Outline 3 Final Paper

Ray Anthony Roderos

APAN K5200

**A Logistic Regression Analysis on the Client Characteristics and Telemarketing Operations on Special Bank Deposit Rates Availment**

# Description of Data

The Portuguese bank’s direct marketing dataset obtained is multivariate with real numbers. There are 45,200 instances with 9 variables. The variables are categorized into 3: bank client data, and details of the call. The research hypothesizes that bank client data and call information from previous contacts on the current and past campaigns have a significant effect on the client for him/her to accept the offer. The study has already excluded some variables deemed not relevant for the study and transformed others to fit the regression model.

# Empirical Model

After data cleansing, exclusion and transformation, this will be the proposed model for the logistic regression:

**YB = YB (AC, JC, MC, EC, DF, HF, LF, MO, DO)**

Where:

**YB** = Special Bank Deposit Rate Availment

**YB** = 0 if didn’t availed the Special Rate

= 1 if availed the Special Rate

**AC** = Age of the Client (Years)

**JC** = Employment of the Client (Dummy Variable)

**JC** = 0 if Unemployed

= 1 if Employed

**MC** = Employment of the Client (Dummy Variable)

**MC** = 0 if not married

= 1 if married

**EC** = Education Level of the Client (Dummy Variable)

**EC** = 0 if Primary or Secondary Graduate

= 1 if Tertiary Graduate

**DF** = History of Credit Default of the Client

**DF** = 0 if Client has not defaulted before

= 1 if Client has defaulted before

**HF** = Client has Housing Loan

**HF** = 0 if Client does not have a housing loan

= 1 if Client has a housing loan

**LF** = Client has a Personal Loan

**LF** = 0 if Client does not have a personal loan

= 1 if Client has a personal loan

**MO** = Month of the Year Call was Made (1-12 months)

**DO** = Day of the Month (in 1-30 days)

Presentation and Analysis of the Results

From the data presented by the bank

Outline

1. Data Processing
   1. The Model that you used
   2. Forward or Backward Stepwise
2. Presentation and Analysis of Results
   1. Analysis of Summarized Data
3. Descriptive Statistics
   1. Log Regression Analysis
4. Conclusion
   1. Policy Implications
   2. Recommendation for Further Studies

Ray Anthony Roderos

C004089342

APAN K5200

**A Logistic Regression Analysis on the Client Characteristics and Telemarketing Operations on Special Bank Deposit Rates Availment**

Table of Contents

[Description of Data 2](#_Toc465437147)

[Missing Values 2](#_Toc465437148)

[Realistic Values 3](#_Toc465437149)

[Correlation 3](#_Toc465437150)

[Data Transformation 3](#_Toc465437151)

[Choosing Variables 4](#_Toc465437152)

[Problem Statement 5](#_Toc465437153)

[Proposed Empirical Model 5](#_Toc465437154)

[Bibliography 7](#_Toc465437155)

[Appendix 7](#_Toc465437156)

[Appendix 1 -Definitions of Variables 7](#_Toc465437157)

[Appendix 2 - Univariate Analysis of the Variables 8](#_Toc465437158)

[Appendix 3 - Initial Regression Run on Data 15](#_Toc465437159)

[Appendix 4 - Correlation 18](#_Toc465437160)

[Appendix 5 – Summary Statistics 20](#_Toc465437161)

[Appendix 6 - Graph Builder 21](#_Toc465437162)

[Appendix 7 – Univariate Analysis of Dummy Variables 25](#_Toc465437163)

[Appendix 8 – SAS JMP Imputation Report 26](#_Toc465437164)

# Description of Data

The Portuguese bank’s direct marketing dataset obtained is multivariate with real numbers. There are 45,200 instances with 17 variables dated 2012. The variables are categorized into 4: bank client data, data on the last contact with the client of the current campaign, social and economic context such as employment rate and the consumer price index and lastly, other attributes on the call such as total number of contacts for this campaign and the outcomes of previous marketing campaigns with the dependent variable being a “yes” or a “no” from the client. The paper hypothesizes that bank client data and call information from previous contacts on the current and past campaigns have a significant effect on the client for him/her to accept the offer.

The definitions of each variable and the appropriate category for each values can be found at Appendix 1 and the univariate analysis at Appendix 2.

# Missing Values

The univariate analysis on all 16 variables has shown that there are no missing values for all the variables. However, some of the variables have a category labelled “unknown,” namely, *job, education, contact* and *poutcome*. To remove the observations with an “unknown” value would entail removing 36,000 values or 81% of the total dataset. These variables with “unknown” values are also multi-category variables that will need transformation into binary parameters or dummy variables. Upon further analysis, many of the variables with missing values which have those “unknown” values are in very small percentages. These are the variables and the number and percentage of values that are categorized as “unknown”:

*Job* – 288 (0.63% of the total dataset)

*Education* – 1857 (4.1% of the total dataset)

*Contact* – 13020 (28.79% of the total dataset)

*Poutcome* – 36959 (81.74% of the total dataset)

Data cannot be obtained to fill these in since this is third-hand data used for machine-learning research which was obtained from a Portuguese bank in 2012. Filling in these values will also be too hard since assumptions cannot be made as there is no data on cellphone and telephone ownership by age group in Portugal in 2012 that can be obtained. *Poutcome* is a result of the bank’s previous marketing campaign which is data that cannot also be obtained easily. Since *contact* and *poutcome* have too many missing values, it is best for the model not to include these variables which will be assessed in the following section. For *job* and *education*, the data were transformed into binary parameters or dummy variables before data imputation was conducted using SAS JMP (Group)[[1]](#footnote-1). The resulting data imputation was rounded off to the nearest numerical category. This can be done because the missing values consist of only less than 5% of the total dataset which is an acceptable margin. Results of the data imputation are on Appendix 8.

# Realistic Values

All of the variables have also shown that the values are within the realistic boundaries such as the maximum of age is 95. *Duration* which means the duration of the last call in seconds, have some of the values in 0 meaning that the call was not performed. Another variable is *pdays* which means the number of days that passed by after the client was last contacted from a previous campaign, many of the values are in -1 which means the client was not previously contacted.

# Correlation

A correlation was done using the new transformed variables with numeric values since it cannot run using the original non-numeric values. Based from the results, there is little correlation amongst the variables. This means that the independent variables don’t have to be removed due to highly correlated variables, and therefore their underlying common factor, affecting the relationship of the dependent variable. The correlation matrix can be found at Appendix 4.

# Data Transformation

An initial run of the dataset using log regression showed that many of the variables with multiple categories were split to accommodate the model. One example is the variable for employment “job” with 11 categories. The model divided these as separate binary parameters (ex. Unemployed – yes/no the administration – yes/no). Because of this, running all 16 variables yielded 43 parameters, too many for the model. The results of the initial run can be found on Appendix 3. For model simplicity, several variables were transformed into binary variables thus employing dummy variables. Variables with more than 2 categories were made into two broader categories. The following are the variables that have been transformed:

*job*: type of job (categorical: "admin.","blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown") – transformed into two categories: employed and unemployed as dummy variables (1 and 0) since the employment categories cannot be categorized as ordinal.

*marital*: marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed) – transformed into two categories: married and not married (1 and 0).

*education*: highest educational attainment (categorical: “primary”,”secondary”, “tertiary”) – transformed into tertiary or non-tertiary as dummy variables (1 and 0).

*poutcome*: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success") – transformed into success and non-success as dummy variable (1 and 0).

These dummy variables simplified but at the same time, retained the necessary information for the model. These also prevented the model from creating more parameters than the 17 variables. Descriptive statistics of these dummy variables can be found at Appendix 7.

# Choosing Variables

For model simplicity, variables were assessed if it is necessary for running the model. 3 elements were used as criteria. 1st is to see if the information the predictor holds is relevant to the dependent variable. Since the dependent variable is about accepting a special bank rate deposit, the client’s financial information will be relevant. And since these are call center operations, operational data and past campaigns may also relevant. The 2nd criterion is that the information must be complete or almost complete with minimal missing values. Since it is hard to fill in large sets of missing values due to the nature of the dataset, variables with lots of missing values is best removed. The 3rd criterion is that there should be little to no correlation with each other so as not to affect relationships with the dependent variable. The following are the variables and the assessment if they should be included or not:

*Age, Job, Marital* and *Education* are basic characteristics of the clients that are not only essential to describe the client but also the basic information used by all banks. The information of all the variables are mostly complete, missing values were less than 5% and was filled up using data imputation.

*Default, Housing* and *Loan* are variables that indicate the financial capability of the client. Since the dependent variable is about special savings rate, fund availability from the side of the client is important. If these indicate that the client doesn’t have money, it is unlikely that the client will avail of the product. The information is complete and with no missing values

*Contact, Month, Day of Week* and *Duration* describe the operational factors of the call. These can be techniques that call center agents can employ on when and how to call potential clients. However, not all the variables are complete. Contact has very a large portion of incomplete data is 28.79% and is very hard to fill in. *Month, Day of Week* and *Duration* are complete. *Duration* is not an independent variable since the length of time of the call cannot be used as a predictor as it will always be a post-operation value meaning that duration can only be determined after the call. It will be hard for the agent to arbitrarily pursue a certain number of seconds regardless of the client’s answer to increase chances of availment. It is determined that Contact and Duration should be removed from the model.

*Campaign, Pdays, Previous* and *Poutcome* indicate the outcome with the client in a previous campaign and its details. *Poutcome* only had 19% of the data with a definite outcome, the rest are “unknown”. *Campaign, Pdays* and *Previous* are predictors of *Poutcome* and had a more direct relationship with the previous campaign rather than the current one. Without knowing the outcomes of the previous campaign, the other 3 variables cannot be used. Because of these, *Campaign, Pdays, Previous* and *Poutcome* are determined not to be included in the model.

From the correlation results, there is a little correlation between the predictors, meeting the 3rd criterion.

# Problem Statement

There is no change of the Problem Statement from the proposal.

1. Client Side: What are the Socio-economic characteristics of the potential client that will entail him/her to avail of the special bank term deposit if telemarketed?
2. Agency Side: What operational and marketing tactics should the bank use to increase their chances of successfully gaining a client through telemarketing?

# Proposed Empirical Model

After cleaning the data for missing values, checking for data that doesn’t make sense, testing for correlation, data transformation for several variables and choosing the variables to include in the model, there will only be 9 predictors for the dependent variable. This will be the proposed model for the logistic regression:

**YB = YB (AC, JC, MC, EC, DF, HF, LF, MO, DO)**

Where:

**YB** = Special Bank Deposit Rate Availment

**YB** = 0 if didn’t availed the Special Rate

= 1 if availed the Special Rate

**AC** = Age of the Client (Years)

**JC** = Employment of the Client (Dummy Variable)

**JC** = 0 if Unemployed

= 1 if Employed

**MC** = Employment of the Client (Dummy Variable)

**MC** = 0 if not married

= 1 if married

**EC** = Education Level of the Client (Dummy Variable)

**EC** = 0 if Primary or Secondary Graduate

= 1 if Tertiary Graduate

**DF** = History of Credit Default of the Client

**DF** = 0 if Client has not defaulted before

= 1 if Client has defaulted before

**HF** = Client has Housing Loan

**HF** = 0 if Client does not have a housing loan

= 1 if Client has a housing loan

**LF** = Client has a Personal Loan

**LF** = 0 if Client does not have a personal loan

= 1 if Client has a personal loan

**MO** = Month of the Year Call was Made (1-12 months)

**DO** = Day of the Month (in 1-30 days)

# Bibliography

Group, IDRE Consulting. *Missing Data Techniques with SAS*. n.d. Powerpoint Slides. 26 October 2016. <http://www.ats.ucla.edu/stat/sas/seminars/missing\_data/Missing%20Data%20Techniques\_UCLA.pdf>.

S. Moro, P. Cortez and P. Rita. A Data­Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22­31, June 2014

UC Irvine Machine Learning Repository**,** https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

# Appendix

## Appendix 1 -Definitions of Variables

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: "primary",”secondary”,”tertiary”)

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")

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") in numbers

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no").

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")

16 - y - has the client subscribed a term deposit? (binary: "yes","no")

## Appendix 2 - Univariate Analysis of the Variables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **age**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 95 | | 99.5% |  | 77 | | 97.5% |  | 61 | | 90.0% |  | 56 | | 75.0% | quartile | 48 | | 50.0% | median | 39 | | 25.0% | quartile | 33 | | 10.0% |  | 29 | | 2.5% |  | 25 | | 0.5% |  | 22 | | 0.0% | minimum | 18 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 40.93621 | | Std Dev | 10.618762 | | Std Err Mean | 0.0499404 | | Upper 95% Mean | 41.034094 | | Lower 95% Mean | 40.838326 | | N | 45211 | | **job**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | admin. | 5171 | 0.11437 | | blue-collar | 9732 | 0.21526 | | entrepreneur | 1487 | 0.03289 | | housemaid | 1240 | 0.02743 | | management | 9458 | 0.20920 | | retired | 2264 | 0.05008 | | self-employed | 1579 | 0.03493 | | services | 4154 | 0.09188 | | student | 938 | 0.02075 | | technician | 7597 | 0.16803 | | unemployed | 1303 | 0.02882 | | unknown | 288 | 0.00637 | | Total | 45211 | 1.00000 |   N Missing 0  12 Levels |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **marital**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | divorced | 5207 | 0.11517 | | married | 27214 | 0.60193 | | single | 12790 | 0.28290 | | Total | 45211 | 1.00000 |   N Missing 0  3 Levels | **education**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | primary | 6851 | 0.15153 | | secondary | 23202 | 0.51319 | | tertiary | 13301 | 0.29420 | | unknown | 1857 | 0.04107 | | Total | 45211 | 1.00000 |   N Missing 0  4 Levels |
| **default**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | no | 44396 | 0.98197 | | yes | 815 | 0.01803 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels | **balance**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 102127 | | 99.5% |  | 18105.3 | | 97.5% |  | 8412.7 | | 90.0% |  | 3574.8 | | 75.0% | quartile | 1428 | | 50.0% | median | 448 | | 25.0% | quartile | 72 | | 10.0% |  | 0 | | 2.5% |  | -375.7 | | 0.5% |  | -868 | | 0.0% | minimum | -8019 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 1362.2721 | | Std Dev | 3044.7658 | | Std Err Mean | 14.319631 | | Upper 95% Mean | 1390.3388 | | Lower 95% Mean | 1334.2053 | | N | 45211 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **housing**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | no | 20081 | 0.44416 | | yes | 25130 | 0.55584 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels | **loan**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | no | 37967 | 0.83977 | | yes | 7244 | 0.16023 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels |
| **contact**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | cellular | 29285 | 0.64774 | | telephone | 2906 | 0.06428 | | unknown | 13020 | 0.28798 | | Total | 45211 | 1.00000 |   N Missing 0  3 Levels | **day**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 31 | | 99.5% |  | 31 | | 97.5% |  | 30 | | 90.0% |  | 28 | | 75.0% | quartile | 21 | | 50.0% | median | 16 | | 25.0% | quartile | 8 | | 10.0% |  | 5 | | 2.5% |  | 2 | | 0.5% |  | 1 | | 0.0% | minimum | 1 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 15.806419 | | Std Dev | 8.3224762 | | Std Err Mean | 0.0391409 | | Upper 95% Mean | 15.883136 | | Lower 95% Mean | 15.729702 | | N | 45211 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **month**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | jan | 1403 | 0.03103 | | feb | 2649 | 0.05859 | | mar | 477 | 0.01055 | | apr | 2932 | 0.06485 | | may | 13766 | 0.30448 | | jun | 5341 | 0.11813 | | jul | 6895 | 0.15251 | | aug | 6247 | 0.13817 | | sep | 579 | 0.01281 | | oct | 738 | 0.01632 | | nov | 3970 | 0.08781 | | dec | 214 | 0.00473 | | Total | 45211 | 1.00000 |   N Missing 0  12 Levels | **duration**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 4918 | | 99.5% |  | 1502.94 | | 97.5% |  | 974 | | 90.0% |  | 548 | | 75.0% | quartile | 319 | | 50.0% | median | 180 | | 25.0% | quartile | 103 | | 10.0% |  | 58 | | 2.5% |  | 19 | | 0.5% |  | 8 | | 0.0% | minimum | 0 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 258.16308 | | Std Dev | 257.52781 | | Std Err Mean | 1.2111616 | | Upper 95% Mean | 260.53698 | | Lower 95% Mean | 255.78918 | | N | 45211 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **campaign**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 63 | | 99.5% |  | 21 | | 97.5% |  | 11 | | 90.0% |  | 5 | | 75.0% | quartile | 3 | | 50.0% | median | 2 | | 25.0% | quartile | 1 | | 10.0% |  | 1 | | 2.5% |  | 1 | | 0.5% |  | 1 | | 0.0% | minimum | 1 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 2.7638407 | | Std Dev | 3.0980209 | | Std Err Mean | 0.0145701 | | Upper 95% Mean | 2.7923983 | | Lower 95% Mean | 2.735283 | | N | 45211 | | **pdays**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 871 | | 99.5% |  | 410.94 | | 97.5% |  | 355 | | 90.0% |  | 185 | | 75.0% | quartile | -1 | | 50.0% | median | -1 | | 25.0% | quartile | -1 | | 10.0% |  | -1 | | 2.5% |  | -1 | | 0.5% |  | -1 | | 0.0% | minimum | -1 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 40.197828 | | Std Dev | 100.12875 | | Std Err Mean | 0.4709087 | | Upper 95% Mean | 41.120817 | | Lower 95% Mean | 39.274839 | | N | 45211 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **previous**    **Quantiles**   |  |  |  | | --- | --- | --- | | 100.0% | maximum | 275 | | 99.5% |  | 12 | | 97.5% |  | 5 | | 90.0% |  | 2 | | 75.0% | quartile | 0 | | 50.0% | median | 0 | | 25.0% | quartile | 0 | | 10.0% |  | 0 | | 2.5% |  | 0 | | 0.5% |  | 0 | | 0.0% | minimum | 0 |   **Summary Statistics**   |  |  | | --- | --- | | Mean | 0.5803234 | | Std Dev | 2.303441 | | Std Err Mean | 0.0108332 | | Upper 95% Mean | 0.6015565 | | Lower 95% Mean | 0.5590902 | | N | 45211 | | **poutcome**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | failure | 4901 | 0.10840 | | other | 1840 | 0.04070 | | success | 1511 | 0.03342 | | unknown | 36959 | 0.81748 | | Total | 45211 | 1.00000 |   N Missing 0  4 Levels |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **y**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | no | 39922 | 0.88302 | | yes | 5289 | 0.11698 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels |  |

## Appendix 3 - Initial Regression Run on Data

**Nominal Logistic Fit for y**

**Effect Summary**

| **Source** | **LogWorth** |  | **PValue** |
| --- | --- | --- | --- |
| duration | 918.939 |  | 0.00000 |
| poutcome | 244.545 |  | 0.00000 |
| month | 216.773 |  | 0.00000 |
| contact | 106.890 |  | 0.00000 |
| housing | 52.755 |  | 0.00000 |
| campaign | 18.469 |  | 0.00000 |
| job | 15.465 |  | 0.00000 |
| loan | 11.875 |  | 0.00000 |
| marital | 8.415 |  | 0.00000 |
| education | 5.281 |  | 0.00001 |
| day | 4.185 |  | 0.00007 |
| balance | 1.898 |  | 0.01265 |
| previous | 0.926 |  | 0.11848 |
| pdays | 0.132 |  | 0.73727 |
| default | 0.037 |  | 0.91841 |
| age | 0.018 |  | 0.95923 |

Converged in Gradient, 6 iterations

**Whole Model Test**

| **Model** | **-LogLikelihood** | **DF** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Difference | 5534.342 | 42 | 11068.68 | <.0001\* |
| Full | 10781.135 |  |  |  |
| Reduced | 16315.477 |  |  |  |

|  |  |
| --- | --- |
| RSquare (U) | 0.3392 |
| AICc | 21648.4 |
| BIC | 22023.2 |
| Observations (or Sum Wgts) | 45211 |

| **Measure** | **Training** | **Definition** |
| --- | --- | --- |
| Entropy RSquare | 0.3392 | 1-Loglike(model)/Loglike(0) |
| Generalized RSquare | 0.4224 | (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n)) |
| Mean -Log p | 0.2385 | ∑ -Log(ρ[j])/n |
| RMSE | 0.2661 | √ ∑(y[j]-ρ[j])²/n |
| Mean Abs Dev | 0.1388 | ∑ |y[j]-ρ[j]|/n |
| Misclassification Rate | 0.0982 | ∑ (ρ[j]≠ρMax)/n |
| N | 45211 | n |

**Lack Of Fit**

| **Source** | **DF** | **-LogLikelihood** | **ChiSquare** |
| --- | --- | --- | --- |
| Lack Of Fit | 45168 | 10781.135 | 21562.27 |
| Saturated | 45210 | 0.000 | **Prob>ChiSq** |
| Fitted | 42 | 10781.135 | 1.0000 |

**Parameter Estimates**

| **Term** |  | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- | --- |
| Intercept |  | 3.04878332 | 0.1557813 | 383.02 | <.0001\* |
| age |  | -0.0001127 | 0.0022052 | 0.00 | 0.9592 |
| job[admin.] |  | -0.1574904 | 0.0573697 | 7.54 | 0.0060\* |
| job[blue-collar] |  | 0.15238214 | 0.0558919 | 7.43 | 0.0064\* |
| job[entrepreneur] |  | 0.19961332 | 0.1073101 | 3.46 | 0.0629 |
| job[housemaid] |  | 0.34651121 | 0.1172163 | 8.74 | 0.0031\* |
| job[management] |  | 0.007788 | 0.0526201 | 0.02 | 0.8823 |
| job[retired] |  | -0.4098531 | 0.0785884 | 27.20 | <.0001\* |
| job[self-employed] |  | 0.14084564 | 0.0939164 | 2.25 | 0.1337 |
| job[services] |  | 0.06630667 | 0.0694439 | 0.91 | 0.3397 |
| job[student] |  | -0.5396262 | 0.094007 | 32.95 | <.0001\* |
| job[technician] |  | 0.0185261 | 0.0514671 | 0.13 | 0.7189 |
| job[unemployed] |  | 0.01922267 | 0.0941595 | 0.04 | 0.8382 |
| marital[divorced] |  | -0.0289853 | 0.039269 | 0.54 | 0.4604 |
| marital[married] |  | 0.15046821 | 0.0271586 | 30.70 | <.0001\* |
| education[primary] |  | 0.20323714 | 0.0511325 | 15.80 | <.0001\* |
| education[secondary] |  | 0.01970889 | 0.0352371 | 0.31 | 0.5759 |
| education[tertiary] |  | -0.1757043 | 0.0427359 | 16.90 | <.0001\* |
| default[no] |  | -0.0083406 | 0.0814195 | 0.01 | 0.9184 |
| balance |  | -1.2835e-5 | 5.1476e-6 | 6.22 | 0.0127\* |
| housing[no] |  | -0.3376922 | 0.0219347 | 237.02 | <.0001\* |
| loan[no] |  | -0.2126858 | 0.0299953 | 50.28 | <.0001\* |
| contact[cellular] |  | -0.5955303 | 0.0363458 | 268.47 | <.0001\* |
| contact[telephone] |  | -0.4321559 | 0.0539572 | 64.15 | <.0001\* |
| day |  | -0.0099689 | 0.0024966 | 15.94 | <.0001\* |
| month[jan] |  | 1.28537544 | 0.1036388 | 153.82 | <.0001\* |
| month[feb] |  | 0.17097761 | 0.0659685 | 6.72 | 0.0095\* |
| month[mar] |  | -1.5662338 | 0.0987462 | 251.58 | <.0001\* |
| month[apr] |  | 0.02365666 | 0.0584578 | 0.16 | 0.6857 |
| month[may] |  | 0.42276806 | 0.05096 | 68.82 | <.0001\* |
| month[jun] |  | -0.4299658 | 0.0693109 | 38.48 | <.0001\* |
| month[jul] |  | 0.85445225 | 0.0557602 | 234.82 | <.0001\* |
| month[aug] |  | 0.71756424 | 0.0533275 | 181.06 | <.0001\* |
| month[sep] |  | -0.8504014 | 0.0978927 | 75.47 | <.0001\* |
| month[oct] |  | -0.8577808 | 0.0873153 | 96.51 | <.0001\* |
| month[nov] |  | 0.89705518 | 0.0637588 | 197.95 | <.0001\* |
| duration |  | -0.0041937 | 6.4533e-5 | 4223.1 | <.0001\* |
| campaign |  | 0.09078174 | 0.0101374 | 80.19 | <.0001\* |
| pdays |  | 0.00010268 | 0.0003061 | 0.11 | 0.7373 |
| previous |  | -0.0101524 | 0.006503 | 2.44 | 0.1185 |
| poutcome[failure] |  | 0.60068522 | 0.0513951 | 136.60 | <.0001\* |
| poutcome[other] |  | 0.39720682 | 0.0634887 | 39.14 | <.0001\* |
| poutcome[success] |  | -1.6903708 | 0.0521372 | 1051.2 | <.0001\* |

For log odds of no/yes

## Appendix 4 - Correlation

**Correlations**

|  | **age** | **balance** | **day** | **duration** | **campaign** | **pdays** | **previous** | **job2** | **marital2** | **education2** | **default2** | **housing2** | **loan2** | **contact2** | **month2** | **poutcome2** | **y2** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 1.0000 | 0.0978 | -0.0091 | -0.0046 | 0.0048 | -0.0238 | 0.0013 | 0.1185 | 0.2863 | -0.1736 | -0.0179 | -0.1855 | -0.0157 | -0.0262 | 0.0929 | 0.0355 | 0.0252 |
| balance | 0.0978 | 1.0000 | 0.0045 | 0.0216 | -0.0146 | 0.0034 | 0.0167 | -0.0096 | 0.0257 | 0.0506 | -0.0667 | -0.0688 | -0.0844 | 0.0273 | 0.0946 | 0.0352 | 0.0528 |
| day | -0.0091 | 0.0045 | 1.0000 | -0.0302 | 0.1625 | -0.0930 | -0.0517 | 0.0172 | 0.0071 | 0.0217 | 0.0094 | -0.0280 | 0.0114 | 0.0279 | 0.1020 | -0.0303 | -0.0283 |
| duration | -0.0046 | 0.0216 | -0.0302 | 1.0000 | -0.0846 | -0.0016 | 0.0012 | -0.0101 | -0.0227 | 0.0026 | -0.0100 | 0.0051 | -0.0124 | 0.0208 | -0.0119 | 0.0424 | 0.3945 |
| campaign | 0.0048 | -0.0146 | 0.1625 | -0.0846 | 1.0000 | -0.0886 | -0.0329 | 0.0256 | 0.0314 | 0.0002 | 0.0168 | -0.0236 | 0.0100 | -0.0196 | 0.0549 | -0.0575 | -0.0732 |
| pdays | -0.0238 | 0.0034 | -0.0930 | -0.0016 | -0.0886 | 1.0000 | 0.4548 | -0.0052 | -0.0276 | 0.0071 | -0.0300 | 0.1242 | -0.0228 | 0.2448 | -0.1089 | 0.2285 | 0.1036 |
| previous | 0.0013 | 0.0167 | -0.0517 | 0.0012 | -0.0329 | 0.4548 | 1.0000 | -0.0072 | -0.0127 | 0.0253 | -0.0183 | 0.0371 | -0.0110 | 0.1478 | -0.0356 | 0.2014 | 0.0932 |
| job2 | 0.1185 | -0.0096 | 0.0172 | -0.0101 | 0.0256 | -0.0052 | -0.0072 | 1.0000 | 0.1128 | 0.0582 | 0.0068 | 0.1054 | 0.0710 | -0.0320 | 0.0411 | -0.0430 | -0.0654 |
| marital2 | 0.2863 | 0.0257 | 0.0071 | -0.0227 | 0.0314 | -0.0276 | -0.0127 | 0.1128 | 1.0000 | -0.1262 | -0.0145 | 0.0177 | 0.0369 | -0.0329 | 0.0629 | -0.0185 | -0.0603 |
| education2 | -0.1736 | 0.0506 | 0.0217 | 0.0026 | 0.0002 | 0.0071 | 0.0253 | 0.0582 | -0.1262 | 1.0000 | -0.0090 | -0.0387 | 0.0043 | 0.1515 | 0.0679 | 0.0372 | 0.0513 |
| default2 | -0.0179 | -0.0667 | 0.0094 | -0.0100 | 0.0168 | -0.0300 | -0.0183 | 0.0068 | -0.0145 | -0.0090 | 1.0000 | -0.0060 | 0.0772 | -0.0154 | 0.0150 | -0.0233 | -0.0224 |
| housing2 | -0.1855 | -0.0688 | -0.0280 | 0.0051 | -0.0236 | 0.1242 | 0.0371 | 0.1054 | 0.0177 | -0.0387 | -0.0060 | 1.0000 | 0.0413 | -0.1881 | -0.1739 | -0.0914 | -0.1392 |
| loan2 | -0.0157 | -0.0844 | 0.0114 | -0.0124 | 0.0100 | -0.0228 | -0.0110 | 0.0710 | 0.0369 | 0.0043 | 0.0772 | 0.0413 | 1.0000 | 0.0109 | 0.0216 | -0.0537 | -0.0682 |
| contact2 | -0.0262 | 0.0273 | 0.0279 | 0.0208 | -0.0196 | 0.2448 | 0.1478 | -0.0320 | -0.0329 | 0.1515 | -0.0154 | -0.1881 | 0.0109 | 1.0000 | 0.1738 | 0.1137 | 0.1484 |
| month2 | 0.0929 | 0.0946 | 0.1020 | -0.0119 | 0.0549 | -0.1089 | -0.0356 | 0.0411 | 0.0629 | 0.0679 | 0.0150 | -0.1739 | 0.0216 | 0.1738 | 1.0000 | 0.0337 | 0.0187 |
| poutcome2 | 0.0355 | 0.0352 | -0.0303 | 0.0424 | -0.0575 | 0.2285 | 0.2014 | -0.0430 | -0.0185 | 0.0372 | -0.0233 | -0.0914 | -0.0537 | 0.1137 | 0.0337 | 1.0000 | 0.3068 |
| y2 | 0.0252 | 0.0528 | -0.0283 | 0.3945 | -0.0732 | 0.1036 | 0.0932 | -0.0654 | -0.0603 | 0.0513 | -0.0224 | -0.1392 | -0.0682 | 0.1484 | 0.0187 | 0.3068 | 1.0000 |

**Scatterplot Matrix**

























## Appendix 5 – Summary Statistics

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | **y** | | | --- | --- | --- | | **job** | **no** | **yes** | | admin. | 4540 | 631 | | blue-collar | 9024 | 708 | | entrepreneur | 1364 | 123 | | housemaid | 1131 | 109 | | management | 8157 | 1301 | | retired | 1748 | 516 | | self-employed | 1392 | 187 | | services | 3785 | 369 | | student | 669 | 269 | | technician | 6757 | 840 | | unemployed | 1101 | 202 | | unknown | 254 | 34 |  |  | **y** | | | --- | --- | --- | | **job** | **no** | **yes** | | admin. |  |  | | blue-collar |  |  | | entrepreneur |  |  | | housemaid |  |  | | management |  |  | | retired |  |  | | self-employed |  |  | | services |  |  | | student |  |  | | technician |  |  | | unemployed |  |  | | unknown |  |  | | |  | **y** | | | --- | --- | --- | | **marital** | **no** | **yes** | | divorced | 4585 | 622 | | married | 24459 | 2755 | | single | 10878 | 1912 |  |  | **y** | | | --- | --- | --- | | **marital** | **no** | **yes** | | divorced |  |  | | married |  |  | | single |  |  | |
| |  | **y** | | | --- | --- | --- | | **education** | **no** | **yes** | | primary | 6260 | 591 | | secondary | 20752 | 2450 | | tertiary | 11305 | 1996 | | unknown | 1605 | 252 |  |  | **y** | | | --- | --- | --- | | **education** | **no** | **yes** | | primary |  |  | | secondary |  |  | | tertiary |  |  | | unknown |  |  | | |  | **y** | | | --- | --- | --- | | **default** | **no** | **yes** | | no | 39159 | 5237 | | yes | 763 | 52 |  |  | **y** | | | --- | --- | --- | | **default** | **no** | **yes** | | no |  |  | | yes |  |  | |
| |  | **y** | | | --- | --- | --- | | **housing** | **no** | **yes** | | no | 16727 | 3354 | | yes | 23195 | 1935 |  |  | **y** | | | --- | --- | --- | | **housing** | **no** | **yes** | | no |  |  | | yes |  |  | | |  | **y** | | | --- | --- | --- | | **loan** | **no** | **yes** | | no | 33162 | 4805 | | yes | 6760 | 484 |  |  | **y** | | | --- | --- | --- | | **loan** | **no** | **yes** | | no |  |  | | yes |  |  | |
| |  | **y** | | | --- | --- | --- | | **contact** | **no** | **yes** | | cellular | 24916 | 4369 | | telephone | 2516 | 390 | | unknown | 12490 | 530 |  |  | **y** | | | --- | --- | --- | | **contact** | **no** | **yes** | | cellular |  |  | | telephone |  |  | | unknown |  |  | | |  | **y** | | | --- | --- | --- | | **poutcome** | **no** | **yes** | | failure | 4283 | 618 | | other | 1533 | 307 | | success | 533 | 978 | | unknown | 33573 | 3386 |  |  | **y** | | | --- | --- | --- | | **poutcome** | **no** | **yes** | | failure |  |  | | other |  |  | | success |  |  | | unknown |  |  | |

## Appendix 6 - Graph Builder

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Appendix 7 – Univariate Analysis of Dummy Variables

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Distributions**  **job2**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | 0 | 2241 | 0.04957 | | 2 | 42969 | 0.95041 | | 22 | 1 | 0.00002 | | Total | 45211 | 1.00000 |   N Missing 0  3 Levels | **marital2**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | 0 | 17997 | 0.39807 | | 1 | 27214 | 0.60193 | | Total | 45211 | 1.00000 |   N Missing 0 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ducation3**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | 0 | 31852 | 0.70452 | | 1 | 13359 | 0.29548 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels |  |

## Appendix 8 – SAS JMP Imputation Report

**Imputation Report**

2145 missing values were replaced by least squares imputation. A shrinkage estimate was used, with off-diagonals scaled by a factor of 0.99739. 2018 rows and 2 columns were affected. There were 3 missing value patterns across columns. Imputed values colored light blue.

1. [↑](#footnote-ref-1)