Predicting Customer Response to Telemarketing Campaigns

**Table of Contents**

1. [Project Background](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#1)
2. [Data Cleaning](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#2)
3. [Exploratory Data Analysis](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#3)
4. [Data Visualization](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#4)
5. [Machine Learning: Classification](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#5)
6. [Machine Learning: Regression](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#6)
7. [Conclusion & Recommendations](https://kkb-production.jupyter-proxy.kaggle.net/k/53966257/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2IiwidHlwIjoiSldUIn0..eGHgcFvbNY5xFPa-FnhR7A.5-_oVN3gyXdbMYlFa3fSYM9P30MRZYo_wwd9oRx-LK93OB2GtTfWDxFSoolwoOntRfRiqSj5EPZdcz9m_hdQGfb8J8KljEezqSmMgkxXsdpVFJAOhA86yqjGyZj-Ab-KTByoREuJ-gMJw9B__6G-nzFDrVVMMVGcRlIpSnloRrTHS2DPeb2d_32fAKceOvDlXdCyxUrEfofu1oiOYb7Nkg.DjPeP_BH_GpNlafm2nnhXQ/proxy/notebooks/__notebook_source__.ipynb#7)

Part 1. Project Background

Nowadays, marketing spending in the banking industry is massive, meaning that it is essential for banks to optimize marketing strategies and improve effectiveness. Understanding customers’ need leads to more effective marketing plans, smarter product designs and greater customer satisfaction.

### Main Objective: increase the effectiveness of the bank's telemarketing campaign

This project will enable the bank to develop a more granular understanding of its customer base, predict customers' response to its telemarketing campaign and establish a target customer profile for future marketing plans.

By analyzing customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviours and identify which type of customers is more likely to make term deposits. The bank can then focus its marketing efforts on those customers. This will not only allow the bank to secure deposits more effectively but also increase customer satisfaction by reducing undesirable advertisements for certain customers.

# Part 2. Data Cleaningclean

## Load the raw data

This dataset is about the direct phone call marketing campaigns, which aim to promote term deposits among existing customers, by a Portuguese banking institution from May 2008 to November 2010. It is publicly available in the UCI Machine learning Repository, which can be retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

There are 41,188 observations in this dataset. Each represents an existing customer that the bank reached via phone calls.

\* For each observation, the dataset records \*\*16 input variables\*\* that stand for both qualitative and quantitative attributes of the customer, such as age, job, housing and personal loan status, account balance, and the number of contacts.

\* There is \*\*a single binary output variable\*\* that denotes “yes” or “no” revealing the outcomes of the phone calls.

**Clean the dataset:**

### 2.1 Deal with missing data

There is no missing value in this dataset. Nevertheless, there are values like “unknown”, “others”, which are helpless just like missing values. Thus, these ambiguous values are removed from the dataset.

### 2.2 Drop outliers in the column 'balance'

In order to capture the general trend in the dataset, outliers in the column “balance” are dropped. Outliers are defined as the values which are more than three standard deviations away from the mean. In sum, 2556 rows of data were removed.

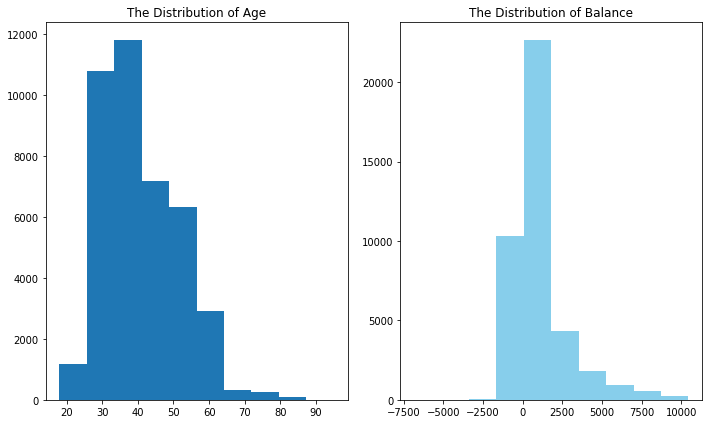
### 2.3 Creating and transforming data

Some changes were made to the column name, units and data types for easier analysis.

# Part 3. Exploratory Data Analysis

To obtain a better understanding of the dataset, the distribution of key variables and the relationships among them were plotted.

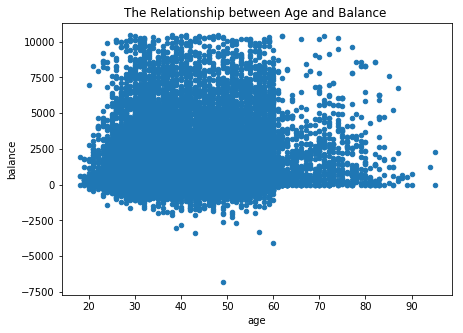
**3.1 Visualize the distribution of 'age' and 'balance'**

****

**The distribution of age**: In its telemarketing campaigns, clients called by the bank have an extensive age range, from 18 to 95 years old. However, a majority of customers called is in the age of 30s and 40s (33 to 48 years old fall within the 25th to 75th percentiles). The distribution of customer age is fairly normal with a small standard deviation.

**The distribution of balance**: After dropping outliers in balance, the range of balance is still massive, from a minimum of -6847 to a maximum of 10443 euros, giving a range of 17290 euros. The distribution of balance has a huge standard deviation relative to the mean, suggesting large variabilities in customers' balance levels.

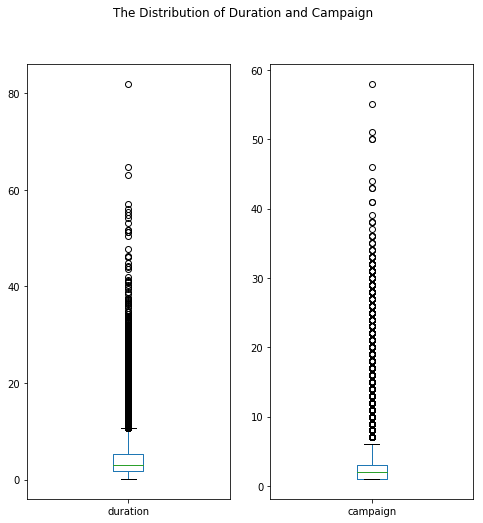
### 3.2 Visualize the relationship between 'age' and 'balance'

****

Based on this scatter plot, there is no clear relationship between client’s age and balance level.

Nevertheless, over the age of 60, clients tend to have a significantly lower balance, mostly under 5,000 euros. This is due to the fact that most people retire after 60 and no longer have a reliable income source.

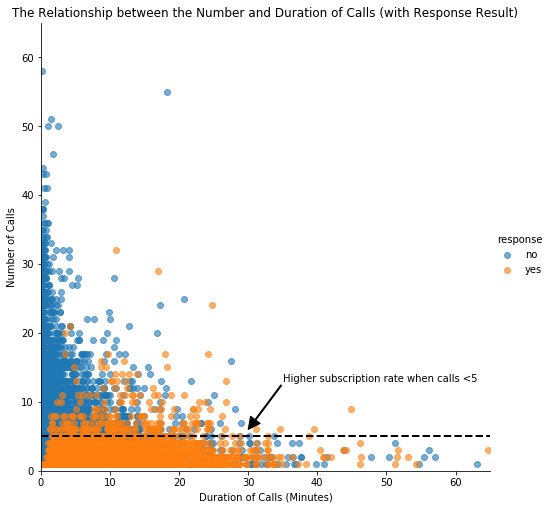
### 3.3 Visualize the distribution of 'duration' & 'campaign'



**The distribution of duration**: As observed from the box plot, the duration of contact has a median of 3 minutes, with an interquartile range of 1.73 minutes to 5.3 minutes. The left-skewed boxplot indicates that most calls are relatively short. Also, there is a large number of outliers ranging from 10 minutes to 40 minutes, which are worth further study.

**The distribution of campagin**: About half of the clients have been contacted by the bank for the second time, while 25% was first introduced to the term deposit. Most clients have been reached by the bank for one to three times, which is reasonable. However, some clients have been contacted by as high as 58 times, which is not normal. These clients may have some special needs that require frequent contact.

### 3.4 Visualize the relationship between 'duration' & 'campaign': with response result

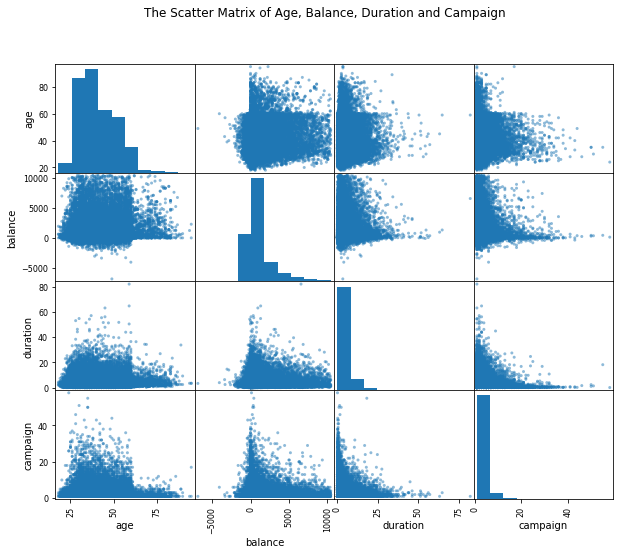


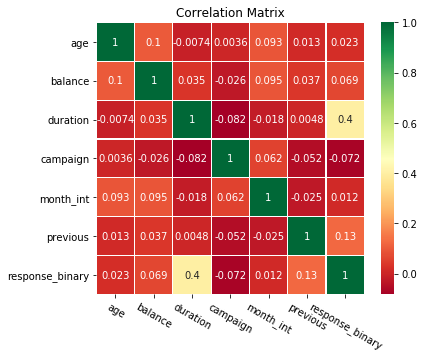
In this scatter plot, clients subscribed to term deposits are denoted as "yes" while those did not are denoted as "no".

As we can see from the plot, “yes” clients and “no” clients are forming two relatively separate clusters. Compared to “no” clients”, “yes” clients were contacted by fewer times and had longer call duration. More importantly, after five campaign calls, clients are more likely to reject the term deposit unless the duration is high. Most “yes” clients were approached by less than 10 times.

This suggests that the bank should resist calling a client for more than five times, which can be disturbing and increase dissatisfaction.

### 3.5 Scatter matrix and Correlation matrix





The scatter matrix does not reveal any clear relationship among age, balance, duration and campaign.

To investigate more about correlation, a correlation matrix was plotted with all qualitative variables. Clearly, “campaign outcome” has a strong correlation with “duration”, a moderate correlation with “previous contacts”, and mild correlations between “balance”, “month of contact” and “number of campaign”. Their influences on campaign outcome will be investigated further in the machine learning part.

The main objective of this project is to identify the most responsive customers before the marketing campaign so that the bank will be able to efficiently reach out to them, saving time and marketing resources. To achieve this objective, classification algorithms will be employed. By analyzing customer statistics, a classification model will be built to classify all clients into two groups: "yes" to term deposits and "no" to term deposits.

## Load the cleaned dataset

## Prepare Data for Classification

### \*Select variables relevant to customers

Only the most relevant customer information is considered, which includes job title, education, age, balance, default record, housing record and loan record. Other information, such as ‘the number of contacts performed before this campaign’, is omitted because it is not directly related to customers themselves.

### Tranform categorical data into dummy variables

Since machine learning algorithms only take numerical values, all five categorical variables (job, education, default, housing and loan) are transformed into dummy variables.

Dummy variables were used instead of continuous integers because these categorical variables are not ordinal. They simply represent different types rather than levels, so dummy variables are ideal to distinguish the effect of different categories.

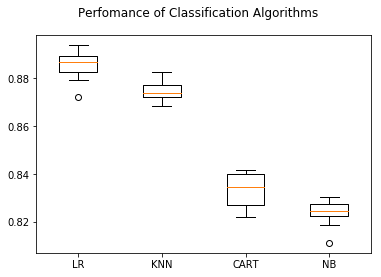
### Feature selection

The values of the first 20 columns, which contain customer statistics, are selected as features while the value of the last column, 'campaign outcome', is set as target.

### Train/ test split

## Compare classification algorithms

Four different classification algorithms (Logistic Regression, K-Neighbors Classifier, Decision Tree Classifier, and Gaussian NB) are run on the dataset and the best-performing one will be used to build the classification model.



**Logistic regression is the best performing model.**

Among all algorithms, logistic regression achieved an accuracy of about 88%, suggesting a high level of strength of this model to classify the customer response given all the defined customer features.

## Test LR model on the test set

## Evalute LR Model

Accuracy score is the percentage of correct predictions out of all predictions made. The LR algorithm achieves an accuracy of 89.08%, suggesting high level of strength of this model to classify the customer response given all the defined customer features.



**However, the result of accuracy score can possibly yield misleading result if the data set is unbalanced, because the number of observations in different classes largely vary.**

A confusion matrix gives a detailed breakdown of prediction result and error types. Each cell in the matrix represents a combination of instances of the predicted response and the actual response. In the test set, the matrix proves that the algorithm performed well because most test results (7277 True Positive predictions) locate on the diagonal cells which represent correct predictions. 891 tests (False negative) predicted the bank’s client would subscribe to the term deposit but they actually did not.

**A problem revealed by this confusion matrix is that the dataset is highly unbalanced, with nearly all client actually decline to subscribe.** This infers that the accuracy score is biased, and further evaluation should be carried out to determine the accuracy of logistic regression model.

Classification report shows the precision, recall, F1 and support scores for the LR classification model.

* Precision of 0 (the client said no) represents that for all instances predicted as no subscription, the percentage of clients that actually said no is 89%.
* Recall is the ability of a classifier to find all positive instances. Recall of 0 indicates that for all clients that actually said no, the model predicts 100% correctly that they would decline the offer.

In general, the report shows that LR model has great predictive power to identify the customers who would not subscribe to the term deposit. However, because of the limited number of clients accepting the term deposit, there is a need for stratified sampling or rebalancing to deal with this structural weakness before we conclude whether LR algorithm can accurately classify those who are more likely to subscribe.

# Part 6. Machine Learning: Regression

Regression analysis is carried out to complement the classification result and help the bank better predict campaign outcome based on customer statistics.

Since the duration of a phone call is positively correlated with the campaign outcome, it can serve as another indicator of the possibility of subscription. In this part, regression algorithms will be used to estimate the duration of a phone call, helping the bank better predict subscription rate.

## Prepare Data for Regression

### 6.1 Feature selection

The values of the first 19 columns, which contain customer statistics, are selected as features while the value of the last column, 'duration', is set as target.

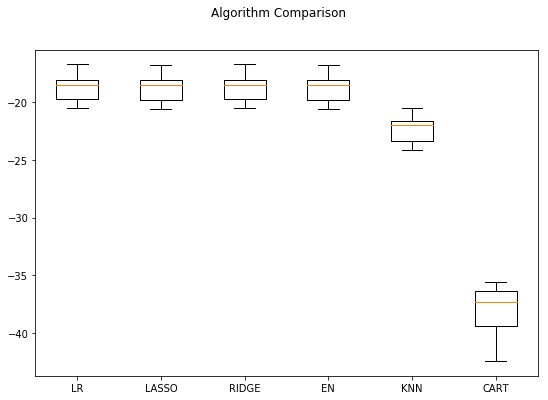
### 6.2 Train/ test split

## Compare regression algorithms

Six different regression algorithms (Linear Regression, Lasso, Ridge, ElasticNet, K Neighbors and Decision Tree) are run on the dataset and the best-performing one will be used to build the estimation model.

**Ridge regression slightly outperforms other models.**

## Stardardize Data



## Test RIDGE model on the test set

## Evaluate RIDGE Model

According to the previous analysis, observations on duration are extremely varied from 0.1 to 81.97 minutes in this dataset. Therefore, a 17.78 MSE testifies that ridge regression is a sound model in predicting the target variable and suggest that the bank can roughly estimate the duration of campaign calls of each client using their customer profiles such as age, job, and loans.

# Part 7. Conclusion & Recommendations

**The main objective of this project is to increase the effectiveness of the bank's telemarketing campaign, which was successfully met through data analysis, visualization and analytical model building. A target customer profile was established while classification and regression models were built to predict customers' response to the term deposit campaign.**

## Conclusion:

According to previous analysis, a target customer profile can be established. The most responsive customers possess these features:

* Feature 1: age < 30 or age > 60
* Feature 2: students or retired people
* Feature 3: a balance of more than 5000 euros

By applying logistic and ridge regression algorithms, classification and estimation model were successfully built. With these two models, the bank will be able to predict a customer's response to its telemarketing campaign before calling this customer. In this way, the bank can allocate more marketing efforts to the clients who are classified as highly likely to accept term deposits, and call less to those who are unlikely to make term deposits.

In addition, predicting duration before calling and adjusting marketing plan benefit both the bank and its clients. On the one hand, it will increase the efficiency of the bank’s telemarketing campaign, saving time and efforts. On the other hand, it prevents some clients from receiving undesirable advertisements, raising customer satisfaction. With the aid of logistic and ridge regression models, the bank can enter a virtuous cycle of effective marketing, more investments and happier customers.

LRLR: 0.885284 (0.005933)

KNN: 0.874938 (0.004055)

CART: 0.833465 (0.006925)

NB: 0.823855 (0.00542LR: 0.885284 (0.005933)

KNN: 0.874938 (0.004055)

CART: 0.885284 (0.005933)

KNN: 0.874938 (0.004055)

CART: 0

833465 (0.00