Lam Lay
EXST 7142
Instructor: Bin Li
Final Individual Project

Bank marketing campaigns dataset analysis

1. Introduction

Successful businesses need to gather information about clients to find the best-selling products and target the right customers. In this paper, I will explore and analyze a dataset that describes Portugal bank marketing campaigns results. This dataset can be found in the UCI Machine Learning Repository [1] and Kaggle [2]. The campaigns were mostly done via direct phone calls, and their goals are to convince the customers to place a bank term deposit. From this dataset, I will find important variables that affect the outcome of the campaign and develop two machine learning models to predict whether a client will subscribe a bank term deposit or not. Because the target is binary ("Yes" and "No"), this is a classification problem. In this paper, I will use Logistic Regression and Decision Tree models to predict the target and compare their predictive results.

2. Dataset Description

The dataset is called "Bank Marketing Data Set" and obtained from UCI Machine Learning Repository [1]. There are 45211 observations and 17 variables in the dataset. The response variable/output is binary with 1 representing the client has subscribed a term deposit and 0 representing the client has not subscribed a term deposit. Therefore, this is a binary classification problem.

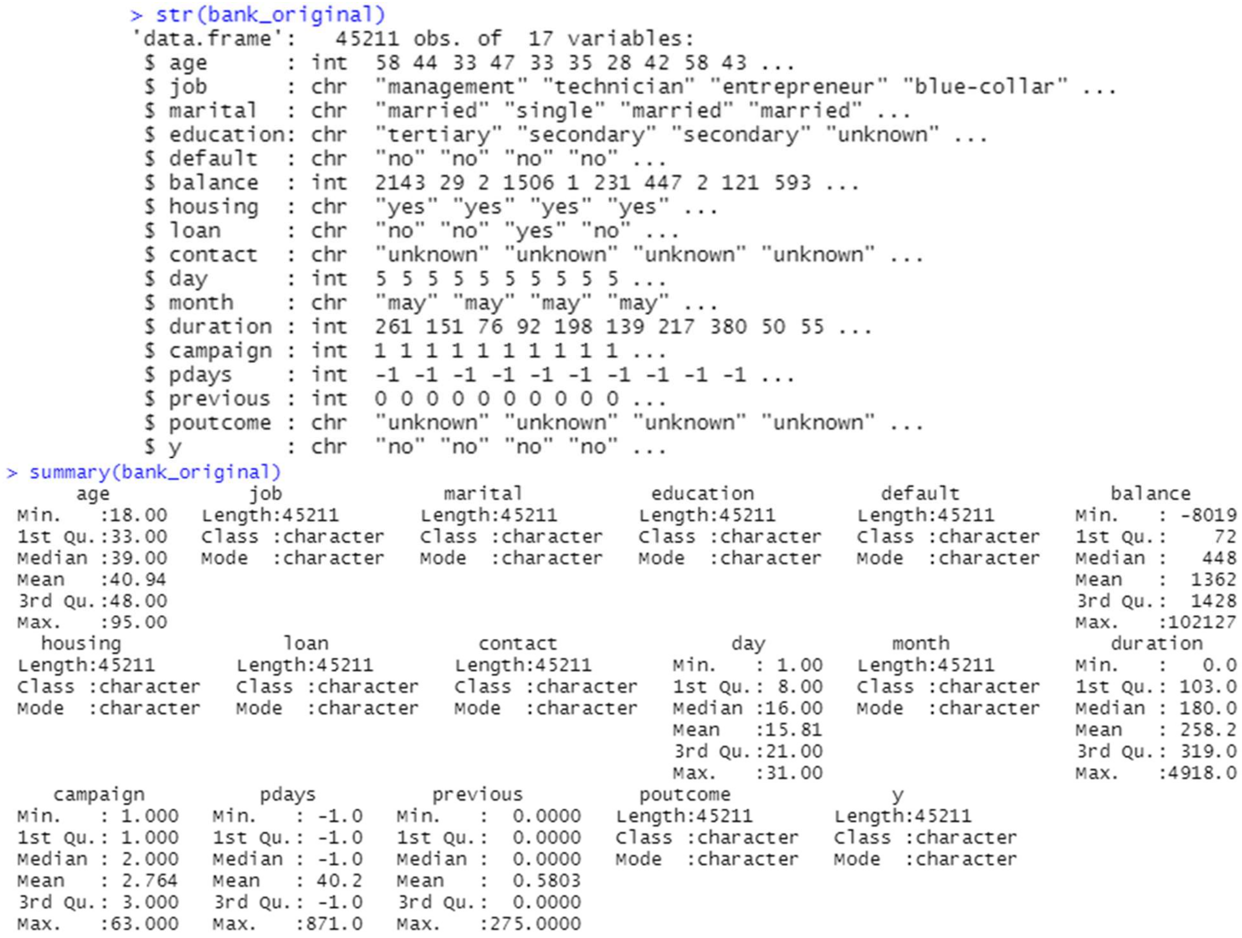
The dataset (*bank-full.csv*) that I downloaded from UCI has less variables and more observations that the one described in Kaggle, but I used the Data Description page in Kaggle to understand the variables [1,2]:

1. **age**: Age (numeric)
2. **job**: Type of job (categorical, 12 job titles)
3. **marital**: Marital status (categorical: "married", "single", "divorced")
4. **education**: Education (categorical: "tertiary", "secondary", "unknown", "primary")
5. **default**: has credit in default? (categorical: "no", "yes")
6. **balance**: has balance loan? (numeric)
7. **housing**: has housing loan? (categorical: "no", "yes")
8. **loan**: has personal loan? (categorical: "no", "yes")
9. **contact**: contact communication type (categorical: "cellular", "telephone", "unknown")
10. **month**: last contact month of year (categorical)
11. **day**: last contact day of a month
12. **duration**: last contact duration, in seconds (numeric)
13. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric: 999 means client was not previously contacted)
15. **previous**: number of contacts performed before this campaign and for this client (numeric)

16. **poutcome**: outcome of the previous marketing campaign (categorical: "unknown", "failure", "other", "success")
17. **y**: target, has the client subscribed a term deposit? (binary: "no", "yes")

3. Data cleaning and exploration:

3.1. First look at the dataset



```
> str(bank_original)
'data.frame':   45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr  "married" "single" "married" "married" ...
 $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
```

```
> summary(bank_original)
      age            job              marital           education           default             balance      
 Min.   :18.00   Length:45211       Length:45211       Length:45211       Length:45211       Min.   : -8019  
 1st Qu.:33.00   Class :character   Class :character   Class :character   Class :character   1st Qu.:    72  
 Median :39.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   448  
 Mean   :40.94                                                                               Mean   :  1362  
 3rd Qu.:48.00                                                                               3rd Qu.:  1428  
 Max.   :95.00                                                                               Max.   :102127  
   housing              loan              contact               day           month              duration     
 Length:45211       Length:45211       Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0  
 Class :character   Class :character   Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0  
 Mode  :character   Mode  :character   Mode  :character   Median :16.00   Mode  :character   Median : 180.0  
                                                          Mean   :15.81                      Mean   : 258.2  
                                                          3rd Qu.:21.00                      3rd Qu.: 319.0  
                                                          Max.   :31.00                      Max.   :4918.0  
    campaign          pdays          previous           poutcome              y            
 Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211       Length:45211      
 1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character   Class :character  
 Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character   Mode  :character  
 Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803                                        
 3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000                                        
 Max.   :63.000   Max.   :871.0   Max.   :275.0000                                        
```

Fig.1. Summary of the dataset

This data set contains 45211 observations and 17 variables as previously mentioned. There are 7 numerical variables (presented as integers) and 10 categorical variables (presented as characters) as shown in figure 1. Only 11.7% of the clients chose to subscribe to the bank term deposit after the campaign; the rest which accounts for 88.2% of the clients did not.

Before cleaning and variable selection, there are no missing values nor duplicated rows in the dataset. However, this may change after data cleaning and removing some variables.

3.2. Graphical Presentation

Before training the models, I analyzed the dataset to understand the subscription distribution for each variable using bar plots and histograms. From figure 2, most customers that subscribed are between age 25 to 45. However, this is probably because most customers who participated in the campaign were in this age range, with the mean age for all customers is 40.94. Based on the ratio of subscribed and not subscribed, people who are younger than 30 and those over 60 years old are more willing to place a deposit than middle aged people. From the same figure, the majority of subscriptions were resulted from less than 5 number of contacts during the campaign, probably because the bank is likely to stop contacting the same person if he/she refused to subscribed after the first few contacts. Most customers that subscribed are blue-collar, management, and technician, but based on the ratio of subscribed and not subscribed, students are the most likely to subscribe. This makes sense because students who attend college far from home need to make new bank accounts for their spending and stipend.

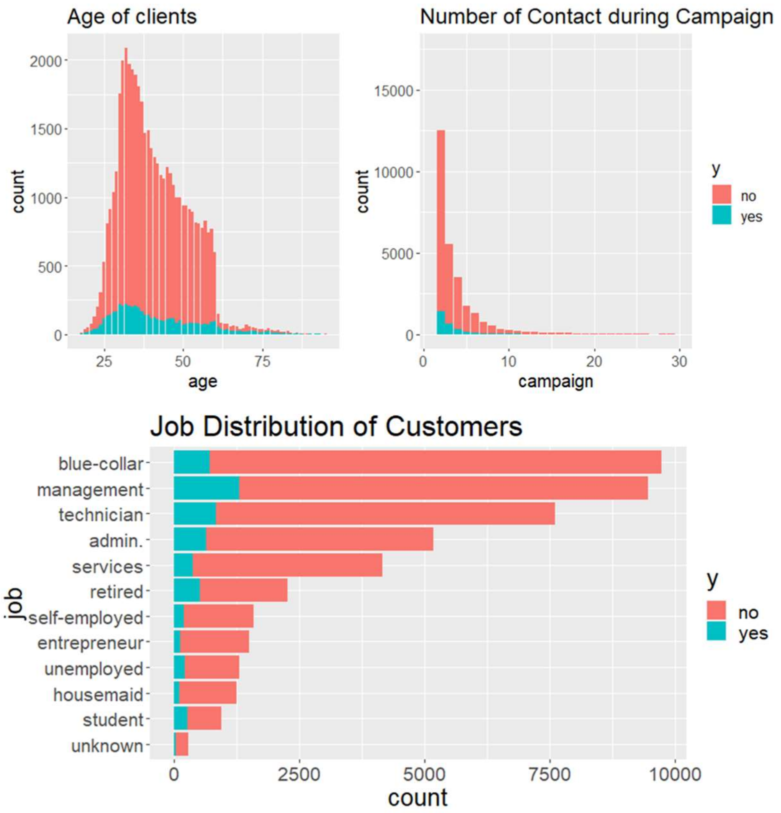


Fig.2. Subscriptions based on age of customers, number of contacts during this campaign, and the job distribution of customers.

Figure 3 shows that majority of customers that subscribe are married or single, have had secondary or tertiary education, have credit in default, and do not have personal and/or housing loans. Their balance loans are low (usually between 0 and 25,000 Euro). Most of them were contacted via cell phone and from April to August. The data description section mentions number 999 for client that was not previously contacted for variable *pdays*, but my dataset has no observations with this number.

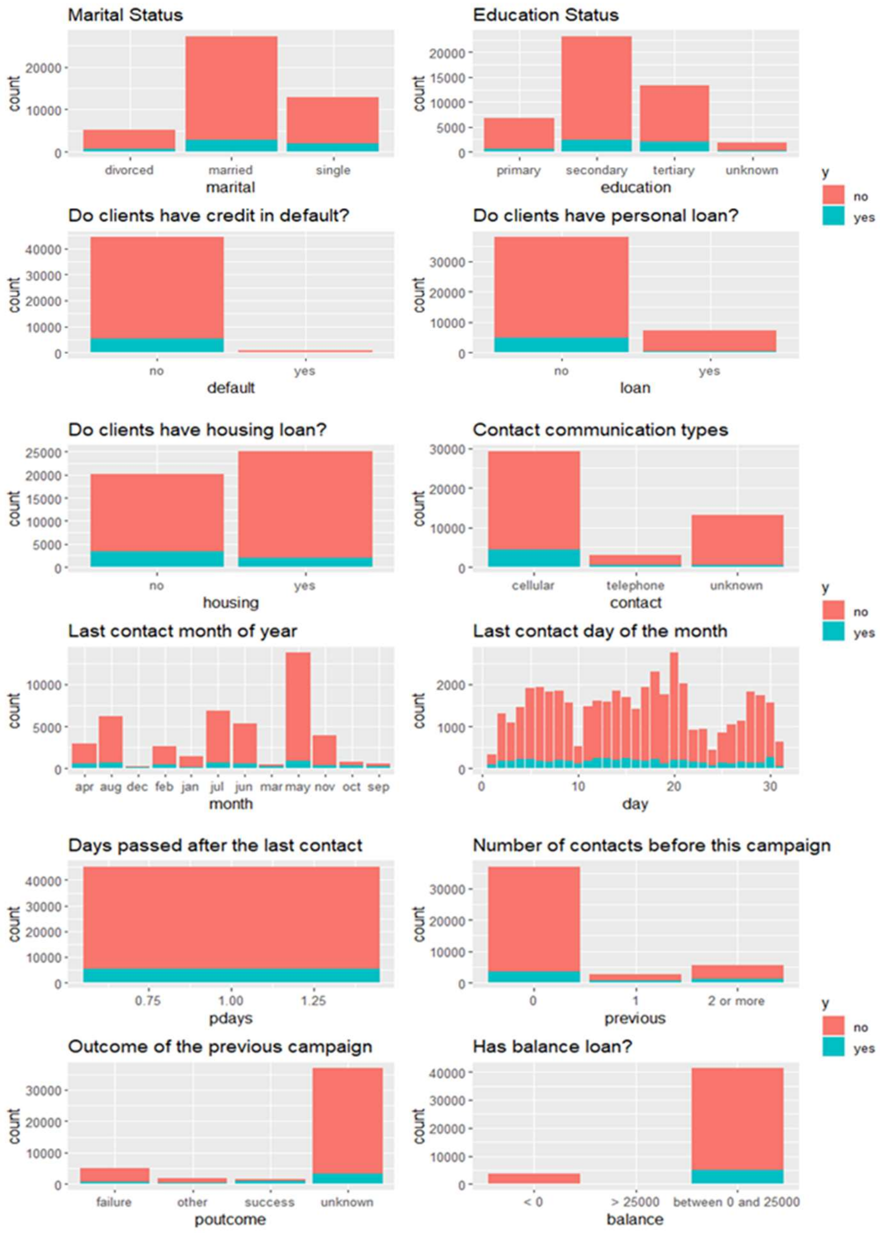
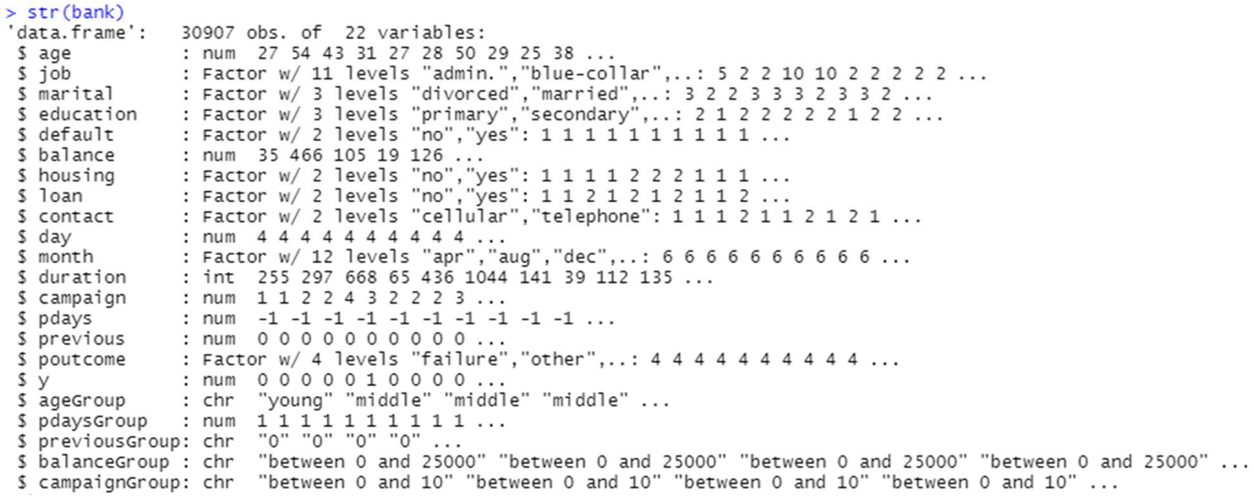


Fig. 3. Subscriptions based on other variables.

4. Methods:

4.1. Data cleaning and variable selection

The next step is cleaning the data to prepare it for training models. First, I removed the rows with “unknown” values for variables *job, education,* and *contact* as shown in figure 2 and 3. I did not remove these rows for variable *poutcome* because 75% of its observations have “unknown” values. For variable selection, I calculated information values to measure of the predictive capability of the independent variables [3]. I also created a logistic regression model on the entire dataset to determine which variables are most significant in predicting the response. Because most of the variables are categorical, I did not create any correlation matrix that is only meaningful for numerical variables. Before training a logistic regression model, I converted the integer variables into numeric and character variables into factor (as shown in figure 4).



```
> str(bank)
'data.frame':	30907 obs. of  22 variables:
 $ age          : num  27 54 43 31 27 28 50 29 25 38 ...
 $ job          : Factor w/ 11 levels "admin.","blue-collar",..: 5 2 2 10 10 2 2 2 2 2 ...
 $ marital      : Factor w/ 3 levels "divorced","married",..: 3 2 2 3 3 3 2 3 3 2 ...
 $ education    : Factor w/ 3 levels "primary","secondary",..: 2 1 2 2 2 2 2 1 2 2 ...
 $ default      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ balance      : num  35 466 105 19 126 ...
 $ housing      : Factor w/ 2 levels "no","yes": 1 1 1 1 2 2 2 1 1 1 ...
 $ loan         : Factor w/ 2 levels "no","yes": 1 1 2 1 2 1 2 1 1 2 ...
 $ contact      : Factor w/ 2 levels "cellular","telephone": 1 1 1 2 1 1 2 1 2 1 ...
 $ day          : num  4 4 4 4 4 4 4 4 4 4 ...
 $ month        : Factor w/ 12 levels "apr","aug","dec",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ duration     : int  255 297 668 65 436 1044 141 39 112 135 ...
 $ campaign     : num  1 1 2 2 4 3 2 2 2 3 ...
 $ pdays        : num  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome     : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ y            : num  0 0 0 0 0 1 0 0 0 0 ...
 $ ageGroup     : chr  "young" "middle" "middle" "middle" ...
 $ pdaysGroup   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ previousGroup: chr  "0" "0" "0" "0" ...
 $ balanceGroup : chr  "between 0 and 25000" "between 0 and 25000" "between 0 and 25000" "between 0 and 25000" ...
 $ campaignGroup: chr  "between 0 and 10" "between 0 and 10" "between 0 and 10" "between 0 and 10" ...
```

Fig.4. Final class of each variable before training predictive models

Although boxplots can be used to find potential outliers, we do not know if they are true outliers or not. By definition, any points that are beyond 2 or 3 standard deviations from the mean are outliers and we can remove them. However, I created diagnostic plots from the previous logistic regression model and found the Cook's distance to determine which points are most likely outliers.

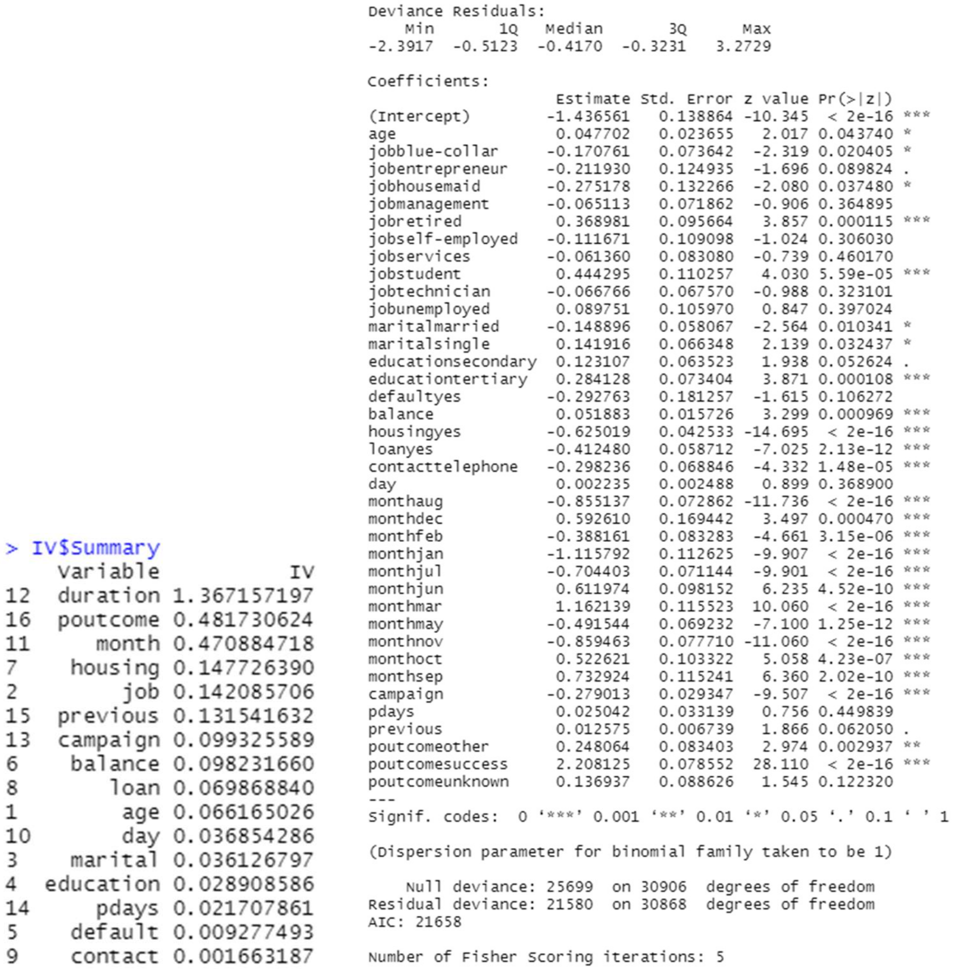
4.2. Predictive models

Before creating any model, I scaled the numerical variables in the dataset. After the selecting significant variables, I separated the dataset into a training set and a test set using 70/30 split ratio. A logistic regression model and a decision tree model were trained on the training set. The test set was used for evaluation purpose. Comparison between models were done using confusion matrix, ROC curve, and AUC. Metrics such as accuracy, sensitivity, specification, misclassification rate, etc. were calculated in Excel based on the confusion matrix [4].
Note that the data description in Kaggle says that the variable *duration* "highly affects the output target…" and "should be discarded if the intention is to have a realistic predictive model" [2]. Therefore, I developed the models with and without variable *duration* and compare their results. Also note that after removing variable *duration* from the dataset, some rows become duplicated, so I removed them before training the models.

5. Results

5.1. Variable Selection

Figure 5 shows that *duration* has the highest information value, indicating it has the most predictive power. Variables *default* and *contact* have information value less than 0.02, so they have almost no predictive power. On the other hand, the logistic models on the entire dataset show that: when *duration* is included, variables *age, default, pdays,* and *previous* are not significant; when *duration* is not included, variables *default, day, pdays,* and *previous* are not significant. Both models show that the "telephone" category of *contact* actually has high predictive power, so I decided to keep the *contact* variable. Because it is more realistic to not include *duration* variable, I chose to remove variables *default, day, pdays,* and *previous* from the dataset regardless whether *duration* was included or not before splitting the dataset into a training set and a test set.



```
> IV$Summary
    Variable          IV
12  duration 1.367157197
16  poutcome 0.481730624
11     month 0.470884718
7    housing 0.147726390
2        job 0.142085706
15  previous 0.131541632
13  campaign 0.099325589
6    balance 0.098231660
8       loan 0.069868840
1        age 0.066165026
10       day 0.036854286
3    marital 0.036126797
4  education 0.028908586
14     pdays 0.021707861
5    default 0.009277493
9    contact 0.001663187
```

```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3917  -0.5123  -0.4170  -0.3231   3.2729  

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -1.436561   0.138864 -10.345  < 2e-16 ***
age                 0.047702   0.023655   2.017 0.043740 *  
jobblue-collar     -0.170761   0.073642  -2.319 0.020405 *  
jobentrepreneur    -0.211930   0.124935  -1.696 0.089824 .  
jobhousemaid       -0.275178   0.132266  -2.080 0.037480 *  
jobmanagement      -0.065113   0.071862  -0.906 0.364895    
jobretired          0.368981   0.095664   3.857 0.000115 ***
jobself-employed   -0.111671   0.109098  -1.024 0.306030    
jobservices        -0.061360   0.083080  -0.739 0.460170    
jobstudent          0.444295   0.110257   4.030 5.59e-05 ***
jobtechnician      -0.066766   0.067570  -0.988 0.323101    
jobunemployed       0.089751   0.105970   0.847 0.397024    
maritalmarried     -0.148896   0.058067  -2.564 0.010341 *  
maritalsingle       0.141916   0.066348   2.139 0.032437 *  
educationsecondary  0.123107   0.063523   1.938 0.052624 .  
educationtertiary   0.284128   0.073404   3.871 0.000108 ***
defaultyes         -0.292763   0.181257  -1.615 0.106272    
balance             0.051883   0.015726   3.299 0.000969 ***
housingyes         -0.625019   0.042533 -14.695  < 2e-16 ***
loanyes            -0.412480   0.058712  -7.025 2.13e-12 ***
contacttelephone   -0.298236   0.068846  -4.332 1.48e-05 ***
day                 0.002235   0.002488   0.899 0.368900    
monthaug           -0.855137   0.072862 -11.736  < 2e-16 ***
monthdec            0.592610   0.169442   3.497 0.000470 ***
monthfeb           -0.388161   0.083283  -4.661 3.15e-06 ***
monthjan           -1.115792   0.112625  -9.907  < 2e-16 ***
monthjul           -0.704403   0.071144  -9.901  < 2e-16 ***
monthjun            0.611974   0.098152   6.235 4.52e-10 ***
monthmar            1.162139   0.115523  10.060  < 2e-16 ***
monthmay           -0.491544   0.069232  -7.100 1.25e-12 ***
monthnov           -0.859463   0.077710 -11.060  < 2e-16 ***
monthoct            0.522621   0.103322   5.058 4.23e-07 ***
monthsep            0.732924   0.115241   6.360 2.02e-10 ***
campaign           -0.279013   0.029347  -9.507  < 2e-16 ***
pdays               0.025042   0.033139   0.756 0.449839    
previous            0.012575   0.006739   1.866 0.062050 .  
poutcomeother       0.248064   0.083403   2.974 0.002937 ** 
poutcomesuccess     2.208125   0.078552  28.110  < 2e-16 ***
poutcomeunknown     0.136937   0.088626   1.545 0.122320    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25699  on 30906  degrees of freedom
Residual deviance: 21580  on 30868  degrees of freedom
AIC: 21658

Number of Fisher Scoring iterations: 5
```

Fig.5. Information values (with *duration*) and significance of each variable in the full model (without *duration*).

5.2. Potential Outliers

The diagnostic plots and Cook’s distance of the full logistic regression models with and without *duration* are shown in figure 6. When *duration* is included, observations 11037 and 10997 are outliers, and when *duration* is not included, observation 5025 is the only outlier. Observation 15853 is the influence point regardless whether *duration* is included or not. Since there are only 3 outliers total and one influence point out of 30,907 observations in the dataset (not including “unknown” values), I decided to not remove them unless my predictive results on the test set were very bad.

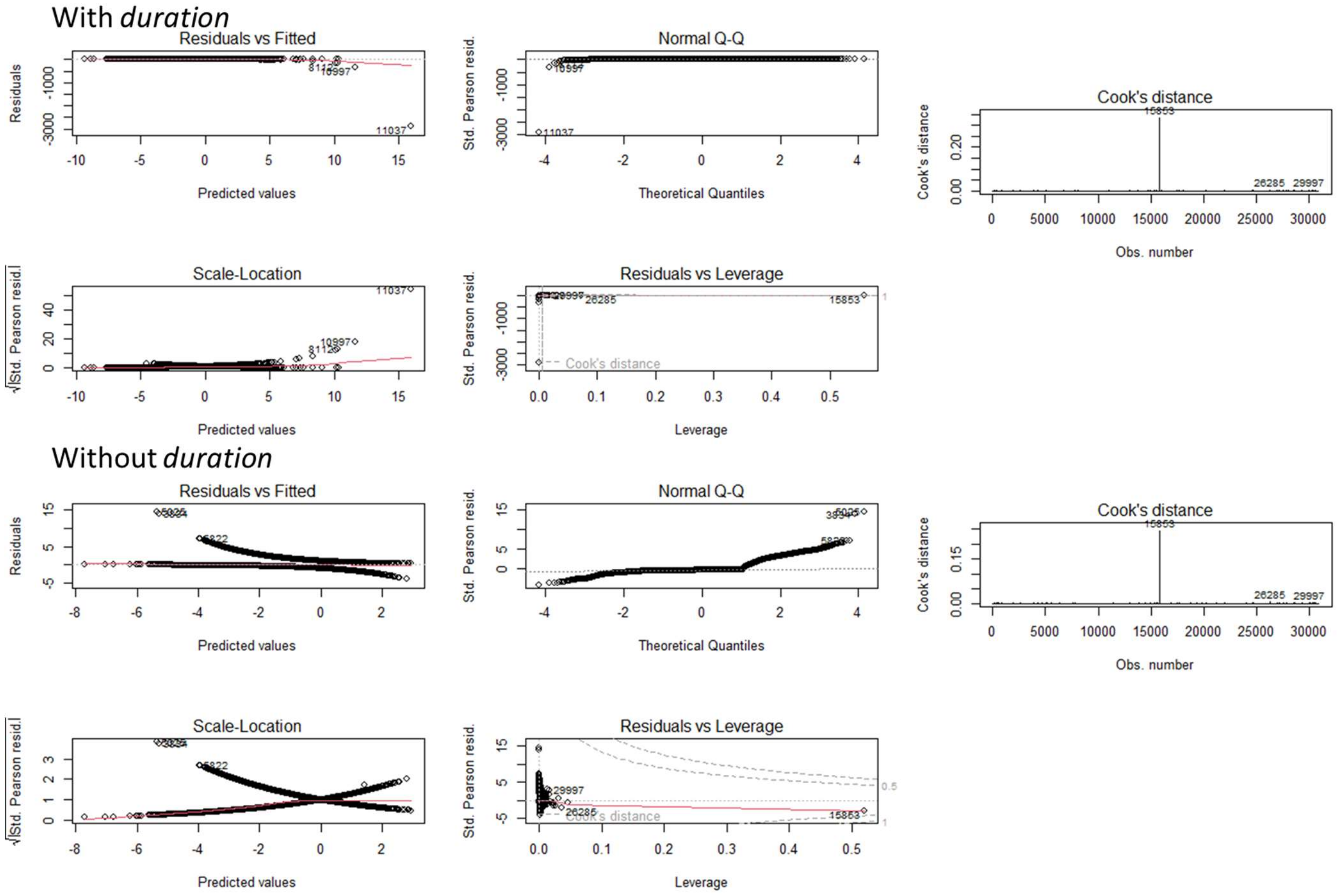


Fig. 6. Diagnostic plots and Cook's distance of logistic regression models on the entire dataset with and without variable *duration*.

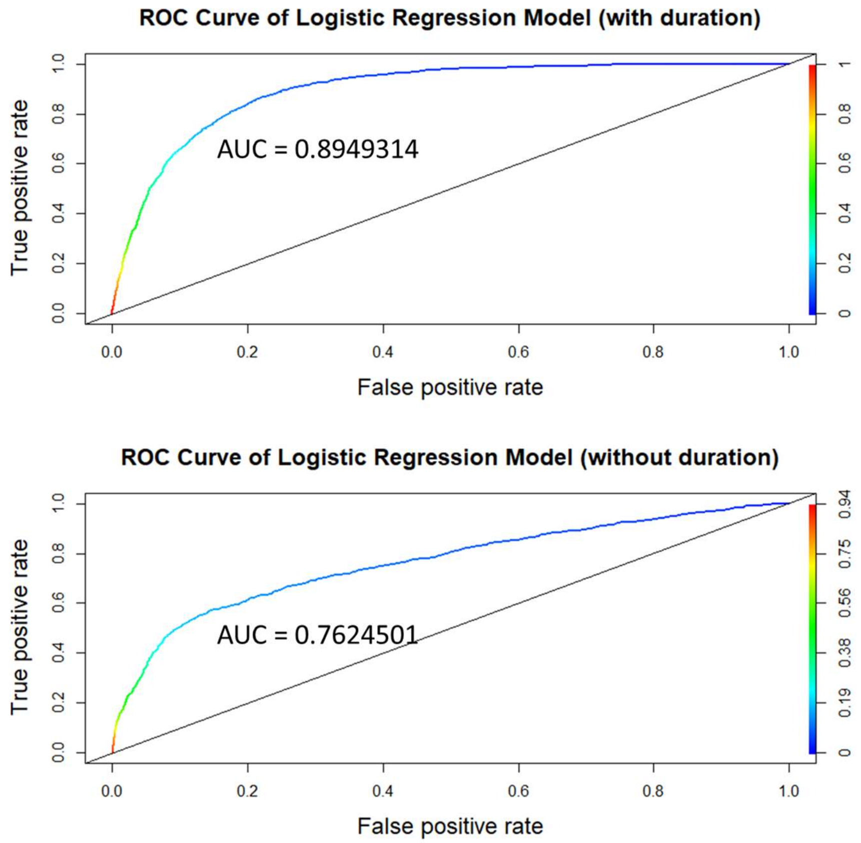
5.3. Predictive models

After the models are trained on the training set, they were used to predict the response on the test set. Table 1 compares the predictive results of all models with and without variable *duration*. All models have accuracy greater than 85% and misclassification rate less than 15%. When *duration* was used, misclassification rate decreases for about 2% and sensitivities increase from 20% to 50%. Specificities and precisions for all models are always greater than 90% and 50%, respectively. Both slightly increase when *duration* variable was not used, but as a tradeoff sensitivity decreases. Figure 7 furthers confirms this with AUC increases from 0.76 to 0.89 when variable *duration* was added to the logistic regression model.

Figure 8 shows that when *duration* variable was not used, the decision tree model only has 1 split using only *poutcome* variable, with the minimum cross-validation error of 0.91. The number of splits increases to 5 and the minimum cross-validation error decreases to 0.82 when *duration* was added to the model. The most predictive feature set based on gini criteria only uses variables *duration* and *poutcome*.

Table 1: Predictive results on the test set of logistic regression models and decision tree models

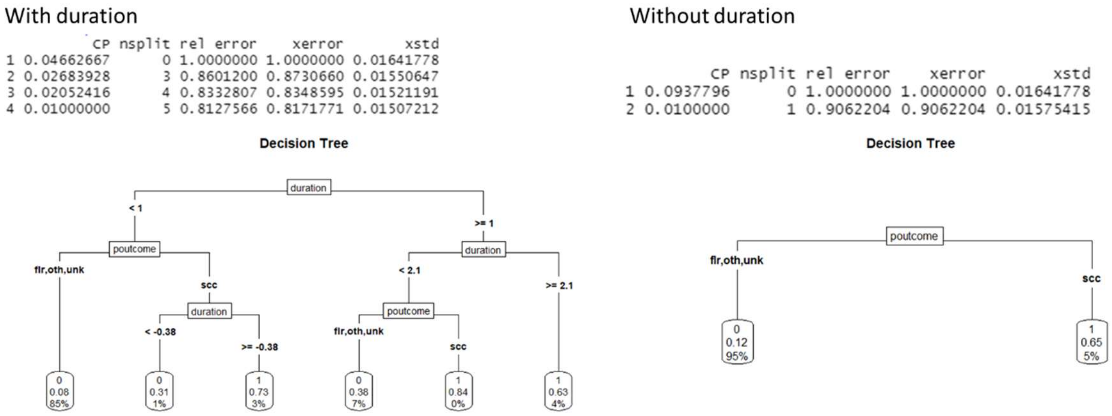
	Logistic Regression		Decision Tree	
	With duration	Without duration	With duration	Without duration
Accuracy	0.8803	0.8683	0.8558	0.8675
Misclassification-rate	0.1197	0.1317	0.1442	0.1325
Precision	0.6080	0.7049	0.5036	0.6418
Sensitivity	0.4955	0.1597	0.5238	0.1984
Specificity	0.9457	0.9886	0.9123	0.9812
F1 score	0.5460	0.2604	0.5135	0.3031



ROC Curve of Logistic Regression Model (with duration)

ROC Curve of Logistic Regression Model (without duration)

Fig.7. ROC curves on the test set of logistic regression models



With duration

```
          CP nsplit rel error    xerror       xstd
1 0.04662667      0 1.0000000 1.0000000 0.01641778
2 0.02683928      3 0.8601200 0.8730660 0.01550647
3 0.02052416      4 0.8332807 0.8348595 0.01521191
4 0.01000000      5 0.8127566 0.8171771 0.01507212
```

Decision Tree

Without duration

```
         CP nsplit rel error    xerror       xstd
1 0.0937796      0 1.0000000 1.0000000 0.01641778
2 0.0100000      1 0.9062204 0.9062204 0.01575415
```

Decision Tree

Fig.8. Cross-validation errors and plots of decision trees

6. Discussion

The diagnostic plots from figure 6 show that the model with *duration* performs better than the one without *duration*. The plot of residuals vs predicted values for the model with *duration* shows that the residuals are somewhat evenly distributed around zero, with a no-apparent-pattern of spread. This means the residuals of the model fairly fits the normal distribution, and we can also assume the error terms have a constant variance. The Q-Q plot of the model tells the same story with the residuals hugging closely to the theoretical quantiles in a straight diagonal line.

Table 1, figures 7 and 8 further confirm that both logistic regression and decision tree perform better when variable *duration* was used to predict the response. The logistic regression models always have lower misclassification rate and higher values for other metrics compared to the decision tree models. This indicates that logistic regression algorithm is more suitable in predicting the responses using this dataset.

I actually also did 10-fold cross-validation with both machine learning algorithms, but neither improve the prediction results. They produce the same confusion matrices as those with no cross-validation. Therefore, I decided to not talk about them in this paper, but you can find the code in the Appendix.

Since the author suggests removing variable *duration* to obtain a more realistic predictive model, we should focus on the results of the models with *duration*. The main problem is that our sensitivity, also known as the true positive rate, will decrease for about 30% when *duration* variable is removed from the model. Because sensitivity equals 1 – Type II error, a decrease in sensitivity results in an increase in Type II error [5]. However, since specificity, also known as true negative rate, increases and equals 1 – Type I error [5], Type I error will decrease when *duration* was not used in the model. Fortunately, "Type I errors are generally considered more serious than Type II errors" [6,7], and this is also true in our case. It is better to underestimate the success of the campaign due to predicting lower number of customers agreeing to place a term deposit (lower number of true positives) because most customers in reality do not subscribe and there will be no severe problems if the bank have more customers agreeing to subscribe than expected. On the other hand, overestimating the success of the campaign due to predicting lower number of true negatives may result in higher cost since the bank will spend more on such campaigns with low return rather than on other more profitable plans.

7. Conclusions

The logistic regression models perform better than decision tree models for this bank marketing dataset. Both machine learning algorithms show that variable *duration* is the most related to the response (customers will subscribe or not subscribe). The next significant variable is *poutcome* according to the decision tree models. However, it is more realistic to not use variable *duration*, so we should focus on the models without it. Even though the performance of the model slightly decreases when variable *duration* is not used, it is not a big problem as it only increases Type II error and not Type I error.

References:

1. Moro, S. (n.d.). UCI Machine Learning Repository: Bank Marketing Data Set. Retrieved November 29, 2022, from https://archive.ics.uci.edu/ml/datasets/bank+marketing
2. VolodymyrGavrysh. (2020, January 12). *Bank Marketing Campaigns Dataset: Opening deposit*. Kaggle. Retrieved November 29, 2022, from https://www.kaggle.com/datasets/volodymyrgavrysh/bank-marketing-campaigns-dataset
3. Prabhakaran, S. (n.d.). InformationValue R Package. Retrieved November 29, 2022, from http://r-statistics.co/Information-Value-With-R.html
4. Mohajon, J. (2021, July 24). *Confusion matrix for your multi-class machine learning model*. Medium. Retrieved November 29, 2022, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826#:~:text=Here%20are%20some%20of%20the,TN%2BFP%2BFN).
5. Some useful statistics definitions. (n.d.). Retrieved November 29, 2022, from https://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html
6. Diong, J. (2021, May 7). *Why type I errors are worse than type II errors*. Scientifically Sound. Retrieved November 29, 2022, from https://scientificallysound.org/2019/05/07/why-type-i-errors-are-worse-than-type-ii-errors/
7. Type I and II errors (2 of 2). (n.d.). Retrieved November 29, 2022, from https://davidmlane.com/hyperstat/A2917.html#:~:text=A%20conclusion%20is%20drawn%20that,is%20set%20by%20the%20experimenter.