NATIONAL UNIVERSITY OF ENGINEERING

COLLEGE OF ECONOMICS AND STATISTICAL ENGINEERING

Statistical Engineering Program

DESIGN OF A PREDICTION SYSTEM FOR EVALUATING THE SUCCESS OF TELEMARKETING CAMPAIGN DEPOSITS FOR CUSTOMERS OF BANKUNI COMPANY USING RANDOM FOREST AND NAÏVE BAYES TECHNIQUES

PREDICCIÓN DEL ÉXITO DE LA CAMPAÑA DE TELEMARKETING MEDIANTE DEPOSITOS A PLAZO PARA LOS CLIENTES DE LA COMPAÑÍA "BANKUNI", MEDIANTE LA COMPARACIÓN DE LA TECNICA DE RANDOM FOREST Y LA TÉCNICA NAIVE BAYES.

COURSE:

Thesis workshop

Author:

- CHERO CAJUSOL CARLOS ENRIQUE

2016

CONTENT

TABLAS

ABSTRACT

The BankUNI Company is a financial institution whose main activities are to provide credit services to customers, receiving savings deposits, etc. In a context of customer loyalty it is necessary to implement relational marketing tools that help improve the company's position in the financial sector in the city of Lima.

It is for this reason that this research has focused on an analysis of the success of the telemarketing campaign offering term deposits to its customers in order to increase customer loyalty.

Thus the study of the problem: What data mining technique, Naive Bayes Random Forests or has better predictive ability to determine the success of the telemarketing campaign term deposits?, This predictability will be measured by indicators such as Specificity, sensitivity and the ROC curve.

KEYWORDS: Random Forest, Naive Bayes, specificity, sensitivity, Gini Index.

Introduction

The present research work aims to compare two classification models to extract a valuable knowledge about the success of the telemarketing campaign applied by a Bank and increase the knowledge about the marketing oriented to the loyalty of the clients. The objective of companies in the financial sector is to generate income by offering various products and services to the market. In order to achieve this, several financial companies have invested in the implementation of the Customer Relationship Management area, in charge of customer management and having as main objective the development of efficient marketing strategies to achieve customer loyalty. The client is not only a data, but a whole world of information (tastes, needs, interests, expectations). For this reason, companies are looking for the creation of appropriate strategies to attract their customers in order to acquire their products, and thus meet their objective, generate income.

To see the success of the strategies presented, the CRM area applies a set of tools and techniques that allow them to extract important information to know if the strategy is efficient or not, through propensity models applied to customers, that is, they adopt the concept of Data Mining in their processes.

On the other hand telemarketing can be defined as the marketing that is done by means of distance communication, such as the telephone, to target a selected set of customers, thus allowing the choice of those customers more likely to acquire the product / service (Tapp, 2008). With the automation of telemarketing through the integration of computers and telephony, it has become more common and easier to generate a wide variety of marketing campaign results reports.

The structure of the thesis work is constituted by the critical analysis of the previously reviewed background, the approach of the problem, the determination of the objectives, the analysis of the variables as well as the hypothesis approach.

CHAPTER I

1. BACKGROUND

For the present thesis about the choice of the best model of Data Mining, as a subject of study were found few works. However, the ones mentioned below are considered relevant since they were the ones that built most valuable information.

Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques-Bart Larivie`re, Dirk Van den Poel, Department of Marketing, Ghent University- August 2005

Objectives:

- Investigate both customer retention and profitability results.
- Explicitly test the differences related to the impact of the same set of explanatory variables in both results.

Conclusions:

- Random forests provide better predictions compared to logistic regression models.
- Random forests shows the importance of each variable with respect to the dependent variable.
- The behavioral variables of the clients play an important role, as well as some of the demographic variables.

Improved Marketing Decision Making in a Customer Churn Prediction Context Using Generalized Additive Models-Kristof Coussement, Dries

Frederik Benoit and Dirk Van den Poel, HUB RESEARCH PAPER- August 2009.

Objectives:

- Comparison in terms of predictive performance for the logistic regression model and the generalized additive model.

Conclusions:

- It is shown that GAM models have better prediction performance.
- It is demonstrated that GAM is able to improve marketing decision making by identifying customers at risk of leakage.

A Data Mining Approach for Bank Telemarketing Using the Rminer Package and R toll- Sergio Moro, Paulo Cortez, Business Research Unit

Objectives:

- It was to gain valuable knowledge in order to guide activities to improve the results of the campaign, to decrease the number of calls and on the other hand to increase the number of subscriptions.
- Use of Data Mining techniques (Decision Tree, Naive Bayes and SVM) in order to extract useful knowledge from the raw data.

Conclusions:

- SVM obtained the best results in ROC curve and elevation curve analysis.

- The duration of the call explains more than 20% success, but after a certain time (50 minutes) the probability of success begins to decrease.

Direct Selling Business Lead Prediction by Social Media Data Mining- Ahmed Balfagih, Dalhousie University Halifax, Nova Scotia -April 2016

Objectives:

- Identify the best technique for classifying social media data that can predict the best sales opportunities (SVM, Naive Bayes and Random Forests)

Conclusions:

- SVM obtained the best results in Curve ROC.
- Random forest yielded the best results compared to Naive by Bayes.

CHAPTER II

2. THE PROBLEM

2.1. TOPIC

Application of Random Forests and Naive Bayes to predict the success of the telemarketing campaign through term deposits in a context of customer loyalty.

2.2. PROBLEM DESCRIPTION

Nowadays financial companies are presenting Customer Relationship Management (CRM) as the main strategy to support marketers by identifying customers most prone to dropping out.
The prediction of customer leakage allows the development of marketing strategies for customer retention aimed at limiting losses and improving marketing decisions, that is, they seek to improve the marketing decision making to avoid customer desertion. It is no use investing in attracting new customers if at the same time nothing is done to keep current customers loyal.
To do this you must know the success of that marketing campaign by developing classification models. They identify a set of significant variables that influence a client's decision to subscribe to a fixed-term deposit and thus evaluate the success of the telemarketing campaign developed by the company.

2.3. PROBLEM FORMULATION

2.3.1. General Problem:

- Which Data Mining, Random Forests or Naive Bayes technique has the best predictive power to determine the success of the term deposit telemarketing campaign?

2.3.2. Specific Problems:

- What are the most significant variables for the development of the Random Forest technique to determine the success of the term deposit telemarketing campaign?
- What are the most significant variables for the development of the Naive Bayes technique to determine the success of the term deposit telemarketing campaign?

2.4. OBJECTIVES

2.4.1. General Objectives

Determine which Data Mining, Random Forest, or Naive Bayes technique has better predictability to determine the success of the telemarketing campaign.

2.4.2. Specific Objectives

- Identify the most significant variables that influence a customer's decision to subscribe to accept a term deposit using the Random Forest technique.
- Identify the most significant variables that influence a customer's decision to subscribe to accept a term deposit using the Naive Bayes technique.

2.5. HYPOTHESIS

2.5.1. General Hypothesis

- The Random Forests model has more predictive power with a 7% higher AUC value compared to the Naive Bayes model to determine the success of the telemarketing campaign.

  Independent Variable: Predictive Capability

  Dependent Variable: Telemarketing Campaign Success.

2.5.2. Specific Hypothesis

- The Random Forest technique showed that the call duration and contact month variables are the variables that most influence a customer's decision to subscribe to a time deposit.

  Independent Variable: Call Duration, Contact Month

  Dependent Variable: Customer's decision

- Using the Naive Bayes technique we found that the variables Number of days that has passed after the customer has been contacted by an earlier campaign, and call duration, are the variables that most influence a client's decision to subscribe to accept a term deposit.

  Independent variable: Call duration, Result of the previous marketing year.

  Dependent Variable: Decision of a client.

2.6. JUSTIFICATION

The present research work justifies its development in the practical importance that this entails, since it will allow putting into practice knowledge acquired during our university education as well as strengthening them. At the same time this research will provide new knowledge of Data Mining techniques little used in companies at the national level and with a high predictive capacity.

Regarding telemarketing is important since in a commercial context as the current, which focuses on customer loyalty. This requires companies to interact more with their clients, to establish personalized relationships with them and to manage all of their information through databases accessible to all employees.

The contribution of Data Mining techniques applied in this research is related to the social impact because the benefits granted by implementing this research will give the Bank Company a better loyalty of its clients from the application of feasible strategies.

Because this research will be based on true company data it can be said that this study will be of great benefit both for the manager, decision makers and for internal and external customers because the service will be improved.

2.7. OPERATIONALIZATION OF THE VARIABLES

Chart 1. Operation of Variables - Independent Variable.

General Hypothesis: The Random Forests model has a greater predictive capacity of 7.5% compared to the Naive Bayes model to determine the success of the telemarketing campaign.			
Independent Variable: Prediction Capability			
Conceptual Definition	**Operational Definition**	**Indicators**	**Values**
It aims to extract knowledge that allows it to predict trends and patterns of behavior. Often an unknown circumstance of interest is going to occur in the future but predictive analysis.	Variable that verifies how accurate the results obtained by the DM models are.	SPECIFICITY	Percentage
		SENSITIVITY	Percentage
		ROC CURVE	Percentage
		LIFT	Percentage

Source: Own Elaboration

2.8. MATRIX OF CONSISTENCY

TEMA: PREDICCIÓN DEL ÉXITO DE LA CAMPAÑA DE TELEMARKETING MEDIANTE DEPOSITOS A PLAZO PARA LOS CLIENTES DE LA COMPAÑÍA "BANKUNI", MEDIANTE LA COMPARACIÓN DE LA TECNICA DE RANDOM FOREST Y LA TÉCNICA NAIVE BAYES.			
PROBLEMA	**OBEJTIVOS**	**HIPOTESIS**	**VARIABLES**
PROBLEMA GENERAL	**OBEJTIVO GENERAL**	**HIPOTESIS GENERAL**	**VARIABLES DE ESTUDIO**
¿Qué técnica de Data Mining, Random Forests o Naive Bayes tiene mejor capacidad de predicción para determinar el éxito de la campaña de telemarketing de depósitos a plazo?	Determinar que técnica de Data Mining, Random Forest o Naive Bayes tiene mejor capacidad de predicción para determinar el éxito de la campaña de telemarketing	El modelo Random Forests presenta mayor capacidad de predicción en 7.5% frente al modelo Naive Bayes para determinar el éxito de la campaña de telemarketing.	**VARIABLE INDEPENDIENTE** Capacidad de predicción **INDICADORES** ESPECIFICIDAD SENSIBILIDAD CURVA ROC LIFT
PROBLEMAS ESPECIFICOS	**OBJETIVOS ESPECIFICOS**	**HIPOTESIS ESPECIFICAS**	**VARIABLE DEPENDIENTE**
1.-¿Cuáles son las variables más significativas para la elaboración de la técnica Random Forest para determinar el éxito de la campaña de telemarketing de depósitos a plazo? 2.- ¿Cuáles son las variables más significativas para la elaboración de la técnica Naive Bayes para determinar el éxito de la campaña de telemarketing de depósitos a plazo?	1.-Identificar las variables más significativas que influyen en la decisión de un cliente a suscribirse a aceptar un depósito a plazo mediante la técnica Random Forest. 2.-Identificar las variables más significativas que influyen en la decisión de un cliente a suscribirse a aceptar un depósito a plazo mediante la técnica Naive Bayes.	1.- Mediante la técnica Random Forest resulto que las variables duración de llamada, Mes de contacto son las variables que más influyen en la decisión de un cliente a suscribirse a aceptar un depósito a plazo. 2.-Mediante la técnica Naive Bayes resulto que las variables duración de llamada, Resultado de la campaña de comercialización Previa son las variables que más influyen en la decisión de un cliente a suscribirse a aceptar un depósito a plazo.	Éxito de la campaña de telemarketing **INDICADORES** TASA DE ACEPTACION

2.9. DEFINITION OF THE VARIABLES

It has 45211 records and 17 variables, which are the result of the direct marketing campaigns during the period May 2013 - November 2015.

Chart 2. Definition of Variables.

Variable	Description	Typo	Example
1. Age	Client's age	Numerical	45
2. Job	Typo of job	Categorical	"Admin", "bluecollar", "entrepreneur", etc.
3. Civil Stat.	Client's civil status	Categorical	"divorced", "married", "single", "unknown", etc.
4. Education	Education level	Categorical	"basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", etc.
5. Credit, arrear	Default: Do you have arrears?	Categorical	"No", "Yes", "unknown"
6. Balance	Average annual balance of all current accounts that the client owns	Numerical	…-10,…,0,1,2,…
7. Housing	Housing: Do you have housing loans?	Categorical	"No", "Yes", "unknown"
8. Loan	Loan: Do you have a personal loan?	Categorical	"No", "Yes", "unknown"
9. Contact	Contact: Type of contact communication.	Categorical	"cellular","telephone"
10. Month	Month: Last month of contacts for years.	Categorical	"jan", "feb", "mar", ..., "nov", "dec"
11. Days	Last contact day of the week.	Categorical	"mon", "tue", "wed", "thu", "fri".
12. Duration	Duration of last call in seconds.	Numerical	1, 2,…
13. Contact.1	Number of contacts made during this campaign and for this client (includes last contact)	Numerical	1, 2,…
14. Pdays	Number of days that have passed after the customer has been contacted for a previous campaign (-1 means customer was not previously contacted)	Numerical	-1, 0, 1, 2,….,
15. Previous cont.	Number of contacts made prior to this campaign and for this client.	Numerical	1,2,…..
16. Previous results	Result of the previous marketing campaign.	Categorical	failure, nonexistent, success
17. Target	Has the customer subscribed a fixed-term deposit?	Binaries	Yes No

Source: Own Elaboration

CHAPTER III

3. THEORETICAL FRAMEWORK

3.1. TECHNIQUES TO USE:

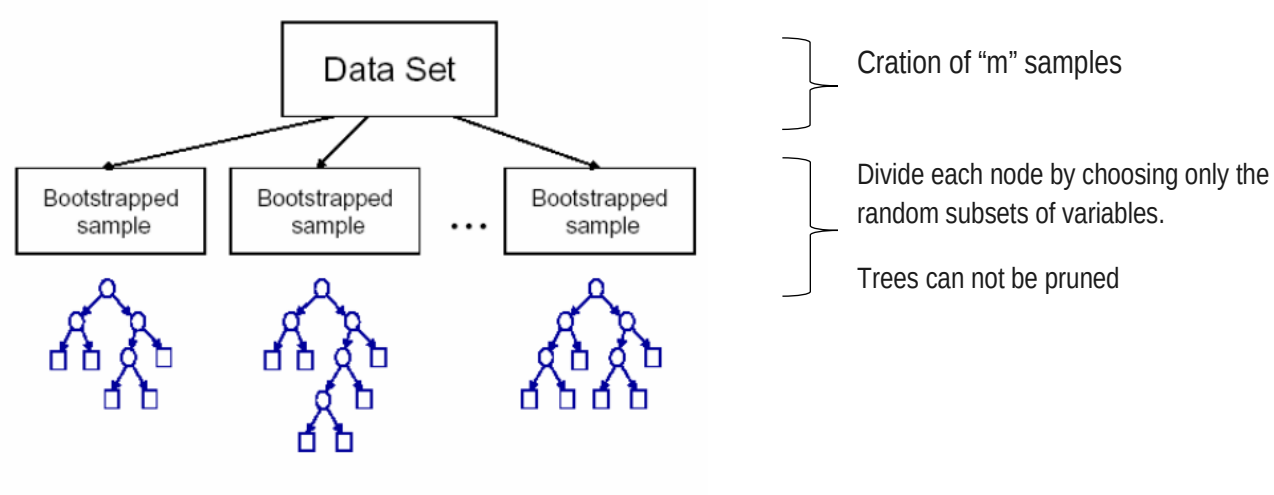
3.1.1. RANDOM FOREST

Random forests, is a classification or regression tool that has recently been developed (Breiman, 2001). It is a combination of tree predictors such that each tree is constructed independently of the others. The method is easy to understand and has proven its effectiveness as a non-linear tool.

Breiman proposes two ways of constructing random forests, both with the same objective, which is to achieve basic, accurate classifiers that are simultaneously as unrelated as possible. The first option, constructs random forests by random selection of input variables, and the second, by random linear combinations of these variables.

Random Forests are considered one of the best classification algorithms, capable of organizing large amounts of data with great accuracy. It is the combination of predictor trees in which each tree depends on the values of a random vector independently tested and with the same distribution for each of these.

Illustration 1. Intuitive form of a random forest



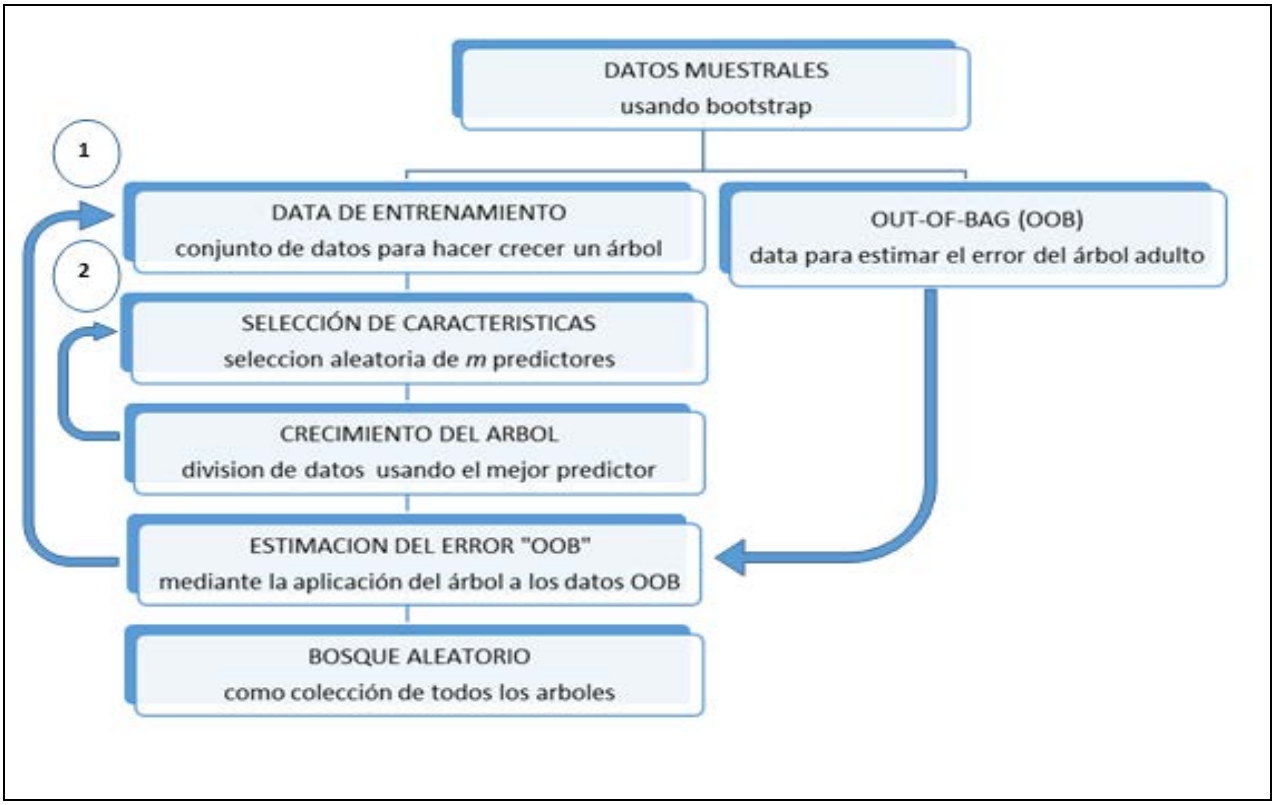
Source: Random Forest-Applied Multivariate Statistics-spring 2012

Random Forest Algorithm

1. Prepare n_{arb} bootstrap samples of the original data.
2. For each of the bootstrap samples, the classification tree will grow without pruning, with the following modification: at each node, instead of choosing the best division among all predictors, the m_{try} = p random samples from the predictors and choosing the best division between These variables. (Bagging can be considered as the special case of the random forests obtained when m_{try} = p, the number of predictors)
3. Predict new data by aggregation of predictors of n_{arb} trees (the majority of votes for the classification, the average for regression)

The number of trees in the forest depends on the number of predictors, so each predictor has sufficient opportunity to be selected. If there are p predictors we will consider $\sqrt{p}$ for classification or p/3 for regression (recommendation according to Breiman, especially with large number of predictors)

Illustration 2. Random Forests Algorithm



Source: Random Forest-Applied Multivariate Statistics-spring 2012

1- Repeat until the specified number of trees is obtained.
2- Repeat until criteria are met to stop tree growth.

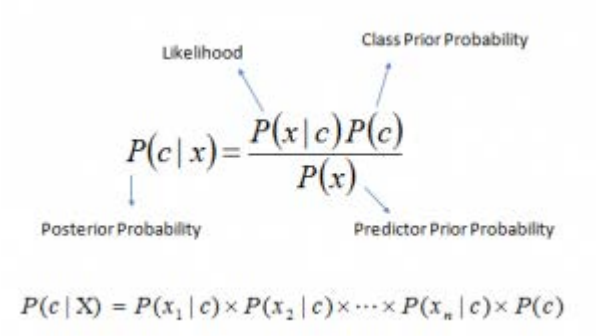
Note that the higher the correlation between tree nodes, the greater the error rate of random forests, it is desirable to have the least correlated trees as possible.

3.1.2. NAIVE BAYES

Naïve Bayes is one of the most efficient classification algorithms. It is a special case of a Bayesian network. The structure and parameters of the Bayesian network without restrictions seem to be a logical means of improvement. It was found by Friedman (1997) as a Bayesian classifier that easily overcomes such an unrestricted Bayesian network in a large sample of reference datasets.

This is a classification technique based on the Bayes' theorem with an assumption of independence among predictors. In simple terms, a Bayes classifier assumes that the presence of a particular characteristic in a class is not related to the presence of any other characteristic. Naive Bayes model is easy to build and particularly useful for large datasets. Along with simplicity, Naive Bayes is known for overcoming even highly sophisticated sorting methods.

Illustration 3. Bayes' Theorem.



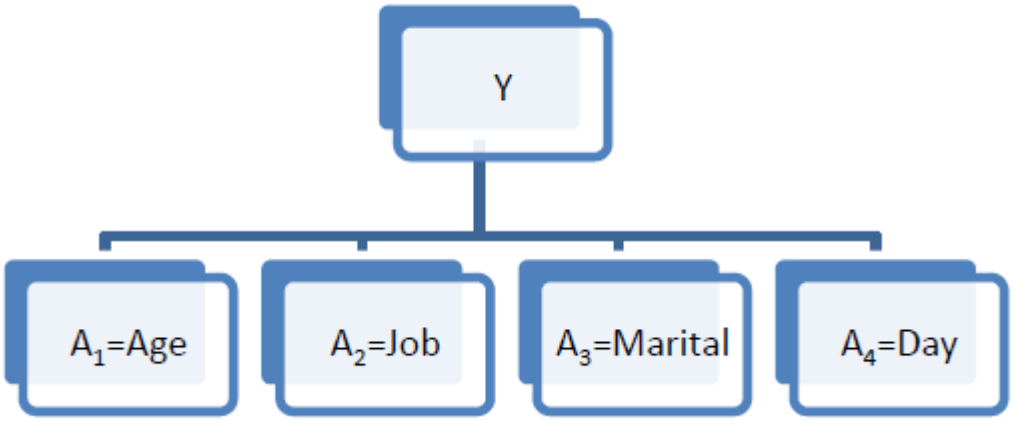
Source: Analytics Vidha-Learn Everything About Analytics

It requires a small amount of training data to estimate the parameters (means and variances of the variables) needed for classification, dealing with real and discrete data, you can also manage the data flow as well. In another form, some often apparent disadvantages of Bayesian analysis are not really problems in practice. Any ambiguity in an earlier election is generally not dangerous, since the various possible prior desirable distributions usually do not strongly disagree within the regions of interest. Bayesian analysis is not limited to what is traditionally considered statistical data, but can be applied to any model space.

The structure of Naïve Bayes in the database is applied in this document is shown graphically in the figure. In this figure, the class node is the parent for each attribute node, but there is no array of attribute nodes. Naïve Bayes is easy to construct because the values can be easily estimated from training instances [8].

In addition, to improve the assumption of conditional independence there is no way, which is to enlarge the structure of Naïve Bayes to represent explicitly attribute dependencies by adding arcs between attributes.

Illustration 4. Naive Bayes example.



Source: Own Elaboration

CHAPTER IV

4. METHODOLOGY

4.1. RESEARCH APPROACH, TYPE AND LEVEL

4.1.1. RESEARCH APPROACH

For the present investigation of the project we used quantitative techniques that were oriented towards the problem that is related to the determination of the model with better predictive capacity of the telemarketing campaign for the clients of the company "Bankuni". To obtain a better study of the problem, we investigate the factors that directly affect the success of the company's telemarketing campaign "BankUni".

4.1.2. RESEARCH TYPE

In the present research, quantitative research was used, with the purpose of different prediction methodologies to the success of a telemarketing campaign.

4.1.3. RESEARCH LEVEL

The level of research in the present study is descriptive and predictive.

4.1.4. RESEARCH DESIGN

The present research work is non-experimental of transversal character, since it is carried out with the information provided by the company of a specific period.

4.2. SAMPLE DESIGN

4.2.1. POPULATION AND SAMPLE

In the present investigation the study population consisted of 45211 clients of the company "BankUNI" each with 17 variables.

4.3. INFORMATION SOURCE

"BankUNI" a leading Peruvian financial services company uses its own contact center to carry out direct marketing campaigns, mainly through telephone calls (telemarketing). Each campaign is managed in an integrated way, the results of all calls and customers are collected in a flat file report. In this context, in September 2015, a research project was conducted to evaluate the efficiency and effectiveness of telemarketing campaigns to sell long-term deposits. The main goal was to achieve valuable knowledge had not been discovered in order to reorient managers activities to improve the results of the campaign.

CHAPTER V

5. RESULTS ANALYSIS AND INTERPRETATION

5.1. PREPARATION OF THE DATA

It was observed that the target variable presents an unbalanced distribution which causes many tools to generate a model whose answer will always be “Unsubscribe” for all clients. Unbalanced data to a classifier will produce undesirable results, such as a much lower throughput in the test training data. In order to deal with this problem a non-event sub-sampling was performed.

Target	Muestra Total	Porcentaje
SI	5289	11.70%
NO	39922	88.30%

Source: Own Elaboration

Target	subMuestreo	Porcentaje
SI	5289	30.0%
NO	12341	70.0%

Source: Own Elaboration

Para la elaboración de los 2 modelos de clasificación se dividió la muestra en 70% y 30% para la entrenar a los diferentes modelos y probar a los modelos respectivamente.

5.2. RANDOM FOREST

5.2.1. DETERMINATION OF THE OPTIMAL NUMBER OF TREES AND VARIABLES

Working with the construction sample, we determined the optimal number of trees in the forest with a default value of 4 (square root of the predictor variables) as the number of random variables in each tree. In random forest, there is no need to perform cross-validation, it is estimated internally, during the run of the model, this is through the OOB ratio.

According to this value the optimum number of trees in the model will be 500 (Figure).

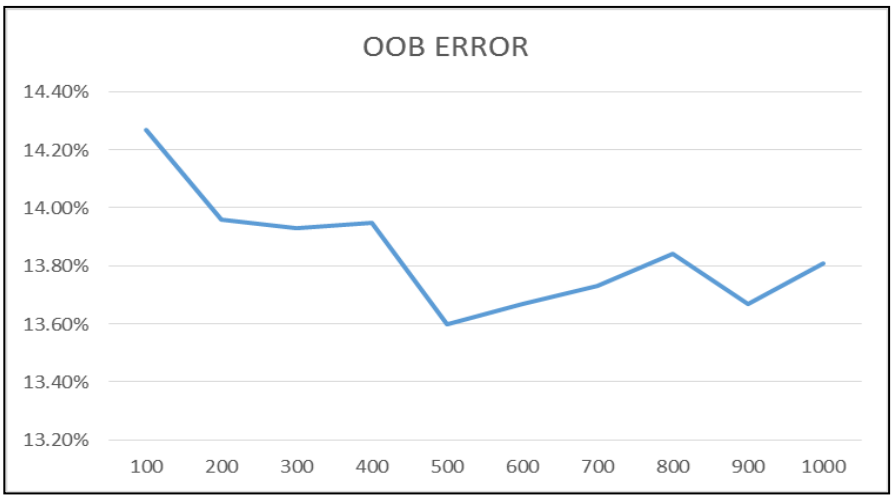


Figura: OOB Error for the number of trees.

Now we determine the number of optimal variables in each tree, using OOB we obtain that 4 is the optimal number of variables for each tree.

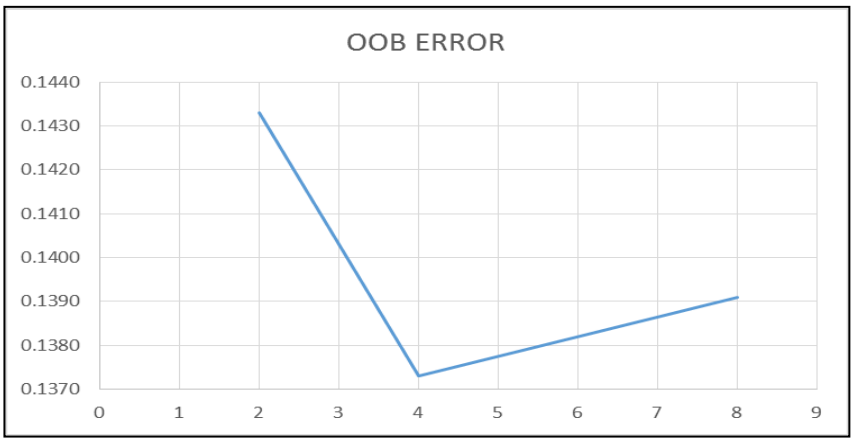


Figura2.OOB Error for the number of variables

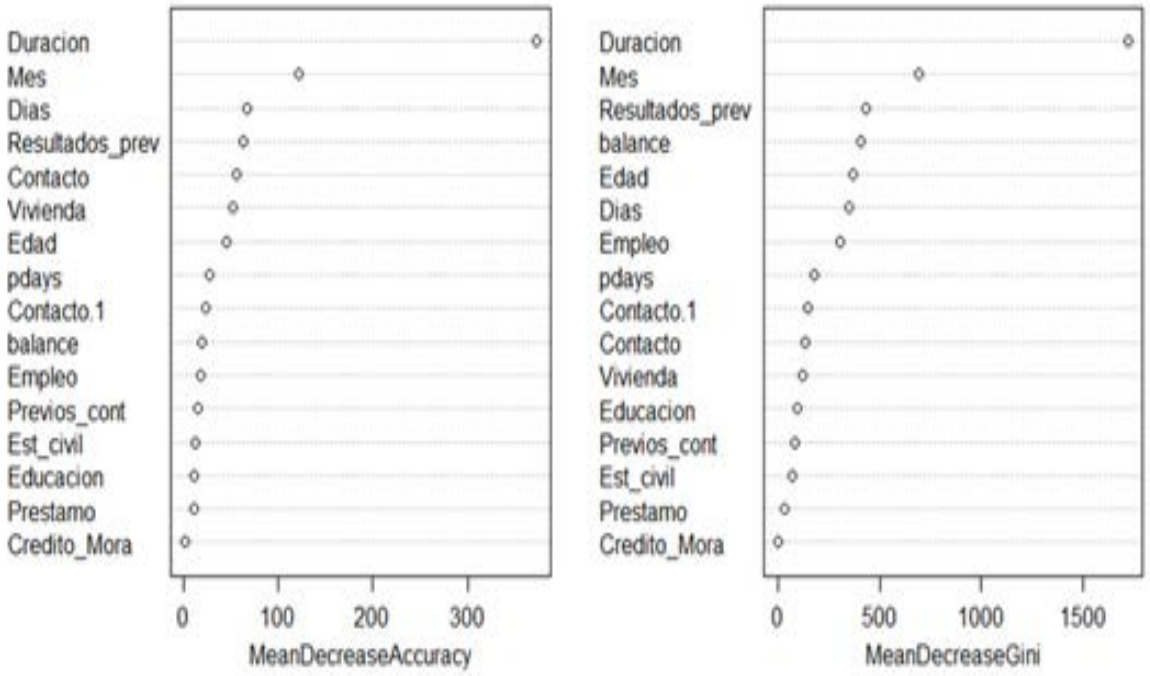
Running the run of the Random Forest model composed of 500 trees and each of 4 variables, we obtain the confusion matrix (classification matrix) that will be detailed later.

5.2.2. IMPORTANCE OF VARIABLES

We now proceed to identify our important variables. To do this we will use the Mean Decrease Gini (MDG) indicators that measure the contribution of each variable in the model construction and the Mean Decrease in Accuracy (MDA), which measures the weight of each variable when making a prediction.

According to the Figure, we have as important variables for both indicators a: Duration and month.

Importance of variables.



5.2.3. ANALYSIS OF THE POWER OF DISCRIMINATION OF THE MODEL

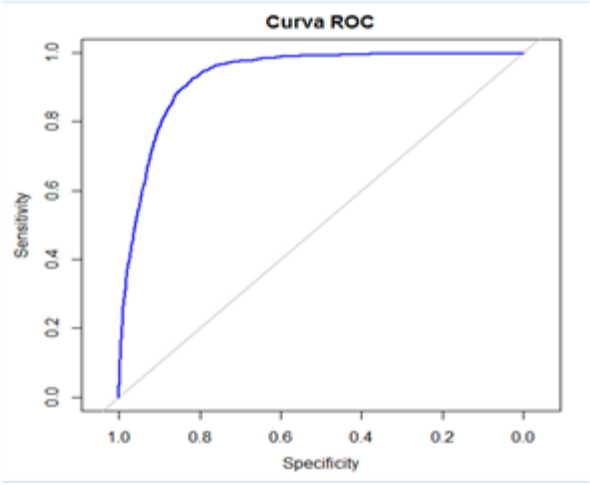
According to the confusion table and the ROC curve, it can be observed that the Random Forest technique generates a good predictor model. The confounding chart shows a 79.13% sensitivity, an overall percentage of 86.24% for the construction sample and 81.18% sensitivity, an overall percentage of 86.70% for the validation sample.

In the following table we have the sensitivities of the construction and validation samples that shows the precision with which they were estimated to the clients who accepted in the campaign.

	RANDOM FOREST					
	MUESTRA DE CONSTRUCCIÓN			MUESTRA DE VALIDACIÓN		
	NO	SI		NO	SI	
NO	2566	307	89.31%	991	122	89.04%
SI	259	982	**79.13%**	89	384	**81.18%**
PORCENTAJE GLOBAL			**86.24%**			**86.70%**

The Random Forest model was adequate because it presented favorable AUC when classifying.

We have an area under the curve of 0.9281, which indicates that the Random forest model gives us a good prediction indicating that an adequate fit of the model is obtained.

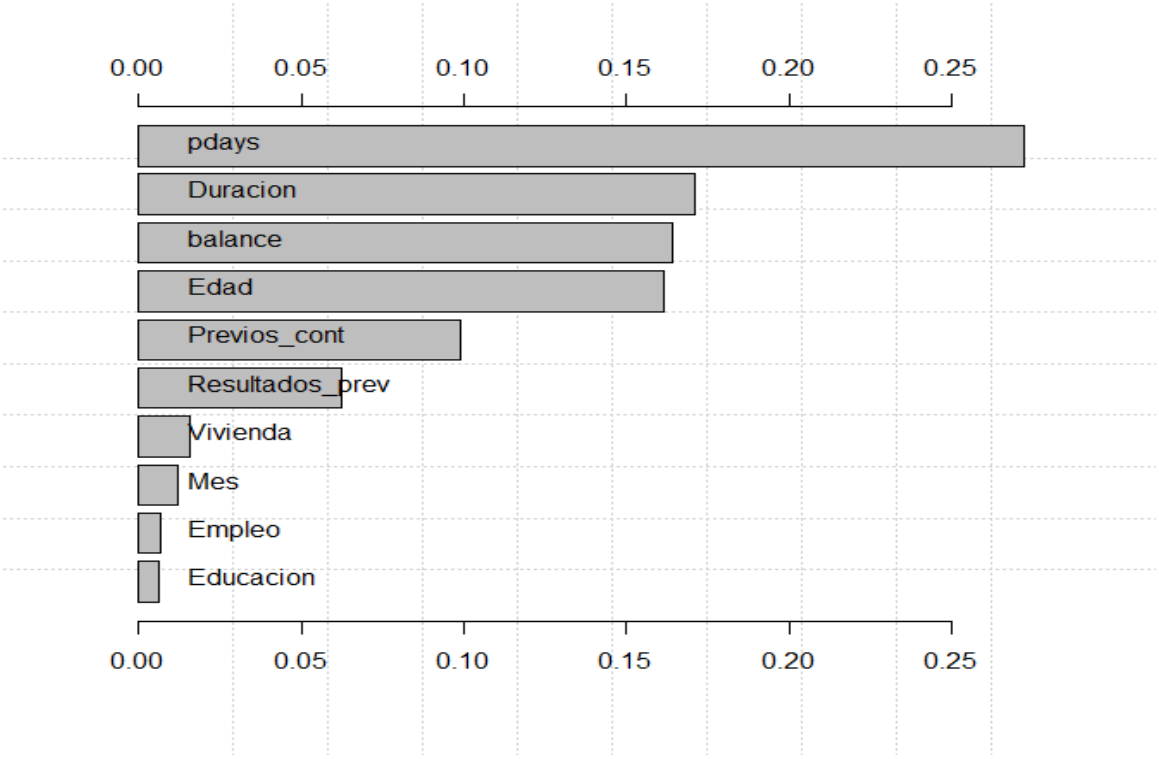


AUC=	0.92814
GINI=	85.63%

5.3. NAIVE BAYES

5.3.1. IMPORTANCE OF VARIABLES

Sensitivity analysis measures how the Naive Bayes model is influenced by each of its input variables as a percentage of the rest. In this way, it is possible to quantify the contribution of a given attribute for the Naive Bayes model. The 10 most relevant input variables (in percent) are shown in the figure.



The two most important attributes are Number of days that passed after the customer has been contacted by a previous campaign (Pdays) and Duration of last call in seconds (Duration).

5.3.2. ANALYSIS OF THE POWER OF DISCRIMINATION OF THE MODEL

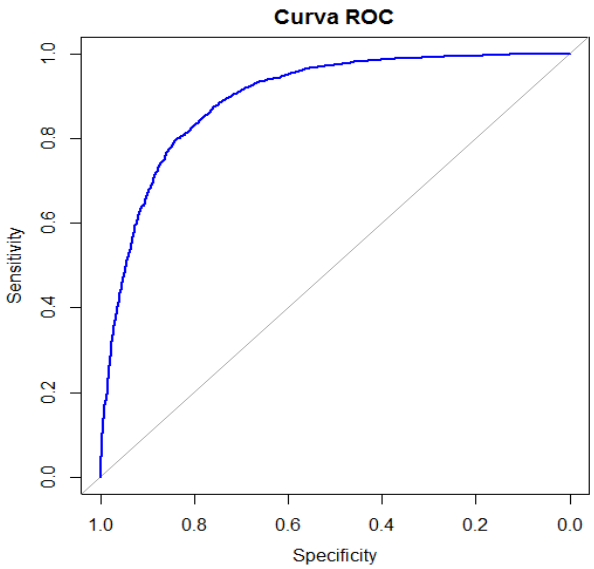
According to the confusion table and the ROC curve, it is observed that the Naive Bayes technique generates a good predictor model. The confusion table shows a 65.67% sensitivity and an overall percentage of 80.65% for the construction sample and 68.21% sensitivity and an overall percentage of 82.92% for the validation sample.

We have an area under the curve of 0.9958, which indicates that the Random forest model gives us a good prediction indicating that an adequate fit of the model is obtained.

The Random Forest model was adequate because it presented favorable AUC when classifying.

In the following table we have the sensitivities of the construction and validation samples that shows the precision with which the clients who accepted in the campaign were estimated.

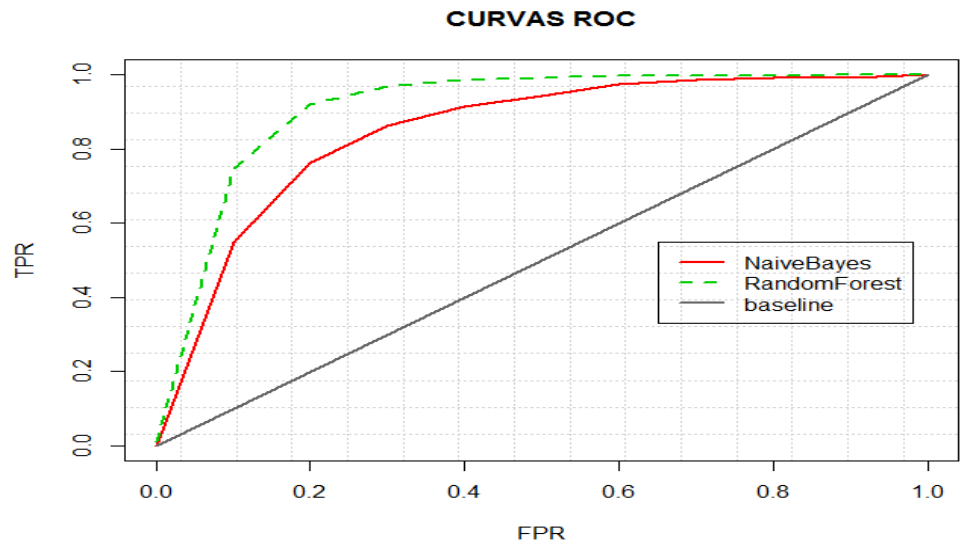
	NAIVE BAYES					
	MUESTRA DE CONSTRUCCIÓN			MUESTRA DE VALIDACIÓN		
	NO	SI		NO	SI	
NO	2503	370	87.12%	1752	212	89.21%
SI	426	815	**65.67%**	267	573	**68.21%**
PORCENTAJE GLOBAL			**80.65%**			**82.92%**

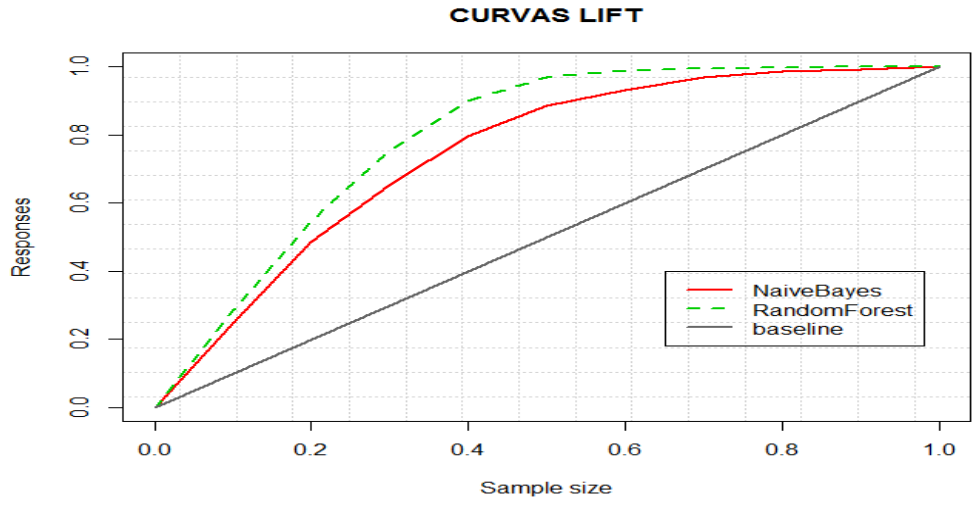


AUC=	0.8582
GINI=	71.64%

5.4. EVALUATION OF MODELS

To evaluate the two models of classification previously seen, the popular metrics are based on the confusion matrix (Kohavi and Provost, 1998) and the popular ROC curve (Fawcett, 2005). Another quite popular evaluation technique in marketing analysis is the cumulative elevation curve (ALIFT), which shows how both positive responses could be achieved from a partial sample selection (Coppock, 2002).





Based on the efficiency metrics of the two models the indicators of both models were compared, it was observed that the sensitivity, ROC curve and the accumulated Lift of the Random Forest model is higher than the Naive Bayes model indicators.

	SENSIBILIDAD	AUC	LIFT
RANDOM FOREST	81.18%	0.9281	3.05
NAIVE BAYES	68.21%	0.8582	2.98

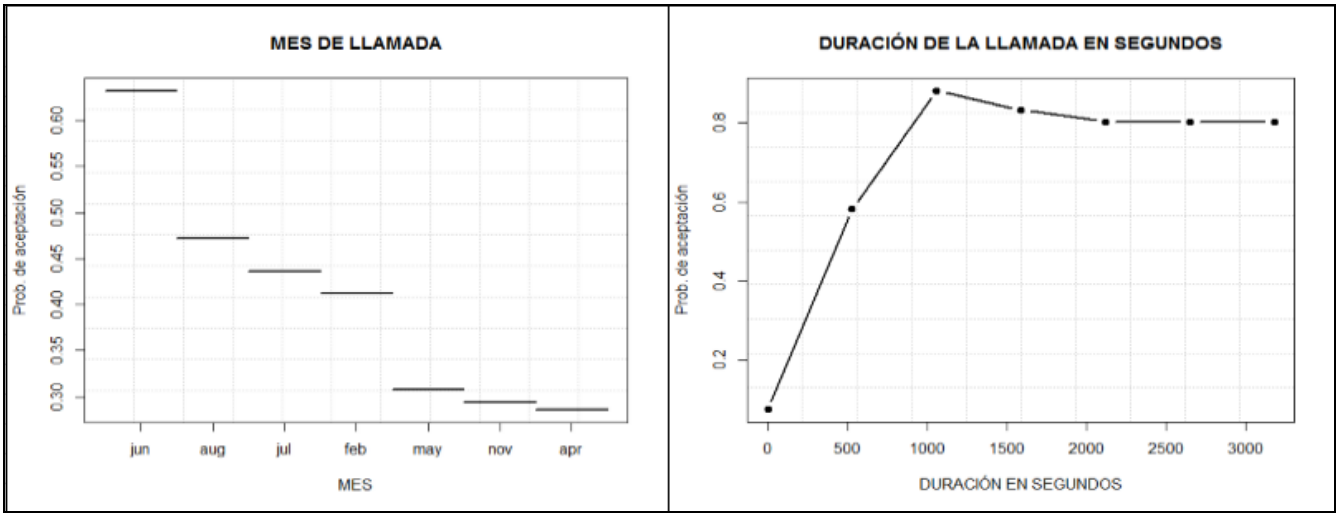
As the lift is a statistical indicator that is calculated through the ratio of the percentage of concentration of elements or facts in a given class, compared to the concentration of the population as a whole; Indicates how many times the model is better, in the capture of the objective fact, than the randomness. For this study, a lift of 3.05 was obtained which indicates that the Random Forest model helps to identify in 3.05 times more cases of clients accepting the term deposit than if we had not made any model.

By means of the Lift curve it is determined that the larger the area between the two lines, the greater capacity of the model to concentrate responders in the higher deciles. At Random Forest's LIFT, the two highest deciles capture about 55% of the people who accept the telemarketing campaign.

Based on these indicators, the model with best predictive capacity is random Forest, so based on this model we will describe what knowledge can be extracted with that model.

To obtain more details of the influence of the input variables to the Randon Forest model, in the figures we plot the variable effect characteristic curve, which shows the average influence of a given attribute (x axis) on the probability of success of the model. (Cortez & Embrechts, 2013).

For the most important variables of the Random Forest model are call duration and call month.



The figure shows that the success of the telemarketing campaign is more likely to occur during the months of June, August, so the call duration alone accounts for more than 20% of success. This result makes sense, since a successful sale requires a deeper dialogue to describe the product.

CHAPTER VI

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. CONCLUSIONS

- The Random Forest model was adequate since it presented favorable indicators when classifying if a customer will access the subscription of a term deposit after having been managed through a telemarketing campaign.
- The most significant variables for the random forest model were duration of the last call in seconds and last month of contact for the telemarketing campaign to succeed.
- In the Naive Bayes model the most significant variables were Number of days that passed after the customer was contacted by a previous campaign and duration of the last call in seconds.

6.2. RECOMMENDATIONS

- The company is encouraged to use Random Forest to determine the success of the telemarketing campaign as it presents better predictability and identifies those variables relevant to the success of the campaign.
- An important factor for the success of the campaign is the duration of the call, so it is recommended to have a deeper dialogue to describe the product (and perhaps create empathy with the customer).
- Similarly an analysis of the influence of the month reveals that success is more likely to occur in the months of July and August. This can be valuable knowledge, since managers can try to change the campaigns for those specific months.

BIBLIOGRAPHY

- **Sergio Moro, Paulo Cortez, Raul A Data Mining Approach for Bank Telemarketing Using the rminer Package and R Tool**
  http://bru-unide.iscte.pt/RePEc/pdfs/13-06.pdf

- **Random Forests. Leo Breiman and Adele Cutler**
  https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr

- **RANDOM FOREST Miguel Cárdenas-Montes**
  http://wwwae.ciemat.es/~cardenas/docs/lessons/RandomForest.pdf

- **RANDOM FOREST DATA**
  http://www.listendata.com/2014/11/random-forest-with-r.html

- **Marcelo R. Ferreyra, JANUARY 6th, 2008**
  http://powerhousedm.blogspot.pe/2008/01/cmo-medir-el-rendimiento-de-un-modelo.html

- **Bart Lariviére, Dirk Van den Poel , Predicting customer retention and profitability by using random forests and regression forests techniques (2005)**

  http://www.sciencedirect.com/science/article/pii/S0957417405000965

- **Christian Diener-** October 21st, 2014, Random Forests

  http://bis.ifc.unam.mx/es/ensenanza/metodos-computacionales-para-clasificacion/presentacion-random-forests

ANNEX

ANNEX A: COST AND BUDGET

Chart 3. Cost and budget.

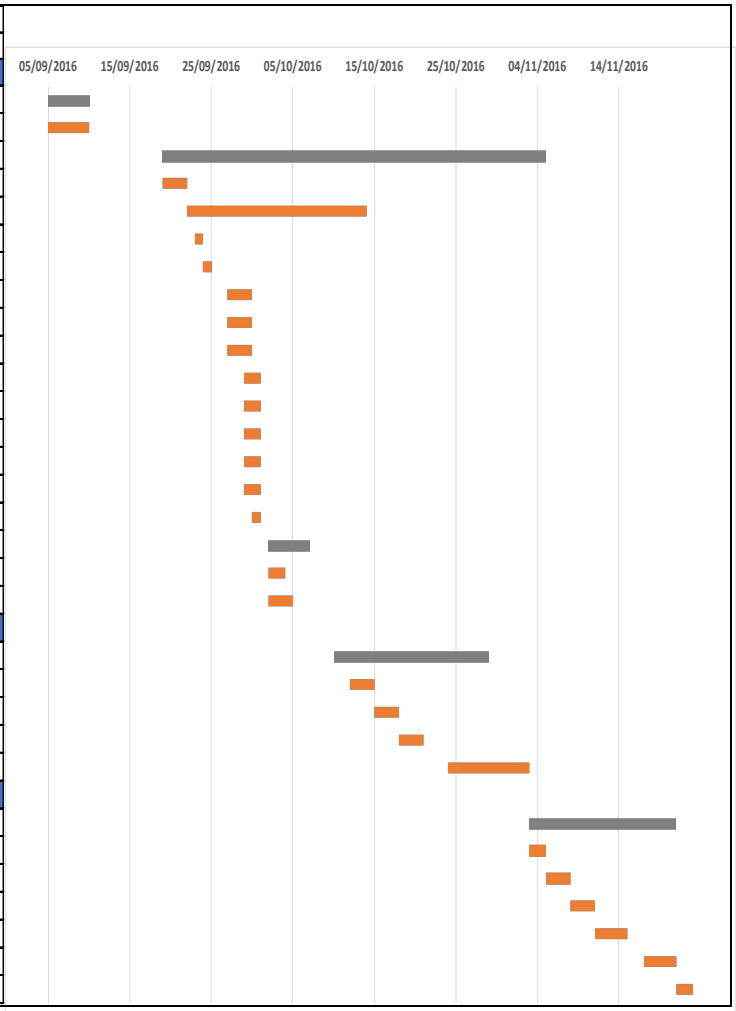
Presupuesto de Inversión		
Componentes	**Actividades**	**Costos**
Mobilidad	- Visita a Bibliotecas	S/ 15.00
Presentaciones	- CD	S/ 10.00
	- Impresión y empastados	S/ 20.00
Imprevistos	- - - - - - - - -	S/ 20.00
Total		S/ 65.00

Source: Own Elaboration

ANNEX B: GANTT DIAGRAM

Chart 4. Gantt Diagram.

	COMIENZO	DURACION	FIN
CRONOGRAMA DE PROYECTO DE TESIS	05/09/2016	**79**	23/11/2016
Primera Revisión	05/09/2016	0	05/09/2016
I.- ANTECEDENTES	**05/09/2016**	**5**	**10/09/2016**
Búsqueda de Investigaciones Pasadas	05/09/2016	5	10/09/2016
II.-PLANTEAMIENTO DEL PROBLEMA	**19/09/2016**	**47**	**05/11/2016**
Descripción del Problema	19/09/2016	3	22/09/2016
Formulación del Problema	22/09/2016	22	14/10/2016
Problema Principal	23/09/2016	1	24/09/2016
Problemas específicos	24/09/2016	1	25/09/2016
Objetivos	27/09/2016	3	30/09/2016
Objetivo General	27/09/2016	3	30/09/2016
Objetivo Especifico	27/09/2016	3	30/09/2016
Hipótesis y Variables	29/09/2016	2	01/10/2016
Hipótesis General	29/09/2016	2	01/10/2016
Hipótesis Especificas	29/09/2016	2	01/10/2016
Determinación de Variables	29/09/2016	2	01/10/2016
Definición de Indicadores de las variables	29/09/2016	2	01/10/2016
Justificación	30/09/2016	1	01/10/2016
III. MARCO TEORICO	**02/10/2016**	**5**	**07/10/2016**
Bases Teóricas	02/10/2016	2	04/10/2016
Definición de Términos	02/10/2016	3	05/10/2016
Segunda Revisión	10/10/2016	0	10/10/2016
IV. METODOLOGIA	**10/10/2016**	**19**	**29/10/2016**
Tipo y Nivel de Investigación	12/10/2016	3	15/10/2016
Determinar la base de Datos	15/10/2016	3	18/10/2016
Preparación de los Datos	18/10/2016	3	21/10/2016
Construcción de los Modelos	24/10/2016	10	03/11/2016
Tercera Revisión	03/11/2016	0	03/11/2016
V. RESULTADOS	**03/11/2016**	**18**	**21/11/2016**
Resultados de las técnicas aplicadas	03/11/2016	2	05/11/2016
Análisis de los Resultados	05/11/2016	3	08/11/2016
Discusión de Resultados	08/11/2016	3	11/11/2016
Conclusiones	11/11/2016	4	15/11/2016
Recomendaciones	17/11/2016	4	21/11/2016
Bibliografía	21/11/2016	2	23/11/2016



Source: Own Elaboration