Contents

"Analysis of estimation approaches for large complex data"

1. Introduction and Overview

1.1. Introduction

In today's fast-paced evolving world, marketing plays a crucial role in a lot of businesses and organizations. Improving the marketing communications process will lead to growth and more profits. A lot of banks are using marketing campaigns in order to promote their products. Telemarketing is when the marketing of products is implemented by means of telephone calls. Contacts can be divided in inbound and outbound, based on who contacted who. For example, if an agent calls a client to sell a product, then it is considered outbound, but if the client calls the contact center, it is considered inbound.

1.2. Overview of the analysis

Before starting our analysis, we prepared our data by removing all the unnecessary variables, by rearranging the levels of the factors in broader categories and by transforming variables. We also did some descriptive and exploratory analysis in order to find relations between the variables.

After the initial processing of the dataset, we fitted a logistic regression classification algorithm to the whole dataset and generated some estimations of the coefficients along with some inferences. Afterwards, we implemented the divide and recombine (D & R) approach by using ten and twenty splits. To divide the data in such a way we used random replicate division and afterwards we fitted a logistic regression model to each one of the random subsets created and then we recombined the results of those computations. That way, we were able to obtain all-data estimates and generate confidence intervals for the cases of using D & R with ten and twenty splits.

After comparing the results of the different estimation approaches, we concluded that the implementation of the D & R approach with ten splits results in relatively good estimations. We also claimed that as the number of splits increases, the effectiveness of the method decreases and the computational time increases and vice versa. Furthermore, we saw that by using twenty splits we faced some other problems such as data multicollinearity and inability to calculate coefficient estimates of categorical variables whose levels corresponded to a very small portion of the dataset when splitting the data.

2. Data preparation

2.1. Description of the dataset

The data refers to telemarketing phone calls and are collected from one of the retail bank, from May 2008 to June 2010. There are 39883 phone contacts and 22 attributes in this data set. We will work with the whole dataset, so the "CODE" variable does not have any meaning, so we removed it.

Our aim was to use a logistic regression model for predicting a successful contact (the client subscribes to the product) and compare the results produced by using all the data and by using the divide and recombine approach with 10 and 20 splits and see how much they agree. Beside our output variable (subscribed), we have 20 attributes which refer to the 4 different groups, as shown in Table 1. The descriptions of these attributes can be found in the Table 1 in Appendix.

Table 1: Variables in the dataset

	Variables Names
Bank client data	age, job, marital, education, default, housing, loan
Last contact of the current campaign	contact, month, day_of_week, duration
Other attributes	pdays, previous, poutcome
Social and economic context attributes	emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

2.2. Data cleansing

Our original dataset is a data frame and at first glance it does not look to have any missing values (na, null, nan) or infinite values. Also, we do not have any duplicate rows in our data. For a better exploratory analysis, we decided to divide our data according to their data type into factors, numeric and integers in order to have a clearer view of the results, as shown in Table 2. We also created some basic descriptive measures for the numeric and integers variables along with some frequency tables for the factors, from which we saw the distribution of counts per category.

Table 2: Variables according to their data type

Variable data type	**Variables Names**
Factor	job, marital, education, default, housing, loan, contact, month, day_of_week, poutcome, SUBSCRIBED
Numeric	emp.var.rate, cons.price.idx, euribor3m, nr.employed
Integer	age, duration, campaign, pdays, previous

2.2.1. Factor attributes

At first we saw that there are some missing records which were kept and labeled as “unknown”. From our analysis conducted later on, we observed that the “unknown” levels of the loan and housing variables created some (multicollinearity) problems, so we decided to remove these values from the dataset.

We also observed that the default attribute was binary (yes/no) and 99.9% of its non-missing values corresponded to only one of the two possible outcomes, which rendered it useless. So, we removed this variable from our dataset.

In the summary of job and education variables we saw a category named “(Other)”. The summary function did not show all the levels of the factors. By further exploration we observed that these variables had also missing values hidden inside them. Specifically, there were 312 (0.8%) and 1606 (4.1%) observations in job and education variables accordingly. We decided to group the levels of some variables into broader categories, which will also be proven useful in terms of model interpretation.

At first, we recoded the levels of education variable into the six different categories (i.e. illiteracy, primary, secondary, professional course, university and unknown). These educational stages, just like the days of week, are considered to have an order in their values. So, for clearer and simpler results in our analysis, we made sure that the levels will be ordered from Monday to Friday and from illiterate to university accordingly.

Next, we reclassified the levels of job variable from 12 to 3, because some of those had very small frequencies compared to others. So, we decided to create broader groups of white and blue collar workers, along with other types of jobs. We also grouped the month column into four groups based on the seasons (i.e. spring, summer, fall, winter).

2.2.2. Integer & Numeric attributes

Next, we continued our analysis by computing some descriptive measures for the integer variables in our dataset. At first, we observed some strange results in the summary of “pdays” variable. The reason was the existence of a value (“999”) which was used as a descriptor for a client who was not previously contacted. Almost 97% of the values of “pdays” variable are equal to “999”, which renders it unusable.

In order to make better use of this variable, we created a new binary coded one, depending on whether the client was previously contacted or not (‘p_contact’). We made sure to assign labels in this new factor variable, in order to give meaning to the coded levels. Then we removed “pdays” from our dataset.

We also observed that the “previous” variable had only six unique values. A lot of them were equal to zero, but due to the fact that we also had a lot of observations with different values (about 12%) we decided that we could make use of this variable and we did not removed it.

Finally, we had a variable that described the duration of the last contact, which of course is only known after a call is performed. But, after the end of the call, the outcome -whether the client had subscribed or not- will obviously be known. Thus, we removed this variable from the data in order to have a realistic predictive model.

We also detected some outliers for the previous and campaign variables, but considering their interpretation, values and descriptive measures we decided not to take any further action. Finally, the numeric variables were all measured in different scales with different ranges. For example, the employment variation rate was used as a quarterly indicator while the euribor 3 month rate was used as a daily indicator. In order to bring them into a common scale (between 0 and 1), without distorting differences in the ranges of values we could normalize them, but the purpose of the project is to compare the different estimation approaches which is why we decided that is was not necessary.

After all these modifications, we created a file with the updated data, which we will use further on for our analysis. Our dataset has now 38,928 rows and 19 columns.

2.3. Bivariate Analysis

At first, we divided the numeric variables in the dataset by the levels of the dependent variable and then computed their mean and the variance for each level. By looking at the results we saw that the variable named “cons.price.idx” had almost equal mean and variances for both levels of the subscribed variable. So, we may assume that this feature will not be proven helpful to our analysis.

Also, we computed correlations and made correlation plots of our dependent variable with the continuous variables to investigate if there are any associations implied by the dataset. We used the empirical values of correlations and we sorted the computed correlations by those who had strong and medium correlation with subscription. In Figure 1 below, we presented the correlation coefficients between our target variable and the numeric attributes.

Figure 1: Correlation coefficients for the numeric variables with the dependent

Based on the results from the correlations we concluded that the employment variation rate, euribor 3 month rate and the number of employees are almost perfectly correlated. So we should include only one of them in our final model, because if we include collinear variables then the estimated coefficients and standard errors will have really big values and thus their interpretation will not have any meaning.

These three attributes appeared to have the higher effect on subscription. They are also negative correlated with our target variable, e.g. the number of clients who subscribe to a term deposit will decrease as the number of employments increases. We also observed a high correlation between the employment variation rate and the consumer price index.

Furthermore, we observed that some variables (i.e. 'campaign' and 'cons.conf.idx') seemed to be irrelevant to our analysis, because they had almost zero correlation with the subscription variable, thus they would not add any information to the subscription of a term deposit.

For the factor variables we decided to create some contingency tables in order to examine their relationships with our target variable. Our conclusions from these tables were based on chi-square test. We observed that the existence of a personal or housing loan and the weekday the last contact was performed, did not affect the outcome of subscription. We also observed that whether a client has a housing or a personal loan is related to each other, just like the previous contact of client is related to the outcome of the previous marketing campaign.

To sum up, from the correlations coefficients and the contingency tables we concluded that the result of the subscription does not seem to be affected by five different variables (i.e. 'campaign', 'cons.conf.idx', 'housing', 'loan' and 'day_of_week').

3. Initial analysis for logistic regression

3.1. Problematic values

Our dependent variable is binary and examines whether a client has subscribed to a term deposit or not. That's why we decided to use a logistic regression model, in order to understand the factors that would influence a client to subscribe.

At first, we fitted a model which contained all the variables in the dataset, along with all the unknown values in some factor's levels. We computed the summary of the model and we observed that the unknown level of loan variable had NA as values. This is usually occurred due to strong correlation between our independent variables and can be avoided by having one less dummy variable. So, we excluded the unknown level from the loan variable.

In the next model we fitted, we observed that the estimations of the coefficients and the standard errors of the unknown level for marital variable and the illiterate level for the education variable had really high values. That means that we cannot estimate them and that their existence in the model does not affect the rest of the coefficients.

The fact that we had grouped some variables into broader categories do not change the results taken from the logistic regression. But we should be careful of the reference level used in those categories. For example, we fitted a model where the illiterate level was used as reference and that resulted into high values for all the other levels of the education variable.

3.2. Hypothesis testing of the model

Regarding the hypothesis testing of the model, we first tested whether a variable has a statistically significant effect in our model. We made use of the Wald test, which corresponds to the same p-value with the one in the summary function, but we can also use it in order to test for an overall effect of a factor. The order in which the coefficients are given in the table of coefficients is the same as the order of the terms in the model.

Furthermore, we compared our model with the saturated model[1] and concluded that our model fits "well". Then we compared our model against a model containing only the intercept (null model). The first approach we used to test this was via the likelihood ratio test, where the difference between the null deviance and the model's deviance is distributed as a chi-squared with degrees of freedom equal to the null df[2] minus the model's df. For the second approach we used analysis of variance (anova). Both approaches concluded that our model as a whole fits significantly better than an empty model.

[1] A saturated model is one in which there are as many estimated parameters as data points. By definition, this will lead to a perfect fit, but will be of little use statistically, as we have no data left to estimate variance.

[2] Degrees of freedom

3.3. Visualization

Before continuing with our analysis, we showed in Figure 2, the distribution of our target variable. From there, we observed that almost 90% of the clients had not subscribed to the term deposit.

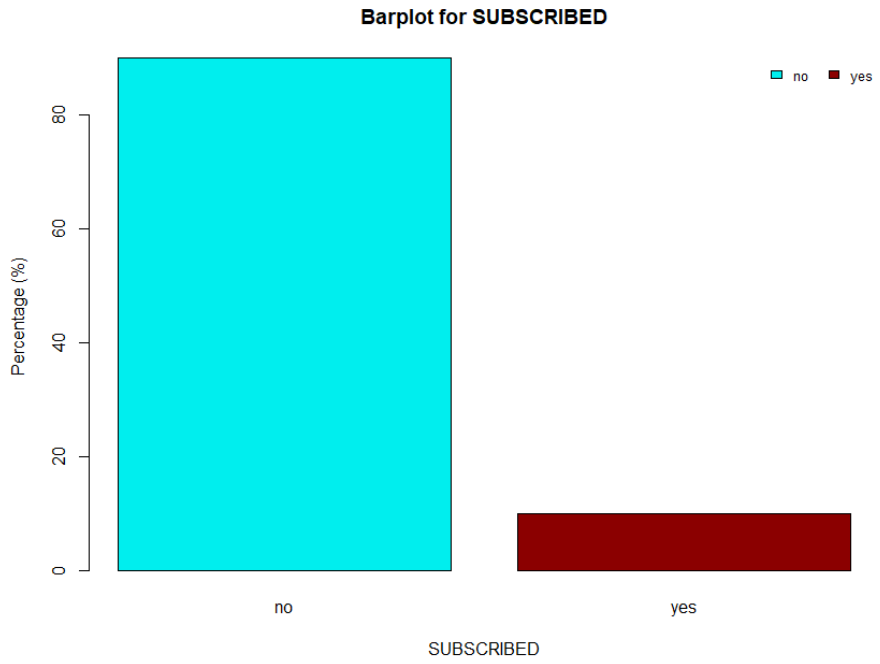

Figure 2: Boxplot for the dependent variable

It has to be noted that logistic regression has some assumptions that should be satisfied in order to be implemented, which will be briefly presented below. First of all, binary logistic regression requires the dependent variable to be binary, just like our subscribe variable is, which has only two possible outcomes ("yes" vs "no"). Secondly, logistic regression requires the observations to be independent of each other.

Thirdly, logistic regression requires there to be little or no multicollinearity among the independent variables. This could be checked the GVIF[3] function. As a rule of thumb, a GVIF value that exceeds 5 or the square root of 10 indicates a problematic amount of collinearity. Fourthly, logistic regression assumes linearity of independent variables and log odds. This could be tested by including in the model interactions between the continuous predictors and their logs via the Box-Tidwell test. Finally, logistic regression typically requires a large sample size.

[3] GVIF stands for Generalized Variance Inflation Factor

4. Divide and Recombine (D & R)

4.1. Introduction

Divide and Recombine (D&R) is a framework mainly used for performing statistical analysis of large complex data (a.k.a. big data) with nearly as much flexibility and ease as with small datasets. It is suited for situations where the number of cases outnumbers the number of variables.

It is accomplished by creating meaningful divisions of the data, applying analytical methods (e.g. fitting a model) to each subset independently and recombining the results of the computations with applying an analytic method independently to each subset (e.g. averaging the model coefficients from each subset) to yield a statistically valid, although not always exact, result for the analytic method.

It has to be noted that there are many forms of divisions as well as recombination's and that D&R differs from MapReduce[4] (although it uses it very much for its operations). In order to perform such an analysis, we decided to use the datadr package in R, which is one component of the DeltaRho environment for the analysis of large complex data.

Each one of the subsets created from the D & R approach is represented as a key[5]-value[6] pair (they are represented as lists of R objects), whose collections form the basic input and output types for all D&R operations.

The collections of key-value pairs are distributed data objects (ddo)[7], or in the case of the value being a data frame, distributed data frames (ddf)[8]. Essentially, these distributed data objects are stored as a list where each one of its elements contained a key-value pair. The *key was the label that uniquely identified each subset and the value was the subset* of the data corresponding to the key.

We wanted to use the divide and recombine approach using 10 and 20 splits. The division of the data can be done in different ways. In our case, we decided to implement a random replicate division, which partitions the data into random subsets based on the approximate desired number of rows for each subset. With random replicate division the observations are seen as exchangeable, with no conditioning variables considered and the results are often approximations.

---

[4] A programming model and an associated implementation used for processing and generating big data sets with a parallel, distributed algorithm on a cluster.
[5] A label that uniquely identifies a subset or an output
[6] The subset or output corresponding to the key
[7] A collection of key-value pairs that constitutes a set of data with arbitrary data structure (but same structure across subsets)
[8] A distributed data object where the value of each key-value pair is a data frame

After specifying a data division, we fitted a logistic regression model to each subset of the division in order to model the output variable as to whether a client has subscribed a term deposit or not and then recombine the results of those computations. This way, we were able to apply such an analytical method across the entire dataset from within the D&R paradigm, thus obtaining all-data estimates.

4.2. Fitting logistic regression to the data

In order to be able to compare the results produced by the D & R approach, we began by fitting a logistic regression model to the whole dataset. Our binary dependent variable (i.e. subscribed) belonged to the binomial exponential family, which is why we decided to use logit as link function[9] (Note: we could use probit as well).

After computing the summary of the model, we plotted the estimations of the coefficients of the logistic regression model as shown in Figure 3. In this plot, we can also see the "neutral" line, which is the vertical intercept that indicates no effect (x-axis position 1).

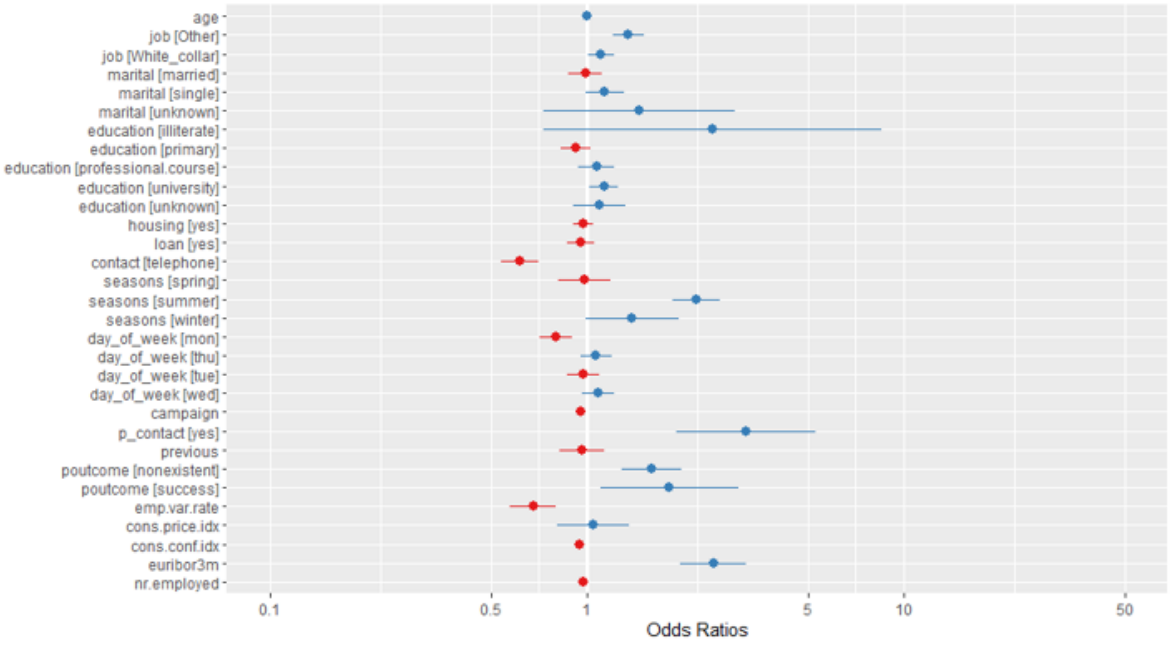

Figure 3: Plot for the coefficient estimations of the logistic regression model

Next, we used the above results to compare them with the D & R approach by using ten and twenty splits.

[9] A link function identifies a function of the mean that is a linear function of the explanatory variables.

4.3. Implementation of Divide and Recombine (10 & 20 splits)

At first, we chose a division that provides about 3,893 rows in each subset in order to split the observations into ten random subsets. Below, we see a plot regarding the distribution of the number of rows in each of the ten subsets.

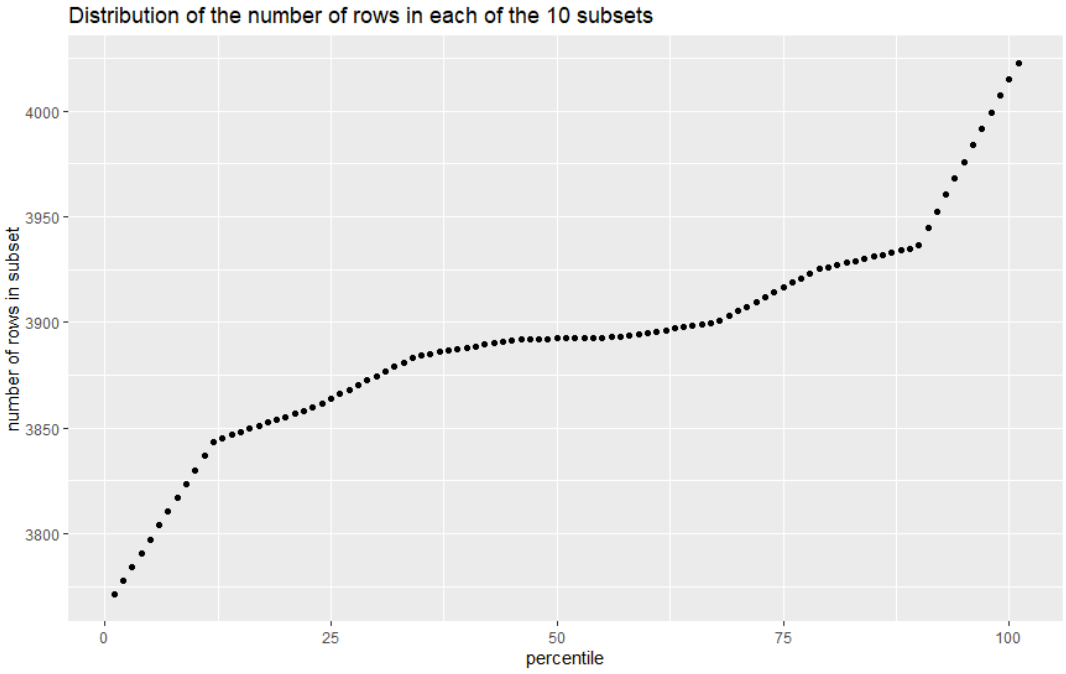

Figure 4: Distribution of the number of rows for the ten subsets

From Figure 4, we observed that there are not exactly 3,893 observations per subset, but this number is on average. The random replicate algorithm simply randomly assigns each row of the input data into the number of bins K determined by the total number of rows n in the data divided by the desired number of rows per subset. Thus, the distribution of the number of rows in each subset is like a draw from a multinomial with number of trials n and event probabilities of being put into one of K bins as pi = 1/K, where i = 1, …, K. We are working on a scalable approach for randomly assigning exactly n/K rows to each subset. As expected, we observed that the number of observations that were inside the ten subsets (key-value pairs) were different and that they ranged from 3,771 to 4,023.

After using random replicate division for creating the ten random subsets we wanted, we fitted a logistic regression model to each subset. The ten coefficients results produced (one for each subset), were used for calculating the weighted average of each coefficient, where the weights refer to the number of observations in each subset (i.e. mean coefficient recombination). That way we have obtained an approximation of the coefficient estimates, but we did not get inference about them.

To do that, we also used the bag of little bootstraps (BLB) transformation method to fit the logistic regression model to the ten subsets created, a procedure that incorporates features of both the bootstrap and subsampling for assessing the quality of estimators. The idea behind this procedure is that the data are split into random subsets and then a bootstrap method is applied to each subset, a bootstrap metric to the result is computed and then the metric across all subsets are averaged.

In our case, we computed the coefficients of the logistic regression model for each bootstrap sample, which were used for creating confidence intervals for the coefficients of each variable in the dataset (Note: We used 100 bootstrap samples, which means that the resampling with a sample of 38,928 rows – as the number of the total observations in the whole dataset- was done 100 times). That way, we were able to create a 95% and a 90% confidence interval for each coefficient in the data.

As previously mentioned, we also wanted to split the observations into twenty random subsets and in order to do that we used about 1,946 rows per subset. In order to grasp a better understanding of the subsets created, we presented below a plot that depicts the distribution of the number of rows in each of the twenty subsets.

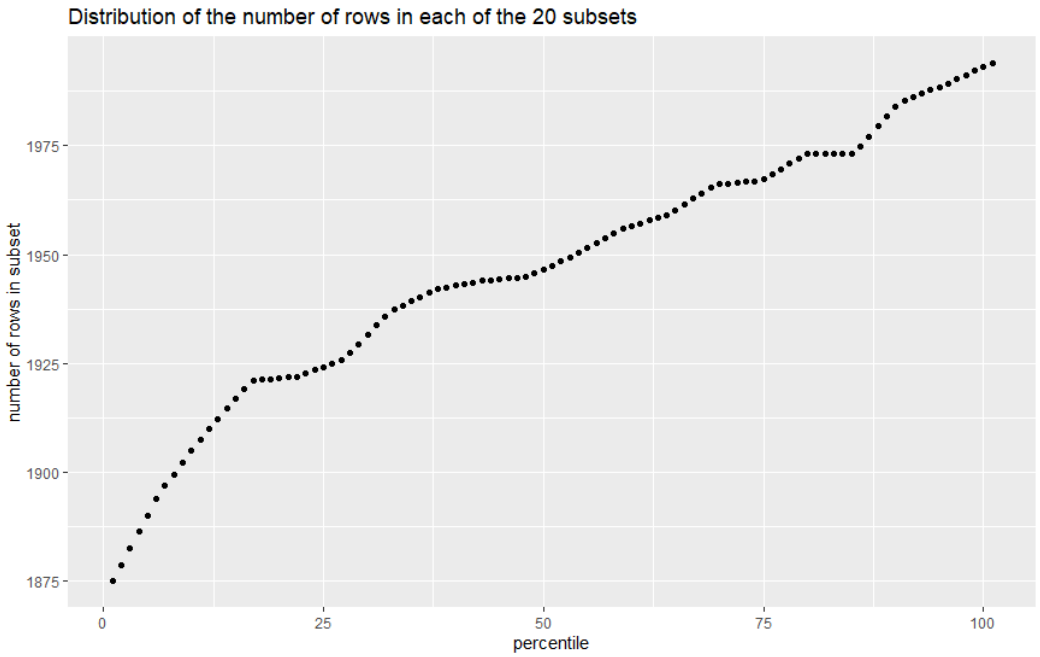

Figure 5: Distribution of the number of rows for the twenty subsets

From Figure 5, we observed that there are not exactly 1,946 observations per subset, but this number is on average. These twenty subsets (key-value pairs) contained observations whose number ranged from 1,875 to 1,985 rows.

Similarly with the implementation of D & R with ten subsets, we fitted a logistic regression model to each one of the twenty (random) subsets created and calculated the weighted average of each of the twenty coefficients produced.

For creating inferences for the coefficients of each variable in the dataset, we used the bag of little bootstraps (BLB) transformation method to fit the logistic regression model to the twenty subsets created. That way, we computed the coefficients of the logistic regression model for each bootstrap sample, thus we were able to create a 95% confidence interval for each coefficient in the data.

5. Evaluation of the estimation approaches

After computing the coefficient estimates along with their corresponding confidence intervals using the D & R approach, it is time to evaluate the results produced by comparing them with the ones created by running the logistic regression model using all the data as is.

In order to evaluate the different estimation approaches we compared the coefficient estimates along with the confidence limits and confidence interval widths produced by each approach. We will first start with the comparison of the coefficient estimations.

5.1. Coefficient estimations comparison

We will begin the evaluation of the implemented approaches by comparing the absolute change of the coefficient estimates produced with those crated when using the whole dataset as is. In Tables 3 and 4 we presented the highest (absolute) changes in the coefficient estimates from the ones produced by the whole data when using ten and twenty splits.

Table 3: Highest absolute change in the coefficient estimates between the models created using the whole data and D & R with ten splits (in descending order of absolute change)

Coefficients	In whole dataset	D & R (with 10 splits)	Absolute Change	Std. Error (whole dataset)
Education (illiterate)	.92	-4.7	5.64	.62
Marital (unknown)	.38	-.68	1.06	.36
(Intercept)	129.2	129.9	.75	24.2
Poutcome (nonexistent)	.47	.53	.06	.11

Table 4: Highest absolute change in the coefficient estimates between the models created using the whole data and D & R with twenty splits (in descending order of absolute change)

Coefficients	In whole dataset	D & R (with 20 splits)	Absolute Change	Std. Error (whole dataset)
Number of employees	-.033	24.5	24.5	.003
Marital (unknown)	.382	-5.48	5.86	.355
Education (illiterate)	.921	-3.42	4.32	.626
(Intercept)	129.2	133.2	3.96	24.2
Poutcome (success)	.604	-.331	.943	.254

The first thing we noticed from the above tables was that when calculating the weighted average for each one of the coefficients produced from the D & R approach with twenty splits, was that there was one coefficient that was not estimated correctly, which is why it changed significantly (more than 24 units) when its standard error was really small (.0026).

Specifically, we observed that the estimation of the coefficient that referred to the number of employees ('nr.employed') had a really high value (it was equal to 24.5 while the estimation using D & R with ten splits or with the whole data as is was equal to -0.27 with a standard error of 0.003 for the latter case). That means that we cannot estimate that coefficient and that its existence in the model does not affect the rest of the coefficients.

We believe that the reason was for this significant change in the value of this coefficient was due to the existence of multicollinearity between our variables. Multicollinearity occurs when independent variables in a regression model are correlated. In general, if two variables are collinear then they carry the same information, thus we should not add them both in our model.

This problem reduces the precision of the estimated coefficients because they become very sensitive to small changes in the model as well as the confidence intervals produced will be wider and may not be dependable.

However, as previously mentioned we did not face such problems in the estimation of the number of employees when using all the data or D & R with ten splits. As a result, we may assume that the problem lies in the way the data split for computing that coefficient, thus there is data multicollinearity. So, if we did variable selection, we would expect to remove this variable from the model when using twenty splits, in contrast with what we would done when using ten splits or the whole data.

From the 32 total coefficients we also observed that when using the D & R approach with ten splits, 30 of them (about 94%) were lower than the standard error. From these thirty instances, we saw that the intercept changed about 0.75 units, but with a high standard error equal to 24 and all the other coefficient estimates had less than 0.06 absolute change. The coefficients with the highest absolute change referred to the illiterate level of the education variable (about 5.6 units change) and to the unknown level of the marital variable (about one unit change).

In contrast, when using twenty subsets, only 30% of the total coefficients were lower than the standard error. The same coefficients as when using ten subsets showed the highest absolute change compared to generated by fitting the logistic regression model to the whole dataset, where the illiterate level of the education variable changed about 5.8 units (instead of 5.6) and the unknown level of the marital variable changed about 4.3 units (instead of one). As previously mentioned, the coefficient that referred to the number of employees had the highest absolute change overall (about 24.5 units). As for the rest of the coefficients, we may observed that 27 of them had only less than 0.5 absolute change, however each one of them had a lot higher value compared to those produced when we used ten subsets.

Therefore, we understand that the approximations of the coefficients of logistic regression created by the D & R procedure with ten splits were accurate enough, while those produced when using twenty splits were not so good.

In order to have a better understanding of the changes in the coefficients, we also reviewed the relative changes, as shown in Tables 5 and 6.

Table 5: Top-5 coefficients with the highest relative change between the models created using the whole data and D & R with ten splits (in descending order of relative change)

Coefficients	Relative Change
Education (illiterate)	616%
Marital (unknown)	279%
Previous	93%
Season (spring)	41%
Education (professional course)	25%

Table 6: Top-10 coefficients with the highest relative change between the models created using the whole data and D & R with twenty splits (in descending order of relative change)

Coefficients	Relative Change
Number of employees	89306.4%
Season (spring)	1747.7%
Marital (unknown)	1536.5%
Campaign	1043.7%
Previous	995.7%
Loan (yes)	614.7%
Consumer Confidence Index	528.9%
Education (illiterate)	472.2%
Consumer Price Index	324.9%
Marital (married)	172.4%

It is clear from the above tables that the relative changes were by far higher when using twenty subsets compared to when we used D & R with ten splits, which is why we decided to further investigate the resulted relative changes in the coefficient estimates. In order to have a better understanding of the distribution of the relative change percnetages we created the following barplots.

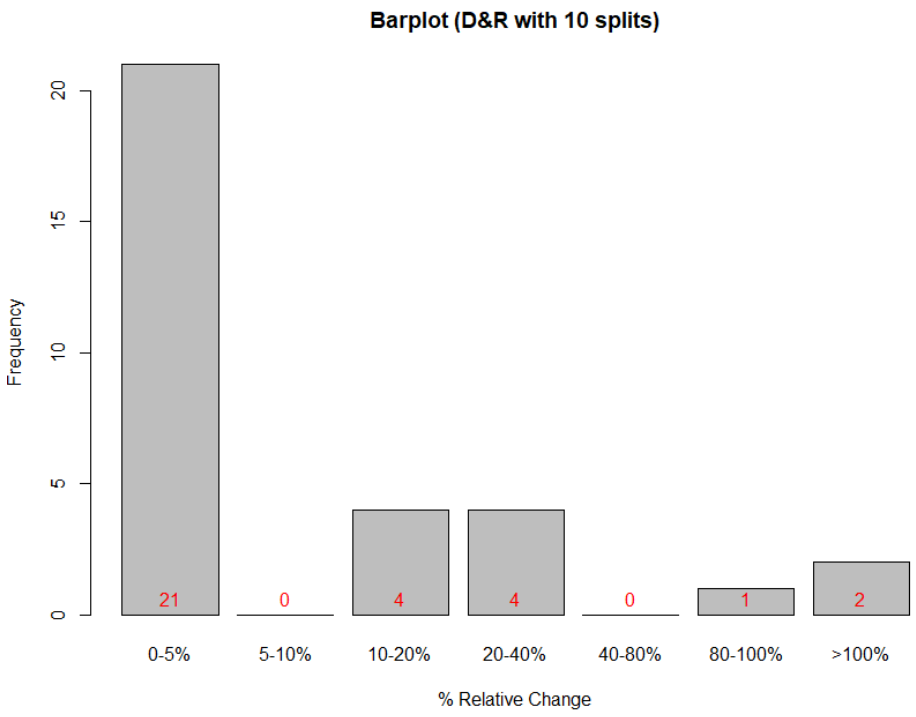

Figure 6: Bar-plot for the distribution of the relative change percentages using D & R with ten splits

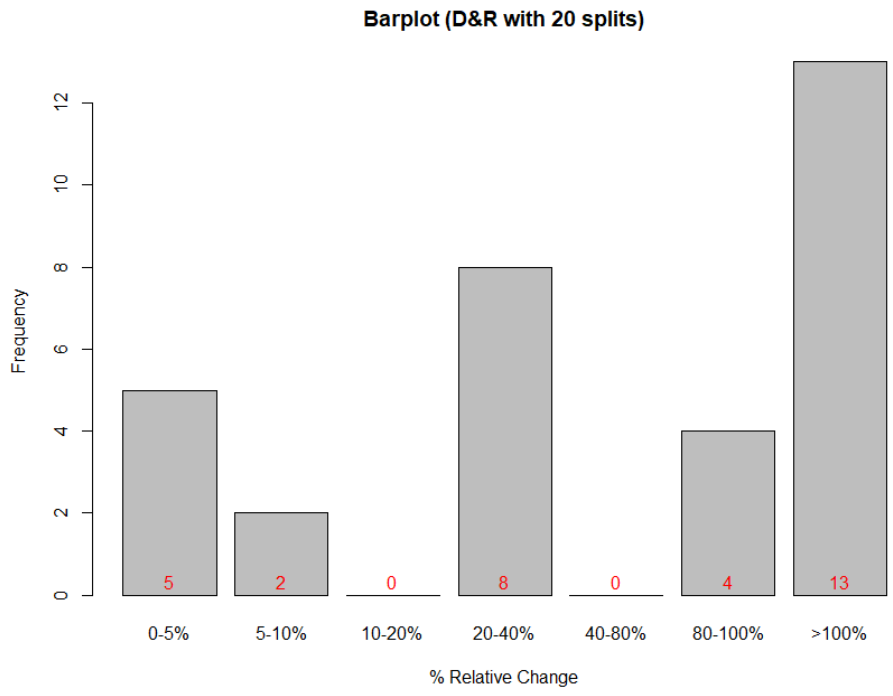

Figure 7: Bar-plot for the distribution of the relative change percentages using D & R with twenty splits

From Figures 6 & 7 along with the exact percentages of the relative changes for each coefficient, our suspicions about the significant higher relative changes occurred when using D & R with twenty splits instead of ten were confirmed.

For instance, we observed that when using ten random subsets only two coefficients estimates had a very high relative change (the illiterate level of the education variable and the unknown level of the marital variable as shown from Table 5), while 21 coefficients estimates had less than 5% relative change.

In contrast, we noticed that when using twenty subsets there were 13 coefficients estimates had a very high relative change (i.e. more than 100%) instead of just two with the D & R with ten splits and only five had less than 5% relative change compared to the 21 when using ten subsets. Note: the samples were randomly chosen so these results may change.

Finally, we computed some metrics in order to compare the difference between the coefficients produced by the logistic regression model using all the data as is and the ones created using D & R with ten and twenty splits. We observed that both the RMSE and MAE were significant higher when using twenty subsets (circa 350% and 460% correspondingly), as shown in Table 7 below.

Table 7: RMSE and MAE values for D & R with ten and twenty splits

	With 10 subsets	**With 20 subsets**
RMSE	1.023	4.583
MAE	.242	1.356

By taking all the above into account, it is clear that by using the divide and recombine approach with ten splits the estimations of the coefficients could be considered as good approximations as most of them would be very close those created by using the whole dataset. On the other hand, when the number of splits increased to twenty, not only the approximations for most of the coefficients differed a lot from their true value, but it also created some other problems (e.g. multicollinearity) as the estimation for some of the coefficients was not anymore possible.

5.2. Confidence Interval and width comparison

Now, we will continue our evaluation by comparing the confidence limits and confidence interval widths produced by each approach for each coefficient of the model. Their evaluation was mainly based on the relative change percentage from the confidence interval created using the whole data as is for fitting the logistic regression model. We started by comparing the 95% confidence intervals created using the D & R approach with ten splits with the one produced after using the whole dataset.

We also computed the 90% confidence intervals when using ten subsets, where the results we observed some slight changes. For example, after using a 95% confidence interval we observed that there were 16 coefficients with relative change for both of their confidence limits less than 6% instead of 12 in the case of the 95% interval. We also saw that there were only five coefficients with relative change for at least one of their confidence limits from 10% to 20% instead of 11 for the 95% confidence interval. Furthermore, the confidence widths with relative change less than 10% corresponded to 28 different coefficients in contrast with the 24 of the 95% confidence interval.

Therefore, we understand that the changes in the coefficients' confidence interval were a lot smaller, thus making them more stable. This behavior was only natural to happen due to the wider confidence limits used. However, the wider the confidence intervals the greater the uncertainty, which in relation to the estimate itself is an indication of instability. This is why we decided to continue our analysis using the 95% confidence interval in order to be as accurate as possible.

For understanding the changes occured in the confidence intervals of the coefficients, we reviewed the relative changes for each one of their confidence limits, as shown in Tables 8 & 9 . We also presented in Tables 10 & 11 the exact confidence intervals for some of the coefficients whose confidence limits presented the highest relative changes, for the cases of the D & R approach with ten and twenty splits. To get a better understanding of the resulted relative changes in the coefficients' confidence limits by each approach, we visualized them by creating bar-plots, as shown in Figures 8 & 9.

Table 8: Highest relative change for one confidence limit of the coefficients' 95% confidence intervals between the models created using the whole data and D & R with ten splits (in descending order of 2.5% percentiles' relative change)

	2.5%	97.5%	Confidence width
Education (illiterate)	1061.6%	294.1%	48.1%
Marital (unknown)	311.5%	112.2%	3.6%
Season (winter)	250.1%	7.2%	3.1%
Marital (single)	172.3%	2.7%	5.8%

Table 9: Highest relative change for one confidence limit of the coefficients' 95% confidence intervals between the models created using the whole data and D & R with twenty splits (in descending order of 2.5% percentiles' relative change)

	2.5%	97.5%	Confidence width
Employees number	21618.8%	190,223.5%	354000%
Season (winter)	5670.8%	67.4%	23.9%
Marital (unknown)	1546.1%	576.5%	32.7%
Poutcome (success)	799.8%	101.6%	29%
Education (illiterate)	717.6%	251.3%	75.5%
Campaign	446.8%	2555.1%	935.1%
Consumer Confidence Index	279.6%	1109.9%	451.4%

Table 10: confidence intervals between the models created using the whole data and D & R with ten splits (in descending order of 2.5% percentiles' relative change)

	In whole dataset			D & R (with 10 splits)		
	2.5%	97.5%	Confidence width	2.5%	97.5%	Confidence width
Education (illiterate)	-.455	2.05	2.51	-5.28	-3.98	1.30
Marital (unknown)	-.359	1.04	1.40	-1.48	-.128	1.35
Season (winter)	-.011	.675	.686	.016	.724	.707
Marital (single)	-.006	.272	.277	.004	.265	.261

Table 11: 95% confidence intervals between the models created using the whole data and D & R with twenty splits (in descending order of 2.5% percentiles' relative change)

	In whole dataset			D & R (with 20 splits)		
	2.5%	97.5%	Confidence width	2.5%	97.5%	Confidence width
Employees number	-.033	-.022	.010	7.02	42.4	35.4
Season (winter)	-.011	.675	.686	.608	1.13	.522
Marital (unknown)	-.359	1.04	1.40	-5.91	-4.96	.942
Poutcome (success)	-.455	2.05	2.51	-3.72	-3.10	.614
Education (illiterate)	.104	1.10	.999	-.727	-.018	.709
Campaign	-.061	-.024	.037	.212	.594	.383
Consumer Confidence Index	-.069	-.032	.037	.124	.328	.204

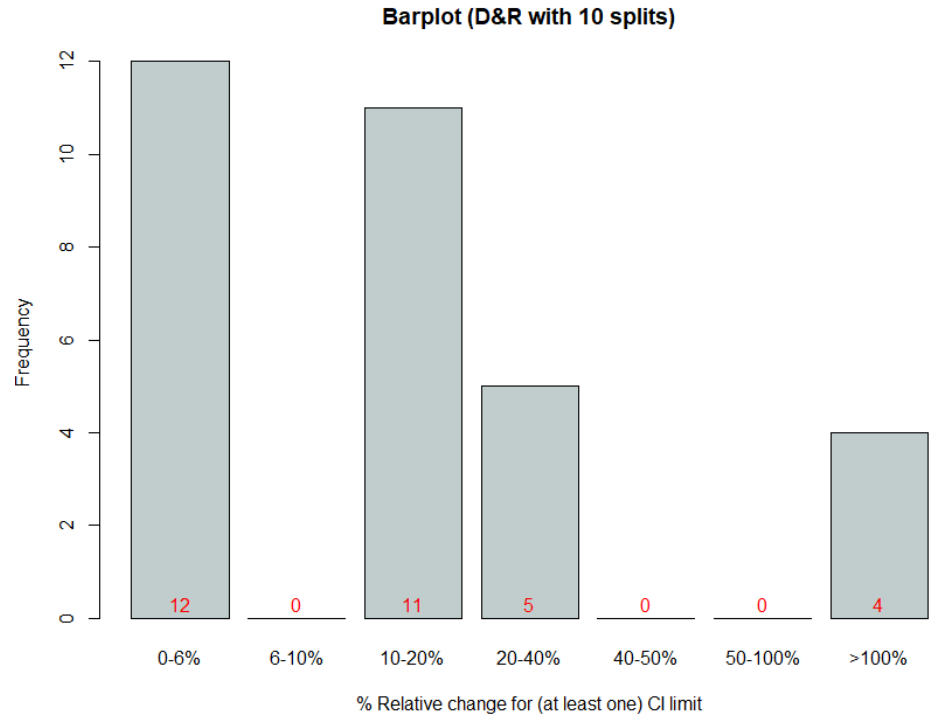

Figure 8: Bar-plot for the distribution of the relative change percentages for the confidence limits using D & R with ten splits

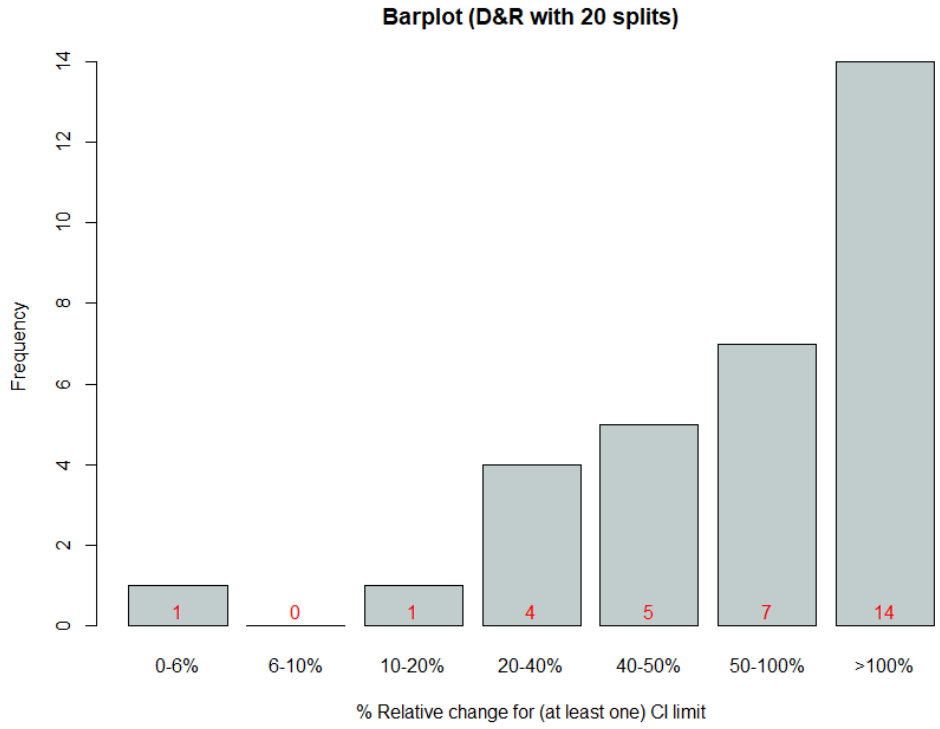

Figure 9: Bar-plot for the distribution of the relative change percentages for the confidence limits using D & R with ten splits

Just by looking at the bar-plots, we observed that the confidence limits were a lot different when using twenty subsets compared to when we used D & R with ten splits. For instance, we saw that in the case of ten subsets for 12 coefficients the relative change for both of their confidence limits was less than 6%, compared to only one in the case of twenty subsets. In addition, while there were only four coefficients with relative change of at least one of their confidence limits more than 100% when implementing D & R with ten splits, this number skyrocketed to 14 when we used twenty splits.

In order to be as thorough as possible we examined in detail the exact percentages of the relative changes for each coefficient confidence limits, from where the difference in the confidence intervals was more than clear.

So, we understood that while for most of the coefficients the confidence limits were shilghtly changed when we used ten subsets, the results produced from the D & R with twenty subsets by far different. By further investigating the results produced by the D & R with ten splits, we observed from Table 8 that the illiterate level from the education variable, the winter level from the season variable and the unknown level of the marital variable presented the highest relative change in their confidence intervals (and things only get worse when we used twenty splits).

However, the previously mentioned levels were the ones with the lowest number of observations overall. Specifically, the illiterate level from the education variable corresponded to 0.04% of the data, the winter level from the season variable to 0.45% of the data and the unknown level from the marital variable to 0.2% of the data. So, we may claim that the coefficient estimates created by the divide and recombine method, faced problems in the correct calculation of coefficients who corresponded to a very small portion of the dataset and this problem only gets worse as the number of splits used increases.

Furthermore, we saw the relative change occurred to the confidence interval widths created using the whole dataset and the ones created using the divide and recombine procedure with ten and with twenty splits. Specifically, we observed that when using ten random subsets only the illiterate level of the education variable had a significant relative change (about 48%) while there were seven coefficients whose confidence widths relative changed were between 10% to 20%. As for the rest 24 coefficients' confidence widths the relative change was less than 10%.

On the other hand, the relative change of the confidence widths when using twenty subsets was significant higher, as expected from the higher confidence limits we previously observed. More specifically, we saw that there were four coefficients who varied significantly, where for three of them the relative change in the confidence width was at least four time bigger and about 75% for the other one. Also, there were 17 confidence widths whose relative change ranged from 10% to 40%, from which only 4 were between 10% and 20% and for the other 11 confidence widths the relative change was less than 10%. In Tables 12 & 13 below, we present the coefficients with the four biggest relative changes in their confidence interval widths for the cases of D & R with ten and twenty subsets.

Table 12: Top-4 relative changes of confidence interval widths between the models created using the whole data and D & R with ten splits (in descending order)

	Confidence width
Education (illiterate)	48.1%
Number of employees	20.0%
Consumer Confidence Index	18.9%
(Intercept)	16.4%

Table 13: Top-4 relative changes of confidence interval widths between the models created using the whole data and D & R with twenty splits (in descending order)

	Confidence width
Number of employees	334000%
Campaign	935.1%
Consumer Confidence Index	451.4%
Education (illiterate)	75.5%

Finally, we computed the RMSE and MAE metrics for the limits of the 95% confidence interval and for the confidence width created with each method. We observed that when using twenty subsets instead of ten, both the RMSE and MAE were higher for the 2.5% percentile[10] (increased by 5% and 60% correspondingly). They were also significantly higher for the 97.5% percentile[11] (RMSE increased by 364% and the MSE increased by 358%) and as expected for the confidence width (RMSE increased by 142% and the MSE increased by 207%). In Table 14 below, we presented the exact values of the RMSE and MAE, from which we can calculate the percentage change for the aforementioned metrics.

Table 14: RMSE and MAE between whole dataset and D & R with 10 splits and with 20 splits (for 95% confidence interval)

	2.5%		97.5%		Confidence width	
	With 10 subsets	With 20 subsets	With 10 subsets	With 20 subsets	With 10 subsets	With 20 subsets
RMSE	1.70	1.79	1.689	7.84	2.76	6.69
MAE	.456	.734	.470	2.15	.542	1.67

[10] 2.5 percent or more is above the referenced value and 97.5 percent is below the referenced value
[11] 97.5 percent or more is above the referenced value and 2.5 percent is below the referenced value

6. Conclusions

When the number of splits in the D & R method increased to twenty, there was a data multicollinearity problem as the estimation for some of the coefficients was not anymore possible, while with the usage of fewer splits we did not face such problems.

The approximations of the coefficient estimates created by the D & R procedure with ten splits outperformed the ones produced when using twenty spits due to the fact that most of them were very close to the ones created after fitting the logistic regression model to the whole dataset as is (in terms of absolute and relative change). For example, in the case of twenty random subsets, we saw that 13 of the coefficient estimates varied significantly (i.e. more than 100% relative change) instead of just two when using ten splits.

Based on the above, there was another problem we observed after implementing the divide and recombine method. Specifically, we saw that there were problems in the correct calculation of the coefficient estimates regarding categorical variables whose levels corresponded to a very small portion of the dataset when splitting the data. This could lead us in wrong results and as we noticed with the usage of twenty splits, this problem only gets worse as the number of subsets in the D & R procedure increases.

The supremacy of this D & R method when using ten splits compared to the usage of twenty splits in terms of generating good estimates, was also clear by the confidence limits and confidence interval widths produced. Specifically, we observed that while for most of the coefficients the confidence limits were shilghtly changed when we used ten subsets, the results produced from the D & R with twenty subsets differed a lot.

In order to be as thorough as possible, we also compared the coefficient estimates and confidence intervals produced by each approach using some metrics (RMSE and MAE) which clearly showed that the lower number of splits produced better results.

Another problem we observed when using the D & R approach is that as the number of splits increases, the recursion becomes slower which resulted in more computation time for creating the estimations of the coefficients. Therefore, if many splits of the data are used then there is the chance that the system may crash.

To conclude with, we understand that the implementation of the D & R approach with ten splits results in relatively good estimations. We can also claim that as the number of splits increases, the effectiveness of the method decreases and the computational time increases and vice versa.

Bibliography

Anon, Assumptions of Logistic Regression, *StatisticsSolutions*. Available at: https://www.statisticssolutions.com/assumptions-of-logistic-regression/

Anon, Logistic Regressions Assumptions and Diagnostics in R, *STHDA*. Available at: http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

Bühlmann P., Drineas P., Kane M., van der Laan M. (eds.), (2016). *Handbook of Big Data*. New York: Chapman and Hall/CRC. Available at: https://www.taylorfrancis.com/books/edit/10.1201/b19567/handbook-big-data-peter-b%C3%BChlmann-petros-drineas-michael-kane-mark-van-der-laan

Hafen R., (January 27, 2015). *Divine and Recombine. A distributed Data Analysis Paradigm*. Workshop on Distributing Paradigm in R. Available at: https://www.yumpu.com/en/document/view/50296869/ryan-talk11

Hafen R., (2016). *Analyze and Visualize Large Complex Data in R*. The DeltaRho Project. Available at http://deltarho.org/index.html

Liu Q., Bhadra A., Cleveland W. S., (2018). *Divide and Recombine for Large and Complex Data: Model Likelihood Functions using MCMC*. Available at: https://arxiv.org/abs/1801.05007

APPENDIX

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')