Logistic regression model

In our binary classification problem, we only want the output to represent probabilities between 0 and 1. The logistic regression model utilizes the logistic function in order to squeeze the output of a linear equation between 0 and 1. Specifically, it takes the right side of the linear regression model and wraps it into the logistic function. Below, we have showcased these formulas and functions between the outcome (y) and features (x) for each instance (i), along with a visual illustration of the logistic function shown in Figure 1.

Logistic function $$logistic(x) = \frac{1}{1 + e^{x}}$$

Linear regression model $$\hat{y}^{(i)} = \beta_0 + \beta_1 {x_1}^{(i)} + \ldots + \beta_p {x_p}^{(i)}$$

Logistic regression model $$P(y^{(i)} = 1) = \frac{1}{1+e^{-(\beta_0 + \beta_1 {x_1}^{(i)} + \ldots + \beta_p {x_p}^{(i)})}}$$

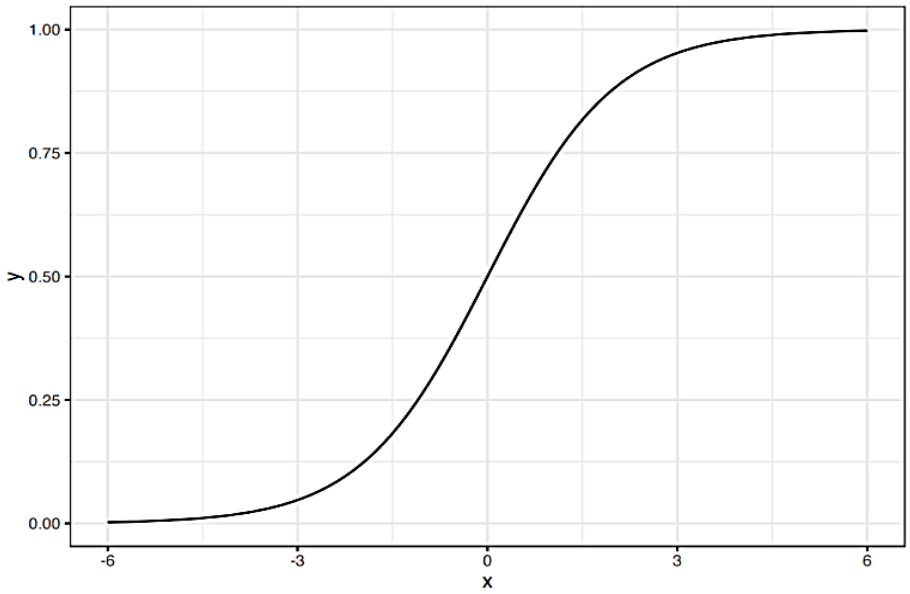

Figure 1: The logistic function, which outputs numbers between 0 and 1 (Note: At input 0, it outputs 0.5)

Model Interpretation

After the data preparation process was over, we concluded to a model with four predictors and one binary response variable named "SUBSCRIBED". The probability of subscribing to the term deposit (i.e. P("SUBSCRIBED"="yes")) was in the unit interval [0,1]. We assumed a linear relationship between the predictor variables and the log-odds of the event where the client had subscribed the term deposit (i.e. "SUBSCRIBED"="yes"). This linear relationship can be expressed mathematically as follows:

$$\text{odds} = \frac{\text{P(SUBSCRIBED)}}{1 - \text{P(SUBSCRIBED)}}$$

$$\log(\text{odds}) = \text{logit(SUBSCRIBED)} = -1.8 + 0.9 * \text{job2} + 0.4 * \text{job3} + 1.5 * \text{duration2} + 3.6 * \text{duration3} + 4.7 * \text{duration4} + 2.3 * \text{p\_contact} - 3.3 * \text{emp.var.rate}$$

Dummy variables:

- $\text{job2} = \begin{cases} 1, \text{if job} = \text{"Other"} \\ 0, \text{if job} \neq \text{"Other"} \end{cases}$
- $\text{job3} = \begin{cases} 1, \text{if job} = \text{"White Collar"} \\ 0, \text{if job} \neq \text{"White Collar"} \end{cases}$
- $\text{duration2} = \begin{cases} 1, \text{if duration} = \text{"5} - \text{10min"} \\ 0, \text{if duration} \neq \text{"10} - \text{15min"} \end{cases}$
- $\text{duration3} = \begin{cases} 1, \text{if duration} = \text{"10} - \text{15min"} \\ 0, \text{if duration} \neq \text{"10} - \text{15min"} \end{cases}$
- $\text{duration4} = \begin{cases} 1, \text{if duration} > \text{"15min"} \\ 0, \text{if duration} <= \text{"15min"} \end{cases}$
- $\text{p\_contact} = \begin{cases} 1, \text{if previous contact} = \text{"yes"} \\ 0, \text{if previous contact} = \text{"no"} \end{cases}$

Obviously job1+ job2, duration1 + duration2 + duration3, p_contact all $\in \{0,1\}$. The baseline category for job was "Blue Collar", for duration was "0-5min" and for previous contact was "no".

Table 1: Summary of the chosen logistic regression model

	Estimate	**Standard Error**	**p-value**
Intercept	-1.80	0.17	< 2e-16 ***
job2	0.91	0.19	2.98e-06 ***
job3	0.41	0.15	.00635 **
duration2	1.53	0.16	< 2e-16 ***
duration3	3.57	0.21	< 2e-16 ***
duration4	4.70	0.25	< 2e-16 ***
p_contact	2.34	0.26	< 2e-16 ***
Employment variation rate (emp.var.rate)	-3.33	0.23	< 2e-16 ***
Residual deviance: 1538.2 on 3910 degrees of freedom			
AIC: 1554.2			

We discovered that the variables “job”, “duration”, and “previous contact” had a positive impact – positive coefficients – on our target variable, implying that an increase in those variables increases the likelihood of a client subscribing to a term deposit. The “employment variation rate”, on the other hand, had a negative relationship with the outcome variable. Because our dependent variable was computed using the logit transformation, we were now working with a different scale. Below, we have provided an interpretation of the model's coefficients.

Coefficient	Interpretation
Intercept β_0	If the employment variation rate does not change, then the log odds of subscription for a client who worked as Blue collar and was not previously contacted are equal to -1.8. This multiplies the actual odds by approximately 0.16 units.
emp.var.rate	One unit increase in employment variation rate will result the log odds of subscription to increase by 3.3 units or multiply the actual odds of subscribing by $e^{3.3} \approx 27.1$ units, assuming that all the other variables remain constant.
job2 ($= \beta_0 + 0.9$)	If the employment variation rate does not change, then the log odds of subscription for a client who was not a blue- or white-collar worker are equal to -0.9, 0.9 units bigger compared to a blue-collar worker.
job3 ($= \beta_0 + 0.4$)	If the employment variation rate does not change, then the log odds of subscription for a client who was a white-collar worker are equal to -1.4, 0.4 units bigger compared to a blue-collar worker.
duration2 ($= \beta_0 + 1.5$)	$\beta_0 + 1.5$: If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was between 5 and 10 minutes are equal to -0.3, 1.5 units bigger compared to one who did not talk at all or talked less than 5 minutes.
duration3 ($= \beta_0 + 3.6$)	If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was between 10 and 15 minutes are equal to 1.8, 1.5 units bigger compared to one who did not talk at all or talked less than 5 minutes.
duration4 ($= \beta_0 + 4.7$)	$\beta_0 + 4.7$: If the employment variation rate does not change, then the log odds of subscription for a client whose last contact duration was bigger 15 minutes are equal to 2.9, 4.7 units bigger compared to one who did not talk at all or talked less than 5 minutes.
p_contact ($= \beta_0 + 2.3$)	If the employment variation rate does not change, then the log odds of subscription for a client who was previously contacted are equal to 0.5, 2.3 units bigger compared to one who was not previously contacted at all.

Furthermore, the residual deviance was used for goodness of fit tests. We noticed from the standard error column of Table 1, how much the estimation of the coefficients can vary. Based on this column, we are able to see how stable each estimate is. In our case, all of the standard errors were small and close to the estimation values, so we assumed that all of our estimations were stable.

From the final column of Table 1, we discovered that the null hypothesis (i.e. a coefficient can be assumed to be equal to zero) was rejected as the p-value of each coefficient was less than our significance level (i.e. a=5%). As a result, all of the model's coefficients were statistically significant in our analysis.

We also created the standard diagnostic plots[1] for our model (i.e. 'residuals vs fitted', 'normal q-q', 'scale-location', and 'residuals vs leverage' plots). However, when combined with a logistic regression model, the interpretations of these plots are not generally valid. For instance, both the 'residuals vs fitted' and the 'scale-location' plots looked like there were problems with the model, but in fact there were none. So, we decided to only use these plots for outlier detection in our analysis.

In terms of evaluating the fit of our model, the residuals against fitted values plot[2] indicated that we had a reasonably good fit. We also observed this by plotting[3] the Pearson or Deviance residuals against each explanatory variable in our model separately, where no unusual pattern was observed. The high value of McFadden's pseudo R-squared reflected this as well.

Assumptions of logistic regression

In addition, we examined whether our model met the basic assumptions of logistic regression. To begin, binary logistic regression requires a **binary dependent variable**, such as our subscribe variable, which had only two possible outcomes ("yes" vs "no").

Secondly, logistic regression requires that the observations be **independent** of one another. In our case this condition was not met, because the observations came from matched data. This was due to the fact that our collected data referred to clients from May 2008 to June 2010, a time period in which multiple contacts with the same client were required. This was a significant issue because, if we assumed that measurements taken in the same client were correlated, the test for a difference in treatments would have a lower residual or error variance than a completely randomized design, thereby increasing the precision of the analysis.

Thirdly, logistic regression requires that the independent variables have little or no **multicollinearity**. Using the square root of the variance inflation factor (GVIF[4]), we verified this during our analysis. Based on the general rule of thumb, a GVIF value greater than 5 (or higher than $\sqrt{10}$=3.16) indicates a high level of collinearity, which is why we decided to remove some variables from the model suggested by AIC. The final selected model showed no issues of multicollinearity.

Fourth, logistic regression is predicated on the **linearity** of independent variables and log odds. The Box-Tidwell test[5] was used to test these assumptions by including interactions between the continuous predictors and their logs in the model. Because the interaction was significant, the assumption was broken. However, because our sample size was large, we should not be overly concerned with just a significant interaction.

Finally, logistic regression typically requires a **large sample size**. In our case, we kept nearly 4,000 observations from our original sample size, which was deemed as an adequate number.

We also did not want any influential values (i.e. outliers or extreme values) in the continuous predictors. From Cook's distance plot[6], we noticed three outliers that could be influential points. By computing the standardized residual error, discovered two data points with absolute standardized residuals greater than three. So, we can assume that these were potentially influential points. As a result, we came to the conclusion that not all of the logistic regression assumptions were satisfied.

---

[1] See Figure 1 in APPENDIX
[2] See Figure 2 in APPENDIX
[3] See Figure 3 in APPENDIX
[4] GVIF stands for Generalized Variance Inflation Factor
[5] See Table 2 in APPENDIX
[6] See Figure 4 in APPENDIX

Link Function

A link function is used to identify a mean function that is a linear function of the explanatory variables. Because our binary dependent variable belonged to the binomial exponential family, we could use either "logit" or "probit" as link functions. From Table 2, we observed that the differences between those two link functions were minor, and the two models produced very similar results. So we could use either one, but due to the fact that our dependent variable was considered to be a truly qualitative and binomial character, we preferred the logit modelling.

Table 2: Comparison between the "probit" and the "logit" link functions

Link Function	Residual Deviance	Residual Deviance/df	AIC	Pseudo R2
Probit	1516.8	0.3879	1532.8	0.4006193
Logit	1538.2	0.3934	1554.2	0.3921716

Conclusions and Discussions

We observed from our data that most of the observations of our target variable (nearly 90%) corresponded to one of its two levels. This indicated that we had an imbalanced dataset, because the classes were not represented equally.

In general, imbalanced classifications pose a challenge for predictive modeling[7], as most of the machine learning algorithms used for classification were built on the assumption of an equal number of examples for each class. This could lead to models with poor predictive performance, particularly for the minority class. Considering that our goal was to investigate which variables contributed to a successful contact, where 90% of the clients in our dataset had not subscribed a term deposit, it could be difficult for our model to predict the likelihood of a client subscribing.

We also noticed that the "yes" values (i.e. coded value=1) of the dependent variable were much more spread out in the plot[8] than the "no" values. That indicated that our model could not predict well the clients who subscribed a term deposit, just as we expected due to the imbalance in our data. As a result, it would be advisable to use our model for descriptive purposes rather than predictions.

Our dataset analysis revealed that we could identify the attributes that contribute to a successful contact. The factors that most influenced a client's subscription to a term deposit were their job type and the existence and duration of a previous contact. In particular, if a client had previously been contacted, he was more likely to subscribe. We also discovered that the more time a client spent talking, the more likely he or she was to subscribe.

On the contrary, we discovered that a decrease in the employment variation rate makes a successful subscription much more difficult. This was to be expected, because when the number of people hired decreases while the number of people fired increases, we understand that unemployment was an impediment to someone subscribing to a term deposit.

[7] https://machinelearningmastery.com/what-is-imbalanced-classification/

[8] See Figure 5 in APPENDIX

We also observed that when we removed the observations that corresponded to the “unknown” level of loan variable, the exact same observations corresponded to the “unknown” level of the housing variable. So, if only housing is "unknown," we can assume that loan is "unknown." Further investigation, based on the contingency tables, revealed that this was due to the fact that these two variables were related to each other.

This work was done considering only a portion of the dataset. For future works we could use the dataset to its fullest potential, by taking into account all of the observations and producing even more accurate results.

APPENDIX

Tables

Table1: Data description

Retail Bank Data
Input variables:
Bank client data 1 - age (numeric) 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) 4 - education (categorical: basic.4y','basic.6y','basic.9y', 'high.school','illiterate','professional.course', 'university.degree','unknown') 5 - default: has credit in default? (categorical: 'no','yes','unknown') 6 - housing: has housing loan? (categorical: 'no','yes','unknown') 7 - loan: has personal loan? (categorical: 'no','yes','unknown') **related with the last contact of the current campaign:** 8 - contact: contact communication type (categorical: 'cellular','telephone') 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec') 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri') 11 - duration: last contact duration, in seconds (numeric). **other attributes:** 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) 14 - previous: number of contacts performed before this campaign and for this client (numeric) 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') **social and economic context attributes** 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric) 17 - cons.price.idx: consumer price index - monthly indicator (numeric) 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric) 19 - euribor3m: euribor 3 month rate - daily indicator (numeric) 20 - nr.employed: number of employees - quarterly indicator (numeric)
Output variable (desired target):
21 - SUBSCRIBED - has the client subscribed a term deposit? (binary: 'yes','no')

Table 2: Box-Tidwell test

```
 MLE of lambda Score Statistic (z)  Pr(>|z|)
      -0.65832             9.0622 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  5
```

We noticed that we rejected the null hypothesis (a=.05>p-value). As a result, the interaction was significant, so the linearity assumption could not be assumed to be true.

Figures

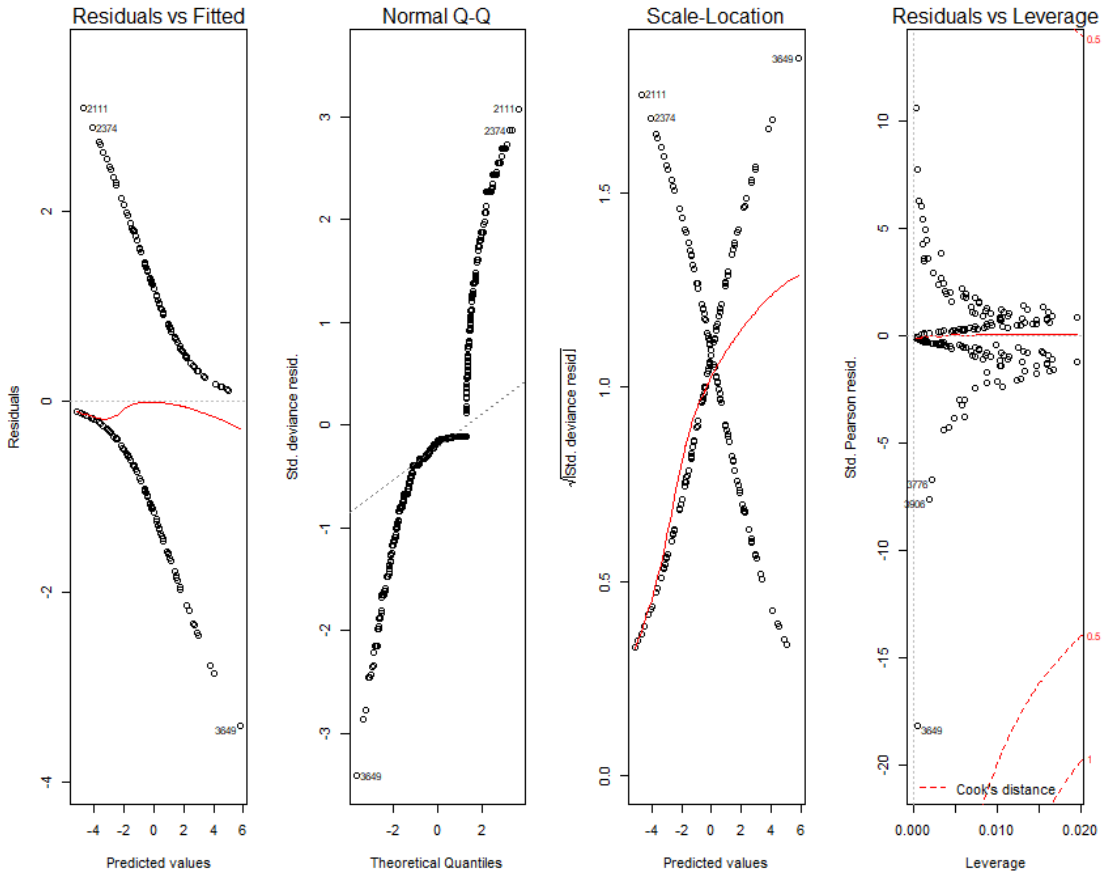

Figure 1: Standard Diagnostic Plots

It seemed like the 2111 and 2374 data points were outliers and could very well affect our model.

Note: We should actually be using Pearson or deviance residuals to gauge our models fit so the fact that we did not observed anything close to a straight line in the upper left plot was fine. Also, Q-Q plots were irrelevant for this kind of model as well.

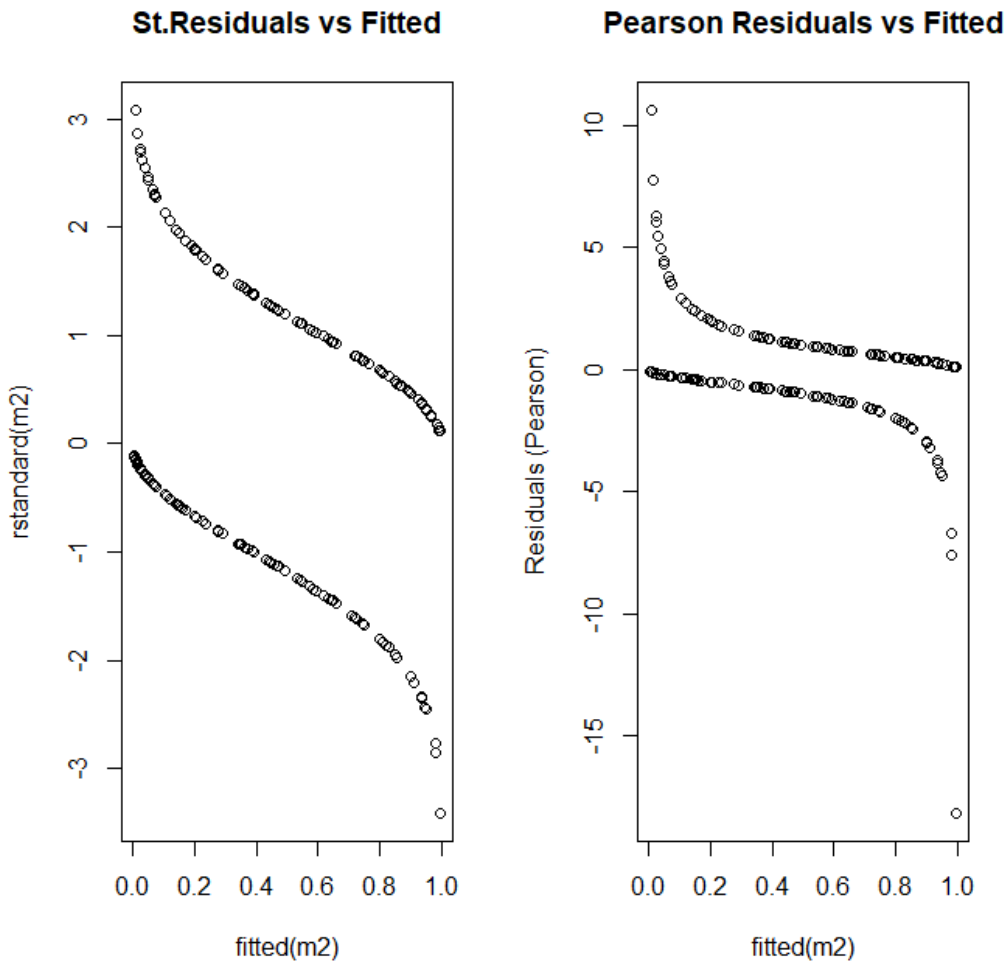

Figure 2: Residuals against Fitted values

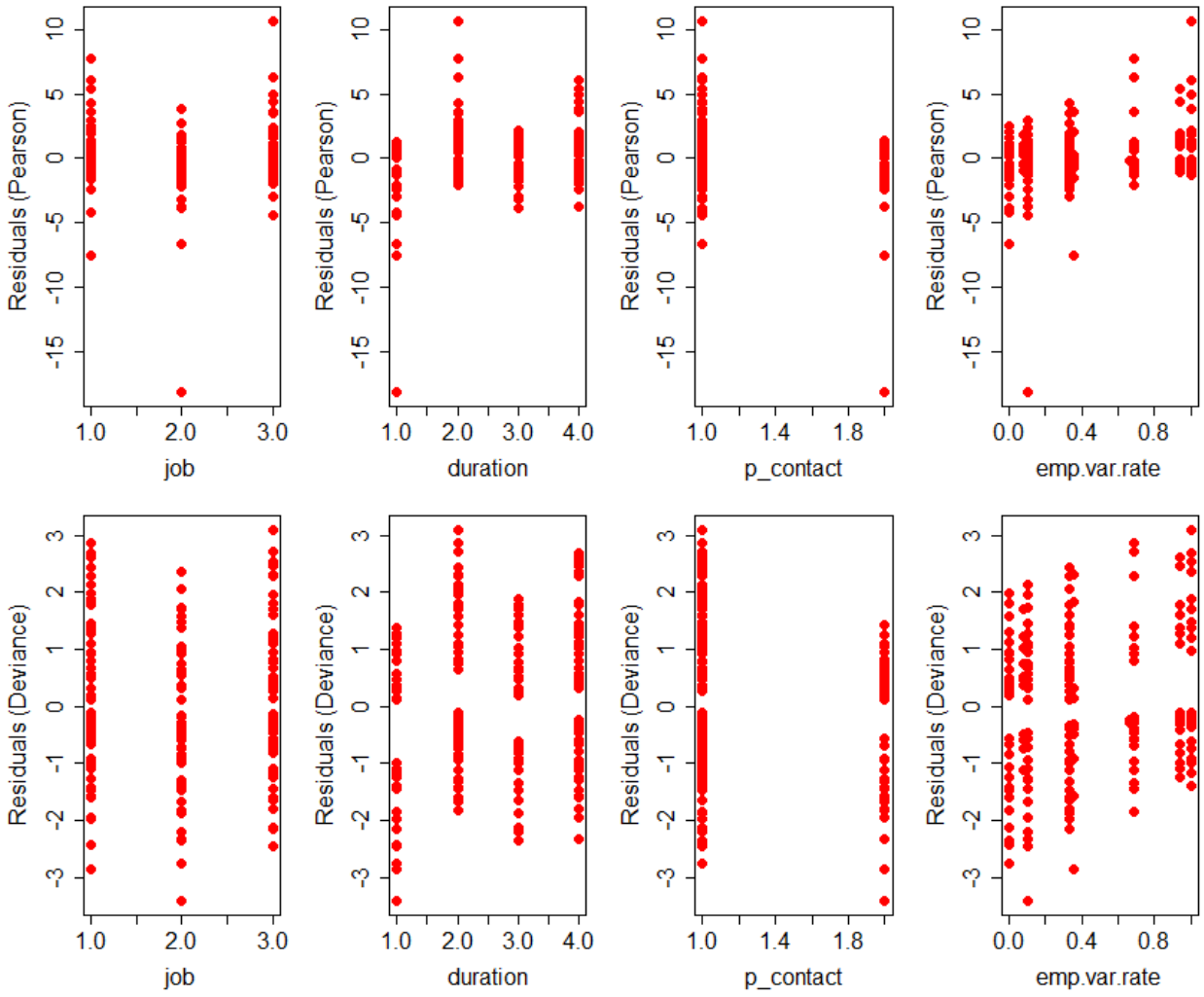

Figure 3: Plot of Pearson and Deviance residuals against the explanatory variables of the model

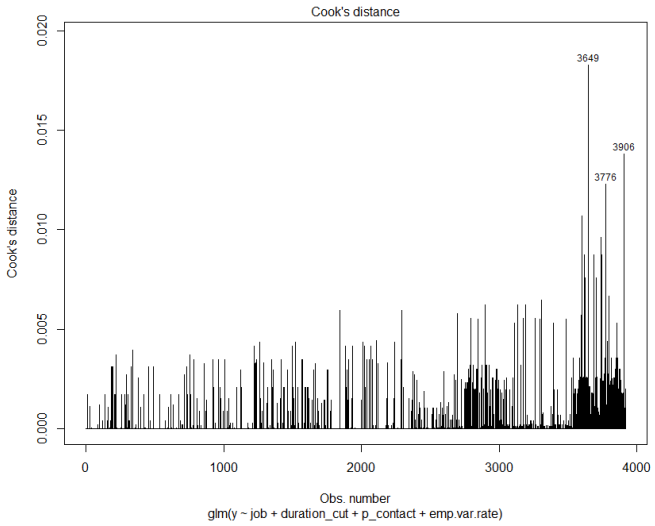

Figure 4: Cook's distance plot for examining for influential values

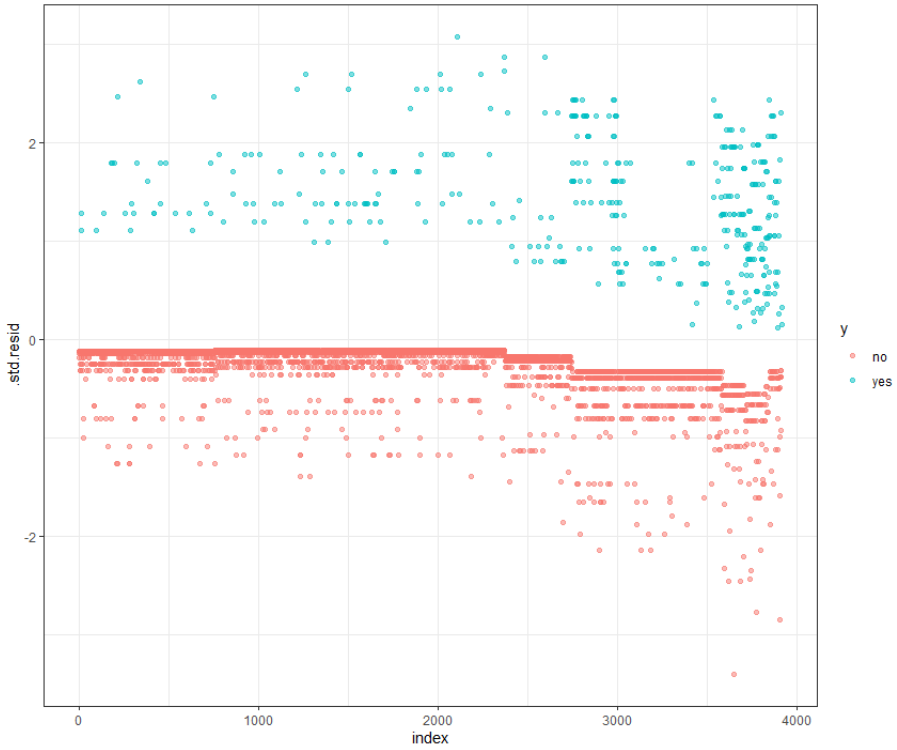

Figure 5: Plot of the Standardized residuals

The "yes" values (coded value=1) of the dependent variable were really spread out in the plot in contrast with the "no" values. That meant that our model could only predict well the "no" values.