Data and Knowledge Engineering

Assessment

Knowledge Discovery and Reporting for Businesses

Submitted by:

Johan Sebastian Ramirez Vallejo

Contents

Table of Figures

Tables

1. Introduction

Direct marketing is a frequent way to promote services and products by companies. It is using a business strategy to increment the outcome and keep their customers also satisfied. Companies as bank invest a lot of effort to hold their clients. The best way is using long terms products to increase their financial assets. For that reason, products such term deposit is an important product and strategy to grow their business.

The purpose of study is increasing the success of further campaign, finding patterns from a dataset that contains the data of a direct marketing strategy to get subscribers of the term deposit. This analysis is focused to discover knowledge to make better decisions and increase the probability of subscription to the product. Therefore, it is analyzed the dataset with different algorithms, methods, and technics to find patterns to support the decisions and archived the goal.

Data mining is a combination of methods, technics, and algorithms to extract meaningful information to be used in advance to make better decisions and avoid extra expenses(Islam et al., 2016). this is done extracting rules that are given by the algorithms, and they are evaluated to identify interestingness, after that the rule is exposed to analysis with the idea to find patterns that can support business decisions.

2. Data mining algorithm/s – technics and methods

- **SMOTE:** it is a technic to oversample an imbalance class, this works synthesizes new examples for the minority class. Those synthetic examples are generated along the class decision boundary(Siers & Islam, 2018).
- **SpreadSubsample:** it is a filter used for imbalance class in WEKA, this randomly under-sampling the majority class.
- **Sysfor:** it is a simple decision forest algorithm, it is focused on knowledge discovery through a forest instead of only focusing on future prediction and high prediction accuracy, so, it uses the original dataset.
- **Forex++:** it is a technic which helps to take the most interesting rules provide by Sysfor. This technic summarizes the top to the bottom 3 important features to find the interestedness high accuracy, high support, and short antecedents
- **J.48:** it is a simple decision tree that can only discover a set of patterns/logic rules. This use a root node which is the attribute qualification.
- **Cross validation:** it is a resampling procedure to evaluate the model on a limited data sample. This use a single parameter k (folds) that refers to the number of groups that a dataset is split. (Siers & Islam, 2018).

3. Software tool/s

- **Excel:** the tool was used for FOUR purposes, the first one was used to call the file and transform it in a valid file to be read it by WEKA, the second was to validate al the rules and ask the dataset different conditions through functions such COUNTIFS, and the third to plot the conditions and data collected to provide meaningful visualization, the four was to provide a statistical data analysis.
- **WEKA:** It is the tool used to analyze the data, this contain different filters and algorithms that can be apply to the dataset. It contains preprocess tap and feature that helps to visualize the data set and apply some filters to enhance the data analysis. I used the smote filter to populate the imbalance class. Also helps to understand the dataset by attributes, instances, records and visualization of them. Classify tap allow to call different algorithms and criteria to apply to the dataset, I used it to extract the rules trough algorithms such as forex++/Sysfore.

4. Exploration of the data set (what the data miner knows about the dataset)

The first step was the dataset exploration and highlight the information that can be useful for the data analysis, therefore, it is created a data dictionary and where were located all important information that where gather. The second step was assumptions of knowledge from the user, this is done taking account the file bank-names.txt to compare the findings against this knowledge from the exploration and applying the nine properties. This stablishes a base of knowledge to increase the expectations during the exploration.

4.1 What the data miner knows? (Assumptions - research)

This section is a compilation of questions and assumption that came to my mind in the way to understand deeply the dataset. This helped me to find some relations and may be see the information from another point of view to find patters.

- **What is term deposit (campaign product)?**
    - A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.
    - The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination—or withdrawal—if they give several days notification. Also, there will be a penalty assessed for early termination(Chen, 2021).

- I assumed that the campaign was made just to the clients already subscribed to any product in the bank
- Duration apparently is an influence factor to client to subscribed to the product.
- The goal is to build a predictive model that can define the pattern to increase the success of the campaign.

4.2 What the user knows?

This is answering with the gathering information provided by the document that contains bank text.

- **Direct marketing campaigns:** Direct marketing campaigns are a form of promotion intended to allow companies to communicate directly with their target audience using a range of media and channels – in this case was on phone calls(Moro et al., 2011).
- General aim of the bank increases a financial asset. Attractive strategy is the product term deposit due keep the client for long term for good interest rates. Therefore, the campaign aim is increase financial asset to the bank.

Call duration is the most relevant feature, meaning that longer calls tend increase success in second place comes the month of contact. Further analysis can show that success is most likely to occur in the last month of each trimester (march, June,

September, and December). Such knowledge can be used to shift can be used to shift campaigns to occur in those months (Moro et al., 2011).

5. Experiments and results

In this section I'm going to explain which the methodology that I followed to increase the success of the model. In the preview section I questioned the dataset according to my knowledge and keep it in mind the relation between findings and this knowledge(Moro et al., 2011). Therefore, this information will support me to analyze the subjective measures for example surprisingness. But first, I must set up a balanced dataset to use the algorithms and get the best results in classification, ROC area and confusion matrix.

5.1 Testing the datasets

The datasets that were given were tested against two algorithms J.48 and Sysfore, those were proved with two ways balanced dataset and imbalance dataset.

5.1.1 Test 1 (First signs)

This first round was to see what the dataset tells me the first sight, the dataset is randomly selected from a full dataset, which for computational processing perform more quickly than the full dataset. I used WEKA to explore the data set and I used 3 important algorithms. The decision tree J.48 says that duration it is the most influence attribute that determine the class "yes". However, the accuracy for this result is not reliable with almost 35%. I tried Sysfore algorithm which results were unsatisfactory because misclassify the class "yes" with almost 80%, this indicate that the class is imbalance. Therefore, I wanted to find the best classification accuracy to TP of class "no" and "yes" and ROC area.

ROC curve is a performance measurement for classification problems at various threshold setting. Roc is a probability curve which should draw a square and it should be tend to 1 for high accuracy.

TP and TN values came from confusion matrix which describe the performance of a model or a classifier. Those tell how many records were correctly classified, otherwise, FP and FN tell how many records were incorrectly classified. Therefore, computing these values against to the totals records for each class give the idea if the model is classifying correctly. Then, the TP of Class "no" and "yes" computed is the percentage of records correctly classify but both should be balanced to create a more reliable model.

WEKA brings a summarize which was compare different dataset that were summarize in the table 1.

5.1.2 Test 2 (Comparison technics)

I wanted to compare the results of test 1, therefore, I tested against the full dataset, I this this one with the same algorithms. I obtain a similar answer to the test 1, the difference was the time to process the classification. It also important that I am using crossing validation which have many interactions against the dataset. I concluded do not use this full dataset for the computational process. I also noticed that the class are not balanced, this cannot provide high accuracy in the classification. The majority class "no" weight is 88% majority class and class "yes" weight 12% minority class, which makes the class "yes" imbalance I asked what happened if under sample the

majority class or oversample the minority class. Therefore, I decided to under-sample the majority class from the full dataset and see what the results are and oversample the random dataset.

To balance the dataset there is 2 ways decreasing the majority class which is call under sample (no common technic) or increase the minority class which is call oversample (common technic).

Random oversampling technic in WEKA is call **SpreadSubsample,** it is not considered a very good technic for these reasons

- Add instances by replication
- No information lost
- Prone to overfitting due to coping same information

Synthetic Minority Oversampling technique (**SMOTE**) it is an effort to make the extra data very real to provide better prediction base on the data set this is done:

- Creates new “synthetic” observations
- SMOTE process
    - Identify the feature vector and its nearest neighbor
    - Take the difference between the two
    - Multiply the difference with a random number between 0 and 1
    - Identify a new point on the line segment by adding the random number to feature vector
    - Repeat the process for identified feature vectors

5.1.3 Test 3 (Under-sample)

I used in WEKA SpreadSubsample to under-sample the full dataset. The idea is randomly decreased the majority class records “no” to the same minority class. The full dataset are 45211 records and using this filter or technic is decreased to 10578. 50%/50% for both classes. This technic is not very reliable, but I’ll compare this against test 4 which use SMOTE technic. I run again the algorithms of decision tree and decision forest and the values of TP and ROC improve considerable.

5.1.4 Test 4 (Oversample)

I used in WEKA SMOTE technique, the minority class “yes” with 521 records was increased until 4000 records which are the same records that majority class “no”, therefore the final dataset is 8000 records. The computational process by the algorithms is good and the TP and ROC improve and behave better than the tests before in average Sysfore decision forest algorithm performance almost to 90% in classification accuracy, therefore, I will use this dataset to extract the interesting rules and make the knowledge discovery.

Dataset test	1		2		3		4	
Instances	**TOTAL** = 4521 **YES** = 521 /**NO**=4000		**TOTAL** = 45211 **YES**=5289/**NO**=3992		**TOTAL** = 10578 **YES** = 5289 /**NO**=5289		**TOTAL** = 8000 **YES**=4000/**NO**=4000	
FILTER					Spreadsubsample		SMOTE	
Algorithm	J.48	Sysfore 20/100	J.48	Sysfore 20/100	J.48	Sysfore 20/100	J.48	Sysfore 20/100
Accurancy "no"	96.0%	98.5%	95.9%	97.1%	80.9%	81.9%	86.3%	88.0%
Accurancy "yes"	35.5%	21.5%	48.1%	37.7%	89.9%	85.4%	89.4%	89.0%
ROC	76.2%	74.0%	84.3%	82.5%	88.0%	89.9%	93.7%	95.6%
	69%	**65%**	**76%**	**72%**	**86%**	**86%**	**90%**	**91%**

Table 1. Dataset Experiment Results

5.2 Attributes that make sense data

This information is gathering from the dataset that was applied the filter SMOTE in the test 4

5.2.1 J.48 simple decision tree algorithm

Interpreting the J.48 which is a simple decision tree indicate that take account the best splitting criteria.

- The best non-class attribute or best splitting criteria is “duration”. It is the attribute with less entropy of entire dataset, and this tell me that duration of a call has a mayor influence on the decision that the customer can take the decision to get the product.
- That also tell me that the agent who make the call have a huge influence on the customer
- the success of the last campaign supports the actual campaign which indicate that the current customers are happy with the company
- Another important attribute is marital status which is likely high that single and divorced people can buy the product.
- There is another pattern found in the certain months where the success of the campaign was important.

5.2.2 Sysfore a decision forest algorithm

This algorithm is a decision forest which is focused on knowledge discovery. It is combined with Forex++ which it is a technic which helps to take the most interesting rules. Attributes that call the attention and are related to the class “yes”. The confidence will be compromised to find more patters. The idea is to discard the obvious ones and have some idea of the others:

- Duration attribute is an influencer to the client subscribed to the term deposit (it is an obvious one).
- Housing attribute is an attribute that indicates that the client has or not loan, therefore is understandable that if the client has loan for a house, it is not going to take a risk. Therefore, the people who have a loan are no going make any investment, but also for the bank is not interesting if they are almost to finish their payment, because the idea is to keep more time the client and increase the financial asset.
- Job attribute have influence on people who their job is management, therefore, may be interest explore more this rule
- Cellular attribute is another the obvious one because the percentage of people that were contacted is high but combined with the attribute pdays can provide useful information
- Previous attribute is a numeric attribute that tell me that if the customer has certain satisfaction can be a potential subscriber
- Pdays attribute tell me that while the bank contacts the client can take more benefits from them

5.3 Data dictionary

This data dictionary describes the bank data(Moro et al., 2011).

#	Field Name	Type	Description	Comments
1	**age**	Numeric	it is a numeric value which indicate the age	possibly to find some patterns to target specific groups
2	**job**	Categorical	The job of people was engaged *** admin** *** student** *** unknown** *** blue-collar** *** unemployed** *** self-employed** *** management** *** retired** *** housemaid** *** technician** *** entrepreneur** *** services**	Different occupation, this attribute is important also to find patterns to target groups
3	**marital**	Categorical	this is the marital status **note:** "divorced" means divorced or widowed *** married** *** divorced** *** single**	Different marital status, this attribute is important also to find patterns to target groups
4	**education**	Categorical	The education level *** unknown** *** secondary** *** primary** *** tertiary**	
5	**default**	Binary	has credit in default? **"yes", "no"**	This means if the person have outstanding debts
6	**balance**	Numeric	average yearly balance, in euros	balance account customer
7	**housing**	Binary	has housing loan? **"yes", "no"**	This means if the person have outstanding debts
8	**loan**	Binary	has personal loan? **"yes", "no"**	This means if the person have
			Related with the last contact of the current campaign	
9	**contact**	categorical	Contact communication type: *** unknown** *** telephone** *** cellular**	what was the means by which communication was contacted
10	**day**	Numeric	last contact day of the month	indicates the last day call to the customer, therefore, have the potential to tell which days are interesting in a moth
11	**month**	Categorical	last contact month of year **Name Month:** "Jan", "Feb.", "mar", ..., "Nov.", "dec")	indicates the last month call to the customer, therefore, have the potential to tell which months are interesting in a year
12	**duration**	Numeric	**duration**: last contact duration, in **seconds**	
			other attributes	
13	**campaign**	Numeric	number of contacts performed during this campaign and for this client (**numeric**, includes last contact) **–**	how many times was contacted the costumer focused to **this campaign**
14	**pdays**	Numeric	number of **days** that passed by after the client was last contacted from a previous campaign (**numeric**, -1 means client was not previously contacted)	previous days last contacted
15	**previous**	Numeric	number of contacts performed before this campaign and for this client	customer follow-up (something the the companies do looking for client satisfaction)
16	**poutcome**	Categorical	Outcome of the previous marketing campaign *** unknown** *** other** *** failure** *** success**	how were the last campaign?
			Output variable (desired target (class)):	
17	**y**	Binary	has the client subscribed a term deposit? **"yes", "no"**	

Table 2. Data Dictionary

6. Interesting Rules

Correlation analysis is one task that can be perform knowledge discovery. The idea is carrying the analysis between class attribute and non-class attribute. To do this require some assumptions according if the correlation is high or low with the class. Therefore, this is done using excel and MS access.

The rules will be extracted using Forex++ with Sysfore. This logic rules are generally useful when they use three to five variables in their antecedents, high accuracy, and high support. Forex++ results bring the rules separate by class. In this case “no” and “yes”, the result summarizes accuracy/confidence, support/coverage, and antecedents (number of attributes related to a rule). These rules will be taken and place in excel and carry a 360-degree analysis in the next section, this is done taken each rule and test every single case against to the dataset that is load in MS access using queries. I’m interested in positive or class “yes” results. Therefore, the rules will be taken from here.

There are rules that repeat information that I already know such the duration of a call which is influential, thus I selected some rules that may have low correlation trying to find a high correlation in a segment, so, I choose a top 10 and after I choose a top 5 which will be analyzed

Rule #	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	calss(y/n)	Confidence	# Records	yes	no	confidence	Lift	Coverage
1	:	:	:	:	:	:			cellular			duration > 210 && duration > 646					yes	93%	1332	1236	96	93%	1.86	17%
2										>7		> 220 && duration > 646					yes	89%	1380	1241	139	90%	1.80	17%
3	:	:	married	:	:	:						duration > 210			> 0		yes	90%	1324	1187	137	90%	1.79	17%
4	>32											>212 duration>645					yes	89%	1269	1135	134	89%	1.79	16%
6	:	:	:	:	:	:	no								> 0		yes	87%	1527	1334	193	87%	1.75	19%
7	:	management	:	:	:	:						duration > 210					yes	83%	1689	1394	295	83%	1.65	21%
8	:	:	:	:	:	:		no			may	duration > 347					yes	83%	1223	1018	205	83%	1.66	15%
9	:	:	:	secondary	:	:			cellular			> 225					yes	82%	2434	1987	447	82%	1.63	30%
10	:	:	:	:	:	:	no			> 18.990675		> 210					yes	80%	1072	944	128	88%	1.76	13%
11	:	:	:	:	:	:							> 1.00388	> -0.9964805			yes	79%	1504	1191	313	79%	1.58	19%
12	:	:	:	:	:	> 390.051232								pdays > -0.9964805			yes	79%	1672	1310	362	78%	1.57	21%

Table 3. Top 12 Interesting Rules

Rule	age	job	marit	educatio	defau	balance	housir	loa	conta	day	mon	duration	campaign	pdays	previo	poutcom	calss(y/	Confiden	# Recor	ye	n	confiden	Lif	Covera
2										>7		> 220 && duration > 646					yes	89%	1380	1241	139	90%	1.80	17%
3	:	:	married	:	:	:						duration > 210			> 0		yes	90%	1324	1187	137	90%	1.79	17%
4	>32											>212 duration>645					yes	89%	1269	1135	134	89%	1.79	16%
7	:	management	:	:	:	:						duration > 210					yes	83%	1689	1394	295	83%	1.65	21%
8	:	:	:	:	:	:		no			may	duration > 347					yes	83%	1223	1018	205	83%	1.66	15%

Table 4. Top 5 Interesting rules

7. Justification of the Interesting Rules

The interesting rules were evaluated with nine properties based on those people who subscribed (class y = “yes”). The goal is to demonstrate the data mining techniques support the knowledge discovery on interesting rules. I observed that duration of a call is important factor that affect many attributes, however, attributes showed up as interesting groups that may demonstrate any pattern that support decision making

7.1 Rule 1: if married and duration >210 and previous > 0: y=yes (confidence 90% and coverage 17%)

Objective measure

- **Conciseness:** If marital = married & duration > 210 & previous > 0 → y = yes (1324/137) → Means we have 1187 “yes” and 137 “no”
- **Coverage/support:** → 1324 (records have the specific leaf)
- **Reliability/confidence:** → (1324-137)/1324 = 90%
- **Lift:** P(y|A) = 1187/1324 = 0.90 / P(Y) (4000/8000) = 0.5 → 0.90/0.5 = 1.8
    - **The probability that obtains a “yes” in the dataset is represented P(Y)= (4000/8000) =50%, therefore, I have 50% chance to get “yes”.** Thus, the probability to get “yes” at the set of rules or leaf or condition is given P(y|A) = 1187/1324 = 90%, so, the probability is 90%. This is the correlation factor represents the reliability of the rule. So if the correlation is greater than 1 is a positive correlation and if it is less than 1 is negative correlation which is not interestingness(Geng & Hamilton, 2006).

Subjective measure

- **Peculiarity:** the peculiarity of this rule is related to the marital status married attribute and the probability of people that subscribed. the other rules are not related to married; therefore, this interestingness apply for this rule.
- **Diversity:** the pattern is diverse against the others taking account the population of single and married, the dataset contains more records of married people, but the probability inside the group of people makes it diverse, and the possibility to find patterns such the probability of single people can get the product.
- **Novelty:** it was unknow that single people also get intertest in the product. It is found through making condition for each group and finding the probability inside of each group.
- **Surprisingness:** single people can become in an interest target group to increase the success of the campaign. Therefore, it is also surprising that there are a small group that get the product at the first contact.
- **Utility:** Exist a pattern where the first contact can turn people in subscribers. The user can use this to improve the campaign, the population of the clients mainly based on married but there is an importance potential on single people who get interest in long term investment

7.2 Rule 2: If age = >32 & duration > 212 & duration >645 → y = yes

Objective measure

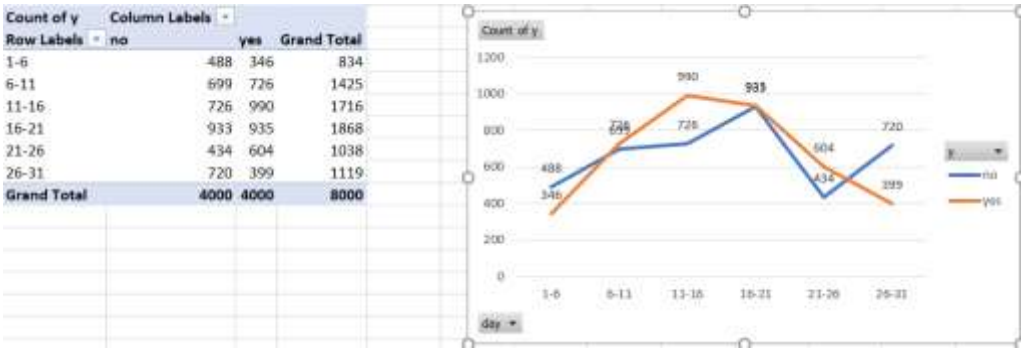
- **Conciseness:** If age = >32 & duration > 212 & duration >645 → y = yes (1269/134) → Means we have 1135 “yes” and 134 “no”
- **Coverage/support:** → 1269 (records have the specific leaf)
- **Reliability/confidence:** → (1269-134)/ 1269= 89%
- **Lift:** P(y|A) = 1135 /1269 = 0.89 / P(Y) (4000/8000) = 0.5 → 0.90/0.5 = 1.78

Lift is the correlation factor represents the reliability of the rule. So if the correlation is greater than 1 is a positive correlation and if it is less than 1 is negative correlation which is not interestingness(Geng & Hamilton, 2006).

Subjective measure

- **Peculiarity:** it is dissimilar taking account the age I can see the age target for the campaign, therefore, it is helpful to know this pattern target this group of people
- **Diversity:** the pattern is diverse in function to the analysis attribute; the age gives a better change to apply during the campaign and it is diverse to other rules.
- **Novelty:** the splitter point was 32 years which was unknown, this is helpful to pay more attention inside the target group
- **Surprisingness:** young people between 19-49 is the most representative groups to target during the campaign
- **Utility:** the rule is useful to target the group of people with specific age.

7.3 Rule 3: If day = >7 & duration > 220 & duration >646 → y = yes



Count of y	Column Labels		
Row Labels	no	yes	Grand Total
1-6	488	346	834
6-11	699	726	1425
11-16	726	990	1716
16-21	933	935	1868
21-26	434	604	1038
26-31	720	399	1119
Grand Total	4000	4000	8000

Figure 1. Day Distribution frequency

- **Conciseness:** If day = >7 & duration > 220 & duration >646 → y = yes (1380 / 139) → Means we have 1241 “yes” and 139 “no”
- **Coverage/support:** → 1380 (records have the specific leaf)
- **Reliability/confidence:** → (1380 -139)/ 1380 = 89%
- **Lift:** P(y|A) = 1135 /1269 = 0.89 / P(Y) (4000/8000) = 0.5 → 0.90/0.5 = 1.78

Lift is the correlation factor represents the reliability of the rule. So if the correlation is greater than 1 is a positive correlation and if it is less than 1 is negative correlation which is not interestingness(Geng & Hamilton, 2006).

Subjective measure

- **Peculiarity:** it is dissimilar because days attribute is not used into the others rules therefore can present patterns that can help for decision making
- **Diversity**: the pattern is diverse in function to the analysis attribute. the rules were discovered in a multiple tree
- **Novelty:** the rule reveals a pattern that can be useful for the campaign, there is a window that can be focused to increase the effort to target the group. The pattern found identify the possible days where there is better chance of a client get the product.
- **Surprisingness:** it is surprising that were found window time where there is a better change to get a positive response of the campaign.
- **Utility:** the rule is useful to intensify the effort during the windows time that was revealed by the rule

7.4 Rule 4: if job = management & duration =>210 → y= yes

- **Conciseness:** if job = management & duration =>210 → y= yes (1689 / 295) → Means we have 1394 “yes” and 295 “no”
- **Coverage/support:** → 1689 (records have the specific leaf)
- **Reliability/confidence:** → (1689 - 295)/ 1689 = 83%
- **Lift:** P(y|A) = 1394/1689 = 0.89 / P(Y) (4000/8000) = 0.5 → 0.89/0.5 = 1.65

Lift is the correlation factor represents the reliability of the rule. So, if the correlation is greater than 1 is a positive correlation and if it is less than 1 is negative correlation which is not interestingness(Geng & Hamilton, 2006).

Subjective measure

- **Peculiarity:** it is dissimilar taking account the specific job which shows more interest in the campaign’s product, therefore, the attribute job = management shows again a target group
- **Diversity:** the pattern is diverse in function to the analysis attribute, the management job audience have more change to get the product, it is the same for other occupations which were analyzed according to his specific group.
- **Novelty:** it was unknown that retired people also have interest in the product
- **Surprisingness:** it is surprisingness that a specific occupation has better acceptance for the product for example retired people.
- **Utility:** the rule is useful in function to target the group of people.

7.5 Rule 5 if loan = no & month = may & duration >347 → y= yes

Row Labels	no	yes	Grand Total	P(n|A)	P(Y|A)	P(Y|C)	Lift
jan	132	48	180	73%	27%	2%	0.53
feb	184	205	389	47%	53%	5%	1.05
mar	28	110	138	20%	80%	2%	1.59
apr	237	485	722	33%	67%	9%	1.34
may	1305	1182	2487	52%	48%	31%	0.95
jun	476	358	834	57%	43%	10%	0.86
jul	645	343	988	65%	35%	12%	0.69
aug	554	543	1097	51%	49%	14%	0.99
sep	35	76	111	32%	68%	1%	1.37
oct	43	462	505	9%	91%	6%	1.83
nov	350	172	522	67%	33%	7%	0.66
dec	11	16	27	41%	59%	0%	1.19
Grand Total	4000	4000	8000	0.5	50%	100%	1.00

Figure 2. Month Distribution Frequency

- **Conciseness:** if loan = no & month = may & duration >347 → y= yes (1223 / 205) → Means we have 1018 “yes” and 205 “no”
- **Coverage/support:** → 1223 (records have the specific leaf)
- **Reliability/confidence:** → (1223 - 205)/ 1223 = 93%
- **Lift:** P(y|A) = 1018 /1223 = 0.83 / P(Y) (4000/8000) = 0.5 → 0.83/0.5 = 1.66

Lift is the correlation factor represents the reliability of the rule. So if the correlation is greater than 1 is a positive correlation and if it is less than 1 is negative correlation which is not interestingness(Geng & Hamilton, 2006).

Subjective measure

- **Peculiarity:** this rule is having the peculiarity that is influential the month when the campaign is released.
- **Diversity**: the pattern is diverse in function to the analysis attribute, see the campaign through the different months shows patterns have not seen in another trees
- **Novelty:** this is a special rule because show a different tend as the rule says. May represents the month where many the bank obtain many subscribers but also where more subscribers decline, however, the rule shows patterns where there is better chance to get the positive response.
- **Surprisingness:** as a patter during the 1 year exist 2 periods where the effort can be improved, between January to May and July to October highlighting that the second period have less resistance to the client turn to acquire the subscription.
- **Utility:** the rule is useful to increase the effort during the periods time where the trend increases the chance.

8. Insights into the Patterns of the Dataset

In this section I will analyze from different point of view and the pattern find in each rule. This is done using excel to plot the data and extracting data from the data set I constructed using smote technic to increase the imbalance class and balance the dataset. In the section above were reported 5 rules which were identified through interestingness measures. I was identified also that the attribute duration is critical influencer on class yes. I will recall the rules in this section to explain different interesting components and observations. The goal is demonstrating that using the data mining techniques can be used to make business decisions, thus, the perspective will be focus on suggest target groups and take actions to increase the chance to get more subscriber to the product.

Duration attribute

I wanted to analyze individual duration non class attribute to observe the general behave on the class.

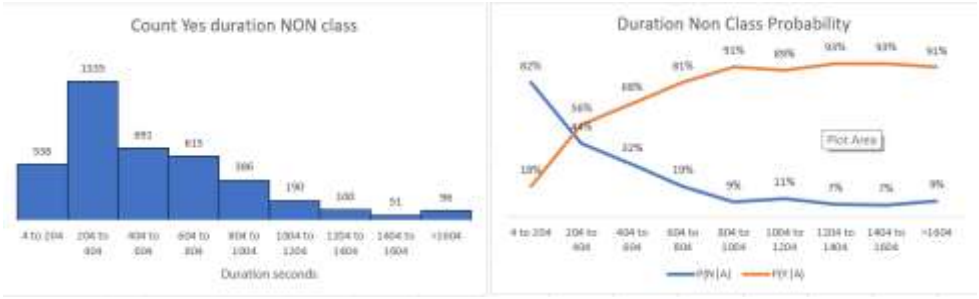


Figure 3. Duration Distribution Frequency

I can observe that clients who had a call duration more than 204 seconds had a chance of 44%, but the probability increase to 91% if the call is extended more, but the number of subscribers decrease with also large calls. Therefore, a call window between 204 to 804 seconds or 4-13 min can be effective and influential on other attributes.

8.1 Rule 1: If marital = married & duration > 210 & previous > 0 → y = yes

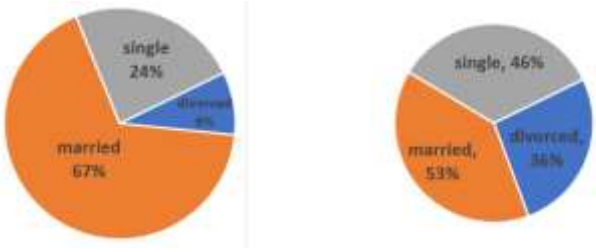


Figure 4. Weight records whole and Class yes

I observed that the married clients represent 67%, single represents 24%, and divorce 9% so married and single people is a significand important therefore, this is the reason why married is interesting, but I also observed that the probability a married client get the product is 53% and single people 46%.

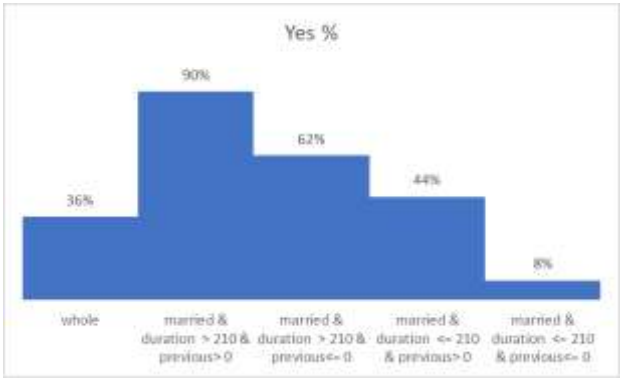


Figure 5. Married Condition Distribution

36% represents the positive response to the product the population who is married, the probability that a married person who get the campaign before and the call is greater than 4 min is 90%, but it is interesting that married people who never get the campaign before at the first contact have the possibility of 62%. However, the chance of people who have previous campaign but the call in less than 4 min is 44%.

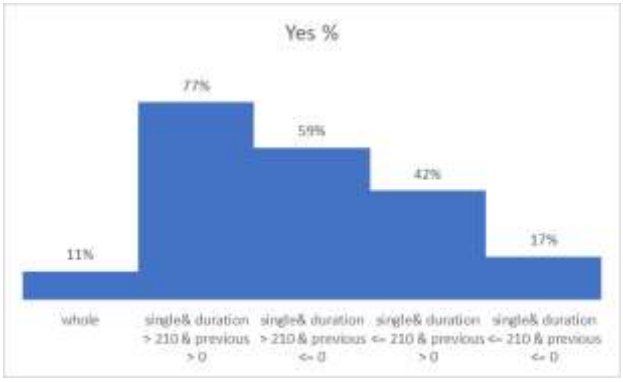


Figure 6. Single status condition distribution

The same pattern is observed with the single people, which represents the 11% of the dataset who have positive answer. Therefore, there is 77% of chance that a

single person gets the product if the call is greater that 4 min and previously contacted, and 59% if was not previously contacted.

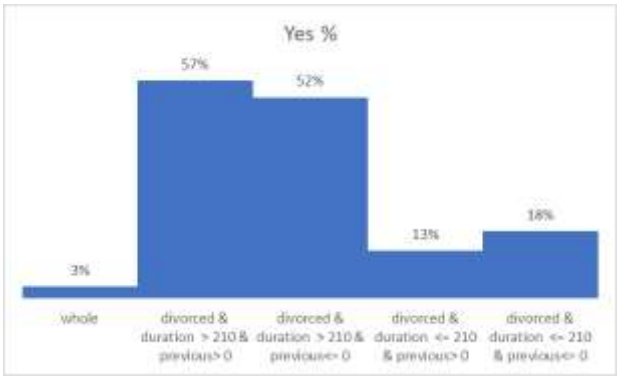


Figure 7. Divorced status condition Distribution

Divorced people represent a lower chance; therefore, they are not going take account in this analysis.

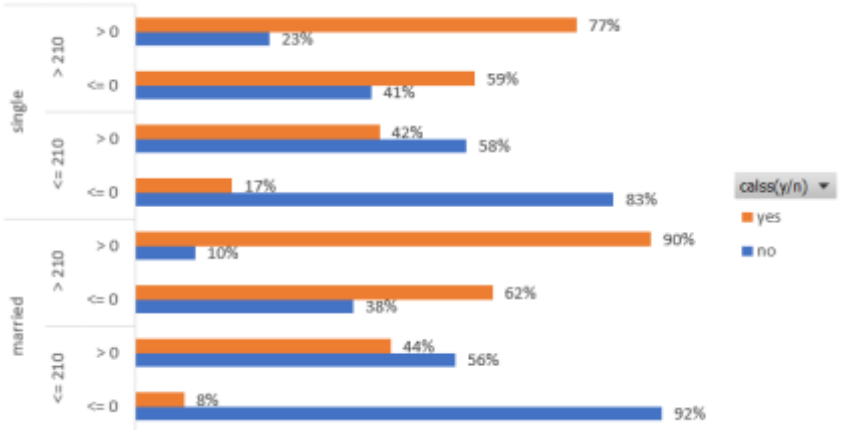


Figure 8. Comparison married and single status condition

On the other hand, duration and previous campaign play an important influence on married and single people. Although generally Married and single clients are likely to subscribed and increase the chance if they receive influence from previous campaign but decrease dramatically if they do not receive previous information.

8.2 Rule 2: If age = >32 & duration > 212 & duration >645 → y = yes

Row Labels	Count of y	Relative frecuency
19-29	635	8%
29-39	3105	39%
39-49	2269	28%
49-59	1448	18%
59-69	366	5%
69-79	141	2%
79-89	36	0%
Grand Total	8000	100%

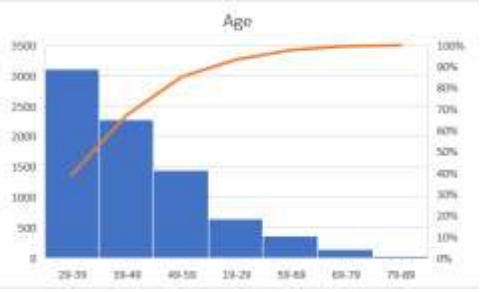


Figure 9. Age relative frequency

Age is an important factor in every campaign for that reason focuses the campaign to the target age is considerable a good technic in marketing because it is easy to identify clients with specific age therefore, I analyzed the age individually and the effect of the rule.

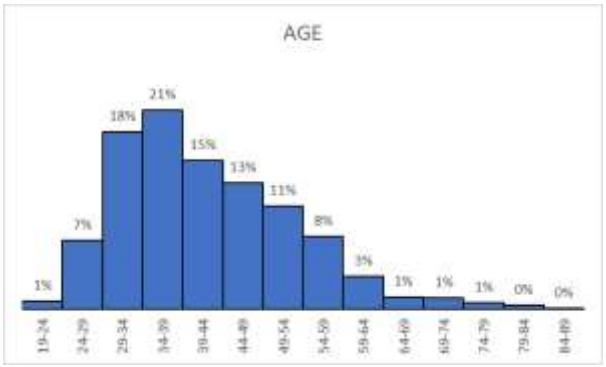


Figure 10. left skew distribution frequency

The probability between 19 to 60 years represents 91% of subscriptions therefore, the 75% represents 19-49, The histogram indicates the distribution of age in the dataset, and the clients between 32-57 indicates the better target age with a probability of 73%.

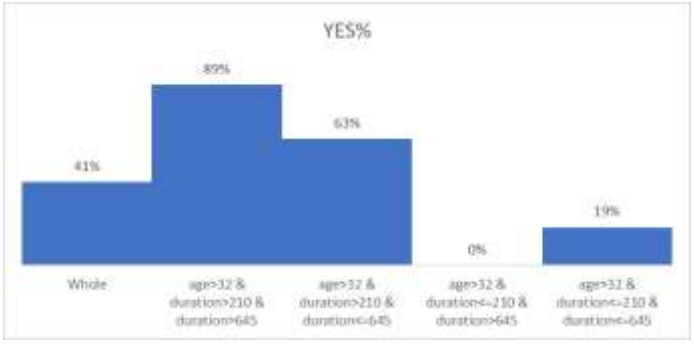


Figure 11. Age condition distribution

People who subscribed greater than 32 years represents the 41% in the whole dataset, which is significant. I observed that the rule makes an important change and increase the probability that the group of people who have more than 32 years and the call was greater than 4 min have more highest probability of 89% while 63% if the call in less than 10 min, the average age is 41 years. The duration attribute makes the difference of subscribing, so following the pattern and target the people between 32-57 have better chance

8.3 Rule 3: If day >7 & duration > 220 & duration >646 → y = yes

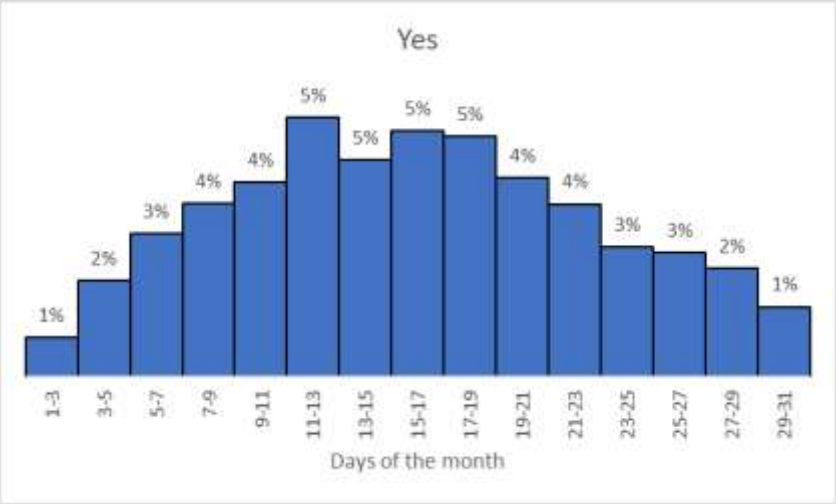


Figure 12. days attribute distribution frequency

The day attribute represents the last day of a month that the customer was contacted. It was found a pattern where exist a window in the middle of the month which is correlated to fortnight payment.

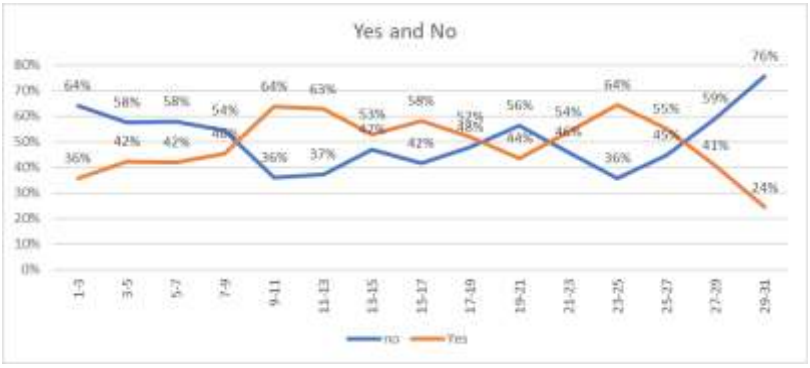


Figure 13. days attribute class yes and no distribution

It's evident that curve obtained, demonstrates the most probable event which indicate that between the day 11 to 19 occurred most subscribers, therefore, it is highly the chance that a client gets the product during this time. The trend between 7 to 11 have the probability of 64% which indicates there is a chance to client get the product. This also happened before to end the month in a lower proportion between the day 23 to 27, therefore the second and the fourth week are more chance that the customer gets the product. it is evident that chance increase before commonly payments in the middle of the month and at the end.

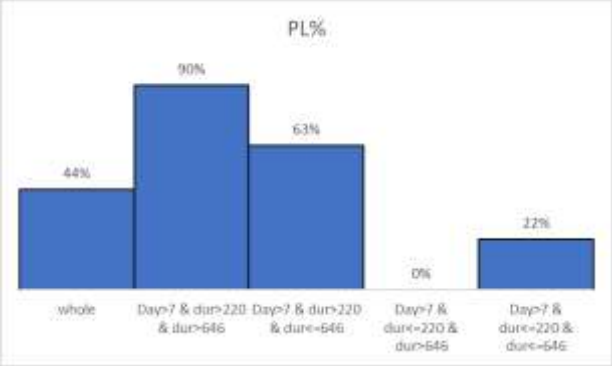


Figure 14. Day condition distribution frequency

There are 6685 clients which represents the 44 % of whole dataset that get the product after the first week. The probability has a positive response after the first week and a duration call up 4 min is 90%, is supported by 1380 clients who get the product, while 63% where calls that concreate the subscription before 10 min, and 22% before 4min. However, the first week and the third indicates a lower among of subscribers instead the second and the last one.

8.4 Rule 4: if job = management & duration =>210 → y= yes

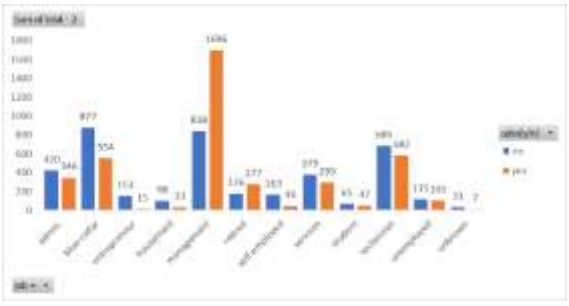


Figure 15. Comparison class yes and no job attribute distribution

The professional communities show in the distribution, management clearly indicates that they are the biggest community in the dataset with 2534 with 67% probability that a client who es manager get the product. Another interesting group is people who is retired, in the dataset, in the dataset we have 453 records with 61% of them get the product. then, manager and Retired people are the most interesting in this stage.

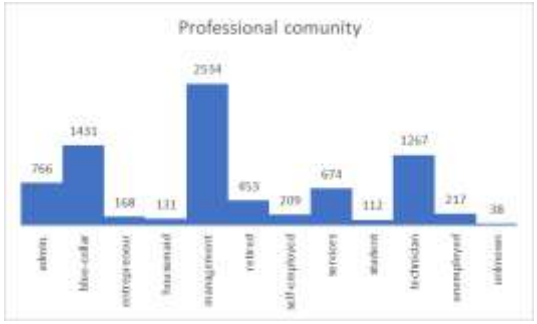


Figure 16. Job attribute distribution frequency

Duration again is a crucial factor which base the difference and provide a pattern where more professional communities demonstrate have interest on the product.

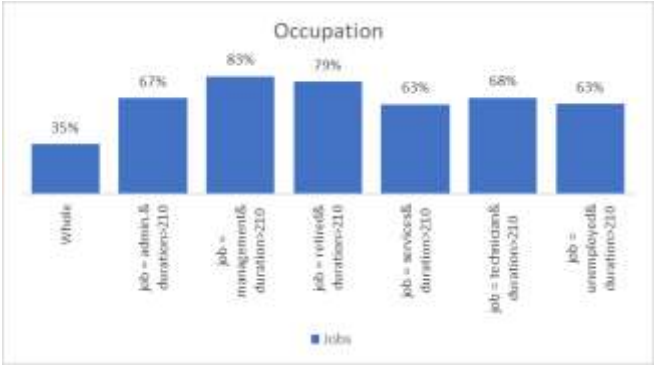


Figure 17. Job attribute distribution frequency

This occupation represents 35% of the entirely dataset. Administration community shows a probability of 67%, Management community still indicates the mayor interest on the product with a probability of 83%, it is followed by retired people with 79%, Technicians surprisingly demonstrates a possibility of 68%, followed by the administration with 67%, while the possibility of subscribers in occupations such services and unemployed with 63%. The other communities do not show differences within the community. Therefore, the chance is high within these communities, therefore target them can bring benefits where is involve the occupation and its environment.

8.5 Rule 5 if loan = no & month = may & duration >347 → y= yes

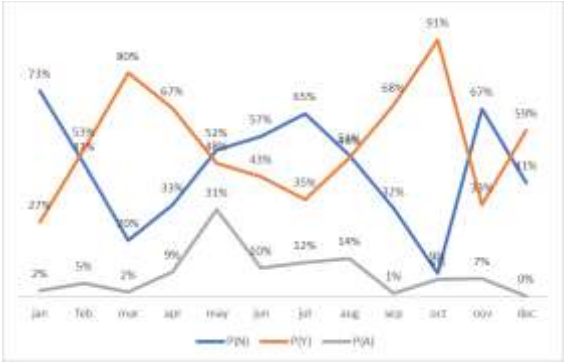


Figure 18. Month attribute distribution

It is observed that the behaved during the year of campaign, there are 2486 records In May which represent the 31% of the dataset. However, the behave of the graph indicates that there is a tendency to get more subscribers in two periods. it is observed that the probability increases after December and June holidays. Thus, holidays are the worst scenario for the campaign. May although collect the majority of records the probability to get a subscriber is 48% while no get a subscriber is 52%.

On the other hand, March has 80% of chance to get a subscriber supported by 138 records where 110 were positive and October rich the highest probability with 91% supported with 505 where 462 were positive.

Figure 19. Month Condition distribution frequency

Interestingly, it is observed that the probability increases if the clients do not possess loans with the bank, and the duration call 6 min. the graph indicate that March and October have the highest probability at the same de 95% and 97% respectively. It shows that rule overall increase the probability during the year except December and January.

Figure 20. Month condition Distribution

9. Conclusion

In this discovery knowledge process on a dataset of a campaign of A term deposit I discovered different pattern through data mining algorithms, models and technics that can be applied to increase the success of the campaign. The patterns found are interesting are useful in different ways such target groups or increase productivity during certain time.

Algorithms such Sysfore was helpful to find rules that were deeply analyzed through statistical methods to provide a meaningful data visualization and easy way to understand the patterns can support decisions and increment the success during the campaign.

Duration attribute and previous campaign were influential on the groups that I analyzed. Where the minimum duration call may have up 4 min to increase the chance, and previous campaign were interesting because it shows that the people who receive previous campaign were more influential than people who did not receive, but it can be apply in the way to increase effort in this first contact to the potential client and create strategies to provide an effective call.

It is important identify this pattern because those can reduce costs to the organization and get better benefits. For example, A large call can cost more to the bank because it can keep busy an agent that can be doing another call. So, determine windows call time can provide a cost effective to the organization.

Patterns find on day and month attributes were interesting, due to that they can be apply during the time that were highlighted. The suggest pattern indicate that there are windows where there is a high chance to get more subscribers. This can be done increasing agent to increment the number of calls during those times. Occupation pattern was important because suggest occupation or jobs that can be targeting, the same for people who is married and single.

Overall, this study provides patterns to the user that can be apply to the market of the bank and increase the possibilities positive respond to the product. the technics, methods and algorithms show again how data mining can offer attractive information base on the history and collected data from different sources to be use in business decision making.

10. Reference

Chen, J. (2021). *Term Deposit*. https://www.investopedia.com/terms/t/termdeposit.asp

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR), 38*(3), 9-es.

Islam, M. Z., D'Alessandro, S., Furner, M., Johnson, L., Gray, D., & Carter, L. (2016). Brand switching pattern discovery by data mining techniques for the telecommunication industry in australia. *Australasian Journal of Information Systems, 20*.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Siers, M. J., & Islam, M. Z. (2018). Novel algorithms for cost-sensitive classification and knowledge discovery in class imbalanced datasets with an application to NASA software defects. *Information Sciences, 459*, 53-70.