Bank Term Deposit Subscription Prediction Using Logistic Regression

By

Prayas Sachdeva

Word Count

2200

Contents

Introduction and Background

The banking industry relies heavily on forecasting to reduce financial losses through the detection of credit fraud. At the same time, predicting customer behaviour can lead to increased profits. To predict whether customers will subscribe to long-term deposits, banks must use features of customers and marketing campaigns to create a two-class classification problem. Logistic regression is used as a classical statistical model for binary outcomes. It is popular for its ability to produce probability estimates, which can be used to make classifications by setting a threshold. The coefficient estimates are also useful for obtaining odds ratios which can be used in business decisions (Yang, 2016).

Hypothesis

Number	Variables	Relationship
H1	Quarterly Indicator of Number of Employees and Subscribed (Golecha, 2017)	Positive
H2	Euribor 3 months rate and Subscribed (Borugadda et al., 2021)	Negative
H3	Employee Variation Rate and Subscribed (Hou et al., 2022)	Positive
H4	Type of Job and Subscribed (Ilham et al., 2019)	Positive

Methodology

1. Business Understanding

The data provided is that of a telemarketing campaign run by a bank to promote subscription to their term deposit product. The data consists of information related to customer attributes, economical factors and some information from the previous campaign. The objective is to explore, understand and build a logistic regression model using R language to predict customer's likeliness to subscribe to the bank's term deposit as this is information will then be utilised for curating more robust marketing campaigns.

2. Data Understanding

The data collected is available in .xlsx format and has 22 variables with 41153 observations. The variables are mix of continuous and categorical variables. Specifically, 11 numerical and 11 character variables. To achieve the laid out of objective, an Analytics Base Table is with selective variables that

may influence the customer's inclination toward subscribing to terms deposits. However, before creating the regression model, the data quality issues are addressed in the following section.

3. Data Preparation

The analytics base table or ABT is created using 16 variables out of the mix which are likely to affect the outcome of the target variables 'Subscribed'.

```
abt                 41153 obs. of 16 variables
  $ age           : num [1:41153] 56 57 37 40 56 45 59 41 24 25 ...
  $ job           : chr [1:41153] "housemaid" "services" "services" "admin." ...
  $ marital       : chr [1:41153] "married" "married" "married" "married" ...
  $ education     : chr [1:41153] "basic.4y" "high.school" "high.school" "basic.6y"
  $ default       : chr [1:41153] "no" "unknown" "no" "no" ...
  $ housing       : chr [1:41153] "no" "no" "yes" "no" ...
  $ loan          : chr [1:41153] "no" "no" "no" "no" ...
  $ campaign      : num [1:41153] 1 1 1 1 1 1 1 1 1 1 ...
  $ previous      : num [1:41153] 0 0 0 0 0 0 0 0 0 0 ...
  $ poutcome      : chr [1:41153] "nonexistent" "nonexistent" "nonexistent" "nonex
  $ emp.var.rate  : num [1:41153] 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
  $ cons.price.idx: num [1:41153] 94 94 94 94 94 ...
  $ cons.conf.idx : num [1:41153] -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4
  $ euribor3m     : num [1:41153] 4.86 4.86 4.86 4.86 4.86 ...
  $ nr.employed   : num [1:41153] 5191 5191 5191 5191 5191 ...
  $ subscribed    : chr [1:41153] "no" "no" "no" "no" ...
```

Using the summary() function, handful of data quality issues are highlighted which are required to be addressed to building an accurate model.

```
> summary(abt)
      age            job              marital           education          default
 Min.   :  4.00   Length:41153       Length:41153       Length:41153       Length:41153
 1st Qu.: 32.00   Class :character   Class :character   Class :character   Class :character
 Median : 38.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 40.03
 3rd Qu.: 47.00
 Max.   :147.00
   housing              loan              campaign        previous        poutcome
 Length:41153       Length:41153       Min.   : 1.000   Min.   :0.0000   Length:41153
 Class :character   Class :character   1st Qu.: 1.000   1st Qu.:0.0000   Class :character
 Mode  :character   Mode  :character   Median : 2.000   Median :0.0000   Mode  :character
                                       Mean   : 2.568   Mean   :0.1731
                                       3rd Qu.: 3.000   3rd Qu.:0.0000
                                       Max.   :56.000   Max.   :7.0000
  emp.var.rate      cons.price.idx  cons.conf.idx      euribor3m      nr.employed
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634   Min.   :4964
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.344   1st Qu.:5099
 Median : 1.10000   Median :93.75   Median :-41.80   Median :4.857   Median :5191
 Mean   : 0.08102   Mean   :93.58   Mean   :-40.51   Mean   :3.620   Mean   :5167
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045   Max.   :5228
  subscribed
 Length:41153
 Class :character
 Mode  :character
```

Following are the data quality issues:

- Outliers

Using the boxplot() and hist(), outliers in numerical variables like ‘age’, ‘campaign’, ‘previous’ and ‘consumer confidence index’ are converted to NA.

For ‘Age’ variable, observations <18 and >85 are assumed to be the outliers and are converted to NA as they were a total of 51 observations.

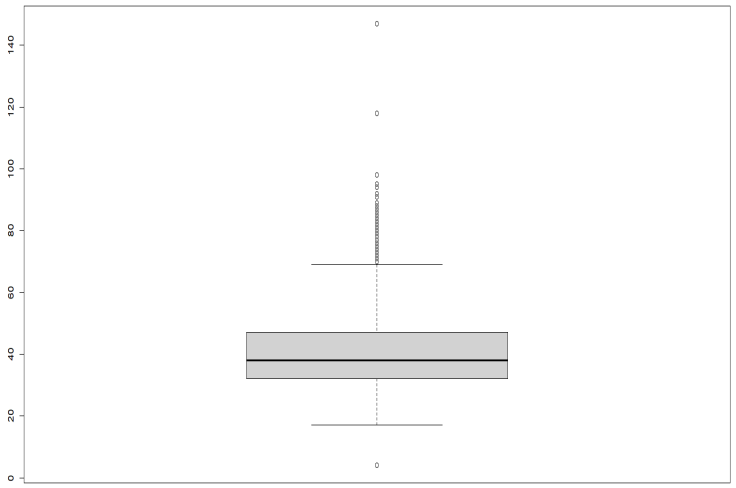

Figure 1 Boxplot of Age

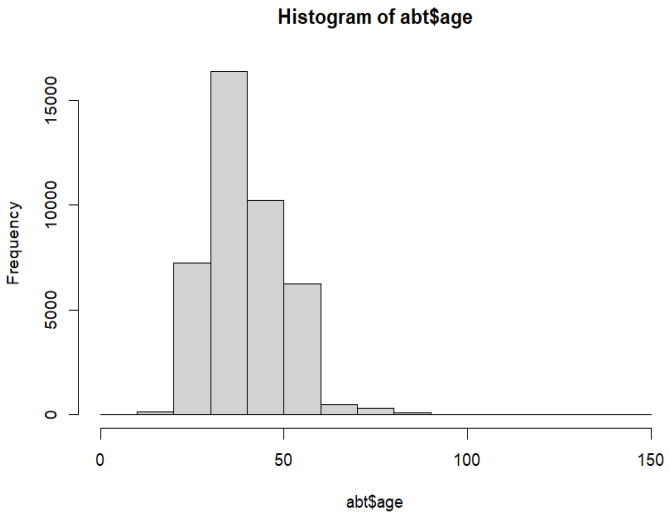

Figure 2 Histogram of Age

For ‘Campaign’ variable, the values above 40 are assumed to be the outliers and thus a total of 6 observations were replaced with NA.

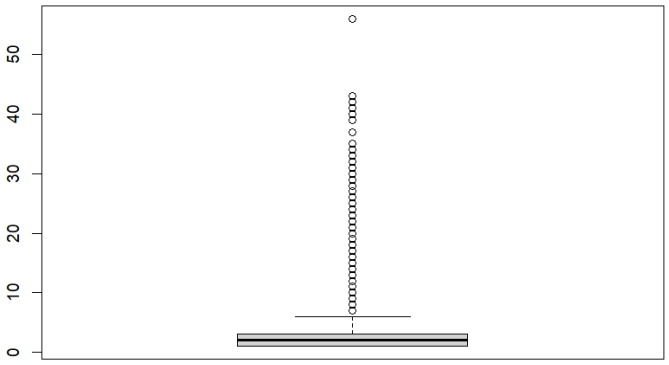

Figure 3 Boxplot of Campaign

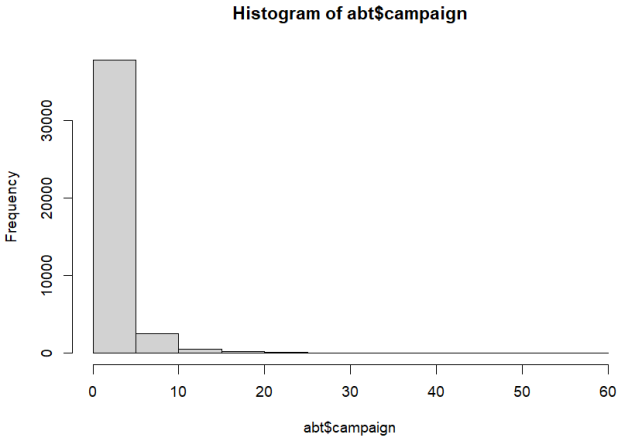

Figure 4 Histogram of Campaign

For 'Cons.Conf.Idx', the values > -30 are identified as outliers and thus a total 714 observations are converted to NA.

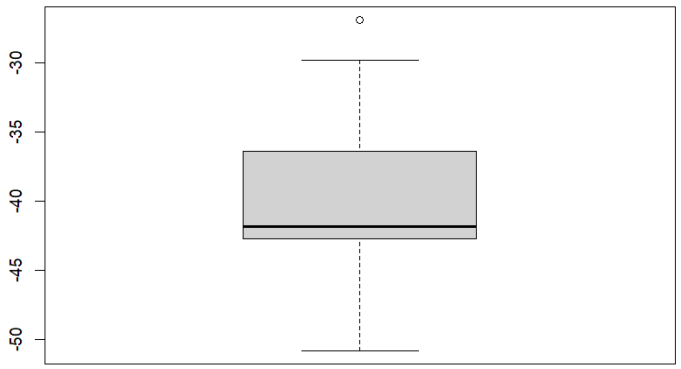

Figure 5 Boxplot of Consumer Confidence Index

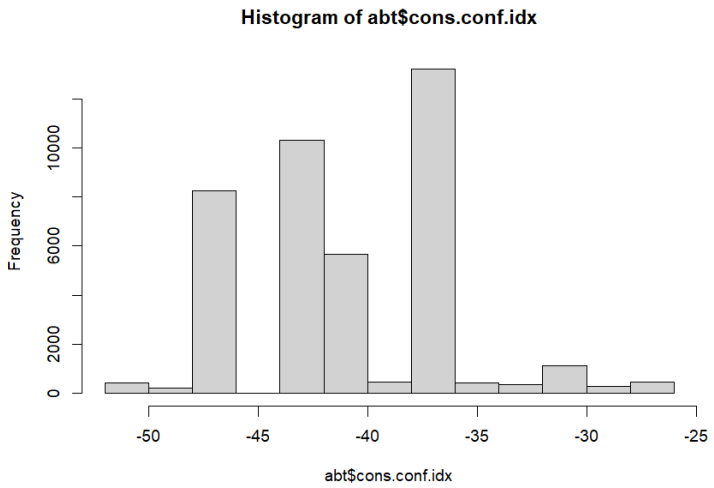

Figure 6 Histogram of Consumer Confidence Index

Lastly, for 'Previous' variable, values >4 are assumed as outliers and thus 24 observations are converted to NA.

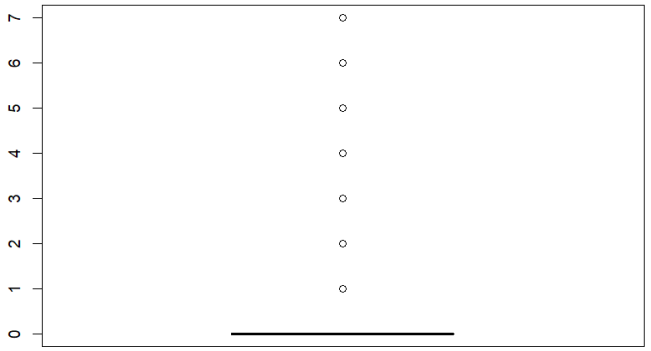

Figure 7 Boxplot of Previous

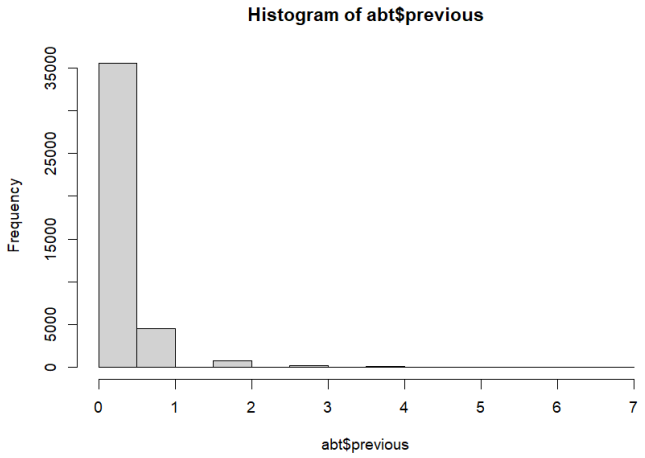

Figure 8 Histogram of Previous

- 'Unknown' observations

Using the summary() function on 'Job', 'Marital', 'Education' and 'Loan', it was found that there are lot of 'Unknown' observation which can affect the accuracy of the model and thus were deemed to be a data quality issues. A total of 330, 80, 1727 and 990 respectively, observation were converted to NA.

However, with 'Default', the 'Unknown' observations were 8583. Since this constituted to be 20% (approx.) of the data, the same were replaced to 'No' along with 221 observations mistyped as 'n' on the basis of Mode Imputation method.

```
> summary(as.factor(abt$job))
       admin.   blue-collar  entrepreneur     housemaid    management       retired self-employed
        10416          9244          1454          1059          2920          1719          1421
     services       student    technician    unemployed       unknown
         3963           875          6739          1013           330
> #Marital
> summary(as.factor(abt$marital))
divorced  married   single  unknown
    4609    24905    11559       80
> #Education
> summary(as.factor(abt$education))
           basic.4y           basic.6y           basic.9y        high.school          illiterate
               4176               2286               6038               9508                 18
professional.course   university.degree            unknown
               5240              12160               1727
> #Default
> summary(as.factor(abt$default))
      n      no unknown     yes
    221   32346    8583       3
> #Loan
> summary((as.factor(abt$loan)))
     no unknown     yes
  33923     990    6240
```

- Dealing with the NA and the incorrect data types

In the process of cleaning the data, a total 3922 observation were converted to NA and thus needs to be omitted from the 'abt' for further processing using the omit() function.

```
> colSums(is.na(abt))
            age             job         marital       education         default         housing
             51             330              80            1727               0               0
           loan        campaign        previous        poutcome    emp.var.rate cons.price.idx
            990               6              24               0               0               0
 cons.conf.idx       euribor3m     nr.employed      subscribed
            714               0               0               0
```

Furthermore, using the mutateif() function, variables with the character datatypes were corrected to as factor datatype for better interpretation of the model.

4. Modelling

- Test of Association

Once the data is prepared for further analysis, appropriate test of correlations were conducted on different sets of input variables and the dependent variable, i.e, 'Subscribed'. Essentially, the cor.test() function is used to derive correlation using 'Spearman' and 'Pearson' method for categorical and continuous variables respectively, along with chisq.test() to identify the statistically significant difference between the expected and the observed frequencies of two ordinal variables. Following tables provides more insight into the same:

Dependent Variable	Input Variable	Function and Method	Output	Interpretation of the output
Subscribed	Age	Cor.test() and Pearson	0.0209743	The test results indicate that there is a statistically significant correlation between these two variables, as the p-value is less than 0.05. The correlation is

				0.0209743, there is a positive correlation, meaning that as age increases, subscribed is more likely to increase.
Subscribed	Job	Chisq.test()	x-squared 673.26	The results show that there is a significant difference between the two variables (X-squared = 673.26, df = 10, p-value < 2.2e-16).
Subscribed	Marital Status	Chisq.test()	x-squared 97.856	There is a statistically significant difference in subscription rate between the different marital statuses, as indicated by the X-squared value of 97.856 and the p-value of less than 2.2e-16.
Subscribed	Education	Chisq.test(), Cor.test() and Pearson	x-squared 157.03, 0.05533752	There is a statistically significant positive correlation between the two variables. The p-value of less than 2.2e-16 indicates that the correlation is highly significant. The sample estimates of the correlation are 0.05533752.
Subscribed	Default	Cor.test() and Spearman	-0.003076673	There is a weak negative relationship between the two variables, but it is not statistically significant because the p-value is 0.5513. The correlation could range from a slight positive to a slight negative relationship.
Subscribed	Housing	Cor.test() and Spearman	0.0110858	The Spearman's rank correlation rho is 0.0110858, which indicates a very weak positive relationship between the two variables. The p-value of 0.03178 indicates that the correlation is statistically significant.
Subscribed	Loan	Cor.test() and Spearman	-0.003066862	In this case, the rho value of -0.003066862 indicates a weak negative relationship between the variables as.numeric(abt$subscribed) and as.numeric(abt$loan). The p-value of 0.5525 suggests that this relationship is not statistically significant.
Subscribed	Campaign	Cor.test() and Pearson	-0.0609361	In this case, the correlation between the variables as.numeric(abt$subscribed) and abt$campaign is -0.0609361. This value indicates a weak negative linear relationship between the two variables, meaning that as the value of as.numeric(abt$subscribed) increases, the value of abt$campaign decreases.
Subscribed	Previous	Cor.test() and Pearson	0.2157882	This indicates that there is a significant correlation between the two variables and that the true correlation is not equal to 0.

Subscribed	Poutcome	Cor.test() and Spearman	0.1256268	This suggests that there is a positive correlation between the two variables, indicating that as one variable increases, the other one is likely to increase as well.
Subscribed	Emp.Var.Rate	Cor.test() and Pearson	-0.2750085	The given values indicate that there is a strong negative linear correlation between the variables "subscribed" (as.numeric) and "emp.var.rate" in the given dataset. The correlation coefficient is -0.275 and the p-value is less than 2.2e-16, which is highly significant.
Subscribed	Cons.Price.Idx	Cor.test() and Pearson	-0.1085067	The correlation is -0.1085067, which indicates that there is a negative correlation between the two variables.
Subscribed	Cons.Conf.Idx	Cor.test() and Pearson	0.01039025	There is a statistically significant correlation between the two variables, as the p-value is less than 0.05. The correlation coefficient is 0.01039025, indicating that the variables are weakly correlated.
Subscribed	Euribor3m	Cor.test() and Pearson	-0.2872919	The test results show that there is a significant negative correlation between the two variables, with a correlation coefficient of -0.2872919 and a p-value of less than 2.2e-16.
Subscribed	Nr.Employed	Cor.test() and Pearson	-0.3344536	This suggests that there is a strong negative correlation between the two variables.

- Visualisations

The relationship is further explored visually using the ggplot() function:

- Nr.Employed and Subscribed

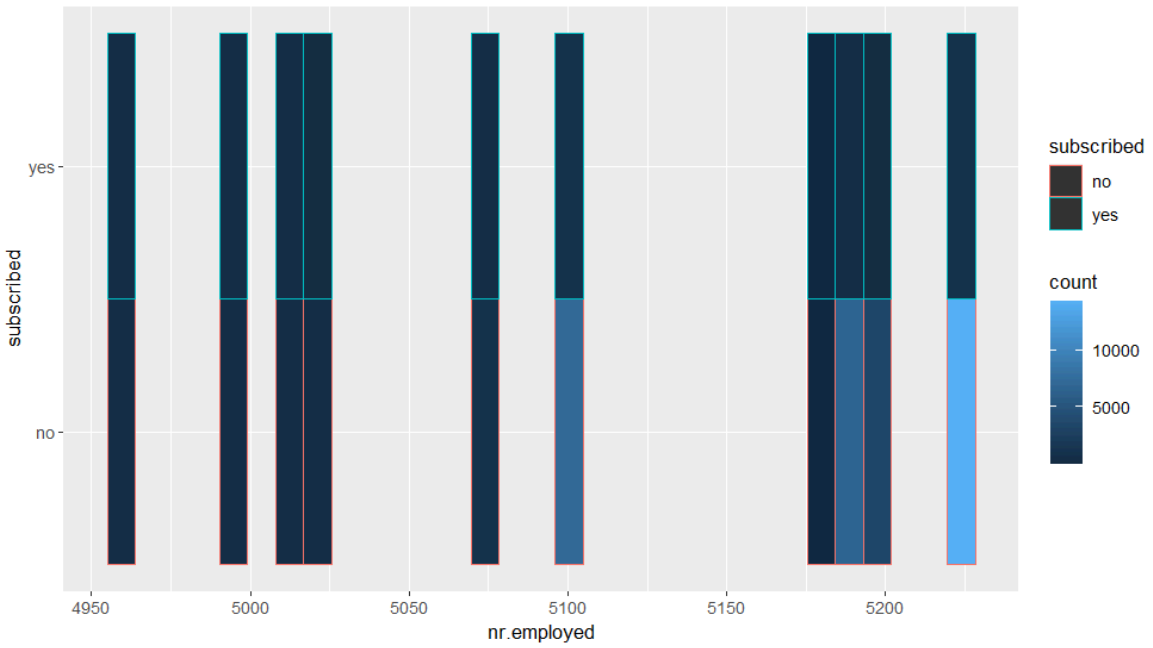

Figure 9 Binplot for Nr.employed and Subscribed

- Euribor3m and Subscribed

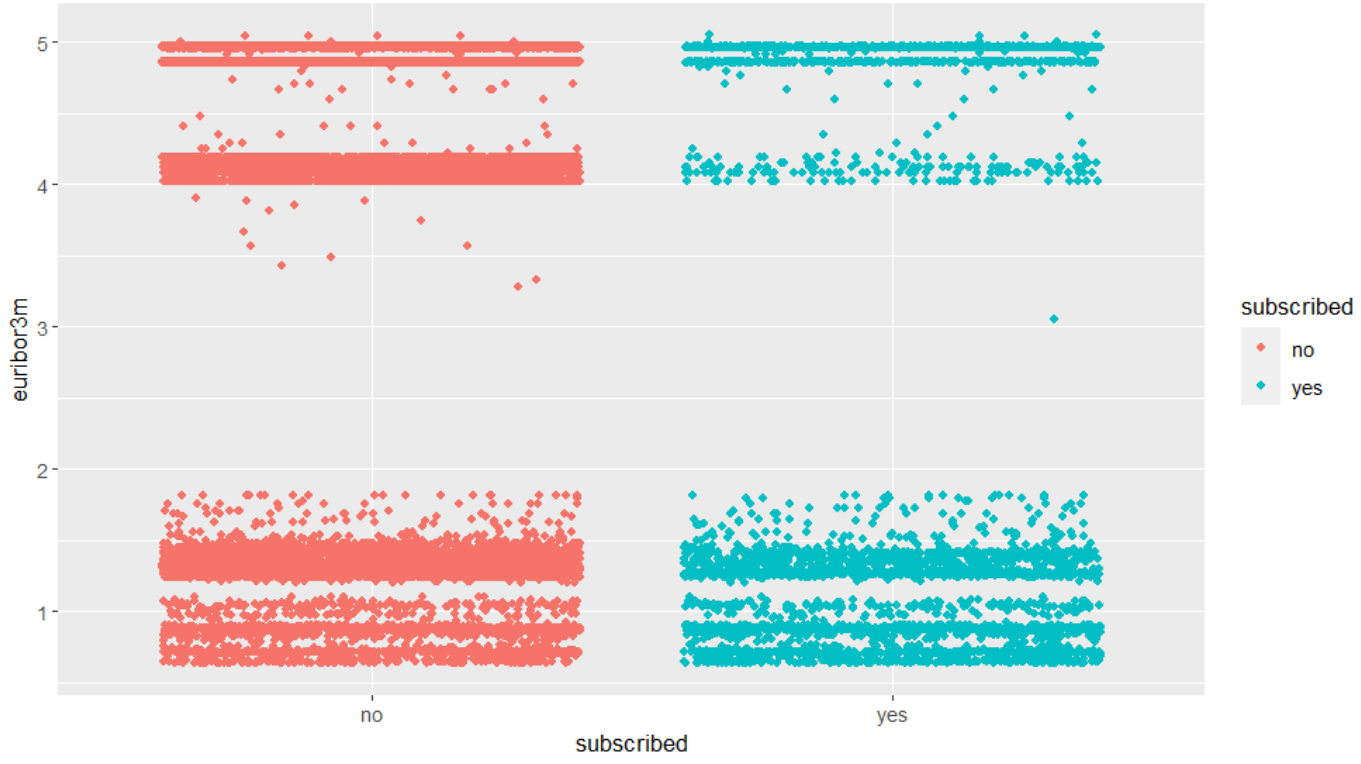

Figure 10 Jitterplot for Euribor3m and Subscribed

- Education and Subscribed

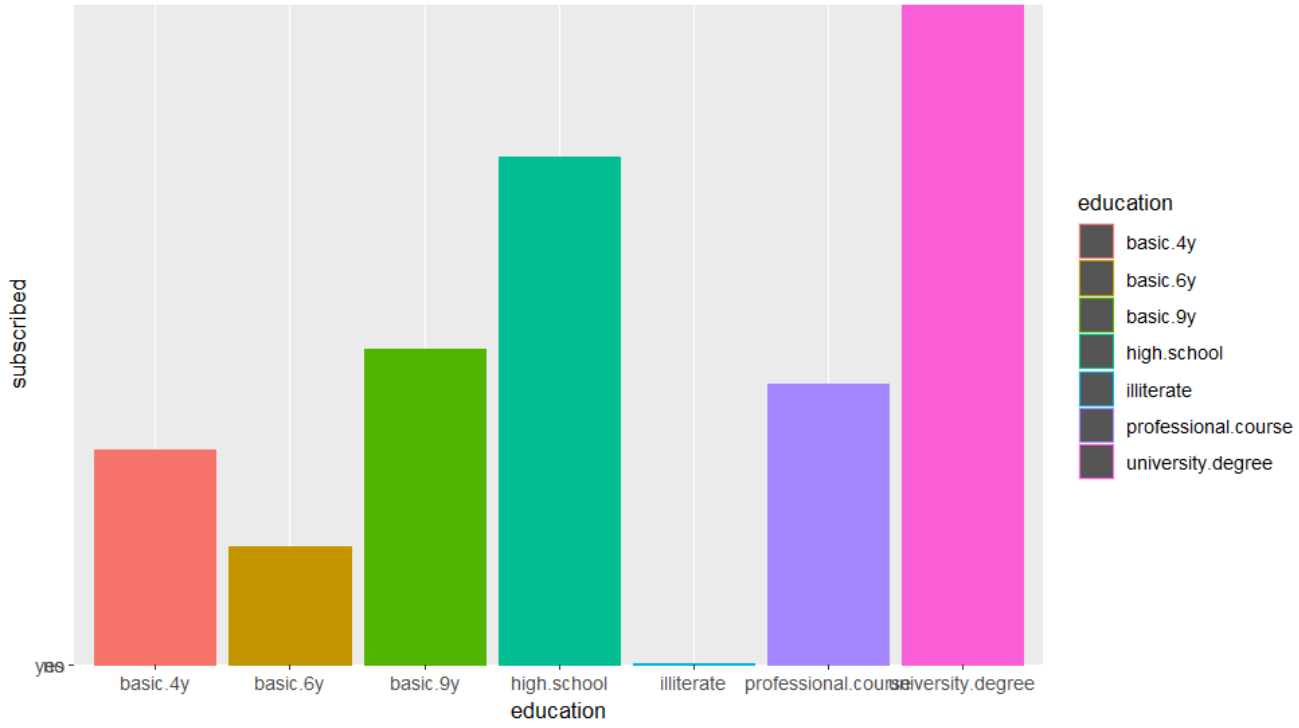

Figure 11 Column plot for Education and Subscribed

- Emp.Var.Rate and Subscribed

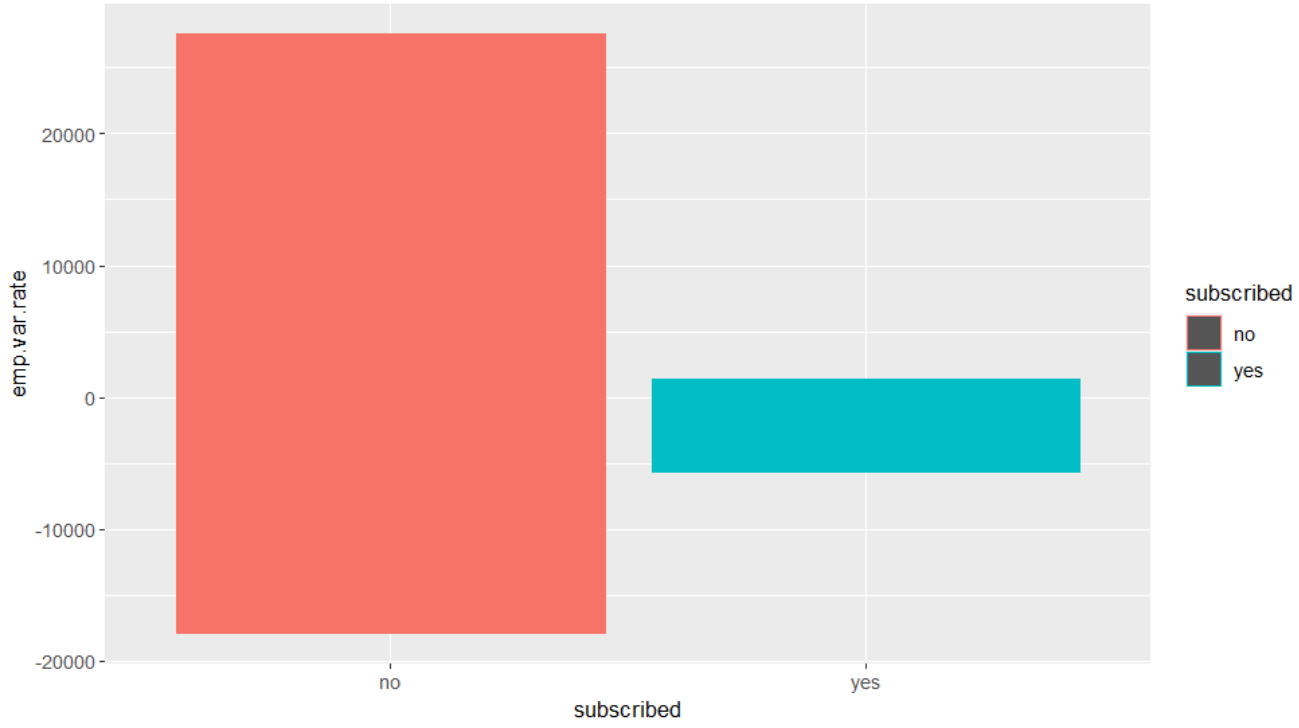

Figure 12 Column plot for Emp.Var.Rate and Subscribed

- Poutcome and Subscribed

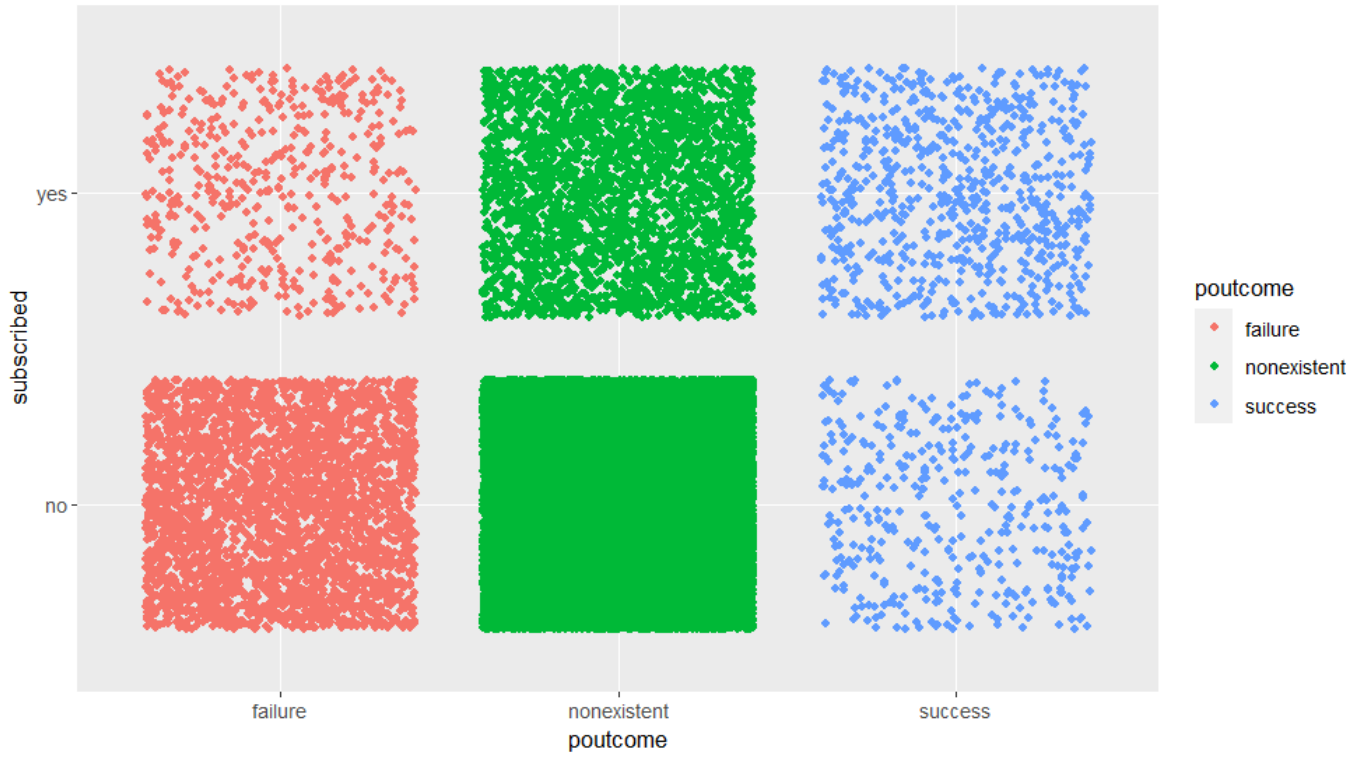

Figure 13 Jitterplot for Poutcome and Subscribed

- Previous and Subscribed

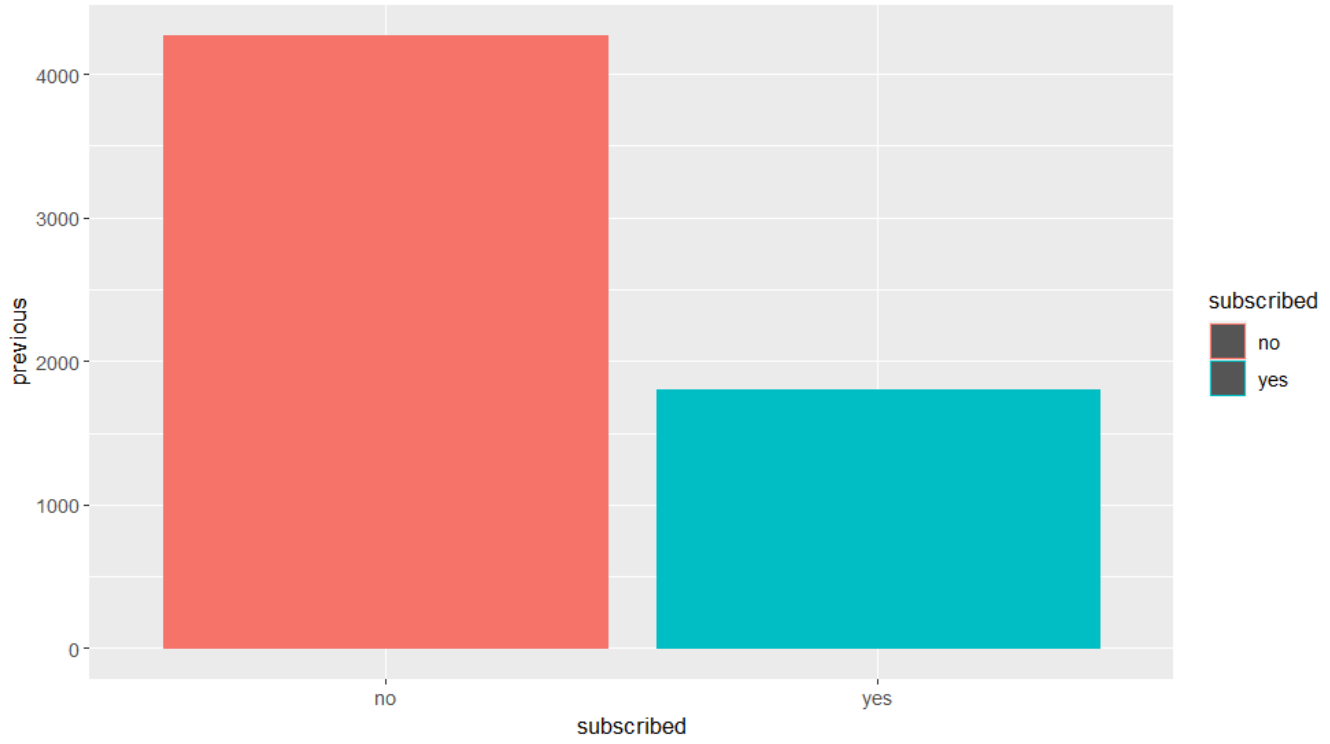

Figure 14 Barplot for Previous and Subscribed

- Job and Subscribed

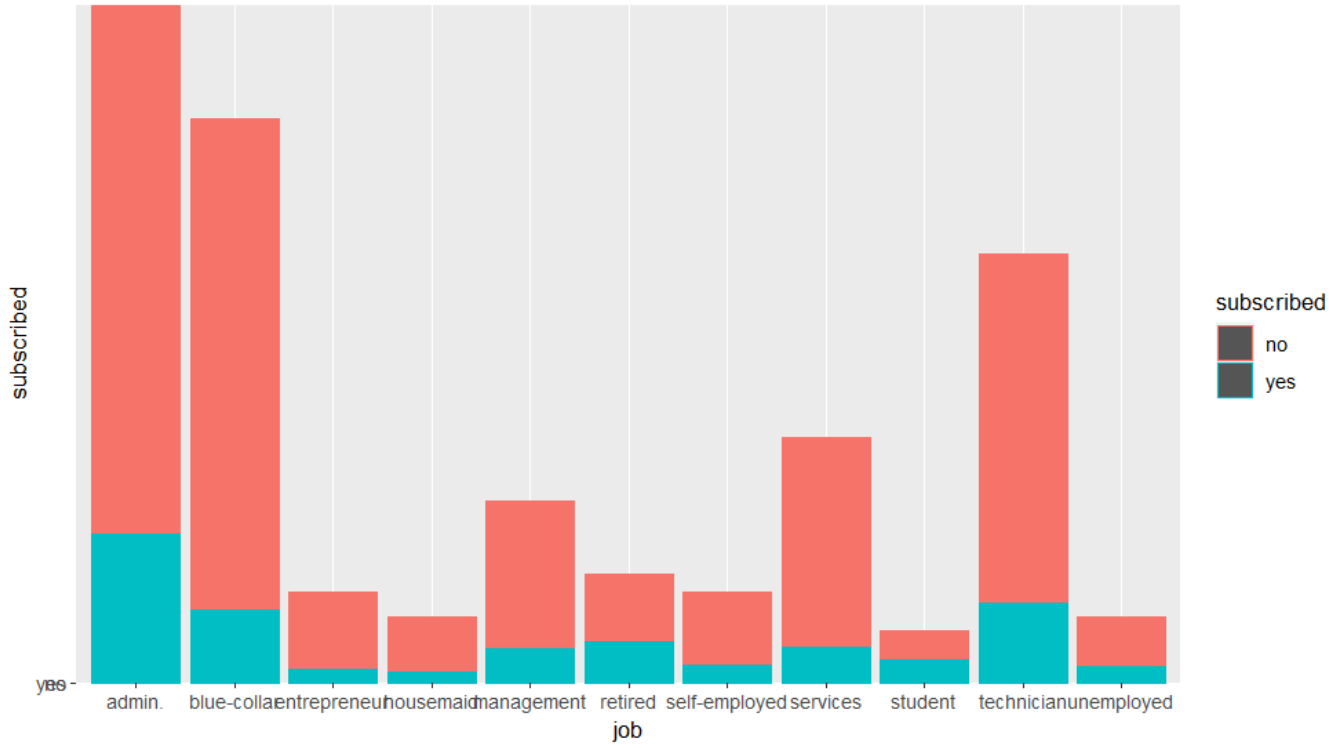

Figure 15 Column plot for Job and Subscribed

- Cons.Price.Idx and Subscribed

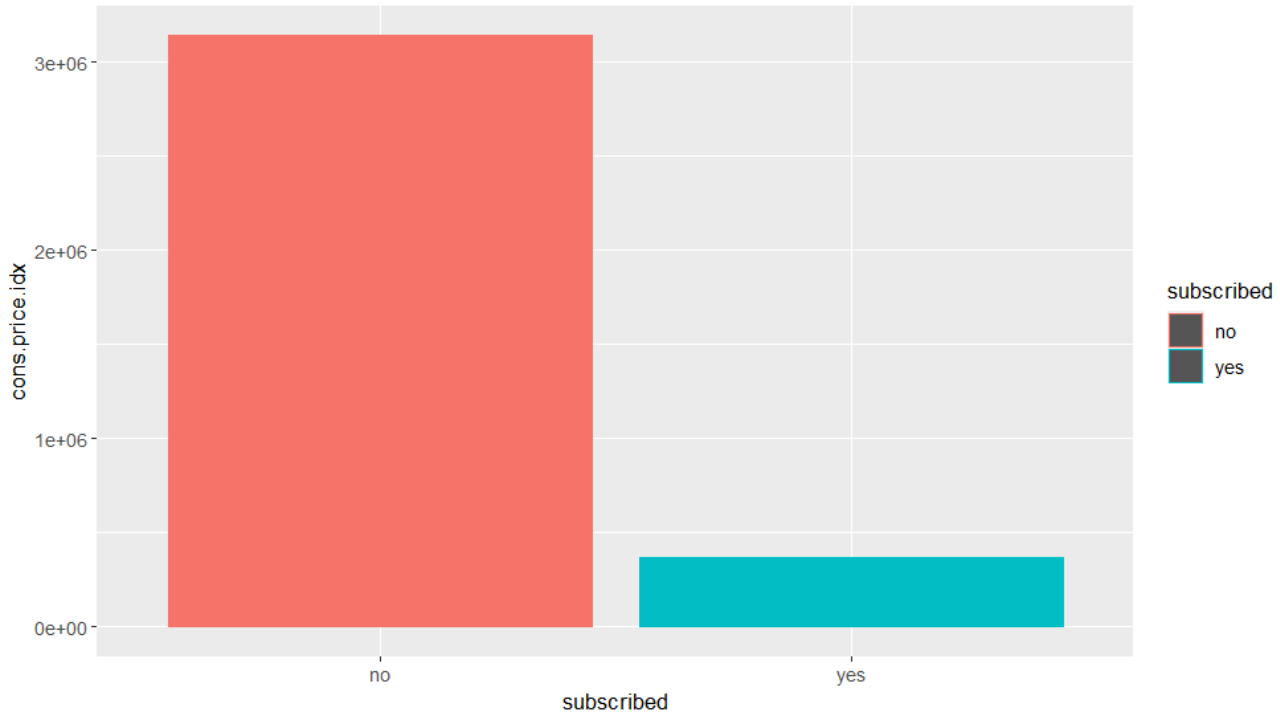

Figure 16 Barplot for Cons.Price.Idx and Subscribed

- Logistic Regression Model

The cleaned and prepared data is first split into train and test datasets using the createDataPartition() function with 80% of data in train dataset and 20% in test dataset. The model is built on the train assetset using the glm() function. A model with 10 different variables is selected with lowest AIC value of 16703.

```
> formula6 <- subscribed ~ poutcome + emp.var.rate + previous + job + campaign + cons.pric
e.idx + education + cons.conf.idx + nr.employed + euribor3m
```

5. Evaluation

Once the model is finalised, the same model is used to make predictions on the test dataset using the predict() function and then the same is inputted into 'class_pred' vector as factor. For deriving the accuracy and the kappa value of the model postResample() function is used. The Kappa value of 0.25 indicated toward the imbalanced data in the dataset (Jiang et al., 2015).

Furthermore, accuracy check is evaluated using the confusionmatrix() function (Aussalet & Hardman, 2010).

```
> postResample(class_pred, test$subscribed)
 Accuracy     Kappa
0.9045588 0.2501188

Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  6645  653
       yes   63  141

               Accuracy : 0.9046
                 95% CI : (0.8977, 0.9111)
    No Information Rate : 0.8942
    P-Value [Acc > NIR] : 0.001604

                  Kappa : 0.2501

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

            Sensitivity : 0.9906
            Specificity : 0.1776
         Pos Pred Value : 0.9105
         Neg Pred Value : 0.6912
             Prevalence : 0.8942
         Detection Rate : 0.8858
   Detection Prevalence : 0.9728
      Balanced Accuracy : 0.5841

       'Positive' Class : no
```

The model's R squared is assessed using the PseudoR2() function, rounded off to 2 decimal places (Field et al., 2012).

```
> round(PseudoR2(multiplemodel6, which = 'all'), 2)
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
           0.18            0.18            0.11            0.23            0.11
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
           0.27            0.18            0.22            0.18        16702.62
            BIC          logLik         logLik0              G2
       16918.66       -8325.31       -10134.79         3618.96
```

6. Assumption Checks

It is essential to identify if the selected model is breaking any assumptions as it may lead the model to spit biased and inaccurate results (James et al., 2022).

- Predicted Probabilities

The predicted values are derived using the fitted() function and dataframe is created. Using the head() function, the first part of the said dataframe is called to view the actual and predicted outcomes of 'Subscribed'.

```
> train$predictedprobabilities <- fitted(multiplemodel6)
> head(data.frame(train$predictedprobabilities, train$subscribed))
  train.predictedprobabilities train.subscribed
1                   0.05947037               no
2                   0.05868168               no
3                   0.05868168               no
4                   0.07237727               no
5                   0.05096696               no
6                   0.07159463               no
```

- Analysing the residuals and isolating influential outliers

Using the rstandard() function, the standardized residuals are accumulated, and it is identified that only 4.46% of observations are falling outside the 1.96 benchmark figure (Pregibon, 2013). The influential cases are determined using the cooks.distance() function.

```
> train$standardisedResiduals <- rstandard(multiplemodel6)
> sum(train$standardisedResiduals>1.96) #4.46% lie oustide the defined range
[1] 1341

> train$cook <- cooks.distance(multiplemodel6)
> sum(train$cook>1)
[1] 0
```

- Multicolinearity

Vif() function calculates the degree of correlation between the independent variables in the model and helps to identify which variables may be causing issues with the model. It is important to check for multicollinearity as it can lead to inaccurate results and bias (Glen, 2020).

To avoid multicollinearity in the model, a benchmark ‘gvif’ value of 10 is set to assess the outcomes. With the ‘GVIF’ of over 10 in emp.var.rate, nr.employed and euribor3m variables indicated the collinearity between the variables and ideally should not used for the creating the model.

```
> vif(multiplemodel6)
                    GVIF Df GVIF^(1/(2*Df))
poutcome        4.630549  2        1.466926
emp.var.rate   28.486350  1        5.337261
previous        4.636432  1        2.153238
job             3.154327 10        1.059120
campaign        1.024355  1        1.012104
cons.price.idx  9.379705  1        3.062630
education       2.934519  6        1.093859
cons.conf.idx   2.330577  1        1.526623
nr.employed    35.566272  1        5.963746
euribor3m      62.967120  1        7.935182
```

- Linearity of logit

The linearity assumption check for the logistic regression model in R is used to ensure that the relationship between the independent and dependent variables is linear. This assumption is important to check as it helps to ensure that the model is working accurately and producing valid results. If the assumption is violated, the results of the model may be biased and inaccurate (Yang et al., 2019).

To do this, only for continuous variables, the model is run including predictors that are the interaction between each predictor and log of itself. Since, some of the variables had negative value, thus log of absolute value is taken.

```
Call:
glm(formula = formula7, family = "binomial", data = abt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0416  -0.4774  -0.3787  -0.1946   2.8396

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.009e+04  1.009e+04  -1.991   0.0465 *
poutcomesuccess               1.745e+00  8.918e-02  19.571   <2e-16 ***
emp.var.rate                  3.740e+00  4.400e+00   0.850   0.3954
euribor3m                     2.441e+00  1.383e+00   1.765   0.0776 .
nr.employed                   4.708e+01  2.228e+01   2.113   0.0346 *
previous                      1.184e+00  5.173e-01   2.290   0.0220 *
jobblue-collar               -9.328e-02  1.677e-01  -0.556   0.5781
jobentrepreneur              -2.634e-01  2.885e-01  -0.913   0.3612
jobhousemaid                  7.691e-02  3.049e-01   0.252   0.8008
jobmanagement                 2.469e-02  1.780e-01   0.139   0.8897
jobretired                    3.333e-01  1.785e-01   1.867   0.0619 .
jobself-employed             -1.872e-01  2.587e-01  -0.724   0.4693
jobservices                  -1.090e-01  1.769e-01  -0.616   0.5376
jobstudent                    7.267e-02  1.956e-01   0.372   0.7103
jobtechnician                 2.099e-01  1.443e-01   1.455   0.1458
jobunemployed                 3.051e-01  2.401e-01   1.271   0.2038
campaign                      2.191e-01  1.953e-01   1.122   0.2619
cons.price.idx               -2.835e+02  1.835e+02  -1.545   0.1223
educationbasic.6y            -2.876e-02  2.620e-01  -0.110   0.9126
educationbasic.9y            -1.743e-01  2.019e-01  -0.863   0.3879
educationhigh.school         -6.548e-02  1.866e-01  -0.351   0.7256
educationilliterate           1.521e+00  2.361e+00   0.644   0.5196
educationprofessional.course  1.691e-01  2.039e-01   0.829   0.4068
educationuniversity.degree    2.327e-02  1.877e-01   0.124   0.9013
cons.conf.idx                -7.023e-01  1.068e+00  -0.657   0.5109
empvarrateLogInt             -1.808e+00  2.205e+00  -0.820   0.4122
euribor3mLogInt              -1.664e-01  7.760e-01  -0.214   0.8302
nr.employedLogInt            -4.944e+00  2.339e+00  -2.114   0.0345 *
previousLogInt               -6.646e-01  3.089e-01  -2.152   0.0314 *
campaignLogInt               -1.470e-01  9.519e-02  -1.544   0.1226
cons.conf.idxLogInt           1.635e-01  2.239e-01   0.730   0.4652
cons.price.idxLogInt          5.124e+01  3.310e+01   1.548   0.1217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

None of the logInt variables are showing as significant. Thus, the assumption is preserved.

Conclusion

To interpret the model and evaluate the coefficients of the model we use the Odds Ratio. These can be calculated using the exp() function on the model’s coefficients. If the value is lower than 1 then there is inverse relationship between the target and independent variables and vice-versa.

Hypothesis 1: The 0.99 value for Nr.employed indicated that a unit increase in it will lead to 0.99 decrease in the odds of subscription to term deposit.

Hypothesis 2: A unit change in Euribor3m will result in 1.18 increase in the odds of user's subscribing to term deposits.

Hypothesis 3: A unit change in Emp.Var.Rate will lead to 0.57 decrease in the odds of subscription.

Hypothesis 4: Customer's with job roles like Retired, Student, Technician and Self Employed are likely to subscribed to terms deposits by the odds of 1.48, 1.24, 1.04 and 1.02 respectively. While other job roles like Unemployed, Management, Housemaid, Services, Blue Collar and Entrepreneur are not like to subscribe to terms deposits with the odds of 0.99, 0.90, 0.85, 0.80, 0.79 and 0.74 respectively.

Additionally, variables like Consumer Confidence Index, Consumer Price Index, Previous and Poutcome and Education levels like Illiterate, University Degree, Professional Course, High School and Basic 6y are observed to have positive odds of subscription to term deposit. Also, inferring from the assumption checks, the model is violating the assumption of multicollinearity.

```
> round(exp(multiplemodel6$coefficients), 2)
                (Intercept)          poutcomenonexistent                poutcomesuccess
                       0.00                         2.18                           6.16
              emp.var.rate                     previous                jobblue-collar
                       0.57                         1.23                           0.79
          jobentrepreneur                  jobhousemaid                 jobmanagement
                       0.74                         0.85                           0.90
               jobretired             jobself-employed                   jobservices
                       1.48                         1.02                           0.80
               jobstudent                jobtechnician                 jobunemployed
                       1.24                         1.04                           0.99
                 campaign              cons.price.idx             educationbasic.6y
                       0.95                         1.66                           1.05
        educationbasic.9y         educationhigh.school          educationilliterate
                       0.90                         1.05                           2.60
educationprofessional.course   educationuniversity.degree                cons.conf.idx
                       1.04                         1.22                           1.02
              nr.employed                     euribor3m
                       0.99                         1.18
```

Reflective Summary

After receiving the positive feedback on the linear regression model, I was confident to undertake the tasks of this assignment more confidently. However, I found that it is different from more ways than one. Interpreting logistic regression model is little difficult along with interpreting the outcomes of the assumption checks, particularly the logit linearity test. Since, the dependent variable is categorical interpreting associations from graph and correlations were different and not easy. In the process of referring to the wider literature, I realised that logistic regression model has more real world application and now that the tasks are carried out in best of my efforts, I feel a little confident in handling it in future.

References

Field, A. (2020) “8.2.4,” in Discovering statistics using IBM SPSS statistics. London, United Kingdom: Sage, pp. 300–303.

Field, A., Miles, J. and Field Zoë (2022) “7.9,” in Discovering statistics using R. Thousand Oaks, California: SAGE/Texts, pp. 287–298.

Glen, S. (2020) Variance inflation factor, Statistics How To. Available at: https://www.statisticshowto.com/variance-inflation-factor/ (Accessed: December 21, 2022).

James, G. et al. (2022) An introduction to statistical learning: With applications in R. Boston, Massachusetts: Springer.

Golecha, Y.S. (2017) Analyzing term deposits in banking sector by performing predictive ..., Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques. Available at: https://trap.ncirl.ie/3100/1/yogeshsanjaygolecha.pdf (Accessed: January 6, 2023).

Borugadda, P., Nandru, D.P. and Madhavaiah, D.C. (2021) Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning Algorithms, View of predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. Available at: https://journal.stic.ac.th/index.php/sjhs/article/view/296/85 (Accessed: January 6, 2023).

Hou, S. et al. (2022) Applying machine learning to the development of prediction models for Bank Deposit Subscription, International Journal of Business Analytics (IJBAN). IGI Global. Available at: https://www.igi-global.com/article/applying-machine-learning-to-the-development-of-prediction-models-for-bank-deposit-subscription/288514 (Accessed: January 6, 2023).

Ilham, A. et al. (2019) Long-term deposits prediction: A comparative framework of ..., Long-term deposits prediction. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1175/1/012035 (Accessed: January 6, 2023).

Yang, C. (2016) Predicting success of bank telemarketing with classification trees and logistic regression, TexasScholarWorks. Available at: https://repositories.lib.utexas.edu/handle/2152/41744?show=full (Accessed: January 6, 2023).

Zhuang, Q.R., Yao, Y.W. and Liu, O. (2018) Application of data mining in term deposit marketing, tion of data mining in term deposit marketing. Available at: http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp707-710.pdf (Accessed: January 6, 2023).

Tekouabou, S.C.K., Cherif, W. and Silkan, H. (2019) A data modeling approach for classification problems: Proceedings of the 2nd International Conference on Networking, Information Systems & Security, ACM Other conferences. Available at: https://dl.acm.org/doi/abs/10.1145/3320326.3320389?casa_token=xTP4YfHzLDEAAAAA%3ANIYmevqkPDJL7EOkrlUDZNqVAVTE6Bxngm99jhluZTcvyF3RC0GNeAzOt4kzCc813QmxG7rcTvkR (Accessed: January 6, 2023).

Rony, M.A.T. et al. (2021) Identifying long-term deposit customers: A machine ... - IEEE xplore, Identifying Long-Term Deposit Customers: A Machine Learning Approach. Available at: https://ieeexplore.ieee.org/abstract/document/9672452/ (Accessed: January 6, 2023).

Moroa, S. et al. (2014) A data-driven approach to predict the success of bank telemarketing, Decision Support Systems. North-Holland. Available at: https://www.sciencedirect.com/science/article/pii/S016792361400061X?casa_token=uMtEtg82m7MAAAAA%3Af17J2db1CbfXk6NXSDrw4q6ok76M6e04rRgDD8aa-AeicrB7iCM7D54iefoIMqdxNrpzN57ynw (Accessed: January 6, 2023).

Desai, R. and Khairnar, V. (1970) Hybrid prediction model for the success of Bank Telemarketing, SpringerLink. Springer Singapore. Available at: https://link.springer.com/chapter/10.1007/978-981-16-2422-3_54 (Accessed: January 6, 2023).

Chen, J. et al. (2014) Who will subscribe a term deposit? - columbia university, Who will subscribe a term deposit? Available at: http://www.columbia.edu/~jc4133/ADA-Project.pdf (Accessed: January 6, 2023).

Abbas, S. (2015) Deposit subscribe prediction using data mining techniques based ... - arxiv, Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset. Available at: https://arxiv.org/pdf/1503.04344.pdf (Accessed: January 6, 2023).

Jiang, W. et al. (2015) Simulating urban land use change by incorporating an autologistic regression model into a clue-S model - journal of geographical sciences, SpringerLink. Science Press. Available at: https://link.springer.com/article/10.1007/s11442-015-1205-8 (Accessed: January 6, 2023).

Sperandei, S. (2018) Understanding logistic regression analysis, Biochemia Medica. Croatian Society of Medical Biochemistry and Laboratory Medicine. Available at: https://www.biochemia-medica.com/en/journal/24/1/10.11613/BM.2014.003 (Accessed: January 6, 2023).

Pregibon, D. (2013) Logistic regression diagnostics, Project Euclid. Institute of Mathematical Statistics. Available at:

https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-4/Logistic-Regression-Diagnostics/10.1214/aos/1176345513.full (Accessed: January 6, 2023).

Aussalet, E.B. and Hardman, L. (2010) Using the confusion matrix for improving ensemble classifiers | IEEE ..., Visualization of Confusion Matrix for Non-Expert Users. Available at: https://ieeexplore.ieee.org/abstract/document/5662159 (Accessed: January 6, 2023).

Appendix

```
#Set WD

setwd("C:/Users/Prayas Sachdeva/Downloads")


#Install and Load the Packages

install.packages('readr')

install.packages('caret')

install.packages('ggplot2')

install.packages('dplyr')

install.packages('psych')

install.packages('DescTools')

install.packages('car')

library(readxl)

library(psych)

library(caret)

library(ggplot2)

library(dplyr)

library(DescTools)

library(car)


#Read the data
```

```
bank <- read_excel("C:/Users/Prayas Sachdeva/Downloads/banksv.xlsx")

summary(bank)


#Creating Analytics Base Table - Using the select() function to pick specific variables from the bank dataset

abt <- bank %>%

  select(age, job, marital, education, default,housing,loan, campaign, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, subscribed)


# data quality issues

glimpse(abt) #glimpse() function helps displays a concise summary of the data frame including the number of rows, number of columns, column names, column classes, and the first few rows of data.

summary(abt)

describe(abt) #describe() function displays the descriptive statistics of  the data


#Subscribed

levels(as.factor(abt$subscribed))


#Age

boxplot(abt$age)

hist(abt$age)

hist(abt$age[abt$age>85])

hist(abt$age[abt$age<18])

sum(abt$age>85 | abt$age<18)

abt$age[abt$age<18 | abt$age>85] <- NA


#Job
```

```
summary(as.factor(abt$job))

abt$job[abt$job=='unknown'] <- NA

levels(abt$job)


#Marital

summary(as.factor(abt$marital))

abt$marital[abt$marital=='unknown'] <- NA


#Education

summary(as.factor(abt$education))

abt$education[abt$education=='unknown'] <- NA


#Default

summary(as.factor(abt$default))

abt$default[abt$default=='n'] <- 'no'

abt$default[abt$default=='unknown'] <- 'no' #on the basis of mode imputation method


#Housing

summary(as.factor(abt$default))


#Loan

summary((as.factor(abt$loan)))

abt$loan[abt$loan=='unknown'] <- NA


#Campaign

boxplot(abt$campaign)
```

```
hist(abt$campaign)

hist(abt$campaign[abt$campaign>40])

sum(abt$campaign>40)

abt$campaign[abt$campaign>40] <- NA


#Previous

boxplot(abt$previous)

hist(abt$previous)

sum(abt$previous>4)

abt$previous[abt$previous>4] <- NA


#Poutcome

summary(as.factor(abt$poutcome))


#Employment Variation Rate

boxplot(abt$emp.var.rate)

hist(abt$emp.var.rate)

summary(as.factor(abt$emp.var.rate))


#Consumer Price Index

boxplot(abt$cons.price.idx)

hist(abt$cons.price.idx)

summary(as.factor(abt$cons.price.idx))


#Consumer Confidence Index

boxplot(abt$cons.conf.idx)
```

```
hist(abt$cons.conf.idx)

hist(abt$cons.conf.idx[abt$cons.conf.idx> -30])

sum(abt$cons.conf.idx> -30)

abt$cons.conf.idx[abt$cons.conf.idx> -30] <- NA


#Euribor 3 months rate

boxplot(abt$euribor3m)

hist(abt$euribor3m)

summary(as.factor(abt$euribor3m))


#Number of employees

boxplot(abt$nr.employed)

hist(abt$nr.employed)


#Dealing with NA

colSums(is.na(abt))

abt <- na.omit(abt)


#Converting Data Types

abt <- abt %>%

  mutate_if(is.character, as.factor)


#Test of association

cor.test(as.numeric(abt$subscribed), abt$age) #cor.test() function provides the p-value whereas cor()
doesn't
```

cor(as.numeric(abt$subscribed), abt$age)#0.0209743 #pearson method ideal for ordinal and numerical variable

#The test results indicate that there is a statistically significant correlation between these two variables, as the p-value is less than 0.05. The sample estimate of the correlation is 0.0209743, and the 95% confidence interval is between 0.01085709 and 0.03108721. This suggests that there is a positive correlation between the two variables, meaning that as age increases, subscribed is more likely to increase.

chisq.test(abt$subscribed, abt$job)#x-squared 673.26 #ideal for two factor variables

cor.test(as.numeric(abt$subscribed), as.numeric(abt$job))#0.01778499

#The test statistic is 3.4451, which suggests a statistically significant correlation at the 0.0005715 level. The 95% confidence interval for the correlation is 0.007666852 to 0.027899483. This suggests that there is a small, positive relationship between the two variables.

chisq.test(abt$subscribed, abt$marital)#x-squared 97.856

#The test results show that there is a statistically significant difference in subscription rate between the different marital statuses, as indicated by the X-squared value of 97.856 and the p-value of less than 2.2e-16.

chisq.test(abt$subscribed, abt$education)#x-squared 157.03

cor.test(as.numeric(abt$subscribed), as.numeric(abt$education)) #0.05533752

#The test results indicate that there is a statistically significant positive correlation between the two variables. The p-value of less than 2.2e-16 indicates that the correlation is highly significant. The 95% confidence interval is 0.04524334 to 0.06542040, meaning that there is 95% certainty that the true correlation lies between these two numbers. The sample estimates of the correlation are 0.05533752.

cor.test(as.numeric(abt$subscribed), as.numeric(abt$default), method = "spearman")

cor(as.numeric(abt$subscribed), as.numeric(abt$default), method = "spearman")

cor(as.numeric(abt$subscribed), as.numeric(abt$default))

cor.test(as.numeric(abt$subscribed), as.numeric(abt$default))#-0.003076673

#This indicates that there is a weak negative relationship between the two variables, but it is not statistically significant because the p-value is 0.5513, which is greater than the significance level of 0.05. The 95% confidence interval for the correlation was between -0.013195778 and 0.007043062, meaning that the correlation could range from a slight positive to a slight negative relationship.

cor(as.numeric(abt$subscribed), as.numeric(abt$housing), method = "spearman")

cor(as.numeric(abt$subscribed), as.numeric(abt$housing))

cor.test(as.numeric(abt$subscribed), as.numeric(abt$housing), method = "spearman") #0.0110858

#Spearman's rank correlation rho is a statistical measure of the strength and direction of a monotonic relationship between two variables. It is calculated by measuring the ranks of the values of each variable. The result of the calculation is a value between -1 and 1. In this case, the Spearman's rank correlation rho between the variables as.numeric(abt$subscribed) and as.numeric(abt$housing) is 0.0110858, which indicates a very weak positive relationship between the two variables. The p-value of 0.03178 indicates that the correlation is statistically significant.

cor(as.numeric(abt$subscribed), as.numeric(abt$loan), method = "spearman")

cor.test(as.numeric(abt$subscribed), as.numeric(abt$loan), method = "spearman") #-0.003066862

#Spearman's rank correlation rho is a statistical measure of the strength and direction of the relationship between two variables. It is calculated by taking the difference in the rankings of two variables and then squaring the result and summing the squared differences. In this case, the rho value of -0.003066862 indicates a weak negative relationship between the variables as.numeric(abt$subscribed) and as.numeric(abt$loan). The p-value of 0.5525 suggests that this relationship is not statistically significant.

cor(as.numeric(abt$subscribed), abt$campaign)

cor.test(as.numeric(abt$subscribed), abt$campaign) #-0.0609361

#Pearson's product-moment correlation is a statistical technique used to measure the strength and direction of the linear relationship between two variables. In this case, the correlation between the variables as.numeric(abt$subscribed) and abt$campaign is -0.0609361. This value indicates a weak negative linear relationship between the two variables, meaning that as the value of

as.numeric(abt$subscribed) increases, the value of abt$campaign decreases. The t-value of -11.824 and the p-value of < 2.2e-16 suggests that the correlation is statistically significant, and the 95% confidence interval of -0.07101183 to -0.05084794 further supports this conclusion.

```
cor(as.numeric(abt$subscribed), abt$previous)

cor.test(as.numeric(abt$subscribed), abt$previous) #0.2157882
```

#The t-value is 42.802, the degrees of freedom is 37511, and the p-value is less than 2.2e-16. This indicates that there is a significant correlation between the two variables and that the true correlation is not equal to 0. The 95 percent confidence interval for the correlation is 0.2061188 to 0.2254155.

```
cor(as.numeric(abt$subscribed), as.numeric(abt$poutcome), method = "spearman")

cor.test(as.numeric(abt$subscribed), as.numeric(abt$poutcome), method = "spearman")

cor.test(as.numeric(abt$subscribed), as.numeric(abt$poutcome)) #0.1256268
```

The t-statistic is 24.525, and the p-value is less than 2.2e-16, indicating a statistically significant relationship between the two variables. The 95% confidence interval is between 0.1156543 and 0.1355739. This suggests that there is a positive correlation between the two variables, indicating that as one variable increases, the other one is likely to increase as well.

```
cor(as.numeric(abt$subscribed), abt$emp.var.rate)

cor.test(as.numeric(abt$subscribed), abt$emp.var.rate) #-0.2750085
```

The given values indicate that there is a strong negative linear correlation between the variables "subscribed" (as.numeric) and "emp.var.rate" in the given dataset. The correlation coefficient is -0.275 and the p-value is less than 2.2e-16, which is highly significant. The 95% confidence interval of the correlation coefficient is between -0.284 and -0.265, indicating that the correlation is highly significant.

```
cor(as.numeric(abt$subscribed), abt$cons.price.idx)

cor.test(as.numeric(abt$subscribed), abt$cons.price.idx) #-0.1085067
```

#The correlation is -0.1085067, which indicates that there is a negative correlation between the two variables. The t-value of -21.14 indicates that the correlation is statistically significant, and the p-value of <2.2e-16 indicates that it is highly significant. The 95% confidence interval is -0.11849614 to -0.09849537, indicating that the correlation is unlikely to be outside of this range.

```
cor(as.numeric(abt$subscribed), abt$cons.conf.idx)

cor.test(as.numeric(abt$subscribed), abt$cons.conf.idx) #0.01039025
```

The test results indicate that there is a statistically significant correlation between the two variables, as the p-value is less than 0.05. The correlation coefficient is 0.01039025, indicating that the variables are weakly correlated. The 95% confidence interval is 0.0002707637 - 0.0205076107, indicating that the true correlation lies within that range.

```
cor(as.numeric(abt$subscribed), abt$euribor3m)

cor.test(as.numeric(abt$subscribed), abt$euribor3m) #-0.2872919
```

#The test results show that there is a significant negative correlation between the two variables, with a correlation coefficient of -0.2872919 and a p-value of less than 2.2e-16. This indicates that as one variable increases, the other decreases. The 95% confidence interval for the correlation is between -0.2965492 and -0.2779805.

```
cor(as.numeric(abt$subscribed), abt$nr.employed)

cor.test(as.numeric(abt$subscribed), abt$nr.employed) #-0.3344536
```

#The test statistic is -68.734 and the p-value is less than 2.2e-16, indicating a strong negative correlation between the two variables. The 95% confidence interval for the correlation is between -0.3434109 and -0.3254356, and the sample estimate is -0.3344536. This suggests that there is a strong negative correlation between the two variables.

```
# Visualisations

ggplot(abt,aes(x = nr.employed, y= subscribed, color = subscribed)) +

  geom_bin2d()
```

```
ggplot(abt,aes(x = subscribed, y= euribor3m, color = subscribed)) +
  geom_jitter()

ggplot(abt,aes(x = education, y= subscribed, color = education)) +
  geom_col()

ggplot(abt,aes(x = subscribed, y= emp.var.rate, color = subscribed)) +
  geom_col()

ggplot(abt,aes(x = poutcome, y= subscribed, color = poutcome)) +
  geom_jitter()

ggplot(abt,aes(x = subscribed, y= previous, color = subscribed)) +
  geom_bar(stat = "identity")

ggplot(abt,aes(x = job, y= subscribed, color = subscribed)) +
  geom_col()

ggplot(abt,aes(x = subscribed, y= cons.price.idx, color = subscribed)) +
  geom_bar(stat = "identity")

#Splitting Data intro Train(80) and Test(20) datasets
set.seed(40386053)
index <- createDataPartition(abt$subscribed, p = 0.8, list = FALSE, times = 1)
train <- abt[index,]
```

```
test <- abt[-index,]


#Simple Regrssion Models
simplemodel1 <- glm(abt$subscribed~abt$age, data = train, family = "binomial")
simplemodel2 <- glm(abt$subscribed~abt$job, data = train, family = "binomial")
simplemodel3 <- glm(abt$subscribed~abt$marital, data = train, family = "binomial")
simplemodel4 <- glm(abt$subscribed~abt$education, data = train, family = "binomial")
simplemodel5 <- glm(abt$subscribed~abt$default, data = train, family = "binomial")
simplemodel6 <- glm(abt$subscribed~abt$housing, data = train, family = "binomial")
simplemodel7 <- glm(abt$subscribed~abt$loan, data = train, family = "binomial")
simplemodel8 <- glm(abt$subscribed~abt$campaign, data = train, family = "binomial")
simplemodel9 <- glm(abt$subscribed~abt$previous, data = train, family = "binomial")
simplemodel10 <- glm(abt$subscribed~abt$poutcome, data = train, family = "binomial")
simplemodel11 <- glm(abt$subscribed~abt$emp.var.rate, data = train, family = "binomial")
simplemodel12 <- glm(abt$subscribed~abt$cons.price.idx, data = train, family = "binomial")
simplemodel13 <- glm(abt$subscribed~abt$cons.conf.idx, data = train, family = "binomial")
simplemodel14 <- glm(abt$subscribed~abt$euribor3m, data = train, family = "binomial")
simplemodel15 <- glm(abt$subscribed~abt$nr.employed, data = train, family = "binomial")


#Assessing Simple regression models
summary(simplemodel1) #AIC: 25325
```

#This analysis investigates the relationship between subscription and age in a dataset. The deviance residuals show that the model is a good fit. The coefficients estimate suggests that there is a positive relationship between age and subscription, and this is supported by the p-value of 4.88e-05, which is highly significant. The null deviance is 25337 and the residual deviance is 25321, showing that the model has reduced the deviance by 16. The AIC score is 25325, indicating that this is a good model. The number of fisher scoring iterations is 4.

summary(simplemodel2) #AIC: 24787

#The results of the logistic regression indicate that different job categories have different effects on the likelihood of an individual subscribing to a service. For example, individuals with a job category of "Blue-Collar" had a log odds of subscribing to the service that was 0.637 lower than individuals in the reference group (Intercept). This suggests that individuals with a job of "Blue-Collar" are less likely to subscribe to the service than individuals in the reference group. Additionally, individuals with a job category of "Retired" had a log odds of subscribing to the service that was 0.767 higher than individuals in the reference group (Intercept). This suggests that individuals with a job of "Retired" are more likely to subscribe to the service than individuals in the reference group.

summary(simplemodel3) #AIC: 25248

#The analysis of the data shows that the subscription rate is affected by marital status. The null deviance is 25337 on 37512 degrees of freedom and the residual deviance is 25242 on 37510 degrees of freedom. There is a significant difference between the two, indicating that the marital status is an important factor in the subscription rate. The coefficient estimate for married is -0.03629 and for single is 0.31860, showing that single people are more likely to subscribe than married people. The AIC value is 25248, indicating that the model is a good fit for the data.

summary(simplemodel4) #AIC: 25191

#This analysis investigates the relationship between the 'subscribed' outcome and the 'education' level of the respondents in the 'abt' dataset. The results indicate that there is a statistically significant relationship between the two variables. The coefficients suggest that respondents with basic education (6y and 9y) have a lower probability of subscribing to the product compared to the illiterate respondents, while those with a high school and professional course education have a slightly higher probability of subscribing. Respondents with a university degree have the highest probability of subscribing. The Null Deviance and Residual Deviance are both relatively high, suggesting that the model is not a good fit for the data.

summary(simplemodel5) #AIC: 25340

#The results of the Generalized Linear Model (GLM) indicate that the variable 'defaultyes' has no statistically significant effect on the variable 'subscribed' with a p-value of 0.934. This is also seen

from the Estimate and Std. Error values for the variable 'defaultyes' which are -9.43208 and 113.71934 respectively. The Null Deviance and Residual Deviance values of 25337 and 25336 on 37512 and 37511 degrees of freedom respectively also indicate that the model is successfully able to capture the variance in the data. The AIC value of 25340 further confirms this.

summary(simplemodel6) #AIC: 25336

#This analysis suggests that housing is a significant predictor of whether or not an individual is subscribed. The coefficient for abt$housingyes was 0.07245, with a standard error of 0.03375 and a z value of 2.147, indicating that the predictor is significant (Pr(>|z|)= 0.0318). The deviance residuals ranged from -0.4804 to 2.1360, with a null deviance of 25337 and a residual deviance of 25332. This suggests that the model fits the data well. The AIC was 25336, indicating a good model fit. The number of Fisher Scoring iterations was 4.

summary(simplemodel7) #AIC: 25340

#This analysis looks at the relationship between the abt$subscribed variable and the abt$loan variable. The null deviance of 25337 on 37512 degrees of freedom indicates that the abt$subscribed variable is not significantly related to the abt$loan variable. The residual deviance of 25336 on 37511 degrees of freedom further supports this conclusion. The AIC of 25340 suggests that the model is a good fit to the data. Additionally, the z-value of -0.594 and the p-value of 0.553 indicate that the abt$loan variable is not statistically significant in predicting the abt$subscribed variable. This means that the abt$loan variable does not have a significant effect on the abt$subscribed variable.

summary(simplemodel8) #AIC: 25154

#This analysis investigated the relationship between the variable 'subscribed' and 'campaign' in the dataset 'abt' using a Generalized Linear Model (GLM). The Deviance Residuals indicated that the model fit the data well, with a minimum value of -0.5076, a 1st quartile of -0.5076, a median of -0.4812, a 3rd quartile of -0.4320, and a maximum value of 2.9977. The Coefficients showed that the 'campaign' variable was significantly associated with the 'subscribed' variable, with a Estimate of -0.113526, a Std. Error of 0.009556, and a z-value of -11.88 (all of which were significant at $p<0.001$). The dispersion parameter for the binomial family was taken to be 1. The null deviance was 25337 on 37512 degrees of freedom, while the Residual deviance was 25150 on 37511 degrees of freedom. The AIC was 25154 and the number of Fisher Scoring iterations was 5. Overall, this analysis showed that the 'campaign' variable had a significant association with the 'subscribed' variable.

summary(simplemodel9) #AIC: 24125

#This analysis used logistic regression to model the likelihood of a customer subscribing to a service based on their previous subscription status. The results of the analysis show that the estimated coefficient of the previous subscription status is 0.956, indicating that those who previously subscribed to the service are more likely to subscribe again. The analysis also showed that the null deviance was 25,337 and the residual deviance was 24,121. The AIC was 24,125 with five Fisher Scoring iterations. These results indicate that the model is a good fit.

summary(simplemodel10) #AIC: 23326

#This analysis evaluated the relationship between the variable 'subscribed' and 'poutcome' in the dataset 'abt'. The Deviance Residuals indicate that the values range from -1.4301 to 2.2202. The Coefficients reveal that the 'poutcomenonexistent' had a negative Estimate of -0.45298 and the 'poutcomesuccess' had a positive Estimate of 2.49979. The results indicate that the 'poutcomenonexistent' was significantly associated with a decrease in the 'subscribed' variable, while the 'poutcomesuccess' was significantly associated with an increase in the 'subscribed' variable. The Null Deviance was 25337 and the Residual Deviance was 23320, indicating that the model was able to explain most of the variance in the data. The AIC was 23326 and the Number of Fisher Scoring iterations was 5.

summary(simplemodel11) #AIC: 22713

#The above glm() command fits a logistic regression model to the data in 'train' with the variable abt$subscribed as the response variable and abt$emp.var.rate as the predictor variable. The deviance residuals range from -0.9855 to 2.5115, with a median of -0.3202. This indicates that the model is a good fit for the data. The coefficient for the predictor variable is -0.55007 with a standard error of 0.01122, and a z-value of -49.01, which is highly significant ($p < 2e-16$). The null deviance is 25337 on 37512 degrees of freedom and the residual deviance is 22709 on 37511 degrees of freedom, suggesting that the model is a good fit for the data. The AIC is 22713 and the number of Fisher Scoring iterations is 5. According to the logistic regression model, a one unit increase in the employee variation rate is associated with a 0.55 decrease in the predicted log-odds of subscribing to a term deposit.

summary(simplemodel12) #AIC: 24902

#This interpretation is based on the results of a Generalized Linear Model (GLM) that was conducted to investigate the effect of the consumer price index (abt$cons.price.idx) on customer subscription (abt$subscribed) using data from a training set. The results indicated that the consumer price index (abt$cons.price.idx) had a statistically significant effect on customer subscription (abt$subscribed). Specifically, for every one-unit increase in the consumer price index (abt$cons.price.idx), customer subscription (abt$subscribed) is expected to decrease by 0.62594. The residual deviance was 24898 on 37511 degrees of freedom and the AIC was 24902. The model converged after 5 iterations of the Fisher Scoring algorithm. The results of the GLM indicate that as the consumer price index increases, the likelihood of a customer subscribing to a term deposit decreases.

summary(simplemodel13) #AIC: 25337

#This analysis was conducted to test the relationship between the dependent variable "Subscribed" and the independent variable "Consumer Confidence Index" in a binomial logistic regression model. The results showed that the Intercept was -1.818247, and the Estimate for the Consumer Confidence Index was 0.007760. This suggests that an increase in the Consumer Confidence Index would lead to a corresponding increase in the likelihood of a customer being subscribed. The Deviance Residuals ranged from -0.4917 to 2.1523 with a median of -0.4707, and the Null Deviance was 25337 with a Residual Deviance of 25333. The AIC was also 25337, and the analysis was completed after 4 iterations of Fisher Scoring. This suggests that the model was successful in predicting the likelihood of a customer being subscribed.

summary(simplemodel14) #AIC: 22481

#This analysis is examining the factors that influence the likelihood of a customer subscribing to a term deposit. The model used was a binomial generalized linear model with the dependent variable being the customer's subscription to the term deposit and the independent variable being the euribor3m rate. The results showed that there is a significant negative correlation between the euribor3m rate and the customer's subscription to the term deposit, with the coefficient being -0.51105. The null deviance was 25337 and the residual deviance was 22477, which indicates that the model is a good fit. The AIC score of 22481 also indicates that the model is accurate. The analysis concluded that the euribor3m rate is a significant factor when considering customer subscription to the term deposit.

summary(simplemodel15) #AIC: 21775

#This analysis is focused on the relationship between the number of employed and subscribed of a customer. The results of the model indicate a significant relationship between the two variables ($p < 2e-16$). The estimate for this relationship is -0.0130247, indicating that for each unit increase in the number of employed, the probability of the customer subscribing decreases by 0.0130247. The Null deviance for this model is 25337 on 37512 degrees of freedom, and the Residual deviance is 21771 on 37511 degrees of freedom. The AIC for this model is 21775. The model was run using 5 Fisher Scoring iterations.

#Accuracy Check for simple regression models

round(PseudoR2(simplemodel1, which = "all"), 2)

round(PseudoR2(simplemodel2, which = "all"), 2)

round(PseudoR2(simplemodel3, which = "all"), 2)

round(PseudoR2(simplemodel4, which = "all"), 2)

round(PseudoR2(simplemodel5, which = "all"), 2)

round(PseudoR2(simplemodel6, which = "all"), 2)

round(PseudoR2(simplemodel7, which = "all"), 2)

round(PseudoR2(simplemodel8, which = "all"), 2)

round(PseudoR2(simplemodel9, which = "all"), 2)

round(PseudoR2(simplemodel10, which = "all"), 2)

round(PseudoR2(simplemodel11, which = "all"), 2)

round(PseudoR2(simplemodel12, which = "all"), 2)

round(PseudoR2(simplemodel13, which = "all"), 2)

round(PseudoR2(simplemodel14, which = "all"), 2)

round(PseudoR2(simplemodel15, which = "all"), 2)

#Creating logisticPseudoR2s function for assessing the model

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")
}

logisticPseudoR2s(simplemodel1)
logisticPseudoR2s(simplemodel2)
logisticPseudoR2s(simplemodel3)
logisticPseudoR2s(simplemodel4)
logisticPseudoR2s(simplemodel5)
logisticPseudoR2s(simplemodel6)
logisticPseudoR2s(simplemodel7)
logisticPseudoR2s(simplemodel8)
logisticPseudoR2s(simplemodel9)
logisticPseudoR2s(simplemodel10)
logisticPseudoR2s(simplemodel11)
logisticPseudoR2s(simplemodel12)
logisticPseudoR2s(simplemodel13)
logisticPseudoR2s(simplemodel14)
```

```
logisticPseudoR2s(simplemodel15)

#predictions using simple model
simple_predictions <- predict(simplemodel15, test, type = "response")
simple_class_pred <- as.factor(ifelse(simple_predictions>0.5,"Yes","No"))
postResample(simple_class_pred, test$subscribed)

#Odds Ratio
exp(simplemodel1$coefficients)
exp(simplemodel2$coefficients)
exp(simplemodel3$coefficients)
exp(simplemodel4$coefficients)
exp(simplemodel5$coefficients)
exp(simplemodel6$coefficients)
exp(simplemodel7$coefficients)
exp(simplemodel8$coefficients)
exp(simplemodel9$coefficients)
exp(simplemodel10$coefficients)
exp(simplemodel11$coefficients)
exp(simplemodel12$coefficients)
exp(simplemodel13$coefficients)
exp(simplemodel14$coefficients)
exp(simplemodel15$coefficients)

#Multiple Regression Model
formula <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed
```

```
multiplemodel1 <- glm(formula = formula, family = "binomial", data = train)
summary(multiplemodel1) #AIC: 16820

formula2 <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed + previous
multiplemodel2 <- glm(formula = formula2, family = "binomial", data = train)
summary(multiplemodel2)

formula3 <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed + job
multiplemodel3 <- glm(formula = formula3, family = "binomial", data = train)
summary(multiplemodel3)

formula4 <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed + cons.price.idx
multiplemodel4 <- glm(formula = formula4, family = "binomial", data = train)
summary(multiplemodel4)

formula5 <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed + previous + job +
cons.price.idx
multiplemodel5 <- glm(formula = formula5, family = "binomial", data = train)
summary(multiplemodel5)

formula6 <- subscribed ~ poutcome + emp.var.rate + previous + job + campaign + cons.price.idx +
education + cons.conf.idx + nr.employed + euribor3m
multiplemodel6 <- glm(formula = formula6, family = "binomial", data = train)
summary(multiplemodel6)

#Accuracy Check for Multiple Regression Model
```

```
logisticPseudoR2s(multiplemodel1)

logisticPseudoR2s(multiplemodel2)

logisticPseudoR2s(multiplemodel3)

logisticPseudoR2s(multiplemodel4)

logisticPseudoR2s(multiplemodel5)

logisticPseudoR2s(multiplemodel6)

round(PseudoR2(multiplemodel6, which = 'all'), 2)


#Predictions using multiple regression model

predictions <- predict(multiplemodel6, test, type = "response")

class_pred <- as.factor(ifelse(predictions >0.5,"yes","no"))#converting predicted probabilites to Yes and No outcomes

postResample(class_pred, test$subscribed)

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


#Odds ratio

round(exp(multiplemodel6$coefficients), 2)


#confidence intervals

exp(confint(multiplemodel6))


#predicted probabilites

train$predictedprobabilities <- fitted(multiplemodel6)

head(data.frame(train$predictedprobabilities, train$subscribed))


#analysing the residuals
```

```
train$standardisedResiduals <- rstandard(multiplemodel6)

train$studentisedResiduals <- rstudent(multiplemodel6)

sum(train$standardisedResiduals>1.96) #4.46% lie oustide the defined range


#checking for influential outliers

train$cook <- cooks.distance(multiplemodel6)

sum(train$cook>1)


#Check for multicolinearity

vif(multiplemodel6)


#linearity of the logit check

abt$empvarrateLogInt <- log(abs(abt$emp.var.rate))*abt$emp.var.rate

abt$euribor3mLogInt <- log(abs(abt$euribor3m))*abt$euribor3m

abt$nr.employedLogInt <- log(abs(abt$nr.employed))*abt$nr.employed

abt$previousLogInt <- log(abs(abt$previous))*abt$previous

abt$campaignLogInt <- log(abs(abt$campaign))*abt$campaign

abt$cons.conf.idxLogInt <- log(abs(abt$cons.conf.idx))*abt$cons.conf.idx

abt$cons.price.idxLogInt <- log(abs(abt$cons.price.idx))*abt$cons.price.idx


formula7 <- subscribed ~ poutcome + emp.var.rate + euribor3m + nr.employed + previous + job +
campaign + cons.price.idx + education + cons.conf.idx + empvarrateLogInt + euribor3mLogInt +
nr.employedLogInt + previousLogInt + campaignLogInt + cons.conf.idxLogInt +
cons.price.idxLogInt

multiplemodel7 <- glm(formula7, data = abt, family = "binomial")

summary(multiplemodel7)
```

#residual plots for linearity check

plot(multiplemodel6, test$subscribed, which = 1)

plot(multiplemodel6, test$subscribed, which = 2)

plot(multiplemodel6, test$subscribed, which = 3)

plot(multiplemodel6, test$subscribed, which = 4)

plot(multiplemodel6, test$subscribed, which = 5)

plot(multiplemodel6, test$subscribed, which = 6)

hist(resid(multiplemodel6))

- Residuals plots

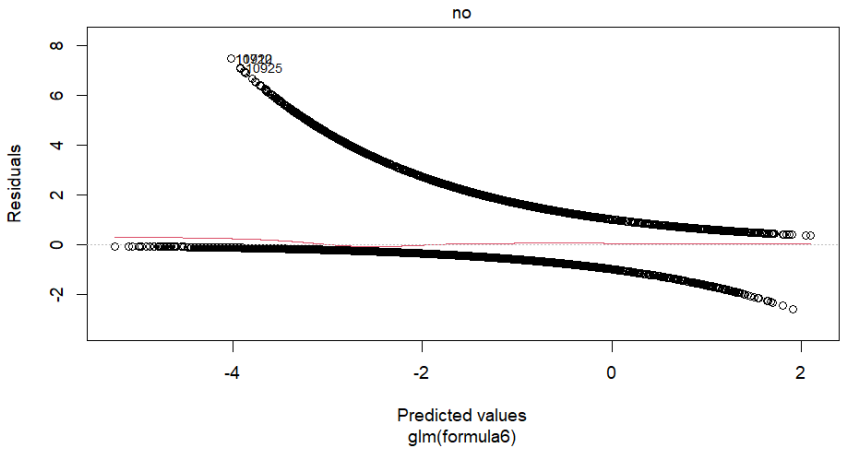

Figure 17 Predicted Values Vs Residuals

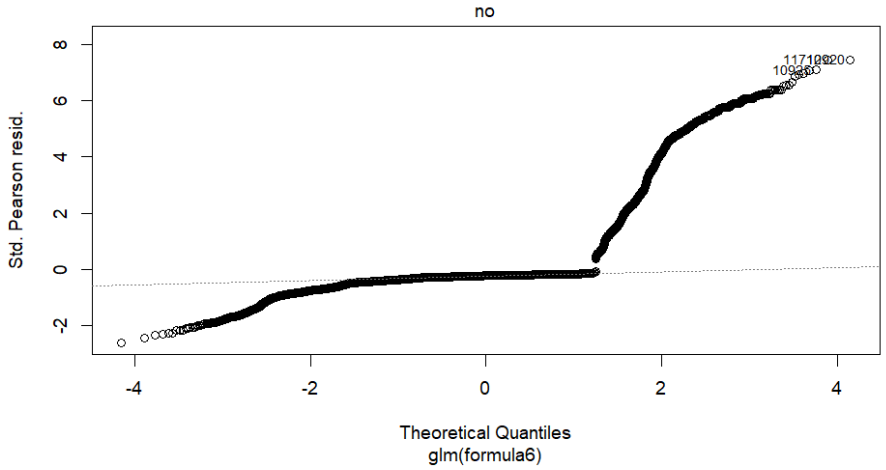

Figure 18 Std. Pearson Resid. Vs Theoretical Quantiles

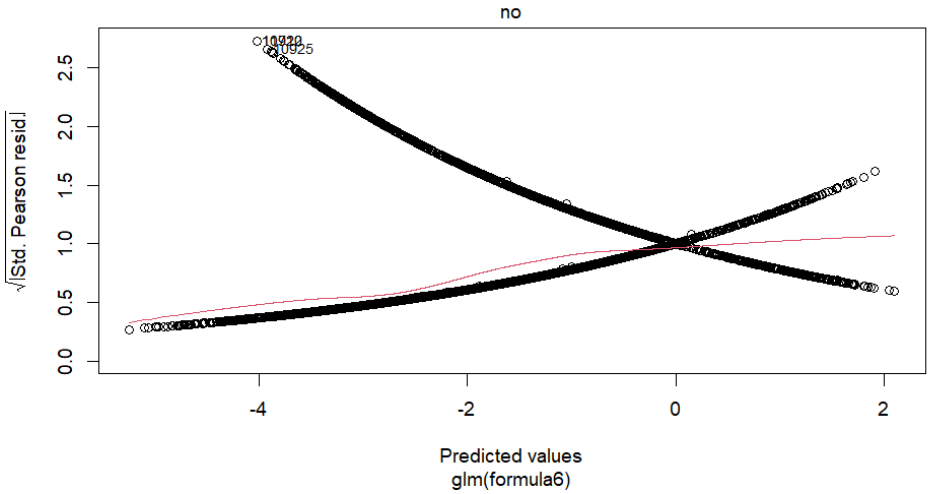

Figure 19 Squared root Std. Pearson Resid. Vs Predicted Values

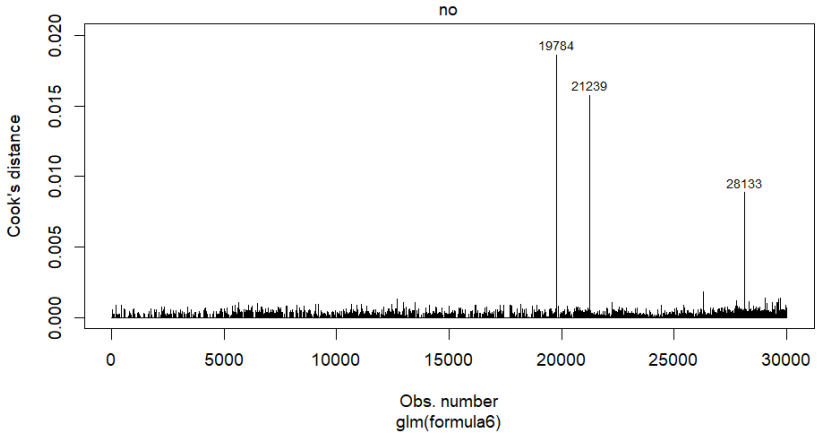

Figure 20 Cook's Distance and Observation Numbers

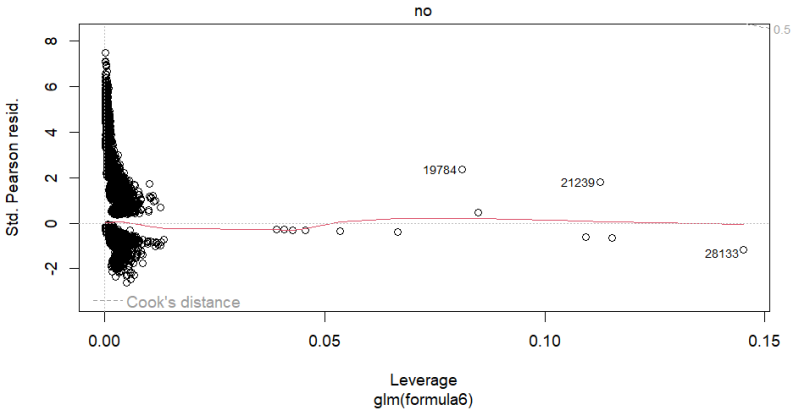

Figure 21 Std. Pearson Resid and Leverage

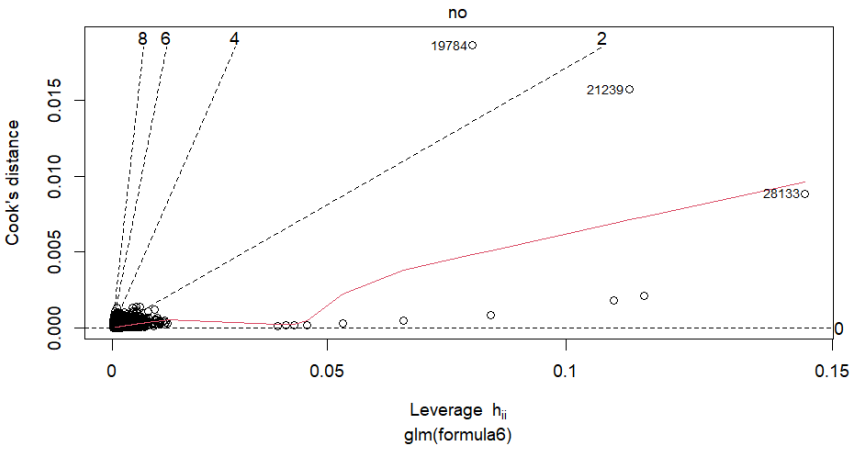

Figure 22 Cook's Distance and Leverage

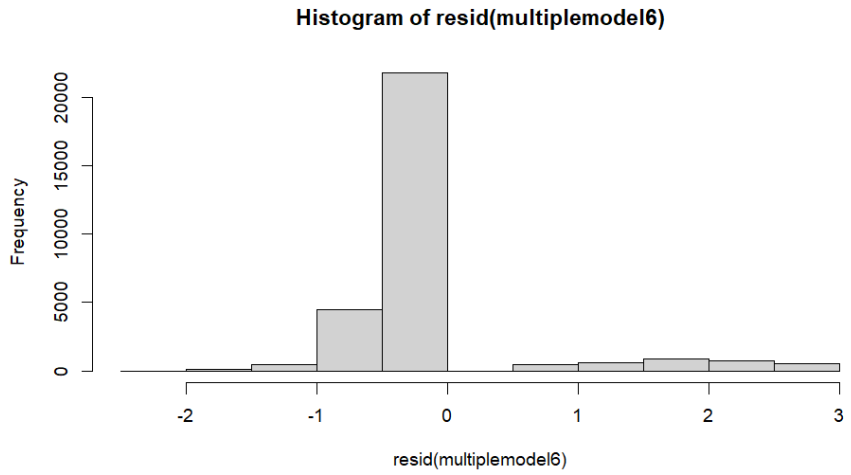

Figure 23 Histogram of Residuals