1. General

The provided training dataset is used as the basis of a predictive binary classification model to determine from the various predictor variables whether a customer will respond to a marketing campaign. The predictor variables may be classified into the following three categories:

- *Demographic* – describe the age, education, marital status, and profession of the subject,
- *Customer Finance* – describe the financial characteristics of the subject,
- *Customer Marketing* – describe previous marketing contacts with the customer,
- *Economic* – describes various economic indicators.

There is a total of 7414 records in the training dataset, all of which are recorded with a value of *yes* or *no* for `responded`. Enumerating the training dataset reveals the `responded` dependent variable to be highly imbalanced, with 11.3% of the values coded as *yes* while the remaining 88.7% were coded as *no*.

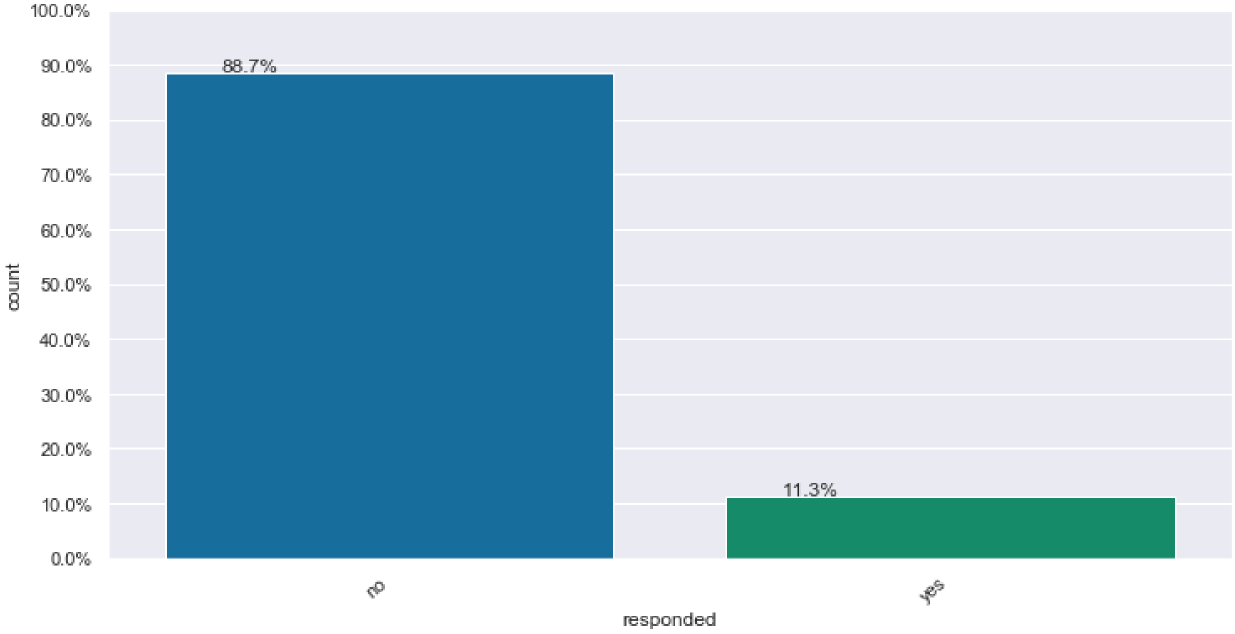


Figure 1 – Distribution of `responded`

2. Exploratory Analysis – Predictor Variables

The predictor variables will be examined for additional insights. For most categorical variables, count plots will be performed to ascertain their distribution

2.1. `custAge`

There are 5610 non-null values for `custAge` out of a total of 7414 total records. A distribution of `custAge` is shown in Figure 2.

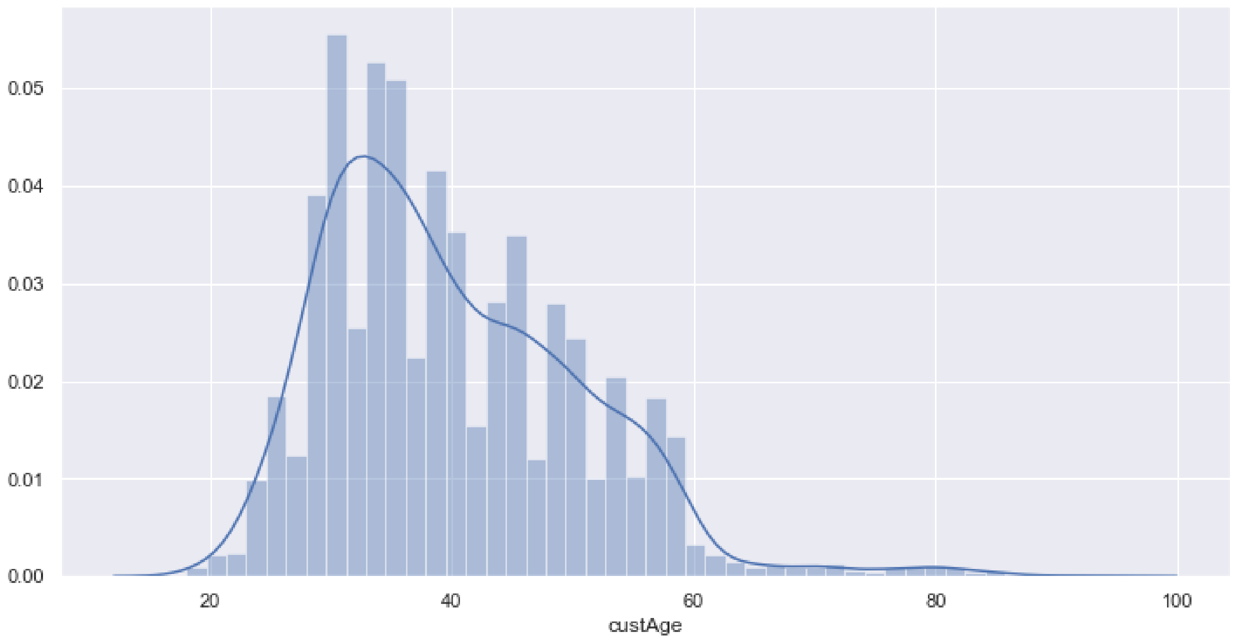


Figure 2 – Distribution of **`custAge`** *(omitting NaN values)*

Because of the large proportion (24.3%) of records with missing values for **`custAge`**, it may be imprudent to simply drop these records. Possible methods of addressing the missing values are:

- Dropping the records which contain missing values for **`custAge`**
- Assign the overall aggregated mean age to null values. To do so assumes that the true values of the missing data are normally distributed about the mean, *e.g.*, that there is no confounding factor resulting in the true mean of the missing data being different than that of the data that is present. As can be seen in Figure 3, this is not the case.
- Assign a *categorical* mean age of subjects matching the same value of another category. For example, subjects with a **`profession`** of *student* for which **`custAge`** is missing can be assigned the mean **`custAge`** of all students. This is likely preferable to the first option due to the fact that there are substantial differences in the mean ages of various classes of profession, as can be seen in Figure 3. Table 1 presents the mean **`custAge`** for each profession.

For the initial analysis, each null/NaN value for **`custAge`** will be replaced by the mean value for that customer's profession.

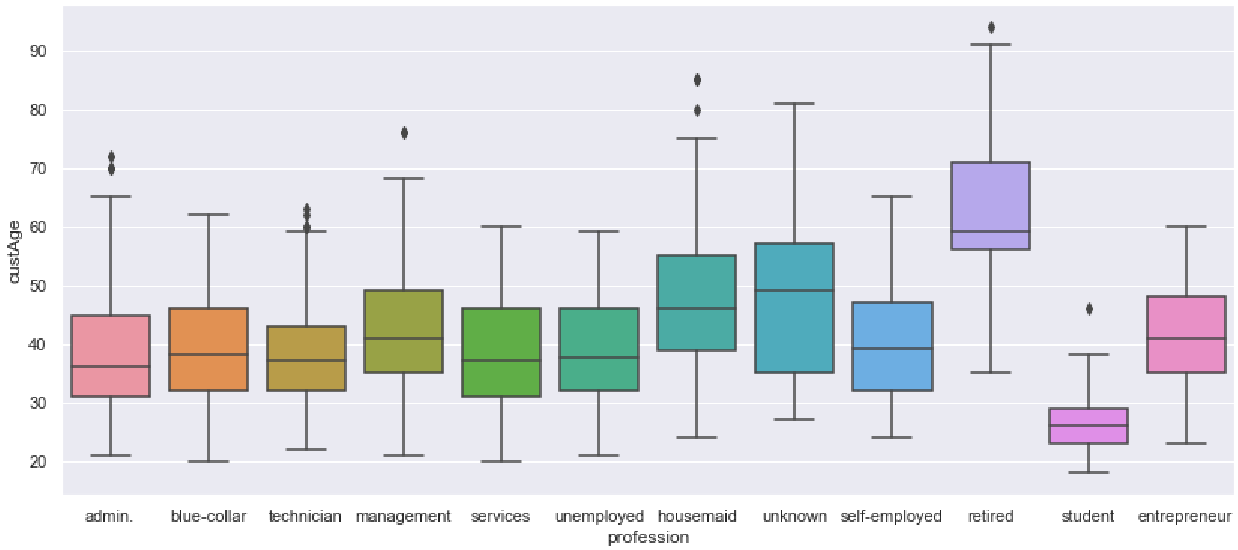


Figure 3 – `custAge` by class of `profession`

Table 1 – mean custAge for each profession class

profession	***mean(custAge)***
admin.	38.07
blue-collar	39.26
entrepreneur	41.34
housemaid	47.21
management	42.04
retired	62.74
self-employed	40.10
services	38.58
student	26.37
technician	38.09
unemployed	38.88
unknown	47.96

2.2. Profession

All 7,414 records are coded with a class value for **`profession`**. A count of each class is shown in Table 2.

Table 2 – Count for each class of profession

profession	***count***
admin.	1885
blue-collar	1665
technician	1212
services	719
management	536
retired	307
entrepreneur	275
self-employed	248
housemaid	187
unemployed	173
student	146
unknown	61

2.3. Marital

All 7,414 records are coded with a class value for `marital`.

Table 3 – Count for each class of marital

marital	***count***
married	4445
single	2118
divorced	843
unknown	8

2.4. Schooling

5,259 of the 7,414 records have non-null values for schooling (70.9%). Of the records with non-null values, the distribution is as follows:

Table 4 – Count for each class of schooling

schooling	***count***
university.degree	1554
high.school	1216
basic.9y	784
professional.course	664
basic.4y	534
basic.6y	275
unknown	231
illiterate	1

It is noted that the null values for `schooling` are not included in the *unknown* class. As these null values truly constitute unknown values, they will be recoded as such for any relevant modeling. The single value of *illiterate* is insubstantial to any model effects; it will be recoded as 'unknown' or may most logically be grouped with the next-lowest level of schooling, *basic.4y*.

Table 5—Count of schooling after modification

schooling	***count***
unknown	2386
university.degree	1554
high.school	1216
basic.9y	784
professional.course	664
basic.4y	535
basic.6y	275

It is also noted that the sub-secondary levels of education (*basic.4y, basic.6y, basic9y*) are candidates for merging, contingent upon further analysis.

2.5. Default

The **`default`** category refers to whether a customer has previously defaulted upon an account. The categorical counts are as follows:

Table 6 – Counts for default

default	***count***
no	5981
unknown	1432
yes	1

The single positive response for this category (out of 7,414 records) diminishes the utility of this categorical variable as a predictor.

2.6. Housing

All 7,414 records are coded with a class value for **`housing`**. This categorical variable refers to whether the customer has a loan for housing.

Table 7—Counts for **`housing`**

housing	***count***
yes	3840
no	3406
unknown	168

2.7. Loan

All 7,414 records are coded with a class value for **`loan`**. This categorical variable refers to whether the customer has a personal loan outstanding.

Table 8—Counts for `loan`

loan	***count***
no	6099
yes	1147
unknown	168

2.8. Contact

Refers to the customer's preferred means of telephone contact; with *telephone* apparently denoting the traditional land-line. Non-null values for this category are populated into all records.

Table 9—Counts for preferred contact

contact	***count***
cellular	4731
telephone	2683

2.9. Month

Month is believed to pertain to the last month of contact. As no context is given relative to the date of the survey, it is generally not possible to deduce the *time that has elapsed* since the contact event recorded. Apart from there being any possible seasonal correlation between the month of contact and **`respond`**, the value of this variable is limited by its lack of context. In such circumstances, the author would normally converse with the customer to develop the appropriate context.

Table 10—Counts for month

month	***count***
mar	93
apr	487
may	2529
jun	939
jul	1229
aug	1112
sep	112
oct	145
nov	741
dec	27

2.10. Day of Week

The `day_of_week` variable represents the day of the week on which the last contact was made. As with the `month` variable, it is not possible to deduce a discrete value of elapsed time from a value of `day_of_week`; however, it may be useful in determining whether contacts made on a specific day of the week are more likely to produce responses.

Table 11—Counts for Day of Week

day_of_week	***count***
mon	1441
tue	1341
wed	1295
thu	1379
fri	1247

Contacts were only made on weekdays, with an approximately uniform distribution among those days. 6,703 of 7,414 records (90.4%) were coded with values for `day_of_week`.

2.11. Campaign

According to the furnished documentation, campaign represents the number times that the customer was contacted. This is a discrete numerical value ranging in values from 1 to 39. All records contain a value for `campaign`.

A histogram (Figure 4) shows the preponderance of customer contact counts to be five or less. As `campaign` is a discrete value, it will be binned into appropriate ranges and treated as a categorical variable.

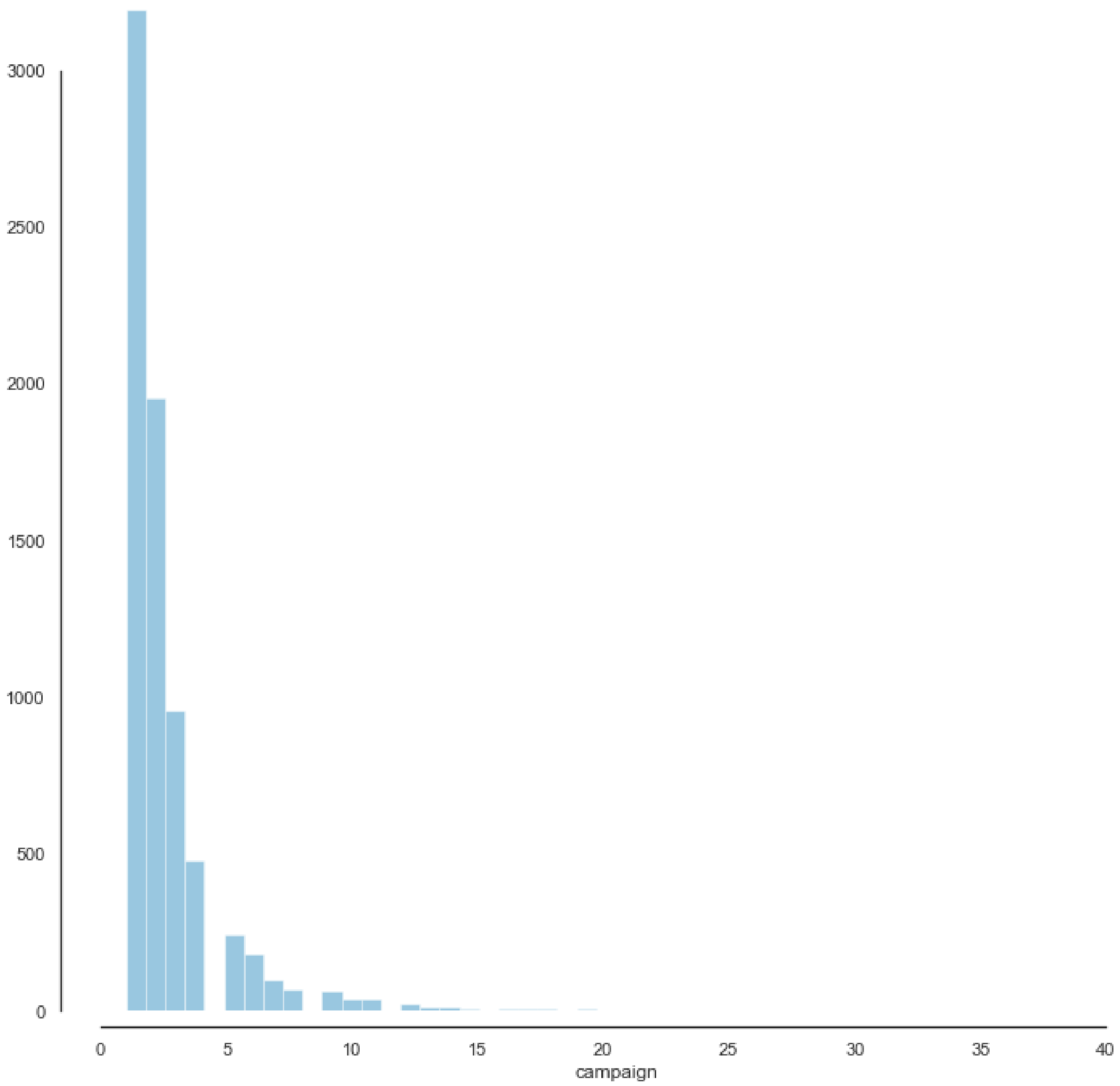


Figure 4—Histogram of the campaign category

Binning the campaign values into four bins of geometrically increasing size shows some potential for using campaign as a predictor:

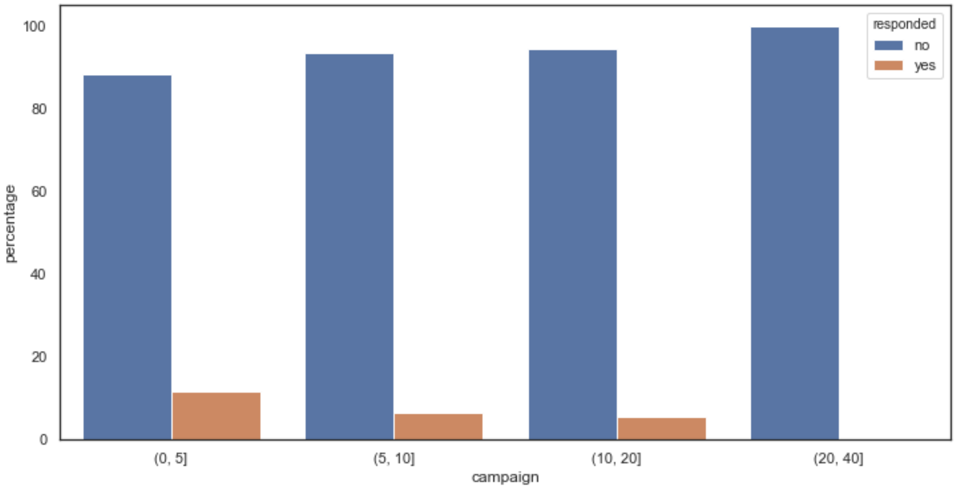


Figure 5—Counts of responded, grouped by campaign counts, when distributed to four bins

2.12. Pdays

The variable **`pdays`** indicates the number of days that passed between the date that the customer was last contacted and the reference date. A value of '999' indicates that the customer was not previously contacted. The '999' flag poses an interesting challenge in that it is essentially a categorical indicator. Counts of the **`pdays`** variable are as follows:

Table 12—Count of `pdays`

pdays	***count***
0	2
1	3
2	13
3	82
4	22
5	7
6	85
7	15
8	3
9	15
10	8
11	4
12	11
13	6
14	5
15	4
16	2
17	2
21	1
22	1
999	7123

It is evident that the overwhelming proportion of customers had never been contacted before. Of those who had, the maximum time that had elapsed was 22 days. To simplify the model and also address the issue of the '999' flag, the pdays values will be binned and treated as categorical. Since the largest quantity of actual days is 22, we will bin from 23 to 1000 to catch the '999' flag, and divide the remainder into approximately equal ranges. A distribution of pdays when binned in this manner is presented in

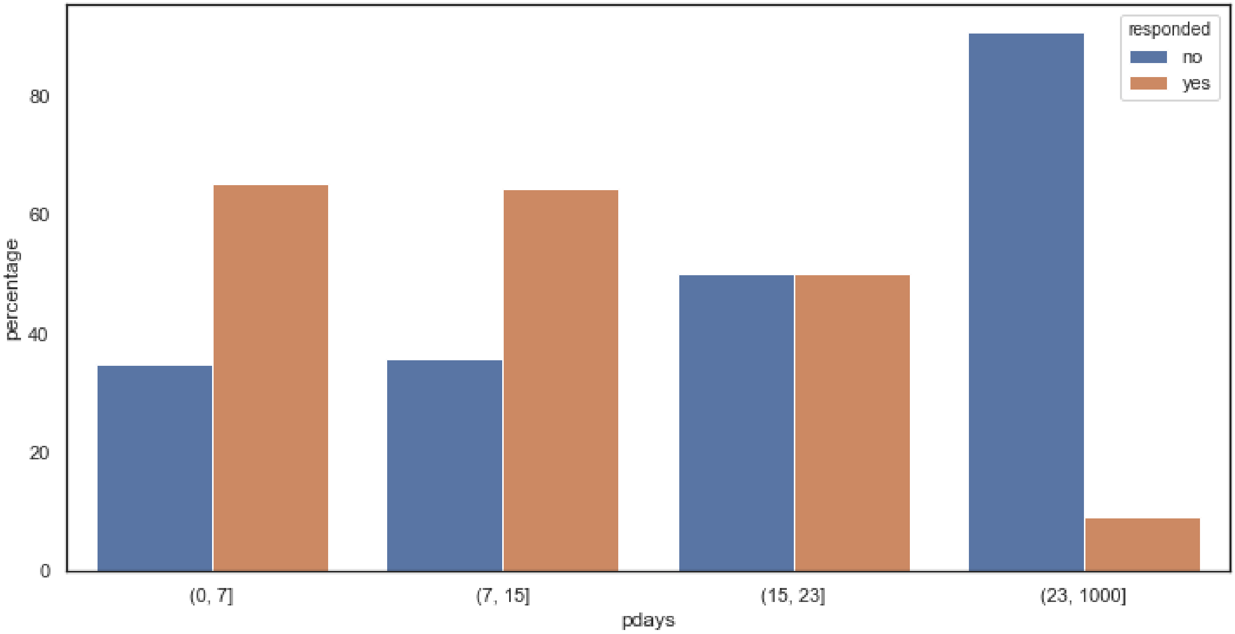


Figure 6—Distribution of values of `pdays` *when binned*

This binning exploration reveals a possible relationship between pdays and responded.

2.13. Previous

The **`previous`** variable indicates the quantity of contacts initiated with the subject prior to the current campaign. Each customer was contacted between zero and 6 times, inclusively prior to the current campaign. A distribution of the outcomes is presented in

Table 13—Count of the variable, `previous`

previous	***count***
0	6350
1	855
2	147
3	41
4	13
5	7
6	1

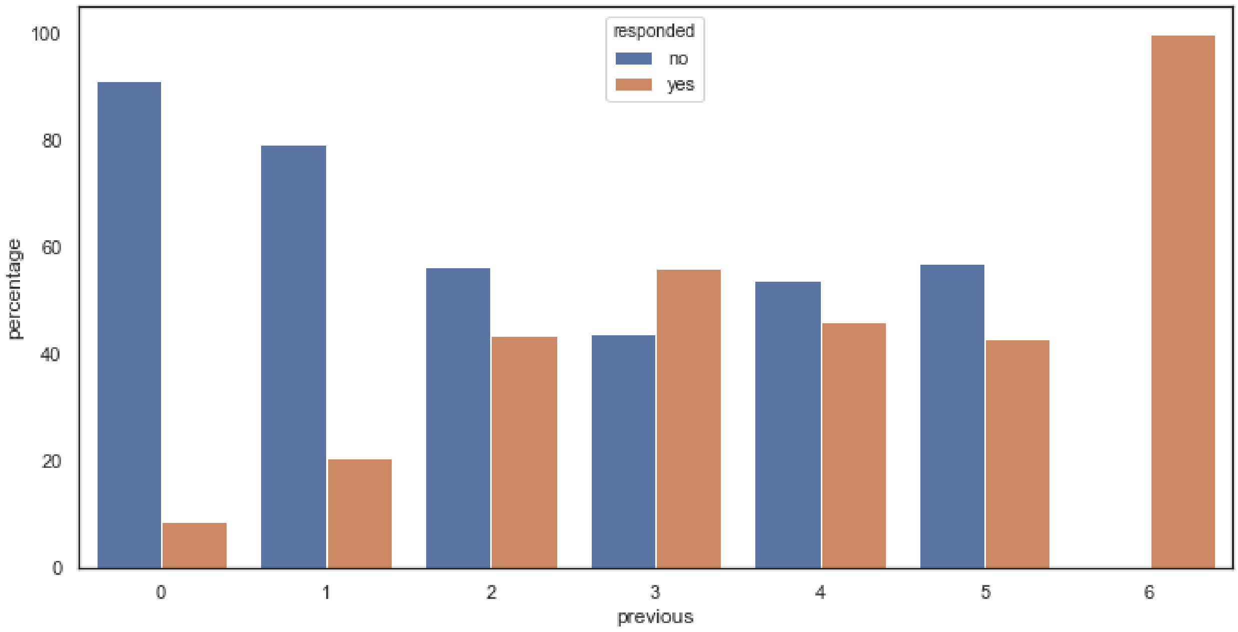


Figure 7—Outcome distribution of the variable, **`previous`**.

2.14. Poutcome

The poutcome variable indicates the outcome of the previous marketing campaign. It is a categorical variable with no missing values. Its context is somewhat nebulous, in that the differentiation between a '*failure*' and a '*nonexistent*' outcome is not clear. In terms of outcome, their proportions are very similar, as shown in Figure 8. For this study, it will be assumed that *failure* and *nonexistent* are synonymous, as neither constitutes a success. In an actual business context, dialogue would be opened with the customer to clarify the context of these values.

Table 14—Count of the poutcome variable

poutcome	***count***
nonexistent	6350
failure	800
success	264

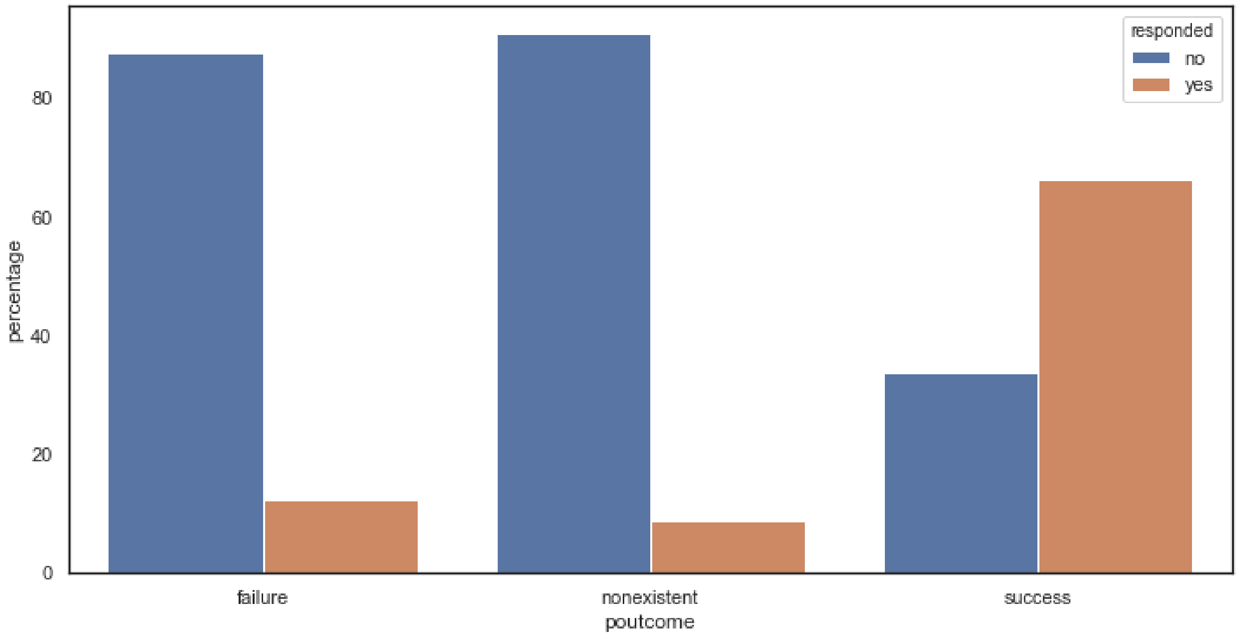


Figure 8—Distribution of **`poutcome`***, grouped by value of* **`responded`***.*

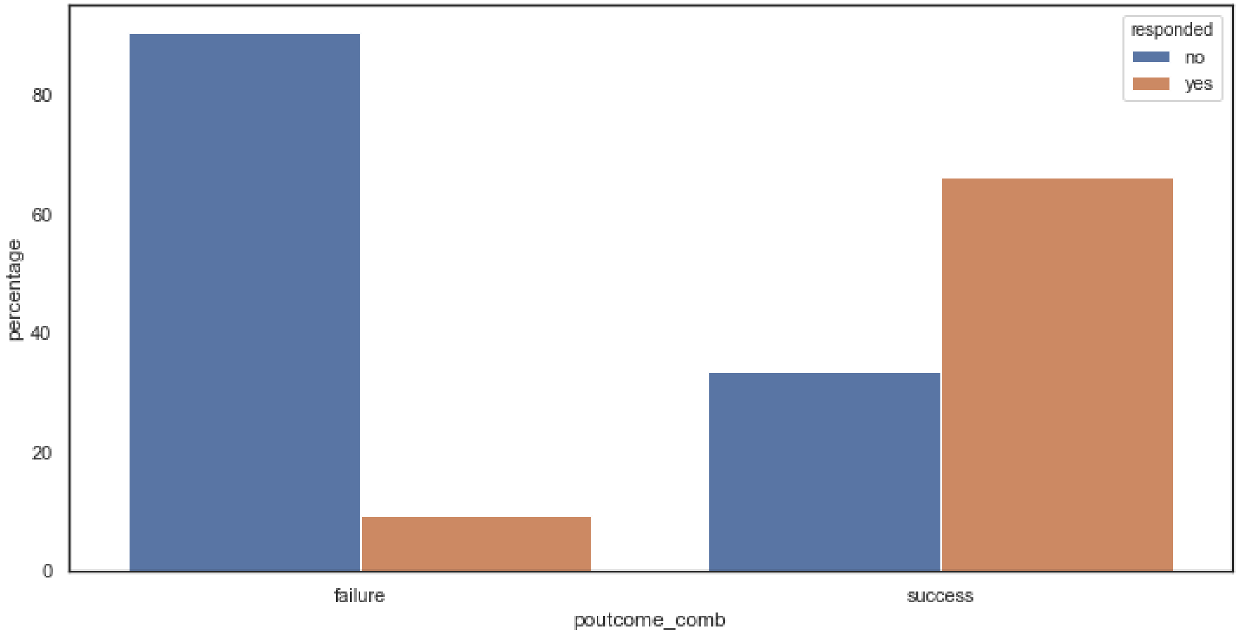


Figure 9—Distribution of **`poutcome`***, with* failure *and* nonexistent *values combined.*

2.15. Pmonths

The **`pmonths`** variable appears to simply be the **`pdays`** feature divided by 30. It will be omitted due to its obvious correlation with **`pdays`**.

2.16. pastEmail

The pastEmail variable quantifies the number of emails sent to the customer prior to the current campaign. Each customer received from zero to 18 e-mails, inclusive; with the vast majority receiving zero.

Table 15—Count of previous e-mails to customer.

pastEmail	***count***
0	6495
1	224
2	296
3	162
4	113
5	34
6	40
7	2
8	18
9	6
10	6
12	10
14	1
15	3
16	2
18	2

2.17. Economic Indicators

The variable emp.var.rate appears to denote the Employment Variation Rate, an economic indicator often used in European countries. Other economic indicators are the Consumer Price Index, the Consumer Confidence Index, the Euribor 3-month indicator, and the number of employees. Each of these variables are continuous and numeric, and there are no missing values among the customers.

3. Correlation Exploration

In order to determine which numeric features share a high degree of correlation, a heatmap plot of Pearson's correlation values is produced, as shown in Figure 10. Notable correlations are:

Feature 1	***Feature 2***	***Pearson's Correlation***
euribor3m	emp.var.rate	0.97
nr.employed	emp.var.rate	0.9
nr.employed	Euribor3m	0.94

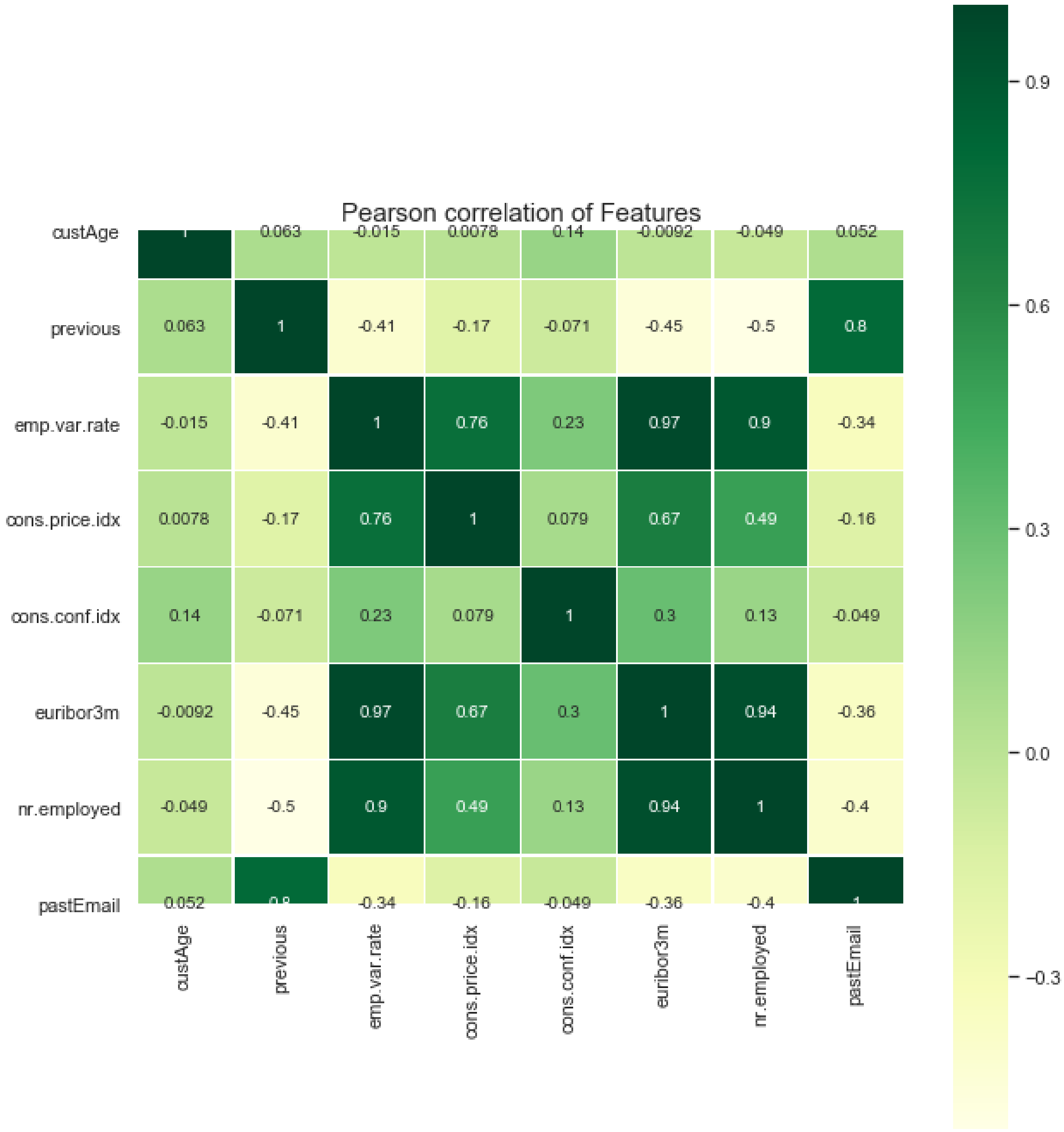


Figure 10—Correlation of Features

4. Further Preprocessing

Executing the `duplicated` method shows there to be 78 redundant rows in the dataframe. These will be dropped. The modifications to the dataframe are:

- Values of `campaign` are binned and converted to categorical
- Values of `pdays` are binned and converted to categorical

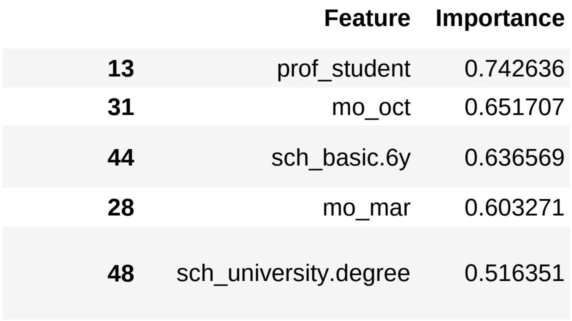
- Values of **poutcome** classified as *nonexistent* are considered synonymous with *failure*, and recoded as such
- The **pmonths** category is dropped to eliminate multicollinearity with **pdays**
- Missing values of **schooling** are consolidated into the extant *unknown* classification
- A single instance of the classification *illiterate* in **schooling** is consolidated into the extant *basic.4y* classification
- Each missing value for **custAge** is replaced with the mean **custAge** for that customer's profession
- After one-hot encoding, there is a 1:1 correlation between the *unknown* class of **housing** and the *unknown* class of **loan**. The *unknown* class of **loan** is dropped
- **emp.var.rate** and **nr.employed** are dropped due to high degree of correlation with **euribor3m**
- Rows which were exact duplicates were assumed to be unintentional duplicates and dropped

5. Modeling and Results

As the response variable is significantly imbalanced, a logistic regression model will be utilized, which handles imbalances well. It will be run with **class_weight ='balanced'** to automatically adjust weights in inverse proportion to the class frequency of the response variable.

After the above preprocessing is performed, the categorical variables are one-hot encoded, and a logistic regression is run. The top five features, ranked by their logistic regression coefficients are:

Table 16 – Feature Importance table, logistic regression



	Feature	Importance
13	prof_student	0.742636
31	mo_oct	0.651707
44	sch_basic.6y	0.636569
28	mo_mar	0.603271
48	sch_university.degree	0.516351

As can be interpreted from this table, the following are the strongest indicators of a positive value of `responded`:

- Customer is identified as a student
- Customers last previously contacted in March or October
- Customers of schooling *basic.6y* (interpretable as those customers who had completed more than 4 years of primary schooling, but no more than 6).
- Customers with a university degree

Receiver Operating Characteristic is used to evaluate relative fit of the model, which, as described, produces an *Area Under Curve* value of 0.7902.

After performing the same transforms to the test dataset, the logistic model was applied, producing an AUC value of 0.7900.