# **Statistics for Big Data**

# **Assignment 2 – Report**

## **Introduction**

We are given a data set related to telemarketing phone calls to sell long-term deposits. Agents make phone calls to a list of clients, or clients call to the call center. In any case, clients are asked to subscribe to the service. We need to create a model that classifies contacts as successful or unsuccessful.

The purpose of this document is to compare two different estimation approaches on the problem mentioned:

1. Run logistic regression on all data.
2. Run logistic regression using “divide and recombine” approach to 10 and 20 splits on data.

## **Data** **Manipulation**

Before running the models, we need to manipulate the data.

**Step 1: Choose columns to use in analysis**

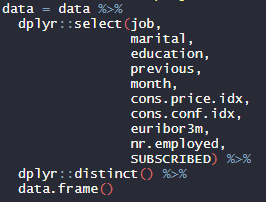


Figure 1: Code for steps 1,2

The independent variables used in analysis are “job”, “marital”, “education”, “previous”, “month”, “cons.price.idx”, “cons.conf.idx”, “euribor3m” and “nr.employed”.

The dependent variable is “SUBSCRIBED”, with values “yes” for successful contact and “no” for unsuccessful.

**Step 2: Remove duplicates**

Dataset contains duplicates that we need to remove in order to proceed.

**Step 3: Transform categorical variables to numeric**

In this step we create dummy columns (columns with values 0/1) that represent the different levels/values of the categorical variables.

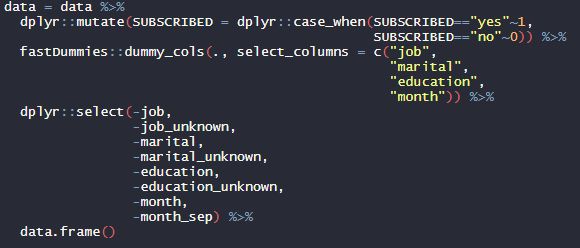


Figure 2: Code for step 3

The dependent variable will have the value 1 for “Yes” and 0 for “No”.

We use library fastDummies for columns "job", "marital", "education" and "month" and we do not forget to remove one level for each variable as well as the original variable. Before dataset had 10 columns, now it has 36.

**Step 4: Create train test splits**

We use an 80 – 20 split schema to create the training and test dataset. There is class imbalance in the data, since 26% of the values of dependent variable are 1 and 74% are 0. These percentages are present in train and test datasets too.

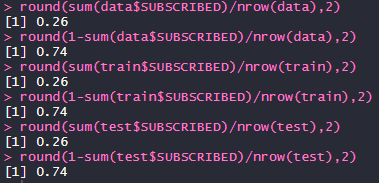


Figure 3: Class imbalance check

## **Modeling**

1. **Logistic Regression on all data**

In this part we fit a glm model to the training set as a whole and get the following coefficients.

|  |  |  |  |
| --- | --- | --- | --- |
| variable | coefficient | variable | coefficient |
| (Intercept) | 97.29 | **marital\_married** | 0.70 |
| previous | 0.02 | **marital\_single** | 0.61 |
| cons.price.idx | 0.04 | **education\_basic.4y** | 0.32 |
| cons.conf.idx | - 0.04 | **education\_basic.6y** | - 0.01 |
| euribor3m | 0.55 | **education\_basic.9y** | 0.17 |
| nr.employed | - 0.02 | **education\_high.school** | 0.39 |
| job\_admin. | 0.75 | **education\_illiterate** | 0.63 |
| job\_blue.collar | 0.79 | **education\_professional.course** | 0.22 |
| job\_entrepreneur | 0.24 | **education\_university.degree** | 0.49 |
| job\_housemaid | 0.05 | **month\_apr** | - 0.09 |
| job\_management | 0.45 | **month\_aug** | 1.00 |
| job\_retired | 0.60 | **month\_dec** | 0.30 |
| job\_self.employed | 0.16 | **month\_jul** | 0.65 |
| job\_services | 0.49 | **month\_jun** | 0.10 |
| job\_student | 0.84 | **month\_mar** | 0.31 |
| job\_technician | 0.78 | **month\_may** | - 0.23 |
| job\_unemployed | 0.11 | **month\_nov** | 0.12 |
| marital\_divorced | 0.28 | **month\_oct** | 0.39 |

We now use the model to predict on the test dataset. The metrics of Figure 4 show the performance of the model. The accuracy is 75%, but as we can see in the confusion matrix, model does not classify class 1 correctly and therefore AUC and F1 are very low.

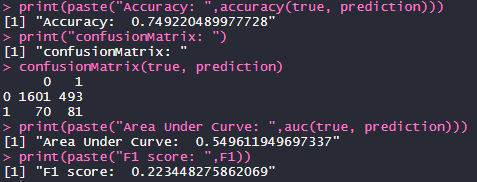


Figure 4: Metrics

1. **Logistic Regression on 10 splits of data**

In this part we try to recreate the results of part 1 by splitting the training data in 10 equal parts, fitting a model in each part and then averaging the coefficients. Then we change the coefficients of the model with the ones we calculated and use the updated model to get predictions. We get the following coefficients:

|  |  |  |  |
| --- | --- | --- | --- |
| variable | coefficient | variable | coefficient |
| (Intercept) | 97.33 | **marital\_married** | 0.70 |
| previous | 0.02 | **marital\_single** | 0.61 |
| cons.price.idx | 0.04 | **education\_basic.4y** | 0.32 |
| cons.conf.idx | - 0.04 | **education\_basic.6y** | - 0.01 |
| euribor3m | 0.55 | **education\_basic.9y** | 0.17 |
| nr.employed | - 0.02 | **education\_high.school** | 0.39 |
| job\_admin. | 0.75 | **education\_illiterate** | 0.62 |
| job\_blue.collar | 0.79 | **education\_professional.course** | 0.22 |
| job\_entrepreneur | 0.24 | **education\_university.degree** | 0.49 |
| job\_housemaid | 0.06 | **month\_apr** | - 0.09 |
| job\_management | 0.45 | **month\_aug** | 1.00 |
| job\_retired | 0.60 | **month\_dec** | 0.30 |
| job\_self.employed | 0.16 | **month\_jul** | 0.65 |
| job\_services | 0.49 | **month\_jun** | 0.10 |
| job\_student | 0.85 | **month\_mar** | 0.31 |
| job\_technician | 0.78 | **month\_may** | - 0.23 |
| job\_unemployed | 0.11 | **month\_nov** | 0.12 |
| marital\_divorced | 0.28 | **month\_oct** | 0.40 |

As we can see the coefficients do not change that much. Intercept differs 0.04, a few coefficients differ 0.01 and all other coefficients do not have differences in the first 2 decimal digits. Metrics are the same as before:

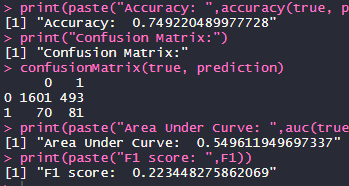


Figure 5: Metrics

1. **Logistic Regression on 20 splits of data**

In this part we try to recreate the results of part 1 by splitting the training data in 20 equal parts, fitting a model in each part and then averaging the coefficients. After, we change the coefficients of the model with the ones we calculated and use the updated model to get predictions. We get the following coefficients:

|  |  |  |  |
| --- | --- | --- | --- |
| variable | coefficient | variable | coefficient |
| (Intercept) | 97.31 | **marital\_married** | 0.70 |
| previous | 0.02 | **marital\_single** | 0.62 |
| cons.price.idx | 0.04 | **education\_basic.4y** | 0.32 |
| cons.conf.idx | -0.04 | **education\_basic.6y** | -0.01 |
| euribor3m | 0.55 | **education\_basic.9y** | 0.17 |
| nr.employed | -0.02 | **education\_high.school** | 0.39 |
| job\_admin. | 0.75 | **education\_illiterate** | 0.62 |
| job\_blue.collar | 0.79 | **education\_professional.course** | 0.22 |
| job\_entrepreneur | 0.24 | **education\_university.degree** | 0.49 |
| job\_housemaid | 0.06 | **month\_apr** | -0.09 |
| job\_management | 0.45 | **month\_aug** | 1.00 |
| job\_retired | 0.60 | **month\_dec** | 0.30 |
| job\_self.employed | 0.16 | **month\_jul** | 0.65 |
| job\_services | 0.49 | **month\_jun** | 0.10 |
| job\_student | 0.84 | **month\_mar** | 0.31 |
| job\_technician | 0.78 | **month\_may** | -0.23 |
| job\_unemployed | 0.11 | **month\_nov** | 0.12 |
| marital\_divorced | 0.28 | **month\_oct** | 0.39 |

As we can see the coefficients do not change that much. Intercept differs 0.02, a few coefficients differ 0.01 and all other coefficients do not have differences in the first 2 decimal digits. Metrics are the same as before:

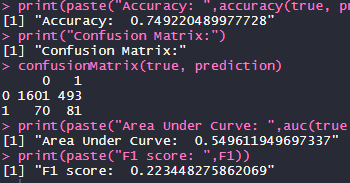


Figure 6: Metrics

## **Conclusion**

To conclude, all three models – whether we used the whole dataset, 10 splits or 20 splits in training phase – where about the same. Therefore, this technic can be considered in cases where data do not fit in memory.