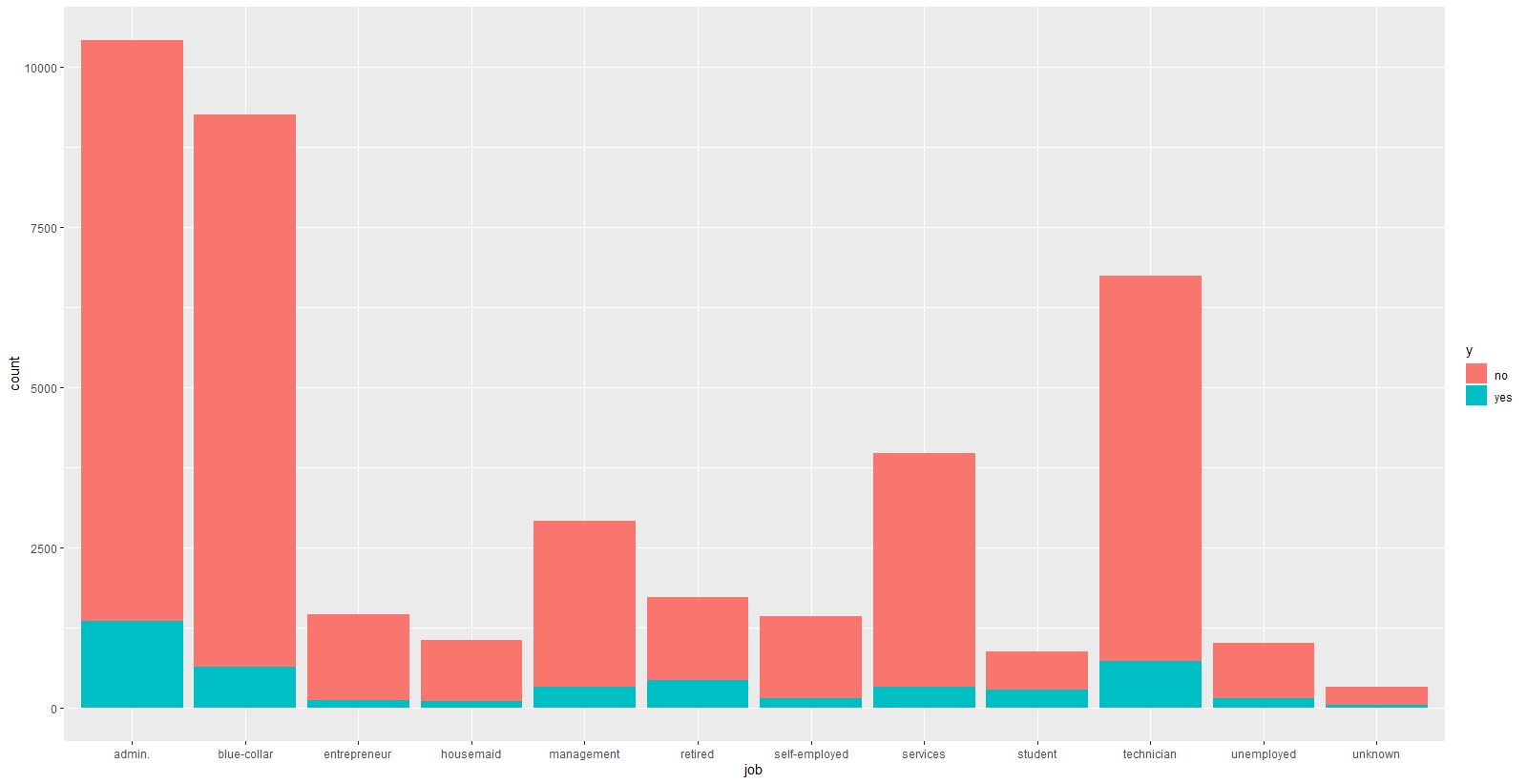
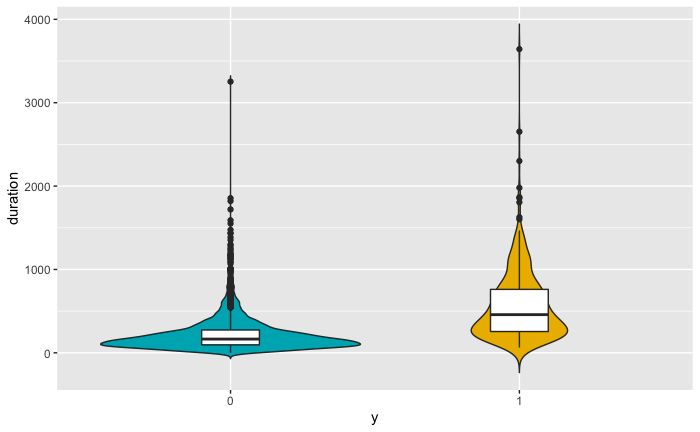
# Results

It can be seen below that the class distribution in the data is extremely unbalanced. This may make accurate parameter estimations difficult.

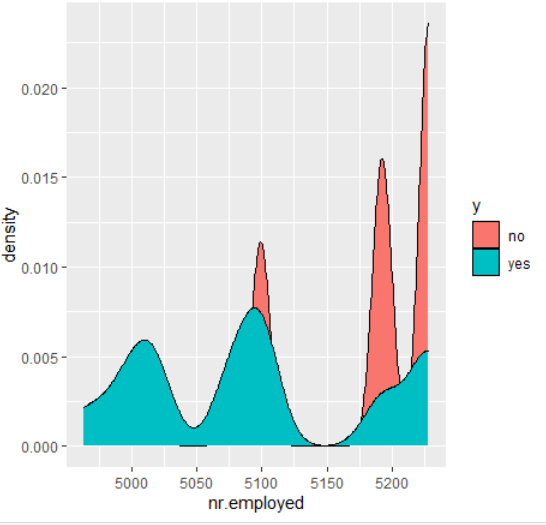
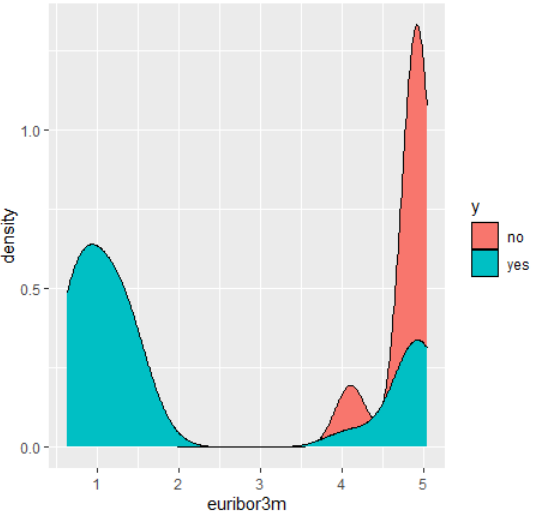
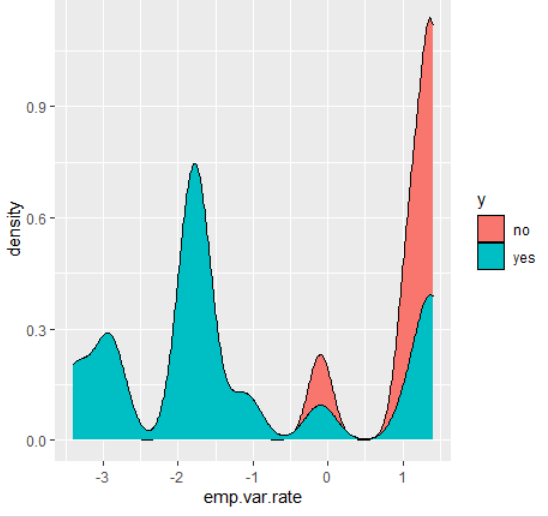


## Exploratory Data Analysis

Our initial guess was the duration is likely to be the most predictive variable because the number of minutes a client is willing to spend talking to a salesperson should indicate their level of interest in the product. From the plot below, we observe that this is indeed the case. As the duration increases, the number of “yes” labels increase. Next, we explore the “job” variable. The plot indicates that people that are retirees, students, unemployed, or in management positions are more likely to sign up for long term deposit accounts. It is surprising that overall, these are some of the least contacted demographics.



Next, we plot the economic variables. The plots of “emp.var.rate”, “eurobor3m” and “nr.employed” are shown below. It is evident from the relative density plots that for all three variables the proportion on “yes” labels is relatively higher at lower values of each variable.



## Main Data Analysis

Based on the above plots in the previous section and the plots in the appendix, we decided to drop the following variables:

|  |  |
| --- | --- |
| **Variable** | **Reason** |
| pdays | Mostly null values |
| poutcome | Mostly “non-existent” |
| day\_of\_week | Seems unpredictive based on proportionality plot |
| default | Entirely populated with “No” |
| housing | Seems unpredictive based on proportionality plot |

Thus, we fit the following model.

And,

This fit results in the following statistics:

* Deviance = 17608
* Null Deviance = 28997
* Difference = 11389

The standard errors of the coefficients of this model (in the appendix) are very large. It is true that some of them may be overestimated. Regardless, we prefer to have a simpler model. We use stepwise AIC to select the best subset of variables from the model. After some iterations, we obtain a smaller model that maximizes the AIC. We can see that using this criterion allows us to drop the variables “age”, “marital”, “nr.employed” and “loan”. Dropping “age” is a surprising result because it seems to make sense logically as a predictor. To explain this, we hypothesize that perhaps aspects of “age” are captured through the “job” and “education” variables, thus making “age” not significant. The linear estimator of this model is of the form:

The AIC of the full model is 17694 and that of the final model is 17685. Therefore, the improvement is just marginal. Next, we test the significance of each of the predictors in this model by dropping them one at a time and using a chi-squared nested model test.

|  |  |
| --- | --- |
| **predictor** | **Pr(>Chi)** |
| job | 3.221e-07 |
| education | 0.02797 |
| month | 2.2e-16 |
| duration | 2.2e-16 |
| campaign | 3.924e-05 |
| emp.var.rate | 2.2e-16 |
| cons.price.idx | 2.2e-16 |
| euribor3m | 1.471e-08 |
| contact | 2.2e-16 |
| Previous | 4.882e-07 |
| cons.conf.idx | 6.441e-05 |

We can see from this exercise that all of the p-values are less that 0.05. Therefore, for the chi-squared test for each predictor, at the 0.05 significance level we can reject the null hypothesis that a model without that preditor is a better fit. Therefore, at this point, the best model choice is to not remove any more predictors.

## Significant Coefficients and Interpretations

In this section we pick four parameters of interest and interpret their coefficients. We take the exponent of the coefficients to get changes in the odds.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coefficient** | **Change in Odds** |
| Duration | 4.71e-03 | 1.005 |
| cons.conf.idx | 2.11e-02 | 1.021 |
| jobstudent | 2.55e-01 | 1.29 |
| Previous | 1.56e-01 | 1.169 |

The interpretation of these coefficients are as follows:

* Holding all other variables fixed, a 1 minute increase in duration results in a 0.5%

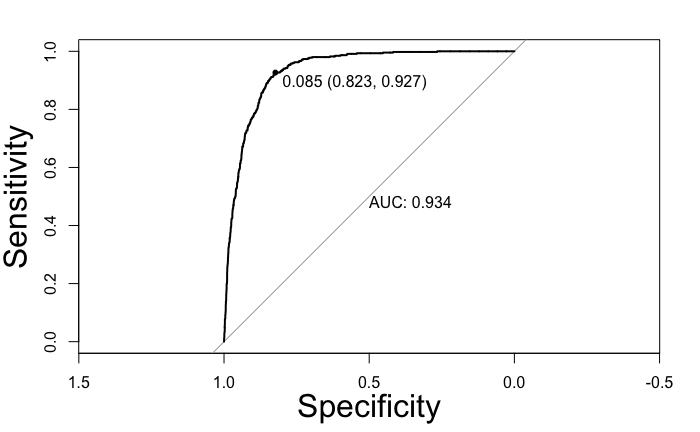
increase in the odds.

* Holding all other variables fixed, a unit increase in consumer confidence index results in a 2.1% increase in the odds
* Holding all other variables fixed, if the customer is a student, the odds of subscribing increase by 29%. This is in line with our initial hypothesis.
* Holding all other variables fixed, for every unit increase in the number of times a customer is contacted, there is a 16.9% increase in the odds.

## 

## Model Performance

It is interesting to note that all of the variables that were considered important in our exploratory analysis, turned out to be important to the model. In the next step, we assess the performance of our finalized model. First we can plot a ROC curve to obtain the optimal threshold. It is surprising that the optimal threshold is a relatively small value (0.085). However, the AUC of 0.934 indicates that our model has a good performance on this dataset.



Using the optimal threshold, we create a classification error table. We can observe from this table that using the optimal threshold, only 304 cases of “yes” were classified as “no” whereas 6957 cases of “no” were classified as “yes”. In our case, this tradeoff is acceptable because, for a bank it is not a significant loss to advertise to customers who may not be interested in the product. On the other hand, it would be a big loss to ignore customers who are potentially interested in the product.

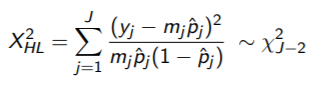
|  |  |  |
| --- | --- | --- |
| Predicted Actual | Yes | No |
| Yes | 4326 | 6897 |
| No | 314 | 29647 |

From this table we can also calculate the following performance statistics.

|  |  |
| --- | --- |
| Threshold | 0.08 |
| Sensitivity | 0.9334 |
| Specificity | 0.8119 |
| Accuracy | 0.8256 |
| AUC | 0.9317 |

Goodness-of-fit and Model Diagnostics

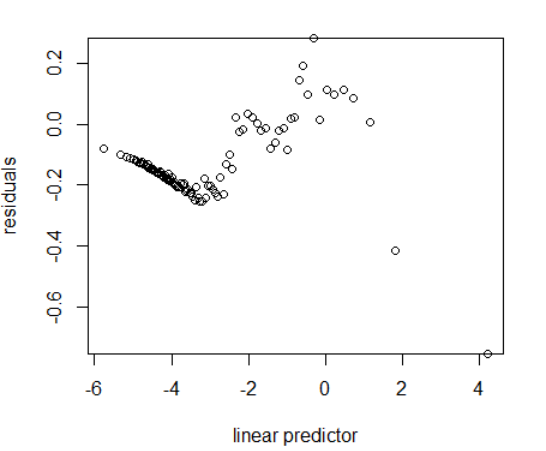
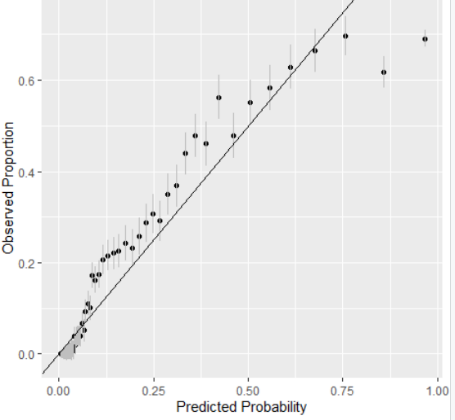
First, we can test for goodness of fit using the Hosmer-Lemeshow statistic which has a chi-squared distribution. It is defined as:



we use the parameter J = 10 and get the following results.

* X-squared = 571.12
* p-value < 2.2e-16

Since the p-value is very small, we can reject the null hypothesis, justifying that there is “no lack of fit”. Thus, we conclude that our model does not fit the data well. Next, we attempt to decipher why this might be the case and whether it is possible to remedy this. In plots below (left) we check if the predicted probabilities closely follow the observed proportions within 100 cuts based on the quantiles of the linear predictor. We can observe that the predicted probabilities are mostly lower than the observed proportions, which means there is something wrong, potentially, in the model specification. From the other plot below (right), we can see that the residuals are not evenly dispersed in their variation with respect to the linear predictor. This is yet another indication that there is something wrong in the model specification.



# Conclusion

We noted in the previous sections that based on the AUC and metrics such as Specificity and Sensitivity, our model seemed to have a decent performance. We also noted that the parameters used in the final model were all significant and contributed to the improvement of the model. However, in the model diagnostics section we noted that under some scrutiny and statistical tests like Hosmer-Lemeshow statistic, or a basic chi-squared test of difference in deviance between a null model and the full model, that our final model seemed to have some error in specification (lack of fit). One reason for this could be that our class distribution is extremely unbalanced. Perhaps the relationship we tried to capture is not linear and therefore we it might be feasible to include higher order terms. It is also possible that it is beneficial to apply transformations to some of the predictors. Moreover, perhaps there is omitted variable bias in our model, for example, in missing some variables that are crucial in predicting the response. Finally, we acknowledge that perhaps there is a better model choice. For example, perhaps we would have had more success with an additive model because of its added flexibility in dealing with non-linear relationships.

# Contributions

**Dario**: Created the formatting from main page and its contents, and separated the sections of work, wrote the introduction, created the variable charts for the descriptions of the data, writing out the descriptions for the abbreviations of predictors. Listed the references, edited the entire document and proof-read / grammar checked. Also created the PowerPoint of slides for the presentation, as well as performed the presentation for group 12.

**Soham:** Wrote most of and rechecked the code, wrote the sections of results, goodness-of-fit, model diagnostics, and the conclusion in the report. Also made the plots and wrote the explanations for all the plots (including variable plots and diagnostic plots). Proofread the document. Made all the tables in the results section. Organized the code in the appendix including section numbers and titles. Added references.

**Sun:** Helped in writing some parts of the code including the Classifications table and ROC curve. Also assisted in the interpretations of coefficients.

**Liu:** Attended office hours once to clarify some point about formatting.

# References

Moro, S., & Cortez, P. (n.d.). UCI machine Learning Repository: Bank marketing data set. Retrieved April 14, 2021, from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Eberly College, P. (n.d.). 4.3 - RESIDUALS vs. Predictor PLOT: Stat 501. Retrieved April 14, 2021, from <https://online.stat.psu.edu/stat501/lesson/4/4.3>

# Appendix

## Section 1: setup

```{r setup, include=FALSE, results='hide', fig.show='hide'}

knitr::opts\_chunk$set(echo = TRUE)

library(ggplot2)

library(ggplot2)

library(dplyr)

library(faraway)

library(tidyr)

library(ResourceSelection)

```

Section 2: Data cleaning and setting categorical variables as factors

```{r 1 }

bank <- read.csv('bank-additional-full.csv')

## transform into some factors

bank$job <- as.factor(bank$job)

bank$marital <- as.factor(bank$marital)

bank$education <- as.factor(bank$education)

bank$default <- as.factor(bank$default)

bank$housing <- as.factor(bank$housing)

bank$loan <- as.factor(bank$loan)

bank$day\_of\_week <- as.factor(bank$day\_of\_week)

bank$contact <- as.factor(bank$contact)

bank$month <- as.factor(bank$month)

bank$poutcome <- as.factor(bank$poutcome)

bank$y <- as.factor(bank$y)

summary(bank)

## check for missing values and we find no missing values

sum(!complete.cases(bank))

```

## converting response to 1/0

## Section 3: descriptive statistics

```{r 2}

## convert y into 0-1 value for further calculations

bank$y <- ifelse(bank$y=='yes',1,0)

barplot(table(bank$y), xlab = 'class') #class distribution

bank$y <- as.factor(bank$y)

#plot numerical variables density

cols <- names(bank)[sapply(bank, is.numeric)]

for(col in cols){

plot(density(bank[,col]),xlab = col)

}

cols2 <- names(bank)[which(!names(bank) %in% cols)]

#plot categorical varibale bar plot.

for(col in cols2){

barplot(summary(bank[,col]), xlab = col)

print(col)

}

```

## Section 4: plotting variables of interest

```{r 3}

ggplot(bank, aes(x=default, fill = y)) + geom\_bar()

ggplot(bank, aes(x=education, fill = y)) + geom\_bar()

ggplot(bank, aes(x=day\_of\_week, fill = y)) + geom\_bar() #might have an impact in combination with other variables

ggplot(bank, aes(x=loan, fill = y)) + geom\_bar()

ggplot(bank, aes(x=previous, fill = y)) + geom\_bar()

ggplot(bank, aes(x=campaign, fill = y)) + geom\_bar()

ggplot(bank, aes(x=contact, fill = y)) + geom\_bar()

ggplot(bank, aes(x=housing, fill = y)) + geom\_bar()

ggplot(bank, aes(x=emp.var.rate, fill = y)) + geom\_density()

ggplot(bank, aes(x=emp.var.rate, fill = y)) + geom\_bar()

ggplot(bank, aes(x=duration, fill = y)) + geom\_bar()

ggplot(bank, aes(x=euribor3m, fill = y)) + geom\_density()

ggplot(bank, aes(x=nr.employed, fill = y)) + geom\_density()

```

## Section 5: Violin plots to understand the distributions.

```{r 4}

e <- ggplot(bank, aes(x = y, y = cons.conf.idx))

# Combine with box plot to add median and quartiles

# Change color by groups

e + geom\_violin(aes(fill = y), trim = FALSE) +

geom\_boxplot(width = 0.2)+

scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))+

theme(legend.position = "none")

e <- ggplot(bank, aes(x = y, y = nr.employed))

# Combine with box plot to add median and quartiles

# Change color by groups

e + geom\_violin(aes(fill = y), trim = FALSE) +

geom\_boxplot(width = 0.2)+

scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))+

theme(legend.position = "none")

```

## Section 6: null values

```{r 5}

bank %>%

summarise\_all(list(~sum(. == "unknown"))) %>%

gather(key = "variable", value = "nr\_unknown") %>%

arrange(-nr\_unknown)

```

## Section 7: First round of variable selection and initial model fit

```{r 6}

##dropping bad variables

drops <- c("pdays","poutcome", "day\_of\_week", "default",'housing' )#Made some improvements.

bank <- bank[ , !(names(bank) %in% drops)]

bank <- bank %>% dplyr::filter(duration != 0)#remove customerswho didnt pick up the calls.

##fiting the model

logistic\_fit <- glm(y~.,data=bank,family=binomial)

sumary(logistic\_fit)

```

## Section 8: model Selection and performance

```{r 7}

#Model selection with AIC

best\_logistic <- step(logistic\_fit, trace=1)

sumary(best\_logistic)

#testing nested models.

drop1(best\_logistic,test="Chi")

#Creating classification table and ROC curve

ypred <- predict(best\_logistic,newdata=bank,type="response")

bank\_2 <- bank

bank\_2$pred <- ypred

testm <- mutate(bank\_2, pred=ifelse(pred < 0.08, 0, 1))

xtabs( ~ y + pred, testm)

suppressMessages(library(pROC))

roc\_obj <- roc(response=bank\_2$y, predictor=bank\_2$pred)

AUC <- auc(roc\_obj)

roc\_logistic <- c(coords(roc\_obj, "b",

ret=c("threshold","se","sp","accuracy"),

best.method="youden"),AUC)

names(roc\_logistic) <- c("Threshold","Sensitivity","Specificity",

"Accuracy","AUC")

t(roc\_logistic)

plot(roc\_obj,legacy.axes=FALSE,print.auc=TRUE,print.thres=TRUE,cex.lab=2)

```

## Section 9: interpreting the coefficients.

```{r 8}

#interpretation of coefficients.

beta <- coef(best\_logistic)

odds <- exp(beta)

```

## Section 10: Model diagnostic plots and Hosmer-Lemeshow statistic

```{r 9}

#Hosmer-lemeshow test

hoslem.test(best\_logistic$y,fitted(best\_logistic),g=10)

#Making plost

linpred = predict(best\_logistic)

quantiles <- quantile(linpred, (1:100)/101)

length(unique(quantiles))

bank\_copy <- bank

bank\_copy$linpred <- linpred

bank\_copy$predprob <- predict(best\_logistic,type="response")

gdf <- group\_by(bank\_copy,cut(linpred, breaks=c(min(linpred),unique(quantiles),max(linpred)),include.lowest = TRUE))

hldf <- dplyr::summarise(gdf, y2 =sum(y == 1 ),

ppred=mean(predprob), count=n())

hldf <- mutate(hldf, se.fit=sqrt(ppred\*(1-ppred)/count))

head(hldf)

#predicted probability vs observed proportion

ggplot(hldf,aes(x=ppred,y=y2/count,ymin=y2/count-2\*se.fit,

ymax=y2/count+2\*se.fit))+

geom\_point()+geom\_linerange(color=grey(0.75))+

geom\_abline(intercept=0,slope=1)+

xlab("Predicted Probability")+ylab("Observed Proportion")

#residuals vs linear predictor plot.

gdf$residuals <- residuals(best\_logistic)

diagdf <- summarise(gdf, residuals=mean(residuals), linpred=mean(linpred))

plot(residuals ~ linpred, diagdf, xlab="linear predictor",cex.lab=1)

```