Banking Customer Data Analysis

Anibal D. Peralta Santiago
Master in Computer Science
Juan M. Ramírez, Ph.D.
Electrical & Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *– This paper focuses on making a Data Mining analysis on customer data from a bank. We try to apply different methodologies and prove some concepts using data. Some of the goals of this analysis is to produce results that my company can use to invent new products that satisfy customers; also to use the findings to help make marketing campaigns for existing products. The beauty about this is that the data, with the help of the applications used, will be the one showing us combined results that are not traditional. We will not be using the same templates and formulas that are generally used in the banking industry; on the contrary, this is a more mathematical (and logical) approach. I hope that these techniques used can align with the organization to improve knowledge management, make advances in business by making the best use of information, enable Data Mining into the business processes (if not already in place), and support with the strategic, tactical, and operational aspects of decision making.*

Key Terms *— Banking Customer, Clustering, Customer Data, Data Mining.*

INTRODUCTION

For this project, I decided to use all of the techniques and skills acquired by the Data Mining and Data Warehousing course to analyze data from the company I work for. Being an Application Administrator managing databases at a local bank, I have been exposed to several complex data warehouses and data marts, focusing primarily in data integration. The usual analysis that my team tends to do is just to provide summarized and detailed reports to view metrics invented by management. The situation is that they are always analyzing the same type of reports invented long ago by their predecessors, which does not necessarily mean that they are getting the full value of the information. What I am proposing is simply "let the data speak for itself". The project consists of extracting a portion of real customer and account data, format the data for use in the Weka and R application software, and develop charts to analyze the data using techniques learned during the Data Mining class. The results of the findings might be used to invent a new product or marketing campaign tailored to the customer's needs.

Data Mining is the process of discovering patterns in large data extracts. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Many organizations (especially the ones that have been operating for a long time), do not know Data Mining. The reality is that it can have an immediate positive impact on the sales of an enterprise, because it improves the ability to make smarter decisions and invent products. One of the most important effects of Data Mining is that increases the value of data. There is often valuable information that is "hidden" in the data; in other words, a human eye cannot associate one thing with the other. This is why a lot of data is never analyzed at all [1].

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis. The key to start is selecting an extract of data [2].

METHODOLOGY

The proposed methodologies are: Data Exploration, Data Classification, Decision Trees, and Data Clustering whereas their descriptions are:

- Data Exploration – This is one of the initial steps in Data Mining. It covers all activities in order to get familiar with the data, from identifying data quality problems to discovering first insights into the data. This is where you look for missing values, error values, etc. It is also where you can set limits to your data to minimize error margins. Additionally, you could convert a numeric data type to a range.
- Data Classification – This is more of a task than a step. Here is where you can create a class (or flag) to identify some type of data. It is somewhat like creating a custom column in the data where, given some attributes, the information can be "classified" as (y/n). This is a little difficult to do because when using applications, a model must be created and samples should be tested first.
- Decision Trees – Making tree nodes is used for Data Classification or for regression. This consists of making models in the form of a tree structure whereas a set is broken into smaller subsets while at the same time an associated decision tree is developed. The output figure created is a tree with decision nodes and leaf nodes. Every node will contain two or more branches. These Decision trees can handle both categorical and numerical data.
- Data Clustering – The process of Clustering starts with dividing a dataset into two groups: one that the members of each group are as similar as possible to one another, while the second group are data sets that are as dissimilar as possible from one another. The benefit of a cluster is that it uncovers previously undetected relationships in a dataset [3].

Basically, I will use the software to present and edit the data so that it can give me tendencies to analyze. For example, I want to be able to tell the difference in account usage of customers between different regions and what type of account leads to another type of account.

CHALLENGES

The biggest development challenges to Data Mining are data-related issues such as:

1. Assuring data quality – applications change overtime, and in those changes and migrations, sometimes there is data that becomes unusable or needs some "cleaning" in order to have trustable data.
2. Supporting highly complex conceptual data models – documentation is very important for this because the business needs to understand the complexity and variations of data created by applications.
3. Supporting access to real-time data – the challenge here could be the cost and implementation effort.
4. Graphs are difficult to analyze – sometimes the graphs that the application shows are not easily interpretable, this makes the task more difficult because one has to keep trying to make sense of what one is looking at.

PROCESS

The process begins with the development of queries using SQL developer. Once the data is in place, it is exported in a single flat file in text format. There will be a possibility that the data will have to be revised manually to maintain error margin at its minimum. The next step is to import the data in both R and Weka applications. Both applications have compatibility with text formats but in my experience, for the data to work well in Weka, its layout has to be a specific one. Going forward, the data mining techniques will be applied, which includes: identifying classifiers, exploring ranges between numeric values, etc. The last step will be to document results, print charts, and hopefully make recommendations based on findings.

In theory, I will analyze the past (Explore the data), so I can try to predict the future (Apply

concepts). Past and future-state analysis should help with building a transformation plan.

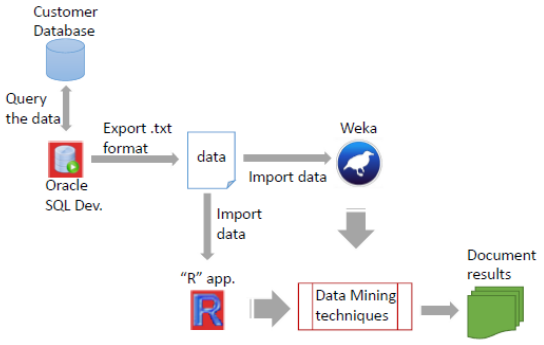

Figure 1
Process Flowchart

DATA SELECTION

As mentioned earlier, data will be collected via an SQL query using Oracle SQL Developer. An overview of the bank's DataMart is the following:

- Customer Level – at customer level, I can find the customer id, along with the demographic information, such as age, Gender, Occupation and Education.
- Account Level – at account level, I can find the account Id, type of account, branch and region.
- Transactional Level – data available at transactional level is the daily transactions (debits, credits) from the accounts.
- Services Level – data available at service level contains the services for the customers' accounts and products. This is mostly for commercial customers who manage their company's accounts and payrolls.

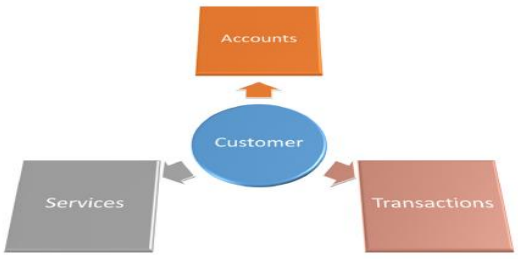

Figure 2
Data Selection

Looking at all available information, a decision had to be made as to what tier of information to collect. I decided to select information at customer level and account level only. The chosen fields from the database are:

- Customer Id – identifies the customer.
- Account_Id – identifies the account.
- Type_Of_Account – account description.
- Account_Balance – month end balance.
- Gender – male or female.
- Occupation – job of customer.
- Education – highest achieved education.
- Region – region that the account belongs.
- Branch – branch that the account belongs.

Since the account data goes up to millions of records, the data needs to be filtered. The following filters were be applied:

- Snapshot as of 10/03/2014.
- Active accounts – accounts that are currently open and have activity (debits and credits).
- Customers that their age range from 25-50 – these customers generally are the working class.
- Accounts with balances greater than $99 and less than $10,000.
- Deposit Accounts – only accounts that are checking or savings.

It all came to a total of 507,586 records. The data was exported to a CSV format in a single flat file.

DATA PREPARATION

The term "Data preparation" defines the step of constructing a dataset from one or more data sources to be used for Data Mining. This step is important because it is when you discover first insights into the data, observe and become aware of any possible data quality issues. Most errors found are: invalid, out-of-range and missing values.

The converted layout of the data selected is as follows:

- Customer Id – integer.
- Account Id – integer.
- Type Of Account – string.

- Account Balance – numeric.
- Gender - nominal with the following possible values: {'Male','Female'}.
- Occupation – string.
- Education – nominal with the following possible values: {'College Graduate', 'Post Graduate', 'No School', 'High School', 'Grade School', 'Associate Degree', 'Elementary School'}.
- Region – string.
- Branch – string.

Upon exploring the data, I came across some invalid and null values that I immediately deleted their corresponding records. Specifically there are 8,727 records with the field value of 'Gender' in blank. Additionally, I came across the 'Occupation' column with an uncategorized value that are contained in 81,189 records. Moreover, the 'Education' column has uncategorized values that are contained in 91,961 records. The 'Branch' column has some incorrect values assigned to 2,516 records. In terms of the 'Region' column, the values were almost correct in every instance, except for 2 values that indicated a subsidiary instead of a region. As you may know, each time I delete bogus records, the other fields or columns seem to have less incorrect values.

After verifying the "cleaned" data, I became aware of duplicated values written differently. This can be an issue because the essence of categorizing data gets affected. For example, the 'Region' column has the value 'Este' and 'Central-Este' which most likely belong in the same group. My course of action is to convert all those values into 'Central-Este'. The same thing happened with 'Aguadilla' and 'Mayaguez', they both belong in the West region and although most of the values have 'Mayaguez', I decided to migrate them into a new value called 'Oeste'.

At the last step of this section I started to look for categories that have too few records. In my opinion, these records are difficult to view in charts because against the other categories, they do not even represent 1% of the data. These categories are in the 'Region' column and the values are 'International' (43 records), 'Tortola' (2 records), 'US Virgin Islands' (28 records). So I proceeded to delete the categories that do not comply with a minimum of records.

R Application

In this step it was time to import the data into the R application. The goal in this task is to get different visualizations of the data. The process to import the data in R is to save a copy of the file in CSV format. The code for importing the data is:

```
DataImportR <- read.table
("C:/DataImportR.csv",header=T,sep=",")
```

Figure 3
R Import Code

An initial summary command was executed in R to try to get to know the limits and variances of the data. It shows the following results (per field):

```
          Type_Of_Account
Acceso Popular     :78662
Multicuenta Popular:57072
Ahorro A Toda Hora :56069
U-Save             :39177
E-Account          :27261
Club Del Ahorro    :20244
(Other)            :44129
```

```
Account_Balance
Min.    : 100
1st Qu.: 266
Median : 609
Mean   :1375
3rd Qu.:1575
Max.   :10000
```

```
   Gender
Female:189072
Male  :133542
```

```
       Occupation_Description
Otros Profes/Univ  : 50666
Obrero En General  : 40425
Empleados Del Ela  : 27139
Supervisor/Manager : 26647
Estudiantes        : 23255
Maestros/Profesores: 19758
(Other)            :134724
```

```
         Education_Description
Associate Degree    : 34194
College Graduate    :109472
Elementary School   :  2534
Grade School        : 29319
High School         :115942
Post Graduate       : 31153
```

Figure 4
R Summary

By looking at figure 1, one can see that we are dealing with the type of account with more customers being "Acceso Popular". In terms of the account balances, the median is 609, which lets us know that there is superior weight of accounts with $609 or more in balances. The population's gender seems to be more female than male. On the other hand, this is already showing us that the top education in the majority of customers is "College Graduate" and "High School"; however, at this stage the balances for each have not been compared.

When looking at the different occupations, notice that the "Other" seems to have the majority of records, followed by university professors. In order to be practical, the classifications such as "Other" need to be avoided because they really do not tell us anything. It just means that the record does not have a clear classification in that field.

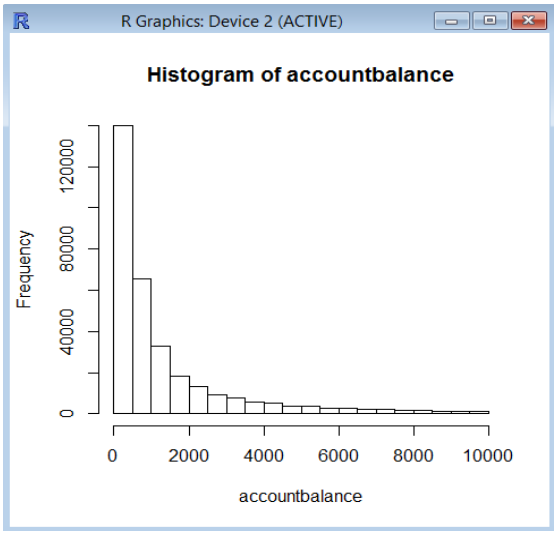

Figure 5
R Balance Histogram

Both Figure 5 and Figure 6 show that the vast majority of the accounts have less than $2,000 balance. Figure 7 shows the distribution of the balances. It seems to be very off the normal distribution, so I decided to concentrate only on the bulk of the data, which means I filtered all of the balances greater than 2,000.

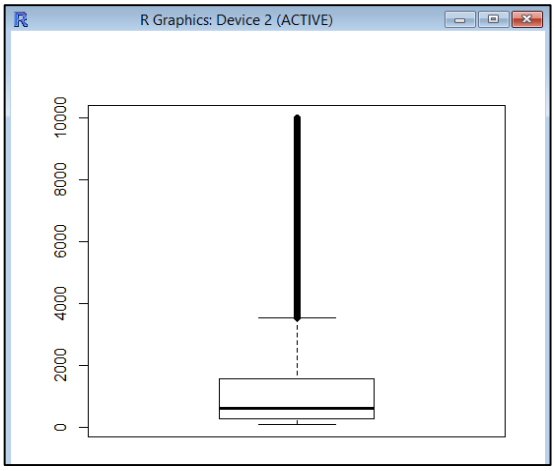

Figure 6
R Balance Boxplot

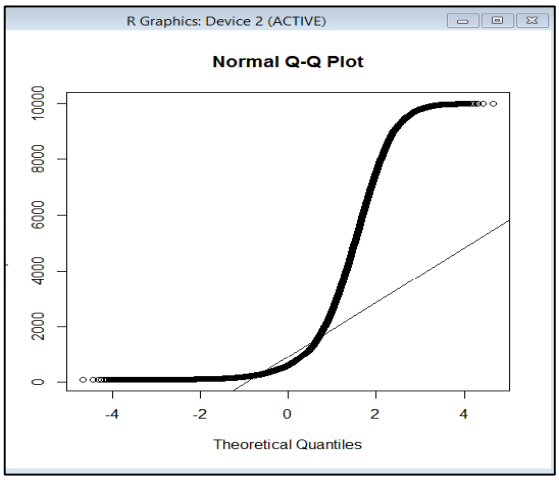

Figure 7
R Balance Distribution (Q-Q plot)

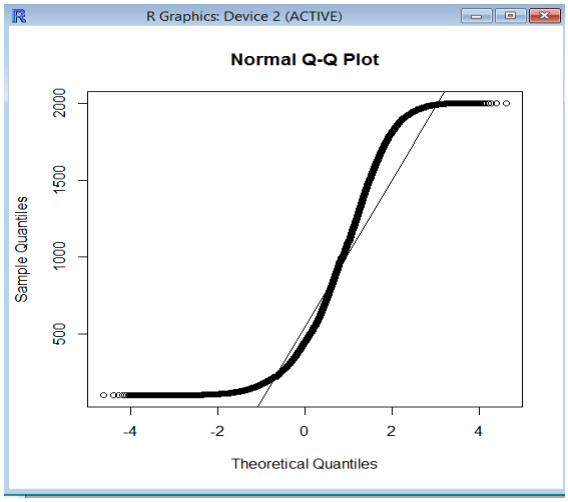

Figure 8
R Balance Distribution v2 (Q-Q plot)

If you look at figure 8, you can appreciate that the volume of data seems more concentrated and normalized.

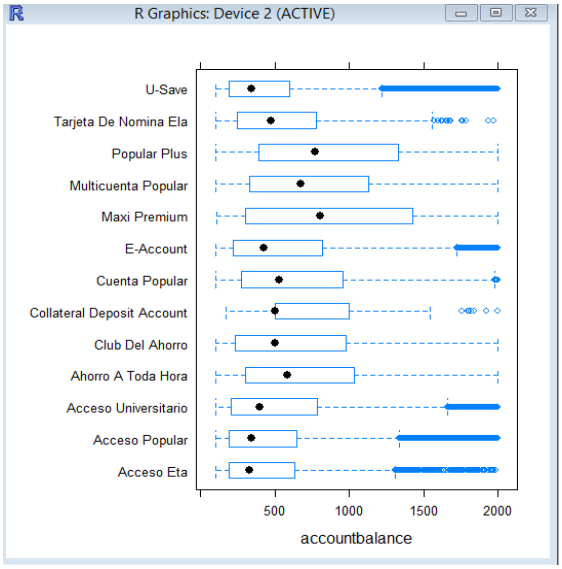

Figure 9
R Balance Distribution by Account Type (BW plot)

Figure 9 shows another distribution, showing the tendencies of customer's balances by account type. I could not help but notice that customers with more balances tend to go for Maxi Premium, Popular Plus and Multicuenta; meanwhile, customers with less balances tend to go for U-Save and Acceso Popular.

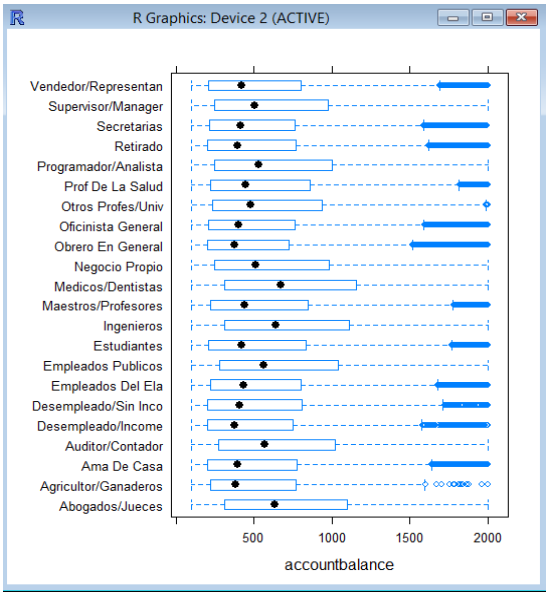

Figure 10
R Balance Distribution by Occupation (BW plot)

Figure 10 shows distribution by occupation. The maximum balances and wider ranges of balances seem to be of doctors (medics and dentists), and engineers; however, minimum balances seem to be for farmers, unemployed and construction workers. Surprisingly, students and teachers have almost the same distribution.

WEKA APPLICATION

Weka is an application used for statistical analysis. This free tool is managed by the University of Waikato. I chose this tool because it is very easy to use and supports a wide array of data formats. The limit to process data depends on the limit of your computer's RAM memory.

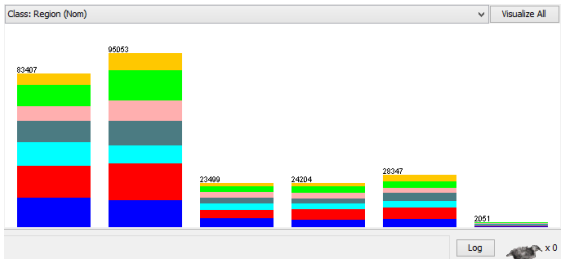

Figure 11
Weka Quantity of Accounts by Region by Type

Continuing the steps for the data analysis, the data was converted to ARFF format and imported to Weka. The first visualization that I decided to do is a plot for quantity of accounts by region by type of education, given that R was giving me some difficulties when managing multiple classifications (see figure 11).

Each column represents the type of education and the colors represent the regions. Education columns (from left to right) represent the following values: College Graduate; High School; Post Graduate; Grade School; Associate Degree; Elementary School. The colors for regions are:

- Blue – Rio Piedras
- Red – Region Norte
- Turquoise – San Juan
- Gray – Ponce
- Pink – Caguas
- Green – Oeste

- Yellow – Central Este

Region Norte and Rio Piedras seem to be very dominant in quantity of accounts for all types of education, with the highest being College Graduate and High School. Most probably these are key locations for students. But perhaps more attention needs to be paid to those customers with higher education.

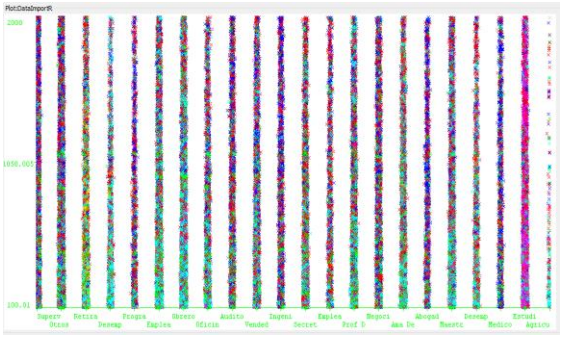

Figure 12
Weka Balance Distribution by Region by Type

All customers seem to be going for the same type of account, regardless of their occupation, except for the students because the majority has "Acceso Universitario". However, there seems to be variety of account types in Engineers, Medics, Auditors and Supervisors. These also have the highest balances in their accounts.

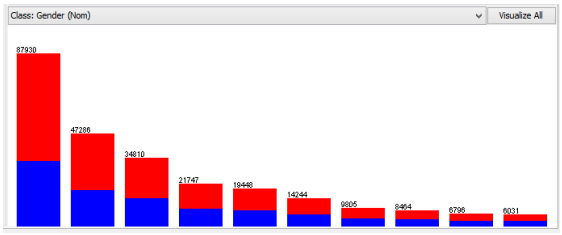

Figure 13
Weka Discretized Balance by Gender

After discretizing the data using the Unsupervised Discretize function my results changed. The field "Balance" assumed the following ranges:

- (0 – 290) - 87,930 records.
- (290 – 480) – 47,286 records.
- (480 – 670) – 34,810 records.
- (670 – 860) – 21,747 records.
- (860 – 1050) – 19,448 records.
- (1050 – 1240) – 14,244 records.
- (1240 – 1430) – 9,805 records.
- (1430 – 1620) – 8,464 records.
- (1620 – 1810) – 6,796 records.
- (1810 – 2000) – 6,031 records.

The red color represents females and blue color represents males. As you can see, females are at the top with a range of 0-290 balance. But as the range of balances are decreasing, the distribution in gender is more even. The fact that there are few accounts with high balances may suggest that we need to attract more customers with capacity to maintain high balance in their checking and savings accounts.

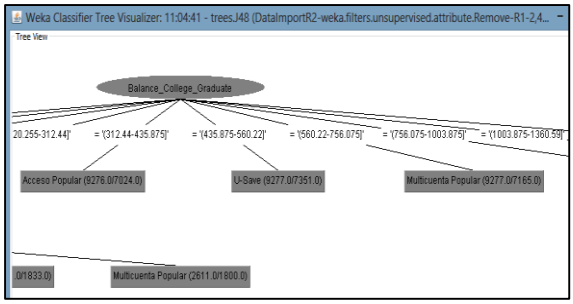

Figure 14
Weka Tree Classifier (Part 1)

The first part of the tree shown in figure 14 shows that if the customer is a "College Graduate" and handles a balance range of $560 or more, he is more likely to go for the "Multicuenta Popular" account. However, balances between $435 and $560 are more likely to have a "U-Save", and balances less than $435 will have "Acceso Popular".

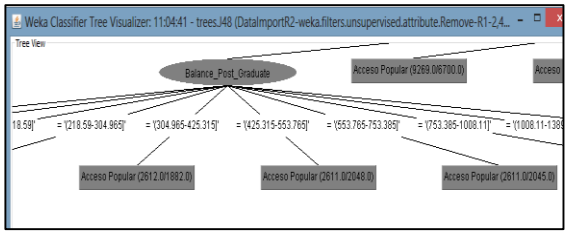

Figure 15
Weka Tree Classifier (Part 2)

For the "Post Graduate" shown in figure 15, the predictions are very similar, with the difference that "U-Save" account is not considered.

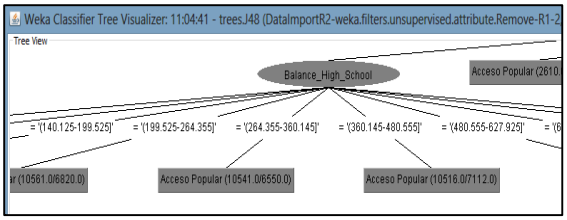

Figure 16
Weka Tree Classifier (Part 3)

For the High School education, the tree shows that customers with highest balances will have "Acceso Popular" and "Ahorro A Toda Hora".

RECOMMENDATIONS

My most important recommendation is to create a new type of account that attracts customers that handle $1,000+ balance. This account could have an interest rate greater than the "Multicuenta Popular" and tailored to the customers that have high education, for example, Post Graduate and College Graduate. The account should have a checking section (similar to Acceso Popular). Another recommendation is to make a marketing campaign targeted towards professionals located in the regions: Caguas, Central-Este and San Juan. The goal of this campaign is to attract more customers with high balances that will increase the customer portfolio for those regions. My last recommendation is to create a savings account with higher interest rate, able to compete with local credit unions, with the goal of making customers with low income, such as, students, construction workers, unemployed and farmers, to make bigger deposits. With this strategy the customers will less likely withdraw their money hoping to gain interests.

CONCLUSION

An essential part about prospecting clients is to become knowledgeable about them using data. All of IT-oriented companies should have a Data Mining concept in place due to the emerging necessity of Data Analysis. The key factor is to have a team with technical skills working together to plan a good solution. Sometimes in big companies the efforts are scattered and you end up with many databases in different areas to analyze the same thing. This also happens with software tools. In the end, if you do not plan, you spend more, and are less efficient.

One of the important things in Data Mining is to keep an open mind about what the data is trying to show. This might not make sense to us but one attribute can be associated with another attribute and at the same time, not have anything to do with one another. We have to use our data to try to know our customer, know their tendencies, and classify them using analysis tools. When you put yourself in your customer's shoes, some things are obvious, but some are not. Customers want to have and use products that satisfy their necessities.

In my opinion, following my recommendations is only a start to improve the country's economy, as well as the bank's reputation. The competitiveness is already high between banks, but what is important is to be competitive, offering products that work best to our customers based on their profile and tendencies.

REFERENCES

[1] PAT. "What is Data Mining", 2014. Retrieved on September 20, 2014 from http://www.predictiveanalyticstoday.com/what-is-data-mining/.

[2] Sayad, S., "An Introduction to Data Mining", 2014. Retrieved on October 1, 2014 from http://www.saedsayad.com.

[3] Tan, Pang-Ning, *et al.*, "Classification: Basic Concepts, Decision Trees, and Model Evaluation", *Introduction to Data Mining*, March 25, 2006.