INST737 – Spring 2016

Digging into Data

Project Report

Data Mining on Bank Data for Predictive Modeling

Submitted by:

Prannoy Banerjee, Neha Chanchlani, Hitesh Gupta

1 Table of content

2 Abstract

In predictive modeling, we use predictive analysis to create a statistical model of future behavior. It consists of a number of predictors, which are variable factors that are likely to influence future behavior or results. The main goal of this type of analysis is to basically go beyond the basic descriptive analysis to report on what has actually happened providing an excellent assessment on what will happen in the future. With the help of this, we can streamline decision-making and can provide new insights that can help us to take better actions. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. We are basically predicting whether a customer will subscribe to "Term Deposit" or not. Along with that we are also trying to predict the amount that bank will suggest the customer to do term deposit.

3 Introduction

Content-Based Systems

Also known as recommender systems, they are used to develop a model based on a user's past behavior. This model is used to predict items that the user may have an interest in. It makes use of content based filtering which utilizes a series of discrete characteristics of an item in order to recommend additional items with similar properties.

Collaborative Filtering Systems

Collaborative filtering (CF) is a technique used by some Content Based systems. Collaborative filtering has two senses, a narrow one and a more general one. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets.

Hybrid

The hybrid approaches combine content-based and collaborative filtering to build a much more robust prediction model. Incorporating both the methods creates the potential for a more accurate prediction

4 Data preparation

The data used for this study is Portugal based Telemarketing data, which is available at - https://archive.ics.uci.edu/ml/datasets/ - the website of UCI Machine Learning Repository. All the data used is in the public domain and does not require any additional license.

The actual dataset can be obtained from the website -https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#
Input Variables:

Age: (Numeric)

Job: type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

Marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

Education (categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course', 'university.degree', 'unknown')

Default: has credit in default? (Categorical: 'no','yes','unknown')

Housing: has housing loan? (Categorical: 'no','yes','unknown')

Loan: has personal loan? (Categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

Contact: contact communication type (categorical: 'cellular','telephone')

Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

Day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')and many more

Our output variables are

Y - has the client subscribed a term deposit? (binary: 'yes','no')

Term Deposit Amount – the amount of term deposit

5 Exploratory Data Analysis

5.1 Distribution of "Term Deposit Accounts" over different Age groups:

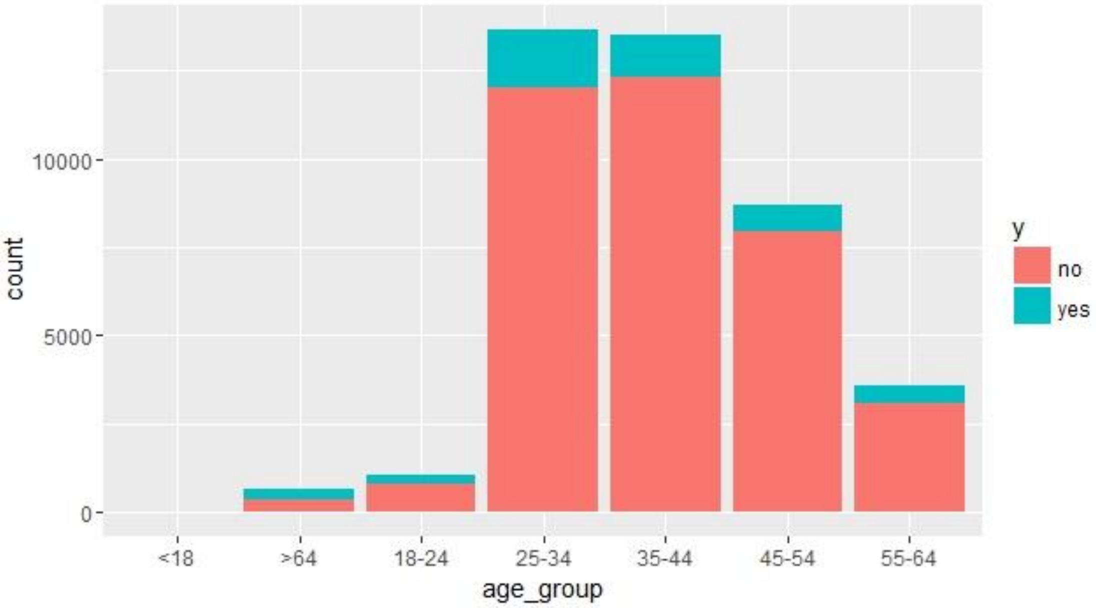


Fig 5.1: Distribution of "Term Deposit Accounts" over different Age groups

5.2 Distribution of “Term Deposit Accounts” over different classes of Education:

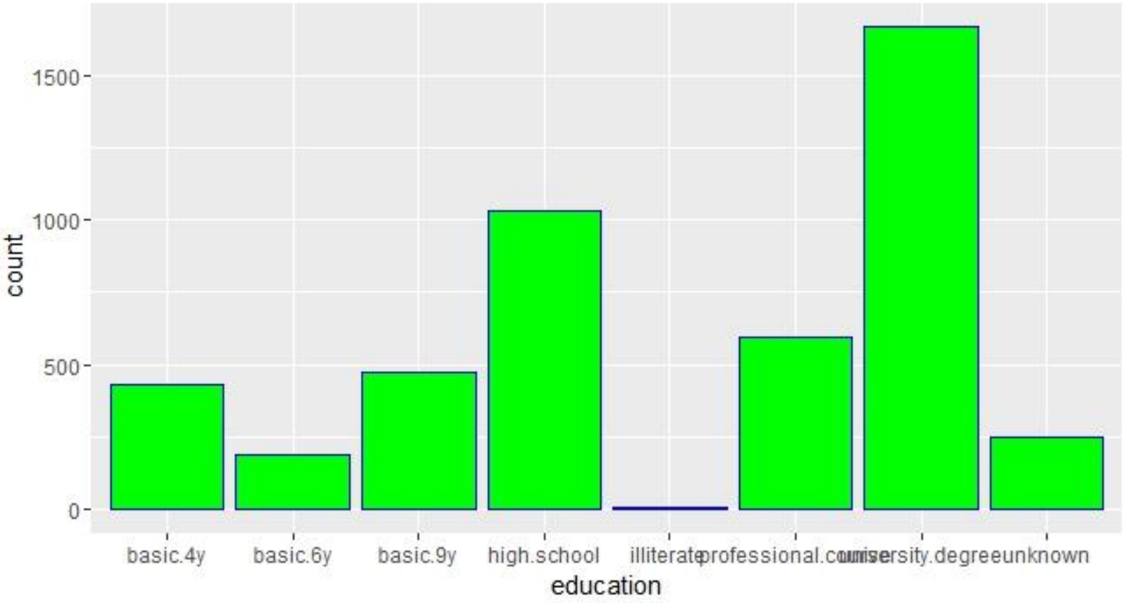


Fig 5.2: Distribution of “Term Deposit Accounts” over different classes of Education

5.3 Distribution of “Term Deposit Accounts” over different job categories:

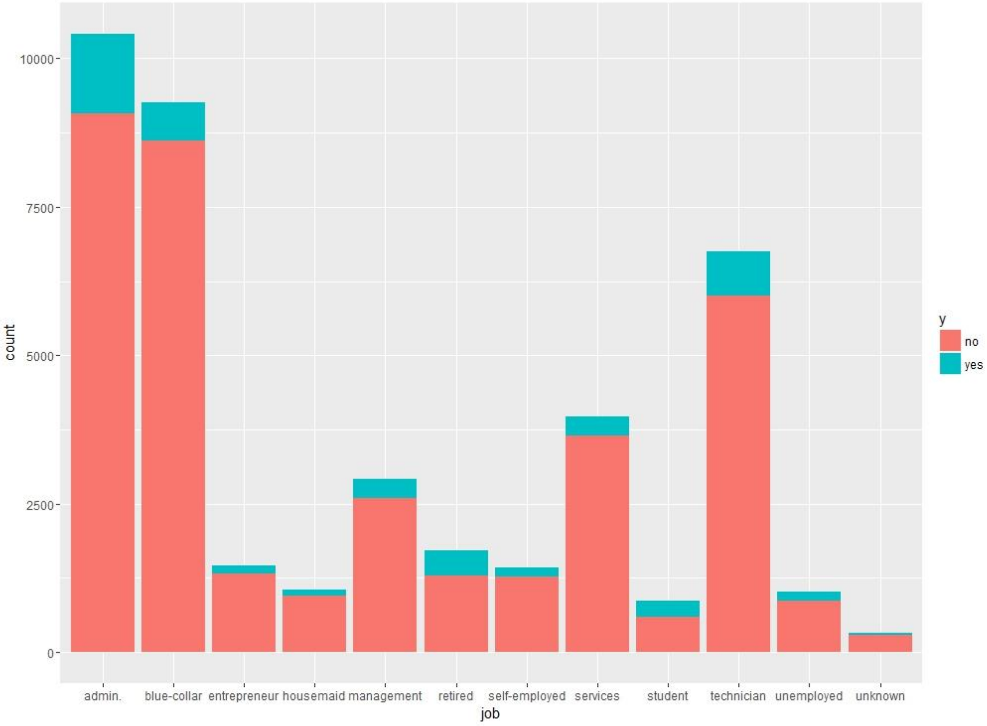


Fig 5.3: Distribution of “Term Deposit Accounts” over different job categories

5.4 Distribution of “Term Deposit Accounts” over different job categories (Marital Status Biased)

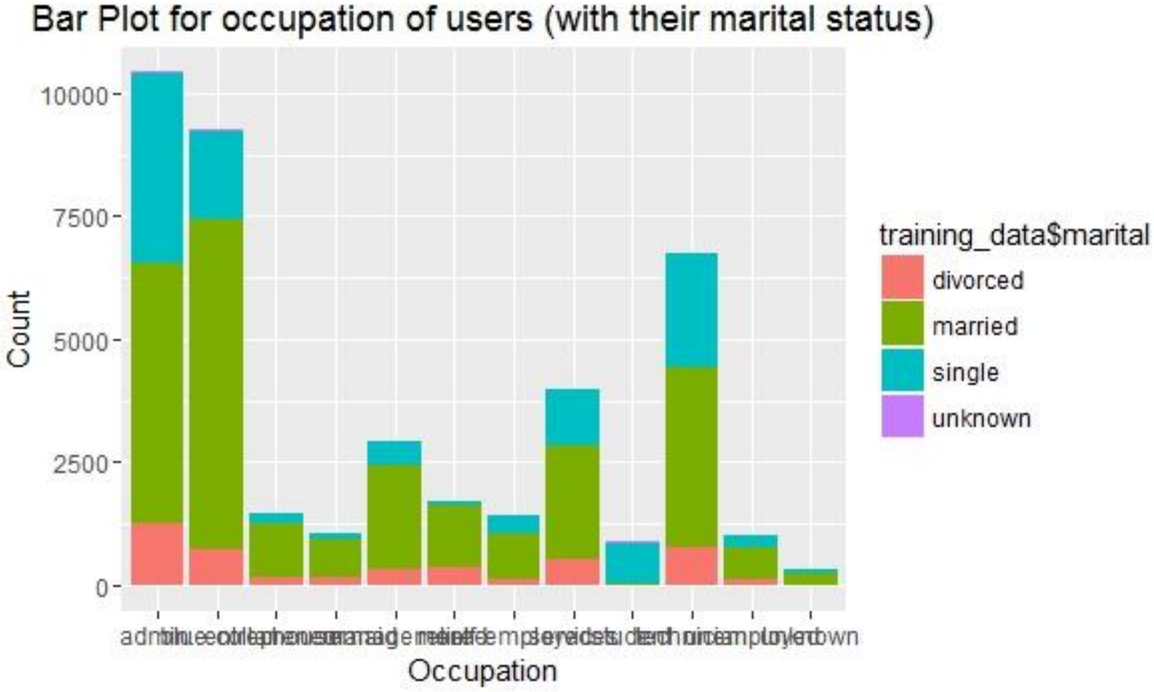


Fig 5.4: Distribution of “Term Deposit Accounts” over different job categories (Marital Status Biased)

5.5 Distribution of the success of the “Term Deposit Account” campaigns over different days of the week

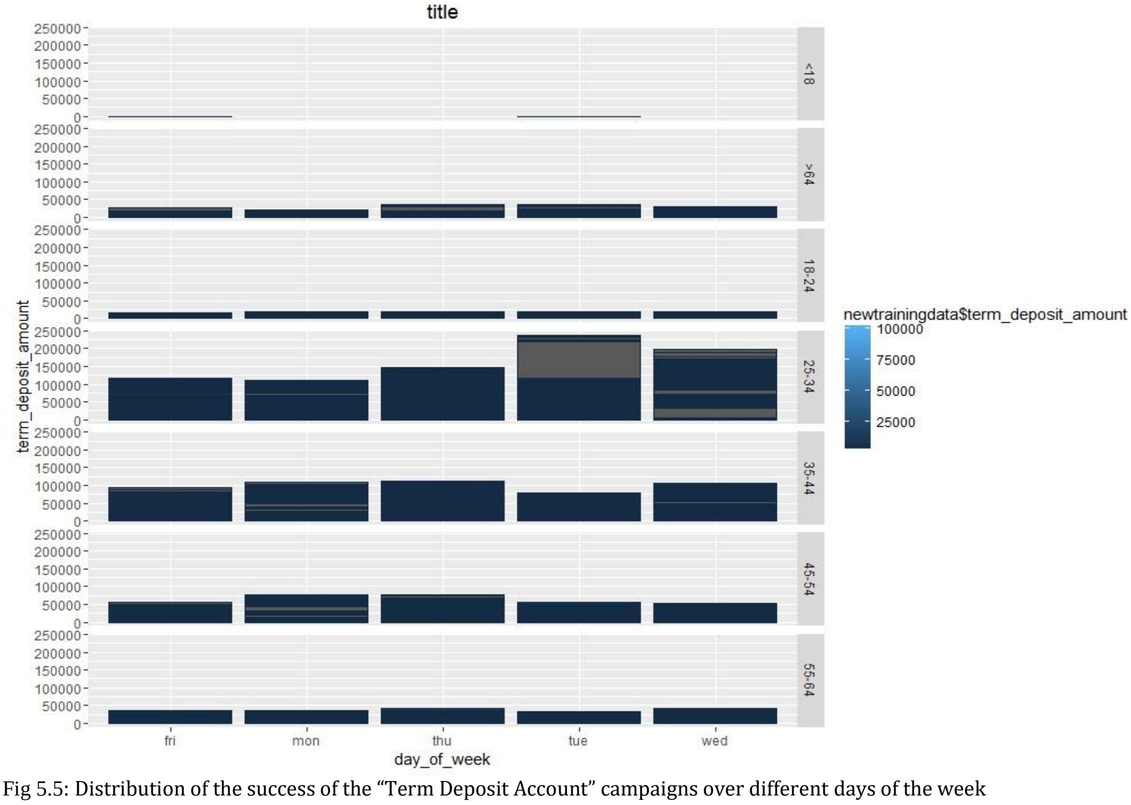


Fig 5.5: Distribution of the success of the “Term Deposit Account” campaigns over different days of the week

5.6 Average Amount of “Term Deposit Amount” over different Age Groups

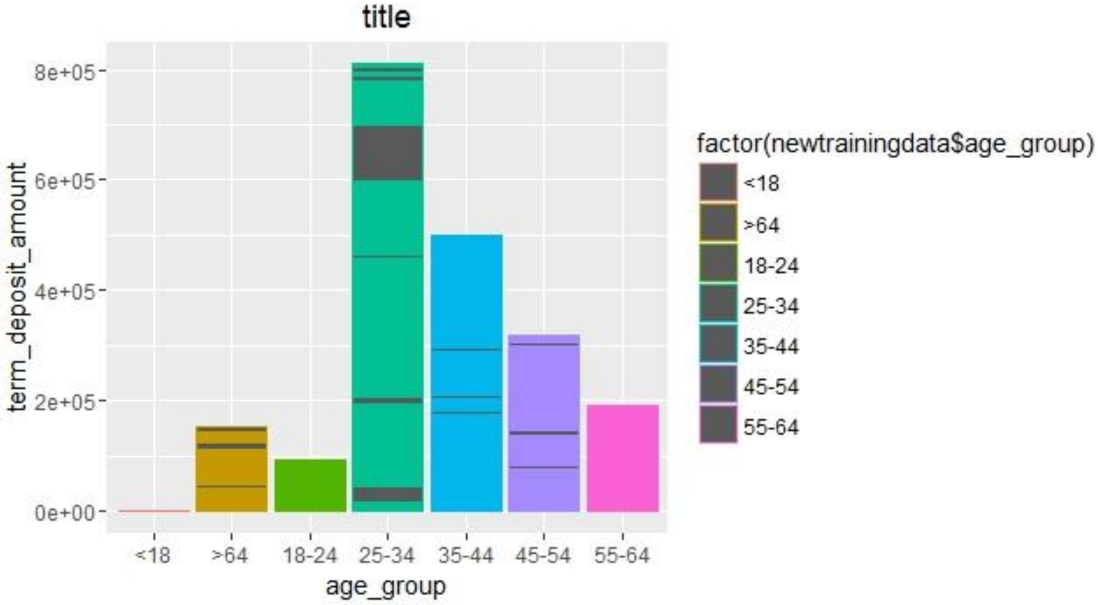


Fig 5.6: Average Amount of “Term Deposit Amount” over different Age Groups

5.7 Average Amount of “Term Deposit Amount” over different Job Categories

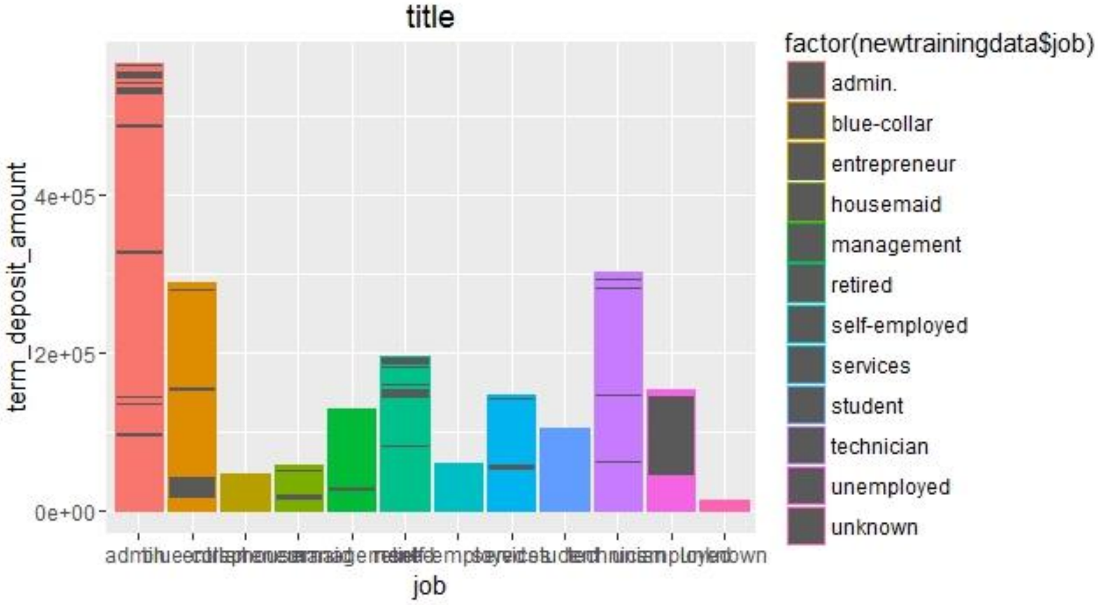


Fig 5.7: Average Amount of “Term Deposit Amount” over different Job Categories

5.8 Number of Users who have subscribed for "Term Deposit Account" for different Age Groups

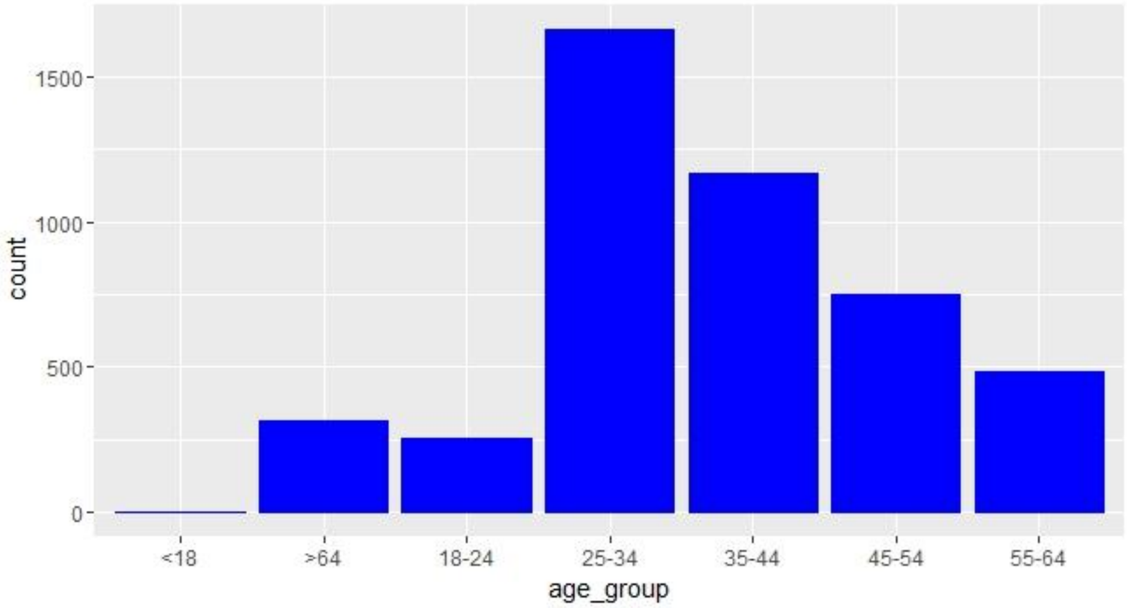


Fig 5.8: Number of Users who have subscribed for "Term Deposit Account" for different Age Groups

5.9 Distribution of the age of the customers of the bank

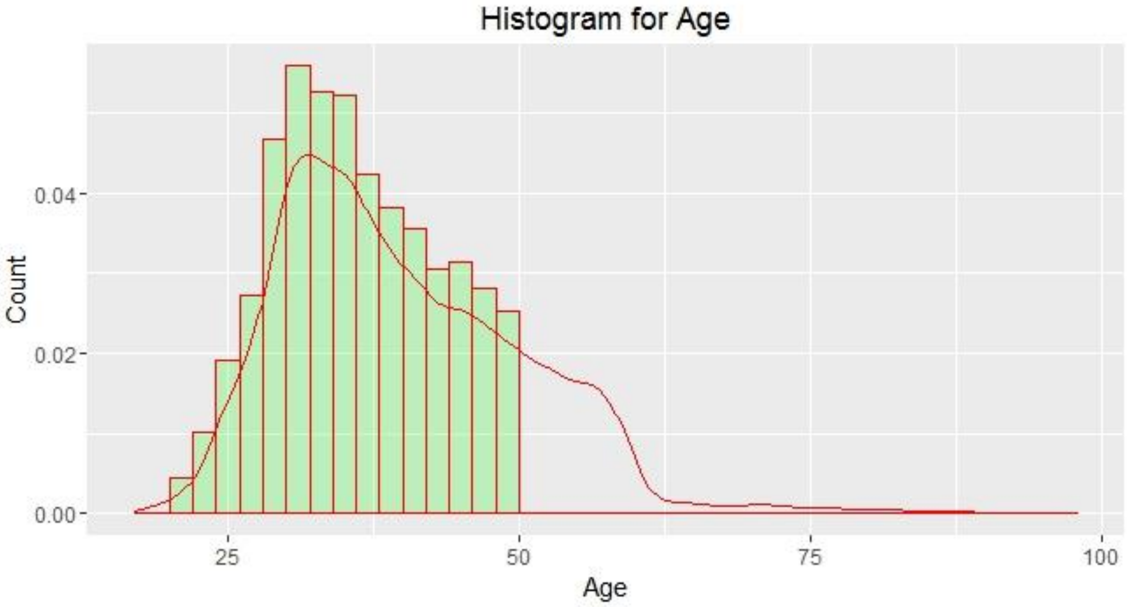


Fig 5.9: Distribution of the age of the customers of the bank

6 Implementation of Algorithms

We used Logistic Regression, Support Vector Machine and Decision Tree Model to predict whether a customer will subscribe for Term Deposit Account or not based on different characteristics of the customer. Later we used Naïve Bayes and Random Forest Algorithms to predict the amount that should be proposed to the customers for the deposit.

To predict whether a customer will subscribe for Term Deposit Account:

6.1 Logistic regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probate regression using similar techniques, with the latter using a cumulative normal distribution curve instead.

However this model is based on a certain set of assumptions:

1. The true conditional probabilities are a logistic function of the independent variables.
2. No important variables are omitted.
3. No extraneous variables are included.
4. The independent variables are measured without error.
5. The observations are independent.
6. The independent variables are not linear combinations of each other.

Following constitutes the output of Logistic Regression Model:

1. **Estimates** describes the relationship between the independent variables and the dependent variable, where the dependent variable is on the logit scale. These estimates tell the amount of increase in the predicted log odds of 'yes' = 1 that would be predicted by a 1 unit increase in the predictor, holding all other predictors constant.
   - For every one unit change in intercept, the log odds of 'yes' (versus 'no') increases by given value.
2. **Standard error** is used for testing whether the parameter is significantly different from 0
   - Z-value is obtained by dividing the parameter estimate by the standard error.
3. Coefficients having **p-values** less than alpha are statistically significant. For example, if alpha is 0.05, coefficients having a p-value of 0.05 or less would be statistically significant (i.e. we can reject the null hypothesis and say that the coefficient is significantly different from 0)
4. **Null deviance** shows how well the response is predicted by the model with nothing but an intercept without taking any explanatory variables.
5. **Residual deviance** shows how well the response is predicted by the model when the predictors (explanatory variables) are included.
6. **Akaike information** criterion estimates the quality of each model.

We considered following combination of features for the logistic regression model:

Feature Combination	Accuracy Percentage
job+education+balance	0.739744848
job+education+balance+age_group+poutcome	0.782335623
job+education+balance+housing	0.733944848
job+education+balance+loan	0.782562334
age+poutcome	0.783022571

age+nr.employed	0.738370952
duration+emp.var.rate	0.769283611
duration+balance_group	0.747791953
duration+ balance_group +poutcome	0.79892051

So the feature combination of “duration+ balance_group +poutcome” gave us maximum accuracy.

Lift Curve for this particular model:

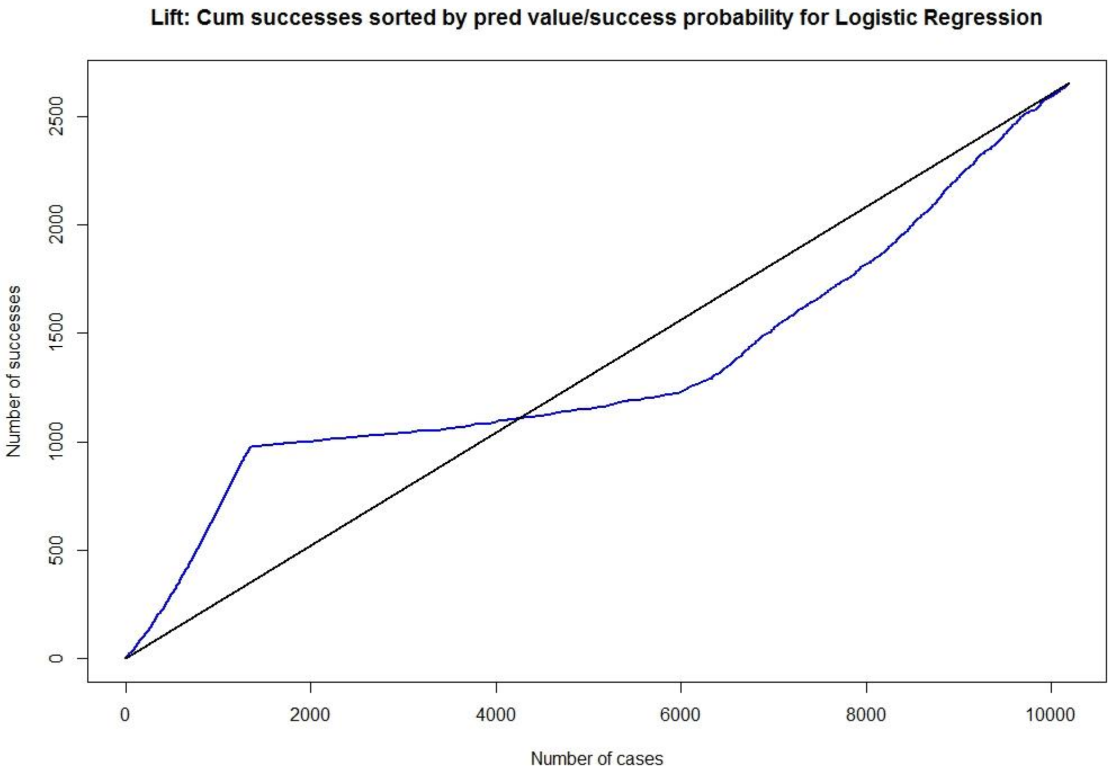


Fig 6.1: Lift Curve

6.2 Decision tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

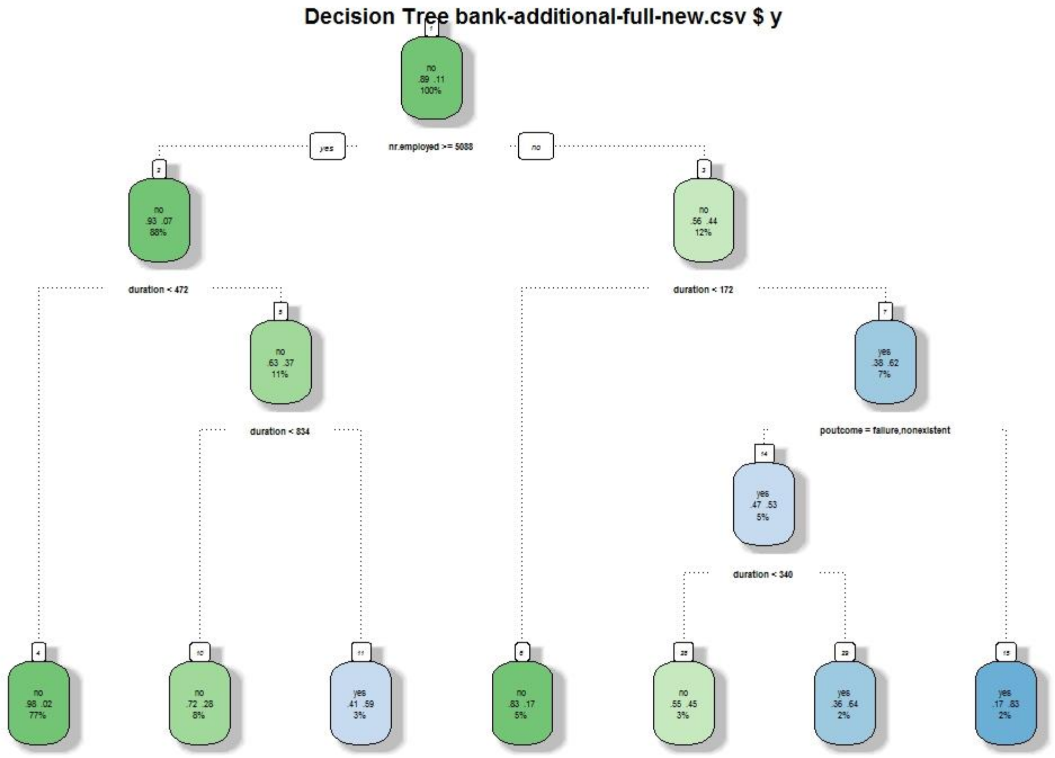
We considered following combination of features for the Decision Tree model:

Feature Combination	**Accuracy Percentage**
job+education+balance	0.880037933
job+education+balance+age_group+poutcome	0.88940256
job+education+balance+housing	0.880037933

job+education+balance+loan	0.880037933
age+poutcome	0.88940256
age+nr.employed	0.881104789
duration+emp.var.rate	0.900426743
duration+balance_group	0.890232338
duration+ balance_group +poutcome	0.903034614
All Features	0.912043623

So the feature combination formed by considering all the feature variables gave us maximum accuracy.

Decision tree for the most accurate model:



Rattle 2016-May-12 00:53:44 Neha

Fig 6.1.1: Decision tree for the most accurate model

Lift curve for the most accurate model:

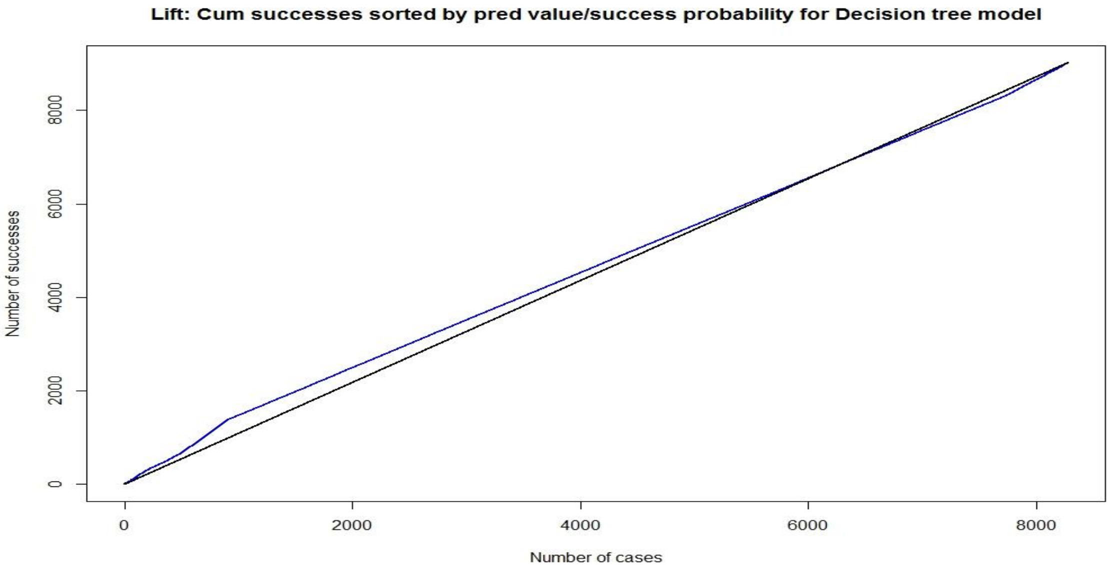


Fig 6.1.2: Lift Curve for the most accurate model

6.3 SVM

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

We considered following combination of features for the SVM model:

Feature Combination	**Accuracy Percentage**
job+education+balance	0.7397448
job+education+balance+age_group+poutcome	0.7829244
job+education+balance+housing	0.7237448
job+education+balance+loan	0.7459244
age+poutcome	0.7830226
age+nr.employed	0.7433759
duration+emp.var.rate	0.7709519
duration+balance_group	0.7504416
duration+ balance_group +poutcome	0.7987242

So the feature combination formed by considering duration, balance_group and poutcome gave us maximum accuracy.

Lift curve for the most accurate model:

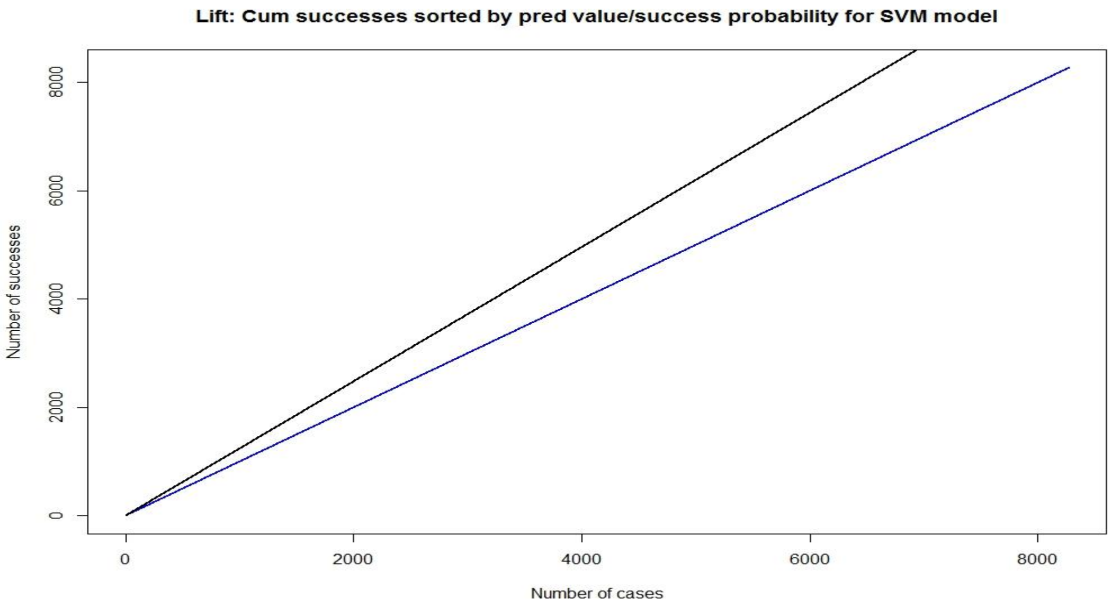


Fig 6.3: Lift curve for the most accurate model

7 To predict the amount to be proposed to the customer for Term Deposit

We had our outcome variable as "Term Deposit Amount" for this classification which is a continuous variable. Thus we divided this outcome variable into groups of amounts, which made this an outcome variable with six levels of classification viz., <5000, 5001-10000, 10001-15000, 15001-20000, 20001-25000 and >25000.

7.1 Naïve Bayes Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They can handle both continuous and discrete predictors; it assumes a distribution (can be chosen) for the continuous predictor.

As a part of the analysis, we used the training dataset to find cases that match exactly the values of the predictor variables of the case in question. The most frequent response can then be used to match cases for deciding whether the prediction for the response should be 0 or 1. The main advantage of Naive Bayes classification is that it evaluates each predictor separately.

In our case, the predictor variables are age_group (7 possibilities), balance_group (6 possibilities), housing (binary), loan (binary) and poutcome (binary).

These categorical predictor variables create (7)(6)(2)(2)(2) = 336 groups. Additionally, most of these groups contain no data. The Naive Bayesian approach looks at each predictor variable separately. Here the probability cutoff is 0.5. We score a case as success (y=1) if its probability is 0.5 or larger, and as failure (y=0), otherwise.

	No TD (0)	Yes TD (1)
No TD (0)	14351	237
Yes TD (1)	1511	369

Observations:

• Among 14588 customers who have not subscribed for a term deposit account, the Naive Bayes classification approach predicts 14351 (98.37%) correctly. But it fails to identify 237 out of 14588 predictions (1.62%) correctly.

• Among 1880 customers who have subscribed for a term deposit account, the Naive Bayes classification approach predicts 369 (19.63%) correctly. But it fails to identify 1511 out of 1880 predictions (80.37%) correctly.

• The misclassification rate of Naive Bayes classification method is 0.1060937 i.e. 10.61 %.

Lift Curve for Naïve Bayes Classifier is as follows:

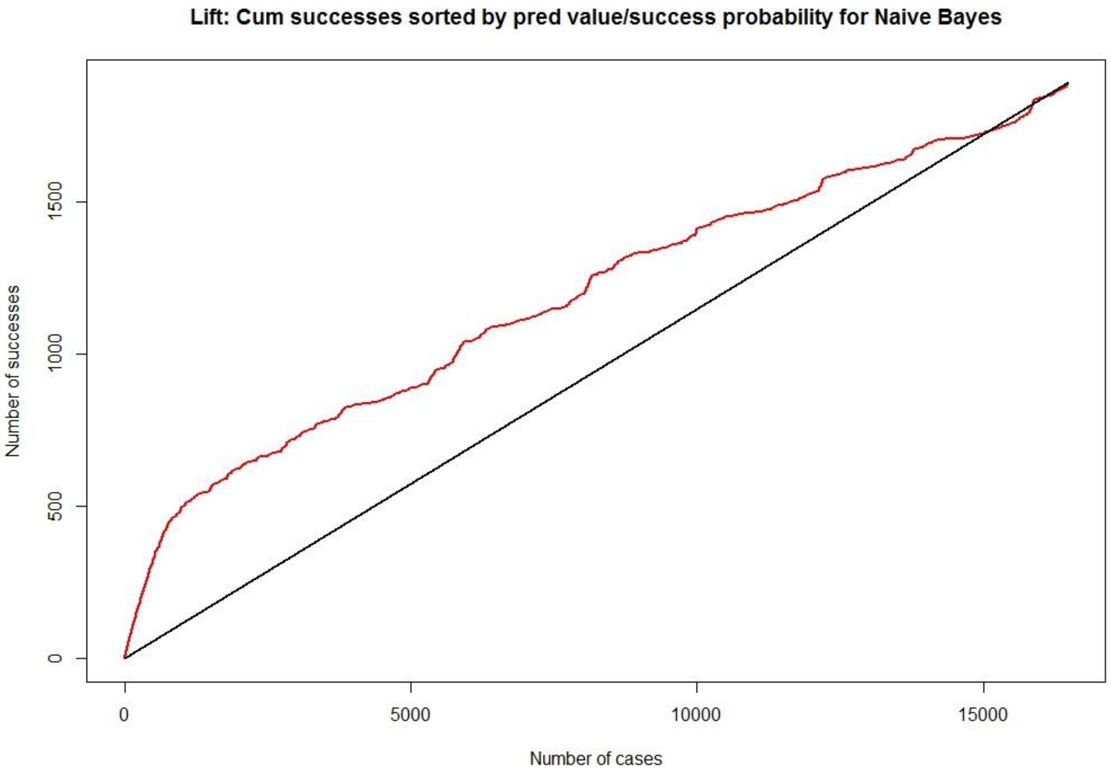


Fig 7.1: Lift Curve for Naïve Bayes Classifier

We observe that the Naïve Bayes method leads to a lift at the beginning of the curve. This indicates that this classification method is good for predicting the response, specifically for a smaller number of data.

7.2 *Random Forest*

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate 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.

Our main objective was to predict an unknown response of a case with 6 levels of classification from the information that is provided by its predictor or feature variables, which, in this application, are assumed categorical.

In our case, the predictor variables are age_group (7 possibilities), balance_group (6 possibilities), housing (binary), loan (binary) and poutcome (binary). Here we have assumed a symmetric cost structure and the probability cutoff is 0.5. i.e. if its probability is 0.5 or greater, we score a case as success (y = 1), or failure (y = 0), otherwise.

Reference	Prediction	
	No TD (0)	Yes TD (1)
No TD (0)	7143	139
Yes TD (1)	722	234

Observations:

- Among 7282 customers who have not subscribed for a term deposit account, the Random Forest classification approach predicts 7143 (98.09%) correctly. But it fails to identify 139 out of 7282 predictions (1.91%) correctly.

- Among 956 customers who have subscribed for a term deposit account, the Random Forest classification approach predicts 234 (24.48%) correctly. But it fails to identify 1511 out of 1880 predictions (75.52 %) correctly.

- The misclassification rate of Random Forest classification method is 0.1045157 i.e. 10.45 %.

Lift curve for Random Forest Classification is as follows:

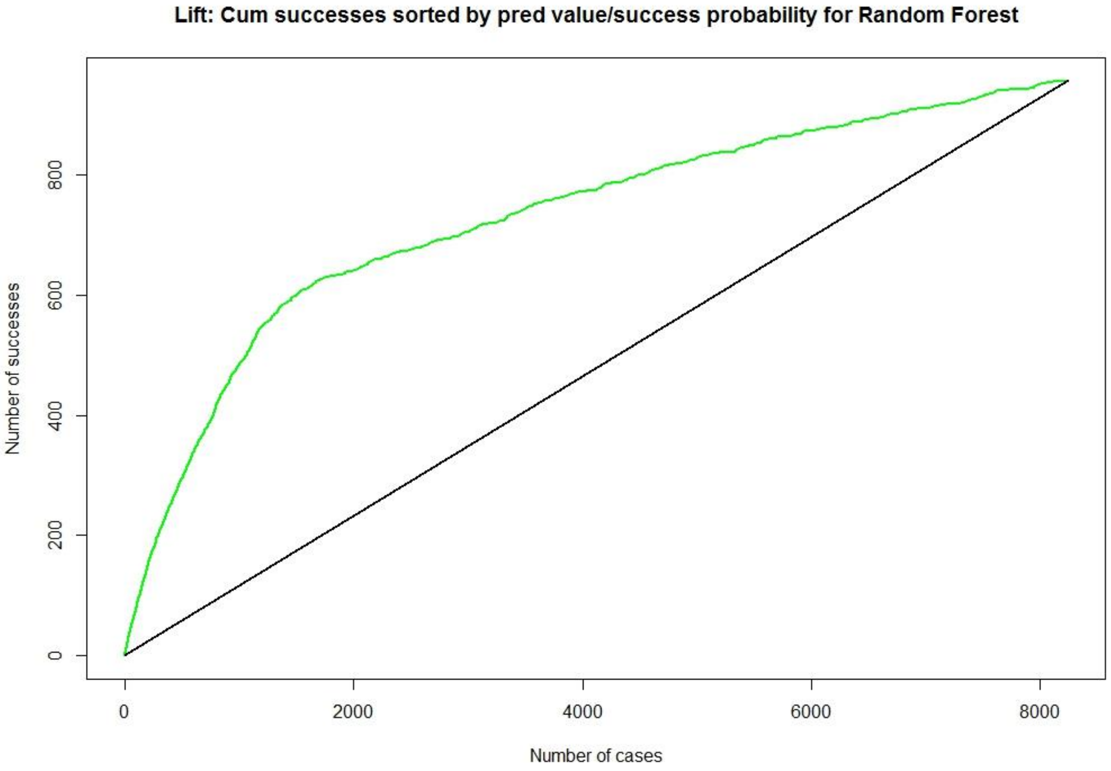


Fig 7.2: Lift Curve for Random Forest Classification

A good lift curve is one that has a very steep incline at the beginning of the curve. If the lift is close to the reference line, then there is not much point in using the estimated model for predicting the response.

It can be observed that, Random Forest method leads to a lift at the beginning of the curve, which is a significantly useful outcome. This indicates that the Random Forest classification method is a good technique for predicting the response.

8 Comparative analysis

8.1 Decision tree, SVM and logistic regression

Feature Combination	Logistic Regression	Decision Tree	SVM
job+education+balance	0.739744848	0.880037933	0.7397448
job+education+balance+age_group+poutcome	0.782335623	0.88940256	0.7829244
job+education+balance+housing	0.733944848	0.880037933	0.7237448
job+education+balance+loan	0.782562334	0.880037933	0.7459244
age+poutcome	0.783022571	0.88940256	0.7830226
age+nr.employed	0.738370952	0.881104789	0.7433759
duration+emp.var.rate	0.769283611	0.900426743	0.7709519
duration+balance_group	0.747791953	0.890232338	0.7504416
duration+ balance_group +poutcome	0.79892051	0.903034614	0.7987242

We used same feature combinations for each of the classification models, from above table we can see that, for each of the combinations, Decision tree model has greater accuracy and the feature combination of “duration, balance_group and poutcome” gives the maximum accuracy.

8.2 Naïve bayes and random forest

The misclassification rate for Naïve Bayes classification method is 0.1060937 (i.e., 10.61 %) whereas, the for Random Forest classification method it is 0.1045157 (i.e., 10.450%). Random Forest method reduces the overall misclassification error in the holdout (evaluation/test) data set hence, this is an improvement over the Naïve rule.

Lift Curve for comparison:

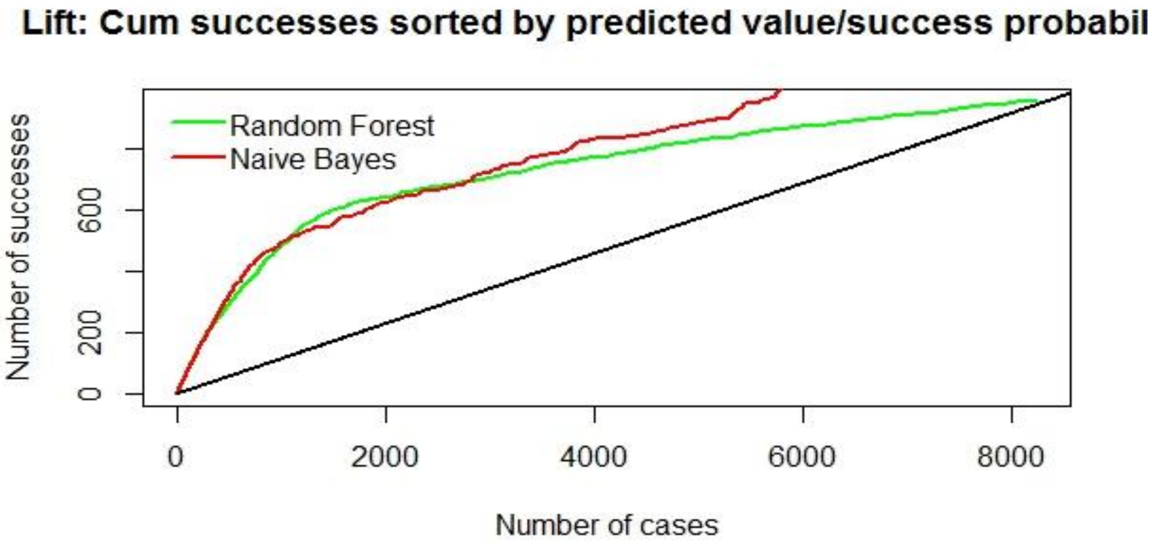


Fig 8.2: Lift Curve

The lift curves show that Random Forest method is more successful in identifying the TDA group than Naïve Bayes classification method. Hence, we can conclude that Random Forest classification method is more successful in identifying the TDA group in comparison with Naïve Bayes classification.

9 References

- https://en.wikipedia.org/wiki/Random_forest
- https://en.wikipedia.org/wiki/Recommender_system
- https://en.wikipedia.org/wiki/Collaborative_filtering
- https://en.wikipedia.org/wiki/Support_vector_machine
- https://cran.rproject.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf 2. Bhalla, D. (2015).
- Random Forests Explained in Simple Terms. Retrieved from: http://www.listendata.com/2014/11/randomforest-with-r.html
- Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Retrieved from: http://eeecon.uibk.ac.at/~zeileis/papers/Hothorn+ Hornik+Zeileis-2006.pdf 5. Jones, M. (2013, December 12).
- Introduction to approaches and algorithms. Retrieved from http://www.ibm.com/developerworks/library/osrecommender1/index.html