**Bank marketing classification**

By Sam Liu

**Abstract**

With the bank marketing campaigns conducted by a Portuguese banking institute, it is very profitable to identify the major factors from all 20 potential independent variables which impact the decision of a client if s/he will make a term deposit. After wrangling and cleaning of the collected data, exploratory data analysis followed by featuring selection is then conducted to choose the variables that are to be used in the final models. At the same time, correlation matrix is also calculated to drop those strongly correlated variables. Finally three models, i.e., logistic regression, random forest and decision tree, are employed and comparison among three models is also studied. The random forest model offers best performance on predicting the purchase rate among all age client groups. It also discovered that using phone survey positively contributes to purchase rate.

1. **Introduction**

Since the incept of mass bank market campaigns, there are overwhelming competitions among marketers, which lead to positive response rate is less than 1%. At the same time, due to the concerns of the privacy the direct marketing sometimes receives negative altitude [1]. Therefore it is important for bank practitioners to analyze the customers data and obtain the important factors which affect the clients decisions. Our goal is to establish a classifier to accurately predict if a customer will make a term deposit. This accuracy will not only improve the campaign efficiency by targeting the import factors the classifier provided, but also more precisely choosing the more potential customers who will make a deposit. All in all, these will reduce cost greatly and achieve more profit.

1. **The bank market dataset**

The data was collected by direct marketing campaigns of a Portuguese bank institution. These campaigns were based on phone calls. This project utilized the dataset which is bank-additional-full.csv with 41188 clients and 20 variables and response variable is yes/no indicating if client will make a deposit or not.

Table 1. Variables of the bank marketing dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number | Variables | Type | Variable Description | Range |
| 1 | Age | Numeric | Age of the client in years | 17--98 |
| 2 | Job | |  | | --- | | Categorical | | Employment status: 'housemaid', 'services', 'admin.', 'blue-collar', 'technician', 'retired', 'management', 'unemployed', 'self-employed', 'unknown', 'entrepreneur', 'student' | NaN |
| 3 | Marital | Categorical | Marital status: 'married', 'single', 'divorced', 'unknown' | NaN |
| 4 | |  | | --- | |  | | Education | | |  | | Categorical | Education status: 'basic.4y', 'high.school', ‘basic.6y', 'basic.9y', 'professional.course', 'unknown', 'university.degree', 'illiterate' | NaN |
| 5 | Default | Categorical | 'no', 'unknown', 'yes' | NaN |
| 6 | Housing | Categorical | 'no', 'unknown', 'yes' | NaN |
| 7 | Loan | Categorical | 'no', 'unknown', 'yes' | NaN |
| 8 | Contact | Categorical | Communication type: 'unknown','telephone','cellular' | NaN |
| 9 | Month | Categorical | 'jan', 'feb', 'mar', ..., 'nov', 'dec' | NaN |
| 10 | Day\_of\_week | Categorical | ‘mon’, ‘tue’,’wed’,’thu’,’fri’ | NaN |
| 11 | Duration | Numeric | last contact duration, in seconds | 0--4918 |
| 12 | Campaign | Numeric | number of contacts performed during this campaign and for this client | 1--56 |
| 13 | Pdays | Numeric | number of days that passed by after the client was last contacted from a previous campaign | 0--999 |
| 14 | Previous | Numeric | number of contacts performed before this campaign and for this client | 0--7 |
| 15 | Poutcome | Categorical | outcome of the previous marketing campaign ( 'failure','nonexistent','success') | NaN |
| 16 | Emp.var.rate | Numeric | employment variation rate | -3.4—1.4 |
| 17 | Cons.price.idx | Numeric | consumer price index | 92.2-94.7 |
| 18 | Cons.conf.idx | Numeric | consumer confidence index | -50.8—26.9 |
| 19 | Euribor3m | Numeric | euribor 3 month rate | 0.634-5.045 |
| 20 | nr.employed | Numeric | number of employees | 4963.6-5228.1 |
| 21 | y | Binary | has the client subscribed a term deposit? 'yes','no’ | NaN |

The dataset was obtained from the University of California at Irvine (UCI) machine learning repository <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

The dataset was collected by S. Moro, P. Cortez and P. Rita [2]. The table 1 lists all variables information including data type (numerical or categorical), description of each variable, name of variable, and range of variable.

It can be seen from Table 1 that the dataset is composed of two different types of data, i.e., numeric and categorical, where categorical includes binary type with two classes (‘yes’ or ‘no’).

The numerical variables (age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m and nr.employed) are either integer or float types; The categorical variables are ones with all types classes (job, marital, education, contact, month, day\_of\_week) and binary ones with two classes of ‘yes’ or ‘no (default, housing, loan, poutcome and y) where y is the response of if the client will make a term deposit. The last column range of table 1 tells the range of numeric variables. For example, the range for age variable is 17-98 meaning all clients age from 17 to 98 years old. Campaign ranges from 1 to 56 contacts during the campaign for clients. Duration ranges from 0 to 4918 seconds since the last contact for clients ( 4918 seconds is less than 1.5 hour, meaning clients took some time to consult friends or family to make decision of deposit). Emp.var.rate represents the employment variation rate ranging from -3.4 to 1.4, where positive rate indicates increased employment due to better profits while negative rate indicates decreased employment due to recession economy.

Fig. 1(a) shows the boxplot of the normalized numerical variables while Fig. 1(b) depicts the histograms for each of these variables. It can be seen that variables emp.var.rate, Cons.price.idx, Cons.conf.idx ,Euribor3m and nr.employed are widely distributed while without outliers or negligible outliers; on the other hand, variables such as Campaign, Pdays and Previous are quite narrowly distributed while having many outliers. The Age variable is the situation between the above two scenarios.

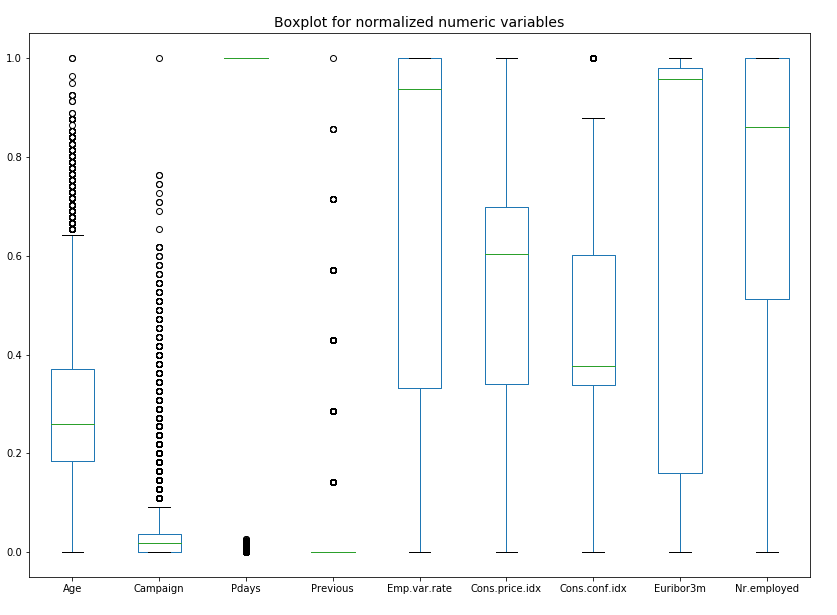


Fig. 1(a) Boxplot of normalized numeric variables

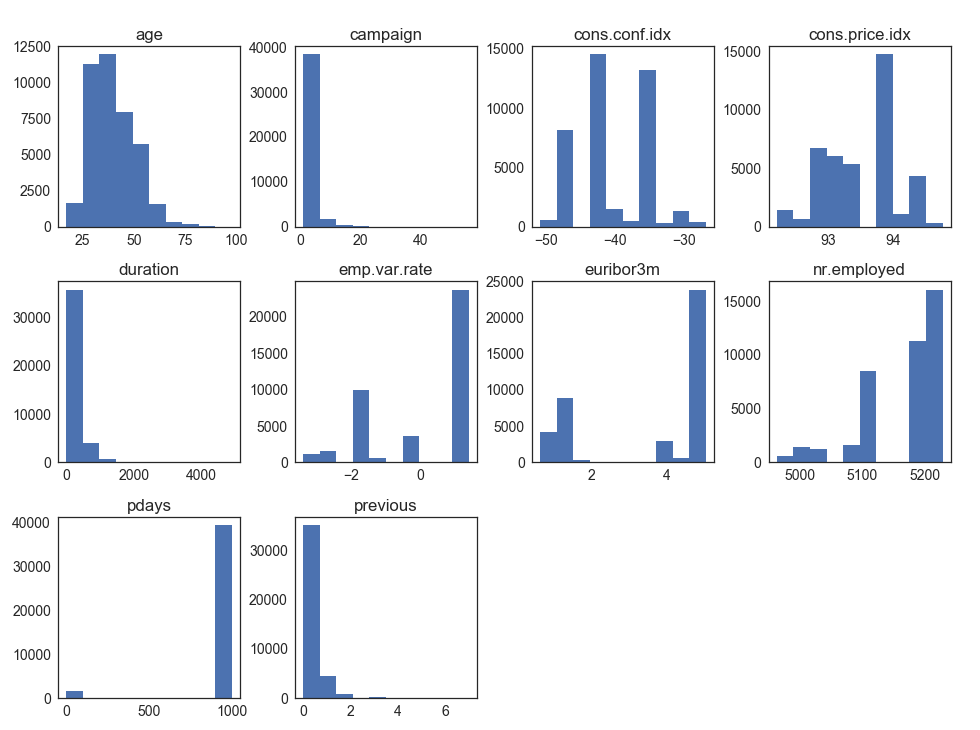


Fig. 1(b) Histograms of numeric variables

1. **Exploratory Data Analysis (EDA)**

**3.1 Missing data treatment (cleaning and wrangling)**

Because of the privacy issues, some clients choose to not disclose their personal information such as marital status, job, default, loan and housing, which produced some data entries with missing values and these are missing data. Missing data may cause misinterpretation of data, therefore they need to be treated, wrangled and prepared to further analysis. For this project, there are two kinds of treatments for the missing data.

The first treatment is removal of missing data. For example, the missing data for variable job, there are about 0.8% data have missing values (labelled as unknown) for job. A 2 x 2 contingency table is constructed to determine the impact of missing values of job on the client deposit.

Table 2. Contingency table for missing data in job

|  |  |  |
| --- | --- | --- |
| Deposit  Job status | Deposit\_No | Deposit\_Yes |
| Job | 36255 | 4603 |
| Job\_missing | 293 | 37 |

Chi-square testing is conducted on the above dataset and p-value of the testing is 0.955 which accept H0 hypothesis, meaning the missing data have no impact on the decision of the client deposits. Therefore for variable job 330 missing entries needs to be removed.

The second treatment is to keep the missing data as a new class. Variable default belongs to this category. Chi-square testing of default status on the clients’ deposits shows p value of 8.42 e-88, which rejects the H0 hypothesis. So the missing value of default is kept as a new class.

* 1. **Dependence of deposit on variables**

Data visualization enables people to communicate information clearly and efficiently [3]. It makes complex data more accessible, understandable and usable. For bank marketing classification, data visualization is employed to discover the relationship between the deposits decision and the variables of interest. It can tell us if a variable has impact on the decision of deposits. For instance, duration is the time in seconds since last contact and it is ranged from 0 to 4918 seconds. Duration is then grouped into 3 classes, i.e., short, medium and long corresponding to (0-102 s), (102-319s) and (319-4918 s). Then the deposits on each class is calculated and plotted as in Fig. 2.

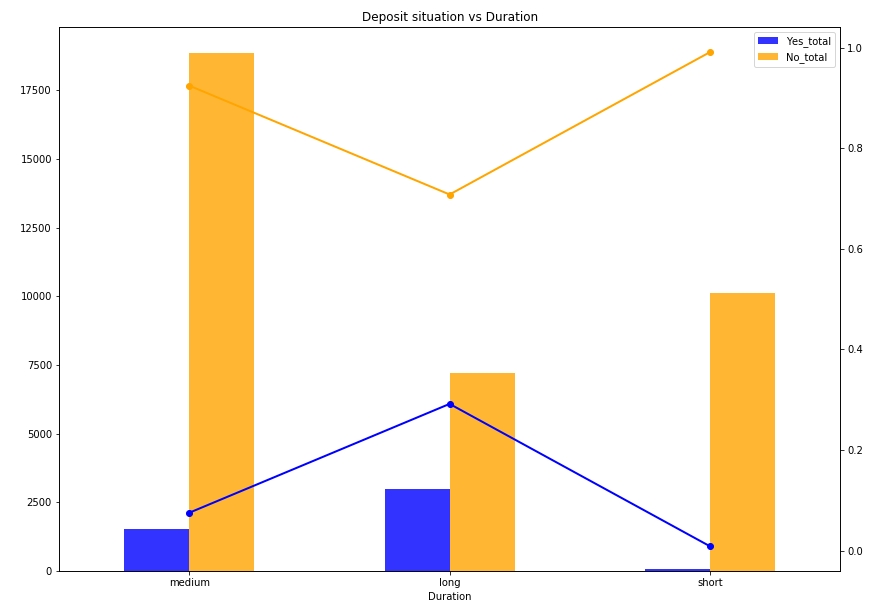


Fig. 2 Deposit situation vs. duration

As shown in Fig. 2, left y axis is the number of clients and right y axis is the fraction for deposit situation. It indicates that most clients took medium duration (102-319s) to make decision of term deposits. Meanwhile, it tells us with longer the duration, more fraction clients made deposits. All variables of interest are studied and respective impact on the deposit situation are plotted as well.

* 1. **Correlation matrix and heatmap**

The correlation between numeric variables of interest is calculated and plotted as in the heatmap in Fig. 3.

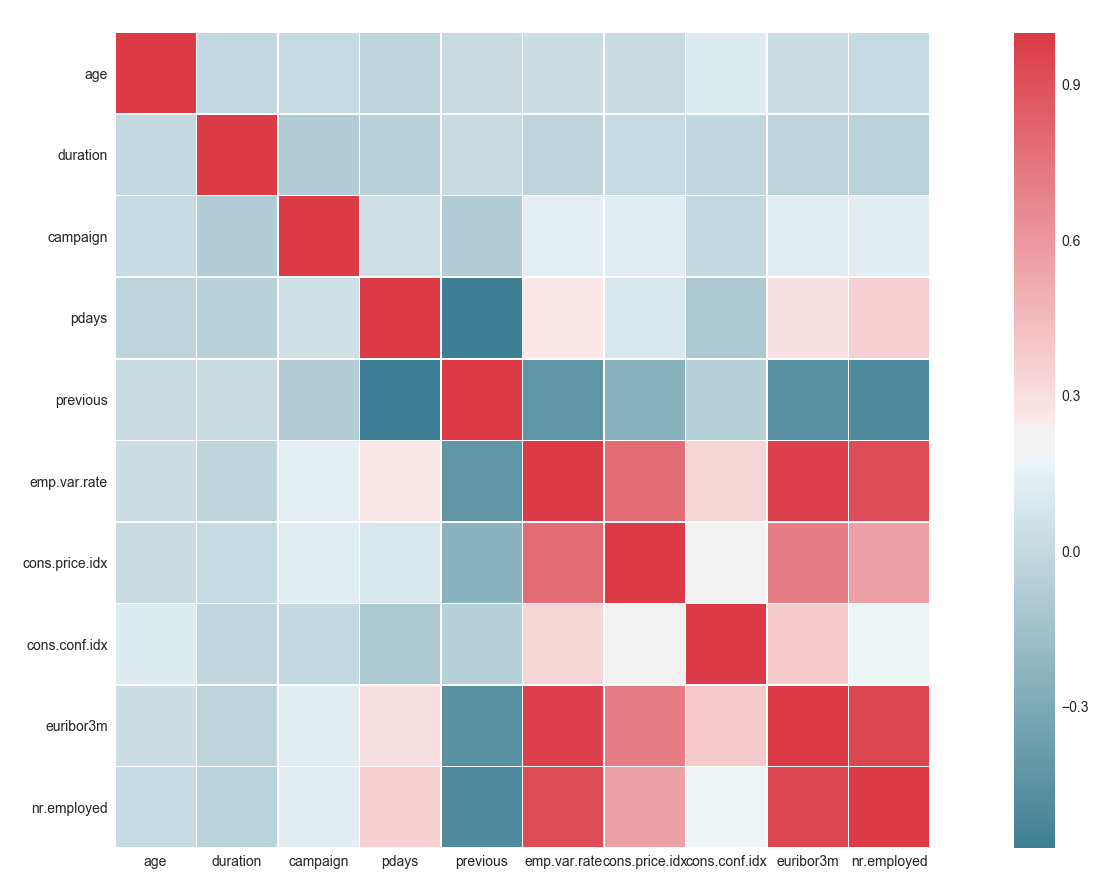


Fig. 3 Heatmap of numeric variables

As shown in Fig. 3, we found that There are three variables strongly correlated to one another, i.e., euribor3m, number of employed and employment variation rate. Cor(euribor3m, number of employed) = 0.949215; Cor(euribor3m, employment variation rate) = 0.975380 and Cor(employment variation rate, number of employed) = 0.914747 . So only one variable is to stay in final model, i.e., employment variation rate.

1. **Feature selection and split of dataset**

In order for categorical variables to be included in machine learning models, it is practical to convert categorical variables into binary integers (0 or 1) vectors using one hot encoding technique. For instance, job variable is converted into dummy variables such as 'job\_blue-collar', 'job\_entrepreneur', 'job\_retired',…, 'job\_services' , if one client’s job is blue collar, then 'job\_blue-collar' is set to 1 while all other dummy job variables are set to zeros.

Following the one-hot encoding, Recursive Feature Elimination(RFE) is employed to select best potential features. RFE is based on the idea to repeatedly construct a model (for example an SVM or a regression model) and choose either the best or worst performing feature (for example based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated [4]. There are 11 variables selected by RFE and they are applied into three classification methods: Logistic Regression, Random Forest and Decision Tree. The data with selected features is then split into two parts, i.e., train and test datasets with ratio of 0.3 (test/train).

1. **Methodologies**
   1. **Logistic Regression**

Logistic Regression is a statistical model to predict a binary response from a bunch of independent variables. The response values are often labeled as integers 1 or 0, which stands for situation with only two outcomes such as success/fail, sick/healthy, win/lose etc.. For some more subtle situation where there are multiple outcomes multinomial Logistic Regression model is used to tackle the issue. Logistic Regression model is easy to implement and a benchmark upon which more complex models are to be built. What’s more it also provides the easy summary of the performance of the model containing *p*-values for independent variables, coefficients and corresponding standard deviations.



Table 3. Summary of Binary Logistic Regression Analysis

As shown in Table 3, all 11 variables of the Logistic Regression model with all p-values less than 0.05, indicates all these variables are significant for the model. All coefficients and respective standard error are in reasonable range. Meanwhile the importance for each variable in the model is calculated and plotted as in Fig. 4. As we may see that the euribor3m(euribor 3 month rate) is the most importance predictor and it contributes negatively to the decision of client on the purchase of term deposit. The cell phone and land phone are next two important predictors and they follow the trend that more cell and land phone contacts used in surveys, more clients tend to purchase the term deposit.

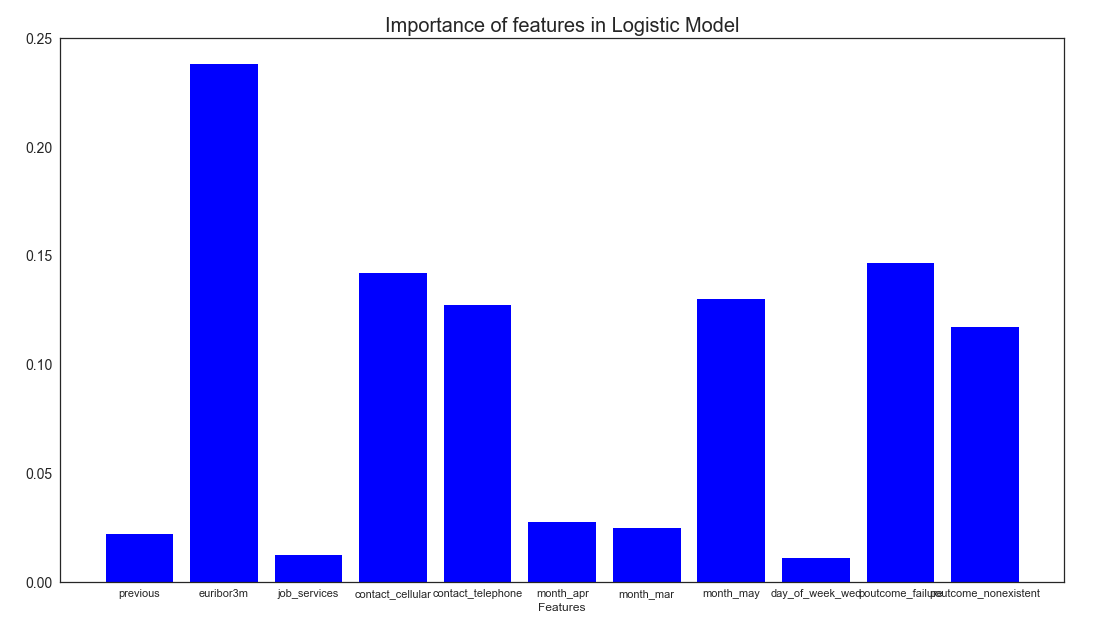


Fig. 4 Importance ranking for 11 variables

* + 1. **Metrics of the Logistic Regression model**

The accuracy of train dataset and of test dataset are reported to be 0.900 and 0.899 respectively. Since the dataset is unbalanced one due to major clients’ turndown the purchase, the logistic regression model tends to fit the data set for major group, thereby the accuracy of whole dataset is not capable of telling the accuracy of minor group-those clients (about 11% of all clients) who purchased deposit.

The Receiving Operating Characteristic (ROC) curve is an appropriate way to offer better description of the performance of the model as shown in Fig. 5. The model also provides the confusion matrix as Table 4.

Table 4. Confusion Matrix of Logistic Regression

|  |  |  |
| --- | --- | --- |
| Prediction  True value | No | Yes |
| No | 10763 | 124 |
| Yes | 1112 | 230 |

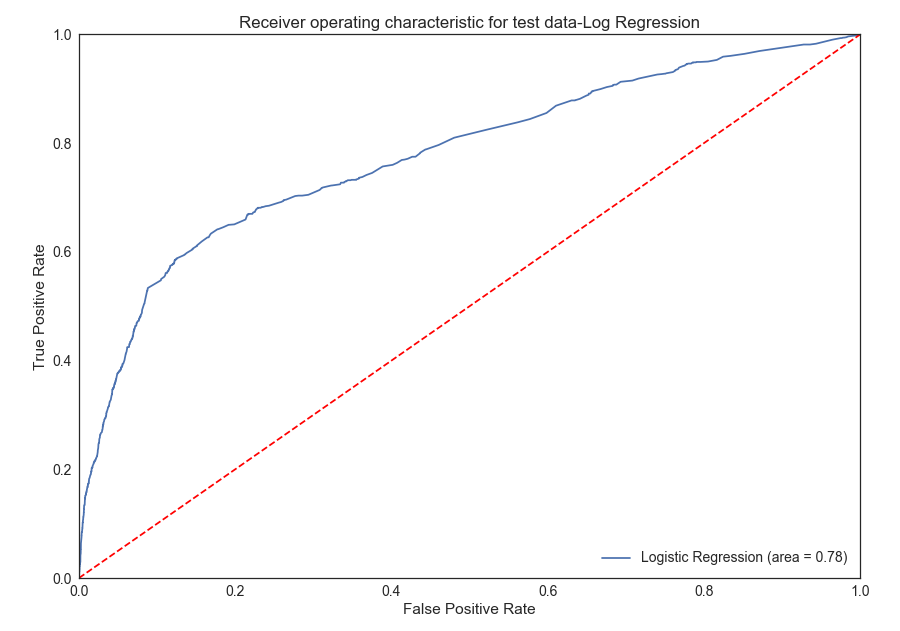


Fig. 5 ROC curve for Logistic Regression –test data

From table 4, we calculate the sensitivity for the model is and the specificity of the model is

* + 1. **Comparison of actual and predicted purchase by logistic regression**

In order to visually compare the actual purchase rate of clients and the predicted purchase rate from the model among age groups, we plot the result shown as in Fig. 6.

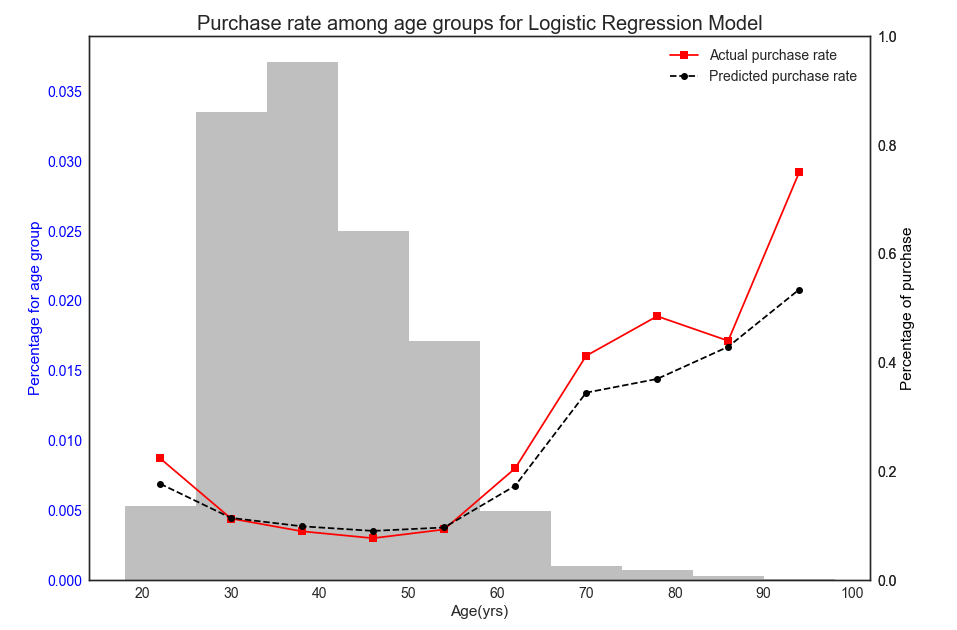


Fig. 6 Comparison between actual and predicted purchase rate among age groups

As shown in Fig. 6, most of the clients belong to age from 28 to 50, the model predicted very close to the actual purchase rate for these age groups. Beyond these groups, the model deviates a bit but those age groups only have small fraction of all groups combined. Thus bankers may focus on these age group and make appropriate measures such as using phone as contact means to promote the campaigns.

* 1. **Random Forest**

The Random Forest is tuned by the RandomizedSearchCV and the optimal hyperparameters are selected for train and test datasets.

The tuned model offers us the importance of selected features as in Fig. 7. It indicates that the euribor3m is the by far most important feature as it takes more than 65% weight of all selected features.

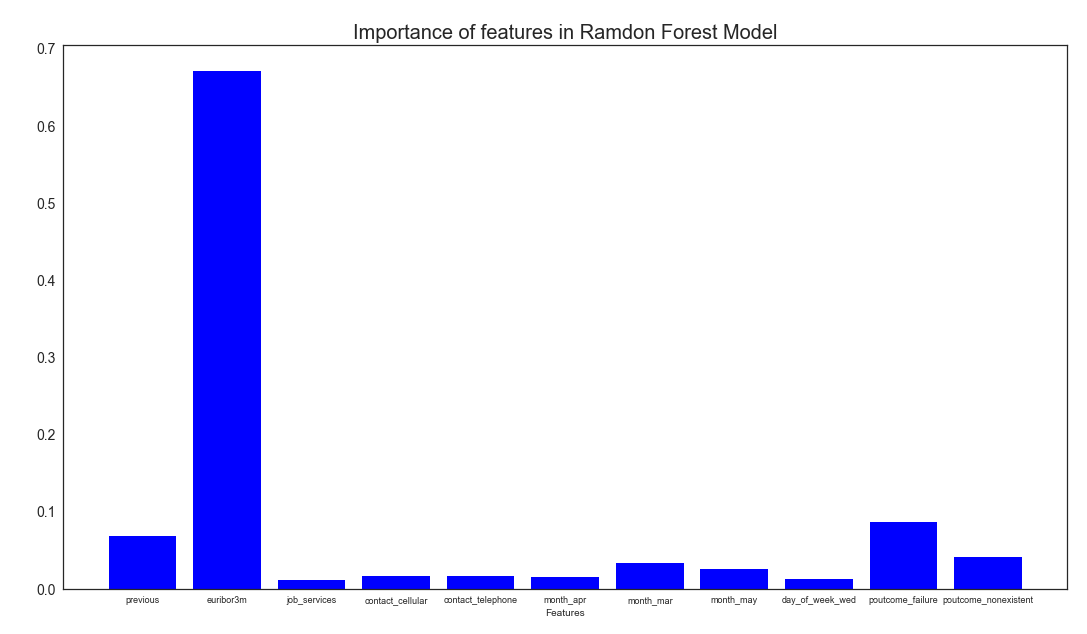


Fig. 7 Importance ranking for 11 variables for Random Forest model

* + 1. **Metrics of Random Forest model.**

The accuracy for train and test datasets is 0.915 and 0.896 respectively for the Random Forest model.

ROC curve for test dataset is shown in Fig. 8 in which the AUC is 0.76 a little less than 0.78 for Logistic Regression.

From table 5, we know sensitivity for the model is and the specificity of the model is

Table 5. Confusion Matrix of Random Forest

|  |  |  |
| --- | --- | --- |
| Prediction  True value | No | Yes |
| No | 10609 | 278 |
| Yes | 991 | 351 |

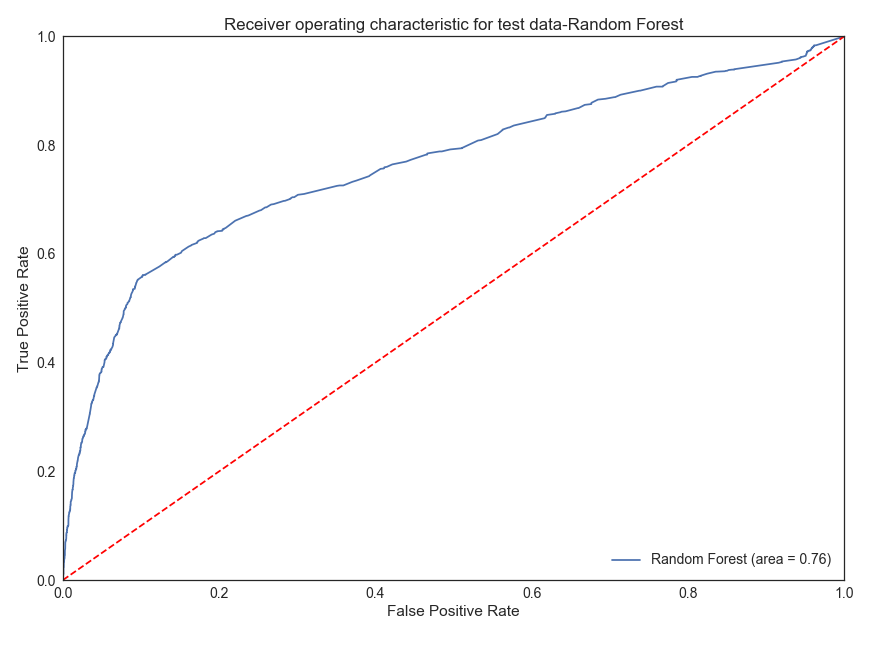


Fig. 8 ROC curve for Random Forest –test data

* + 1. **Comparison of actual and predicted purchase by Random Forest**

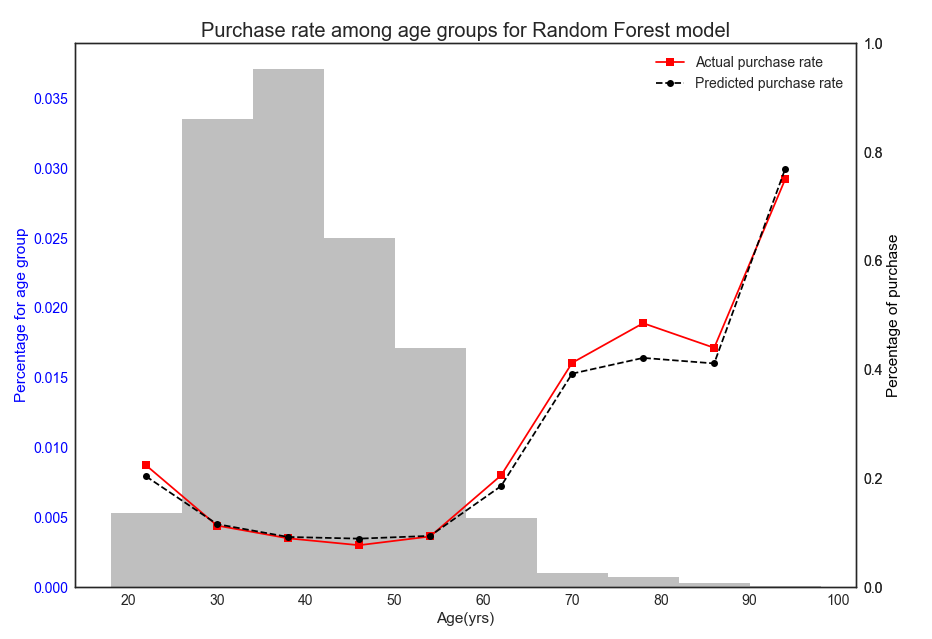
****

Fig. 9 Comparison between actual and predicted purchase rate among age groups

Random Forest model provides a little better performance in terms of the closeness match between actual and predicted purchase rate especially for age group from 28 to 50 years old. It even provides better performance in other age groups as well.

1. **Recommendations**

Both logistic regression and Random Forest model provide good accuracy for both train and test datasets. It also can be seen that in general case, the Random Forest gives better performance to predict purchase rates among all age groups when compared to the Logistic Regression model.

To banking business practitioners, from logistic regression model, we know the contact using cell and land phone are important to positively contribute the final decision of deposit. Therefore for bankers it may profitable to hire more people to conduct phone surveys.

**References:**

[1]. https://en.wikipedia.org/wiki/Direct\_marketing

[2]. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[3]. <https://en.wikipedia.org/wiki/Data_visualization>

[4]. <https://blog.datadive.net/selecting-good-features-part-iv-stability-selection-rfe-and-everything-side-by-side/>