**Analysis Overview:**

* Data preparation (handle missing values, group variables, normalization)
* Bivariate Analysis (correlation coefficients, boxplots, contingency tables)
* Model Selection:
  + Remove problematic values (coefficients with “NA” values, coefficients whose value and standard error is very “high”)
  + Hypothesis testing (Wald test, Goodness of fit test)
  + Variable Selection (Stepwise procedures, Multi-collinearity testing)
  + Model Interpretation
  + Assumptions of logistic regression
  + Conclusions and Discussions

**Model Interpretation:**

Mathematical formulation of the model:

{\displaystyle \ell =\log \_{b}{\frac {p}{1-p}}=\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}We concluded to a model with four predictors and one binary response variable named “SUBSCRIBED”, which we denoted as p=P(“SUBSCRIBED”=”yes”). This referred to the probability of subscribing to the term deposit, which was in the unit interval [0,1]. We assumed a linear relationship between the predictor variables and the log-odds of the event that “SUBSCRIBED”=”yes”. This linear relationship could be written in the following mathematical form:

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

Dummy variables:

* job2 =
* job3 =
* duration2 =
* duration3 =
* duration4 =
* p\_contact =

Obviously job1+ job2, duration1 + duration2 + duration3, p\_contact all ∈ {0,1}.

Baseline category for job was “Blue Collar”, for duration is “0-5min” and for previous contact was “no”.

Table 1: Summary of the chosen logistic regression model

|  |  |  |  |
| --- | --- | --- | --- |
|  | Estimate | Std.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 observed that job, duration and previous contact variables had a positive impact – positive coefficients - to our target variable, meaning that a rise in those variables will increase the probabilities of a client to subscribe to a term deposit. In contrary, the employment variation rate had a negative relationship with the outcome variable.

Interpretation of the coefficients of the model

Our dependent variable was computed using the logit transformation; thus, we were now using a different scale.

Intercept β0: 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.

Coefficient 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 e3.3 ≈ 27.1 units, assuming that all the other variables remain constant.

Coefficient job2: β0 + 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.

Coefficient job3: β0 + 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.

Coefficient duration2: β0 + 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.

Coefficient duration3: β0 + 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.

Coefficient duration4: β0 + 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.

Coefficient p\_contact: β0 + 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.

Extra interpretation of the model

We used the residual deviance for goodness of fit tests.

From the standard error column we saw how much the estimation of the coefficients can change. We looked at this column as indication of how stable my estimation is. In our case, all the standard errors were small and close to the estimation values, so assumed that all the estimations were stable.

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

**Further analysis**

We also created the standard diagnostic plots[[1]](#footnote-1) for our model ('residuals vs fitted', 'normal q-q', 'scale-location', and 'residuals vs leverage' plots). But the interpretations of these plots are not generally valid when used with a logistic regression model. For example, both the 'residuals vs fitted' and the 'scale-location' plots looked like there were problems with the model, but in fact there were not any. So, we decided not to use these plots in our analysis, except only for outlier detection.

As for the evaluation of the fit of our model, we could say from the residuals against fitted values plot[[2]](#footnote-2) that we had a pretty decent fit. We also saw that by plotting[[3]](#footnote-3) the Pearson or Deviance residuals against each explanatory variable in our model separately, where we did not observe any unusual pattern. This was also seen by the high value of McFadden's pseudo R-squared.

**Assumptions of logistic regression**

Firstly, binary logistic regression requires the dependent variable to be **binary**, just like our subscribe variable was, which has only two possible outcomes (“yes” vs “no”)

Secondly, logistic regression requires the observations to be **independent** of each other.  In our case this was not the case, because the observations came from matched data. This was because our collected data referred to clients from May 2008 to June 2010, during which period more than one contact to the same client was required. This was a major problem, because if we assumed that measurements taken in the same client were correlated, the test for a difference in treatments will involve a smaller residual or error variance than that based on a completely randomized design, thereby increasing precision in the analysis.

Thirdly, logistic regression requires there to be little or no **multicollinearity** among the independent variables.  We checked that during our analysis using the GVIF[[4]](#footnote-4) function. As a rule of thumb, a GVIF value that exceeds 5 or the square root of 10 indicates a problematic amount of collinearity and this was the reason we decided to drop some variables from the model suggested by AIC. The final selected model did not have any multicollinearity problems.

Fourthly, logistic regression assumes **linearity** of independent variables and log odds. We tested this assumptions by including in the model interactions between the continuous predictors and their logs via the Box-Tidwell test[[5]](#footnote-5). The interaction was significant so the assumption has been violated. However, the size plays a role here too, so we should not be very concerned with just a significant interaction, because our sample size was large.

Finally, logistic regression typically requires a large sample size. Here we kept almost 4,000 observations from our original sample size, which was considered as an adequate number.

We also did not want to have any influential values (extreme values or outliers) in the continuous predictors. From Cook’s distance[[6]](#footnote-6) we saw that there were 3 outliers which may be influential points. By computing the standardized residual error we saw that there were two data points with absolute standardized residuals above 3. So, we could assume that these may be influential points.

As a result, we concluded that not all the logistic regression assumptions were satisfied.

**Link functions**

A link function identifies a function of the mean that is a linear function of the explanatory variables. Our binary dependent variable belongs to the binomial exponential family, so we could use logit or probit as link functions.

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 |

We observed that the differences between those two link functions were minor and the two models led to very similar results. So, we could choose either one of them, but due to the fact that our dependent variable was considered to be a truly qualitative and binomial character we preferred the logit modelling.

**Conclusions and Discussions**

**Comment on whether the model can be used for prediction**

From Figure 3, we observed that most of the observations of our target variable, almost 90%, corresponded to one of its two levels. This indicated that we had an imbalanced dataset, because the classes were not represented equally.

Generally, imbalanced classifications pose a challenge for predictive modeling[[7]](#footnote-7), as most of the machine learning algorithms used for classification were based on the assumptions of an equal number of examples for each class. This could result in models with poor predictive performance, especially for the minority class. Considering that our goal was to examine which variables contributed to a successful contact, where 90% of the clients in our dataset had not subscribed a term deposit, it could be challenging for our model to predict the likelihood of a client subscribing.

We also observed that the “yes” values (i.e. coded value=1) of the dependent variable were really spread out in the plot[[8]](#footnote-8) in contrast with 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 of our data.

So, it would be advisable to use our model for descriptive purposes instead of predictions.

**Conclusions**

The analysis of our datasets, showed us that we could identify the attributes that contribute to a successful contact.

The factors which contributed the most to the subscription of a term deposit from a client were their job type and the existence and duration of a previous contact. Specifically, if a client was previously contacted then he was more likely to subscribe. We also noticed that the more time a client spent talking, the better the chances of subscription were.

In contrary, we observed that a drop in the employment variation rate will make it a lot more difficult for a successful subscription. This was expected because when the number of people who are hired decreases and the number of those who are fired increases, then we understand that unemployment was an inhibit factor for someone to subscribe to a term deposit.

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, we may assume that loan is “unknown” if only housing is “unknown”. By further investigation we saw from the contingency tables that the reason for this was because these two variables were related to each other.

This work has been done considering a part of the dataset. For future works we could use the dataset to its fullest potential, by taking all the observations into consideration and create even more accurate results.

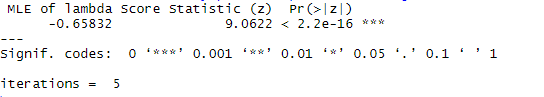
**APPENDIX:**

**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



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**

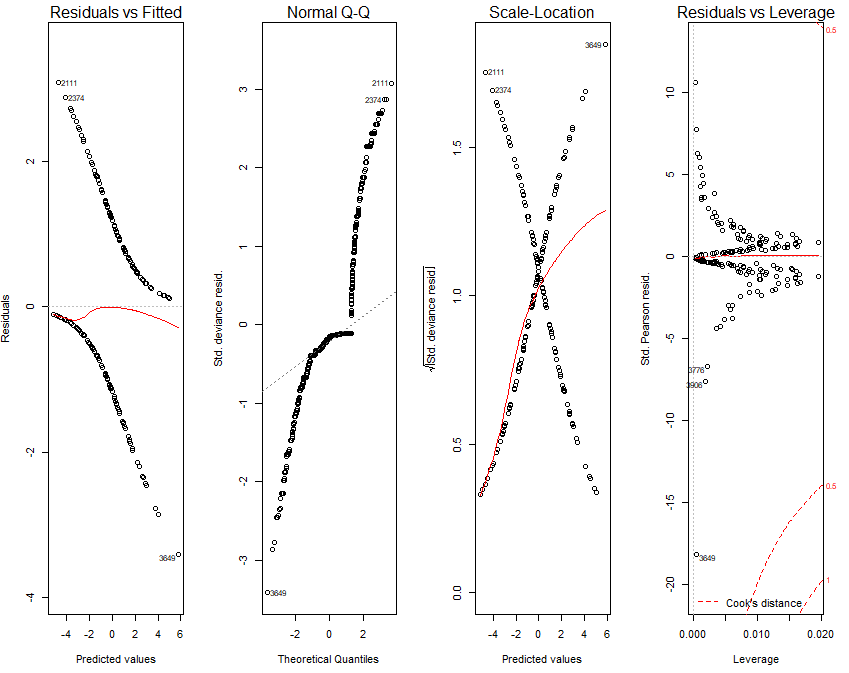


Figure 1: Standard Diagnostic Plots

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.

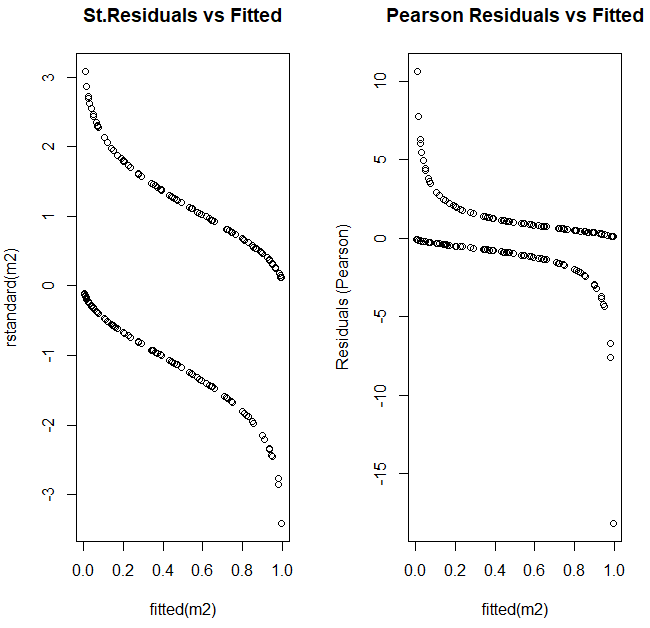


Figure 2: Residuals against Fitted values

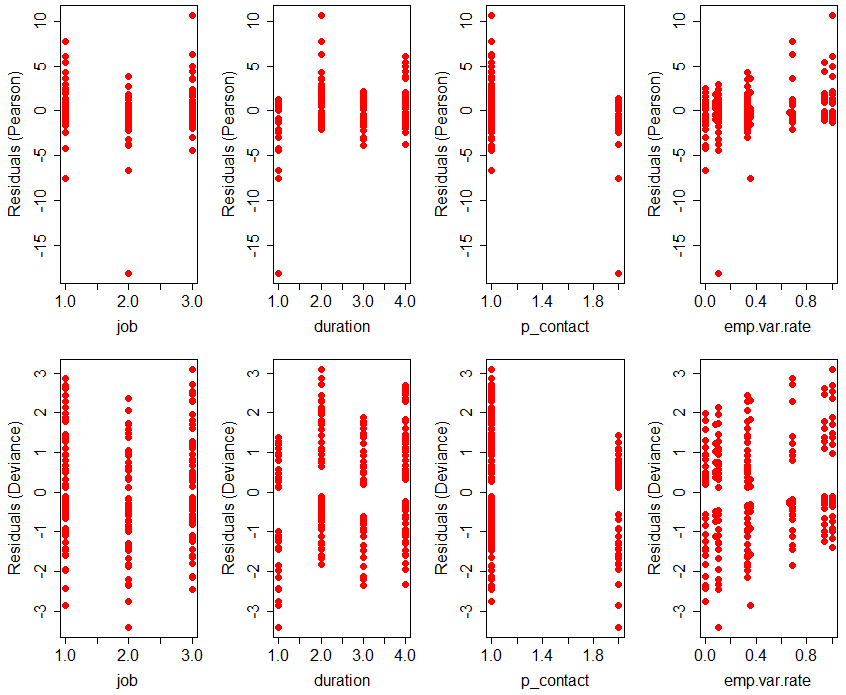


Figure 3: Plot of Pearson and Deviance residuals against the explanatory variables of the model

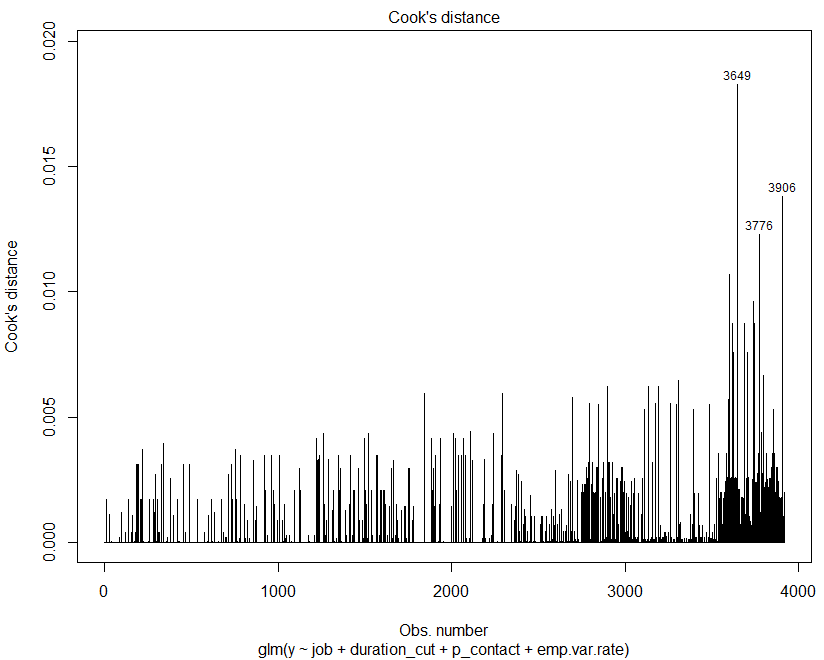


Figure 4: Cook’s distance plot for examining for influential values

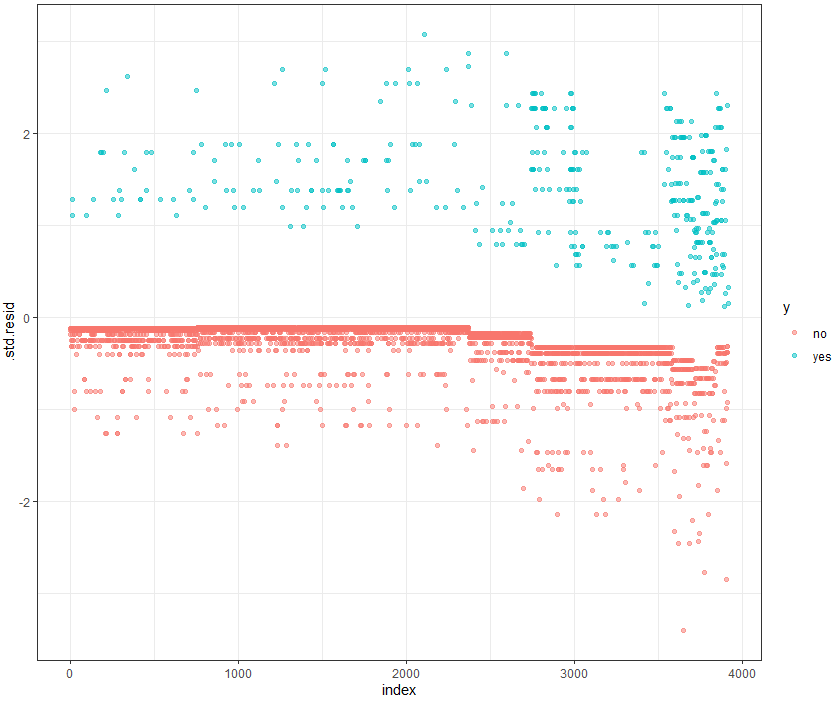


Figure 5: Plot of the Standardized residuals

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.

1. See Figure 1 in APPENDIX [↑](#footnote-ref-1)
2. See Figure 2 in APPENDIX [↑](#footnote-ref-2)
3. See Figure 3 in APPENDIX [↑](#footnote-ref-3)
4. GVIF stands for Generalized Variance Inflation Factor [↑](#footnote-ref-4)
5. See Table 2 in APPENDIX [↑](#footnote-ref-5)
6. See Figure 4 in APPENDIX [↑](#footnote-ref-6)
7. <https://machinelearningmastery.com/what-is-imbalanced-classification/> [↑](#footnote-ref-7)
8. See Figure 5 in APPENDIX [↑](#footnote-ref-8)