DATA MINING PROJECT 1

1. BUSINESS UNDERSTANDING

Supermarket chain “Coop” provided us with some data in order to discover new insights in their products and customers’ purchasing behavior. For this we can use data about products, the marketing hierarchy, sales, customers and dates.

2. DATA UNDERSTANDING

Before we can start with the data understanding, we need to convert the csv-files to arff-files, using the tool “arff-writer” in Knime. Then we load each table into Weka to perform some basic statistics and histograms to understand the data. For the more sophisticated analysis of the data, we load the tables to an SQL server database. This enables us to perform queries and transformations on the data (in linked tables) in the database using the business intelligence tools of Visual Studio and the SQL server management studio software.

2.1. DATA EXPLORATION

As a first step, we look at some statistics and distributions that might contain useful information. The table that contains data about customers can be analyzed using distributions and histograms. Figure 1 shows the distribution of the profession of the customers. Other histograms show more information about the birth year, anno_socio of customers but also about the quality of the data. For example for the birth year there are some outliers, which are mistakes (birth year 1260 is impossible).

Figure 2 shows the transaction statistics. These basic statistics can help us to get an idea of the data and enable us to see if there are some peculiar data entries in the table. For example the negative values for “QTA_PESO” and “QTA_PEZZI”. The negative values of “IMPORTO” are interpreted as the value of a (special) promotion and therefore are not “wrong” measurements.

In further data exploration we want to learn more about products and their link with the product hierarchy. Using Visual Studio (figure 14), we join the tables of articles with the table of marketing structure. The attribute “cod_mkt_id” is used to link the tables. We group by sector and count the distinct products for each sector (figure 3). From the figure we can tell for instance that the sector “FRESCHISSIMI” contains 19389 products.

In the next table (figure 4) we show how many of these products per sector were actually sold in the 6 months of sales data (computed with Visual Studio, figure 15). If we compare the number of products in the sector “FRESCHISSIMI”: 19389 products in the products table (see figure 3) of which only 857 (4,4%) were actually sold in 6 months (figure 4).

From the table with transactions, we compute for each article how many times it was sold (figure 5 shows a small extract from the whole table of results). The table was computed using the Visual Studio software (figure 16). It performs a join of the sales table with the products and then computes the number of products sold (in descending order). In total, 5502 products out of 345208 different products were sold one time or more, which accounts for just 1,6% of all products from the table products (articoli). The most sold item is the plastic bag of Coop bought at the cash register.

The last step in our data exploration may be of great interest to the supermarket. It is the ordering of products sold sorted by revenue. It was computed with a simple query summing the revenue, grouping by "articolo_id". The top part of the result set is displayed in figure 6. The product 'B.A. macinato scelto…' seems to be the product that has generated the most revenue (not profit!) over the 6 months of sales recorded in the sales table.

2.2. CUSTOMER PROFILING

A supermarket like Coop can benefit with regard to revenue and competitors if it can distinguish between profitable and occasional customers. To be able to perform customer segmentation, we compute for each customer (i) monetary volume of purchases made by the customer, (ii) number of transactions (visits) performed and (iii) number of single products bought including all products using the following query in Microsoft SQL Server:

```
SELECT CLIENTE_ID, SUM(IMPORTO) AS [MONETARY VOLUME], COUNT(DISTINCT
SCONTRINO) AS [NUMBER OF VISITS], COUNT(DISTINCT ARTICOLO_ID) AS
[NUMBER OF PRODUCTS]
FROM   SSET_VEN_CORSODM
GROUP BY CLIENTE_ID
```

Which generates the table, a part of it in figure 7. We think tha these 3 measures are the most important one can compute out of the provided data.

(i) shows how much revenue the customer generates
(ii) the number of visits helps us distinguish frequent customers from infrequent customers
(iii) number of products shows the basket size of a customer, does (s)he buy only few products or does (s)he buy regularly a decent amount of poducts?

Next, we computed the same aggregates for each customer, but only including the products of the category "FRESCHISSIMI", using the following query in Microsoft SQL Server:

```
SELECT CLIENTE_ID, SUM(IMPORTO) AS [MONETARY VOLUME], COUNT(DISTINCT
SCONTRINO) AS [NUMBER OF VISITS], COUNT(DISTINCT
SSET_VEN_CORSODM.ARTICOLO_ID) AS [NUMBER OF PRODUCTS]
FROM   SSET_VEN_CORSODM INNER JOIN
            SSET_ART_CORSODM ON SSET_VEN_CORSODM.ARTICOLO_ID =
     SSET_ART_CORSODM.ARTICOLO_ID INNER JOIN
            SSET_MKT_CORSODM ON SSET_ART_CORSODM.COD_MKT_ID =
SSET_MKT_CORSODM.COD_MKT_ID
WHERE SETTORE='FRESCHISSIMI'
GROUP BY CLIENTE_ID
```

Resulting in the table, part of it displayed in figure 8.

3. INDIVIDUAL EVENTS DETECTION

3.1. CHURN ANALYSIS

Out of the transaction data we can extract some other useful information for each individual customer. One of the possible analyses that can be done is “churn analysis”. We compute for each customer the number of visits (s)he has done over the months the sales data covers. We say that a customer is churning when the customer has bought something in a month, but nothing in the next n months. This is slightly different from the objective for this project that stated the **last** n months. We think that the churn analysis is more complete when we look at all cases, not only the last n months. So for example if a customer buys something in April and nothing in the **next** n months, (s)he is churning.

In this analysis we choose to set n equal to 2 because supermarkets benefit from frequent visits of customers and so we want to limit the time frame to consider a customer “churning” (this means a customer can churn more than once over 6 months!). The following SQL-query computes the number of visits in each month for each customer.

```
SELECT CLIENTE_ID, MESE_N, COUNT(DISTINCT V.SCONTRINO) AS [Number of visits]
FROM tj.dbo.SSET_VEN_CORSODM AS V, tj.dbo.SSET_DATA_DORSODM AS D
WHERE V.DATA_ID = D.DATA_ID
GROUP BY CLIENTE_ID, MESE_N
ORDER BY CLIENTE_ID, MESE_N
```

The first rows of the result set is shown in figure 9. Although the result contains all the data needed to determine if a customer is churning or not, it displays only the months in which the customer has visited the store (not the others). The analyst now has to look for himself if the customer is churning in a particular month.

Alternatively we can compute for each customer the month(s) in which (s)he is churning in the program Visual Studio using a data flow and transformations (figure 17). In this analysis we include also the tipologia of that customer in the month (s)he is churning (mind that if the customer changes tipologia within the month (s)he is churning, the highest one is chosen). In this way the supermarket can see if there are also loyal customer (tiplogia 2-6) who are churning and find out why the customer left. An extract of the result set of the churn analysis is shown in figure 10.

3.2. CUSTOMERS FOCUSING ON PRODUCT(-SUBCATEGORIES)

Another interesting question one can ask, provided with this data, is whether individual customers focused their purchases on specific kinds of products. For this question we are focusing on the purchases of products within the sector “FRESCHISSIMI”. For each customer we compute the number of products bought, within a subcategory (reparto) of the sector “FRESCHISSIMI”, as a percentage of the total number of different products bought within the sector (settore) “FRESCHISSIMI”. We do this computation separately for two periods, two quarters actually. From the resulting data set (figure 11 shows first 20 rows) we can see if a customer is focusing on a particular subcategory if (s)he buys more than double in percentage in the second period than the first. An analysis of this kind may be very useful to marketing practices and more specifically personalized advertising. The results were obtained using Visual Studio software, the data flow is transformed via a process shown in figure 18.

In figure 12 one can see an example of a “focusing” customer. Customer 7837 buys more than double in percentage of red meat (carni rosse) products in the last 3 months compared to the first 3 months. So maybe in the future we can try and advertise these red meat products together with other products (cross-selling).

4. TIME SERIES

Using the following query, we obtain the number of products bought by a customer in a particular week for the segment “pesce”. The result is (partly) visualized in figure 13.

```
SELECT V.CLIENTE_ID, D.SETTIMANA_ANNO, COUNT(*) AS Number_of_purchases
FROM [tj].[dbo].[SSET_VEN_CORSODM] AS V INNER JOIN [tj].[dbo].[SSET_ART_CORSODM] AS A ON
V.ARTICOLO_ID=A.ARTICOLO_ID INNER JOIN [tj].[dbo].[SSET_MKT_CORSODM] AS M ON
A.COD_MKT_ID= M.COD_MKT_ID INNER JOIN [tj].[dbo].[SSET_DATA_DORSODM] AS D ON
V.DATA_ID = D.DATA_ID
WHERE SEGMENTO='PESCE'
GROUP BY V.CLIENTE_ID, D.SETTIMANA_ANNO
ORDER BY V.CLIENTE_ID, D.SETTIMANA_ANNO
```

First, the tables “SSET_VEN_CORSODM”, “SSET_ART_CORSODM”, “SSET_MKT_CORSODM” and “SSET_DATA_DORSODM” are joined. By grouping on “CLIENTE_ID” and “SETTIMANA_ANNO”, for each customer the number of products bought in each week is counted. The where-clause selects only the products of the segment “pesce”.

From the obtained table, we can tell that most people don’t buy a product of the segment “pesce” often. For most people who bought a product in different weeks, there are a few weeks between each purchase. We’ve chosen for this particular segment, because it holds information that can be useful to manage inventory. Because products of the segment “pesce” rot fast, it’s important to harmonize supply and demand for these products in order to avoid unsold items. All items that don’t get sold in time contribute to an extra loss.

5. EVALUATION/CONCLUSION

The data analysis performed in this project is only a small part of interesting analysis one can perform using the data of the supermarket Coop. The results of the customer analysis (2.2 Customer profiling) can be used for clustering and finding segments of customers. Furthermore one can analyze in more detail a customer using the results from churn analysis (3.1), focusing (3.2) and time series (4) for example to look for frequent patterns or association rules.

6. FIGURES

Frequent Value Distribution (0.1000 %) - PROFESSIONE

Value	Count	Percentage
AGRICOLTORE	5051	0.4260 %
ALTRE PROFESSIONI	26654	2.2479 %
ARTIGIANO	16926	1.4275 %
CASALINGA	215702	18.1919 %
COMMERCIANTE	5399	0.4553 %
COOP CONSUMO	3663	0.3089 %
DIRIGENTE DI AZIEN...	3228	0.2722 %
DISOCCUPATO	16002	1.3496 %
ENTE PRIVATO	6587	0.5555 %
ENTE PUBBLICO	42155	3.5553 %
IMPIEGATO	172015	14.5074 %
INSEGNANTE	7839	0.6611 %
LAVORATORE AUTO...	30148	2.5426 %
LIBERO PROFESSIO...	39649	3.3439 %

Figure 1

Column Statistics Profiles - [dbo].[SSET_VEN_CORSODM]

Column	Minimum	Maximum	Mean	Standard Deviation
ARTICOLO_ID	105	311816	88250.1990589356	91532.4292779957
CLIENTE_ID	1146	1025026	557950.162886771	254618.323551186
DATA_ID	1187	1369	1280.42706867296	51.6648686241127
IMPORTO	-114,1399993...	156,240005493164	1.95761291427029	2.28325867407314
QTA_PESO	-0,873000025...	69,4469985961914	0.164945443104629	0.467793695000522
QTA_PEZZI	-4	504	1.30583427276398	1.51410508573637
SCONTRINO	11870070115...	1369007029999	1280434087877.92	51664868840.315
TIPOLOGIA_ID	0	7	5.67829460078282	1.3776916922992

Figure 2

ProductCount	SETTORE
1491	CONFEZIONATO PER VENDITA
89068	STAGIONALI E BRICO
2	EROGAZIONE CARBURANTI
19389	FRESCHISSIMI
825	INIZIATIVE SPECIALI
43	NON DEFINITO
16341	FRESCHI
90155	PERSONA
33430	MULTIMEDIA
22295	CHIMICA
1316	SALUTE
44231	CASA
26622	GROCERY ALIMENTARI

Figure 3

ProductCount	SETTORE
3	CONFEZIONATO PER VENDITA
125	STAGIONALI E BRICO
857	FRESCHISSIMI
5	INIZIATIVE SPECIALI
1076	FRESCHI
104	PERSONA
49	MULTIMEDIA
1000	CHIMICA
118	CASA
2165	GROCERY ALIMENTARI

Figure 4

ARTICOLO_ID	DES_ART	TimesSold
135625	SHOPPERS COOP COEX PICCOLI DEGRADAB.18MESI CM32+11+11X59 CA	4007
31560	BANANE COOP ES 19+ I^ SF	3820
56786	PANE PIZZICATO GR.500	3443
27227	BAGUETTE AGRITECH SACCHETTO 300G	2814
14075	MORTADELLA BOLOGNA IGP C/PISTACCHI SIGARO 1/2 CASA MODENA KG 14 CA ^	2714
65123	PANE COMUNE TIPO 00 PANIFICIO BATANI KG 0.500	2339
65616	MELONE SEMIRETATO IT 800-1200 I^ SF	2160
306158	PROSCIUTTO CRUDO S/OSSO SPUNTATO C/AGLIO PIACENTI KG 5 CA	1886
40812	SCHIACCIA SALATA	1844
40808	SCHIACCIA ASSORTITA	1832
56784	PANE DA GR.500	1750
48067	ZUCCHERO SEMOLATO ITALIA ZUCCHERI SCATOLA KG 1	1722
35993	B.A. MACINATO SCELTO G.200/300	1721
40809	PIZZA MOZZARELLA	1677
15786	LATTE UHT PS COOP BRICK 1 L	1653

Figure 5

ARTICOLO_ID	IMPORTO	DES_ART
35993	6748	B.A. MACINATO SCELTO G.200/300
27664	5773	POLLO PETTO COOP A FETTE CF
40368	5244	VIT/NE FETTINE PER GRIGLIA
27761	5126	VIT/NE MAGRO SCELTO FETTE GR.200/300
69420	4298	TACCHINO FESA FETTE FEMMINA COOP CF CA 200G
28864	4245	VIT/NE BISTECCHE I TAGLIO
31560	4196	BANANE COOP ES 19+ I^ SF
37009	4131	SUINO BISTECCA I TAGLIO X2
15670	4100	ACQUA MINERALE LEVISSIMA NATURALE PET LT.1
26017	4083	SCOTTONA BISTECCA I TAGLIO
306158	3963	PROSCIUTTO CRUDO S/OSSO SPUNTATO C/AGLIO PIACENTI KG 5 CA
65616	3531	MELONE SEMIRETATO IT 800-1200 I^ SF
27966	3404	SCOTTONA FETTINE PER GRIGLIA
14075	3339	MORTADELLA BOLOGNA IGP C/PISTACCHI SIGARO 1/2 CASA MODENA KG 14 CA
56786	3161	PANE PIZZICATO GR.500
35995	3150	B.A.HAMBURGER I TAGLIO
75954	3127	VIT/NE CARPACCIO
15695	3069	ACQUA EFFERVESCENTE NATURALE ULIVETO PET 1
92461	3060	RICARICA 30 E. TIM
15775	3057	PESCHE GIALLE IT AA I^ SF
34402	2949	SUINO ARISTA SENZ'OSSO A FETTE

Figure 6

	CLIENTE_ID	MONETARY VOLUME	NUMBER OF VISITS	NUMBER OF PRODUCTS
1	160172	13,4200001321733	1	11
2	914170	32,1100000105798	1	18
3	101809	11,3500000517815	1	7
4	69106	8,04000022821128	1	3
5	921925	18,549999833107	1	9
6	573052	28,4300000350922	1	19
7	675264	834,340003557503	82	172
8	933071	35,2499999366701	1	20
9	1009078	3,4100000243634	1	4
10	710126	691,379998575896	37	216
11	174876	29,1899996958673	1	15

Figure 7

	CLIENTE_ID	MONETARY VOLUME	NUMBER OF VISITS	NUMBER OF PRODUCTS
1	675264	195,53000125289	57	55
2	776741	95,4199995994568	18	31
3	692505	209,110000029206	52	65
4	914170	8,83999997377396	1	5
5	933071	12,7999998033047	1	8
6	904352	4,79000008106232	1	2
7	874305	44,0499999523163	10	23
8	982351	16,8900001645088	1	6
9	682758	303,109999001026	23	57
10	740457	423,349999159575	46	86
11	921925	15,5399998426437	1	8

Figure 8

	CLIENTE_ID	MESE_N	Number of visits
1	1146	5	1
2	1207	4	12
3	1207	5	11
4	1207	6	9
5	1207	7	22
6	1207	8	15
7	1207	9	14
8	3563	4	3
9	3563	5	4
10	3563	6	6
11	3563	7	5
12	3563	8	8
13	3563	9	2
14	5379	5	1
15	5379	6	1
16	5379	7	2
17	5379	8	1
18	5379	9	5
19	5801	9	1
20	5892	4	1
21	5892	5	8
22	5892	6	3
23	5892	7	8
24	5892	8	5
25	5892	9	5
26	6228	4	11
27	6228	5	2
28	6228	6	4
29	6228	7	5
30	6228	8	4
31	6228	9	1
32	7191	5	3
33	7191	6	1
34	7191	7	1
35	7191	8	1
36	7191	9	4
37	7409	4	1
38	7409	5	2
39	7409	6	3
40	7409	7	4
41	7409	8	3
42	7409	9	1
43	7758	6	1
44	7758	7	2
45	7758	8	3

Figure 9

CLIENTE_ID	MESE_N	Number of visits	TIPOLOGIA_ID
1146	5	1	7
8363	5	1	7
8504	7	1	7
9632	7	1	7
9678	5	4	2
10340	7	1	7
12696	7	1	7
12697	6	1	7
13159	5	1	7
13656	7	1	7
13718	4	1	7
13737	5	1	7
14090	4	1	7
14594	7	1	7
14628	4	1	7
14690	6	1	7
15907	4	1	7
17249	7	4	7
17332	5	1	7
17337	5	1	7
17437	5	2	7
18744	4	1	7
19793	6	1	7

Figure10

CLIENTE_ID	REPARTO	Count_first_3	Count_last_3	Percentage_first_3	Percentage_last_3
1146	AVICUNICOLO	1	0	11,1	0
1146	ORTOFRUTTA	6	0	66,7	0
1146	PANE	2	0	22,2	0
1207	AVICUNICOLO	0	1	0,0	2
1207	CARNI ROSSE	8	10	40,0	20
1207	ORTOFRUTTA	4	19	20,0	38
1207	PANE	8	20	40,0	40
3563	AVICUNICOLO	2	2	5,7	6,2
3563	CARNI ROSSE	6	10	17,1	31,2
3563	ORTOFRUTTA	19	10	54,3	31,2
3563	PANE	8	10	22,9	31,2
5379	CARNI ROSSE	4	1	50,0	8,3
5379	ORTOFRUTTA	4	8	50,0	66,7
5379	PANE	0	3	0,0	25
5801	ORTOFRUTTA	0	4	0,0	80
5801	PANE	0	1	0,0	20
5892	AVICUNICOLO	0	1	0,0	7,7
5892	CARNI ROSSE	4	2	36,4	15,4
5892	ORTOFRUTTA	0	5	0,0	38,5
5892	PANE	7	5	63,6	38,5

Figure 11

CLIENTE_ID	REPARTO	Count_first_3	Count_last_3	Percentage_first_3	Percentage_last_3
7837	AVICUNICOLO	2	1	9,5	6,2
7837	CARNI ROSSE	6	11	28,6	68,8
7837	ORTOFRUTTA	7	0	33,3	0
7837	PANE	6	4	28,6	25

Figure 12

	CLIENTE_ID	SETTIMANA_ANNO	Number_of_purchases
1	1207	18	1
2	1207	20	1
3	7837	26	1
4	8035	34	1
5	14844	21	1
6	31792	31	2
7	31792	33	2
8	51575	32	1
9	51575	37	1
10	54262	32	1
11	68693	18	1

Figure 13

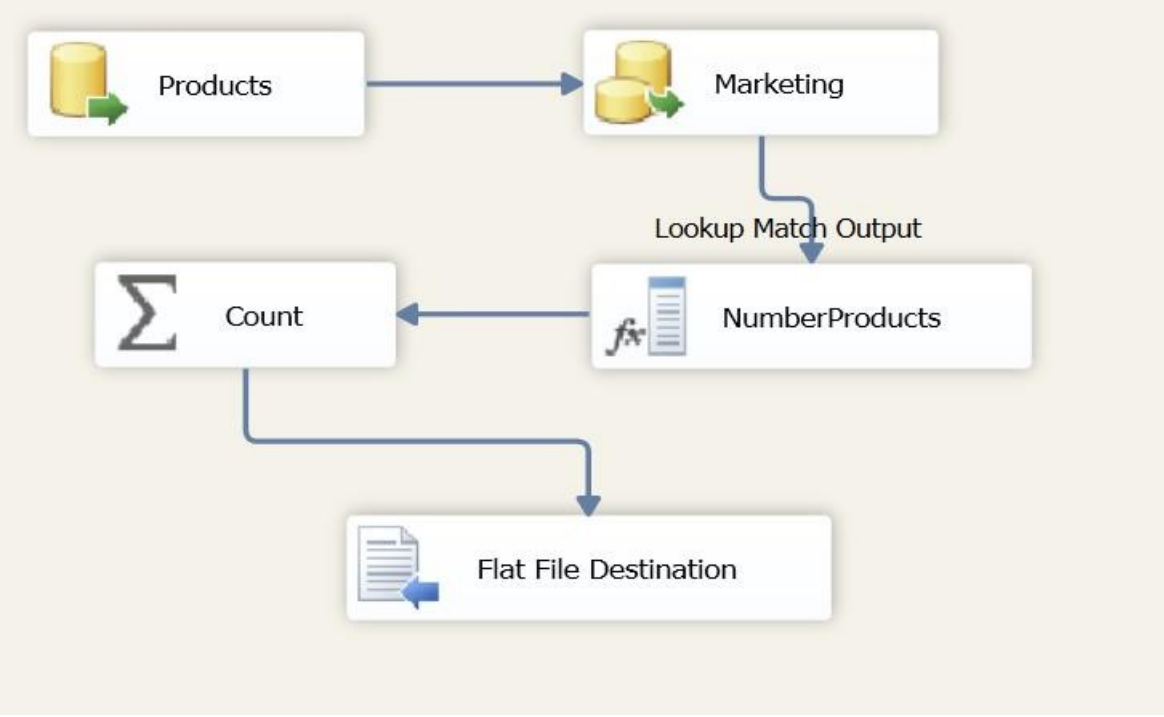


Figure 14

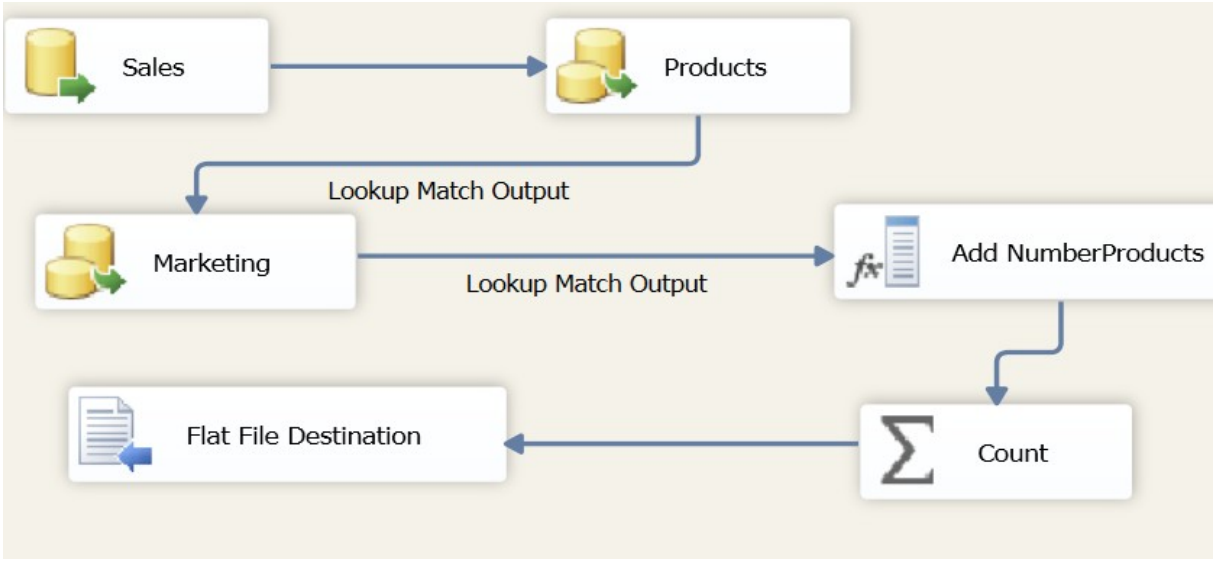


Figure 15

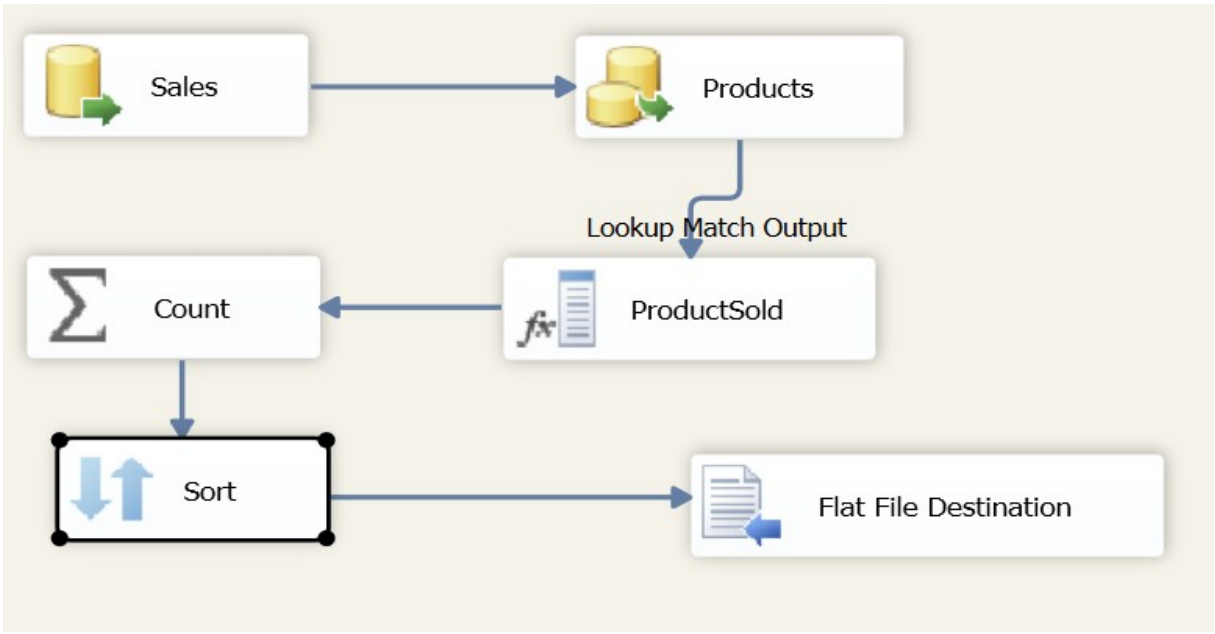


Figure 16

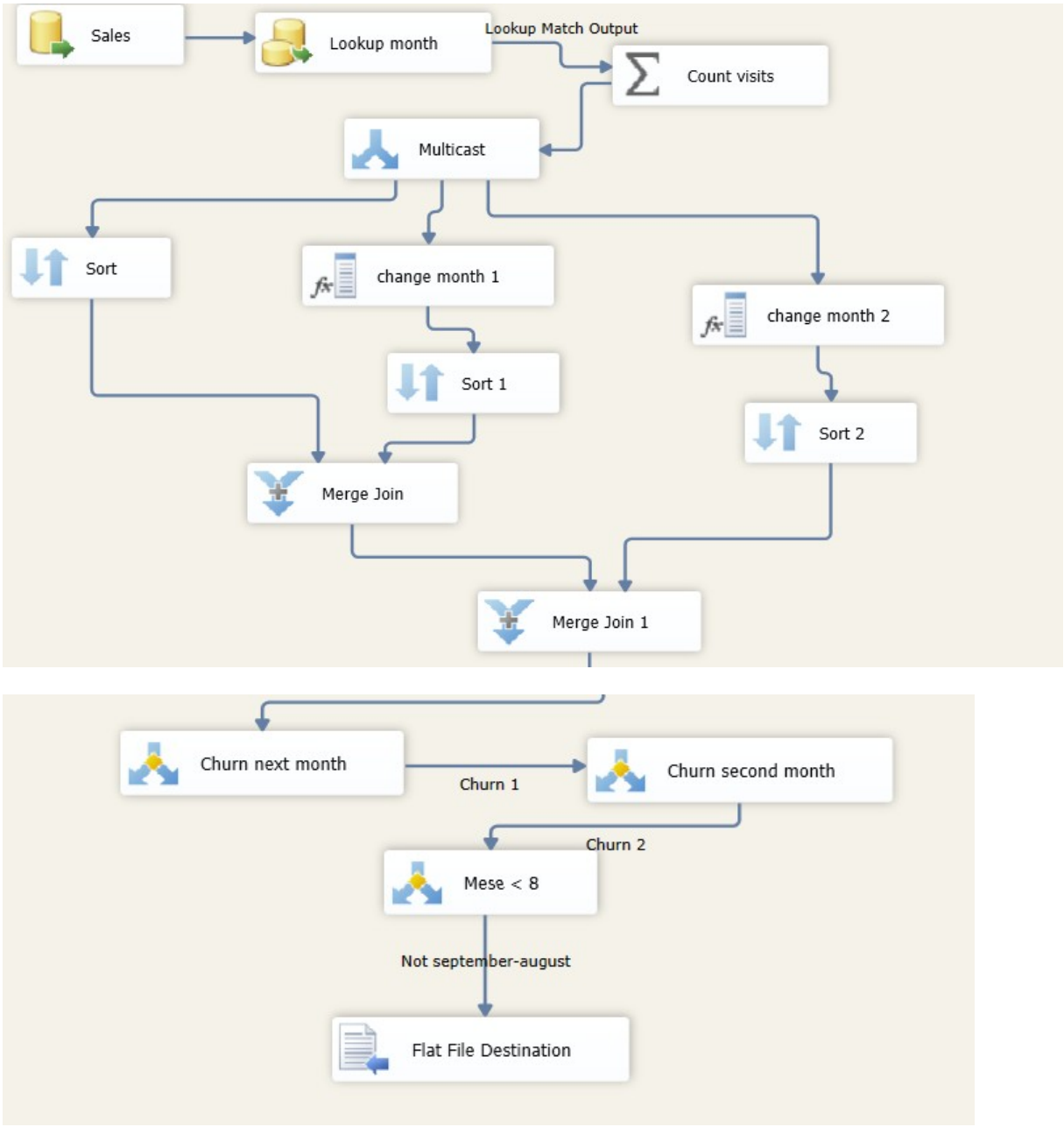

Sales
Lookup month
Lookup Match Output
Count visits
Multicast
Sort
change month 1
change month 2
Sort 1
Sort 2
Merge Join
Merge Join 1
Churn next month
Churn 1
Churn second month
Churn 2
Mese < 8
Not september-august
Flat File Destination

Figure 17

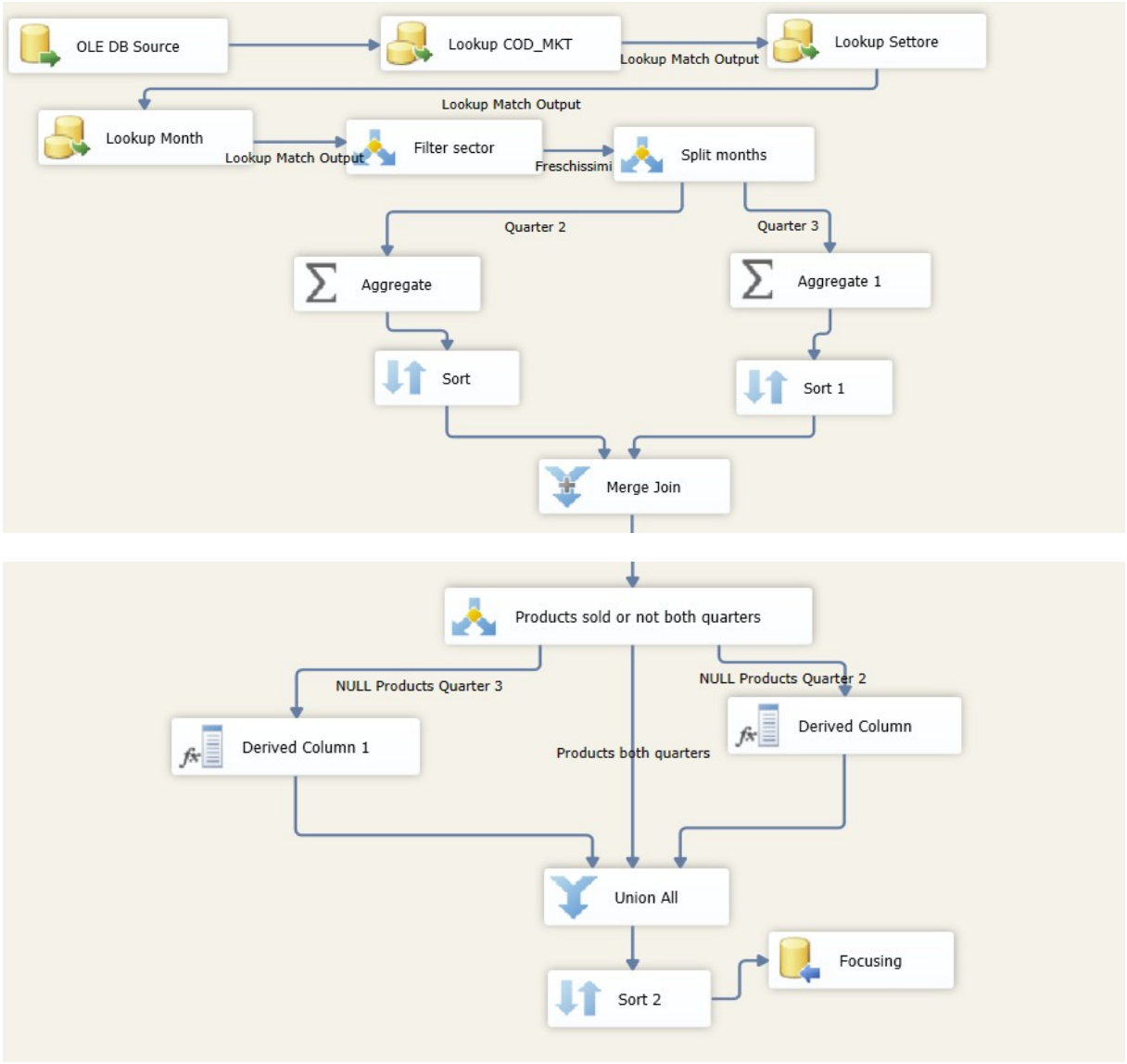

OLE DB Source
Lookup COD_MKT
Lookup Match Output
Lookup Settore
Lookup Match Output
Lookup Month
Lookup Match Output
Filter sector
Freschissimi
Split months
Quarter 2
Quarter 3
Aggregate
Aggregate 1
Sort
Sort 1
Merge Join
Products sold or not both quarters
NULL Products Quarter 3
NULL Products Quarter 2
Derived Column 1
Products both quarters
Derived Column
Union All
Focusing
Sort 2

Figure 18

Data Mining Project Part 2

1 CLUSTERING

Before starting the clustering analysis, we need to do some preprocessing. For the attribute 'ANNO_NASCITA', all the cases with value '0' are replaced by '?', which corresponds to a missing value. Furthermore, all cases who were born in 1905 or before were marked as missing values. Also for other attributes ('SESSO', 'STATO CIVILE', 'PROFESSIONE') we replaced the values 'ND' or 'Non disponible' with a '?'.

Now the preprocessing is done, we can start to perform clustering by using the Simple K Means algorithm in Weka. By comparing the SSE of different clusterings on the same attributes and by taking into account how much the obtained clusters differ from one another, we selected the clustering with k equals 6 as the best one.

```
kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 2389.198509615827
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3           4           5
                             (4184)      (368)       (527)       (335)       (706)       (116)      (2132)
=============================================================================================================================
ANNO_NASCITA               1955.4091   1968.3234   1957.1025   1938.0054    1955.296   1947.8966   1955.9422
SESSO                          Donna       Uomo       Donna       Donna       Uomo        Uomo       Donna
STATO_CIVILE               Coniugato      Celibe      Celibe   Coniugato   Coniugato   Coniugato   Coniugato
PROFESSIONE                CASALINGA  IMPIEGATO  ALTRE PROFESSIONI  PENSIONATO    OPERAIO     OPERAIO  CASALINGA
ANNO_SOCIO                 1997.7089   2001.4565    1990.888   1991.7731   1991.7436   1989.8276   2002.0849
MONETARY VOLUME              62.7421     35.4523    127.7947     84.3991     27.7433    411.3679     40.5908
NUMBER OF VISITS              9.4328       6.087     18.9583      12.606      4.8782     64.2241      5.6843
NUMBER OF PRODUCTS           15.6656     11.5679      27.592     20.5104      9.2025     76.6207     11.4873



Time taken to build model (full training data) : 0.26 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       368 (  9%)
1       527 ( 13%)
2       335 (  8%)
3       706 ( 17%)
4       116 (  3%)
5      2132 ( 51%)
```

In the figure above the raw clustering results are shown. In the table on the next page once can find the description of the different clusters based on the distributions of the attributes within each cluster.

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Size	368	527	335	706	116	2132
Anno nascita	1971-1976	No particular interval that contains the majority of the cluster	1936-1941	No particular interval that contains the majority of the cluster	<1954	1953-1968
Sesso	Male	Female	Slightly more females	Male	Male	Female
Stato_civile	Bachelor	Bachelor	Married	Married	Married	Married
Professione	Employee	Housewife or other profession	Retired	Worker	Worker or other profession	Housewife
Anno_socio	2002-2005	1981-1989 or 1996-2004	1979-1985 or 1996-2008	No particular interval that contains the majority of the cluster	No particular interval that contains the majority of the cluster	2005-2008
Fl_invio_revista	Yes	Yes	Yes	Yes	Yes	No
Monetary volume	€0.38-€35.33	-6 – 93.91	0.57-76.8	0.29-17.99	119.87-295.68	0.22-29.43
Number of visits	1-6	1-15	1-12	1-3	38-65	1-5
Number of products	1-8	1-16	1-15	1-5	47-81	1-6

Cluster 4 seems to be the most interesting one, because of its high value for monetary volume. The cluster is characterized by married male workers, who seem to visit the supermarket very often. The majority of the people in this cluster is interested in receiving discounts, so it might be a good idea to send them more personalized offers in the future. Note that cluster 4 is the smallest one among all clusters, so there's only a small group of loyal customers.

Cluster 2 is characterized by a particular type of people: most of them are married and retired. They are less loyal customers as the ones contained in cluster number 4, but they may be an interesting target for promotional offers given their specific characteristics.

The last interesting cluster is cluster number 1. This group of customers contains mostly female bachelors of different ages. Based on the number of visits and monetary volume, they are in the top three of the best clusters.

2 EVENT CHARACTERIZATION

In this paragraph we analyze the event of churning in more detail. For the supermarket it is important to learn what the typical behavior of churning customers is. This information can be used to prevent customers from churning in the future. To do this, we build a classifier which is able to predict if a customer is going to churn or not.

First we collect the data that is needed to build the classifier. In the previous part of the project the churning customers were put in a file as follows:

CLIENTE_ID	MESE_N	Number of visits	TIPOLOGIA_ID
1146	5	1	7
8363	5	1	7
8504	7	1	7

We say that a customer is churning when the customer has bought something in a month, but didn't buy anything in the next 2 months [1]. The customer with ID 1146 churns in month 5 because he doesn't buy anything in months 6 and 7. In this table only customers who churned were displayed. The classifier, however, needs to be built using data for all the months for each customer, not only the months in which (s)he churned. For this reason we developed a new SSIS project to compute for each customers the aggregates for each month and then joined them with the table of the previous project by adding a column to each row 'IsChurning'. IsChurning can contain only two different values 'yes' and 'no'. If the customer is churning in a certain month the values for 'IsChurning' is set to 'yes' otherwise to 'no'.

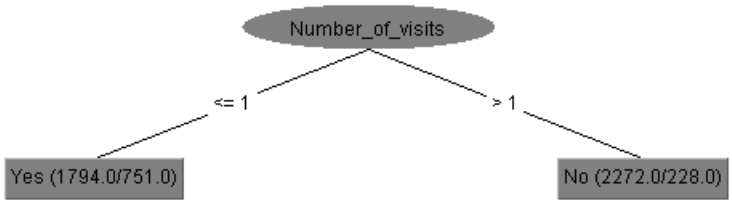
Next, we can load the data into Weka for performing the classification task and further preprocessing. The first preprocessing task is to remove the 'CLIENTE_ID' column, because this will be useless for classifying. An important remark needs to be made concerning the data used for classification. A customer churns if he buys something in a month, but not in the next two months. This means that a customer who is churning in month 7, doesn't buy anything the next two months. The data provided by Coop contained information about the sales of months 4 to 9. Hence we can't tell if a customer churns in month 8 or 9. In Weka we apply the 'remove with values' filter to remove the months 8 and 9. After the filtering 6099 instances remain.

The next step is to split the data set into training set and test set. First we randomize the data and then we split the data set in two parts: a training set, which contains 66% of the data and a test set, which contains 33% of the data by using the following preprocessing steps in Weka:

- Randomize
- Stratified remove folds (2 times)

After the preprocessing, we use the J48 algorithm for decision trees providing our test set to evaluate the outcome. The classification is done on the class variable 'isChurning' provided with 3 other variables: Number of visits, Monetary volume, Number of different products. The resulting tree contains only one split based on the attribute 'Number of visits' (shown in the figure below). We can interpret the tree using the decision rule: if a customer visits the shop more than once a month, (s)he is not churning, otherwise (s)he is churning. The decision tree is rather trivial and indicates that the number of visits is a fairly good classifier. As one can see in the summary below, almost 75% of the instances were classified correctly. It is important that we can recognize the churning customes in order to prenvent them from leaving. The recall for churning is 81% in this model, indicating that 81% of the people who churn are classified this way by the model.

[1] This is slightly different from the objective for project 1 that stated the **last** n months. We think that the churn analysis is more complete when we look at all cases, not only the last n months. So for example if a customer buys something in April and nothing in the **next** 2 months, (s)he is churning.



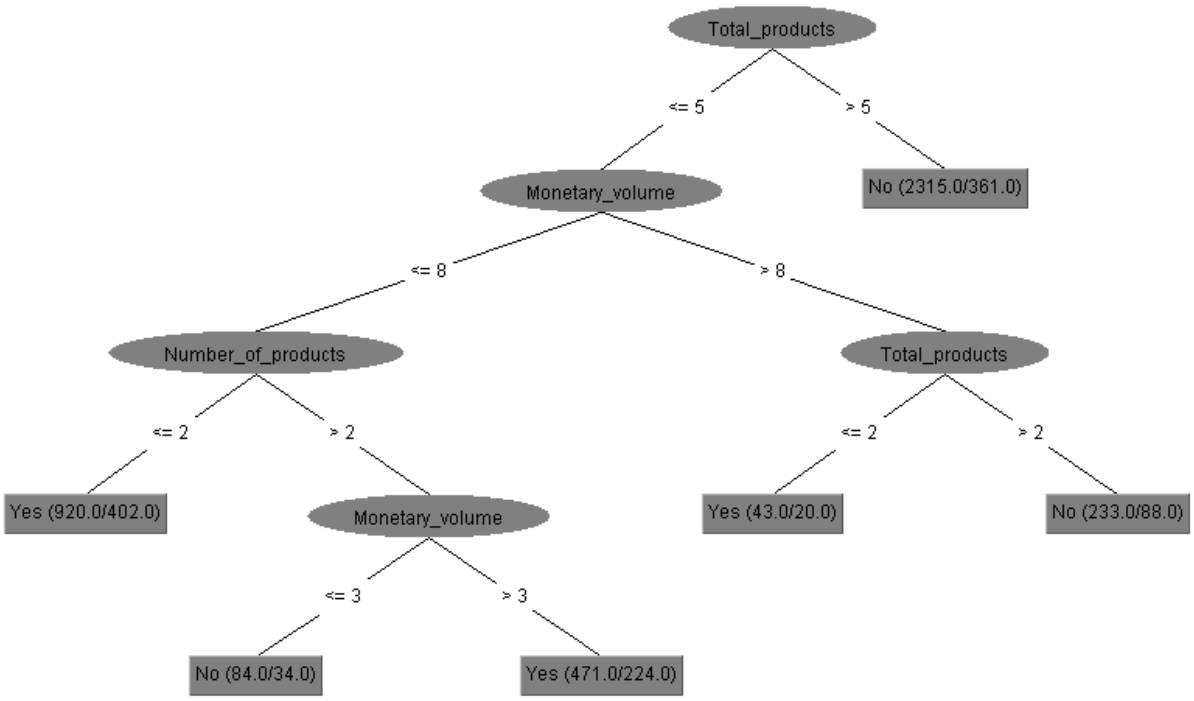

```
=== Summary ===

Correctly Classified Instances        1523               74.9139 %
Incorrectly Classified Instances       510               25.0861 %
Kappa statistic                          0.4767
Mean absolute error                      0.3204
Root mean squared error                  0.4025
Relative absolute error                 74.5714 %
Root relative squared error             86.8469 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances             2033

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,811    0,279    0,569      0,811    0,669      0,496    0,766     0,521     Yes
                 0,721    0,189    0,894      0,721    0,798      0,496    0,766     0,836     No
Weighted Avg.    0,749    0,217    0,792      0,749    0,758      0,496    0,766     0,738

=== Confusion Matrix ===

    a    b   <-- classified as
  515  120 |    a = Yes
  390 1008 |    b = No
```

If we remove the variable number of visits, we can try to extract a different decision tree. This effort yields no better result and so is not shown here.

Finally we try to classify using a slightly different data set. Here another aggregate is included: the total number of products bought (different from the already included 'Number of products' which indicates the number of **different** products). The new aggregate is called 'Total products'. Next we try to build a decision tree using:

- Number of products
- Monetary volume
- Total products
- Class isChurning

We have set the parameters of the J48 algorithm in such a way that the tree is pruned such that no leaf contains less than 10 instances. As a result (figure below) a tree is obtained with six leaves and so six decision rules, starting at the root 'Total products'. One can see some trivial relationships, for example: the more products a customer buys and the more money he spends, the less likely it is for the customer to churn. The evaluation of the tree on the test set shows (figure below) that this model performs worse than the previous one. Accuracy dropped to 69,5% and recall of churning customers dropped even more (to 58%).




```
=== Summary ===

Correctly Classified Instances        1413               69.5032 %
Incorrectly Classified Instances       620               30.4968 %
Kappa statistic                          0.3165
Mean absolute error                      0.3675
Root mean squared error                  0.4314
Relative absolute error                 85.5293 %
Root relative squared error             93.0753 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances             2033

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,583    0,254    0,510      0,583    0,544      0,318    0,705     0,463     Yes
                 0,746    0,417    0,797      0,746    0,771      0,318    0,705     0,807     No
Weighted Avg.    0,695    0,366    0,708      0,695    0,700      0,318    0,705     0,699

=== Confusion Matrix ===

   a    b   <-- classified as
 370  265 |    a = Yes
 355 1043 |    b = No
```

Based on the accuracy and recall, we can conclude the first simple model was the best. According to that model we can tell based on the number of visits if a customer churns or not. Still this makes us only distinguishing between sporadic customers who visit the supermarket only once and customers who visit more than once. In this last group there are also some customers who churn, wrongly classified by the model as not churning. These particular cases might be of interest to the supermarket that wants to prevent customers from leaving, but especially the ones who are not only sporadic customers. The second model suggests to look first at the total number of products, and later on to monetary volume of purchases and number of different products to distinguish between churning and non churning customers. Yet again a significant part of customers who buy a greater total number of products and buy for a greater total monetary volume are wrongly classified as not

churning. Finally we can conclude we lack other data about the behavior specifically observed in case customers churn.

3 INNOVATORS

To be able to detect the early adopters for different products, we start by writing a query that gives us an overview of the number of times the product was sold in each week:

```
SELECT V.ARTICOLO_ID,A.DES_ART,D.SETTIMANA_ANNO, COUNT (DISTINCT CLIENTE_ID) AS
Number_of_purchasers
FROM [tj].[dbo].[SSET_VEN_CORSODM] AS V INNER JOIN [tj].[dbo].[SSET_ART_CORSODM] AS A ON
V.ARTICOLO_ID=A.ARTICOLO_ID INNER JOIN [tj].[dbo].[SSET_MKT_CORSODM] AS M ON
A.COD_MKT_ID= M.COD_MKT_ID INNER JOIN [tj].[dbo].[SSET_DATA_DORSODM] AS D ON
V.DATA_ID = D.DATA_ID
GROUP BY V.ARTICOLO_ID,A.DES_ART,SETTIMANA_ANNO
ORDER BY V.ARTICOLO_ID,SETTIMANA_ANNO
```

First we join the tables SSET_VEN_CORSODM, SSET_ART_CORSODM, SSET_MKT_CORSODM and SSET_DATA_DORSODM to have access to all information (customer, transactional, product and date) we need. Next, we select the ARTICOLO_ID, the DES_ART, the SETTIMANA_ANNO and the number of distinct purchasers for each article in each week, by grouping on ARTICOLO_ID, DES_ART and SETTIMANA_ANNO. The result of the query is (partly) displayed in the figure below.

	ARTICOLO_ID	DES_ART	SETTIMANA_ANNO	Number_of_purchasers
1	105	SUS.PRESIDENT IT 40-50 I^CT K2	36	12
2	105	SUS.PRESIDENT IT 40-50 I^CT K2	37	11
3	105	SUS.PRESIDENT IT 40-50 I^CT K2	38	1
4	235	CONF. 8 PZ. LUMINO VOTIVO MOD. 5B	39	1
5	488	BIDET 40/60 UNICOLOR CON APPEND.	38	2
6	488	BIDET 40/60 UNICOLOR CON APPEND.	39	1
7	489	TELO BAGNO UNICOLOR C/APPEND..	39	1
8	587	BORSA TERMICA ST.COOP CM50X50+8 G91,06 COMPRESA M...	14	1
9	587	BORSA TERMICA ST.COOP CM50X50+8 G91,06 COMPRESA M...	15	1
10	587	BORSA TERMICA ST.COOP CM50X50+8 G91,06 COMPRESA M...	20	1
11	587	BORSA TERMICA ST.COOP CM50X50+8 G91,06 COMPRESA M...	21	2

For instance for the product with ARTICOLO_ID 105 we denote a decreasing trend. By investigating the resulting table, we can search for a specific pattern: if the number of distinct purchasers is low in the first week that the product was sold, and increases in the weeks after, then the customers who already bought the product in the first week are early adopters. Scanning the table for this pattern identifies some products that are adopted early by some customers and adopted later by others.

The products with ARTICOLO_ID 1745 and 9865 are adopted by more and more customers over time.

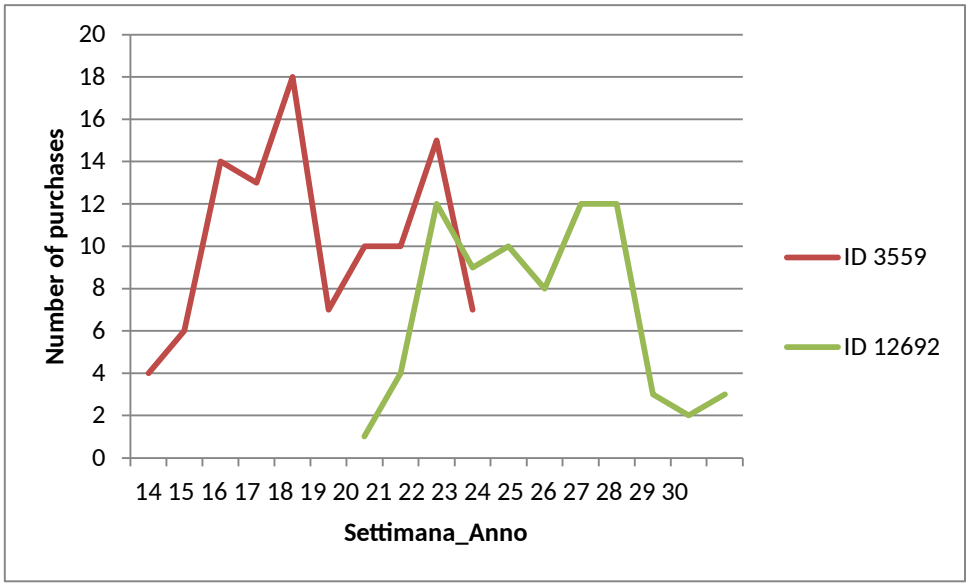
	ARTICOLO_ID	DES_ART	SETTIMANA_ANNO	Number_of_purchasers
1	1745	GUANCIALE IN MEMOFOAM -SCHIUMA VISCOELASTICA	35	2
2	1745	GUANCIALE IN MEMOFOAM -SCHIUMA VISCOELASTICA	36	3
3	1745	GUANCIALE IN MEMOFOAM -SCHIUMA VISCOELASTICA	37	5

	ARTICOLO_ID	DES_ART	SETTIMANA_ANNO	Number_of_purchasers
355	9865	BIRRA HEINEKEN 5 GRADI 6X500ML	26	2
356	9865	BIRRA HEINEKEN 5 GRADI 6X500ML	27	12

The products with ARTICOLO_ID 3559 and 12692 show a different trend: first, the number of people who buy these products increases over time. Later on, there is a decrease, that can be explained as follows: in a first stage, more and more people buy the product, maybe influenced by an advertisement or promotion. Once they have used the product, they might experience that the product is not what they expected and decide to not buy the product again.

	ARTICOLO_ID	DES_ART	SETTIMANA_ANNO	Number_of_purchasers
184	3559	BIBITA COCA COLA LATTINA ML.330 X6	20	1
185	3559	BIBITA COCA COLA LATTINA ML.330 X6	21	4
186	3559	BIBITA COCA COLA LATTINA ML.330 X6	22	12
187	3559	BIBITA COCA COLA LATTINA ML.330 X6	23	9
188	3559	BIBITA COCA COLA LATTINA ML.330 X6	24	10
189	3559	BIBITA COCA COLA LATTINA ML.330 X6	25	8
190	3559	BIBITA COCA COLA LATTINA ML.330 X6	26	12
191	3559	BIBITA COCA COLA LATTINA ML.330 X6	27	12
192	3559	BIBITA COCA COLA LATTINA ML.330 X6	28	3
193	3559	BIBITA COCA COLA LATTINA ML.330 X6	29	2
194	3559	BIBITA COCA COLA LATTINA ML.330 X6	30	3

	ARTICOLO_ID	DES_ART	SETTIMANA_ANNO	Number_of_purchasers
788	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	14	4
789	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	15	6
790	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	16	14
791	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	17	13
792	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	18	18
793	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	19	7
794	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	20	10
795	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	21	10
796	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	22	15
797	12692	NETTARE COOP DI PESCA CLUSTER ML.200X3	23	7



Now that we have identified some products that exhibit a trend pointing to early adopting, it's easy to identify the customers who are the early adopters of these products. We write a query of the form:

```
SELECT V.CLIENTE_ID,C.ANNO_NASCITA,C.SESSO,C.STATO_CIVILE,C.PROFESSIONE,C.ANNO_SOCIO,
V.ARTICOLO_ID,A.DES_ART,D.SETTIMANA_ANNO
FROM [tj].[dbo].[SSET_VEN_CORSODM] AS V INNER JOIN [tj].[dbo].[SSET_ART_CORSODM] AS A ON
V.ARTICOLO_ID=A.ARTICOLO_ID INNER JOIN [tj].[dbo].[SSET_CLIENTE_CORSODM] AS C ON
C.CLIENTE_ID=V.CLIENTE_ID INNER JOIN [tj].[dbo].[SSET_MKT_CORSODM] AS M ON
A.COD_MKT_ID= M.COD_MKT_ID INNER JOIN [tj].[dbo].[SSET_DATA_DORSODM] AS D ON
V.DATA_ID = D.DATA_ID
WHERE V.ARTICOLO_ID=? AND D.SETTIMANA_ANNO=?
```

Where the first question mark denotes the ARTICOLO_ID of the product bought and the second question mark denotes the SETTIMANA_ANNO in which the early adopter bought the product. The query also makes it

possible to see the characteristics of the early adopters. For instance if we want to look up the customer who was an early adopter of article 2559 by buying it in week 20 already, we get the result below:

	CLIENTE_ID	ANNO_NASCITA	SESSO	STATO_CIVILE	PROFESSIONE	ANNO_SOCIO	ARTICOLO_ID	DES_ART
1	494273	1983	Uomo	Celibe	OPERAIO	2002	3559	BIBITA COCA COLA

The query result gives us the customer id, as well as the characteristics of the customer (ANNO_NASCITA, SESSO, STATO_CIVILE, PROFESSIONE and ANNO_SOCIO) and the article for which the customer is an early adopter.

4 FREQUENT PATTERNS

We started by performing some preprocessing steps: we selected the top cluster, cluster 4 and for the clients in this cluster, we selected all the transactions. In order to perform frequent pattern analysis in Weka, we need to apply the association rule algorithm that is called 'Apriori'. Because the loaded date was in the tabular form, there were only 2 columns: 'SCONTRINO' and 'SEGMENTO'. Since the Apriori algorithm can't handle data in tabular form, we applied the filter called 'denormalize' to transform the data into a binary format. Next, we removed the 'SCONTRINO' column and applied the association rule algorithm to the remaining 231 columns, each representing a different segment to find association rules. We chose to mine for patterns at the level of segments, because a segment contains multiple products and so it's more likely to find patterns at the level of segments than at the level of products.

After the preprocessing, we needed to decide on the parameters of the Apriori algorithm. We want to find rules with medium or low support, for a confidence level as high as possible. After some trial and error, it seems that there can only be extracted rules with low support (6% or lower) and a confidence level not higher than 55%. The best ten rules are shown below:

```
Apriori
=======

Minimum support: 0.01 (395 instances)
Minimum metric <confidence>: 0.3
Number of cycles performed: 99

Generated sets of large itemsets:

Size of set of large itemsets L(1): 66

Size of set of large itemsets L(2): 53

Size of set of large itemsets L(3): 4

Best rules found:

 1. SEGMENTO_COMUNE=t SEGMENTO_SALSICCE/INSACCATI FRESCHI=t 847 ==> SEGMENTO_LAVORAZIONI INTERNE=t 466    <conf:(0.55)> lift:(2.26) lev:(0.01) [259] conv:(1.68)
 2. SEGMENTO_SALSICCE/INSACCATI FRESCHI=t 2342 ==> SEGMENTO_LAVORAZIONI INTERNE=t 1247    <conf:(0.53)> lift:(2.18) lev:(0.02) [676] conv:(1.62)
 3. SEGMENTO_PIZZA BIANCA - FOCACCIA=t SEGMENTO_LAVORAZIONI INTERNE=t 1024 ==> SEGMENTO_COMUNE=t 534    <conf:(0.52)> lift:(1.35) lev:(0) [137] conv:(1.28)
 4. SEGMENTO_PIZZA BIANCA - FOCACCIA=t 4886 ==> SEGMENTO_COMUNE=t 2477    <conf:(0.51)> lift:(1.31) lev:(0.01) [586] conv:(1.24)
 5. SEGMENTO_CONDITO=t 1060 ==> SEGMENTO_COMUNE=t 536    <conf:(0.51)> lift:(1.31) lev:(0) [125] conv:(1.24)
 6. SEGMENTO_PIZZA INTERA=t 1623 ==> SEGMENTO_COMUNE=t 757    <conf:(0.47)> lift:(1.21) lev:(0) [128] conv:(1.15)
 7. SEGMENTO_COSCE=t 1094 ==> SEGMENTO_COMUNE=t 446    <conf:(0.41)> lift:(1.05) lev:(0) [22] conv:(1.03)
 8. SEGMENTO_POMODORO OBLUNGO VERDE=t 1254 ==> SEGMENTO_COMUNE=t 510    <conf:(0.41)> lift:(1.05) lev:(0) [24] conv:(1.03)
 9. SEGMENTO_LAVORAZIONI INTERNE=t SEGMENTO_BANANE=t 1156 ==> SEGMENTO_COMUNE=t 466    <conf:(0.4)> lift:(1.04) lev:(0) [18] conv:(1.03)
10. SEGMENTO_POMODORO ROSSO A GRAPPOLO=t 1722 ==> SEGMENTO_COMUNE=t 691    <conf:(0.4)> lift:(1.04) lev:(0) [24] conv:(1.02)
```

We can tell from the figure that the two most frequent segments in the dataset, 'lavorazioni interne' and 'comune' always appear in the consequent. So we can't really extract useful information from the obtained rules. When we try to obtain more useful rules by removing those two frequent segments, the algorithm is not able to find any rules anymore. Unfortunately, we have to conclude that the clusters don't contain useful information about frequent patterns.

5 PRIVACY EVALUATION

To evaluate possible privacy issues in this data mining project, we need to describe what personal data we have at our disposal. First of all there is a table containing data about specific customers with the following attributes:

- Client-id
- Birth year
- Sex
- Marital status
- Profession
- Year of membership
- Interest in promotion

We assume the data is already made anonymous by removing names, addresses and place of birth to protect the privacy of customers. This of course, makes it impossible for a data analyst to **directly** identify a specific customer with the data available. Next we can ask ourselves if there are other quasi-identifiers in the data set provided. The birth year, sex, marital status and profession qualify, but seem to be too unspecific to possibly identify a person: many people in the data set have the same age, profession, marital status and sex. Even if you find a person with a unique combination of these attributes, it seems unlikely that you can identify this person without information about postal codes or names. We can conclude that the supermarket provided us with a data set precise enough to perform data mining tasks, but reasonable imprecise to prevent a spy from extracting personal shopping habits.