# Analysis of Marketing Campaign of personal loan and deposit in bank XYZ

Group S: 404 NOT FOUND

May 9, 2019

Team member: Cheng Zhang, Zhaoyang Wang, Tianyi Zhang

Email address: cz2532@columbia.edu,zw2551@columbia.edu,tz2390@columbia.edu

If you have any question or interested in more detailed information about this project, please don't hesitate to contact us!

## Introduction of the project

This project is to analysis a marketing campaign of personal loan and a marketing campaign of personal deposit. Although datasets for loan and deposit are kind of independent, we assume that they are project for one bank, the bank XYZ. The fictitious situation is that bank XYZ launch a marketing campaign for months and we get many detailed data for this campaign, we are working on this project in order to review and improve the campaign effectiveness and meet requirement of limited budget of bank XYZ. Besides, since the bank faced a liquidity risk, in other words, it's lack of money, it also launched a deposit campaign. If bank XYZ can have enough money, we no longer need to set a budget for the profitable personal loan. We already get some historical data of the marketing campaign of deposit, because the bank has already aware of the liquidity risk and want to get more money. Unfortunately, Since the campaign just started, we just get the preliminary data of this campaign But we can still analysis that dataset to increase the efficiency of the deposit campaign. More detail introduction of the explored questions are in the "Our investigation" part in our report.

## Sources of Data

Actually, the personal loan campaign data set is from HSBC, they are data sets for an onsite competition about data mining. Here is the link:https://github.com/alan-chengzhang/Applied-Data-Science-Final. All the csv file start with the name "DUMMY" are the personal loan datasets. The deposit dataset is from UCI machine learning database. Link: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, you can also find it in that GitHub link. Only the bank-full-additional.csv is used for our project. Two dictionaries of datasets are in previous GitHub link, with the name starts with "Dictionary".

HSBC is an international bank whose headquarter. Since the original data sets are used for a data mining competition and the company is seeking for solutions from competitors, the data should be reliable and accurate. Besides, after union, merged and clean the data (after data processing), there are more than 600,000 rows and this international bank has loan service all around the world, it should be representative of the population we are studying.

The deposite dataset is from UCI machine learning database and is widely used, wcich indicate the data is useful for this kind of topic. After the data cleaning and feature selection process, the data should be reliable and accurate. Also the owner of the dataset, Portuguese banking institution, is also an international bank and has the service all around the world, it should be representative of the population we are studying.

## Structure of this report

There will be two big part in this report. The first part is about the personal loan campaign and the second part is about the deposit campaign.

In each part, the following subtitles will included: "examination of the data", "Our investigation", "The results", "Interpretion". The content will be as follows: the datasets and the variables will be briefly introduced and then the application (shiny app) will be introduced. After that, we will introduce the data cleaning, feature selection process and how we make the datasets reliable. The last and the major part is the analysis of the data. The analysis part will mainly focus on giving the message to our business partner, not the data scientist. So, the model will just be introduced intuitively and most part will be the inference and explanation of our models and the results.

The “assumptions”, “limitations and uncertainties”, “area of future investigation” and “references” will be the last four parts of this report.

Please note that many important plots are in the slides but not in this report, since the page is limited.

# Part one: Analysis of personal loan marketing campaign

## Examination of Data

There are three data sets: Campaign List: Contains all the information of this campaign.

Lending Product Information: Historical loan usages and credit information of clients.

Customer Information: (6 tables for 6 months) Detailed features of clients which are somewhat relevant to the campaign.

The page is limited, all the detail information can be found in the dictionary.

### Data cleaning and feature selection.

Please refer to “data_cleaning.docx” for the part of handling missing values.

After that we find some columns have extremely large values in some columns. If data is large than mean plus 3 times of standard deviation then it is defined as outlier and that roll should be deleted.

After that we find some columns have extremely large values in some columns. If data is large than mean plus 3 times of standard deviation then it is defined as outlier and that roll should be deleted.

There are still 650564 rows of cleaned and filtered data, which is more than 90% of original dataset.We just delete two features and less than 10% rows in these process, which is good. Now, the data is reliable for the following analysis. We also build a shiny app to our business partner and the brief intriduction is as follows:

### Introduction of shiny application

Please refer to the “into_of_shiny.docx”

## Our investigation

As for a marketing campaign, what we care more is how to use current data to make it more effecient and more porfitable. Here are three topics and related models we are going to use in this part.

### Topics for data analysis

#### Topic 1: Effectiveness of loan campaign: increase the response rate

Methods: logistic regression (with stepwise selection), random forest, adaboost

The response rate is not high (shown in slides), we should find a way to improve that. In this topic, we care about both inference and accuracy(prediction result). So, we first use a logistic regresion with stepwise selection to see the influence of features to the response. Then random forest model and adaboost model are used for better prediction result. Finally, a case will be used (in “Interpretation”) to show the influence of our model on the selection of target clients.

#### Topic 2: Channel selection: Choose the best channel to contact selected clients.

Methods: random forest, adaboost

After select the target user, the following question is how to connect them.

In this part, the only thing we care about is the accuracy, so we don't use any regression model.Just as the first topic, randome forst and adaboost will be used.

Topic 3: Loan amount: Choose top clients with the highest loan amounts

Methods: linear regression(with stepwise selection), random forest

Regardless of the risk management process, banks always prefer larger amount of single loan. Sometimes the loan campaign has a budget, we just can lend money to some of the target clients. However, we won't know how much our clients want to borrow unless we make future discussion with them, this further discussion also cost a lot for the bank. So, we cannot blindly choose clients when there are limit budget, but need to predict loan amount for each target clients before head.

In this topic, we care about both inference and accuracy (prediction result). So, we first use a regression with stepwise selection to see the influence of features to the loan amount. Then a random forest model is used for better prediction result. Finally, a case will be used to show the influence of our model on the clients selection.

## Standard to check the outcomes models

When we have multiple models and we want to select the base one, some standards are needed.

In the marketing situation, what we care more is the recall, since we are trying to select as many as the target clients who are intersted in the loan. Definition of recall: True positive / (True positive + False Positive). In other word, it's a porpotion of number of target user comparing to number of all target user in the data sets.

Besides, we also care about the accuracy of our prediction (selection). So, the overall accuracy will be shown together with the recall.

To follow a standard procedure of machine learning model, the ROC curve and the AUC value will be shown. The ROC curve is to seek for a balance between ture positive and false positive, which shows the accuracy of clssify into true under certain level of mistake. The AUC value is the area under the curve. The more the ROC curve close to upperleft side, the higer AUC value will be, which means a better model for the prediction.

Finally, for a prediction problem for numeric variable, rathe than the categorical variable, we can no longer use all the previous methods. We will use r-squared to test how good the model fit the data. It explains how much information of data are captured by the model. Of course, all the previous standards are calculated in the test set. The original dataset is seperated into training and test set by 7:3.

# The results

Some dignostic plots about the training process are not included in this report, since it's more about story telling. Please refer to the original code the these results.

## Topic1: Effectiveness of loan campaign

### Handle imbalance response

```
##             type    num
## 1:     response  22214
## 2: not.response 627987
```

In the original dataset, we want to make number of responsed clients and number of not response to be balanced, so we will randomly select 22214 rows in the dataset of not responsed clients to make the data balanced. Since the number of rows of very large, there is no need to use the resampling strategy, which may distroy the original distribution.

### logistic regression model for inference

We will chase for the accuracy for the prediction model, in other words, select as many target clients in the original dataset using some machine learning model. But before that, we need first focus on the inference of the variables. We want to now the influence of some variables to the response. In this case, logistic regression is the best choice. Besides, some of the feature may represent similar pattern of clients, so we use a stepwise selection procedure to handle this problem and do the feature selection. The model is not for the good prediction outcome, but for the inference based on the odds ratio.

Firstly, we show the performance of the model. The accuracy(0.754248), recall(0.7515379), and AUC(0.8338 762) value are not very high. But that is acceptable for a logistic regression model.
Odds ratios for all the selcted features by the stepwise procedure are in slides. If the odds ratio is greater than 1, the feature will have a positive influence on response. If it is less than 1, the feature wll have negative influence on response.

If we don't get enough data about a client, we can get a basic idea of him based on the inference results.

The most positive influential feature if the msk_ever_loan_user. It means if the customer ever had any loan product before, the probability of response will be 3.365 times than before. It's reasonable since that group of clients are more interested in the personal loan.

The most negative influential feature if the msk_ever_loan_user. It means if the customer's Total deposit balance by month-end of the data month increase by one unit, the probability of response will be 0.5279 times than before. It's reasonable since that if clients like to save money and they have enough saving, they don't need and are also not willing to have a personal loan.

## Random forest model for better prediction.

Random forests is an ensemble learning method for classification and regression by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

It's better than regression since it can capture the non-linear trend. Besides, it's also better than the decision tree since it using ensemble technic to deal with overfitting to their training set.The tunning of hyperparameters and training the model are very carefully done and result of model is as follows:

The recall(0.8907727), accuracy(0.8853814) and AUC(0.918467) value are all around 0.9. It's much better than the logistic regresion model and are all good prediction outcomes. ROC curve is in the slides.

Sone inference can be drawn from the feature importance (please refer to slides), the higher the number, the model important the feature will be. The msk_total_saving_bal is the most influential feature which is same as the result in logistic regression. But the msk_ever_loan_user is not that importance. The difference is caused by the nonlinear trend in the features.

One weaknees of the model is that we need to carefully tune the hyperparameters in order to get a better result, which is time consuming and not very friendly to business partner.

## Adaboost model for avioding tuning parameters

As mentioned before, the weakness of random forest is that it needs us to tune the parameters carefully. It makes our code not completely reproducible for new datasets, and some business partners may be confused with that.

In this part, we want to build a model that makes the whole process to be easy without parameters. In this case, the adaboost model is the optimal choice.

AdaBoost, short for Adaptive Boosting, can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. Since it's formed by a group of weak learners, it no longer needs those hyperparameters. Besides, if the data sets are well cleaned without outliers, theoretical the adaboost will avoid the overfitting.

We can choose the round of training in this model. The larger the number, the better the prediction will be. However, the prediction accuracy will become stable in a large number of round, i.e. 200-300 rounds.

The “fastAdaboost” package enable us to apply the method 40-50 times faster than the traditional “adabag” package. However, it’s still relatively slow. For now, it need 20-30 minutes to finish the training process. Time consuming for training the model is a major weakness for the adaboost. But tuning the parameters in random forest needs longer time. So, this time is acceptable.

The round of trainning is set to be 100 (in order to be faster). The recall(0.9294824) and accuracy(0.9250877) of adaboost are relatively high that random forest, but the AUC(r response.auc.ada) value is lower. It shows that adaboost is better at selecting target clients but make more mistakes in the same time. It’s hard too tell which one between adaboost and random forest is better.

## Topic 2: Channel selection for target clients

We select those responded clients (with response=1, contact_ind=Y) and determine which channel are the best to contact those clients. We only care about prediction result, so no infernce part will be included in thie topic. Besides, the following analysis are based on this assumption：  In the historical data. If the clients response, then the channel is regarded as best channel for him.

### *Handle the imbalance of channel*

Since the total number of rows is small, we use resampling to make them balance. To be more specific, after getting the class largest number of rows, we sample the rows of other classes with replacement and make the row number to be the same with the largest number. Then, different classes of channel are balanced.

```
##       channel    N
## 1: Phone Call 7680
## 2:      Email 1351
## 3:        SMS 4000
```

### Random forest for channel selestion

The parameters are carefully tuned. Please refer to the original code for the training process and some dignostic plots about the training. Since we want to control the error of each class, what we care is the recall and accuracy of each class. So, we use confusion matrix to calculate them. The following is the recall for each class and the overall accuracy.

```
dt.recall
```

```
##                  recall
## recall.email 0.9995656
## recall.phone 0.9789880
## recall.sms   0.9907814
```

The recall of Phone call and SMS are relatively smaller.The accuracy is 0.989728. In the confusion matrix, we also find there are more misclssification of these two classes. It may because we resample the data and there are some repetitive rows. If the original dataset is balanced or we make the raw number to be the same as the smallest one, the recall and accuracy tends to be higher.

### Adaboost for channel selestion

Also, we use adaboost model, which is a more robust and user friendly model without hyperparameters. With a limitation of the package, which can only handle 0-1 dependent variable, we need to train three model for three channel and combine the result together. In order to finish it faster, the round is set to be 100. It will take 20-30 minutes to train three models, which is similar to the time of tunning parameters in random forest.

The recall and accuracy for each channel are as follows

```
##              phone     email       sms
## recall   0.9839410 0.9978299 0.9683160
## accuracy 0.9893125 0.9985541 0.9790681
```

The outputs are almost the same with the random forest, but slightly worse. Besides, they also show the same pattern cased by resampling. If we use more round, the result should be the same or better than the ramdon forest.

That is to say, when you want to choose the best channel for target clients, you are choose either of these two models, which are all lead to a good result.

## Topic3: Choose top clients with the highest loan amounts

In this topic, since the dependent variable is numeric, we use r-squared on the testing set to evaluated the how good the prediction is. r-squared is a standard of how much information in the data sets are explained by the model. In other words, it's a percentage of the dependent variables variation that model explains.

### Regresion model with stepwise selection.

As mentioned before, some of the feature may represent similar pattern of clients, so we use a stepwise selection procedure to handle this problem and do the feature selection. The model is not for the good prediction outcome, but for the inference based on the coefficient of selected features.

Firstly, we show the performance of the model. The r-squred(0.7708401) is not very high, but good enough for a model for inference.

The coefficient (showed in slides) means that, when the value of one feature increase 1, the amount of loan will change according to the number of that coefficient.

The most positive feature for the loan amount is msk_total_deposit_trn_count. It shows if the customer's total transaction count increase one unit, the loan amount will increase 891.3494. It may because this client group have enough money but need some working cash, so they tend to borrow more money and have the ability to pay them back.

The most negative feature for the loan amount is msk_instl_loan_userY. It shows that if customer ever had personal loan product, the loan amount will decrease 39832.9774. It tells that even though this group of clients are the most positive for the response in topic 1, the loan amount of them is relatively low. It may because they are not willing to borrow too much money, since they are in shortage of money and don't have ability to pay large amount loan in the future or may have some other loans. Besides, some risk management process in the bank may control the amount of loan to this client group.

### Random forest model for better prediction

Since the regression model just take the linear trend into account, the prediction result is not very good. Now, we use the random forest which can capture the non-linear trend to get a better prediction accuracy. Now that the respond variable is the numerical data, the type of random forest becomes to be "regression", which means that it uses a set of tree based regression functions to predict then result.We can get the r-squared of each tree model. So, we use average r-squared to evaluate the model and compare it to the stepwise regression. (Please refer original code about the result in training process)

In this model, the importance of features(showed in slides) shows that the most influential feature is the msk_unauth_over_card_limit. It shows that whether clients ever had other unsecured lending product is most influential to the loan amount. It might because of the risk management system in the bank. The difference of these result and result from regression is because of the non-linear for the features.

The r-squred is 0.9622969, which is extremely high. It means that most information of features are explained by this random forest model.

# Interpretation

## Topic 1: Effectiveness of loan campaign

In "The results", some inferences are drawn from the odds ratio from the logistic regression model. If we don't have enough data, we can get an overview of clients given the odds ratio.

Besides, the recall, accuracy and AUC value are all good for random forest and adaboost model. When we have similar problem with same columns of data, we can always use these two models to select clients and increase the effectiveness. If we understand about tuning parameters, we can use random forest. If not, adaboost is a better choice with similar results. We use a case to show the influence of these models.

### A case to show the influence of models

We apply our model to the whole dataset and check the improvement of the efficiency. When our target is the response rate, what we care more is how much the efficiency is increased comparing to random select. So, we use the term "lift" to show that. The definition of lift is the recall of the model divided by accuracy for response clients of random select.

It shows that in the whole dataset, the lift is 28.3080036, which mean it is 28.3080036 times compare to the ramdom select. It means our model is very useful to selct the target clients. If you use our model to do this selction, the effecicy will be 28.3080036 than contact clients blindly.

The lift is 28.3093212, shows that the prediction quality of adaboost model is similar with the random forest model.

You can use either of them to select target clients for higher response. If you don't want to spend time to tune the parameters of cannot understand the meaning of parameters, the adaboost is a better choice.

## Topic 2: Channel selection for target clients

In this topic, we only care about the prediction output and the recall and accuracy are all very high.

As for random forest mode. If you use this model to choose the best channel to contact with target clients, the accuracy will be 99% percent. The outputs are almost the same with the random forest, but slightly worse. Besides, the result will even better if there is no imbalance situation or we use more rounds in our adaboost model.

So, don't hesitate to use the model in real cases and you can choose either of these models.

Of course, the recalls and accuracy are too high that is even counterintuitive. It may because of the assumption we make at the beginning of this topic. However, even though the assumption is sometimes not very valid in real cases, we still find proper way to contact with clients, which must be better than the random select (with accuracy of 33%).

## Topic 3: Choose top clients with the highest loan amounts

In "The results", some inferences are drawn from the coefficients from the regression model. If we don't have enough data, we can get an overview of clients given the coefficients.

As for the random forest model, the r-squred extremely high. It means that most information of features are explained by this random forest model. So, When you have similar task for selecting the clients with the higest loan amount, you can always trust the random forest model.

We also use a case to show the influence of the model to the selection precess.

### A case to show the influence of model

Now, since the amount of deposit is limited, we only allowed to give 1000 clients the loan in this campaign. We need to find a way to target the clients who will borrow the largest amount of money. We will use the model in the whole unique dataset of loan_amount which contain all unique information of clients who get the loan and see the increase comparing to the ramdon select. The increment is as follows. Besides, The less the number of chosen clients, the higher the increment will be.

The result shows that, if we use the model to select the clients, the total increase will be 37.434776% comparing to the random select. You can expect a high increment if the volumn of the campaign is higher. This randome forest model is very useful for the real cases.

# The next step: Analysis of marketing campaign for deposit

The limited budget in personal loan shows the bank may in shortage of amount of liquidity and want to avoid protential liquidity risk. It show the bank also need a marketing campaign for deposit.

Since the bank is already aware of the liquidity risk, it has launched a deposit campaign. If bank XYZ can have enough money, we no longer need to set a budget for the profitable personal loan.

We already get some historical data of the marketing campaign of deposit, because the bank has already aware of the liquidity risk and want to get more money. Unfortunately, Since the campaign just started, we just get the preliminary data of this campaign But we can still analysis that dataset to increase the efficiency of the deposit campaign.

So, the second part of this article will be the analysis of a marketing campaign dataset of the deposit.

# Part two: Analysis of deposit marketing campaign

## Introduction

Marketing selling campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. In order to explore the customer decision of deposit, we use Logistic Regression model and support vector machines (SVMs) model to predict the decision of customers. LR have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks. SVM are more flexible when compared with classical statistical modeling, presenting learning capabilities that range from linear to complex nonlinear mappings. Due to such flexibility, SVM tend to provide accurate predictions, but the obtained models are difficult to be understood by humans. After building model, we can compare the model result and discuss which model is better for market campaign of deposit dataset and discuss the further issue in our process.

## Sources of Data

We use market campaign dataset which is related with direct marketing campaigns of a Portuguese banking institution (the summary is in original code). The scenario is that clients contact customers to ask whether customer what to subscribe long term deposit or not. The result must be binary value which is “yes” or “no”. Meanwhile, after each connection, several information will be recorded in dataset including client basic information, connection information and economic information.

## Examination of the Data

### Issues of values “unknown”

Some information cannot be recorded in dataset and by default these values are set to “unknown” in original dataset and we should change these values to NA value.

### Unused variable

More than eight thousand record in “default”" feature are NA and compare the total record of our dataset, nearly a quarter of “default”" variable is NA. We assume that this variable has little impact to outcome and just ignore it.

```
##    default     N
## 1:      no 32588
## 2:    <NA>  8597
## 3:     yes     3
```

The last step is ignoring the record which has at least one NA value. After finishing this the size of our dataset is below.

```
dim(dat)

## [1] 38245    21
```

# Our Investigation

Using market deposit data, we try to build machine learning model to predict customer decision without any cost of connection. According to the basic information of customers, if bank can predict the probability of success. Some customer that would not subscribe service will be filtered before connection. To finish this process, we need to do variable selection and modeling. We decide to choose logistic regression and SVM model to predict outcomes.

## Vriable Selection

Market campaign is based on phone call and customers are not interested the background of clients. The basic information about clients have little impact to results. So variable about background information about client can be ignored in model. Please refer to the plot in the original code.

We do variable selection and plot result. The “duration”" variable has high impact to customer decision and if this feature is included in model, other feathers’ contribution to model result will decrease. On the other hand, for future prediction, duration is unknow value before connections happen. According the reason above, “duration” variable should be dropped from model.

In conclusion, features in model are contact, day_of_week, campaign, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m,nr.employed.

## Modeling

### Logistic Regression

Multinomial logistic regression (MLR) is a classification method that generalizes logistic regression to multiclass problems. It can be used when the response variable in question is nominal. Rather than modeling this response variable directly, MLR models the probability that response variable belongs to a particular category. We fit logistic regression model in R and get prediction accuracy and ROC curve.

### SVM Model

Support vector machine (SVM) aims to find the hyperplane that has the max margin to separate the classes. The points that support the max margin hyperplane are called support vectors. In order to avoid the no-separation cases, slack variables are introduced for more general cases.

In a binary classification case, for a set of$n$ training observations,$x_1, x_2, \dots x_n \in \mathbb{R}^p$,and class labels $y_1, y_2, \dots, y_n \in \{-1,1\}$. The support vector machine can be derived by operating the following optimization:

$$max_{\beta_j,\epsilon_j,M} M \ subject \ to \sum \beta_j^2 = 0$$

$$y_i(\beta_0 + \sum_{i \in S} \alpha_i \, k(x, x_i)) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Here $\epsilon_i$ are slack variables, $C$ is the budget of slackness, $k(x, x_i)$ is the kernel of the SVM, $\beta_j$are variables in the kernel. The RBF (radial basic function) kernel used in this paper is:

$$k(x, x_i) = e^{-\frac{\gamma(x - x_i)^2}{\sigma}}$$

## The result

### Logistic Regression

Please refer to original code and slides for the detailed information of the model

All coefficients in logistic regression model are great and the accuracy(0.9028432) of this model is nearly 90%. It is obvious that logistic regression fit our original data well. But when analyze ROC curve, it is not as great as we expect.

### SVM Model

The accuracy of SVM model is also beyond 90% and ROC curve of SVM is better than logistic model. As we mentioned on the introduction, SVM are more flexible when compared with classical statistical modeling, presenting learning capabilities that range from linear to complex nonlinear mappings. It trends to produce more accurate prediction.

## Interpretation

According the result above, all the variable in market campaign deposit dataset are meaningful and bank can use this dataset to predict customer decision in the future. For the deposit service, bank can make some prediction and filter the customers with low probability to subscribe deposit service, which will decrease the cost of connection. As we mentioned in introduction part, this is achievable.

# Assumptions

Assumption 1: The columns (features) in the dataset are all relevant to our target variables. Reason: Since the datasets are from competition and machine learning database, the provider of these datasets already gives the some topics of analysis, which show that the features are related to our topic for analysis.

Assumption 2: The datasets are reliable and representative. Reason: Please refer to the "sources of data" part of this report.

Assumption 3: For the topic two (Channel selection) of first part, in the historical data, if the clients response, then the channel is regard as best channel for him. Reason: Since the client's response in that channel, based on the common sense, we choose the prorate channel to contact with him. Besides, only in this way can we do the analysis in this topic.

# Limitations and Uncertainties

## The first part

1. The feature selection process is not very strict. We just assume that features in the dataset are all relevant to our target variables and do the selection based on the result of data cleaning. Since this project is faced with our business partners and the result of the models are good, we don't go future in the feature selection process. However, it may affect the results of inferences draw from regression model, even though the stepwise selection is used. Besides, if this project is more technical, we must try some better ways for the feature selection.

2. Approach when handling the imbalance of data is relatively simple and it shows some potential problem in the analysis of channel selection. When use the sampling, we simple discard many rows in the original dataset. When we do the resampling, the original distribution of features are destroyed, which leads to a worse prediction result. (As shown in channel selection).

3. Since we need to train the adaboost model again when we knit the report, it will take a very long time. Besides, tuning the parameters of random forest model for a new dataset is even more time consuming. It is not convenient for our business patterners.

### The second part

The main issue is variable selection, according discussion above, the features in model are selected by our common sense and because the result of our modeling is not bad. We just assume that variable selection by common sense is reasonable. But if our modeling result is bad, what should we do to improve our modeling?

```
##                    Overall
## age              0.7583850
## jobblue-collar   2.3338617
## jobentrepreneur 1.4766854
## jobhousemaid     0.2540938
## jobmanagement    0.2386233
## jobretired       1.8957490
```

Using variance importance based on logistic regression using all variable in the dataset. We can choose the variable with bigger variance importance to fit our target model.

### About the data sets

Even though we regard the datasets as they are from the bank XYZ, they are actually independent with each other. Besides, the data set for deposit campaign just has one target variable.

## Areas of Future Investigation

In the future, we hope to get data sets with more information from same organization, it will enable us to combine the results of analysis together. Also, we hope to get data from multiple bank. It will enable us to analysis different banks, which may show some interesting pattern related to the difference of some markets.

Besides, we need to better handling the feature engineering and processing of imbalance data mentioned before. It will enable us to get better inferences and prediction results based on our model.

What’s more, in order to let everyone to use the random forest model, we need to build some application that directly to use the tuned parameters to get the results. Then, they are able to use both random forest and ad boost model.

Finally, we need to use some cloud computing platform, such as AWS, to make the training process of models to be faster.

## Reference

[1] James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R. Springer, 2014.

[2] https://en.wikipedia.org/wiki/Random_forest

[3] https://en.wikipedia.org/wiki/AdaBoost

[4] https://en.wikipedia.org/wiki/Support-vector_machine