Universiteit Leiden
The Netherlands

Computer Science & Economics

Analytics using OLAP and Data Mining Techniques

for Transactional Data.

Mitch Bos

Supervisors:
Dr. rer. nat. Michael Emmerich & Kaifeng Jang

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 24, 2017

Abstract

This thesis explores different analysis techniques for transactional data on two data sets. The first data set is an open data set from an online retail warehouse. The second data set is a non-disclosed data set from a telecommunication company. In particular, OLAP and association rule mining are used and it is investigated what insights can be found by these techniques and how they should be implemented on transactional data. Based on these insights a pricing scheme is suggested that can be used by the telecommunication company to set prices based on historical data.

Acknowledgements

I would like to thank Michael Emmerich for his support and advise on the different subjects and techniques in this paper. Also I would like to thank TeleFuture for their data and their internship. In particular, I would like to thank Armand Krijgsman for his help and support at TeleFuture.

Contents

Part I

General

Chapter 1

Introduction

Firstly a background of the subject will be given, followed by an outline of the structure of the thesis.

1.1 Background

Data Mining is becoming increasingly popular in businesses from every sector. In order to obtain and keep a competitive advantage over their competitors, it is becoming increasingly important to not only store their customer information, but also analyze it. The amount of data that companies receive from their customers keeps growing by the day. But the main problem for most businesses is that they do not know how the exploit this data in order to increase revenue, target a specific customer group or even detect fraudulent transactions.

This thesis will consist of two cases of transaction data sets. By using different OLAP and Data Mining techniques this thesis will explore and exploit the large data and try to find patterns in the data. Which can provide interesting insight and opportunities for service improvements.

The main research question is: **Given large data volumes from past transactions: are there patterns in the transaction data that can be used to determine pricing schemes or other strategies for clients?**

1.2 Thesis Outline

Since the research has been done during an internship at TeleFuture and the information from the data set needs to be kept secret, the decision has been made to write the thesis in three different parts The first part will be the general part. The second part will be about the analysis of the TeleFuture data set, which will only be available for the company and the supervisors who grade the thesis. The third part will be about the analysis of the transaction data from an unnamed online retail company. The theory will be the same but the results can be different. The general outline of the different theses will also be the same. First, there will be a general part of the thesis. In Chapter 2 the theory will be explained in relation to the used techniques. In Chapter 3 the tools used to perform the analyses will be discussed. Lastly for this part, in Chapter 4, the used Data will be discussed and how it is modified to be able to perform analyses on it. Secondly, the thesis will be divided into two cases with the same basic outline. First, the results of the research will be explained. Secondly, the conclusion will conclude the parts and take a look at future work.

Chapter 2

Theory

In this chapter, the techniques used to analyze the data will be discussed and the research already done in these fields.

2.1 OLAP

Online Analytical Processing, OLAP in short, is a basic approach to answering analytical queries swiftly [4]. It is part of the Business Intelligence subject, and also falls under the subjects of relational databases, reporting, and data mining. An OLAP system typically involves around an OLAP cube. OLAP applications basically consist of four operations: *roll-up*, *drill-down* and *slice* and *dice* [6].

2.1.1 OLAP Cube

An OLAP cube is a multidimensional cube that consists of numerical *measures* which are categorized by *dimensions*. The data from the cube is typically stored in a *fact table* and different *dimension tables*. The fact table consists of the useful and aggregated data extracted from the raw database. These will be the measures in the OLAP cube. In the dimension tables, the information about the different labels of the data in the fact table is stored. The dimensions of the cube are extracted from these tables. In Figure 1 an example of such an OLAP cube is given. As seen in the Figure, the different dimension are on the sides of the cube. These dimensions describe the different facts. The facts are shown in the OLAP cube as the smaller cubes, or in other words, the *fact cells*.

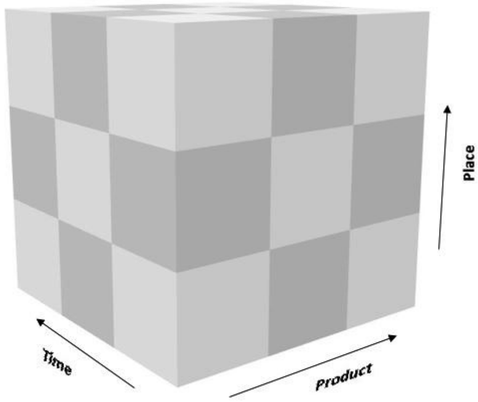


Figure 1: Example of an OLAP cube

2.1.2 Roll-up

The Roll-up operations are the aggregations that are executed on the data. The roll-up operation lets the user ascend a step in the hierarchy of the dimensions. For instance, first, the data is grouped by city. After a roll-up the data is grouped by country. In Figure 2 the principle is visualized.

2.1.3 Drill-down

The drill-down operation is exactly the opposite. This operation lets the user view the data more specialized. This increases the number of dimensions. For instance first, the data is aggregated by quarter. After using the drill-down operation, the information is aggregated per month. In Figure 2 this operation is also shown.

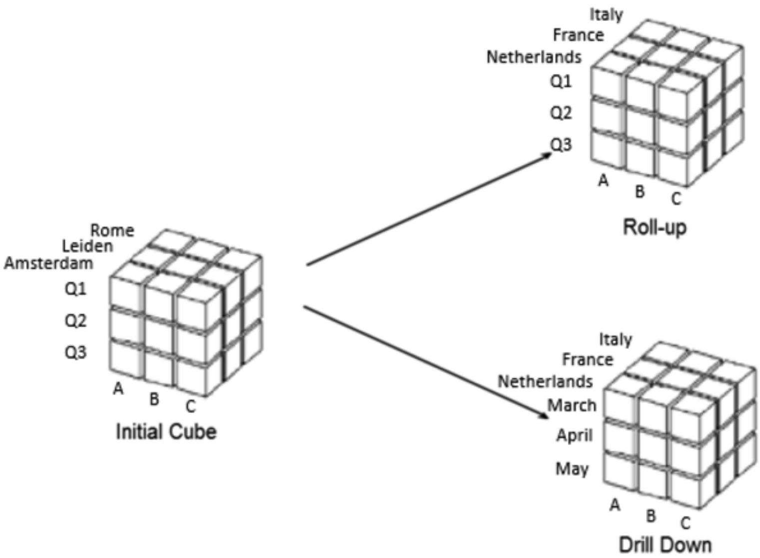


Figure 2: The Roll-up and Drill-down operations

2.1.4 Slice

Slicing is an operation that lets the user specify one specific dimension. Then it provides the user with a sub cube with the information from this specific dimension. For instance, the user can select a specific country and then the slice operation shows all the information from the initial cube that corresponds with this specific country. This operation is shown in Figure 3.

2.1.5 Dice

The dicing operation is almost the same as the slicing operation but for dicing, multiple dimension are chosen and then a sub cube is created that can be viewed from multiple perspectives by the user. For instance, the user wants to see all the information from two specific countries during two specific months. The dice operation will then return a sub cube of the initial OLAP cube with only the given dimensions. This principle is also shown in Figure 3.

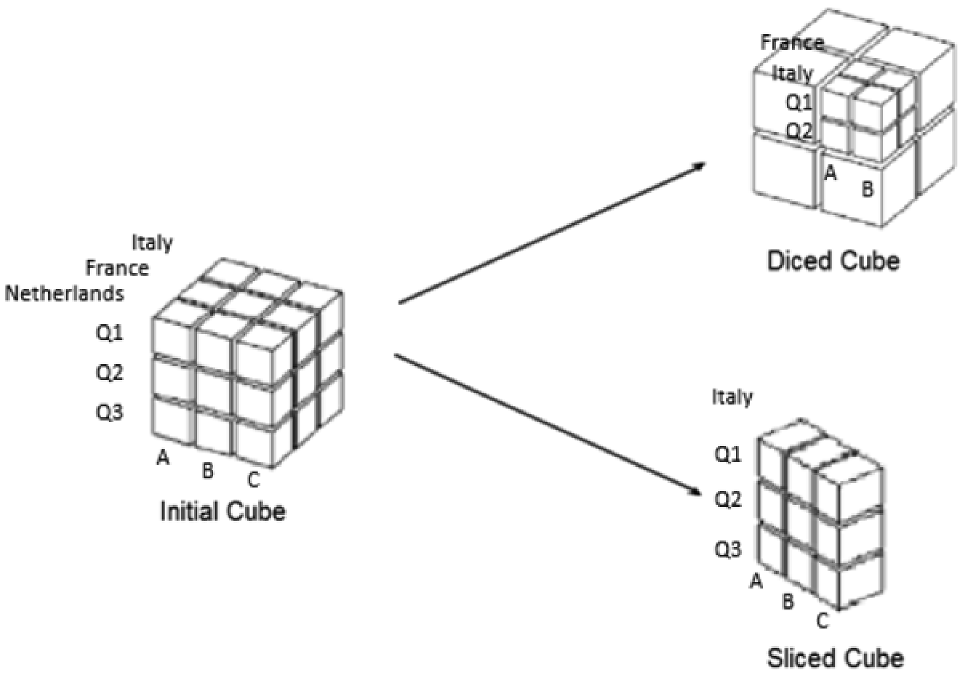


Figure 3: The Dice and Slice Operations

2.2 Association Rule Mining

During this research the analysis method of Association Rule Mining will be used. The concept of Associated Rule Mining was first introduced in [1] for discovering patterns between different products in a large-scale data set of transactions. The most known example is the rule: $\{onions, potatoes\} \implies \{burgers\}$. This rule indicates that if a customer buys onions and potatoes, it is also likely that the customer buys hamburger meat. This information can be used for different marketing and pricing schemes.

In [1] a formal model is given as follows:

Given is $I = \{\iota_1, \iota_2, \ldots, \iota_m\}$ a set of binary attributes called items and $D = \{t_1, t_2, \ldots, t_n\}$ a database with n transactions t_i and each transaction being a set of m items. With every transaction containing a set of items which take the binary value 1 if this item was purchased and 0 if it was not.

An association rule can be defined as an implication of the form $X \implies Y$ with $X \subseteq I$ and $Y = I_k$. In other words, Y is a single item from the set I and is implied by a subset of items X. Y is called the *consequent* (or right-hand-side) and X the *antecedent* (or left-hand-side)

This way many rules can be derived from a data set but not every rule is equally interesting. In order to test the significance of different rules, certain constraints are used. The best-known constraints, also explained in [1], is the *minimum support* and *confidence* constraint. The support is a measurement of how many times a specific itemset appears in the database. In other words, the support of set X, containing items from database T, is the proportion of transactions t that contain X. The support can be formally defined as:
$supp(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$. For example, if the database consists of nine transactions and itemset X can be found three times in the database, then $supp(X) = \frac{3}{9} = \frac{1}{3} \approx 33\%$

The confidence is a measurement of how many times the rule is found to be true. The confidence of a rule $X \implies Y$ can be defined as the proportion of all the transactions in T that contains both X and Y to transactions that contain X. Formally formulated as:
$conf(X \implies Y) = \frac{supp(X \cup Y)}{supp(X)}$. For example from the example above, if the support of $X \cup Y = 0.2$ then the $conf(X \implies Y) = \frac{0.2}{\frac{1}{3}} = 60\%$.

Two more metrics can be defined, *lift* [3] and *conviction* [5]. Lift is the measurement which defines the ratio between the support of the set of items and the support of the item if they were completely independent. Formerly written as: $lift(X \implies Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$. If the lift is > 1, it means that a rule might be useful for predicting the consequent in future data sets.
The last measurement, the conviction can be defined as:
$conv(X \implies Y) = \frac{1 - supp(Y)}{1 - conf(X \implies Y)}$. The downside of confidence is that it does not take the probability of the occurrence of the consequent. So this way a rule that is completely independent of the antecedent can still have a high confidence. For example from [5], if 80% of customers buy milk but the buying of milk is completely unrelated to buying smoked salmon, then the rule *salmon* $\Rightarrow$ *milk* can still hold and so according to the confidence measure, this rule will be a legal rule. By using the conviction definition, these rules will not

hold. The range of conviction is $[1, \infty\rangle$. With 1 indicating independence and ∞ indicating that the consequent and antecedent are completely dependent of each other.

2.2.1 Apriori

The first step of Association Rule Mining is finding frequent item sets and is considered the most difficult. But Agrawal and Srikant came up with an algorithm to find these itemsets fast [2]. In this research, they introduced the Apriori and AprioriTid algorithms. In Figure 4 the Apriori Algorithm is shown.

The Apriori algorithm states that if a certain set is not frequent, then this means that an extension of this set is not frequent as well.

```
1)  L_1 = {large 1-itemsets};
2)  for ( k = 2; L_{k-1} ≠ ∅; k++ ) do begin
3)      C_k = apriori-gen(L_{k-1});  // New candidates – see Section 2.1.1
4)      forall transactions t ∈ D do begin
5)          C_t = subset(C_k, t);  // Candidates contained in t – see Section 2.1.2
6)          forall candidates c ∈ C_t do
7)              c.count++;
8)      end
9)      L_k = {c ∈ C_k | c.count ≥ minsup}
10) end
11) Answer = ⋃_k L_k;
```

Figure 4: Apriori Algorithm [1]

1. Find the frequent itemsets with one item. This is seen on line 1 in Figure 4.
2. Using the itemsets from step 1. Find the frequent itemset with k items. Seen in step 2 until 7 in Figure 4.
3. Keep the itemset, if the support $\geq$ minimum support.
4. Repeat steps 2 and 3 until it is not possible to find new frequent itemsets.

2.2.2 FP Growth Algorithm

A second fast way to find association rules is with the Frequent Pattern Growth Algorithm (FP Growth) introduced in [8]. The FP Growth algorithm first converts the input dataset to an FP Tree structure. An FP Tree structure as defined by Han [8] is shown in Figure 5 and is described as follows:

1. The tree consists of one root labeled *NULL*, a set of item subtrees as the children and a Header Table
2. The subtrees are the items of the frequent items connected with each other to form the frequent sets.
3. The header table which consists of a specific item-name and a link which points to the first node with this item name

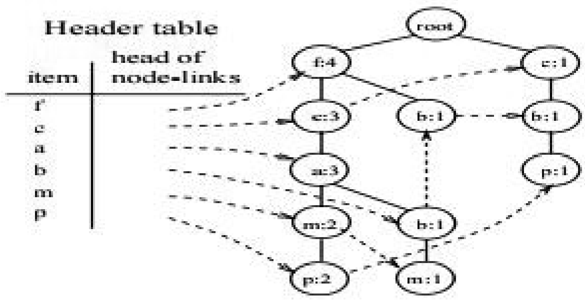


Figure 5: Example of FP-tree construction [8]

The algorithm to create the tree is explained in [8]. After the tree has been created the algorithm for the creation of the association rules can be started. The algorithm as given in [8] is shown in Figure 6.

```
Procedure FP-growth (Tree, α)
{
(1) if Tree contains a single path P
(2) then for each combination (denoted as β)
    of the nodes in the path P do
(3)     generate pattern β ∪ α with support =
            minimum support of nodes in β;
(4) else for each a_i in the header of Tree do {
(5)     generate pattern β = a_i ∪ α with
            support = a_i.support;
(6)     construct β's conditional pattern base and
            then β's conditional FP-tree Tree_β;
(7)     if Tree_β ≠ ∅
(8)     then call FP-growth (Tree_β, β)          }
}
```

Figure 6: FP Growth Algorithm [8]

The algorithm is difficult to interpret. From the constructed FP Tree, the algorithm follows a bottom to top approach. Starting with the node that has the least frequency, we have to find the conditional pattern base for this. By using the tree in Figure 5 we start with node p and then we find the conditional patterns: $fcam$:2 and cb:1. Now we make the frequent list with the nodes in the conditional pattern and then we get c:3, f:2, a:2, m:2, b:1. If we have a minimum support of 3 then only c is the node that meets this requirement. C is added to the frequent item list with a frequency c and we move on to the next node. This is the first step of the algorithm, go through the entire tree and find paths and for all combinations find the different frequent item sets.

Step two starts if we find a path with more nodes that meet the minimum support requirement. If for instance, the minimum support of the tree in Figure 5 was 2, then we have to first set up the conditional FP Tree of node p. This is a just a new FP Tree but then with the item set that contains p. Then from this tree, we start the algorithm again until all the nodes have been processed.

Chapter 3

Software

To give an impression of the way the work has come about, the used software will be explained in this chapter.

3.1 Excel

The software used for the more basic OLAP analysis is Excel. Excel is a spreadsheet program from Microsoft with OLAP analysis tools integrated. By using Excel, it is possible to create a *Pivot Table*. A pivot table is nothing more than an OLAP Cube displayed as a table. In Excel, it is possible to use the slice, dice, roll-up and drill-down operations in order to view the data from different perspectives [9]. Excel is also a nice tool to immediately visualize the derived views in a table or graph for further visual analysis.

3.2 Weka

In order to use the exploratory algorithms that are specified in Chapter 2 the analysis program Weka 3: Data Mining Software in Java will be used. Weka is a collection of different algorithms that can be applied on different data sets. In The WEKA Workbench [7] these different algorithms and the way Weka is used are explained.

Chapter 4

Dataset

The Data of the different data sets are explained in this Chapter. Also, the conversion and modifications that were done on the data in order to be able to conduct analysis are explained.

4.1 Modifying Data

4.1.1 Data Sets

4.1.1.1 TeleFuture

TeleFuture uses two applications to monitor their data. One application to monitor payments and one to monitor the amount of subscribers and views of their services. These applications have been created and implemented by TeleFuture themselves but can not be used in order to perform an analysis. In order to perform a correct analysis, the raw data is needed and should be modified in order to perform the algorithms mentioned in Chapter 2.

For the analysis, different tables have to be linked. The *customers* and *mcb_users* tables are linked with the *payments* table. This is necessary to find all the subscriptions and payments a specific user has made. These are needed for further analyzing the data and finding the relevant association rules.

Because the data set is, after linking all the tables, over 500 million records, it was decided to focus attention on one specific country for the **association analysis**. Furthermore, because the services change a lot, the payments of the earlier years were not used for this analysis.

4.1.1.2 Online Retail Company

The data set of the Online Retail Company was only one table that contains the different invoices customers made. Per item, it is recorded who bought the item, how many were bought, how much it costs, and which country it was bought in. For this Data Set, no tables had to be linked. A part of the data set can found in Appendix A.

This data set was used because of the confidentiality of the data from TeleFuture so this data set is part of the non-confidential part.

4.1.2 OLAP aggregation

In order to use OLAP analysis, first, the Data has to be generalized. For OLAP, we are not looking for customer based analysis but a more generalized analysis. The data is aggregated on the number and revenue of sales per product. This is done for the different time periods and the different countries. In other words, the dimensions of the OLAP Cube are the time (by quarter, month & year), the products and the country. The measures are the number of sales and the revenues. In Figure 7 below an example is shown in a pivot table made in Excel.

Rijlabels	Som van UnitPrice	Som van Quantity
2010		
Kwrt4		
dec	260520.85	342228.00
2011		
Kwrt1		
jan	172752.80	308966.00
feb	127448.77	277989.00
mrt	171486.51	351872.00
Kwrt2		
apr	129164.96	289098.00
mei	190685.46	380391.00
jun	200717.34	341623.00
Kwrt3		
jul	171906.79	391116.00
aug	150385.68	406199.00
sep	199235.21	549817.00
Kwrt4		
okt	263434.09	570532.00
nov	327149.85	740286.00
dec	133915.66	226333.00
Eindtotaal	**2498803.97**	**5176450.00**

Figure 7: Example pivot table made in Excel

The derived OLAP Cube will be suitable for further OLAP analysis. Since it is now possible to choose specific values for the different dimensions and use OLAP operations on the cube.

4.1.3 Association Analysis

For Association Rule Mining, the data set needs to be formatted the right way. The data should be in a so called, *denormalized form*. For every user there will be a new row. For each user then will be noted if he/she bought a certain product or not. An example of such a set can be found in Table 1.

The values in the columns of apple, orange, banana, and peach are binary and specify if the user has bought the item or not. For the TeleFuture data set it works the same way but then the column names are the different

User	Apple	Orange	Banana	Peach
1	0	1	0	1
2	1	1	1	1
3	1	1	0	0
4	0	0	0	1

Table 1: Example Denormalized Data

services provided. And for the Online Retail data set the columns are the different items that are bought. A part of the denormalized data set of TeleFuture as seen in the Weka reader is found in Appendix B

Part II

Online Retail Company Data Set

Chapter 5

Results and Discussion

In this Chapter, the results of the OLAP analysis of the Online Retail Company Data Set will be reported.

5.1 OLAP Analysis

5.1.1 Results

By creating an OLAP Cube of the available data, a more understandable visualization of the data can be found. The first noticeable information that can be interpreted from the aggregated information, is the wide range of different countries in which orders have been placed. In order to analyze the information more efficient, first the data is *diced*, so we only have the information of the Top-10 countries and only the year 2011.

In Figure 8 the percentages per quarter in 2011 of the total of this year are shown.

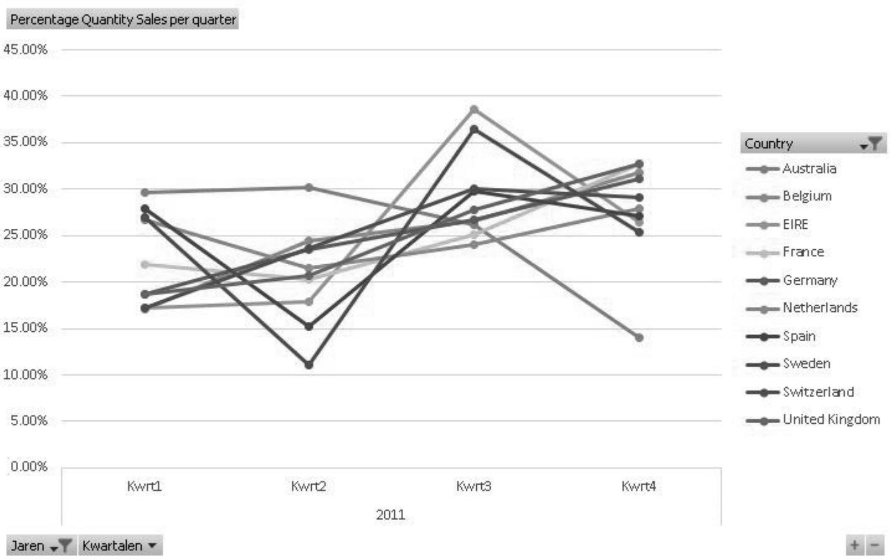


Figure 8: Percentage quantity of sales per quarter of total 2011

It is shown in percentages because the United Kingdom has around 10 times the amount of sales of the rest and this way the growth per quarter can be seen as well.

The data in Figure 8 only shows the number of sales. But when we look at Figure 9, which shows the top-10 countries with highest sales revenue, other noticeable things might stick out.

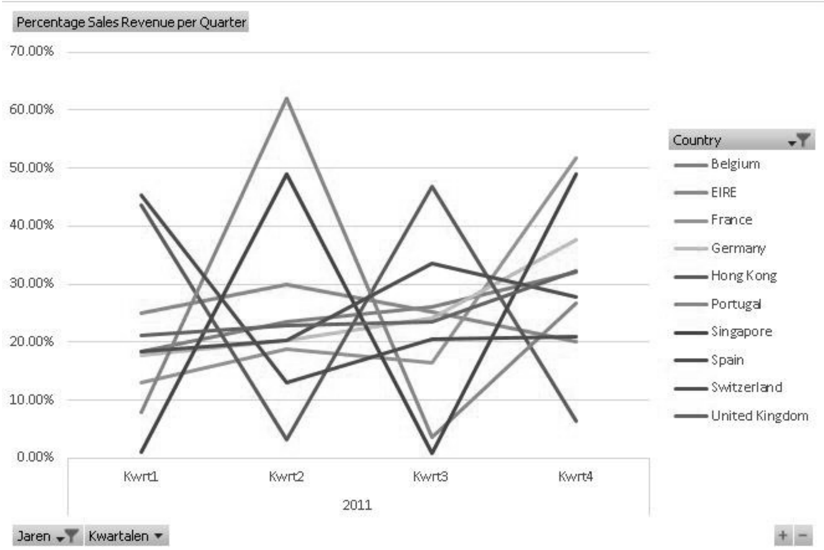


Figure 9: Percentage sales per quarter of total 2011

In order to find more information about the different trends, in Figure 10 the price per sale of the countries found in Figure 9 are shown.

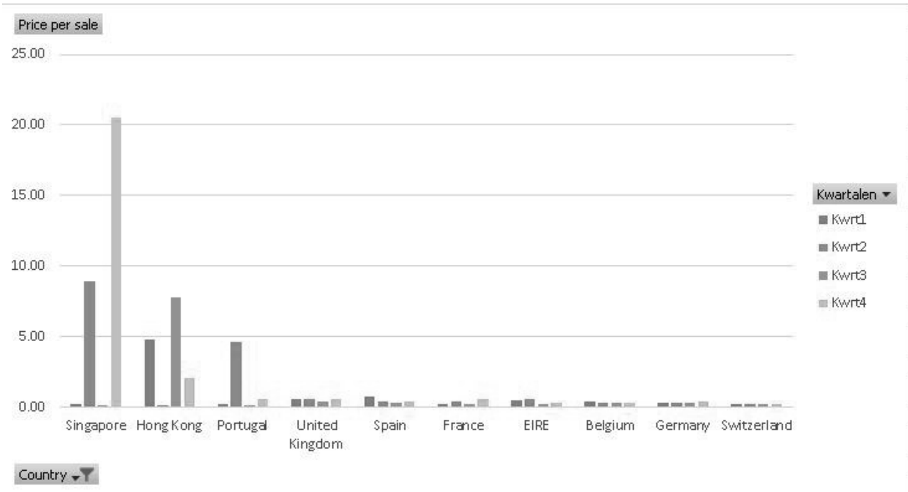


Figure 10: The price per sales

5.1.2 Discussion

Figure 8 shows that for Sweden, Ireland and Germany there was a significant increase in sales during the third quarter of 2011, but it decreased again in the fourth. OLAP does not give an explanation for this so we could only speculate what the reason is for this. There could be an event during this time or a sale which specifically target these three countries. From the data, we can also see that countries as Belgium, France, Netherlands and The United Kingdom have grown steady over 2011 with in increase of around 10% compared to the first quarter.

In the information from Figure 9, we can find that there are different countries in here. New countries are Hong-Kong, Portugal & Singapore which replaced Australia, Netherlands & Sweden. The new countries can not be specified as stable countries. The revenue of sales changes drastically per quarter. This is even more clear in Figure 10. In the figure, these countries show that Singapore, Hong-Kong & Portugal are outliers during some quarters when it comes to the price per sale. A reason for this can't be found but it does mean that, whenever customers from these countries buy something, it is always the more expensive products.

In Figure 11 we focus on these three countries. In the Figure, the three countries are separated and we look at their respective sales revenue and their sales amounts.

From the information in Figure 11, it is still not clear what is happening in these countries. The revenue and quantity of sales do not seem to have a correlation with each other in these countries.

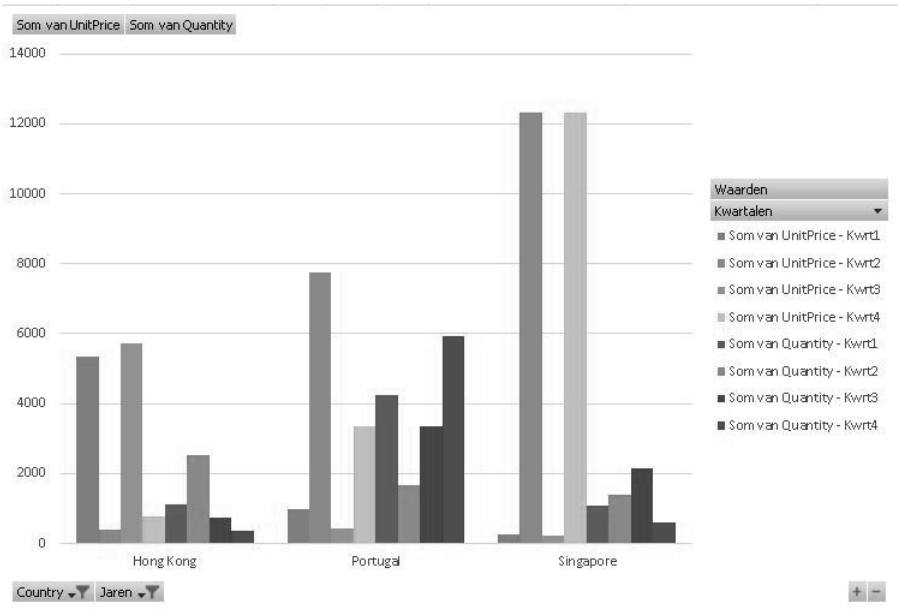


Figure 11: Quantity of sales and revenue

The last thing we can find from the pivot table in excel is that the product with the highest price per sale is a postage for each country.

5.2 Association Analysis

5.2.1 Results

For the Online Retail Company data set association rules were found and shown in Figure 12. The different products are shown as StockCode_id. Pruning was an issue for this data set. A lot of redundant rules were produced that were so-called super rules. These had to be removed and then the Association Rules, shown in Figure 12, were found with their corresponding confidence and lift metrics in Table 2.

1. {StockCode_22698, StockCode_23171} ⇒ {StockCode_22697}
2. {StockCode_22697, StockCode_23170} ⇒ {StockCode_22699}
3. {StockCode_22699, StockCode_22698} ⇒ {StockCode_22697}
4. {StockCode_22697, StockCode_23173} ⇒ {StockCode_22699}
5. {StockCode_22423, StockCode_22698} ⇒ {StockCode_22697}
6. {StockCode_22745, StockCode_22746} ⇒ {StockCode_22748}
7. {StockCode_22699, StockCode_23171} ⇒ {StockCode_23170, StockCode_22697}
8. {StockCode_22386, StockCode_23202} ⇒ {StockCode_85099*B*}
9. {StockCode_22386, StockCode_85099*C*} ⇒ {StockCode_85099*B*}
10. {StockCode_21733, StockCode_22804} ⇒ {StockCode_85123*A*}

Figure 12: Top 10 Rules based on Confidence

RuleNumber	Confidence	Lift
1	0.99	192.56
2	0.98	192.41
3	0.96	188.16
4	0.96	188.64
5	0.95	184.83
6	0.94	405.09
7	0.93	819.56
8	0.93	114.23
9	0.92	113.27
10	0.92	117.21

Table 2: Confidence and Lift Metrics per rule

The corresponding product description can be found in Appendix C. The association rules from Figure 12 are based on the confidence metric. The rules based on conviction will provide the same set of rules. In Figure 13 the Association Rules based on the lift metric are shown with their corresponding confidence and lift in Table 3

1. {StockCode_22697, StockCode_23170} ⇒ {StockCode_22699, StockCode_23171}
2. {StockCode_22697, StockCode_23171} ⇒ {StockCode_22699, StockCode_23170}
3. {StockCode_23173, StockCode_23175} ⇒ {StockCode_23174}
4. {StockCode_23173, StockCode_23174} ⇒ {StockCode_23175}
5. {StockCode_22423, StockCode_23175} ⇒ {StockCode_23174

Figure 13: Top 5 Rules based on Lift

RuleNumber	Confidence	Lift
1	0.91	819.56
2	0.87	646.78
3	0.88	492.02
4	0.91	473.93
5	0.84	471.3

Table 3: Confidence and Lift Metrics per rule

For the lift metric, the direction of the rule does not matter. Only the combination of the items is relevant. We removed all the rules with the same item but in a different direction.

5.2.2 Discussion

When we compare the rules found in Figure 12 and 13 with the descriptions from Appendix C, the relation between the different products becomes clear instantly. The descriptions of the products of the rules show that similar products are bought together. Different kinds of teacups bought together with a specific saucer or different parts of a complete set of a playhouse.

For pricing schemes and marketing strategies, these rules can help to put certain products together for sale or place them on the same shopping page. Because of the fact that the data is based on an online retail company, the different rules can also be used to generate corresponding advertisements.

The lift metrics also support the rules. Since the rules from Figure 12 have a lift above 100 and the rules from Figure 13 even above 400, it means that these specific items occur more often together than expected based on their separate occurrence. Based on the lift of the rules, a lift above 400 tells us that the rule has a confidence of 400 times the expected confidence. This means that the rules are not by random chance.

Chapter 6

Conclusion

This part of the research is the non-confidential part of the research and tried to find an answer to the question: **Is it possible to find patterns in the data from an Online Retail Company to find pricing schemes or other strategies for clients?** First, an OLAP analysis is conducted in order to make the date more understandable and gain a better understanding of general aggregations. Then a more detailed analysis has been conducted, which is an association analysis that makes it possible to find relations between different products and identify buying habits.

The research is based on the open data set of an online retail warehouse which sells all kind of products. The data, however, is limited because it is up until 2011. The results of the analysis will be somewhat outdated but will still give a better understanding of the different kinds of analysis methods.

The OLAP analysis shows that mostly the quantity of sales is growing steadily each quarter. Most countries have a stable revenue growth, with some countries that are unstable over the different quarters. Another interesting observation is that for three countries, Hong-Kong, Singapore, and Portugal, the number of sales was way lower than the revenue return of those sales. The revenue per sales in these countries was in certain quarters extremely high, compared to the other countries.

Looking at the association analysis, buying habits of customers have been found. Customers tend to buy complete collections of certain play toys or tea sets. The measures of the rules indicate that these collections of products are bought together more often than separate.

The warehouse is growing and could, based on the OLAP analysis, focus more on and expand more to the countries other than the United Kingdom. By employing pricing schemes and marketing strategies on the collections of products found in the association rules, they could further improve their revenue and increase their sales volume.

Part III

TeleFuture Data Set

Chapter 7

Results and Discussion

This Chapter will not contain specific data; only keys and graphs. Also because the data that is available from TeleFuture is so large, we focus our attention for the Association Analysis to a specific country, South-Africa. This is the newest and most interesting country in which TeleFuture also has the most difficulties. Furthermore, this Chapter will have an extra section about the model created for TeleFuture to analyze the data and give a prediction for future costs and profits.

7.1 OLAP

7.1.1 Results

In order to get a more clear view of the available data, OLAP analysis is conducted. The data will become more understandable and certain trends can be identified instantly. Please note that the revenues in the tables are in eurocents.

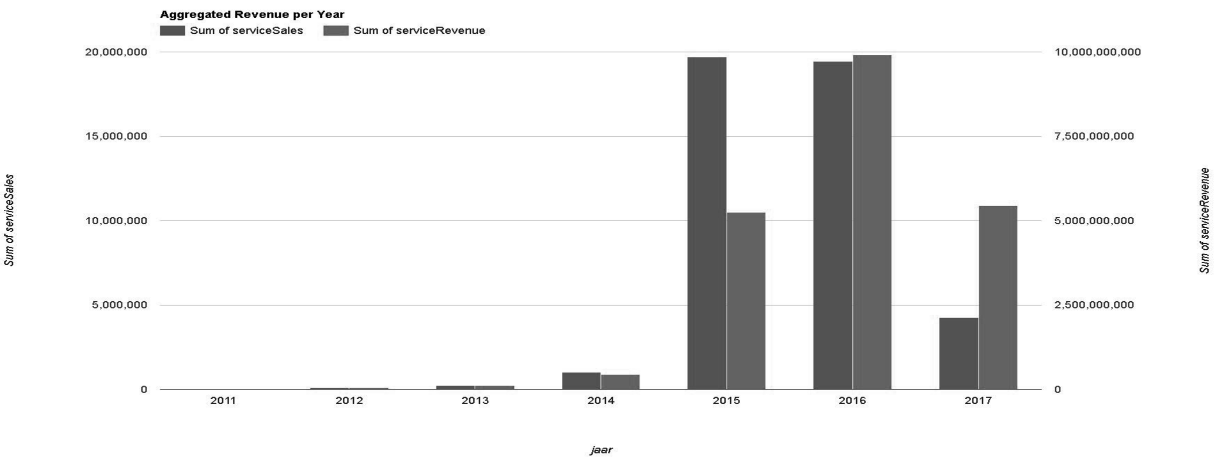


Figure 14: Revenue TeleFuture vs sales per year

Figure 14 shows the revenue versus the number of sales per year. And if we slice the cube in a different way, we can show the revenue per sale. These are found in Figure 15.

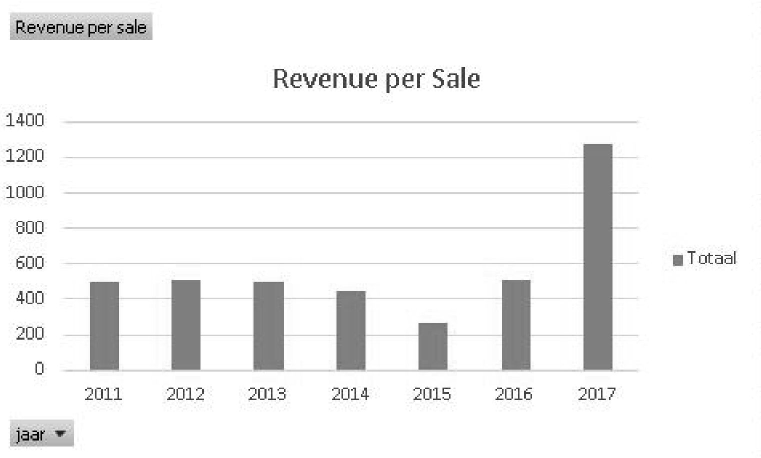


Figure 15: Revenue TeleFuture per sale

The changes in sources that provide the revenue for TeleFuture are provided in Figure 16. It shows the revenue per mobile operator (as MCCMNC number[1]).

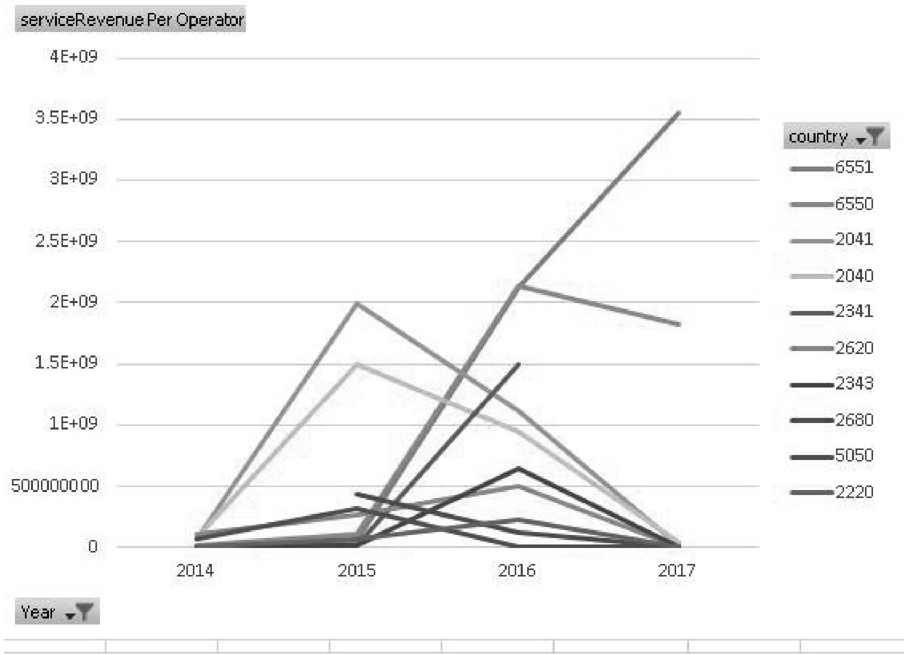


Figure 16: Revenue per operator

[1]MCCMNC is a combined number of Mobile Country Code (MCC) & Mobile Network Code (MNC) which is unique for each Mobile Operator

7.1.2 Discussion

By looking at the data obtained from the OLAP Cube, it can be seen in Figure 14 that the sales revenue is growing fast every year. Even with 2017 only just been started[2], TeleFuture has now already achieved more than 50% of the revenue of 2016.

But a second thing can be noticed, in particular more recently from 2016 and 2017. The amount of revenue per sale has grown. This means that every new subscriber of a service becomes more valuable. In other words, customers cancel their subscriptions on services provided by TeleFuture later.

This is also supported by Figure 15. We can study this further and more clearly. The data shows that there was a decline in revenue per sale but that since 2016 TeleFuture has improved their revenue per sale again. And it is now at almost € 14,00. As seen in Figure 16, the main source of income in 2015 were operator 2041 and operator 2040 which are operators in the Netherlands. This changed drastically since South Africa (operator 6550 and 6551) has made an impressive growth. While only having started in 2015, South-Africa already has a revenue of over €$3,5 \times 10^7$ in just a few years.

In addition Figure 16 it is clearly visible that TeleFuture stops and starts connections with operators suddenly which is typical for the kind of business.

7.2 Association Rule Analysis

7.2.1 Results

As said above, for this research the focus was placed on South-Africa. Firstly Association Rules were found for the different services provided. The different attributes are of the form service_id_*n* with *n* as the unique identifier. The full top 40 association rules are shown in Appendix D. In Figure 17 the top 10 of the rules is shown based on the confidence of the rules. Their measures are shown in Table 4.

1. {service_id_2441, service_id_2189, service_id_1777, service_id_1613, service_id_1969} ⇒ {service_id_1649}
2. {service_id_2189, service_id_1777, service_id_1613, service_id_1969} ⇒ {service_id_1649}
3. {service_id_2441, service_id_2189, service_id_1777, service_id_1613, service_id_1997} ⇒ {service_id_1649}
4. {service_id_2189, service_id_1777, service_id_1613, service_id_1997} ⇒ {service_id_1649}
5. {service_id_2441, service_id_2189, service_id_1777, service_id_1969} ⇒ {service_id_1649}
6. {service_id_2189, service_id_1969, service_id_1997} ⇒ {service_id_1649}
7. {service_id_2441, service_id_2189, service_id_1777, service_id_1997} ⇒ {service_id_1649}
8. {service_id_2189, service_id_1841, service_id_2421} ⇒ {service_id_1649}
9. {service_id_2441, service_id_2189, service_id_1613, service_id_1997} ⇒ {service_id_1649}
10. {service_id_2441, service_id_2189, service_id_1613, service_id_1969} ⇒ {service_id_1649}

Figure 17: Top 10 Association Rules by Confidence Metric

The entire top 10 rules lead to the same *consequent* with almost identical *antecedents*.

[2] Used data from March 2017

RuleNumber	Confidence	Lift	Conviction
1	0.99	1.94	50.21
2	0.99	1.94	41.39
3	0.9	1.94	37.31
4	0.99	1.93	34.73
5	0.99	1.93	34.82
6	0.99	1.93	30
7	0.98	1.93	27.7
8	0.98	1.92	27.54
9	0.98	1.92	25.7
10	0.98	1.92	25.34

Table 4: Confidence, Lift and Conviction Metrics per rule

But the information from Figure 17 and Appendix D do show that the rules can be used for fraud detection. Since the rules are so identical and because of the kind of services the company supplies, it is possible to identify the services that are more susceptible to fraud. In Figure 18 the top 5 rules are shown based on the *lift* metric of the rules. The lift measurement tells us more about the dependency of the items on each other. Tabel 5 shows that the support for these 5 rules is a factor 100 times higher than the support if these items were independent.

1. {service_id_2189, service_id_1997} ⇒ {service_id_1649, service_id_2441, service_id_1777, service_id_1613}
2. {service_id_1837, service_id_2317} ⇒ {service_id_1841, service_id_2525}
3. {service_id_1649, service_id_2189, service_id_1997} ⇒ {service_id_2441, service_id_1777, service_id_1613}
4. {service_id_2441, service_id_2189, service_id_1997} ⇒ {service_id_1649, service_id_1777, service_id_1613}
5. {service_id_2441, service_id_2317} ⇒ {service_id_2525, service_id_1837}

Figure 18: Top 5 Association Rules by Lift Metric

RuleNumber	Confidence	Lift	Conviction
1	0.48	100.44	1.92
2	0.72	99.29	3.48
3	0.51	98.61	2.01
4	0.6	96.64	2.47
5	0.6	96.43	2.48

Table 5: Confidence, Lift and Conviction Metrics per rule

Because of the fraud, it is not possible to analyze these rules further like this. In order to find more useful rules, the service with id 1649 has been deleted from the instances list. This was the consequent for most of the rules. If we delete this instance we find 7 different rules which might prove more useful for analysis. The rules are found in Figure 19 with their corresponding measurements in Table 6.

1. {service_id_1841, service_id_2421} ⇒ {service_id_2189}
2. {service_id_2441, service_id_1837, service_id_2317} ⇒ {service_id_2525}
3. {service_id_2189, service_id_2317} ⇒ {service_id_1841}
4. {service_id_2189, service_id_1613, service_id_1997} ⇒ {service_id_2441}
5. {service_id_1613, service_id_1969, service_id_1997} ⇒ {service_id_2441}
6. {service_id_2441, service_id_2525, service_id_2317} ⇒ {service_id_1837}
7. {service_id_2189, service_id_1777, service_id_1613, service_id_1969} ⇒ {service_id_2441}

Figure 19: 7 rules

RuleNumber	Confidence	Lift	Conviction
1	0.96	5.99	23.28
2	0.96	23.1	21.74
3	0.94	9.9	15.46
4	0.93	4.6	11.26
5	0.93	4.61	11.75
6	0.91	25.2	11.01
7	0.91	4.5	8.71

Table 6: Confidence, Lift and Conviction Metrics per rule

Lastly, also rules based on the metric *conviction* were produced. The rules based on conviction were the same rules as those produced before by confidence so no new insights were obtained from these.

7.2.2 Discussion

The first set of association rules contained many fraudulent services. The way the fraud works is as follows. A user sees an advertisement of a specific service, a fraudulent one, on a website or application. When this user subscribes to this service, he or she automatically subscribes to a number of other services without their knowledge. Normally a subscriber has to verify and confirm each subscription separately but these services skip this process of verification. Normally a higher lift would mean that the items have a strong dependency on each other. But because of the way the services are provided and because we know how the fraud works, it can be stated that these rules are almost certainly related to fraud. So when we know this it is clear that these fraudulent methods affect the association rule sets and will not give a clear result set for pricing schemes and marketing strategies.

By analyzing the rules found, the different fraudulent services could be identified and terminated. The rules from Figures 17 and 18 lead to the following list of services which have a good chance to be involved in fraud.

- service_id_2441
- service_id_2189
- service_id_1777
- service_id_1613
- service_id_1969
- service_id_1649
- service_id_1997

Another support and check for the fact that these services are involved in fraud is that 81% of the subscribers to any of these services never completed a single payment.

It was not possible to retrieve and interpret the kind of services available. Because of this, it is not possible to categorize and interpret the different rules found in Figure 19 in a way that these could be used for pricing schemes.

However, it is possible to draw some conclusions based on the different metrics of the rules. The confidence of the different rules tells us how often a rule is found to be true. Since the confidence of all the rules, based on the confidence measure, is 0.9 or higher, it is safe to state that the rules that were found are indeed correct rules and the services have a relationship with each other. This can be because of fraudulent transactions or because they have a positive effect on one another.

The lift is the metric which informs us about the expected support for a rule. The rules found all have a lift metric of > 1 which means that the rules, even the ones with fraudulent services, might provide us with enough information to make a prediction on a future data set with the same services. When the lift is > 1, which is the case, it tells us that the occurrence of these combined services is higher than expected based on their separate occurrences. So this means that the rules are not produced by random chance and there is definitely a relation between the different services.

Although the lift of the rules seen in Figure 18 is really high, the confidence of the top rules based on the lift is below 0.8. This means that these rules are less than 80% of the times correct. So these rules do have a higher support together than expected if they would have been independent but the rules themselves are not accurate enough to use for future prediction.

The conviction of the different rules also gives us more information about the rules. Since the rules based on this metric are the same as the rules based on the confidence measure, and the conviction of these rules is between 8 and 24, this tells us that the rules will hold and also that the consequent and the antecedent are not independent of one another.

7.3 Analysis Tool

The main difference between the research part on behalf of TeleFuture and the University was that a *future prediction model* had to be created for TeleFuture. This model should help them rate a specific country and predict if and when a profit is made in this country.

7.3.1 Theoretical Model

The idea of the model is to give the user a set of parameters that can be filled in. The model will then calculate among other things, the costs, revenues, the number of subscriptions and number of payments made. The theoretical model is the blueprint for the software implementation of the analysis tool. The theoretical model is made in Microsoft Excel and the different parameters can be found in Appendix E.

7.3.1.1 Billing Succes

By using historic data in the model of the amount of failed and successful payments, a specific billing strategy can be found. In Figure 7 per month, the best day to initialize a payment is shown. There should be noted that the month July and August are not shown in the figure. This is because the services were stopped during this period of time. The first thing that stands out, is the fact that the successful billing ratio is going down by the month. There are a number of reasons for this.

Ratio	Day	Month	Year
26.7425	13	1	2016
24.9511	1	2	2016
28.3594	4	3	2016
21.4754	22	4	2016
36.8595	2	5	2016
22.9549	1	6	2016
5.2960	17	9	2016
6.6328	20	10	2016
9.2992	18	11	2016
7.9036	1	12	2016
6.3371	2	1	2017
6.3923	1	2	2017

Table 7: Per month best payment day

The first reason is that TeleFuture decided to start billing every subscriber, even if the first attempt of payment did not succeed. In other words, when previously the first ever tried payment failed, the subscriber would not be acknowledged as a new member, so future payments would not be attempted anymore. In later months, future payment tries were attempted.

The second reason for this phenomenon is that the number of payments per day has been increased. When a payment fails, a try is done 6 hours later again. The chances are that, whenever a payment fails once, it will also fail a second and third time on the same day.

It could be stated that by analyzing the payment success ratio, it is possible to find the periods per month to attempt a payment. Most of the months it is clear that the most successful period is at the start of each month. This can be concluded from Figure 7. But this is only true for the months up until June 2016. After the small break of providing the services in South-Africa, the ratio´s do not provide enough information anymore. The number of failed payments is too large and the difference in ratio per day is not significant enough to come to a clear result. The reasons for this are the same as mentioned above for the decline in ratio.

7.3.2 Implemented Software

After completing the theoretical model an *analysis tool* was implemented. This tool provides the user with a prediction of future profits and costs. The user fills in the parameters and the program calculates per day: the number of new members, the number of members that quit, the number of successfully billed members and the revenue and costs per category. The layout of the program is found in Appendix F.

The relevant information for the analysis of the payment success ratio had to be retrieved from the database and calculated by hand. In order to make this process more streamlined and efficient, a program was written and implemented in the tool to calculate for a given starting day, the percentage of subscribers that have quit and the ratio of payments that were successful. The predictions that are computed by the analysis tool are based on historic data. This historic data could be of a country in which TeleFuture is already situated and shares similarities with the country that the user wants to analyze. For instance, if the user wants to look at a prediction for Ireland, he can use historic data from England. It is also possible of course to make a prediction based on historic data of the country to be analyzed, if available. Based on this data, and the input given in by the user, predictions are made. These prediction are supported by the different OLAP aggregations and data mining techniques.

The information from this program in combination with the analysis tool gives the user the means to predict the future status of a specific country and also analyze how related countries will succeed. An example result table is found in Appendix G.

Chapter 8

Conclusions

This part of the research tried to find an answer to the question: **Is it possible to find patterns in the data to find pricing schemes or other strategies for clients of TeleFuture?** First, an OLAP analysis was conducted in order to understand the data and visualize it. Secondly, an association Analysis was conducted to find more detailed relations and rules in the data that could help answer this question. Lastly, a predictive analytical software was implemented to help TeleFuture in the future to forecast the success in specific countries and markets based on historic data.

The research had some limits. The association analysis is conducted on a case of one specific country, South-Africa. The rules found for the services in that country might be different from other countries. Furthermore, it was difficult to have consistent data. One month, there were running 100 services in a specific country and the next month there were none. This was especially a limitation for the OLAP analysis.

The OLAP analysis showed that TeleFuture its revenue is growing steadily. At the start of 2017, it had already reached 50% of the revenue of entire 2016. It also showed that the revenue per sale is growing which is a good thing since TeleFuture aims to have long lasting subscribers for their services. Different countries were separated and so the main sources of revenue gain were identified by the OLAP.

The main result from the Association Analysis was that it did not return the preferred relations between provided services. The rules did result in suspicious services that could be fraudulent. When these services were removed from the data set, other rules were found that could provide more information about customer behavior. Based on the corresponding metrics of the rules, these rules prove useful for further testing and analysis.

Based on the conducted research, TeleFuture is growing steadily and should, based on the OLAP Analysis, continue to distribute attention across multiple different countries. Also, they could, based on the association rules found, identify and put warnings on or eliminate the fraudulent services to improve billing success rates and limit future costs. Lastly, TeleFuture could, by analyzing the rules further, improve their customer targeting and marketing.

Bibliography

[1] Agrawal, R., Imieliński, T., and Swami, A. Mining association rules between sets of items in large databases. In *Acm sigmod record* (1993), vol. 22, ACM, pp. 207–216.

[2] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[3] Berry, M. J., and Linoff, G. *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[4] Berson, A., and Smith, S. J. *Data warehousing, data mining, and OLAP*. McGraw-Hill, Inc., 1997.

[5] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record* (1997), vol. 26, ACM, pp. 255–264.

[6] Chaudhuri, S., and Dayal, U. An overview of data warehousing and olap technology. *ACM Sigmod record 26*, 1 (1997), 65–74.

[7] Frank, E., Hall, M. A., and Witten, I. H. *The Weka Workbench*. Morgan Kaufmann, 2016.

[8] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In *ACM Sigmod Record* (2000), vol. 29, ACM, pp. 1–12.

[9] Microsoft. Create pivottables. `http://support.office.com/en-gb/article/Create-a-PivotTable-to-analyze-worksheet-data`.

Appendices

Appendix A

Online Retail Company Data Set

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1-12-2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	1-12-2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1-12-2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1-12-2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1-12-2010 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	1-12-2010 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	1-12-2010 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	1-12-2010 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	1-12-2010 8:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	1-12-2010 8:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	1-12-2010 8:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	1-12-2010 8:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	1-12-2010 8:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MUG COSY	6	1-12-2010 8:34	1.65	13047	United Kingdom
536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	1-12-2010 8:34	4.25	13047	United Kingdom
536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	1-12-2010 8:34	4.95	13047	United Kingdom
536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	1-12-2010 8:34	9.95	13047	United Kingdom
536367	21754	HOME BUILDING BLOCK WORD	3	1-12-2010 8:34	5.95	13047	United Kingdom
536367	21755	LOVE BUILDING BLOCK WORD	3	1-12-2010 8:34	5.95	13047	United Kingdom
536367	21777	RECIPE BOX WITH METAL HEART	4	1-12-2010 8:34	7.95	13047	United Kingdom
536367	48187	DOORMAT NEW ENGLAND	4	1-12-2010 8:34	7.95	13047	United Kingdom
536368	22960	JAM MAKING SET WITH JARS	6	1-12-2010 8:34	4.25	13047	United Kingdom
536368	22913	RED COAT RACK PARIS FASHION	3	1-12-2010 8:34	4.95	13047	United Kingdom
536368	22912	YELLOW COAT RACK PARIS FASHION	3	1-12-2010 8:34	4.95	13047	United Kingdom
536368	22914	BLUE COAT RACK PARIS FASHION	3	1-12-2010 8:34	4.95	13047	United Kingdom
536369	21756	BATH BUILDING BLOCK WORD	3	1-12-2010 8:35	5.95	13047	United Kingdom
536370	22728	ALARM CLOCK BAKELIKE PINK	24	1-12-2010 8:45	3.75	12583	France
536370	22727	ALARM CLOCK BAKELIKE RED	24	1-12-2010 8:45	3.75	12583	France
536370	22726	ALARM CLOCK BAKELIKE GREEN	12	1-12-2010 8:45	3.75	12583	France
536370	21724	PANDA AND BUNNIES STICKER SHEET	12	1-12-2010 8:45	0.85	12583	France
536370	21883	STARS GIFT TAPE	24	1-12-2010 8:45	0.65	12583	France
536370	10002	INFLATABLE POLITICAL GLOBE	48	1-12-2010 8:45	0.85	12583	France
536370	21791	VINTAGE HEADS AND TAILS CARD GAME	24	1-12-2010 8:45	1.25	12583	France
536370	21035	SET/2 RED RETROSPOT TEA TOWELS	18	1-12-2010 8:45	2.95	12583	France
536370	22326	ROUND SNACK BOXES SET OF4 WOODLAND	24	1-12-2010 8:45	2.95	12583	France
536370	22629	SPACEBOY LUNCH BOX	24	1-12-2010 8:45	1.95	12583	France
536370	22659	LUNCH BOX I LOVE LONDON	24	1-12-2010 8:45	1.95	12583	France

Part of Online Retail Company Data Set

Appendix B

Denormalized TeleFuture Data Set

No.	1: service_id_116 Nominal	2: service_id_1002 Nominal	3: service_id_1126 Nominal	4: service_id_1178 Nominal	5: service_id_1259 Nominal	6: service_id_1262 Nominal	7: service_id_1306 Nominal	8: service_id_1554 Nominal	9: service_id_1609 Nominal	10: service_id_1613 Nominal	11: service_id_1649 Nominal	12: service_id_1749 Nominal	13: service_id_1777 Nominal	14: service_id_1809 Nominal	15: service_id_1833 Nominal	16: service_id_1837 Nominal
1	f	f	t	f	t	t	f	t	t	t	t	t	t	f	t	t
2	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
3	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
4	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
5	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
6	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
7	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
8	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
9	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
10	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
11	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
12	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
13	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
14	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
15	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f
16	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
17	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
18	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f
19	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
20	f	f	f	f	t	t	f	f	f	f	f	f	f	f	f	f
21	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f
22	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
23	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
24	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
25	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
26	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
27	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
28	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
29	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
30	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
31	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
32	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
33	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
34	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
35	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
36	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
37	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
38	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
39	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
40	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
41	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
42	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
43	f	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f
44	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
45	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
46	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
47	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
48	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
49	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
50	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
51	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
52	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
53	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
54	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
55	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f
56	f	f	f	f	t	t	f	f	f	f	f	f	f	f	f	f
57	f	f	f	f	t	f	f	f	f	f	f	f	f	f	f	f

Part of the Denormalized Data Set of TeleFuture

Appendix C

Descriptions Online Retail Company

StockCode	Description
22697	GREEN REGENCY TEACUP AND SAUCER
23170	REGENCY TEA PLATE ROSES
22699	ROSES REGENCY TEACUP AND SAUCER
22698	PINK REGENCY TEACUP AND SAUCER
23171	REGENCY TEA PLATE GREEN
23173	REGENCY TEAPOT ROSES
22697	GREEN REGENCY TEACUP AND SAUCER
22423	REGENCY CAKESTAND 3 TIER
22745	POPPY'S PLAYHOUSE BEDROOM
22746	POPPY'S PLAYHOUSE LIVINGROOM
22748	POPPY'S PLAYHOUSE KITCHEN
22386	JUMBO BAG PINK POLKADOT
23202	JUMBO BAG VINTAGE LEAF
85099B	JUMBO BAG RED RETROSPOT
85099C	JUMBO BAG BAROQUE BLACK WHITE
22804	CANDLEHOLDER PINK HANGING HEART
85123A	WHITE HANGING HEART T-LIGHT HOLDER
21733	RED HANGING HEART T-LIGHT HOLDER

Figure 20: StockCode descriptions

Appendix D

Association Rules TeleFuture

```
1. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1969=t]: 616 ==> [service_id_1649=t]: 611   <conf:(0.99)> lift:(1.94) lev:(0) conv:(50.21)
2. [service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1969=t]: 677 ==> [service_id_1649=t]: 670   <conf:(0.99)> lift:(1.94) lev:(0) conv:(41.39)
3. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1997=t]: 534 ==> [service_id_1649=t]: 528   <conf:(0.99)> lift:(1.94) lev:(0) conv:(37.31)
4. [service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1997=t]: 568 ==> [service_id_1649=t]: 561   <conf:(0.99)> lift:(1.93) lev:(0) conv:(34.73)
5. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1969=t]: 712 ==> [service_id_1649=t]: 703   <conf:(0.99)> lift:(1.93) lev:(0) conv:(34.82)
6. [service_id_2189=t, service_id_1969=t, service_id_1997=t]: 552 ==> [service_id_1649=t]: 544   <conf:(0.99)> lift:(1.93) lev:(0) conv:(30)
7. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1997=t]: 623 ==> [service_id_1649=t]: 613   <conf:(0.98)> lift:(1.93) lev:(0) conv:(27.7)
8. [service_id_2189=t, service_id_1841=t, service_id_2421=t]: 1633 ==> [service_id_1649=t]: 1605   <conf:(0.98)> lift:(1.92) lev:(0) conv:(27.54)
9. [service_id_2441=t, service_id_2189=t, service_id_1613=t, service_id_1997=t]: 683 ==> [service_id_1649=t]: 671   <conf:(0.98)> lift:(1.92) lev:(0) conv:(25.7)
10. [service_id_2441=t, service_id_2189=t, service_id_1613=t, service_id_1969=t]: 829 ==> [service_id_1649=t]: 814   <conf:(0.98)> lift:(1.92) lev:(0) conv:(25.34)
11. [service_id_2189=t, service_id_1777=t, service_id_1997=t]: 708 ==> [service_id_1649=t]: 694   <conf:(0.98)> lift:(1.92) lev:(0) conv:(23.09)
12. [service_id_2189=t, service_id_1613=t, service_id_1997=t]: 734 ==> [service_id_1649=t]: 719   <conf:(0.98)> lift:(1.92) lev:(0) conv:(22.44)
13. [service_id_2189=t, service_id_1777=t, service_id_1969=t]: 860 ==> [service_id_1649=t]: 842   <conf:(0.98)> lift:(1.92) lev:(0) conv:(22.14)
14. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1613=t]: 1493 ==> [service_id_1649=t]: 1454   <conf:(0.97)> lift:(1.91) lev:(0) conv:(18.26)
15. [service_id_2189=t, service_id_1613=t, service_id_1969=t]: 956 ==> [service_id_1649=t]: 930   <conf:(0.97)> lift:(1.9) lev:(0) conv:(17.32)
16. [service_id_1649=t, service_id_1841=t, service_id_2421=t]: 1655 ==> [service_id_2189=t]: 1605   <conf:(0.97)> lift:(6.02) lev:(0) conv:(27.23)
17. [service_id_2441=t, service_id_2189=t, service_id_1969=t]: 1044 ==> [service_id_1649=t]: 1010   <conf:(0.97)> lift:(1.89) lev:(0) conv:(14.59)
18. [service_id_1649=t, service_id_2189=t, service_id_2317=t]: 916 ==> [service_id_1841=t]: 884   <conf:(0.97)> lift:(10.15) lev:(0) conv:(25.12)
19. [service_id_1841=t, service_id_2421=t]: 1693 ==> [service_id_2189=t]: 1633   <conf:(0.96)> lift:(5.99) lev:(0) conv:(23.29)
20. [service_id_2441=t, service_id_2189=t, service_id_1997=t]: 882 ==> [service_id_1649=t]: 850   <conf:(0.96)> lift:(1.89) lev:(0) conv:(13.07)
21. [service_id_2189=t, service_id_1777=t, service_id_1613=t]: 1771 ==> [service_id_1649=t]: 1705   <conf:(0.96)> lift:(1.88) lev:(0) conv:(12.93)
22. [service_id_2189=t, service_id_1997=t]: 1092 ==> [service_id_1649=t]: 1044   <conf:(0.96)> lift:(1.87) lev:(0) conv:(10.9)
23. [service_id_2441=t, service_id_2189=t, service_id_1613=t]: 2124 ==> [service_id_1649=t]: 2027   <conf:(0.95)> lift:(1.87) lev:(0) conv:(10.6)
24. [service_id_2189=t, service_id_2421=t]: 2886 ==> [service_id_1649=t]: 2740   <conf:(0.95)> lift:(1.86) lev:(0) conv:(9.6)
25. [service_id_1841=t, service_id_2421=t]: 1693 ==> [service_id_1649=t, service_id_2189=t]: 1605   <conf:(0.95)> lift:(13.82) lev:(0) conv:(17.72)
26. [service_id_2189=t, service_id_1969=t]: 1440 ==> [service_id_1649=t]: 1359   <conf:(0.94)> lift:(1.85) lev:(0) conv:(8.59)
27. [service_id_2189=t, service_id_2317=t]: 1948 ==> [service_id_1841=t]: 1835   <conf:(0.94)> lift:(9.91) lev:(0) conv:(15.46)
28. [service_id_1649=t, service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1997=t]: 561 ==> [service_id_2441=t]: 528   <conf:(0.94)> lift:(4.65) lev:(0) conv:(13.16)
29. [service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1997=t]: 568 ==> [service_id_2441=t]: 534   <conf:(0.94)> lift:(4.65) lev:(0) conv:(12.95)
30. [service_id_1649=t, service_id_2189=t, service_id_1613=t, service_id_1997=t]: 719 ==> [service_id_2441=t]: 671   <conf:(0.93)> lift:(4.62) lev:(0) conv:(11.71)
31. [service_id_2441=t, service_id_2189=t, service_id_1777=t, service_id_1837=t]: 669 ==> [service_id_1649=t]: 623   <conf:(0.93)> lift:(1.82) lev:(0) conv:(6.96)
32. [service_id_2189=t, service_id_1613=t, service_id_1997=t]: 734 ==> [service_id_2441=t]: 683   <conf:(0.93)> lift:(4.6) lev:(0) conv:(11.26)
33. [service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1997=t]: 568 ==> [service_id_1649=t, service_id_2441=t]: 528   <conf:(0.93)> lift:(12.69) lev:(0) conv:(12.84)
34. [service_id_2189=t, service_id_1613=t]: 2714 ==> [service_id_1649=t]: 2504   <conf:(0.92)> lift:(1.81) lev:(0) conv:(6.29)
35. [service_id_2189=t, service_id_1841=t, service_id_1777=t, service_id_1837=t]: 617 ==> [service_id_1649=t]: 568   <conf:(0.92)> lift:(1.8) lev:(0) conv:(6.04)
36. [service_id_2189=t, service_id_1613=t, service_id_1997=t]: 734 ==> [service_id_1649=t, service_id_2441=t]: 671   <conf:(0.91)> lift:(12.48) lev:(0) conv:(10.63)
37. [service_id_1649=t, service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1969=t]: 670 ==> [service_id_2441=t]: 611   <conf:(0.91)> lift:(4.51) lev:(0) conv:(8.91)
38. [service_id_2189=t, service_id_1777=t, service_id_1613=t, service_id_1969=t]: 677 ==> [service_id_2441=t]: 616   <conf:(0.91)> lift:(4.5) lev:(0) conv:(8.71)
39. [service_id_2441=t, service_id_2189=t, service_id_1841=t, service_id_1777=t]: 682 ==> [service_id_1649=t]: 618   <conf:(0.91)> lift:(1.77) lev:(0) conv:(5.13)
40. [service_id_2441=t, service_id_2189=t, service_id_1777=t]: 2963 ==> [service_id_1649=t]: 2678   <conf:(0.9)> lift:(1.77) lev:(0) conv:(5.07)
```

Top 40 Association Rules TeleFuture

Appendix E

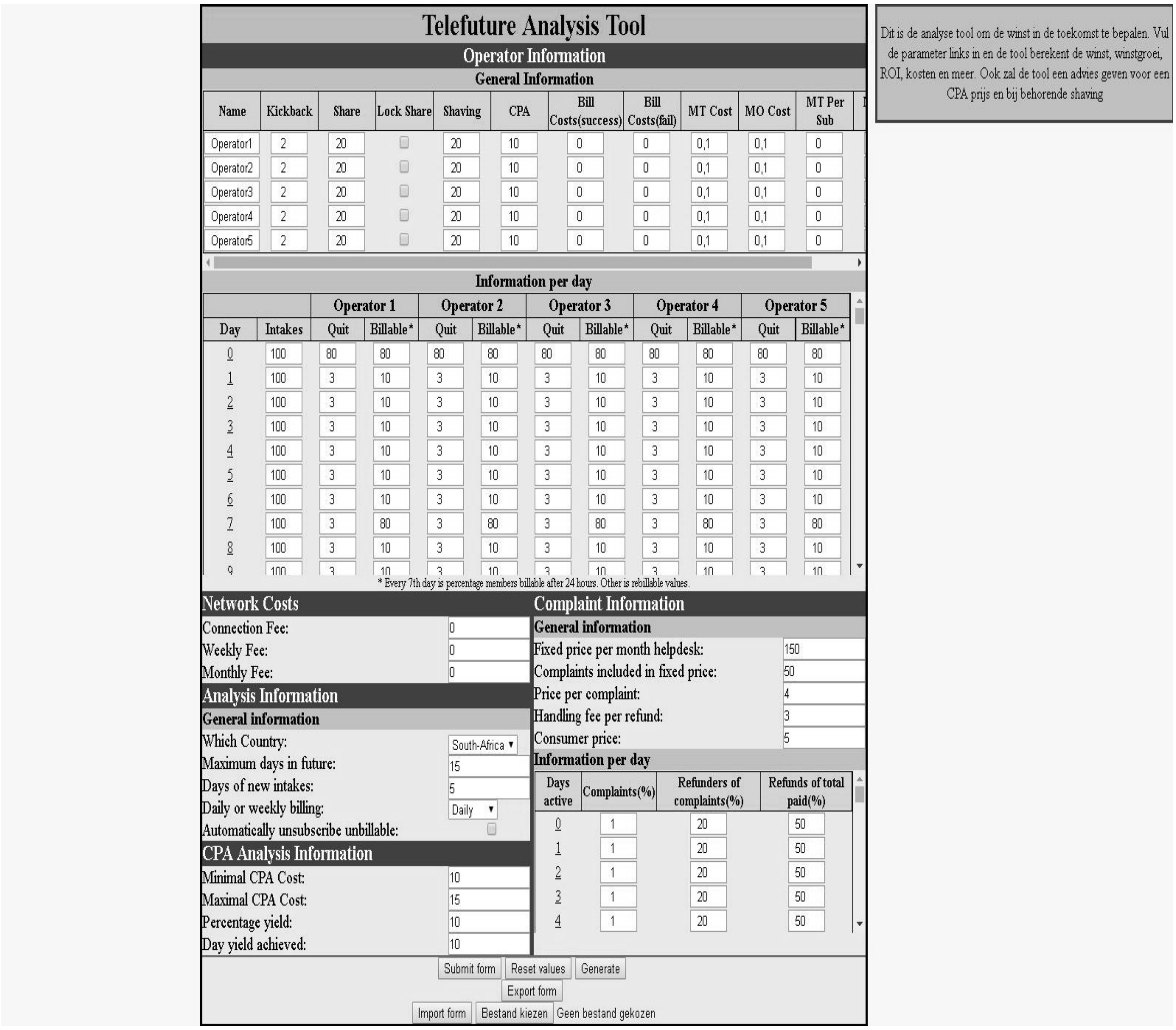
Parameters Analysis Tool

Parameter	Description
Kickback	The payout for TeleFuture per subscription
Shaving	Percentage that TeleFuture doesn't have to pay of CPA
CPA	Costs per subscriber
Billing Costs	Costs per payment try
MT sms Costs	Costs of a sent sms to customer
MO sms Costs	Costs of a recieved sms from customer
Quiting	Percentage of people that stops subscription after x days
Billing	Percentage of subscribers that can pay
Monthly Network Costs	Costs TeleFuture pays per month to network
Connection Fee	Starting fee to initialize a network
Complaints	Percentage of subscribers that complain about subscription
Refunds	Percentage of complaints that recieve a refund
Helpdesk Costs	The monthly costs of the helpdesk which handles complaints

Paramaters of Analysis Tool TeleFuture

Appendix F

Analysis Tool



Main Screen Analysis Tool TeleFuture

Appendix G

Result Table Analysis Tool

Intakes information					Intakes/ Billing Costs						
Day	Intakes	Total Members	Active Members	Billed	CPA Costs	Network Costs	Investment Cost	Cumulative Costs	SMS Costs	Billing Costs	Total Billing Costs
0	100	100	87	16	940.8 ZAR	1500 ZAR	2440.80 ZAR	2440.8 ZAR	38 ZAR	8.04 ZAR	2486.84 ZAR
1	100	200	171	24	940.8 ZAR	0 ZAR	940.80 ZAR	3381.6 ZAR	38 ZAR	15.39 ZAR	994.19 ZAR
2	100	300	253	31	940.8 ZAR	0 ZAR	940.80 ZAR	4322.4 ZAR	38 ZAR	22.53 ZAR	1001.33 ZAR
3	100	400	334	37	940.8 ZAR	0 ZAR	940.80 ZAR	5263.2 ZAR	38 ZAR	29.55 ZAR	1008.35 ZAR
4	100	500	414	42	940.8 ZAR	0 ZAR	940.80 ZAR	6204 ZAR	38 ZAR	36.54 ZAR	1015.34 ZAR
5	100	600	493	47	940.8 ZAR	0 ZAR	940.80 ZAR	7144.8 ZAR	38 ZAR	43.44 ZAR	1022.24 ZAR
6	100	700	573	52	940.8 ZAR	0 ZAR	940.80 ZAR	8085.6 ZAR	38 ZAR	50.25 ZAR	1029.05 ZAR
7	100	800	652	56	940.8 ZAR	0 ZAR	940.80 ZAR	9026.4 ZAR	38 ZAR	57.21 ZAR	1036.01 ZAR
8	100	900	730	60	940.8 ZAR	0 ZAR	940.80 ZAR	9967.2 ZAR	38 ZAR	64.08 ZAR	1042.88 ZAR
9	100	1000	808	64	940.8 ZAR	0 ZAR	940.80 ZAR	10908 ZAR	38 ZAR	70.86 ZAR	1049.66 ZAR
10	100	1100	886	66	940.8 ZAR	0 ZAR	940.80 ZAR	11848.8 ZAR	38 ZAR	77.76 ZAR	1056.56 ZAR
11	100	1200	964	70	940.8 ZAR	0 ZAR	940.80 ZAR	12789.6 ZAR	38 ZAR	84.54 ZAR	1063.34 ZAR
12	100	1300	1043	74	940.8 ZAR	0 ZAR	940.80 ZAR	13730.4 ZAR	38 ZAR	91.32 ZAR	1070.12 ZAR
13	100	1400	1122	77	940.8 ZAR	0 ZAR	940.80 ZAR	14671.2 ZAR	38 ZAR	98.25 ZAR	1077.05 ZAR
14	100	1500	1200	81	940.8 ZAR	0 ZAR	940.80 ZAR	15612 ZAR	38 ZAR	105.12 ZAR	1083.92 ZAR
15	100	1600	1279	84	940.8 ZAR	0 ZAR	940.80 ZAR	16552.8 ZAR	38 ZAR	111.96 ZAR	1090.76 ZAR
16	100	1700	1354	87	940.8 ZAR	0 ZAR	940.80 ZAR	17493.6 ZAR	38 ZAR	118.89 ZAR	1097.69 ZAR
17	100	1800	1430	92	940.8 ZAR	0 ZAR	940.80 ZAR	18434.4 ZAR	38 ZAR	125.34 ZAR	1104.14 ZAR
18	100	1900	1506	94	940.8 ZAR	0 ZAR	940.80 ZAR	19375.2 ZAR	38 ZAR	132.06 ZAR	1110.86 ZAR
19	100	2000	1583	98	940.8 ZAR	0 ZAR	940.80 ZAR	20316 ZAR	38 ZAR	138.66 ZAR	1117.46 ZAR
20	100	2100	1636	101	940.8 ZAR	0 ZAR	940.80 ZAR	21256.8 ZAR	38 ZAR	145.41 ZAR	1124.21 ZAR
21	100	2200	1647	105	940.8 ZAR	0 ZAR	940.80 ZAR	22197.6 ZAR	38 ZAR	149.94 ZAR	1128.74 ZAR
22	100	2300	1656	103	940.8 ZAR	0 ZAR	940.80 ZAR	23138.4 ZAR	38 ZAR	150.15 ZAR	1128.95 ZAR
23	100	2400	1662	104	940.8 ZAR	0 ZAR	940.80 ZAR	24079.2 ZAR	38 ZAR	150.09 ZAR	1128.89 ZAR

	Complaint/Refund Costs									Total Cost		Revenue			Ratio's		
Day	Complaints	Complaints by helpdesk	Complaint to pay for	Complaint Costs	Refunds	Refund Costs	Handling Costs	Total Costs	Cumulative Costs	Total Overall Costs	Cumulative Overall Costs	Revenue	Cumulative Revenue	Profit	ROI	Revenue Growth	Profit Growth
0	1	0	1	0 ZAR	0	0 ZAR	0 ZAR	0 ZAR	0.00 ZAR	2486.84 ZAR	2486.84 ZAR	176 ZAR	176 ZAR	-2310.84 ZAR	-94.68	0 %	0 %
1	2	0	2	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	13.00 ZAR	1007.19 ZAR	3494.03 ZAR	264 ZAR	440 ZAR	-3054.03 ZAR	-90.31	50.00 %	-32.16 %
2	2	0	2	0 ZAR	0	0 ZAR	0 ZAR	0 ZAR	13.00 ZAR	1001.33 ZAR	4495.36 ZAR	341 ZAR	781 ZAR	-3714.36 ZAR	-85.93	29.17 %	-21.62 %
3	3	0	3	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	26.00 ZAR	1021.35 ZAR	5516.71 ZAR	407 ZAR	1188 ZAR	-4328.71 ZAR	-82.24	19.35 %	-16.54 %
4	5	0	5	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	39.00 ZAR	1028.34 ZAR	6545.05 ZAR	462 ZAR	1650 ZAR	-4895.05 ZAR	-78.90	13.51 %	-13.08 %
5	5	0	5	0 ZAR	0	0 ZAR	0 ZAR	0 ZAR	39.00 ZAR	1022.24 ZAR	7567.29 ZAR	517 ZAR	2167 ZAR	-5400.29 ZAR	-75.58	11.90 %	-10.32 %
6	6	0	6	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	52.00 ZAR	1042.05 ZAR	8609.34 ZAR	572 ZAR	2739 ZAR	-5870.34 ZAR	-72.60	10.64 %	-8.70 %
7	7	0	7	0 ZAR	0	0 ZAR	0 ZAR	0 ZAR	52.00 ZAR	1036.01 ZAR	9645.35 ZAR	616 ZAR	3355 ZAR	-6290.35 ZAR	-69.69	7.69 %	-7.15 %
8	7	0	7	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	65.00 ZAR	1055.88 ZAR	10701.23 ZAR	660 ZAR	4015 ZAR	-6686.23 ZAR	-67.08	7.14 %	-6.29 %
9	9	0	9	0 ZAR	3	40 ZAR	9 ZAR	49 ZAR	114.00 ZAR	1098.66 ZAR	11799.89 ZAR	704 ZAR	4719 ZAR	-7080.89 ZAR	-64.91	6.67 %	-5.90 %
10	10	0	10	0 ZAR	2	30 ZAR	6 ZAR	36 ZAR	150.00 ZAR	1092.56 ZAR	12892.45 ZAR	726 ZAR	5445 ZAR	-7447.45 ZAR	-62.85	3.13 %	-5.18 %
11	12	0	12	0 ZAR	4	40 ZAR	12 ZAR	52 ZAR	202.00 ZAR	1115.34 ZAR	14007.79 ZAR	770 ZAR	6215 ZAR	-7792.79 ZAR	-60.93	6.06 %	-4.64 %
12	11	0	11	0 ZAR	3	30 ZAR	9 ZAR	39 ZAR	241.00 ZAR	1109.12 ZAR	15116.91 ZAR	814 ZAR	7029 ZAR	-8087.91 ZAR	-58.91	5.71 %	-3.79 %
13	9	0	9	0 ZAR	5	70 ZAR	15 ZAR	85 ZAR	326.00 ZAR	1162.05 ZAR	16278.96 ZAR	847 ZAR	7876 ZAR	-8402.96 ZAR	-57.28	4.05 %	-3.90 %
14	11	0	11	0 ZAR	2	30 ZAR	6 ZAR	36 ZAR	362.00 ZAR	1119.92 ZAR	17398.88 ZAR	891 ZAR	8767 ZAR	-8631.88 ZAR	-55.29	5.19 %	-2.72 %
15	15	0	15	0 ZAR	3	50 ZAR	9 ZAR	59 ZAR	421.00 ZAR	1149.76 ZAR	18548.64 ZAR	924 ZAR	9691 ZAR	-8857.64 ZAR	-53.51	3.70 %	-2.62 %
16	14	0	14	0 ZAR	3	60 ZAR	9 ZAR	69 ZAR	490.00 ZAR	1166.69 ZAR	19715.33 ZAR	957 ZAR	10648 ZAR	-9067.33 ZAR	-51.83	3.57 %	-2.37 %
17	17	0	17	0 ZAR	3	70 ZAR	9 ZAR	79 ZAR	569.00 ZAR	1183.14 ZAR	20898.47 ZAR	1012 ZAR	11660 ZAR	-9238.47 ZAR	-50.12	5.75 %	-1.89 %
18	15	0	15	0 ZAR	1	30 ZAR	3 ZAR	33 ZAR	602.00 ZAR	1143.86 ZAR	22042.33 ZAR	1034 ZAR	12694 ZAR	-9348.33 ZAR	-48.25	2.17 %	-1.19 %
19	15	0	15	0 ZAR	4	80 ZAR	12 ZAR	92 ZAR	694.00 ZAR	1209.46 ZAR	23251.79 ZAR	1078 ZAR	13772 ZAR	-9479.79 ZAR	-46.66	4.26 %	-1.41 %
20	16	0	16	0 ZAR	0	0 ZAR	0 ZAR	0 ZAR	694.00 ZAR	1124.21 ZAR	24376 ZAR	1111 ZAR	14883 ZAR	-9493 ZAR	-44.66	3.06 %	-0.14 %
21	16	0	16	0 ZAR	2	40 ZAR	6 ZAR	46 ZAR	740.00 ZAR	1174.74 ZAR	25550.74 ZAR	1155 ZAR	16038 ZAR	-9512.74 ZAR	-42.85	3.96 %	-0.21 %
22	17	0	17	0 ZAR	1	10 ZAR	3 ZAR	13 ZAR	753.00 ZAR	1141.95 ZAR	26692.69 ZAR	1133 ZAR	17171 ZAR	-9521.69 ZAR	-41.15	-1.90 %	-0.09 %
23	17	0	17	0 ZAR	3	70 ZAR	9 ZAR	79 ZAR	832.00 ZAR	1207.89 ZAR	27900.58 ZAR	1144 ZAR	18315 ZAR	-9585.58 ZAR	-39.81	0.97 %	-0.67 %

Result Table of Analysis Tool