# **Part 1. Project Background**

Nowadays, marketing expenditures in the banking industry are 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 Objectives: predict customers' responses to future marketing campaigns & 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 behaviors 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. About the Data**

The 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.](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

There are 41,188 observations in the dataset, with no missing values. 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 marketing phone calls. "Yes" means that a customer subscribed to term deposits.

### **Data Cleaning**

1)Several changes were made to the dataset to prepare it for analysis.

2)Drop ambiguous values, such as "others" and "unknown".

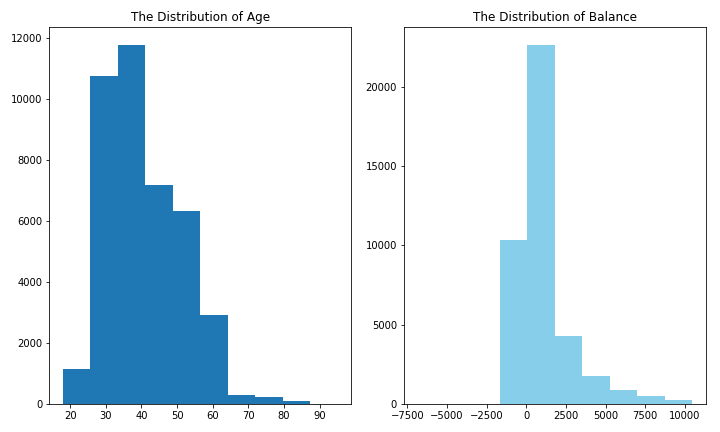
3)Drop outliers to capture the general trend. Outliers are defined as customers whose balance levels are either too high or too low (i.e., more than three standard deviations away from the mean).

4)Change the "response" variable (yes/no) to binary values (1/0) 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 customer age and balance levels:**



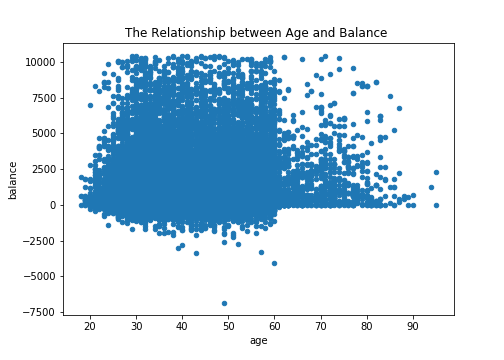
****The distribution of age:****

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:****

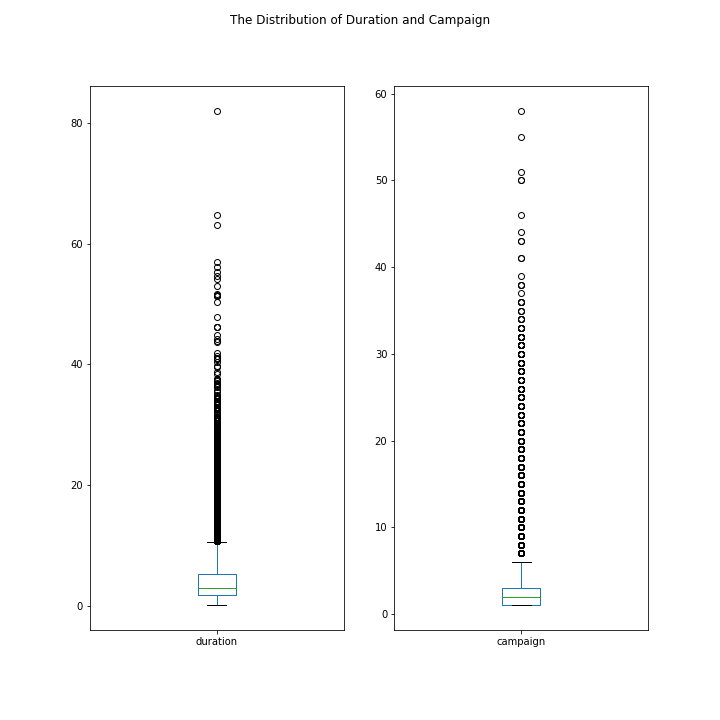
After dropping outliers in balance, the range of balance is still massive, from a minimum of -6847 to a maximum of 10443 euros, giving a range of 17290 euros. The distribution of balance has a huge standard deviation relative to the mean, suggesting large variabilities in customers' balance levels.

### **3.2 Visualize the relationship between customer 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 phone call duration & the number of campaigns**



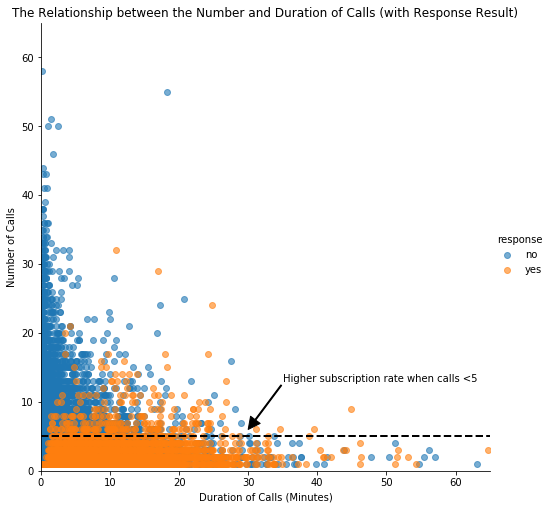
**The distribution of duration**

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 campaign**

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.

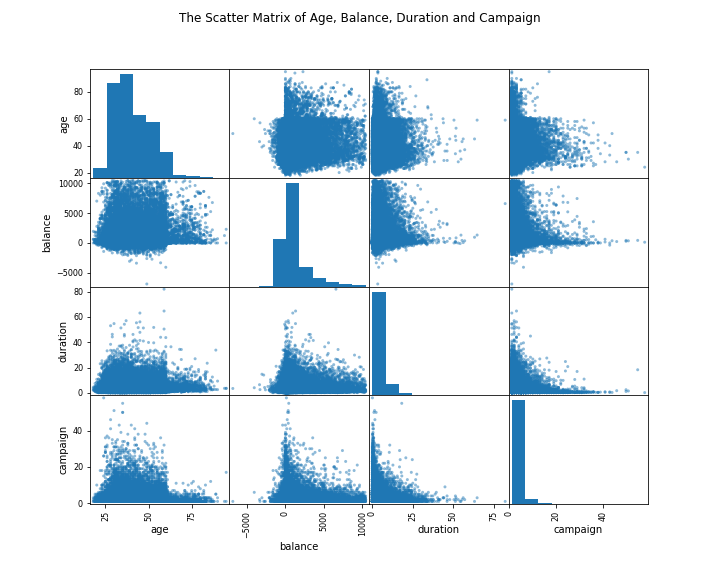
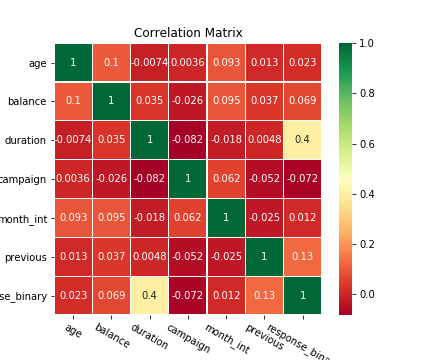
### **3.4 Visualize the relationship between phone call duration & the number of campaigns**



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 & Correlation matrix**

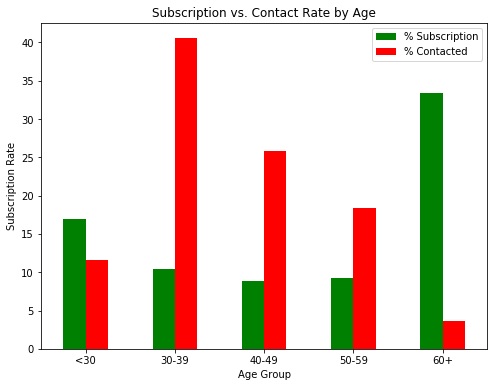
The scatter matrix does not reveal any clear relationship among age, balance, duration and campaign.

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.

# **Part 4. Data Visualization**

Now we have a good understanding of the distribution of key variables. Five plots will be generated to further investigate the influence of different customer characteristics on the subscription rate.

### **4.1 Visualize the subscription and contact rate by customer age**



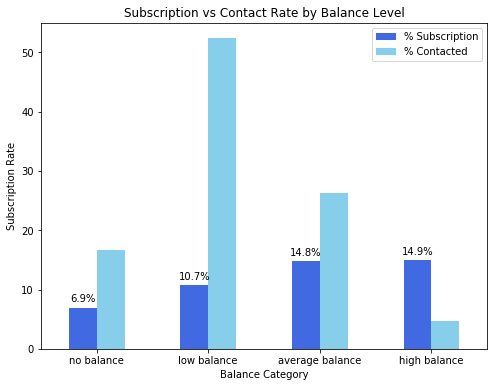
****Insights: target the youngest and the oldest instead of the middle-aged****

Green vertical bars indicate that clients with a age of 60+ have the highest subscription rate. About 17% of the subscriptions came from the clients aged between 18 to 29. More than 50% of the subscriptions are contributed by the youngest and the eldest clients.

* It is not surprising to see such a pattern because the main investment objective of older people is saving for retirement while the middle-aged group tend to be more aggressive with a main objective of generating high investment income. Term deposits, as the least risky investment tool, are more preferable to the eldest.
* The youngest may not have enough money or professional knowledge to engage in sophisticated investments, such as stocks and mutual funds. Term deposits provide liquidity and generate interest incomes that are higher than the regular saving account, so term deposits are ideal investments for students.

However, red vertical bars show that the bank focused its marketing efforts on the middle-aged group, which returned lower subscription rates than the younger and older groups. Thus, to make the marketing campaign more effective, the bank should target younger and older clients in the future.

### **4.2 Visualize the subscription rate by balance level**



****Insights: target clients with average or high balance****

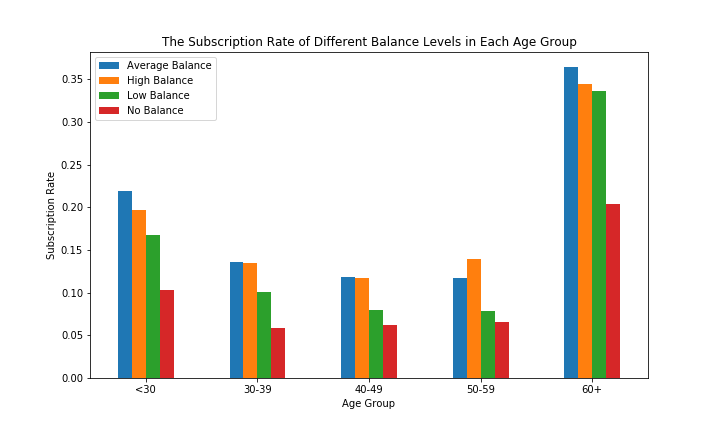
To identify the trend more easily, clients are categorized into four groups based on their levels of balance:

* No Balance: clients with a negative balance.
* Low Balance: clients with a balance between 0 and 1000 euros
* Average Balance: clients with a balance between 1000 and 5000 euros.
* High Balance: clients with a balance greater than 5000 euros.

Unsurprisingly, this bar chart indicates a positive correlation between clients’ balance levels and subscription rate. Clients with negative balances only returned a subscription rate of 6.9% while clients with average or high balances had significantly higher subscription rates, nearly 15%.

However, in this campaign, more than 50% of clients contacted only have a low balance level. In the future, the bank should shift its marketing focus to high-balance customers to secure more term deposits.

### **4.3 Visualize the subscription rate by age and balance**



****Insights: target older clients with high balance levels****

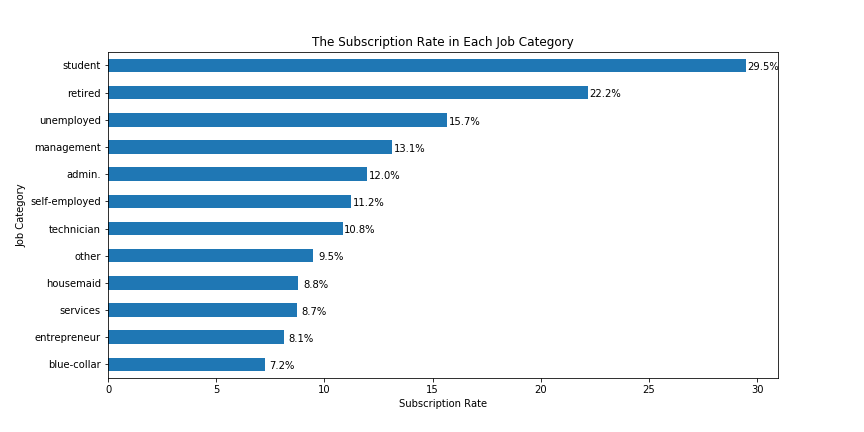
While age represents a person’s life stage and balance represents a person’s financial condition, jointly evaluating the impact of these two factors enables us to investigate if there is a common trend across all ages, and to identify which combination of client features indicates the highest likelihood of subscription.

In order to investigate the combined effect of age and balance on a client’s decision, we performed a two-layer grouping, segmenting customers according to their balance levels within each age group.

* The graph tells the same story regarding the subscription rate for different age groups: ****the willingness to subscribe is exceptionally high for people aged above 60 and younger people aged below 30**** also have a distinguishable higher subscription rate than those of other age groups.
* Furthermore, ****the effect of balance levels on subscription decision is applicable to each individual age group****: every age group shares a common trend that the percentage of subscription increases with balance.

In sum, the bank should ****prioritize its telemarketing to clients who are above 60 years old and have positive balances****, because they have the highest acceptance rate of about 35%. The next group the bank should focus on is ****young clients with positive balances****, who showed high subscription rates between 15% and 20%.

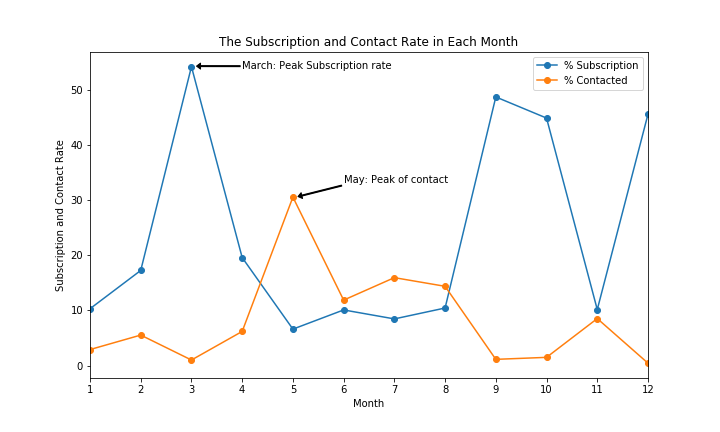
### **4.4 Visualize the subscription rate by job**



****Insights: target students and retired clients****

As noted from the horizontal bar chart, students and retired clients account for more than 50% of subscription, which is consistent with the previous finding of higher subscription rates among the younger and older.

### **4.5 Visualize the subscription and contact rate by month**



****Insights: initiate the telemarketing campaign in fall or spring****

Besides customer characteristics, external factors may also have an impact on the subscription rate, such as seasons and the time of calling. So the month of contact is also analyzed here.

This line chart displays the bank’s contact rate in each month as well as clients’ response rate in each month. One way to evaluate the effectiveness of the bank's marketing plan is to see whether these two lines have a similar trend over the same time horizon.

* ****The bank contacted most clients between May and August.**** The highest contact rate is around 30%, which happened in May, while the contact rate is closer to 0 in March, September, October, and December.
* However, the subscription rate showed a different trend. ****The highest subscription rate occurred in March****, which is over 50%, and all subscription rates in ****September, October, and December**** are over 40%.

Clearly, ****these two lines move in different directions which strongly indicates the inappropriate timing of the bank’s marketing campaign.****To improve the marketing campaign, the bank should consider initiating the telemarketing campaign in fall and spring when the subscription rate tends to be higher.

Nevertheless, the bank should be cautious when analyzing external factors. More data from previous marketing campaign should be collected and analyzed to make sure that this seasonal effect is constant over time and applicable to the future.

# **Part 5. Machine Learning: Classification**

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.

### **Prepare Data for Classification**

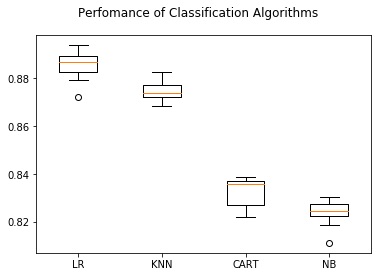
Select the most relevant customer information: job title, education, age, balance, default record, housing record and loan record

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: all customer statistics were selected as features while the campaign outcome was set as target. 80% of the data was used to build the classification model and 20% was reserved for testing the model.

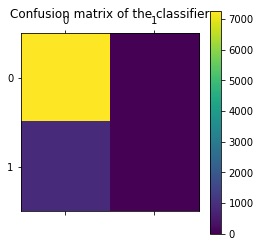
### **Build Classification Model**

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



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

Among all algorithms, logistic regression had the highest accuracy, about 88%, so it would be used to predict customers' responses. The test of logistic regression model successfully achieved an accuracy of 89.08%, suggesting high level of strength of this model to classify customers' responses given all the defined customer features. To evaluate the performance of the logistic regression model, a confusion matrix was created.

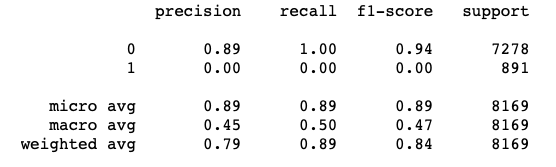


****However, the result of accuracy score can possibly yield misleading result if the data set is unbalanced, because the number of observations in different classes largely vary.****

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 ****the 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. 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.

### **Build Regression Model**

Six different regression algorithms (Linear Regression, Lasso, Ridge, ElasticNet, K Neighbors and Decision Tree) were run on the dataset and the best-performing one would be used to build the regression model. As we can see, ridge regression slightly outperformed other models, and the same was true after standardization.

The ridge regression model had an MSE of 17.78. 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**

****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.****

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.

# **Part 8. Recommendations**

****1. More appropriate timing****

When implementing a marketing strategy, external factors, such as the time of calling, should also be carefully considered. The previous analysis points out that March, September, October and December had the highest success rates. Nevertheless, more data should be collected and analyzed to make sure that this seasonal effect is constant over time. If the trend has the potential to continue in the future, the bank should consider initiating its telemarketing campaign in fall and spring.

****2. Smarter marketing design****

By targeting the right customers, the bank will have more and more positive responses, and the classification algorithms would ultimately eliminate the imbalance in the original dataset. Hence, more accurate information will be presented to the bank for improving the subscriptions. Meanwhile, to increase the likelihood of subscription, the bank should re-evaluate the content and design of its current campaign, making it more appealing to its target customers.

****3. Better services provision****

With a more granular understanding of its customer base, the bank has the ability to provide better banking services. For example, marital status and occupation reveal a customer's life stage while loan status indicates his/her overall risk profile. With this information, the bank can estimate when a customer might need to make an investment. In this way, the bank can better satisfy its customer demand by providing banking services for the right customer at the right time.

# **Part 1. Project Background**

Nowadays, marketing expenditures in the banking industry are 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 Objectives: predict customers' responses to future marketing campaigns & 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 behaviors 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. About the Data**

The 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.](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

There are 41,188 observations in the dataset, with no missing values. 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 marketing phone calls. "Yes" means that a customer subscribed to term deposits.

### **Data Cleaning**

Several changes were made to the dataset to prepare it for analysis.

Drop ambiguous values, such as "others" and "unknown".

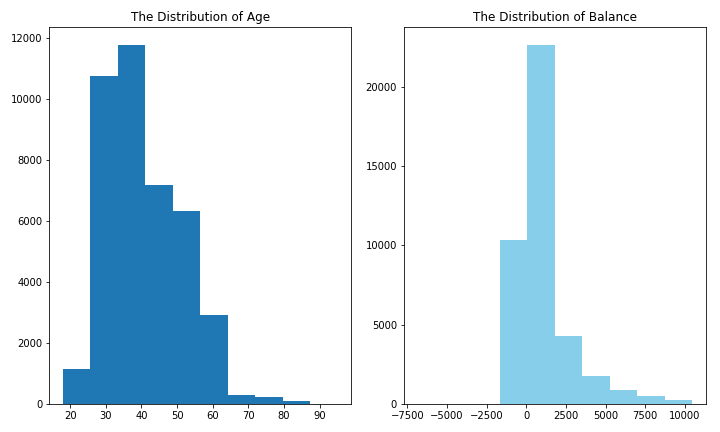
Drop outliers to capture the general trend. Outliers are defined as customers whose balance levels are either too high or too low (i.e., more than three standard deviations away from the mean).

Change the "response" variable (yes/no) to binary values (1/0) 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 customer age and balance levels:**



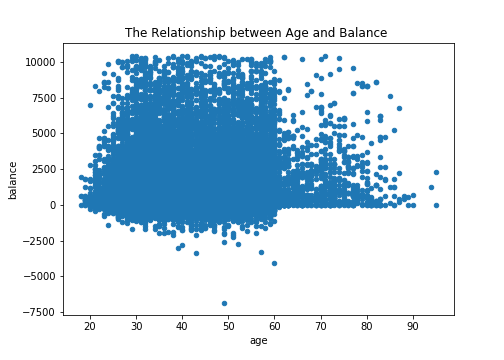
****The distribution of age:****

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:****

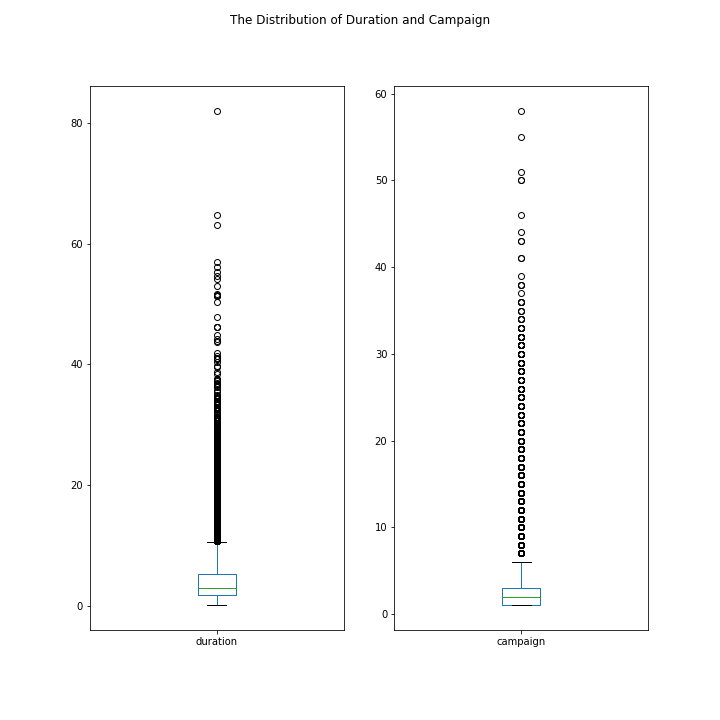
After dropping outliers in balance, the range of balance is still massive, from a minimum of -6847 to a maximum of 10443 euros, giving a range of 17290 euros. The distribution of balance has a huge standard deviation relative to the mean, suggesting large variabilities in customers' balance levels.

### **3.2 Visualize the relationship between customer 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 phone call duration & the number of campaigns**



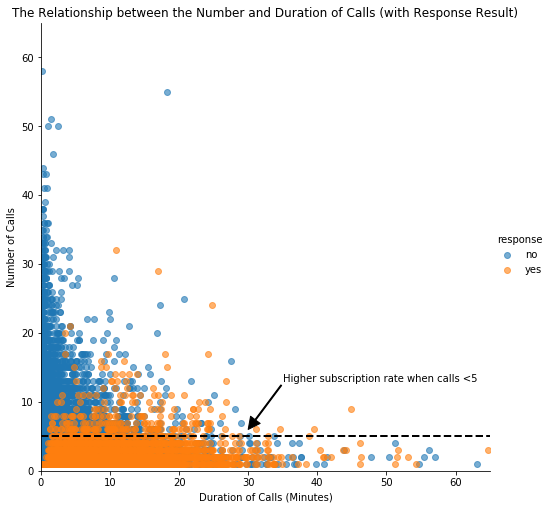
**The distribution of duration**

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 campaign**

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.

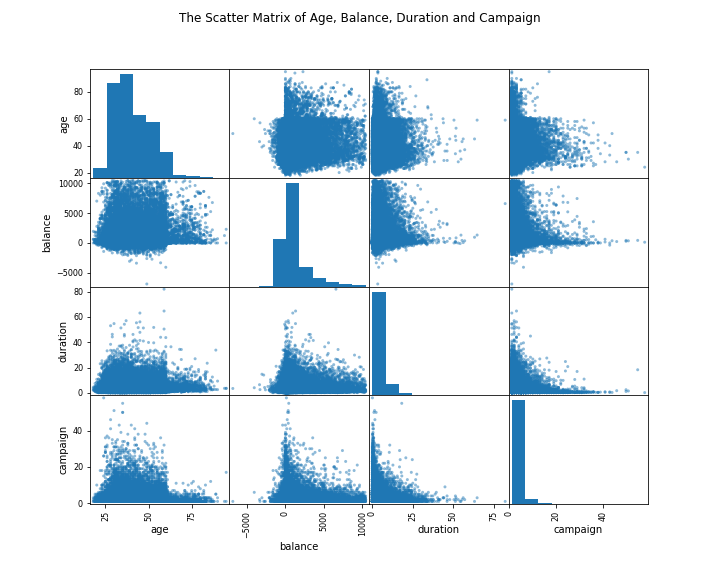
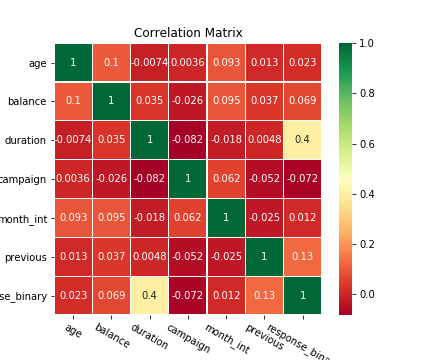
### **3.4 Visualize the relationship between phone call duration & the number of campaigns**



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 & Correlation matrix**

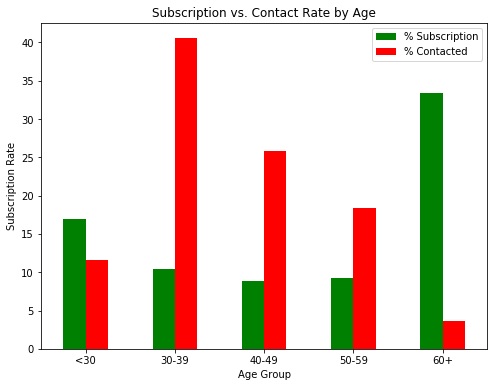
The scatter matrix does not reveal any clear relationship among age, balance, duration and campaign.

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.

# **Part 4. Data Visualization**

Now we have a good understanding of the distribution of key variables. Five plots will be generated to further investigate the influence of different customer characteristics on the subscription rate.

### **4.1 Visualize the subscription and contact rate by customer age**



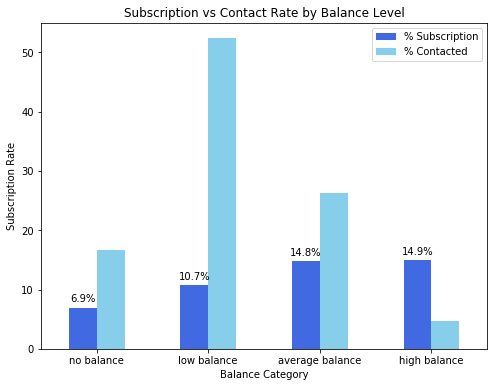
****Insights: target the youngest and the oldest instead of the middle-aged****

Green vertical bars indicate that clients with a age of 60+ have the highest subscription rate. About 17% of the subscriptions came from the clients aged between 18 to 29. More than 50% of the subscriptions are contributed by the youngest and the eldest clients.

* It is not surprising to see such a pattern because the main investment objective of older people is saving for retirement while the middle-aged group tend to be more aggressive with a main objective of generating high investment income. Term deposits, as the least risky investment tool, are more preferable to the eldest.
* The youngest may not have enough money or professional knowledge to engage in sophisticated investments, such as stocks and mutual funds. Term deposits provide liquidity and generate interest incomes that are higher than the regular saving account, so term deposits are ideal investments for students.

However, red vertical bars show that the bank focused its marketing efforts on the middle-aged group, which returned lower subscription rates than the younger and older groups. Thus, to make the marketing campaign more effective, the bank should target younger and older clients in the future.

### **4.2 Visualize the subscription rate by balance level**



****Insights: target clients with average or high balance****

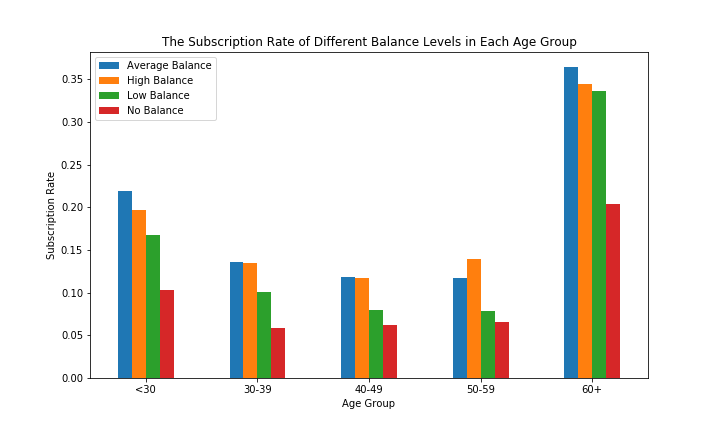
To identify the trend more easily, clients are categorized into four groups based on their levels of balance:

* No Balance: clients with a negative balance.
* Low Balance: clients with a balance between 0 and 1000 euros
* Average Balance: clients with a balance between 1000 and 5000 euros.
* High Balance: clients with a balance greater than 5000 euros.

Unsurprisingly, this bar chart indicates a positive correlation between clients’ balance levels and subscription rate. Clients with negative balances only returned a subscription rate of 6.9% while clients with average or high balances had significantly higher subscription rates, nearly 15%.

However, in this campaign, more than 50% of clients contacted only have a low balance level. In the future, the bank should shift its marketing focus to high-balance customers to secure more term deposits.

### **4.3 Visualize the subscription rate by age and balance**



****Insights: target older clients with high balance levels****

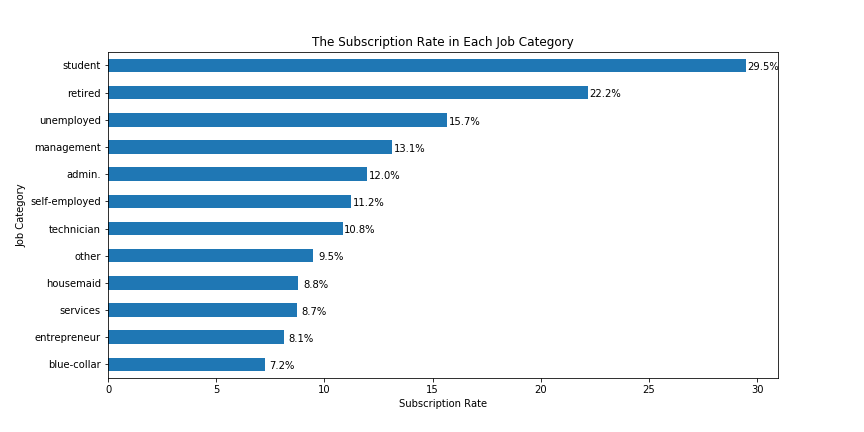
While age represents a person’s life stage and balance represents a person’s financial condition, jointly evaluating the impact of these two factors enables us to investigate if there is a common trend across all ages, and to identify which combination of client features indicates the highest likelihood of subscription.

In order to investigate the combined effect of age and balance on a client’s decision, we performed a two-layer grouping, segmenting customers according to their balance levels within each age group.

* The graph tells the same story regarding the subscription rate for different age groups: ****the willingness to subscribe is exceptionally high for people aged above 60 and younger people aged below 30**** also have a distinguishable higher subscription rate than those of other age groups.
* Furthermore, ****the effect of balance levels on subscription decision is applicable to each individual age group****: every age group shares a common trend that the percentage of subscription increases with balance.

In sum, the bank should ****prioritize its telemarketing to clients who are above 60 years old and have positive balances****, because they have the highest acceptance rate of about 35%. The next group the bank should focus on is ****young clients with positive balances****, who showed high subscription rates between 15% and 20%.

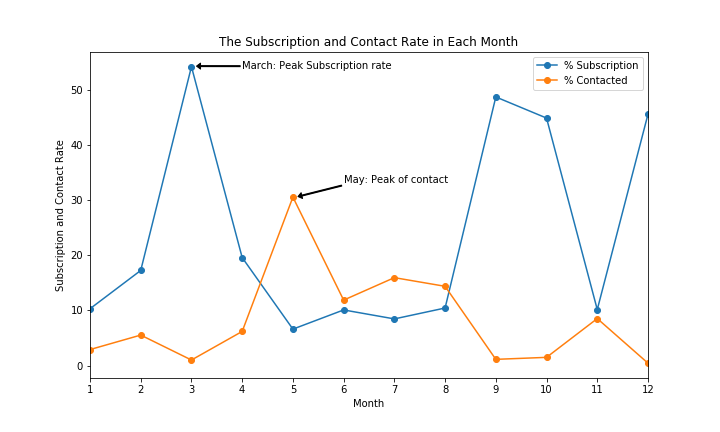
### **4.4 Visualize the subscription rate by job**



****Insights: target students and retired clients****

As noted from the horizontal bar chart, students and retired clients account for more than 50% of subscription, which is consistent with the previous finding of higher subscription rates among the younger and older.

### **4.5 Visualize the subscription and contact rate by month**



****Insights: initiate the telemarketing campaign in fall or spring****

Besides customer characteristics, external factors may also have an impact on the subscription rate, such as seasons and the time of calling. So the month of contact is also analyzed here.

This line chart displays the bank’s contact rate in each month as well as clients’ response rate in each month. One way to evaluate the effectiveness of the bank's marketing plan is to see whether these two lines have a similar trend over the same time horizon.

* ****The bank contacted most clients between May and August.**** The highest contact rate is around 30%, which happened in May, while the contact rate is closer to 0 in March, September, October, and December.
* However, the subscription rate showed a different trend. ****The highest subscription rate occurred in March****, which is over 50%, and all subscription rates in ****September, October, and December**** are over 40%.

Clearly, ****these two lines move in different directions which strongly indicates the inappropriate timing of the bank’s marketing campaign.****To improve the marketing campaign, the bank should consider initiating the telemarketing campaign in fall and spring when the subscription rate tends to be higher.

Nevertheless, the bank should be cautious when analyzing external factors. More data from previous marketing campaign should be collected and analyzed to make sure that this seasonal effect is constant over time and applicable to the future.

# **Part 5. Machine Learning: Classification**

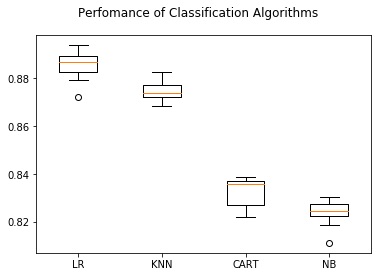
The main objective of this project is to identify the most responsive customers before the marketing campaign so that the bank will be able to efficiently reach out to them, saving time and marketing resources. To achieve this objective, classification algorithms will be employed. By analyzing customer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

### **Prepare Data for Classification**

* Select the most relevant customer information: job title, education, age, balance, default record, housing record and loan record
* 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: all customer statistics were selected as features while the campaign outcome was set as target. 80% of the data was used to build the classification model and 20% was reserved for testing the model.

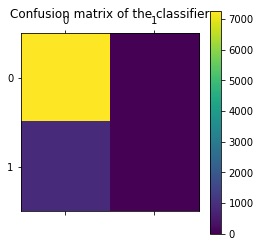
### **Build Classification Model**

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



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

Among all algorithms, logistic regression had the highest accuracy, about 88%, so it would be used to predict customers' responses. The test of logistic regression model successfully achieved an accuracy of 89.08%, suggesting high level of strength of this model to classify customers' responses given all the defined customer features. To evaluate the performance of the logistic regression model, a confusion matrix was created.

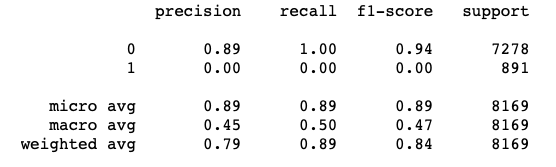


****However, the result of accuracy score can possibly yield misleading result if the data set is unbalanced, because the number of observations in different classes largely vary.****

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 ****the 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. 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.

### **Build Regression Model**

Six different regression algorithms (Linear Regression, Lasso, Ridge, ElasticNet, K Neighbors and Decision Tree) were run on the dataset and the best-performing one would be used to build the regression model. As we can see, ridge regression slightly outperformed other models, and the same was true after standardization.

The ridge regression model had an MSE of 17.78. 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**

****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.****

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.

# **Part 8. Recommendations**

****1. More appropriate timing****

When implementing a marketing strategy, external factors, such as the time of calling, should also be carefully considered. The previous analysis points out that March, September, October and December had the highest success rates. Nevertheless, more data should be collected and analyzed to make sure that this seasonal effect is constant over time. If the trend has the potential to continue in the future, the bank should consider initiating its telemarketing campaign in fall and spring.

****2. Smarter marketing design****

By targeting the right customers, the bank will have more and more positive responses, and the classification algorithms would ultimately eliminate the imbalance in the original dataset. Hence, more accurate information will be presented to the bank for improving the subscriptions. Meanwhile, to increase the likelihood of subscription, the bank should re-evaluate the content and design of its current campaign, making it more appealing to its target customers.

****3. Better services provision****

With a more granular understanding of its customer base, the bank has the ability to provide better banking services. For example, marital status and occupation reveal a customer's life stage while loan status indicates his/her overall risk profile. With this information, the bank can estimate when a customer might need to make an investment. In this way, the bank can better satisfy its customer demand by providing banking services for the right customer at the right time.