# BANKING PRODUCT PURCHASE PREDICTION

DEC520-101-Team06-
Aditya Agrawal (ama131), Parth Gairola (pg155), Ruth George (rag83), Mihir Joglekar (mj248) ,
Kelly Zhang(xz324), Lei Zheng(lz219)

## Business Understanding

The Banking Sector is increasingly more competitive in today's world as research shows that 45% of dissatisfied customers will discourage their friends from using that bank's service. In the process of achieving competitiveness, it is insufficient for banks to only focus on customer service, instead, being able to predict purchase decisions and customize products will optimize the chances of successfully attracting new customers while retaining old ones. In this study, we are attempting to identify the relevancy and combinations of customers' characteristics and predict if they will subscribe to a term deposit. This will help the banking institutions in classifying customers and have a more targeted marketing approach to sell their products, increasing market campaign efficiency and improving market resource utilization.

## Data Understanding

This data set is published by the University of California Irvine Machine Learning Repository. This Bank Marketing Data Set "is related with direct marketing campaigns (phone calls) of a Portuguese banking institution". It contains 17 attributes and 45211 instances recorded from May 2008 to November 2010. (Variable definitions can be found in the appendix)

The target variable is a binary variable indicating whether the client subscribed to the term deposit (yes = 1, no = 0). Here the term deposit is the product being sold and relevant to the marketing strategy. The rest of the variables include useful demographic information about customers such as their age, their current job, their marital status and their education background. Other information surrounding the customers'

previous banking decisions are also available and likely to be useful. These variables and attributes are numerical, categorical, or binary.

**Data Preparation**

To clean up the data, we first dropped the 'ID' column. We also removed the "duration" variable as this information is not accessible before any sales call, meaning that it won't be useful in our prediction and would lead to insignificant results as it would not contribute towards formulating an effective marketing strategy. Then we created dummy variables for the categorical data. We converted the following variables into dummy variables for the purpose of classification: jobs, marital status, default, housing, loan, month, contact, pdays, poutcome. In addition, In order to capture seasonality, months are converted into 4 seasons, January ~ March is season 1, April ~ June is season 2, July ~ September is season 3, and October ~ December is season 4.

The dataset is highly unbalanced with 88.7% of *no* outcomes and only 11.3% of *yes* outcomes. To balance it, we tried three different approaches: under-balancing, over-balancing, and the ROSE function from the ROSE library – a bootstrap technique that helps with the classification of highly unbalanced datasets. The three approaches were tested using AUC. The best AUC performance was achieved under the ROSE approach. This is the data used for modeling and future analysis.

**Exploratory Data Analysis**

We started the EDA by looking at a correlation matrix (Figure 12) of all the fields and discovered the following correlations in the dataset:

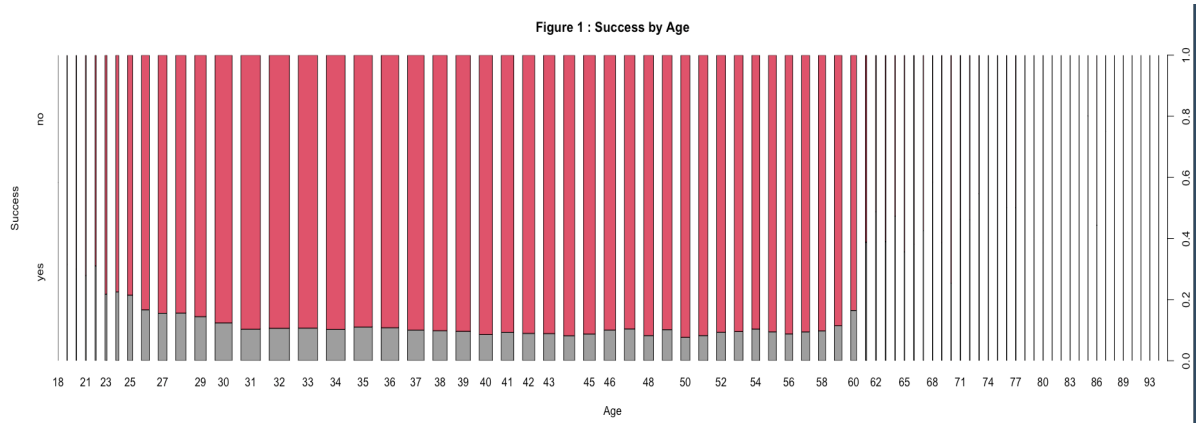
1. There is a positive correlation between *pdays* and *poutcome_success*, meaning that people who were contacted less recently are more likely to purchase a product. This could indicate a threshold towards customer engagement. This information could be very useful for marketing campaigns as often institutions 'over-market' and end up not-converting the potential customer.

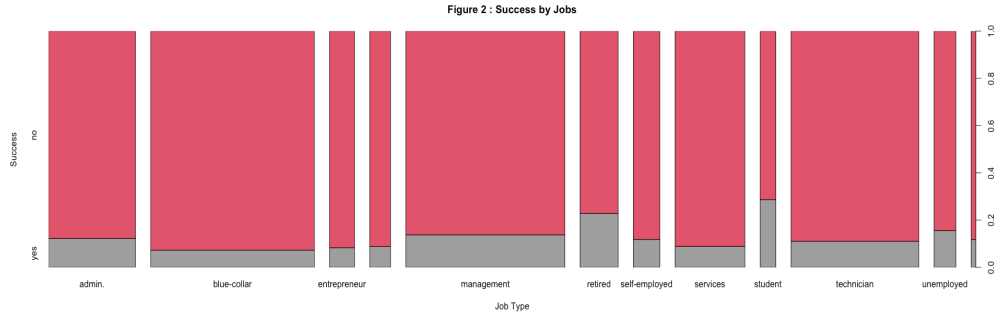
2. *Education* has a positive correlation with the *job type* people have. More education leads to a higher chance of management positions, and less education leads to blue-collar jobs.

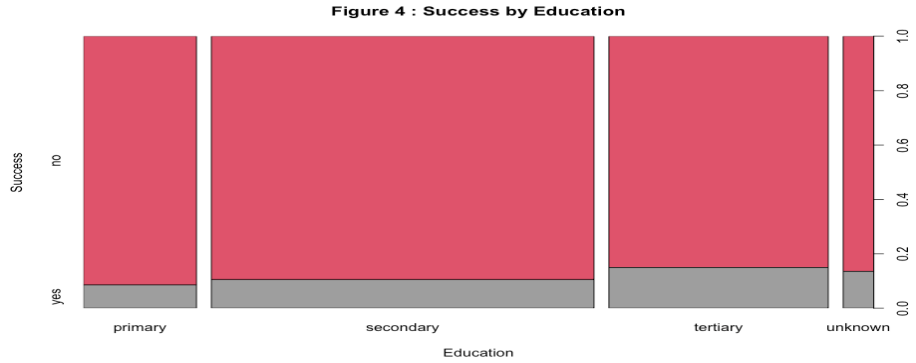
We then looked at the influence of the specific critical fields with our target variable. We categorized these fields into three domains.

1. Demographics of potential customers
2. Financials of potential customers
3. Specifics of the Marketing campaign

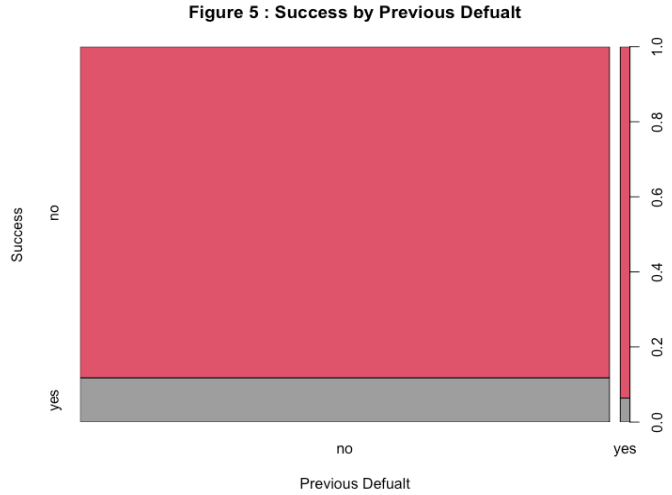
Based on demographics, we found that the campaign was more successful among certain age groups – in young adults and senior citizens (Figure 1). These findings were corroborated when we explored a relationship with jobs. Customers in administrative and management positions, students, and retirees are more likely to buy new products (Figure 2). We believe that this is because young people are more susceptible to changes. This could point towards ethical concerns as both, young adults, and senior citizens are more likely to be influenced into taking less-informed financial decisions. The success rate is highest among single and marginally higher in divorced than married couples based on marital status (Figure 3). Also, the success rate increases as the level of education received goes up (Figure 4), which could be due to the fact that people with higher education tend to earn a higher income, therefore more open to product purchasing.

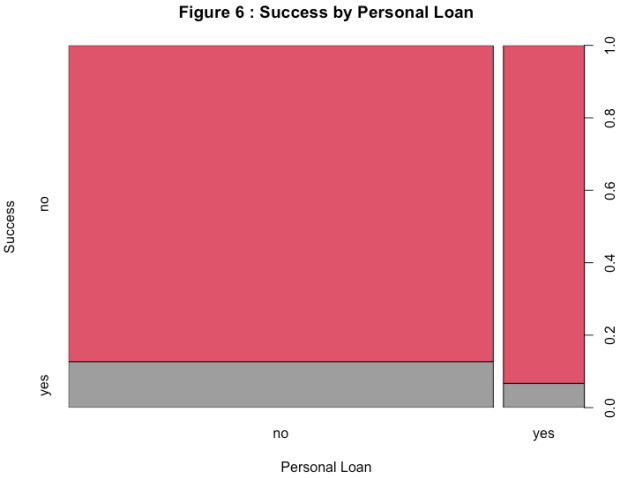




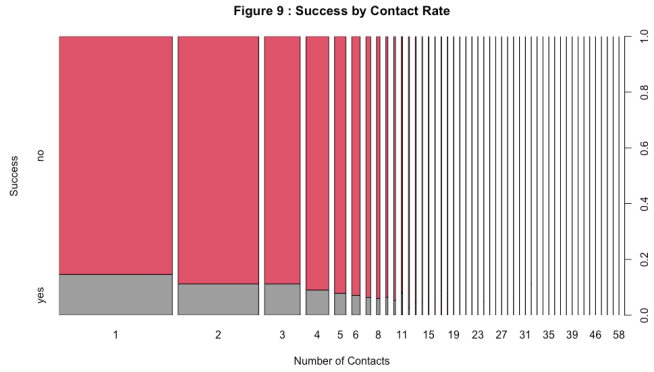



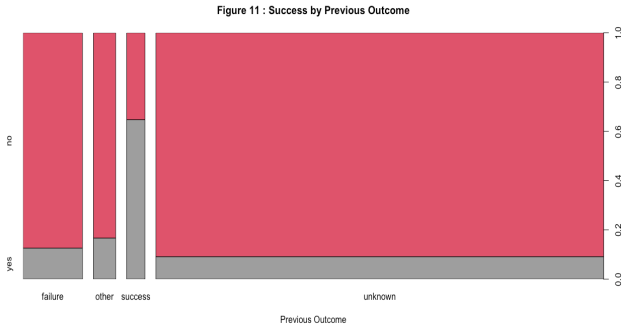

Other information like previous loans, housing, and defaults helped us understand customers' financial situation and how that impacts their decision to buy another product offered by the bank. The absence of previous defaults substantially increased the probability of success (Figure 5). Similarly, people without personal loans (Figure 6) and housing loans (Figure 7) are more likely to buy the product. Not having loans could indicate that they are more financially stable, meaning that they are more likely to purchase other financial products.



The specifics of the marketing campaign also had implications for the success rate. For example, the distribution of success rate differs across months. It appears in the summer season, from May through August, there are the most customers. March, September & October have a higher success rate compared to other months (Figure 8). If we explore the success rate based on the number of contacts, we see a declining rate as contacts increase (Figure 9). Also, when contacted via cellular than any other means is more likely to have a positive outcome (Figure 10), which is reasonable as most people own a cellphone. The success of previous campaigns with a customer correlates with the outcome of this campaign. Customers who bought products in an earlier campaign are more likely to buy a new product in this campaign (Figure 11).




**Modeling**

This is a classification problem and thus we used the following supervised modelling techniques

- Logistic regression with GLM
- Random Forests
- Gradient Boosting Machine (GBM)

We recognize that each modelling technique has its advantages and disadvantages, so we provide analysis and interpretation about each technique we have used. Logistic regression model is easier to implement and train, but it has the assumption of linearity between the dependent variable and independent variable; Random Forests can analyze linear and non-linear relationships well, but it is

unable to provide complete visibility into the coefficients; Gradient Boosting Machine often require many trees, but it often provides better predictive accuracy. Due to the varied nature of all techniques, it provides the institution with different options to select from based on their specific requirements.

The data set was divided into training and test using the 75-25 approach in which 75% of data points are used for training the model and 25% are used for testing its efficacy. Every technique was accompanied with 10-fold Cross validation so as to increase the out-of-sample accuracy and make the most of the data we started with, whilst helping in parameter tuning for complex models such as number of trees in Gradient Boosting classifier and the number of variables randomly sampled as candidates at each split in Random Forest.

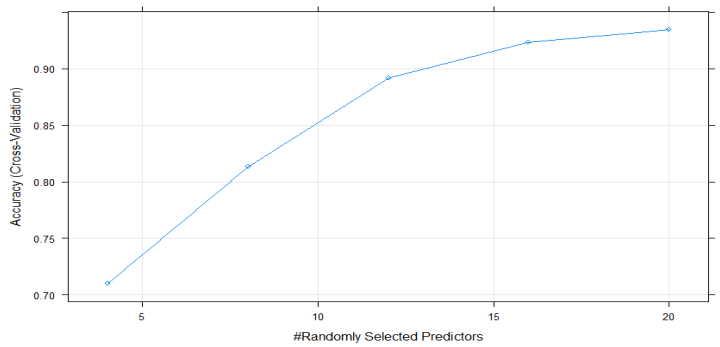
*Logistic regression with GLM* - The regression was run with all the variables and we used 10 fold cross validation. The model was concurrent with our findings in exploratory data analysis:

- Students and retired people are more likely to purchase
- Likelihood to purchase increases with increase in education level
- People with housing loans are more likely to purchase
- People who were part of successful campaigns previously are more likely to purchase

The model yielded the following figures (Figure 16):

- Accuracy- 68.5%; Sensitivity- 0.71; Specificity- 0.66

*Random Forests* - This supervised learning technique builds an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.




The model was run with different values of mtry. Mtry is the number of variables randomly sampled as candidates at each split. We started with 4 as our initial mtry and kept increasing the value which led to constant increase in accuracy. The evaluation parameters plateaued at 20 and thus we kept this value in our final model (Figure 19). The optimum level was found at mtry = 20 which gave the following results.

- Accuracy- 93.5%; Sensitivity- 0.91; Specificity- 0.96

*Gradient Boosting Model (GBM)*- The GBM package implements the generalized boosted modeling framework. Boosting is the process of iteratively adding basis functions in a greedy fashion so that each additional basis function further reduces the selected loss function. We set the value of n.trees as 100, as this indicates the number of tree iterations this model runs. The interaction.depth indicates the maximum depth of each tree and this was set to 4. Shrinkage or learning rate specifies the rate at which the model learns patterns in the data was set to 0.1. The model with the above parameters gave the following results

- Accuracy- 72%; Sensitivity- 0.801; Specificity- 0.63

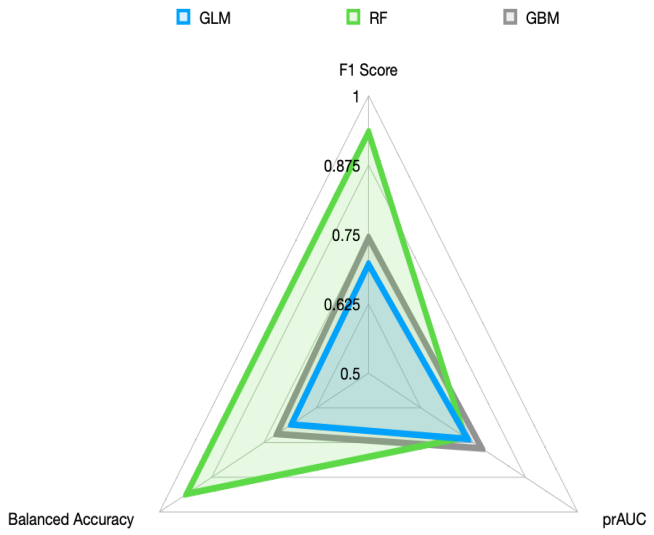
**Evaluation**

The following table shows the confusion matrices for all 3 models:

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Logistic regression with GLM | 0.6856 | 0.7198 | 0.6505 |
| Random Forest | 0.9392 | 0.9168 | 0.9622 |
| Gradient Boost Machine | 0.72 | 0.8018 | 0.6362 |

The Random Forest model performed the best out of all three. In this particular business case our target was to achieve maximum specificity to reduce the false positive rates so that the banking institution does not end up targeting the wrong customers and fail to optimize the marketing budget. Though the above metrics tell a lot about the model and its effectiveness, we must consider other metrics such as the F1 score, the prAUC value and the Balanced Accuracy.

F1 Score is the harmonic mean of Precision and Recall, and prAUC tells us about the ability of the model to classify between different classes, especially for unbalanced data. Balanced Accuracy measures the ratio of correctness balancing for binary classes.




```
Confusion Matrix and Statistics

          Reference
Prediction   X0   X1
        X0 5245  211
        X1  476 5371

               Accuracy : 0.9392
                 95% CI : (0.9347, 0.9436)
    No Information Rate : 0.5061
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8785

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9168
            Specificity : 0.9622
         Pos Pred Value : 0.9613
         Neg Pred Value : 0.9186
             Prevalence : 0.5061
         Detection Rate : 0.4640
   Detection Prevalence : 0.4827
      Balanced Accuracy : 0.9395

       'Positive' Class : X0
```

The figure above, reiterates the above conclusions about the effectiveness of the Random Forest model over the Logistic model and the Gradient Boost classifier. Though the Random Forest is better than the Gradient Boost Model in terms of Balanced Accuracy and F1 score, it does not match up in terms of prAUC. This points us towards diving deeper into modeling this data in order to further identify the optimum model. A possible option would be to try the XGBoost model, which improves on the GBM in terms of accuracy, and could possibly challenge the Random Forest model.

```
Random Forest

33908 samples
   29 predictor
    2 classes: 'X0', 'X1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 30517, 30516, 30518, 30518, 30518, 30517, ...
Resampling results across tuning parameters:

  mtry  logLoss    AUC        prAUC      Accuracy   Kappa      F1         Sensitivity  Specificity  Pos_Pred_Value
   4    1.0796034  0.7793192  0.6555218  0.7098619  0.4184415  0.7350472  0.7950935    0.6224929    0.6835941
   8    0.5157396  0.8945502  0.7479564  0.8132890  0.6261600  0.8223811  0.8538222    0.7717384    0.7933200
  12    0.3561105  0.9541986  0.7882855  0.8917660  0.7835445  0.8921799  0.8847016    0.8990081    0.8998413
  16    0.2822716  0.9735889  0.7647043  0.9235284  0.8471125  0.9228683  0.9037528    0.9438005    0.9428811
  20    0.2591234  0.9796450  0.7305349  0.9347055  0.8694673  0.9339661  0.9123169    0.9576563    0.9567129
  Neg_Pred_Value  Precision  Recall     Detection_Rate  Balanced_Accuracy
  0.7478509       0.6835941  0.7950935  0.4024708       0.7087932
  0.8375076       0.7933200  0.8538222  0.4321991       0.8127803
  0.8838591       0.8998413  0.8847016  0.4478297       0.8918548
  0.9054019       0.9428811  0.9037528  0.4574734       0.9237766
  0.9142405       0.9567129  0.9123169  0.4618085       0.9349866

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 20.
```

**Deployment**

We apply various data mining approaches to find a model that can explain and predict the success of a marketing campaign on the bank clients. The analysis of our modeling indicates that the random forest has better predictive performance than the other two models(logistic model and gradient boosting machine model). The high TPR (true-positive rate) means that clients who will subscribe the deposit are classified as the subscribers correctly by the random forest model. The high specificity also prevents the bank from targeting wrong customers.

The most important attributes as indicated by the Random Forest Model are - *Balance* and *Age.* Interpretation of how these attributes affect the rate of signing up for the term deposit is, however, difficult solely using the random forest model. This aligns with our original exploratory data analysis and those interpretations continue to be important towards understanding customer behaviour.

The bank could collect further data in order to learn more about its customers, and use the existing model as a framework to build upon. This will help predict which client the bank should launch the marketing campaign on and target these identified clients in order to increase the campaign efficiency. The model can be trained and improved on different data sets based on the geographical location of customers and the type of target product. The current model is very specific to a particular banking product in a specific

location but it provides a well structured skeleton which can be used for deployment across the banking industry.

Any shortcomings in the complexity of the model can be overcome with effective business analysts who can communicate the model clearly to stakeholders and by using the domain expertise of branch managers to implement and test for other variables that may have a significant impact in predicting customer behaviour. Random Forest model could further be improved by fine-tuning its parameters further or by trying the XGBoost model.

**Potential Issues**

There exists a few concerns with the application of this dataset. Firstly, since this data is collected specifically for a Portuguese banking institution, it may be difficult to apply to a US bank. Customers behave differently in different cultural backgrounds and social situations. Although this model may not be directly applied, we could use new data collected in the US to train this model. Secondly, surveys are conducted through phone calls, there potentially exists sampling bias and non-response bias. Particular portions of the population have a higher likelihood to answer their phone and are willing to answer the survey. Since multiple phone calls are required to complete the survey, there is a risk that some participants did not complete the entire survey, and those who did may fall into a specific character type.

One important ethical issue to consider for this modelling exercise is the use of customers personal data. The anonymity of the data has to be maintained throughout and it must be made sure that the data is legally available for the organisation to be used and deployed. The data is collected through marketing campaigns and previous purchases. The terms and conditions of these should clearly specify the context in which data will be used.