Predicting Customers' Subscription to Term Deposit using Machine Learning – A Classification Approach

Chukwuekwu Musei

1.0 Introduction

1.1 Background

The banking sector has substantially improved the level of its services. One of such services is the term deposit, a primary source of instrument that the financial institutions use for stabilising its capital base (Zhuang et al., 2018,). As a result, banks analyse their customer's information using data mining approaches for important decision making strategies for their companies (Hung *et al*., 2019), particularly when faced with the pressures in the economy(Zhuang *et al*., 2018; Moro *et al*. 2011) and competitive marketing at low cost(Lau *et al*., 2004; Moro *et al.* 2011).

The purpose of this paper is to use logistic regression (LR) to predict customer subscription to term deposit using a bank's dataset, which could assist in future marketing potential. The paper has been organised in the following way. Session 2 provides the methodology used in carrying out the task. In session 3, the results and findings are interpreted and discussed. Session 4 concludes by highlighting the benefits and limitations of the model, while session 5 ends the paper by discussing the reflective commentary.

1.2 Related Work

The era of big data has made it possible for financial institutions to make sense of customer data (Chen *et al*. 2014). Moro *et al*. (2011) performed predictive analysis on a Portuguese's bank dataset to predict which customer will subscribe to a term deposit. They used three classification methods, Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM). It was observed that SVM model performed better than NB and DT, with a AUC of 90% and ALIFT at 50% respectively.

Miguéis *et al.* (2017) and Asare-Frempong and Jayabalan (2017) show the prediction of customer response to bank subscription using random forest (RF), which did better compared to LR, NN and SVM. Their work focused on the imbalance of the data class using synthetic minority oversampling technique and easy ensemble to balance the distribution (Wankhede *et al.*, 2019).

Wankhede *et al*. (2019) predicted customers who are likely to subscribe to term deposit on a Portuguese bank's dataset. They used four classification methods of LR, RF, SVM, and extreme gradient boost (XGBoost). Findings show that XGBoost model has highest AUC at 79% compared to SVM, RF and LR at 71%, 69% and 61% respectively, however, in terms of the test accuracy, RF performed better at 87.71%. .

This paper will use the LR model for prediction because it has the ability to fit models that human beings can comprehend for interpretation (Moro *et al*., 2014), particularly where the accuracy of a simple model like LR is near that of a complex model (Kuhn and Johnson, 2013). The choice of variables in this paper were based on extant literatures where similar analyses have been done with the same dataset using chi-square and information gain for feature selection (Parlar and Acaravci, 2017). Five hypotheses will be tested in this task, and they are as follows:

h1 Poutcome is positively related to subscription
h2 Month is positively related to subscription
h3 Pdays is positively related to Subscription
h4 Contact is positively related to Subscription
h5 Previous is positively related to Subscription

2.0 Methodology

The CRISP-DM structure in Figure 1 below was used to execute this task. During summary statistics of the bank dataset in R, there was a total of 41153 observations and 22 variables. The dataset had some data quality issues, such as outliers, missing data and some categorical data errors.

A combination of histogram and boxplot in ggplot2 aided to spot the appropriate range to subset outliers without allowing it to affect the entire analyses (see Appendix 2), including the NAs that were coded back to the dataset after been replaced with the value of the respective means.

Additionally, the measures of association of the target variable and the predictors, particularly the variables in the hypotheses were observed as a prelude to see how the predictor variables may likely influence the target variable. including the use ggplot2 to plot bivariate relationships (see Figures 5-8).

Multiple LR method was used to predict the customers that are likely to subscribe to the bank's term deposit. The cleaned dataset was partitioned into train and test data at 80% and 20% respectively. The test data was used to check for the accuracy of the predicted train dataset in order to evaluate its performance by safeguarding it from overfitting to produce optimally good models (Graham *et al.,* 2018). In the end, the best model built using the forward stepwise method (Field et al., 2012), was subjected through various assumption checks for validation.

3.0 Results and Discussions

3.1 Descriptive Statistics

Tables 1 and 2 show the summary statistics of the selected variables in the dataset before and after data cleaning. It can be seen in Table 1 that the *age* variable has a minimum of 4 years and maximum of 147 years. It was logical to exclude the outliers (118 and 147 years old) which were above 100 years in the *age* variable because (OECD, 2021) claim that the life expectancy of humans at birth is between 84 and 87 years old, including the age for employment which stands at 15-64 years old (OECD, 2022). Similarly, the 4 year old in the dataset was left because it is possible for a parent to open a term deposit for their children towards education savings, as banks now have products for children. On the other hand, *pdays* number of observation has increased from 41113 to 41153 after the 40 missing values were coded as mean.

TABLE 1: Descriptive statistics of selected unclean data

	vars	n	mean	sd	median	min	max	range
poutcome*	1	41153	1.93	0.36	2.00	1.00	3.00	2.00
month*	2	41153	5.70	2.76	5.00	1.00	11.00	10.00
pdays	3	41113	962.41	187.07	999.00	0.00	999.00	999.00
contact*	4	41153	1.36	0.48	1.00	1.00	2.00	1.00
previous	5	41153	0.17	0.50	0.00	0.00	7.00	7.00
default*	6	41153	2.20	0.42	2.00	1.00	4.00	3.00
job*	7	41153	4.72	3.59	3.00	1.00	12.00	11.00
day_of_week*	8	41153	3.01	1.40	3.00	1.00	5.00	4.00
cons.price.idx	9	41153	93.58	0.58	93.75	92.20	94.77	2.57
Cons.conf.idx	10	41153	-40.51	4.63	-41.80	-50.80	-26.90	23.90
euribor3m	11	41153	3.62	1.73	4.86	0.63	5.04	4.41
campaign	12	41153	2.57	2.77	2.00	1.00	56.00	55.00
age	13	41153	40.03	10.44	38.00	4.00	147.00	143.00

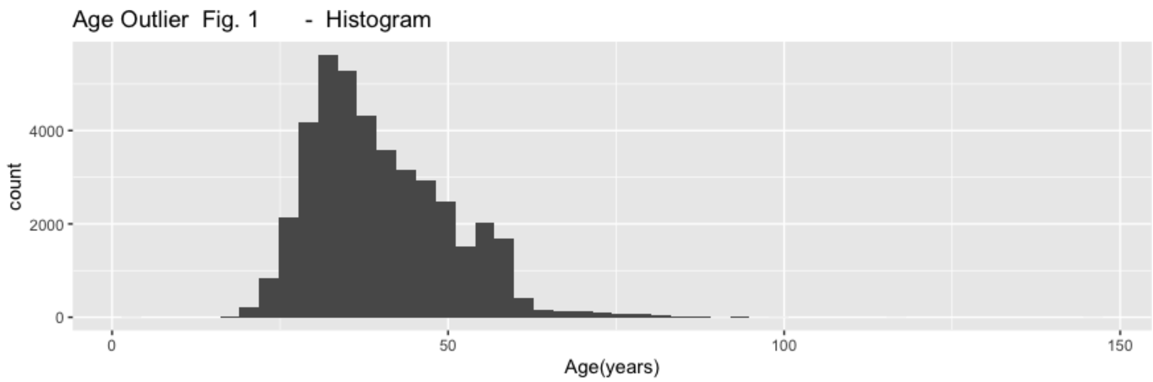
*categorical variables

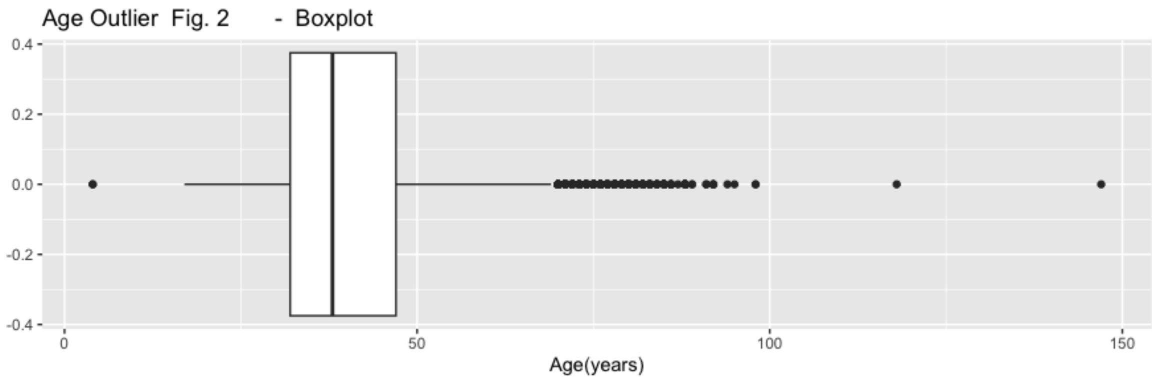
TABLE 2: Descriptive statistics of selected clean data

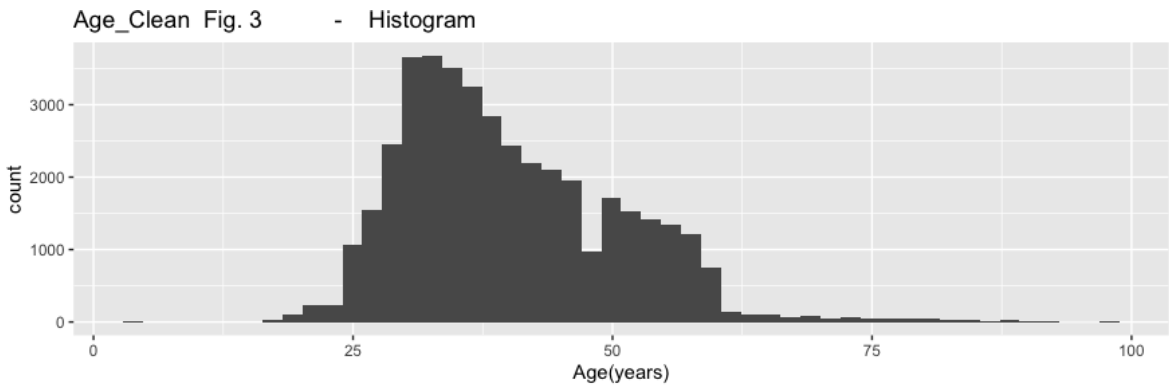
	vars	n	mean	sd	median	min	max	range
poutcome*	1	41153	1.93	0.36	2.00	1.00	3.00	2.00
month*	2	41153	5.70	2.32	5.00	1.00	10.00	9.00
pdays	3	41153	962.41	186.98	999.00	0.00	999.00	999.00
contact*	4	41153	1.36	0.48	1.00	1.00	2.00	1.00
previous	5	41153	0.17	0.50	0.00	0.00	7.00	7.00
default*	6	41153	1.21	0.41	1.00	1.00	3.00	2.00
job*	7	41153	4.72	3.59	3.00	1.00	12.00	11.00
day_of_week*	8	41153	3.01	1.40	3.00	1.00	5.00	4.00
cons.price.idx	9	41153	93.58	0.58	93.75	92.20	94.77	2.57
Cons.conf.idx	10	41153	-40.51	4.63	-41.80	-50.80	-26.90	23.90
euribor3m	11	41153	3.62	1.73	4.86	0.63	5.04	4.41
campaign	12	41153	2.57	2.77	2.00	1.00	56.00	55.00
age	13	41153	40.02	10.42	38.00	4.00	98.00	94.00

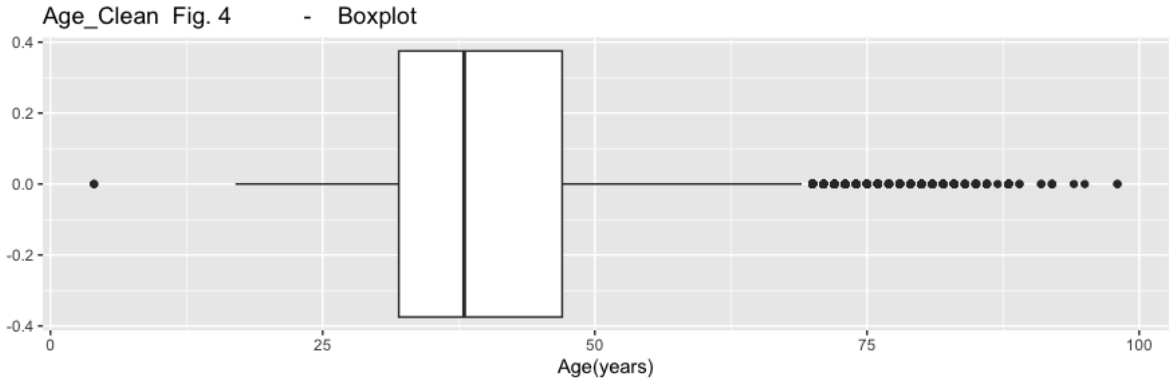
*categorical variables

3.2 Final Visualisations









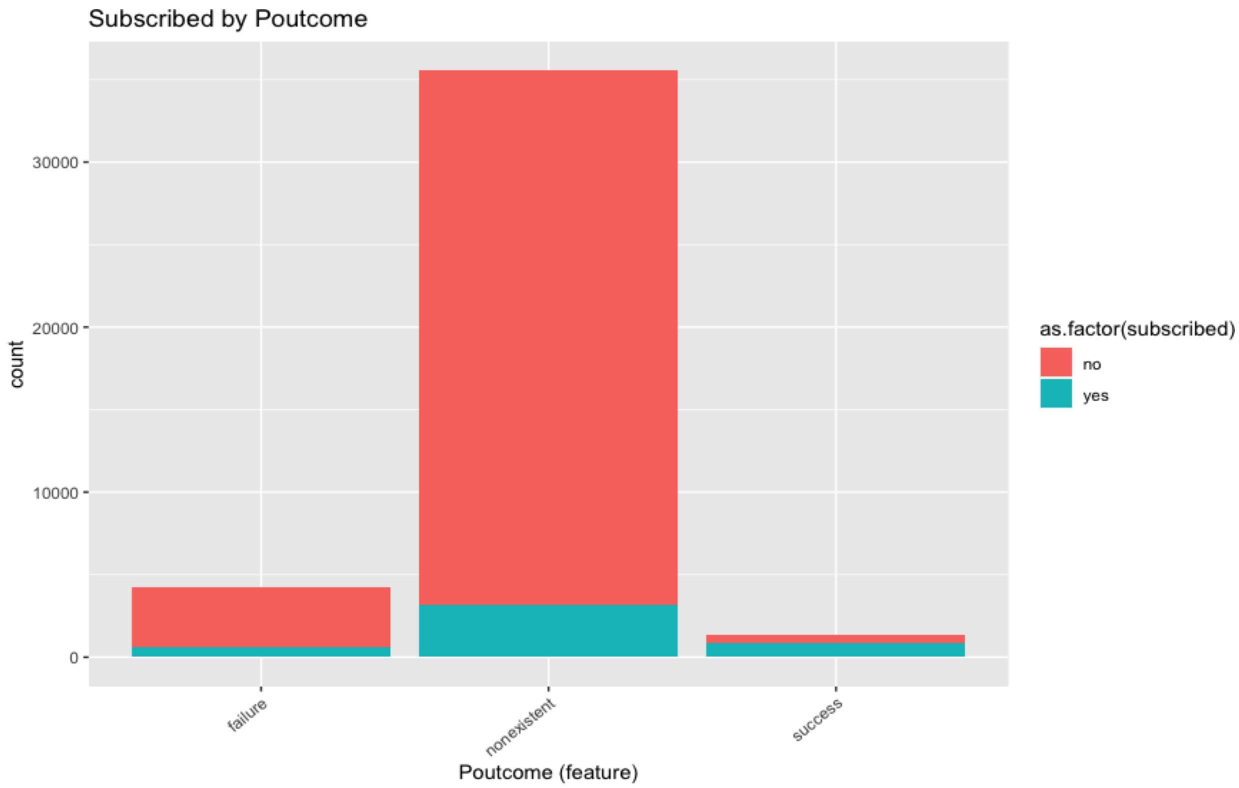


Fig. 5: Bivariate relationship between subscribed and poutcome

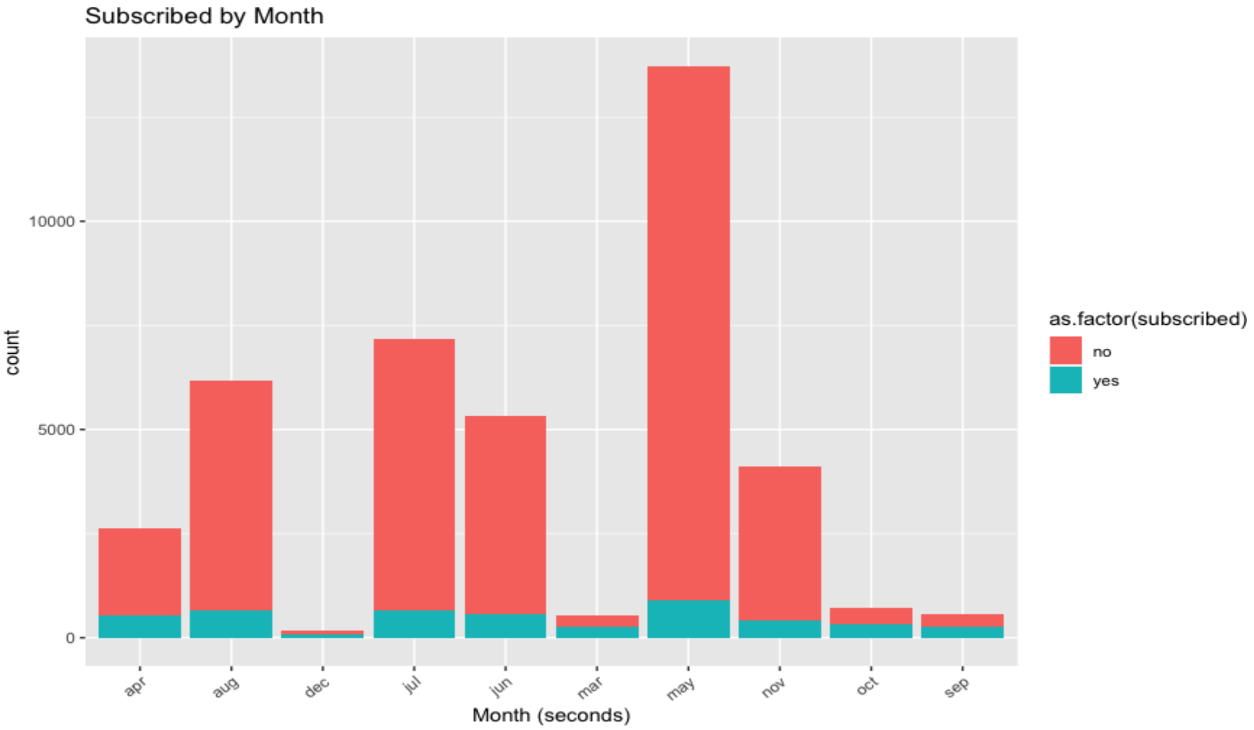


Fig. 6: Bivariate relationship between subscribed and month

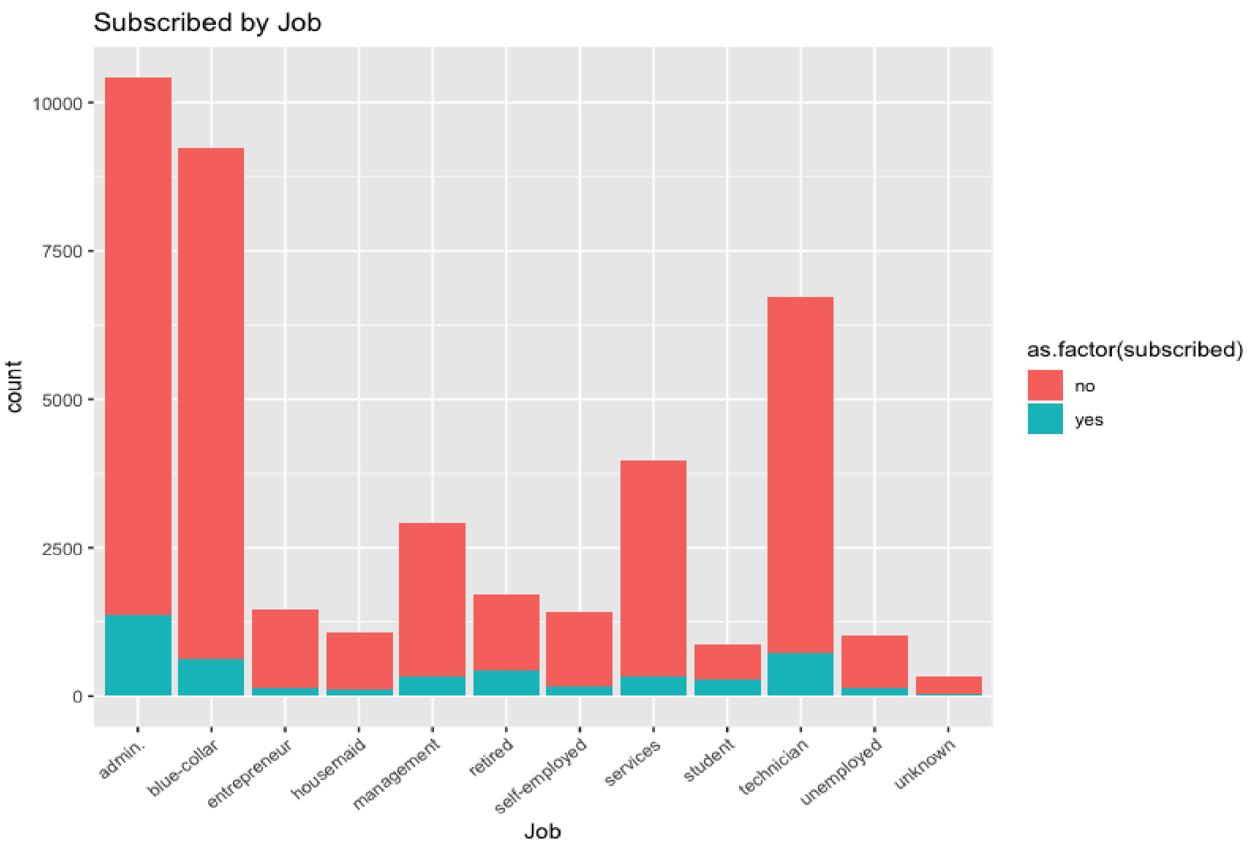


Fig. 7: Bivariate relationship between subscribed and job

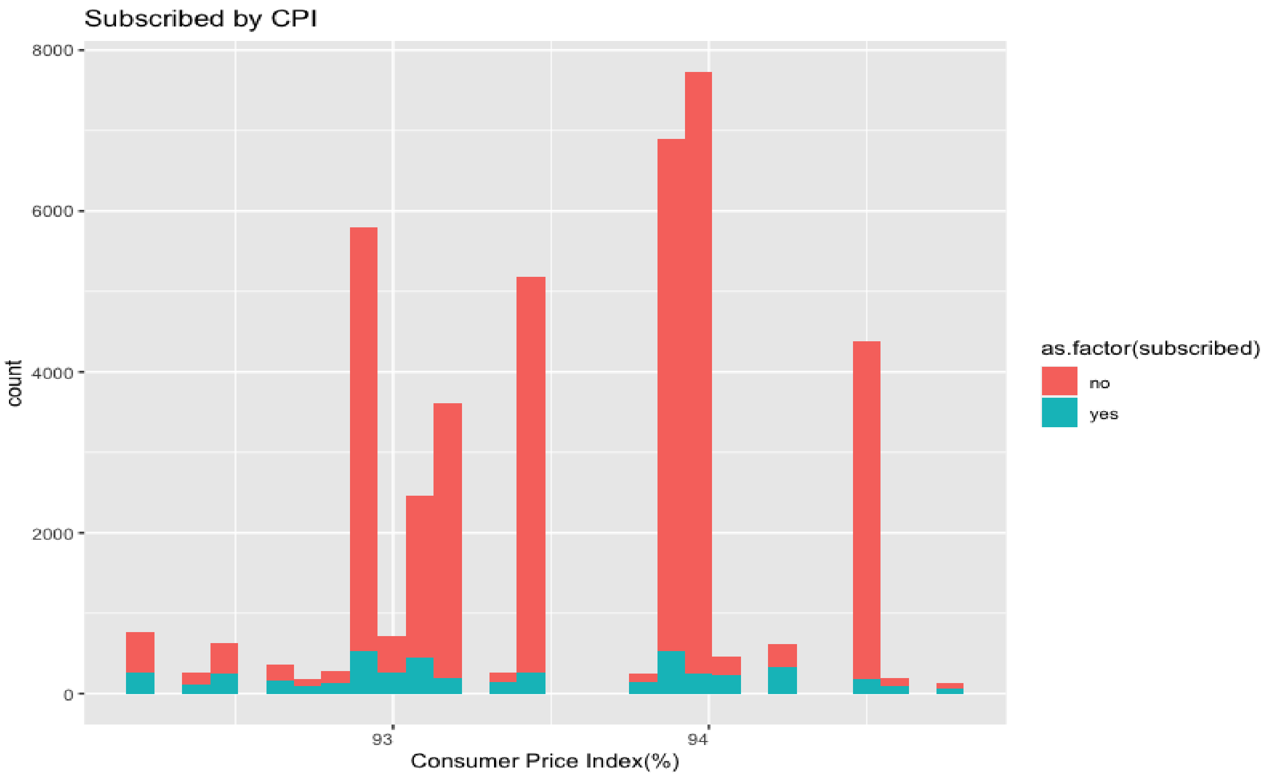


Fig. 8: Bivariate relationship between subscribed and consumer price index

3.3 Table of measures of association

Bivariate Relationship	P- value	Test
subscribed and poutcome	p-value < 2.2e-16	Chisq.test
subscribed and month	p-value < 2.2e-16	Chisq.test
subscribed and pdays	p-value < 2.2e-16	T.test
subscribed and contact	p-value < 2.2e-16	Chisq.test
subscribed and previous	p-value < 2.2e-16	T.test
subscribed and default	p-value < 2.2e-16	Chisq.test
subscribed and job	p-value < 2.2e-16	Chisq.test
subscribed and day of week	p-value < 4.675e-05	Chisq.test
subscribed and cons.price.idx	p-value < 2.2e-16	T.test
subscribed and cons.conf.idx	p-value < 2.2e-16	T.test
subscribed and euribor3m	p-value < 2.2e-16	T.test
subscribed and campaign	p-value < 2.2e-16	T.test

The results of the bivariate measures of association indicate that all predictors are correlated to the target variable with p-value < 0.05.

3.4 Table of multiple logistic regression models

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H	Model I
AIC	19714	19609	19496	19462	19388	19316	18369	18345	18346
Null Deviance	23192	23192	23192	23192	23192	23192	23192	23192	23192
Residual Deviance	19684	19575	19440	19398	19322	19248	18299	18273	18272

Table 3.4 show the results of the multiple logistic regression models. There was a consistent drop in the value of AIC from models A-H, which indicates model improvement (Field et al., 2012), however, there was an increase in the value of AIC from model H to model I, thus, the reason for dropping the model (see Appendix 3).

3.5 Final multiple logistic regression model (Model H)

```
> summary(modelH)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1657  -0.4025  -0.3279  -0.2447   3.0402

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -4.522e+01  4.401e+00 -10.276  < 2e-16 ***
poutcomenonexistent  4.835e-01  9.472e-02   5.105 3.31e-07 ***
poutcomesuccess      6.728e-01  2.183e-01   3.083  0.00205 **
monthaug            -1.049e-01  1.007e-01  -1.042  0.29746
monthdec             4.368e-01  2.001e-01   2.183  0.02904 *
monthjul             2.053e-01  9.186e-02   2.235  0.02539 *
monthjun             1.292e-01  8.956e-02   1.443  0.14914
monthmar             9.304e-01  1.213e-01   7.673 1.69e-14 ***
monthmay            -5.987e-01  7.245e-02  -8.263  < 2e-16 ***
monthnov            -1.083e-01  9.663e-02  -1.120  0.26255
monthoct             2.152e-01  1.225e-01   1.757  0.07892 .
monthsep            -1.960e-01  1.323e-01  -1.482  0.13842
pdays               -1.131e-03  2.233e-04  -5.065 4.09e-07 ***
contacttelephone    -5.362e-01  6.771e-02  -7.919 2.39e-15 ***
previous            -4.058e-02  6.092e-02  -0.666  0.50535
defaultunknown      -2.672e-01  6.355e-02  -4.205 2.61e-05 ***
defaultyes          -7.634e+00  8.375e+01  -0.091  0.92737
```

```
jobblue-collar      -1.812e-01  6.272e-02  -2.889  0.00387 **
jobentrepreneur     -9.941e-02  1.192e-01  -0.834  0.40412
jobhousemaid        -1.056e-01  1.360e-01  -0.777  0.43737
jobmanagement        2.248e-02  8.130e-02   0.276  0.78217
jobretired           1.810e-01  8.417e-02   2.150  0.03154 *
jobself-employed    -7.474e-02  1.142e-01  -0.655  0.51263
jobservices         -1.478e-01  7.906e-02  -1.870  0.06153 .
jobstudent           2.177e-01  1.017e-01   2.140  0.03234 *
jobtechnician       -2.561e-03  6.148e-02  -0.042  0.96677
jobunemployed       -1.105e-01  1.262e-01  -0.875  0.38136
jobunknown          -1.997e-01  2.393e-01  -0.835  0.40383
day_of_weekmon      -2.635e-01  6.474e-02  -4.070 4.70e-05 ***
day_of_weekthu       6.046e-02  6.173e-02   0.979  0.32738
day_of_weektue       3.837e-02  6.371e-02   0.602  0.54702
day_of_weekwed       1.519e-01  6.310e-02   2.406  0.01611 *
cons.price.idx       5.122e-01  4.835e-02  10.595  < 2e-16 ***
cons.conf.idx        4.724e-02  5.119e-03   9.227  < 2e-16 ***
euribor3m           -5.680e-01  1.789e-02 -31.741  < 2e-16 ***
campaign            -5.009e-02  1.032e-02  -4.854 1.21e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18273  on 32887  degrees of freedom
AIC: 18345

Number of Fisher Scoring iterations: 9
```

Poutcome: It can be seen that *poutcome (nonexistence) and poutcome(success)* are statistically significant and have a positive relationship to the probability to getting subscription.

Month: The months of December, July, and March are statistically significant positively related to subcribed. Compared to the month of April, customers are more likely to get subscribed in the month of March than April. Similarly, moving from April to May, customers are less likely to get subscribed in the month of May.

Pdays: Pdays is statistically significant and it is negatively related to the possibility for customer to getting subscribed. However as pdays increases the likelihood of customers getting subscribed to term deposit decreases.

Contact: Table 3.5 show that contact is statistically significant and has a negative relationship with the possibility of customers getting subscribed to term deposit. It can be seen that moving from cellular to telephone, customers are less likely to get subscribed to telephone than cellular.

Default: The variable *default* is statistically significant and negatively related to the possibility of customers to getting subscribed.

Job: Job(blue collar), job(retired) and Job(student) are statistically significant and positive related to subscribed but negatively related to blue collar jobs. Compared to admin, job(blue collar) is less likely to getting subscribed to term deposit than admin, while compared to admin, Job(retired) and Job(student) are more likely to getting subscribed than admin.

Days of week: Days of the week on Monday and Wednesday are statistically significant. Day of the week(mon) is negatively related to subscribed while day of the week(wed) has a positive relationship. Compared to day of the week(fri), customers are less likely to getting subscribed on day of the week(mon) but more likely to getting subscribed on day of the week(wed).

Consumer Price Index (CPI) and Consumer Confidence Index (CCI): CPI and CCI variables are statistically significant and positively related to the possibility of customers to getting subscribed.

Campaign: Campaign variable is statistically significant and is negatively related to subscribed.

3.6 Odds ratio

```
> exp(modelH$coefficients)
      (Intercept) poutcomenonexistent     poutcomesuccess           monthaug            monthdec
     2.293916e-20        1.621818e+00        1.959660e+00       9.004300e-01        1.547811e+00
         monthjul            monthjun            monthmar           monthmay            monthnov
     1.227935e+00        1.137919e+00        2.535398e+00       5.495354e-01        8.973955e-01
         monthoct            monthsep               pdays   contacttelephone            previous
     1.240091e+00        8.220482e-01        9.988694e-01       5.849536e-01        9.602335e-01
   defaultunknown          defaultyes      jobblue-collar    jobentrepreneur        jobhousemaid
     7.655104e-01        4.837735e-04        8.342722e-01       9.053707e-01        8.997399e-01
    jobmanagement          jobretired    jobself-employed        jobservices          jobstudent
     1.022732e+00        1.198396e+00        9.279812e-01       8.625813e-01        1.243184e+00
    jobtechnician       jobunemployed          jobunknown    day_of_weekmon      day_of_weekthu
     9.974421e-01        8.954145e-01        8.189407e-01       7.683782e-01        1.062322e+00
   day_of_weektue      day_of_weekwed      cons.price.idx      cons.conf.idx           euribor3m
     1.039115e+00        1.163993e+00        1.668964e+00       1.048370e+00        5.666701e-01
         campaign
     9.511416e-01
```

Poutcome: Table 3.6 show that moving from failure to success, the odds of marketing campaign success is more likely to getting subscribed customers for term deposit at 1.9596 times higher than those of the failure outcomes. It can be argued that positive response to marketing campaigns by customers is a plus to any business

Month: In Table 3.6, the odds of customers to getting subscribed from April to March is 2.53554. Also compared to the month of April, the month of May have less chances of customer subscription at 0.549 odds. Also, one unit increase in reaching out to customers in the month of December increase subscription by 1.5478 odds. According to Moro *et al.* (2011) the months in the last trimester such as March and December receive higher subscriptions, while the month of May recorded one of the lowest.

Pdays: As pdays increases by one unit, the odds of subscribing to term deposit decreases by 0.998.

Contact: Compared to cellular, telephone have less chances of getting subscription at 0.584 odds. It can be claimed that customers who use cellular move around with it compared to telephone that is stationary at home, leading to missing calls in an attempt to reach customers.

Default: Compared to ‘no’, ‘unknown’ customers are less likely to getting subscribed at 0.765 odds.

Job: Compared to admin in Table 3.6, job(blue collar) is less likely to getting subscribed to term deposit at 0.834 odds. Also, compared to admin, Job(retired) and Job(student) are more likely to getting subscribed at 1.198 and 1.243 odds respectively.

Days of week: Compared to day of the week(fri), customers are less likely to getting subscribed on day of the week(mon) at 0.768 odd but more likely to get subscription on day of the week(wed) at 1.164 odds. It can be claimed that engaging in marketing campaign first working day of the week when people are scheduling work plan for the week will not yield as much compared to when it is done in the middle of the week.

Consumer Price Index (CPI) and Consumer Confidence Index (CCI): As CPI and CCI variables increases by one unit of price, the odds of subscription increases by 1.669 and 1.05 respectively. CPI assesses the rate of inflation (IMF, 2023), while CCI measure future economic assumptions (OECD, 2022). This is in line with the logistic regression analysis carried out by Wankhede *et al.* (2019) and Miguéis *et al.* (2017) which indicate that social economic indices are dependent on whether customers are likely to subscribe to term deposits or not.

Campaign: It can be seen in Table 3.6 that one unit increase in *campaign* variable results to a 0.951 odds decease is subscription.

3.7 Table of model accuracy using postResample

	Model A	**Model B**	**Model C**	**Model D**	**Model E**	**Model F**	**Model G**	**Model H**
Accuracy	0.8973	0.8982	0.8974	0.8974	0.8981	0.8974	0.9006	0.9002
Kappa	0.2123	0.2522	0.2485	0.2485	0.2518	0.2627	0.3061	0.3031

Table 3.7 show the values of the model accuracies. Model H with 90% accuracy close to (Desai and Khairnar, 2022) 91.02% accuracy in a similar logistic regression analysis conducted. The confusion matrix below give a detailed result which indicate that 215 customers have been predicted correctly to subscribe to term deposit and as such, more targeted campaign should be directed to them so that they do not stop subscribing.

```
> confusionMatrix(data = class_pred, test$subscribed)
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  7194  713
       yes  108  215

               Accuracy : 0.9002
                 95% CI : (0.8936, 0.9066)
    No Information Rate : 0.8872
    P-Value [Acc > NIR] : 8.09e-05

                  Kappa : 0.3031

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9852
            Specificity : 0.2317
         Pos Pred Value : 0.9098
         Neg Pred Value : 0.6656
             Prevalence : 0.8872
         Detection Rate : 0.8741
   Detection Prevalence : 0.9608
      Balanced Accuracy : 0.6084

       'Positive' Class : no
```

Confusion matrix showing model accuracy

3.8 Pseudo R2s

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Hosmer and Lemeshow	0.151	0.156	0.162	0.164	0.167	0.17	0.211	0.212
Cox and Snell	0.101	0.104	0.108	0.109	0.111	0.113	0.138	0.139
Nagelkerke	0.2	0.206	0.213	0.215	0.219	0.223	0.273	0.275

Table 3.8 show 3 types of R-squared. It can be seen that the values are increasing from models A-H, hence it is a good model.

3.9 Assumption Checks and things that could go wrong

3.9.1 Analysing the residuals

The sum of the standardised residuals was calculated as 1464 data > 1.96 which is < 5% of 32923 train data = 1646.15, thus, it satisfies the assumption.

3.9.2 Check for examining influential cases using cook distance

Findings show sum of cooks distance not > 1.

3.9.3 Multicollinearity Check

```
> vif(modelH) ### variables poutcome and pdays have multicolinearity issues since they are > 10
                   GVIF Df GVIF^(1/(2*Df))
poutcome      24.888905  2        2.233580
month          5.144236  9        1.095262
pdays         10.890733  1        3.300111
contact        1.896261  1        1.377048
previous       4.561378  1        2.135738
default        1.100759  2        1.024290
job            1.237484 11        1.009733
day_of_week    1.043148  4        1.005294
cons.price.idx 2.538718  1        1.593335
cons.conf.idx  2.310424  1        1.520008
euribor3m      2.770449  1        1.664467
campaign       1.038084  1        1.018864
```

Result show that *poutcome* and *pdays* failed the multicollinearity test. Efforts to hold each of the variables at constant at different intervals did not improve the model (see Appendix 5).

3.9.4 Testing for the Linearity of the Logit

The model built to test for the linearity of the logit did not pass the test since one of the variables is significant (see Appendix 6).

3.9.5 Durbin-Wanton Test(dwtest)

DW= 1.9. A value between 1.5 and 2.5 is good, meaning assumption satisfied.

```
> dwtest(modelH) ### DW = 1.9 This falls in between 1.5 and 2.5

        Durbin-Watson test

data:  modelH
DW = 1.8891, p-value < 2.2e-16
```

4.0 Conclusions

This paper was developed to predict the number of bank customers who are likely to subscribe to term deposit using LR. Based on related works, an appropriate hypothesis was set to investigate the customer variables in the dataset that will influence the target variable. Findings show that both customer profile and social-economic indices have a role in influencing whether customers will subscribe to term deposit. Despite some limitation in class imbalance, the model will help financial institution to target telemarketing campaigns to prospective customers who are likely to subscribe to term deposit rather than spend so much on general promotion, hence, significantly increasing cost savings in the competitive financial sector.

5.0 Reflective Commentary

Performing technical part of data analysis tasks on R software package as well as writing a concise report has now become natural. This has improved my confidence level on what to do whenever a business problem needs solution, particularly how LR is used to solve classification problems prevalent in the society. The knowledge garnered so far will aid as foundation to learning more advanced analytics techniques in the second semester.

Conclusion

In this paper, through the online housing platform to grab Chengdu housing rental data as a data set, through the processing of abnormal data and missing data, 33111 pieces of data were visualized and analyzed, and 12 characteristic factors needed to build the prediction model were obtained. It is found that there is more market for houses less than 90 ㎡ .And 96% of the rental area is less than 30 ㎡ in the mode of joint rental, while 70% of the total rental area is less than 90 ㎡ . From the above analysis, it can be seen that Chengdu tenants have a greater demand for joint rental and small area houses. The houses are mainly located in the area within the Third Ring Road and along the Metro Line 1 in Tianfu new area. For rent prediction, this paper puts the training set data into three models: RandomForestRegressor,XG**B**oost and LightGBM. It is found that XG**B**oost model has better prediction effect than the other two models, which makes a good contribution to the research of housing rent prediction in Chengdu.

(Udoekanem et al. 2014; Udoekanem et al., 2015).

References

Chen, K., Hu, Y. H. & Hsieh, Y. C. (2015) ‘Predicting customer churn from valuable B2B customers in the logistics industry: a case study’, *Inf Syst E-Bus Manage*, 13, pp. 475–494.Available at: https://doi.org/10.1007/s10257-014-0264-1.

Desai, R., & Khairnar, V. (2022) ‘Hybrid prediction model for the success of bank telemarketing. in: Raj, J.S., Palanisamy, R., Perikos, I., Shi, Y. (eds) Intelligent Sustainable Systems. Lecture Notes in Networks and Systems, vol 213. Springer, Singapore. https://doi.org/10.1007/978-981-16-2422-3_54.

Graham, B., Bond, R., Quinn, M., & Mulvenna, M. (2018) ‘Using data mining to predict hospital admissions from the emergency departmen*t. IEEE Access, 6, 10458–10469.* Available at: https://doi/10.1109/access.2018.2808843. (Accessed 28 December 2022)

Field, A., Miles, J. and Field Zoë (2012) *Discovering statistics using R.* Thousand Oaks: SAGE/Texts.

Gupta, A., Raghav, A. and Srivastava, S. (2021) ‘Comparative Study of Machine Learning Algorithms for Portuguese Bank Data’, *International Conference on Computing, Communication, and Intelligent Systems*, pp. 401-406, Available at: https://doi/10.1109/ICCCIS51004.2021.9397083.

Guyon, I. & Elisseeff, A. (2003) ‘An introduction to variable and feature selection’ *Journal of Machine Learning Research*, 3, pp. 1157-1182.

Hung, P. D., Hanh, T. D. & Tung, T. D. (2019) ‘Term deposit subscription prediction using Spark MLlib and ML packages*’, Proceedings of the 5th International Conference on E-Business and Applications.Available at:* https://doi/10.1145/3317614.3317618 (Accessed: 30 December 2022).

IMF (2023) *Consumer price index* [Online] Available at: https://data.imf.org/?sk=4FFB52B2-3653-409A-B471-D47B46D904B5 (Accessed: 2 January 2023).

Kuhn, M. and Johnson, K (2013) *Applied Predictive Modelling*. London, U.K.: Springer.

Lau, Kn., Chow, H. & Liu, C. (2004) 'A database approach to cross selling in the banking industry: practices, strategies and challenges', *J Database Mark Cust Strategy Manag,* 11, pp. 216–234. Available at: https://doi.org/10.1057/palgrave.dbm.3240222.

Lu, X. Y., Chu, X. Q., Chen, M. H., Chang, P. C. & Chen, S. H. (2016) 'Artificial immune network with feature selection for bank term deposit recommendation', *Journal of Intelligent Information Systems, 47(2), 267–285.* Available at: https://doi/10.1007/s10844-016-0399-2.

Miguéis, V.L., Camanho, A.S. & Borges, J. (2017) 'Predicting direct marketing response in banking: comparison of class imbalance methods', *Serv Bus*, 11, pp.831–849 https://doi.org/10.1007/s11628-016-0332-3.

Moro, S., Laureano, R. M.S. & Cortez, P. (2011) 'Using data mining for bank direct marketing: an application of the crisp-dm methodology' [Online]. Available at: MoroCortezLaureano_DMApproach4DirectMKT.pdf.

Moro, S., Cortez, P., & Rita, P. (2014). 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems, 62, 22–31.*Available at: https://doi/10.1016/j.dss.2014.03.001.

OECD (2021) Life expectancy at birth. Available at: *https://data.oecd.org/healthstat/life-expectancy- at-birth.htm#indicator-chart* (Accessed: 19 November 2022).

OECD (2022) Labour force statistics: summary tables, *OECD Employment and Labour Market Statistics* (database), https://doi.org/10.1787/data-00286-en (Accessed:19 November 2022).

OECD (2022) *Consumer confidence index* [Online] Available at: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm Accessed 5 December 2023

Parlar, T. & Acaravcı, S. K. (2017) 'Using data mining techniques for detecting the important features of the bank direct marketing data', *International Journal of Economics and Financial Issues,* 7(2), pp. 692-696.

Wankhede, P., Singh, R., Rathod, R., Patil, J. & Khadtare T.D. (2019) 'Improving prediction of potential clients for bank term deposits using machine learning approaches', [Online]. Available at: IRJET-V6I5100520190814-54045-zhs1m6-libre.pdf (d1wqtxts1xzle7.cloudfront.net).

Zhuang, Q. R., Yao, Y. W. and Liu, O. (2018) 'Application of data mining in term deposit marketing', *Proceedings of the International MultiConference of Engineers and Computer* [Online]. Available at: (iaeng.org) (Accessed: 30 December 2022).

Appendix 1: R Code

```
### Set working directory ##
setwd("/Volumes/GoogleDrive/My Drive/Documents/MGT_7177/ST-Assignment 2")


### load in libraries
library(tidyverse)
library(readxl)
library(psych)
library(gridExtra)
library(factoextra)
library(dplyr)
library(vtable)
library(car)
library(caret)
library(lmtest)


### load in data
musei1 <- read_excel("banksv.xlsx")



                    ### UNDERSTANDING BANK SAVINGS DATASET


### Understanding data, looking for outliers, missing values
glimpse(musei1)

### looking at the first 10 rows of the dataset as well as the last 7 rows of the observations
head(musei1, 10)
tail(musei1, 7)
names(musei1) ### observing variables in the banksv dataset

### summarise data
summary(musei1)


## understanding the selected variables for the analyses
SH1 <- c('poutcome','month','pdays', 'contact', 'previous', 'default','job',
'day_of_week','cons.price.idx','cons.conf.idx','euribor3m','campaign','age')
summary(musei1[SH1])

## Checking summary statistics of the unclean selected variables above
describe(musei1[SH1])
```

DATA QUALITY AND FORMATTING ISSUES

```
### Checking for outliers in Age using ggplot  ###

histo1<- ggplot(musei1) +
  geom_histogram(aes(age), bins = 50) +
  labs(title = "Age Outlier  Fig. 1       - Histogram", x= "Age(years)")

boxplt1<- ggplot(musei1) +
  geom_boxplot(aes(age))+
  labs(title = "Age Outlier  Fig. 2       - Boxplot", x= "Age(years)")

### combining the above visualisation histo1 and boxplt1
grid.arrange(histo1, boxplt1)

### subset to remove outliers in Sale.Price ###
histo2 <- ggplot(musei1[musei1$age < 100,]) +
geom_histogram(aes(age), bins = 50) +
  labs(title = "Age_Clean  Fig. 3          - Histogram", x= "Age(years)")

boxplt2 <- ggplot(musei1[musei1$age< 100,]) +
  geom_boxplot(aes(age)) +
  labs(title = "Age_Clean  Fig. 4          -    Boxplot", x= "Age(years)")

### combining the above visualisation histo2 and boxplt2
grid.arrange(histo2, boxplt2)

### Assigning age outlier as NA
musei1$age[musei1$age > 100] <- NA

### summarise Age
summary(musei1$age) ### 2 NAs observed

### Remove 2 NAs from age outlier and replace with mean value
musei1$age[is.na(musei1$age)] <- mean(musei1$age, na.rm=TRUE)
summary(musei1$age)

### Checking for DQ issues in pdays ###
summary(musei1$pdays) ### 40 NAs observed


### Remove 40 NAs from pdays variable and replace with mean value
musei1$pdays[is.na(musei1$pdays)] <- mean(musei1$pdays, na.rm=TRUE)
summary(musei1$pdays)


###Converting character variables to as.factor in preparation for analyses
musei1 <- musei1 %>% mutate_if(is.character, as.factor)

### Checking for DQ issues in defaults  ###
```

```
summary(musei1$default) ### error in categorical variable

### Convert n to no
musei1$default[musei1$default == "n"] <- "no"
summary(musei1$default) ### n remains with a value of 0

### Remove n with a value of zero using droplevel function
musei1$default<- droplevels(musei1$default)
summary(musei1$default) ### error tectified in default variable


### Checking for DQ issues in month ###
summary(musei1$month) ### error in categorical variable

### Convert march to mar
musei1$month[musei1$month == "march"] <- "mar"
summary(musei1$month) ### march remains with a value of 0

### Remove march with a value of zero using droplevel function
musei1$month<- droplevels(musei1$month)
summary(musei1$month) ### error rectified in default variable


### Summary of data after fixing data quality issues

### summarise data
summary(musei1)


## checking out the descriptive statistics of the selected variables after cleaning
SH1 <- c('poutcome','month','pdays', 'contact', 'previous', 'default','job',
'day_of_week','cons.price.idx','cons.conf.idx','euribor3m','campaign','age')
summary(musei1[SH1])

## Creating summary statistics of the clean selected variables above
describe(musei1[SH1])



          ## HYPOTHESES ##

## h1 Poutcome is positively related to subscription
## h2 Month is positively related to subscription
## h3 Pdays is positively related to Subscription
## h4 Contact is positively related to Subscription
## h5 Previous is positively related to Subscription
```

VISUALISATION

```
## subscribed by duration using ggplot2
ggplot(musei1, mapping=aes(x=poutcome)) +
geom_bar() +
  labs(title="Subscribed by Poutcome",x="Poutcome (feature)", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x= poutcome,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title="Subscribed by Poutcome",x="Poutcome (feature)")


## subscribed by Month using ggplot2
ggplot(musei1, mapping=aes(x=month)) +
geom_bar(bins=50) +
  labs(title="Subscribed by Month",x="Month(seconds)", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x= month,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title="Subscribed by Month",x="Month (seconds)")


## subscribed by contact using ggplot2
ggplot(musei1, mapping=aes(x=contact)) +
geom_bar(bins=100) +
  labs(title="Subscribed by Contact",x="Contact(feature)", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x= contact,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title="Subscribed by Contact",x="Contact (feature)")



## subscribed by previous using ggplot2
ggplot(musei1, mapping=aes(x=previous)) +
geom_bar(bins=20) +
  labs(title="Subscribed by Previous",x="Previous (No. of contacts)", y="Subscribed (no/yes)") +
```

```
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x= previous,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 360,hjust = 1)) +
  labs(title="Subscribed by Previous",x="Previous (No. of contacts)")




## subscribed by default using ggplot2
ggplot(musei1, mapping=aes(x=default)) +
geom_bar(bins=50) +
  labs(title="Subscribed by Default",x="Default", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x= default,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title="Subscribed by Default",x="Default")




## subscribed by job using ggplot2
ggplot(musei1, mapping=aes(x=job)) +
geom_bar(bins=50) +
  labs(title="Subscribed by Job",x="Job", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x=job,
           fill = as.factor(subscribed))) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 40,hjust = 1)) +
  labs(title="Subscribed by Job",x="Job")




## subscribed by campaign using ggplot2
ggplot(musei1, mapping=aes(x=campaign)) +
geom_bar(bins=50) +
  labs(title="Subscribed by Campaign",x="Campaign(No. of contacts)", y="Subscribed (no/yes)")
+
  facet_wrap(~subscribed)
```

```
ggplot(musei1,
       aes(x= campaign,
           fill = as.factor(subscribed))) +
  geom_histogram(position = "stack") +
  theme(axis.text.x = element_text(angle = 360,hjust = 1)) +
  labs(title="Subscribed by campaign",x="Campaign (No. of contacts)")




## subscribed by cons.price.idx using ggplot2
ggplot(musei1, mapping=aes(x=cons.price.idx)) +
geom_histogram(bins=20) +
  labs(title="Subscribed by Consumer Price Index",x="Consumer Price Index(%)", y="Subscribed
(no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x=cons.price.idx,
           fill = as.factor(subscribed))) +
  geom_histogram(position = "stack") +
  theme(axis.text.x = element_text(angle = 360,hjust = 1)) +
  labs(title="Subscribed by CPI",x="Consumer Price Index(%)")




## subscribed by age using ggplot2
ggplot(musei1, mapping=aes(x=age)) +
geom_histogram(bins= 40) +
  labs(title="Subscribed by Age",x="Age(Years)", y="Subscribed (no/yes)") +
  facet_wrap(~subscribed)

ggplot(musei1,
       aes(x=age,
           fill = as.factor(subscribed))) +
  geom_histogram(position = "stack") +
  theme(axis.text.x = element_text(angle = 360,hjust = 1)) +
  labs(title="Subscribed by Age",x="Age(years)")
```

```
## MEASURES OF ASSOCIATION (MA)

## subscribed and poutcome
chisq.test(musei1$poutcome, musei1$subscribed)

## subscribed and month
chisq.test(musei1$month, musei1$subscribed)

## subscribed and pdays
t.test(musei1$pdays, as.numeric(musei1$subscribed))

## subscribed and previous
t.test(musei1$previous, as.numeric(musei1$subscribed))

## subscribed and default
chisq.test(musei1$default, musei1$subscribed)

## subscribed and job
chisq.test(musei1$job, musei1$subscribed)

## subscribed and day of week
chisq.test(musei1$day_of_week, musei1$subscribed)

## subscribed and cons.price.idx
t.test(musei1$cons.price.idx, as.numeric(musei1$subscribed))

## subscribed and cons.conf.idx
t.test(musei1$cons.conf.idx, as.numeric(musei1$subscribed))

## subscribed and euribor3m
t.test(musei1$euribor3m, as.numeric(musei1$subscribed))

## subscribed and contact
chisq.test(musei1$contact, musei1$subscribed)

## subscribed and age
chisq.test(musei1$age, as.numeric(musei1$subscribed))
```

```
## MULTIPLE LOGISTIC REGRESSION

### Check for missing values on selected data to avoid NA during check for model accuracy on
test data (20%)
musei1 <- musei1[!is.na(musei1$poutcome),]
musei1 <- musei1[!is.na(musei1$month),]
musei1 <- musei1[!is.na(musei1$pdays),]
musei1 <- musei1[!is.na(musei1$contact),]
musei1 <- musei1[!is.na(musei1$previous),]
musei1 <- musei1[!is.na(musei1$default),]
musei1 <- musei1[!is.na(musei1$job),]
musei1 <- musei1[!is.na(musei1$day_of_week),]
musei1 <- musei1[!is.na(musei1$cons.price.idx),]
musei1 <- musei1[!is.na(musei1$cons.conf.idx),]
musei1 <- musei1[!is.na(musei1$euribor3m),]
musei1 <- musei1[!is.na(musei1$campaign),]

### Set seed to keep values of regression constant
set.seed(1846)
index <- createDataPartition(musei1$subscribed, times =1, p =0.8, list= FALSE)

train <- musei1[index,]
test <- musei1[-index,]



### model A
formula <- subscribed ~ poutcome + month + pdays + contact + previous ### Setting the formula
modelA <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelA)


### model B
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default ### Setting the
formula
modelB <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelB)


### model C
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job ###
Setting the formula
modelC <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelC)
```

```
### model D
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job +
day_of_week### Setting the formula
modelD <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelD)


### model E
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job +
day_of_week + cons.price.idx ### Setting the formula
modelE <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelE)


### model F
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job +
day_of_week + cons.price.idx + cons.conf.idx ### Setting the formula
modelF <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelF)


### model G
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job +
day_of_week + cons.price.idx + cons.conf.idx + euribor3m ### Setting the formula
modelG <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelG)


### model H
formula <- subscribed ~ poutcome + month + pdays+ contact + previous + default + job +
day_of_week + cons.price.idx + cons.conf.idx + euribor3m + campaign ### Setting the formula
modelH <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelH)
```

model I
formula <- subscribed ~ poutcome + month + pdays + contact + previous + default + job + day_of_week + cons.price.idx + cons.conf.idx + euribor3m + campaign + age### Setting the formula
modelI <- glm(formula, data = train, family = "binomial")

summarise the model
summary(modelI) ### This model is dropped because it doesnt improve the previous model(modelH)

Checks for model accuracies

check modelA accuracy of the test data (20%)
predictions <- predict(modelA, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelB accuracy of the test data (20%)
predictions <- predict(modelB, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelC accuracy of the test data (20%)
predictions <- predict(modelC, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelD accuracy of the test data (20%)
predictions <- predict(modelD, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelE accuracy of the test data (20%)
predictions <- predict(modelE, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelF accuracy of the test data (20%)
predictions <- predict(modelF, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

check modelG accuracy of the test data (20%)
predictions <- predict(modelG, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)

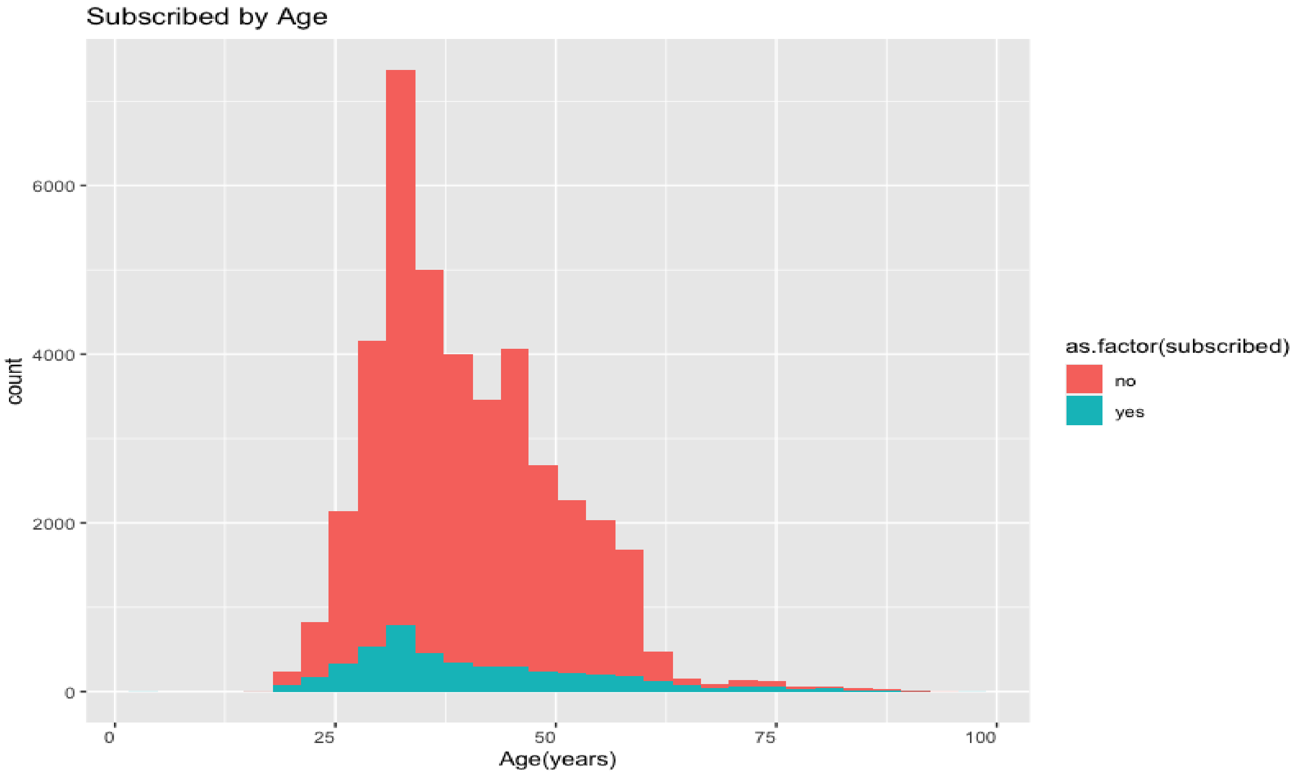
```
### check modelH accuracy of the test data (20%)
predictions <- predict(modelH, test, type = "response")
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)


### model accuracy using confusion matrix
confusionMatrix(data = class_pred, test$subscribed)

### View the odds ratio for ModelH
exp(modelH$coefficients)


#check the R squared value on the training data

logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n") #Function source: Field et al., 2012
}



### run logisticPseudoR2s for all models (A-H)
logisticPseudoR2s(modelA)
logisticPseudoR2s(modelB)
logisticPseudoR2s(modelC)
logisticPseudoR2s(modelD)
logisticPseudoR2s(modelE)
logisticPseudoR2s(modelF)
logisticPseudoR2s(modelG)
logisticPseudoR2s(modelH)
```

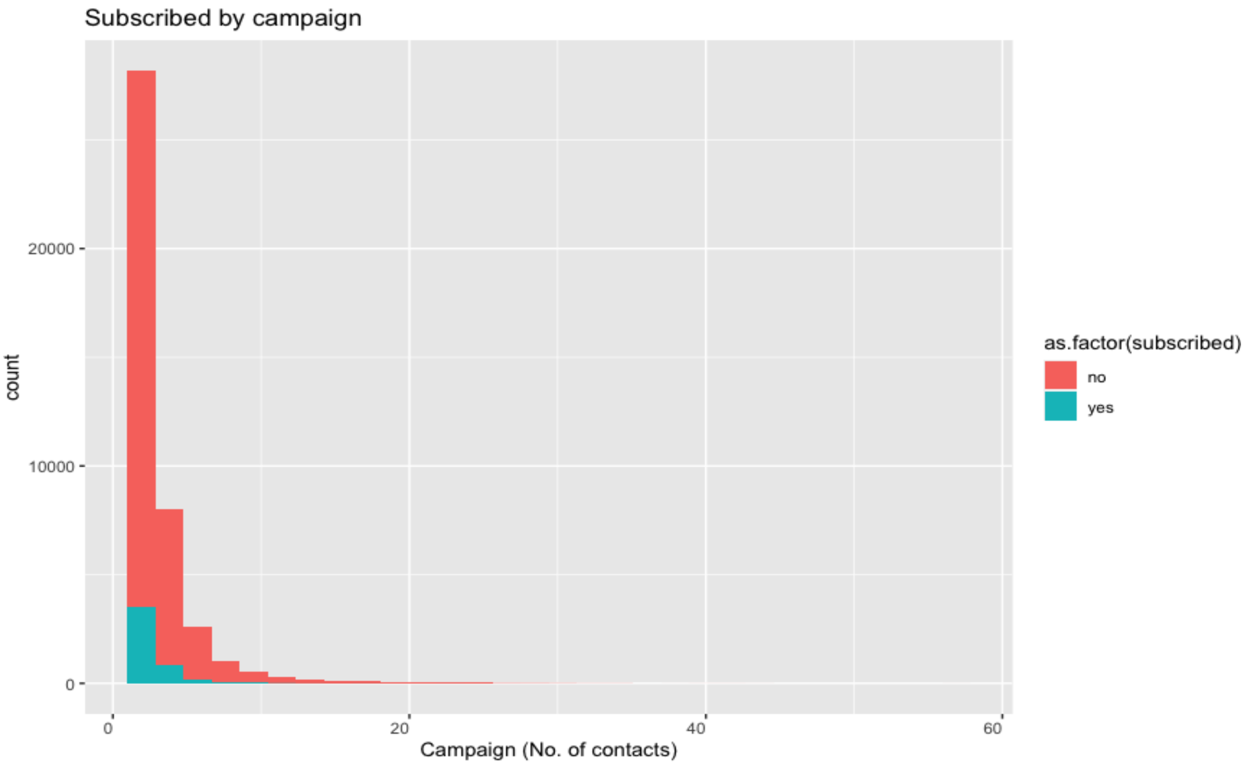
ASSUMPTION CHECKING AND THINGS THAT COULD GO WRONG

```
### Evaluate the model assumption
### Add the predicted probability to the dataframe using the fitted() function
train$predictedProbabilities <- fitted(modelH)

### checking probability of subscribed, and the actual outcome
head(data.frame(train$predictedProbabilities, train$subscribed))
tail(data.frame(train$predictedProbabilities, train$subscribed))

### Analysing the Residuals using the standardised residuals to check the model fit.
### As a rule of thumb only 5% should lie outside of ± 1.96
### and about 1% should lie outside of ± 2.58. Cases above 3 are a cause for concern.
train$standardisedResiduals <- rstandard(modelH) ### Also known as the errors in unit standard deviations

### counting how many of the standardised residuals is/are above 1.96. Only 5% above is acceptable
sum(train$standardisedResiduals > 1.96) ### 1464 data is > 1.96 which is < 5% of 32923 train data = 1646.15, thus satisfies assumption

### summarise standardisedResiduals to check for values above 3
summary(train$standardisedResiduals) ### No value above 3.0, thus satisfies assumption

### Examining Influential Cases using cooks distance
train$cook <- cooks.distance(modelH)

### check for cooks distance greater that 1
sum(train$cook > 1) ### assumption satisfied since none is greater than 1

### Checking for Multicolinearity having in mind value above 10 is not good.
vif(modelH) ### variables poutcome and pdays have multicolinearity issues since they are > 10
    ### each variable was removed at different intervals to see if modelH will impprove
    ### None of the variables improved modelH, each having a VIC of 18372 and 18368 respectively
    ### greater than ModelH whose VIC is 18345(when both variables are kept in the model)
    ### see Appendix 5
```

```
### Testing for the Linearity of the Logit

### There are 4 continous variables in ModelH and checking that they are linearly related
### to the log of the outcome variable (subscribed).
### Run the logistic regression model including predictors that are the interaction
### between each predictor and the log of itself.
### This test is essential in order to know how the model performs in a business environment. It
performs
### better than the accuracy in the residuals training dataset which is not reliable on a data it hasnt
seen before

### Create the interaction terms of the variable(numeric) with its log.
train$cpiLogInt <- log(train$cons.price.idx)*train$cons.price.idx
train$eb3LogInt <- log(train$euribor3m)*train$euribor3m
train$camLogInt <- log(train$campaign)*train$campaign
train$pdyLogInt <- log(train$pdays)*train$pdays



### build formula
formula<- subscribed ~ poutcome + month + pdays+ contact + previous + default + job +
day_of_week + cons.price.idx + cons.conf.idx + euribor3m + campaign + cpiLogInt + eb3LogInt
+ camLogInt + pdyLogInt ### Setting the formula
modelLogInt <- glm(formula, data = train, family = "binomial")

### summarise the model
summary(modelLogInt)

### Independent residuals : run dwtest . value between 1.5 and 2.5 is good
dwtest(modelH) ### DW = 1.9 This falls in between 1.5 and 2.5
```

Appendix 2: Other Visualisations





Appendix 3: Other multiple logistic regression models

```
> summary(modelA)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2756  -0.4413  -0.4296  -0.2634   2.6507

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.1937459  0.2654582  -0.730 0.465479
poutcomenonexistent  0.2860046  0.0953236   3.000 0.002697 **
poutcomesuccess     0.7827905  0.2259954   3.464 0.000533 ***
monthaug           -0.8517931  0.0747524 -11.395  < 2e-16 ***
monthdec            1.1070304  0.1981191   5.588 2.30e-08 ***
monthjul           -0.8579263  0.0742603 -11.553  < 2e-16 ***
monthjun           -0.0458330  0.0863636  -0.531 0.595628
monthmar            1.1626243  0.1160205  10.021  < 2e-16 ***
monthmay           -0.8233175  0.0712179 -11.561  < 2e-16 ***
monthnov           -0.9083685  0.0830282 -10.940  < 2e-16 ***
monthoct            1.0034488  0.1106620   9.068  < 2e-16 ***
monthsep            0.6839312  0.1227174   5.573 2.50e-08 ***
pdays              -0.0015220  0.0002312  -6.582 4.63e-11 ***
contacttelephone   -1.0923022  0.0583453 -18.721  < 2e-16 ***
previous            0.2319352  0.0619728   3.743 0.000182 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19684  on 32908  degrees of freedom
AIC: 19714

Number of Fisher Scoring iterations: 5
```

```
> summary(modelB)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2522  -0.4649  -0.4096  -0.2826   2.8064

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.1456663  0.2641775  -0.551 0.581362
poutcomenonexistent  0.2875860  0.0950410   3.026 0.002479 **
poutcomesuccess      0.7682577  0.2246767   3.419 0.000628 ***
monthaug            -0.8224314  0.0748942 -10.981  < 2e-16 ***
monthdec             1.0566322  0.1970410   5.362 8.21e-08 ***
monthjul            -0.8108462  0.0744625 -10.889  < 2e-16 ***
monthjun            -0.0395153  0.0861159  -0.459 0.646334
monthmar             1.1129903  0.1159235   9.601  < 2e-16 ***
monthmay            -0.8040522  0.0712535 -11.284  < 2e-16 ***
monthnov            -0.9231896  0.0830970 -11.110  < 2e-16 ***
monthoct             0.9509557  0.1103611   8.617  < 2e-16 ***
monthsep             0.6388916  0.1223036   5.224 1.75e-07 ***
pdays               -0.0015029  0.0002299  -6.536 6.32e-11 ***
contacttelephone    -1.0372113  0.0580661 -17.863  < 2e-16 ***
previous             0.2222724  0.0616779   3.604 0.000314 ***
defaultunknown      -0.5985049  0.0608284  -9.839  < 2e-16 ***
defaultyes          -8.3033389 84.4267720  -0.098 0.921655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19575  on 32906  degrees of freedom
AIC: 19609

Number of Fisher Scoring iterations: 9
```

Appendix 3 contd: Other logistic regression models

```
> summary(modelC)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4278  -0.4573  -0.3977  -0.2792   2.8406

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.1648434  0.2653303  -0.621 0.534418
poutcomenonexistent  0.2688381  0.0951771   2.825 0.004734 **
poutcomesuccess      0.8000543  0.2238511   3.574 0.000352 ***
monthaug            -0.8472568  0.0759503 -11.155  < 2e-16 ***
monthdec             0.8900704  0.1976103   4.504 6.66e-06 ***
monthjul            -0.7806909  0.0748539 -10.430  < 2e-16 ***
monthjun            -0.0304278  0.0861214  -0.353 0.723854
monthmar             1.0286771  0.1167769   8.809  < 2e-16 ***
monthmay            -0.7568824  0.0717080 -10.555  < 2e-16 ***
monthnov            -0.9060710  0.0838357 -10.808  < 2e-16 ***
monthoct             0.8361757  0.1112867   7.514 5.75e-14 ***
monthsep             0.5299419  0.1229307   4.311 1.63e-05 ***
pdays               -0.0014162  0.0002291  -6.182 6.31e-10 ***
contacttelephone    -1.0196652  0.0578559 -17.624  < 2e-16 ***
previous             0.1993744  0.0617343   3.230 0.001240 **
defaultunknown      -0.5583410  0.0618001  -9.035  < 2e-16 ***
defaultyes          -8.2657869 84.4685349  -0.098 0.922046
jobblue-collar      -0.2807358  0.0607408  -4.622 3.80e-06 ***
jobentrepreneur     -0.2144143  0.1167221  -1.837 0.066215 .
jobhousemaid        -0.0718344  0.1300894  -0.552 0.580817
jobmanagement       -0.0317415  0.0786206  -0.404 0.686412
jobretired           0.5011673  0.0813448   6.161 7.23e-10 ***
jobself-employed    -0.1180202  0.1107153  -1.066 0.286433
jobservices         -0.2386071  0.0771566  -3.093 0.001985 **
jobstudent           0.6327379  0.1007721   6.279 3.41e-10 ***
jobtechnician       -0.1139320  0.0592071  -1.924 0.054317 .
jobunemployed       -0.0426676  0.1229119  -0.347 0.728486
jobunknown          -0.1788917  0.2368483  -0.755 0.450069
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19440  on 32895  degrees of freedom
AIC: 19496

Number of Fisher Scoring iterations: 9
```

```
> summary(modelD)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3255  -0.4598  -0.3823  -0.2736   2.9037

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.1337484  0.2689715  -0.497 0.619007
poutcomenonexistent  0.2644745  0.0952798   2.776 0.005507 **
poutcomesuccess      0.7859334  0.2246652   3.498 0.000468 ***
monthaug            -0.8776322  0.0765622 -11.463  < 2e-16 ***
monthdec             0.8895010  0.1981111   4.490 7.13e-06 ***
monthjul            -0.8079628  0.0754496 -10.709  < 2e-16 ***
monthjun            -0.0403855  0.0866560  -0.466 0.641184
monthmar             1.0313519  0.1173072   8.792  < 2e-16 ***
monthmay            -0.7794388  0.0723332 -10.776  < 2e-16 ***
monthnov            -0.9363773  0.0843123 -11.106  < 2e-16 ***
monthoct             0.8082205  0.1116901   7.236 4.61e-13 ***
monthsep             0.4947314  0.1235713   4.004 6.24e-05 ***
pdays               -0.0014308  0.0002299  -6.223 4.86e-10 ***
contacttelephone    -1.0263717  0.0579581 -17.709  < 2e-16 ***
previous             0.1975363  0.0618331   3.195 0.001400 **
defaultunknown      -0.5541002  0.0618480  -8.959  < 2e-16 ***
defaultyes          -8.3196937 84.4680872  -0.098 0.921539
jobblue-collar      -0.2822173  0.0608164  -4.640 3.48e-06 ***
jobentrepreneur     -0.2109546  0.1168034  -1.806 0.070908 .
jobhousemaid        -0.0705707  0.1302267  -0.542 0.587883
jobmanagement       -0.0301595  0.0787516  -0.383 0.701742
jobretired           0.4961445  0.0815608   6.083 1.18e-09 ***
jobself-employed    -0.1153341  0.1109101  -1.040 0.298392
jobservices         -0.2350568  0.0772008  -3.045 0.002329 **
jobstudent           0.6242706  0.1008969   6.187 6.12e-10 ***
jobtechnician       -0.1128289  0.0592647  -1.904 0.056934 .
jobunemployed       -0.0451147  0.1229265  -0.367 0.713615
jobunknown          -0.1919842  0.2379919  -0.807 0.419849
day_of_weekmon      -0.2314838  0.0631923  -3.663 0.000249 ***
day_of_weekthu       0.0630049  0.0601591   1.047 0.294959
day_of_weektue       0.0715115  0.0618141   1.157 0.247321
day_of_weekwed       0.1352596  0.0613029   2.206 0.027355 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19398  on 32891  degrees of freedom
AIC: 19462

Number of Fisher Scoring iterations: 9
```

Appendix 3 contd: Other logistic regression models

```
> summary(modelE)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2627  -0.4532  -0.3779  -0.2698   2.9105

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        35.6575161  4.0958771   8.706  < 2e-16 ***
poutcomenonexistent  0.4334683  0.0971090   4.464 8.05e-06 ***
poutcomesuccess     0.8696481  0.2227580   3.904 9.46e-05 ***
monthaug           -0.8469523  0.0765958 -11.057  < 2e-16 ***
monthdec            0.7043070  0.1980243   3.557 0.000376 ***
monthjul           -0.5398347  0.0809435  -6.669 2.57e-11 ***
monthjun            0.1112381  0.0875147   1.271 0.203700
monthmar            0.9913891  0.1168574   8.484  < 2e-16 ***
monthmay           -0.8121254  0.0718290 -11.306  < 2e-16 ***
monthnov           -0.9190402  0.0842485 -10.909  < 2e-16 ***
monthoct            0.7528983  0.1111932   6.771 1.28e-11 ***
monthsep            0.5206245  0.1232504   4.224 2.40e-05 ***
pdays              -0.0013689  0.0002278  -6.009 1.86e-09 ***
contacttelephone   -0.7446823  0.0644115 -11.561  < 2e-16 ***
previous            0.3136809  0.0631375   4.968 6.76e-07 ***
defaultunknown     -0.5175009  0.0620721  -8.337  < 2e-16 ***
defaultyes         -8.2274480 84.4566489  -0.097 0.922396
jobblue-collar     -0.2794192  0.0609609  -4.584 4.57e-06 ***
jobentrepreneur    -0.2193147  0.1170340  -1.874 0.060939 .
jobhousemaid       -0.0597537  0.1301820  -0.459 0.646233
jobmanagement      -0.0430162  0.0790897  -0.544 0.586516
jobretired          0.4609821  0.0818091   5.635 1.75e-08 ***
jobself-employed   -0.1238008  0.1114187  -1.111 0.266512
jobservices        -0.2296308  0.0772793  -2.971 0.002964 **
jobstudent          0.5874125  0.1004447   5.848 4.97e-09 ***
jobtechnician      -0.0964831  0.0593773  -1.625 0.104181
jobunemployed      -0.0634689  0.1230466  -0.516 0.605986
jobunknown         -0.2118292  0.2398163  -0.883 0.377075
day_of_weekmon     -0.2254882  0.0632576  -3.565 0.000364 ***
day_of_weekthu      0.0722290  0.0602655   1.199 0.230717
day_of_weektue      0.0659262  0.0620107   1.063 0.287717
day_of_weekwed      0.1419735  0.0614122   2.312 0.020788 *
cons.price.idx     -0.3868874  0.0441703  -8.759  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19322  on 32890  degrees of freedom
AIC: 19388

Number of Fisher Scoring iterations: 9
```

```
> summary(modelF)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4091  -0.4458  -0.3732  -0.2812   2.8925

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        25.1646075  4.3191156   5.826 5.67e-09 ***
poutcomenonexistent  0.4296104  0.0978483   4.391 1.13e-05 ***
poutcomesuccess     0.8473847  0.2241009   3.781 0.000156 ***
monthaug           -1.3560686  0.0968820 -13.997  < 2e-16 ***
monthdec            0.2356583  0.2076823   1.135 0.256499
monthjul           -0.8166613  0.0874192  -9.342  < 2e-16 ***
monthjun           -0.0580631  0.0894840  -0.649 0.516425
monthmar            0.8986803  0.1190782   7.547 4.45e-14 ***
monthmay           -0.9049015  0.0737412 -12.271  < 2e-16 ***
monthnov           -1.1724541  0.0900846 -13.015  < 2e-16 ***
monthoct            0.2301478  0.1284872   1.791 0.073259 .
monthsep           -0.0343507  0.1393039  -0.247 0.805227
pdays              -0.0013426  0.0002291  -5.862 4.58e-09 ***
contacttelephone   -1.0185894  0.0746328 -13.648  < 2e-16 ***
previous            0.3024985  0.0631783   4.788 1.68e-06 ***
defaultunknown     -0.5196673  0.0620947  -8.369  < 2e-16 ***
defaultyes         -8.1991999 84.4448642  -0.097 0.922651
jobblue-collar     -0.2545910  0.0611970  -4.160 3.18e-05 ***
jobentrepreneur    -0.1999992  0.1172031  -1.706 0.087927 .
jobhousemaid       -0.0815304  0.1309906  -0.622 0.533669
jobmanagement      -0.0417723  0.0794264  -0.526 0.598941
jobretired          0.4237774  0.0824443   5.140 2.74e-07 ***
jobself-employed   -0.1147428  0.1118213  -1.026 0.304832
jobservices        -0.2105498  0.0775422  -2.715 0.006622 **
jobstudent          0.5902717  0.1011964   5.833 5.45e-09 ***
jobtechnician      -0.0852710  0.0595793  -1.431 0.152368
jobunemployed      -0.0609269  0.1235571  -0.493 0.621937
jobunknown         -0.2387869  0.2399426  -0.995 0.319647
day_of_weekmon     -0.2193343  0.0634978  -3.454 0.000552 ***
day_of_weekthu      0.0740180  0.0605213   1.223 0.221327
day_of_weektue      0.0536796  0.0622984   0.862 0.388878
day_of_weekwed      0.1428729  0.0615877   2.320 0.020350 *
cons.price.idx     -0.2515582  0.0473746  -5.310 1.10e-07 ***
cons.conf.idx       0.0460421  0.0053609   8.588  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 19248  on 32889  degrees of freedom
AIC: 19316

Number of Fisher Scoring iterations: 9
```

Appendix 3 contd: Other logistic regression models

```
> summary(modelG)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1516  -0.3990  -0.3304  -0.2456   2.9004

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -4.521e+01  4.397e+00 -10.284  < 2e-16 ***
poutcomenonexistent  4.791e-01  9.466e-02   5.061 4.17e-07 ***
poutcomesuccess      6.834e-01  2.181e-01   3.134  0.00173 **
monthaug            -1.246e-01  1.006e-01  -1.238  0.21570
monthdec             4.004e-01  2.002e-01   2.000  0.04551 *
monthjul             1.747e-01  9.165e-02   1.907  0.05658 .
monthjun             1.156e-01  8.958e-02   1.290  0.19693
monthmar             9.093e-01  1.210e-01   7.515 5.68e-14 ***
monthmay            -6.141e-01  7.233e-02  -8.491  < 2e-16 ***
monthnov            -1.039e-01  9.671e-02  -1.075  0.28246
monthoct             2.159e-01  1.226e-01   1.761  0.07823 .
monthsep            -2.068e-01  1.323e-01  -1.564  0.11790
pdays               -1.127e-03  2.231e-04  -5.051 4.41e-07 ***
contacttelephone    -5.477e-01  6.748e-02  -8.117 4.77e-16 ***
previous            -4.169e-02  6.085e-02  -0.685  0.49333
defaultunknown      -2.685e-01  6.351e-02  -4.228 2.36e-05 ***
defaultyes          -7.594e+00  8.388e+01  -0.091  0.92786
jobblue-collar      -1.761e-01  6.268e-02  -2.809  0.00497 **
jobentrepreneur     -9.730e-02  1.192e-01  -0.816  0.41434
jobhousemaid        -1.039e-01  1.362e-01  -0.763  0.44559
jobmanagement        2.668e-02  8.130e-02   0.328  0.74276
jobretired           1.794e-01  8.410e-02   2.133  0.03294 *
jobself-employed    -7.353e-02  1.140e-01  -0.645  0.51889
jobservices         -1.457e-01  7.906e-02  -1.843  0.06529 .
jobstudent           2.231e-01  1.017e-01   2.194  0.02821 *
jobtechnician        3.716e-04  6.143e-02   0.006  0.99517
jobunemployed       -1.082e-01  1.261e-01  -0.858  0.39087
jobunknown          -2.031e-01  2.391e-01  -0.849  0.39565
day_of_weekmon      -2.641e-01  6.473e-02  -4.080 4.51e-05 ***
day_of_weekthu       6.883e-02  6.168e-02   1.116  0.26450
day_of_weektue       5.300e-02  6.362e-02   0.833  0.40482
day_of_weekwed       1.652e-01  6.302e-02   2.622  0.00875 **
cons.price.idx       5.115e-01  4.830e-02  10.591  < 2e-16 ***
cons.conf.idx        4.812e-02  5.116e-03   9.407  < 2e-16 ***
euribor3m           -5.745e-01  1.787e-02 -32.149  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18299  on 32888  degrees of freedom
AIC: 18369
```

Appendix 3 contd: Other logistic regression models

```
> summary(modelI) ### This model is dropped because it doesnt improve the previous model(modelH)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1634  -0.4023  -0.3276  -0.2445   3.0407

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.519e+01  4.400e+00 -10.270  < 2e-16 ***
poutcomenonexistent  4.836e-01  9.472e-02   5.106 3.30e-07 ***
poutcomesuccess     6.696e-01  2.182e-01   3.068  0.00215 **
monthaug           -1.028e-01  1.007e-01  -1.021  0.30721
monthdec            4.321e-01  2.002e-01   2.158  0.03090 *
monthjul            2.099e-01  9.201e-02   2.281  0.02255 *
monthjun            1.339e-01  8.973e-02   1.493  0.13557
monthmar            9.314e-01  1.213e-01   7.682 1.57e-14 ***
monthmay           -5.945e-01  7.262e-02  -8.186 2.70e-16 ***
monthnov           -1.075e-01  9.663e-02  -1.113  0.26587
monthoct            2.166e-01  1.225e-01   1.768  0.07705 .
monthsep           -1.943e-01  1.323e-01  -1.469  0.14176
pdays              -1.134e-03  2.233e-04  -5.079 3.79e-07 ***
contacttelephone   -5.362e-01  6.771e-02  -7.918 2.41e-15 ***
previous           -4.126e-02  6.093e-02  -0.677  0.49831
defaultunknown     -2.749e-01  6.417e-02  -4.284 1.84e-05 ***
defaultyes         -7.638e+00  8.368e+01  -0.091  0.92728
jobblue-collar     -1.838e-01  6.278e-02  -2.927  0.00342 **
jobentrepreneur    -1.082e-01  1.196e-01  -0.904  0.36584
jobhousemaid       -1.245e-01  1.380e-01  -0.903  0.36679
jobmanagement       1.411e-02  8.186e-02   0.172  0.86318
jobretired          1.303e-01  1.025e-01   1.271  0.20375
jobself-employed   -7.568e-02  1.141e-01  -0.663  0.50714
jobservices        -1.471e-01  7.904e-02  -1.861  0.06269 .
jobstudent          2.420e-01  1.054e-01   2.295  0.02173 *
jobtechnician      -3.076e-03  6.147e-02  -0.050  0.96009
jobunemployed      -1.134e-01  1.262e-01  -0.898  0.36902
jobunknown         -2.149e-01  2.400e-01  -0.896  0.37050
day_of_weekmon     -2.638e-01  6.473e-02  -4.076 4.58e-05 ***
day_of_weekthu      6.103e-02  6.173e-02   0.989  0.32280
day_of_weektue      3.776e-02  6.371e-02   0.593  0.55338
day_of_weekwed      1.523e-01  6.310e-02   2.413  0.01582 *
cons.price.idx      5.110e-01  4.836e-02  10.567  < 2e-16 ***
cons.conf.idx       4.691e-02  5.133e-03   9.140  < 2e-16 ***
euribor3m          -5.678e-01  1.790e-02 -31.721  < 2e-16 ***
campaign           -5.022e-02  1.032e-02  -4.865 1.15e-06 ***
age                 1.857e-03  2.134e-03   0.870  0.38426
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18272  on 32886  degrees of freedom
AIC: 18346

Number of Fisher Scoring iterations: 9
```

Appendix 4: Additional Descriptive Statistics

Descriptive statistics of selected unclean data

```
> describe(musei1[SH1])
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
poutcome*	1	41153	1.93	0.36	2.00	2.00	0.00	1.00	3.00	2.00	-0.88	3.97
month*	2	41153	5.70	2.76	5.00	5.77	4.45	1.00	11.00	10.00	-0.19	-1.30
pdays	3	41113	962.41	187.07	999.00	999.00	0.00	0.00	999.00	999.00	-4.92	22.18
contact*	4	41153	1.36	0.48	1.00	1.33	0.00	1.00	2.00	1.00	0.56	-1.68
previous	5	41153	0.17	0.50	0.00	0.05	0.00	0.00	7.00	7.00	3.83	20.09
default*	6	41153	2.20	0.42	2.00	2.14	0.00	1.00	4.00	3.00	1.25	0.26
job*	7	41153	4.72	3.59	3.00	4.48	2.97	1.00	12.00	11.00	0.45	-1.39
day_of_week*	8	41153	3.01	1.40	3.00	3.01	1.48	1.00	5.00	4.00	0.01	-1.27
cons.price.idx	9	41153	93.58	0.58	93.75	93.58	0.56	92.20	94.77	2.57	-0.23	-0.83
cons.conf.idx	10	41153	-40.51	4.63	-41.80	-40.61	6.52	-50.80	-26.90	23.90	0.31	-0.36
euribor3m	11	41153	3.62	1.73	4.86	3.80	0.16	0.63	5.04	4.41	-0.71	-1.41
campaign	12	41153	2.57	2.77	2.00	1.99	1.48	1.00	56.00	55.00	4.76	36.95
age	13	41153	40.03	10.44	38.00	39.30	10.38	4.00	147.00	143.00	0.81	1.11

*categorical variables

Descriptive statistics of selected clean data

```
> describe(musei1[SH1])
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
poutcome*	1	41153	1.93	0.36	2.00	2.00	0.00	1.00	3.00	2.00	-0.88	3.97
month*	2	41153	5.23	2.32	5.00	5.31	2.97	1.00	10.00	9.00	-0.31	-1.03
pdays	3	41153	962.41	186.98	999.00	999.00	0.00	0.00	999.00	999.00	-4.92	22.20
contact*	4	41153	1.36	0.48	1.00	1.33	0.00	1.00	2.00	1.00	0.56	-1.68
previous	5	41153	0.17	0.50	0.00	0.05	0.00	0.00	7.00	7.00	3.83	20.09
default*	6	41153	1.21	0.41	1.00	1.14	0.00	1.00	3.00	2.00	1.44	0.07
job*	7	41153	4.72	3.59	3.00	4.48	2.97	1.00	12.00	11.00	0.45	-1.39
day_of_week*	8	41153	3.01	1.40	3.00	3.01	1.48	1.00	5.00	4.00	0.01	-1.27
cons.price.idx	9	41153	93.58	0.58	93.75	93.58	0.56	92.20	94.77	2.57	-0.23	-0.83
cons.conf.idx	10	41153	-40.51	4.63	-41.80	-40.61	6.52	-50.80	-26.90	23.90	0.31	-0.36
euribor3m	11	41153	3.62	1.73	4.86	3.80	0.16	0.63	5.04	4.41	-0.71	-1.41
campaign	12	41153	2.57	2.77	2.00	1.99	1.48	1.00	56.00	55.00	4.76	36.95
age	13	41153	40.02	10.42	38.00	39.30	10.38	4.00	98.00	94.00	0.78	0.80

*categorical variables

Appendix 4 contd: Additional Descriptive Statistics

Descriptive statistics of all variables

```
> describeBy(musei1)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
ID	1	41153	20577.00	11879.99	20577.00	20577.00	15252.99	1.00	41153.00	41152.00	0.00
age	2	41153	40.03	10.44	38.00	39.30	10.38	4.00	147.00	143.00	0.81
job*	3	41153	4.72	3.59	3.00	4.48	2.97	1.00	12.00	11.00	0.45
marital*	4	41153	2.17	0.61	2.00	2.21	0.00	1.00	4.00	3.00	-0.06
education*	5	41153	4.75	2.14	4.00	4.88	2.97	1.00	8.00	7.00	-0.24
default*	6	41153	2.20	0.42	2.00	2.14	0.00	1.00	4.00	3.00	1.25
housing*	7	41153	2.07	0.99	3.00	2.09	0.00	1.00	3.00	2.00	-0.14
loan*	8	41153	1.33	0.72	1.00	1.16	0.00	1.00	3.00	2.00	1.82
contact*	9	41153	1.36	0.48	1.00	1.33	0.00	1.00	2.00	1.00	0.56
month*	10	41153	5.70	2.76	5.00	5.77	4.45	1.00	11.00	10.00	-0.19
day_of_week*	11	41153	3.01	1.40	3.00	3.01	1.48	1.00	5.00	4.00	0.01
duration	12	41153	258.23	259.17	180.00	210.57	139.36	0.00	4918.00	4918.00	3.26
campaign	13	41153	2.57	2.77	2.00	1.99	1.48	1.00	56.00	55.00	4.76
pdays	14	41113	962.41	187.07	999.00	999.00	0.00	0.00	999.00	999.00	-4.92
previous	15	41153	0.17	0.50	0.00	0.05	0.00	0.00	7.00	7.00	3.83
poutcome*	16	41153	1.93	0.36	2.00	2.00	0.00	1.00	3.00	2.00	-0.88
emp.var.rate	17	41153	0.08	1.57	1.10	0.27	0.44	-3.40	1.40	4.80	-0.72
cons.price.idx	18	41153	93.58	0.58	93.75	93.58	0.56	92.20	94.77	2.57	-0.23
cons.conf.idx	19	41153	-40.51	4.63	-41.80	-40.61	6.52	-50.80	-26.90	23.90	0.31
euribor3m	20	41153	3.62	1.73	4.86	3.80	0.16	0.63	5.04	4.41	-0.71
nr.employed	21	41153	5167.02	72.28	5191.00	5178.41	55.00	4963.60	5228.10	264.50	-1.04
subscribed*	22	41153	1.11	0.32	1.00	1.02	0.00	1.00	2.00	1.00	2.45

*categorical variables

Appendix 5: Removed poutcome and pdays variables that failed Multicollinearity test at different intervals with one variable held constant did not improve Model H

```
> summary(modelH)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2115  -0.4055  -0.3285  -0.2446   3.0407

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.738e+01  4.361e+00 -10.866  < 2e-16 ***
monthaug         -7.944e-02  1.004e-01  -0.791  0.42885
monthdec          4.510e-01  1.994e-01   2.262  0.02368 *
monthjul          2.097e-01  9.189e-02   2.282  0.02249 *
monthjun          1.297e-01  8.957e-02   1.447  0.14778
monthmar          9.463e-01  1.215e-01   7.787 6.86e-15 ***
monthmay         -5.960e-01  7.240e-02  -8.232  < 2e-16 ***
monthnov         -1.033e-01  9.646e-02  -1.071  0.28435
monthoct          2.199e-01  1.223e-01   1.798  0.07214 .
monthsep         -1.749e-01  1.319e-01  -1.326  0.18487
pdays            -1.711e-03  8.903e-05 -19.212  < 2e-16 ***
contacttelephone -5.373e-01  6.775e-02  -7.931 2.18e-15 ***
previous         -2.886e-01  3.954e-02  -7.298 2.92e-13 ***
defaultunknown   -2.692e-01  6.352e-02  -4.239 2.25e-05 ***
defaultyes       -7.719e+00  8.413e+01  -0.092  0.92689
jobblue-collar   -1.827e-01  6.267e-02  -2.915  0.00355 **
jobentrepreneur  -1.007e-01  1.190e-01  -0.846  0.39775
jobhousemaid     -1.038e-01  1.359e-01  -0.764  0.44469
jobmanagement     2.289e-02  8.119e-02   0.282  0.77802
jobretired        1.808e-01  8.408e-02   2.151  0.03150 *
jobself-employed -7.338e-02  1.139e-01  -0.644  0.51955
jobservices      -1.520e-01  7.906e-02  -1.922  0.05458 .
jobstudent        2.305e-01  1.015e-01   2.272  0.02311 *
jobtechnician    -6.059e-03  6.144e-02  -0.099  0.92145
jobunemployed    -1.066e-01  1.258e-01  -0.847  0.39695
jobunknown       -2.018e-01  2.390e-01  -0.844  0.39844
day_of_weekmon   -2.631e-01  6.470e-02  -4.066 4.79e-05 ***
day_of_weekthu    6.194e-02  6.169e-02   1.004  0.31537
day_of_weektue    3.995e-02  6.367e-02   0.627  0.53041
day_of_weekwed    1.574e-01  6.302e-02   2.497  0.01251 *
cons.price.idx    5.461e-01  4.781e-02  11.423  < 2e-16 ***
cons.conf.idx     4.705e-02  5.112e-03   9.204  < 2e-16 ***
euribor3m        -5.645e-01  1.783e-02 -31.666  < 2e-16 ***
campaign         -4.987e-02  1.031e-02  -4.837 1.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18304  on 32889  degrees of freedom
AIC: 18372
```

```
> summary(modelH)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1788  -0.4040  -0.3285  -0.2444   3.0412

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.621e+01  4.390e+00 -10.526  < 2e-16 ***
poutcomenonexistent 5.460e-01  9.462e-02   5.770 7.93e-09 ***
poutcomesuccess     1.691e+00  8.765e-02  19.287  < 2e-16 ***
monthaug           -9.454e-02  1.005e-01  -0.941  0.34690
monthdec            4.258e-01  1.997e-01   2.132  0.03298 *
monthjul            2.149e-01  9.172e-02   2.343  0.01912 *
monthjun            1.357e-01  8.943e-02   1.518  0.12906
monthmar            9.305e-01  1.211e-01   7.684 1.54e-14 ***
monthmay           -5.987e-01  7.236e-02  -8.275  < 2e-16 ***
monthnov           -9.668e-02  9.643e-02  -1.003  0.31606
monthoct            2.337e-01  1.223e-01   1.911  0.05597 .
monthsep           -1.954e-01  1.321e-01  -1.479  0.13921
contacttelephone   -5.313e-01  6.763e-02  -7.856 3.97e-15 ***
previous            6.522e-02  5.827e-02   1.119  0.26300
defaultunknown     -2.655e-01  6.350e-02  -4.180 2.91e-05 ***
defaultyes         -7.654e+00  8.384e+01  -0.091  0.92725
jobblue-collar     -1.820e-01  6.268e-02  -2.904  0.00369 **
jobentrepreneur    -1.027e-01  1.191e-01  -0.862  0.38844
jobhousemaid       -1.050e-01  1.360e-01  -0.772  0.44009
jobmanagement       2.327e-02  8.126e-02   0.286  0.77458
jobretired          1.814e-01  8.411e-02   2.157  0.03098 *
jobself-employed   -7.757e-02  1.141e-01  -0.680  0.49669
jobservices        -1.495e-01  7.895e-02  -1.894  0.05819 .
jobstudent          2.255e-01  1.016e-01   2.220  0.02645 *
jobtechnician      -9.068e-04  6.141e-02  -0.015  0.98822
jobunemployed      -1.100e-01  1.263e-01  -0.871  0.38351
jobunknown         -1.930e-01  2.386e-01  -0.809  0.41859
day_of_weekmon     -2.633e-01  6.467e-02  -4.071 4.69e-05 ***
day_of_weekthu      6.004e-02  6.165e-02   0.974  0.33017
day_of_weektue      3.829e-02  6.363e-02   0.602  0.54740
day_of_weekwed      1.508e-01  6.302e-02   2.393  0.01670 *
cons.price.idx      5.101e-01  4.826e-02  10.570  < 2e-16 ***
cons.conf.idx       4.749e-02  5.117e-03   9.280  < 2e-16 ***
euribor3m          -5.694e-01  1.787e-02 -31.871  < 2e-16 ***
campaign           -4.995e-02  1.032e-02  -4.841 1.29e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23192  on 32922  degrees of freedom
Residual deviance: 18298  on 32888  degrees of freedom
AIC: 18368
```

Appendix 6: Test for the Linearity of the Logit

```
> summary(modelLogInt)

Call:
glm(formula = formula, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1759  -0.4014  -0.3263  -0.2457   3.2726

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.411e+03  1.144e+03  -1.233  0.21753
poutcomenonexistent 4.864e-01  9.586e-02   5.074 3.89e-07 ***
poutcomesuccess     5.811e-01  2.360e-01   2.462  0.01381 *
monthaug           -3.398e-02  1.098e-01  -0.310  0.75694
monthdec            5.115e-01  2.049e-01   2.497  0.01254 *
monthjul            2.950e-01  1.025e-01   2.877  0.00401 **
monthjun            1.779e-01  9.673e-02   1.840  0.06584 .
monthmar            9.652e-01  1.237e-01   7.805 5.95e-15 ***
monthmay           -5.740e-01  7.387e-02  -7.770 7.84e-15 ***
monthnov           -8.845e-02  1.027e-01  -0.861  0.38917
monthoct            3.035e-01  1.359e-01   2.233  0.02553 *
monthsep           -1.549e-01  1.408e-01  -1.100  0.27136
pdays              -3.712e-02  3.169e-02  -1.171  0.24147
contacttelephone   -5.070e-01  7.465e-02  -6.792 1.11e-11 ***
previous           -3.493e-02  6.227e-02  -0.561  0.57482
defaultunknown     -2.631e-01  6.359e-02  -4.137 3.52e-05 ***
defaultyes         -7.645e+00  8.385e+01  -0.091  0.92736
jobblue-collar     -1.748e-01  6.281e-02  -2.782  0.00540 **
jobentrepreneur    -9.822e-02  1.192e-01  -0.824  0.40990
jobhousemaid       -1.013e-01  1.361e-01  -0.745  0.45640
jobmanagement       2.146e-02  8.140e-02   0.264  0.79208
jobretired          1.913e-01  8.440e-02   2.267  0.02338 *
jobself-employed   -7.448e-02  1.142e-01  -0.652  0.51410
jobservices        -1.475e-01  7.923e-02  -1.862  0.06258 .
jobstudent          2.239e-01  1.019e-01   2.197  0.02805 *
jobtechnician      -2.400e-04  6.151e-02  -0.004  0.99689
jobunemployed      -1.065e-01  1.264e-01  -0.843  0.39923
jobunknown         -2.020e-01  2.402e-01  -0.841  0.40044
day_of_weekmon     -2.661e-01  6.480e-02  -4.107 4.01e-05 ***
day_of_weekthu      6.200e-02  6.186e-02   1.002  0.31623
day_of_weektue      3.919e-02  6.381e-02   0.614  0.53915
day_of_weekwed      1.538e-01  6.322e-02   2.433  0.01497 *
cons.price.idx      8.118e+01  6.784e+01   1.196  0.23150
cons.conf.idx       5.639e-02  1.004e-02   5.615 1.97e-08 ***
euribor3m          -5.517e-02  4.610e-01  -0.120  0.90475
campaign            5.472e-02  5.231e-02   1.046  0.29561
cpiLogInt          -1.456e+01  1.225e+01  -1.188  0.23486
eb3LogInt          -2.687e-01  2.362e-01  -1.138  0.25525
camLogInt          -4.218e-02  2.119e-02  -1.991  0.04650 *
pdyLogInt           5.177e-03  4.557e-03   1.136  0.25596
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23152  on 32908  degrees of freedom
Residual deviance: 18252  on 32869  degrees of freedom
  (14 observations deleted due to missingness)
AIC: 18332

Number of Fisher Scoring iterations: 9
```