CHURN ANALYSIS ON LastFM LISTENINGS

ADVANCED TOPICS ON DATA MINING AND CASE STUDIES

Prof. Mirco Nanni

Prof. Anna Monreale

Simay KOCAN-525804

Hüseyin Varol ERDEM-525822

2015-2016

1. BUSINESS UNDERSTANDING

In LastFM data, as business objective, it is aimed to increase number of users who listen to a chosen genre. “Rock” genre was chosen to analyze. Data mining objectives were finding which kind of users listen to it and which kind of users abandoned it also changed their preferences, finding a group of users with similar behavior.

2. DATA UNDERSTANDING

There are 3 csv files; listening, genre and network.

Listening (user_id, date, track, artist, album):

This file includes 14.650.594 rows with some noisy and missing values. There are 4.041.345 duplicated rows (all attributes values are same). In addition, some data have same user_id and date with different tracks and also some rows have no date values. All of them were cleaned up data preparation phase.

Genre (artist, genre of artist):

This file includes 138.033 rows with 613 duplicates. There are no missing values.

Network (user_id1, user_id2):

There are 2.543.288 rows and also 11.257 duplicated data. Some of user_id2 are in the listening table some of them are not. In addition, there are no missing values.

3. CHURN ANALYSIS

Before applying the churn analysis, the transactional data must be prepared. For all aggregation to create new variable, MSSQL Management Studio was used. The new data set that is result of queries was used in R to build a model.

3.1 DATA PREPARATION

Firstly, duplicate rows were eliminated in all three tables for better performance by using a program (Delimit) which allows us to handle with lots of rows. After elimination in listening file, there remained 10.609.249 rows, in genre file, there remained 137.420 rows and in network file, there remained 2.532.031 rows. When csv. file were reading into a table, there were some problems when the data was separated.

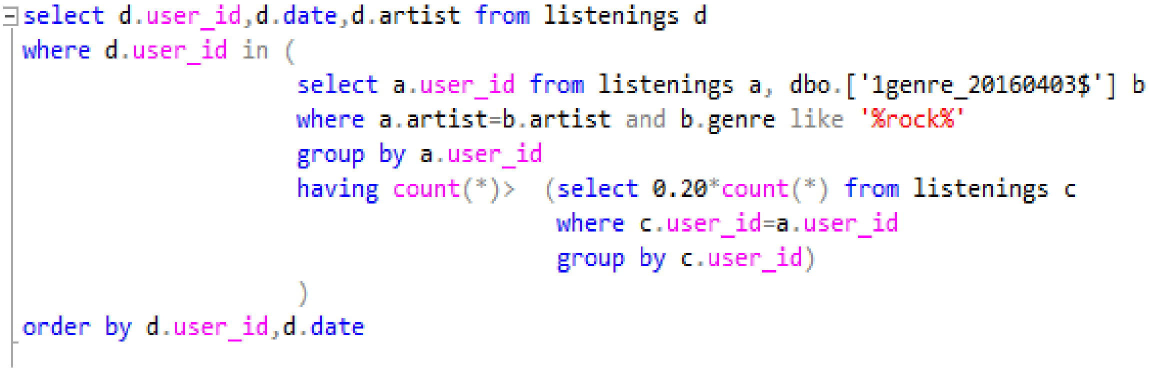
In listening file when file was separated with comma, some files were separated wrong because, e.g the track name or artist name may include “,” (In 736.838 rows). These types of commas have a space character after themselves. So, “,” remained same but “,(space)“ were changed with “-“ for a true reading into a table. The second problem is in listening files, the commas inside of numbers. They were changed with “.”. E.g. 10,000 to 10.000. The third one, since some data have null date, they were eliminated. Some rows have same user_id and date but track names are different. Thus, they were not eliminated from program and could not be deleted. It may be different meaning. For example, maybe a user can open different songs in same time on different computers. It is not known that it is allowed or not. After these arranges, the data was split on 11 excel file with each of them has 1.000.000 except the last one to import to Microsoft Sql Server for handling data.

In genre file, after duplicate elimination, the same process (changing “,(space)” with “-“ was performed only and import to Microsoft Sql Server with one file because of the number of rows.

In network file, after duplicate elimination, the data was split in 3 excel files and they were imported to Microsoft Sql Server.

3.2 DATA SELECTION

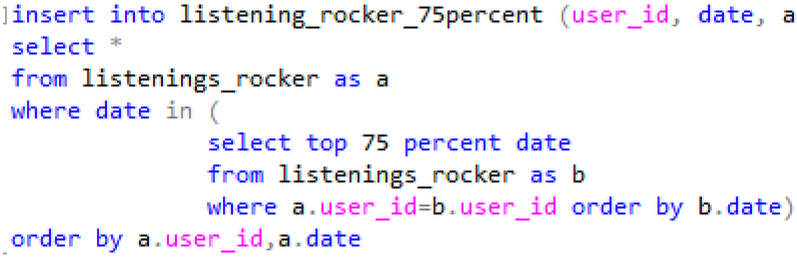
According to the definition of our churn analysis, firstly users who listened rock tracks were selected. In our analysis, the users who have never listened rock track are not important. Then, it was thought that the users who listened very few number of rock tracks also are not so important. Thus, the users who listened rock tracks which are 20% of total listenings, were chosen. After this step, 39.927 users were remained. To make this, the query is shown below was executed and the result of this query was inserted to new table (listenings_rocker).



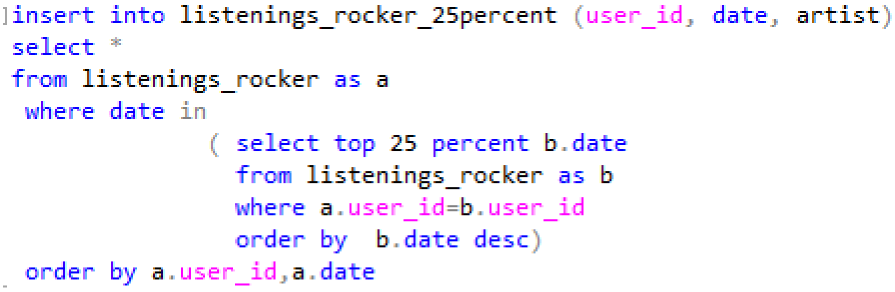
```
select d.user_id,d.date,d.artist from listenings d
where d.user_id in (
                    select a.user_id from listenings a, dbo.['1genre_20160403$'] b
                    where a.artist=b.artist and b.genre like '%rock%'
                    group by a.user_id
                    having count(*)>  (select 0.20*count(*) from listenings c
                                       where c.user_id=a.user_id
                                       group by c.user_id)
                    )
order by d.user_id,d.date
```

Predictive and Target Variables Selection

For every user, the date is in different range. Thus, when date range was separated, predictive time was calculated by taking first 75% of date and target time was calculated by taking last 25% of date for every user. The data was split according to these time ranges. The query is shown below.



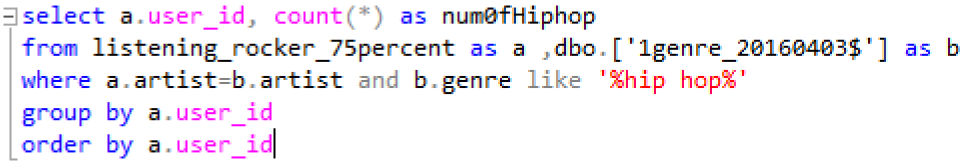
```
insert into listening_rocker_75percent (user_id, date, artist)
select *
from listenings_rocker as a
where date in (
              select top 75 percent date
              from listenings_rocker as b
              where a.user_id=b.user_id order by b.date)
order by a.user_id,a.date
```



```
insert into listenings_rocker_25percent (user_id, date, artist)
select *
from listenings_rocker as a
 where date in
              ( select top 25 percent b.date
                from listenings_rocker as b
                where a.user_id=b.user_id
                order by  b.date desc)
 order by a.user_id,a.date
```

Predictive variables are:

- User_id's , number of total listenings, number of rock tracks, number of pop tracks, number of metal tracks, number of electronic tracks, number of punk tracks, number of hip-hop tracks, number of friends who listen rock in rate 20% of total listenings. Below, there is a query to aggregate the data group by user_id.



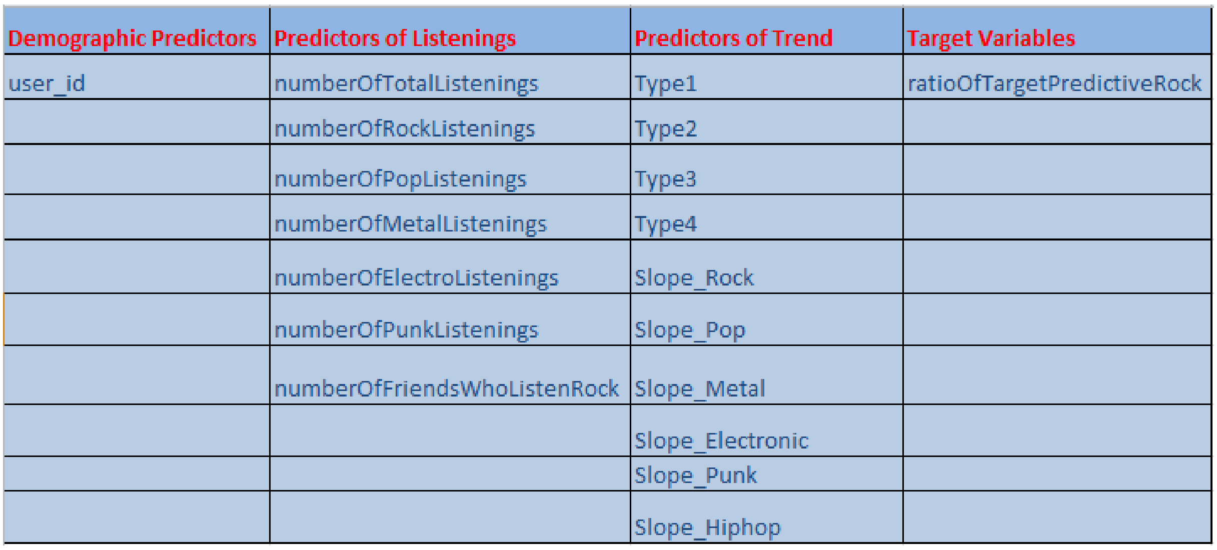
```
select a.user_id, count(*) as num0fHiphop
from listening_rocker_75percent as a ,dbo.['1genre_20160403$'] as b
where a.artist=b.artist and b.genre like '%hip hop%'
group by a.user_id
order by a.user_id
```

- For regression analysis (to understand user's trend to genres) predictive period was split in four time ranges. Then in these ranges, number of listenings of every genre was calculated for every user. According to these numbers regression lines and their slopes were calculated for understanding users tendency (increasing if slope is positive, decreasing if it is negative)
- The last predictive variables are for understanding the type of the users in these four periods. Type1, type2, type3 and type 4 attributes shows that ratio of number of rock listenings and total listenings in these periods.

It was tried to learn user's friends tendency, however from this data, it could not be inferred this information because, for every user, the timestamp is different. It is not known any information about in these periods they are friends or not.

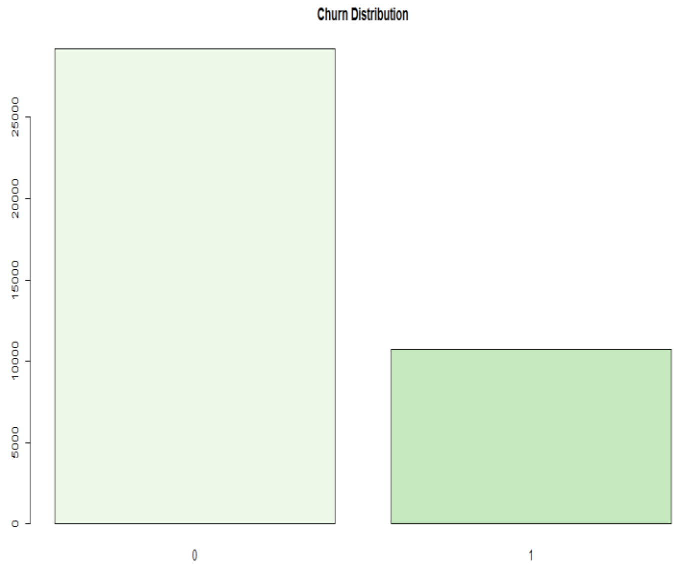
Ratio between number of rock listening in the target period and number of rock listening in the predictive period was taken for alarm value. If this values is smaller than 0.20 (40% decay), it shows us there are dramatic decrease for user's listenings on rock tracks. Alarm values were discretized to "0" (is not alarm) and "1" (alarm).

The table shows that all variables for churn analysis.



Demographic Predictors	Predictors of Listenings	Predictors of Trend	Target Variables
user_id	numberOfTotalListenings	Type1	ratioOfTargetPredictiveRock
	numberOfRockListenings	Type2	
	numberOfPopListenings	Type3	
	numberOfMetalListenings	Type4	
	numberOfElectroListenings	Slope_Rock	
	numberOfPunkListenings	Slope_Pop	
	numberOfFriendsWhoListenRock	Slope_Metal	
		Slope_Electronic	
		Slope_Punk	
		Slope_Hiphop	

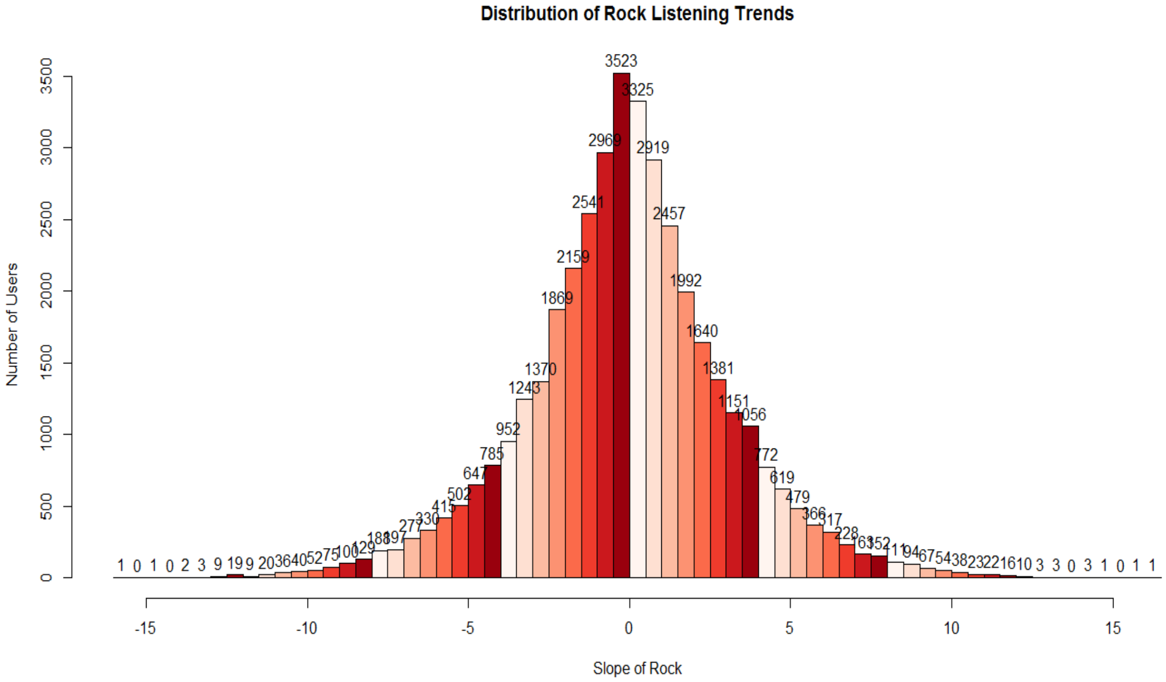
The table shows churn distribution of the discretized variables,



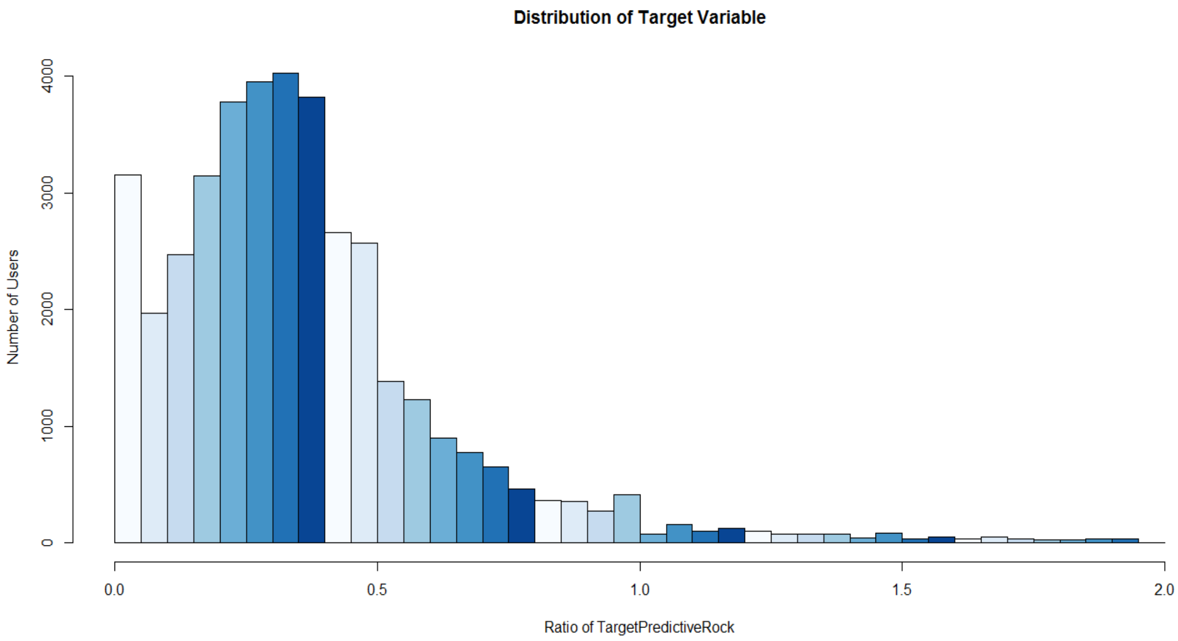
Number of users who are churn : 10735

Number of users who are not churn:

29192



Above table shows rock listeners trend. When looking for distribution of slope, there are many user whose number of listenings rock trends are decreasing time by time. Thus churn analysis can be so important for rock genre.



In our churn analysis, users who are less than %20 were chosen as a churn.

3.3 MODELLING

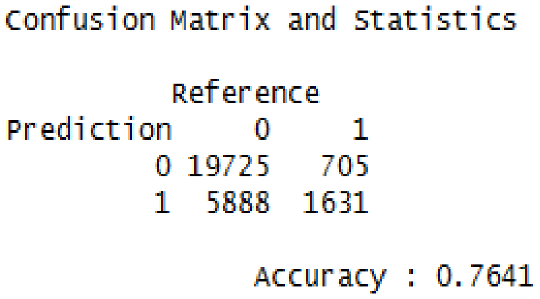
In this project, two types of classification model were used. 1) with Decision tree and 2)with Neural Network

DECISION TREE

When the decision trees were produced, as a classification algorithm, CART algorithm was used. After many tries, the best performance tree was selected. All predictive variables were used for modeling. The data was split two parts: 70% train set and 30% test set. This process was made by using R.

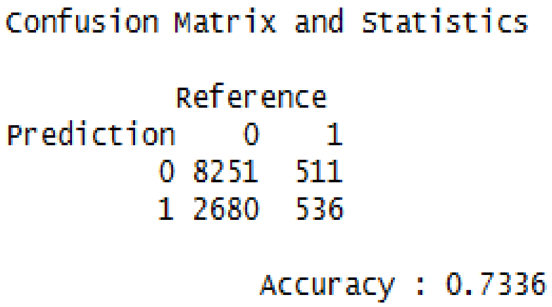
Performance of Classifier

The model has accuracy 0.7641 on training set



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19725   705
         1  5888  1631

               Accuracy : 0.7641
```

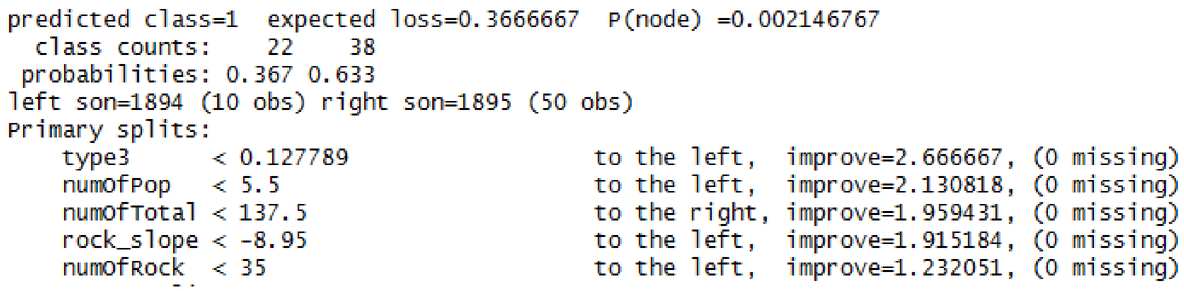
The model has accuracy 0.7336 on test set



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8251  511
         1 2680  536

               Accuracy : 0.7336
```

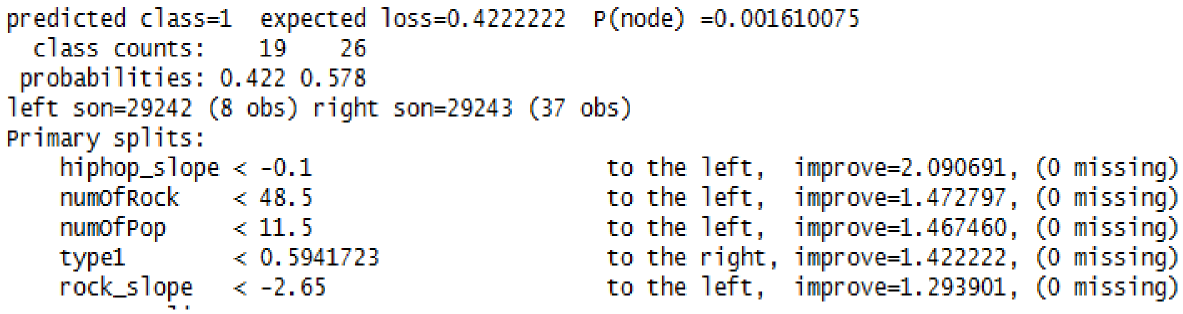
Gain is = 536/ (536+2680) = 16.7%

Sample Classification Rules



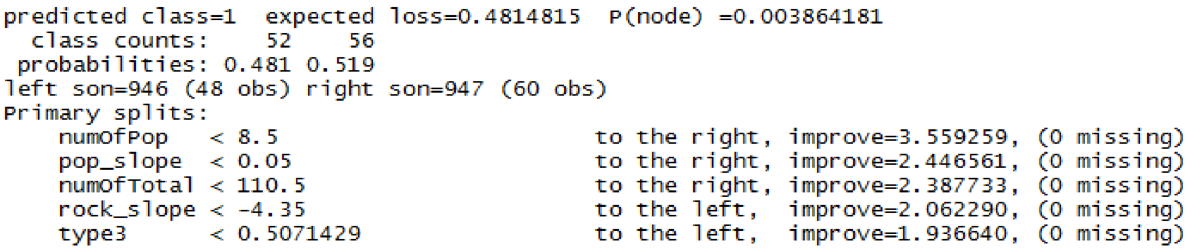
```
 predicted class=1  expected loss=0.3666667  P(node) =0.002146767
   class counts:    22    38
  probabilities: 0.367 0.633
 left son=1894 (10 obs) right son=1895 (50 obs)
 Primary splits:
     type3      < 0.127789                to the left,  improve=2.666667, (0 missing)
     numOfPop   < 5.5                     to the left,  improve=2.130818, (0 missing)
     numOfTotal < 137.5                   to the right, improve=1.959431, (0 missing)
     rock_slope < -8.95                   to the left,  improve=1.915184, (0 missing)
     numOfRock  < 35                      to the left,  improve=1.232051, (0 missing)
```

If type3 < 0.127789 & numOfPop <5.5 & numOfTotal <137.5 & rock_slope <-8.95 & numOfRock <35 then it is churn, (Confidence = 63.3%)



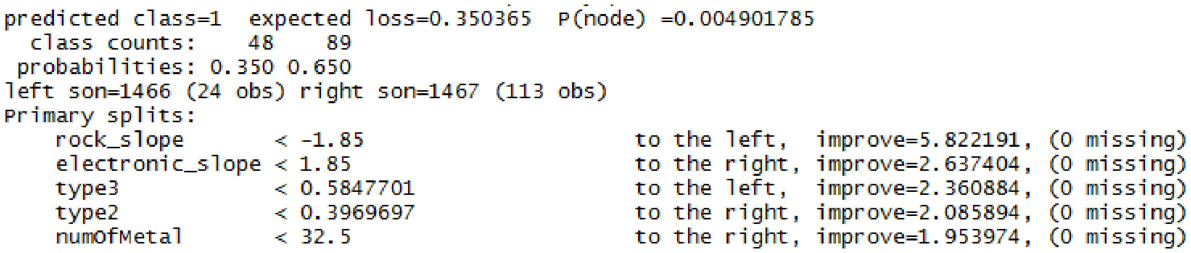
```
 predicted class=1  expected loss=0.4222222  P(node) =0.001610075
   class counts:    19    26
  probabilities: 0.422 0.578
 left son=29242 (8 obs) right son=29243 (37 obs)
 Primary splits:
     hiphop_slope < -0.1                  to the left,  improve=2.090691, (0 missing)
     numOfRock    < 48.5                  to the left,  improve=1.472797, (0 missing)
     numOfPop     < 11.5                  to the left,  improve=1.467460, (0 missing)
     type1        < 0.5941723             to the right, improve=1.422222, (0 missing)
     rock_slope   < -2.65                 to the left,  improve=1.293901, (0 missing)
```

If hiphop_slope <0.1 & numOfRock <48.5 & numOfPop <11.5 & type1<0.5941723 & rock_slope < -2.65 then it is churn (Confidence = 57.8 %)



```
 predicted class=1  expected loss=0.4814815  P(node) =0.003864181
   class counts:    52    56
  probabilities: 0.481 0.519
 left son=946 (48 obs) right son=947 (60 obs)
 Primary splits:
     numOfPop   < 8.5                     to the right, improve=3.559259, (0 missing)
     pop_slope  < 0.05                    to the right, improve=2.446561, (0 missing)
     numOfTotal < 110.5                   to the right, improve=2.387733, (0 missing)
     rock_slope < -4.35                   to the left,  improve=2.062290, (0 missing)
     type3      < 0.5071429               to the left,  improve=1.936640, (0 missing)
```

If numOfPop < 8.5 & pop_slope <0.05 & numOfTotal <110.5 & rock_slope < -4.35 & type3 < 0.5071429 then it is churn (Confidence = 51.9%)



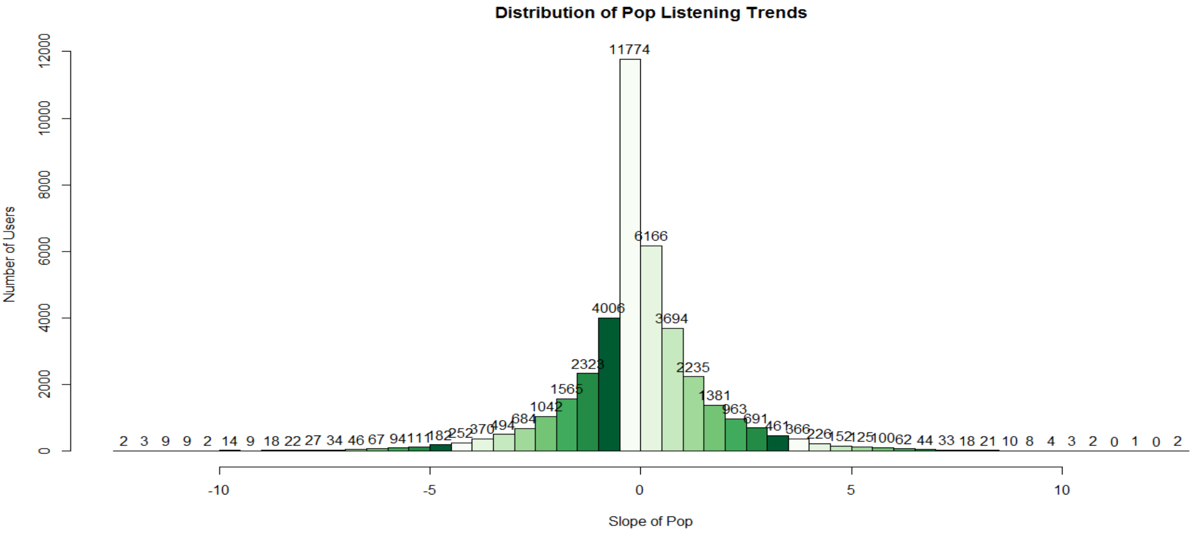
```
 predicted class=1  expected loss=0.350365  P(node) =0.004901785
   class counts:    48    89
  probabilities: 0.350 0.650
 left son=1466 (24 obs) right son=1467 (113 obs)
 Primary splits:
     rock_slope       < -1.85             to the left,  improve=5.822191, (0 missing)
     electronic_slope < 1.85              to the right, improve=2.637404, (0 missing)
     type3            < 0.5847701         to the left,  improve=2.360884, (0 missing)
     type2            < 0.3969697         to the right, improve=2.085894, (0 missing)
     numOfMetal       < 32.5              to the right, improve=1.953974, (0 missing)
```

If rock_slope < -1.85 & electronic < 1.85 & type3 < 0.5847701 & type2 < 0.3969697 & numOfMetal < 32.5 then it is churn (Confidence = 65%)

It seems that, rock_slope (users tendency to rock time by time) is important factor when decision trees were built. Generally negative rock_slopes (regression line is decreasing) decided who is churn or not. It shows that churn users tend to decrease listening rock track in predictive period.

Another important predictive variable is type3 which is ratio of number of rock listenings and total listenings in 3rd quarter of predictive period. It shows that, in this quarter, users whose rock listenings is fewer than about 0.50 of total listenings tend to be churn.

The last important variables are number of rock listenings and number of pop listenings. When user's rock listenings is less than 48.5 (users who listens rock sometimes), they tend to be churn. Another interesting variable is number of pop listening. It seems that, when it is less than about 10 people tend to be a churn. Below gaph shows that distribution of pop_slope.



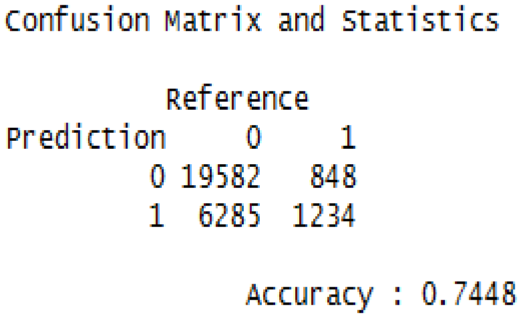
Firstly, number of friends is thought that, they are significant measure. However, when decision tree were built, they were not used.

NEURAL NETWORK

When neural network models were produced, as a function of output, softmax was used. The data was split two parts : 70% train set and 30% test set again. After many tries, the best performance tree was selected with these parameters: number of hidden neurons = 40 and maximum iteration = 150 with 842 weights. All predictive variables were used for modeling. This process was made by using R.

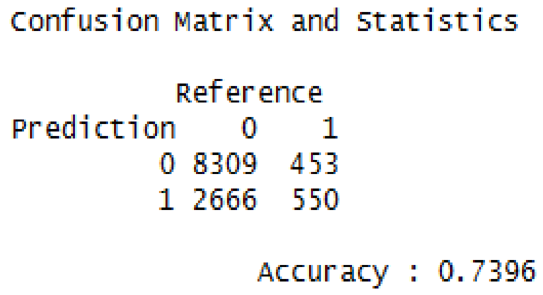
Performance of Classifier

The model has accuracy 74.4% on training set:



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19582   848
         1  6285  1234

               Accuracy : 0.7448
```

The model has accuracy 73.96% on test set:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8309  453
         1 2666  550

               Accuracy : 0.7396
```

Gain is = 550/ (550+2666) = 17.11 %

4. CUSTOMER SEGMENTATION

LastFM data was used to find groups of users with similar behavior. In order to make this, firstly, transaction data was prepared and then K-Means clustering algorithm was applied. For the all part of customer segmentation, KNIME Analytics Platform was used.

4.1 DATA PREPARATION

Before aggregating the transaction data, there are some date values in the listenings table with the value of NULL and some date values had incorrect type (such as 3, 4265) also some values corresponded to future date time. Therefore, these data were removed from the analysis. By joining the listening- network-genre and aggregating the transaction data, "shareholder "value variables were created:

- Number of total listenings
- Number of "rock" listenings
- Number of "pop" listenings
- Number of "jazz" listenings
- Number of "electronic" listenings
- Number of "punk" listenings
- Number of "metal" listenings
- Number of "hip-hop" listenings
- Number of friends
- Number of Distinct albums (that the user listen)
- Number of Distinct artist (that the user listen)
- Difference of timestamp (max-min)

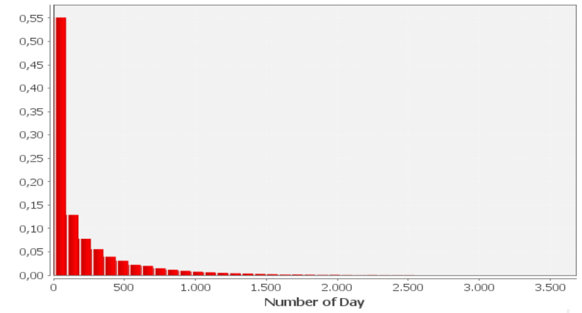
After then, some transformations were applied to create new variables with above ones. To calculate number of days since the user first listened in the LastFM, difference of timestamp was divided by the value of 'one day' (86400). The results that are smaller than 1 were transformed to 1. The second transformation was average of total listenings over number of days. It was ratio between number of total listenings and number of day. This ratio can give us profitability of the users. The last one is ratio between total listenings and distinct artist values for each user. The result of these transformations was used as shareholder variables.

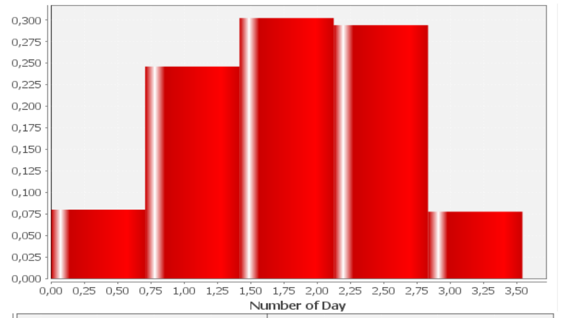
- Number of Day
- Average listening
- Ratio of Distinct Artist and Listenings (#Total Listenings/#Distinct Artist)

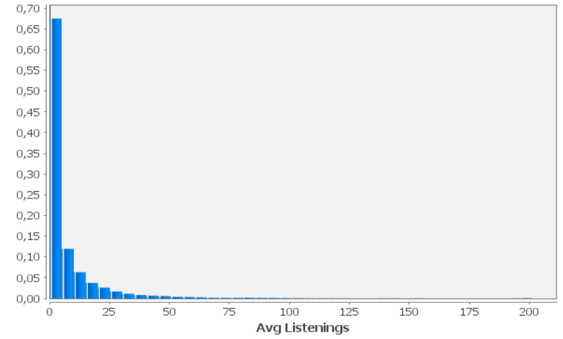
Discretization of Data

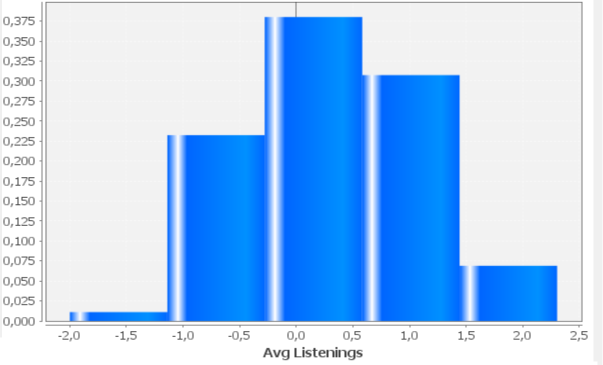
In this analysis, unsupervised discretization technic was used. As break points, quantiles were used (0.0, 0.25, 0.50, 0.75, 1.0) and also for very skewed distributions (Average listenings and Number of day) log transformation was applied.

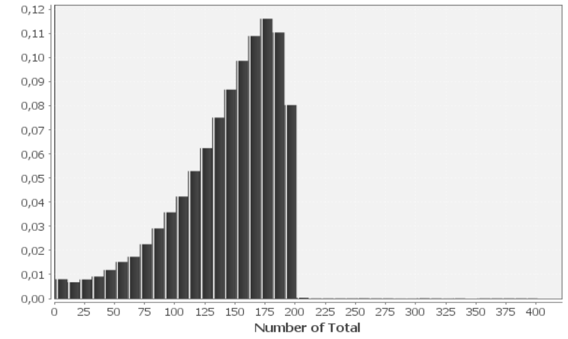
Before/after the discretization, some data distributions are:

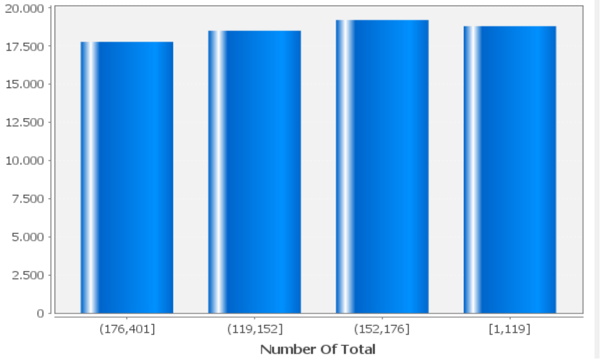


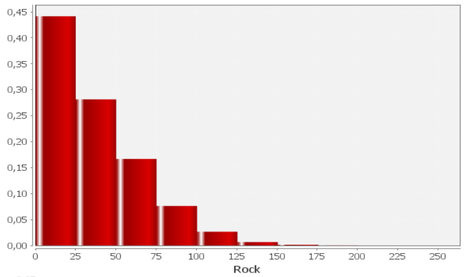


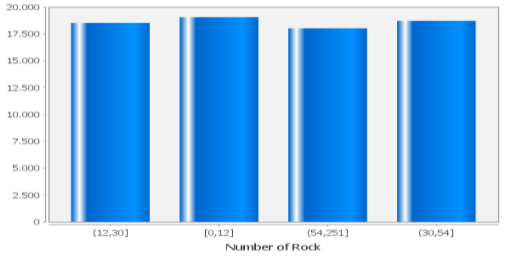


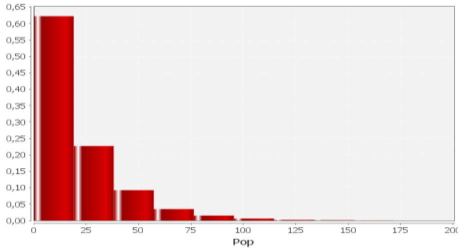


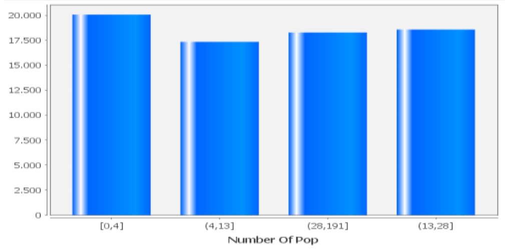


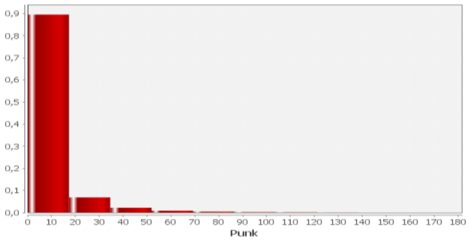


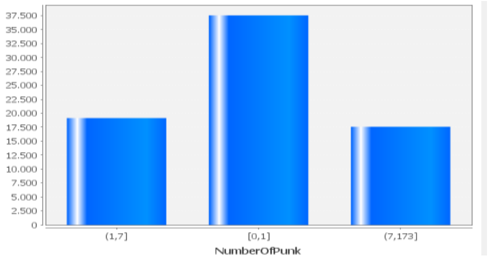


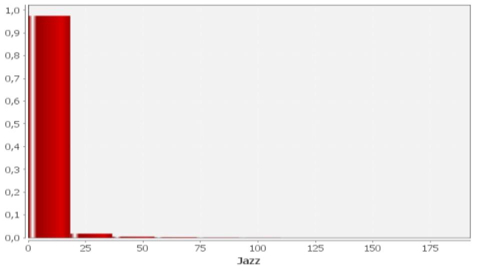


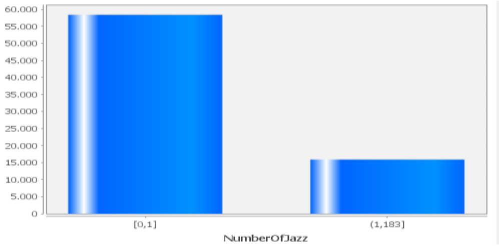






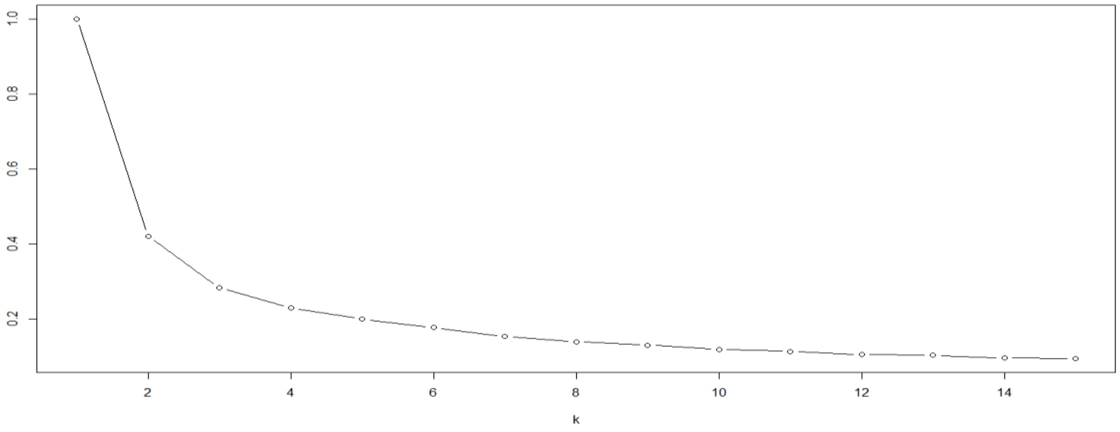




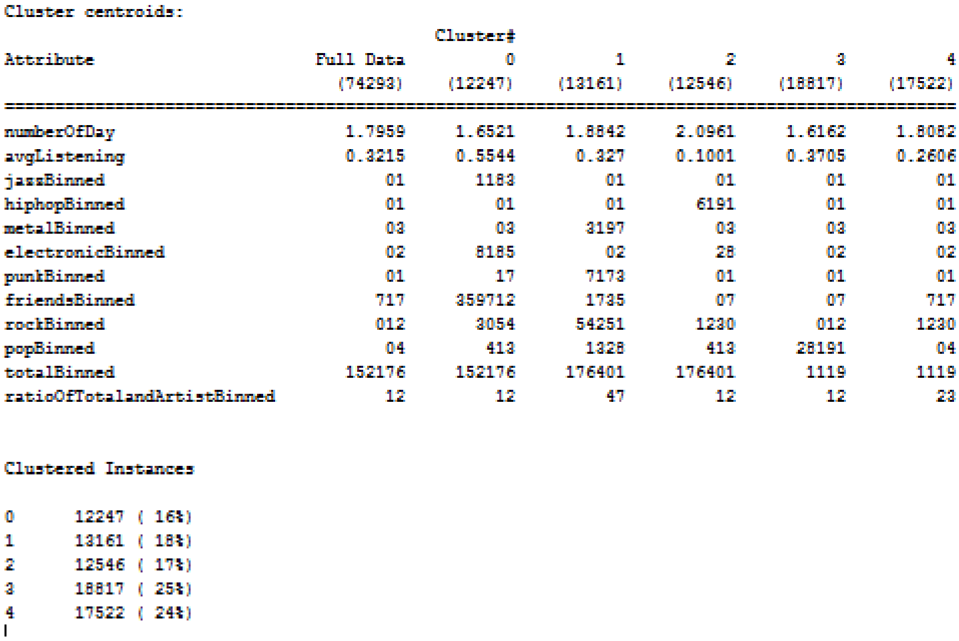


4.2 CLUSTERING

Before applying K-Means, user-defined parameter ‘k’ has to be chosen. To determine best value of k, some k values were tried and within sum of squared errors of these values were plotted. It is shown below.



As a result, k=5 was chosen and clustering algorithm was applied.



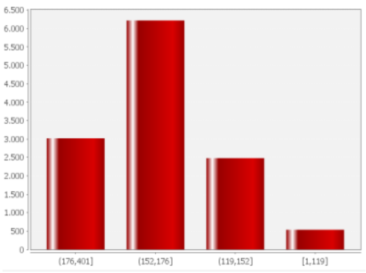
```
Cluster centroids:
                                                Cluster#
Attribute                          Full Data          0          1          2          3          4
                                     (74293)    (12247)    (13161)    (12546)    (18817)    (17522)
====================================================================================================
numberOfDay                           1.7959     1.6521     1.8842     2.0961     1.6162     1.8082
avgListening                          0.3215     0.5544      0.327     0.1001     0.3705     0.2606
jazzBinned                                01       1183         01         01         01         01
hiphopBinned                              01         01         01       6191         01         01
metalBinned                               03         03       3197         03         03         03
electronicBinned                          02       8185         02         28         02         02
punkBinned                                01         17       7173         01         01         01
friendsBinned                            717     359712       1735         07         07        717
rockBinned                               012       3054      54251       1230        012       1230
popBinned                                 04        413       1328        413      28191         04
totalBinned                           152176     152176     176401     176401       1119       1119
ratioOfTotalandArtistBinned               12         12         47         12         12         23


Clustered Instances

0       12247 ( 16%)
1       13161 ( 18%)
2       12546 ( 17%)
3       18817 ( 25%)
4       17522 ( 24%)
```

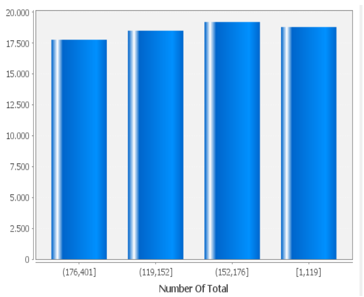
Visualization and Characterization of Clusters

Distributions of variables were shown with bar chart and to understand interesting clusters, some comparisons were done between distributions of the variable in the clusters and whole data set.

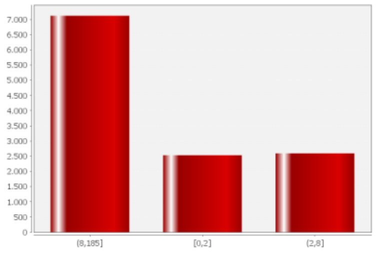
Cluster 0:

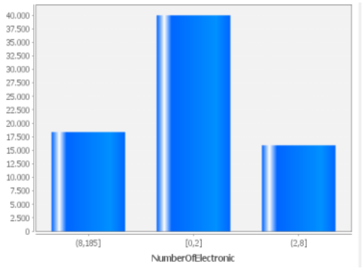
- 16% of all data set
- When the total listenings of cluster 1 is compared with total of all data set, it seems that the more than half of users with the range [152-176) of total listenings are in Cluster 0.



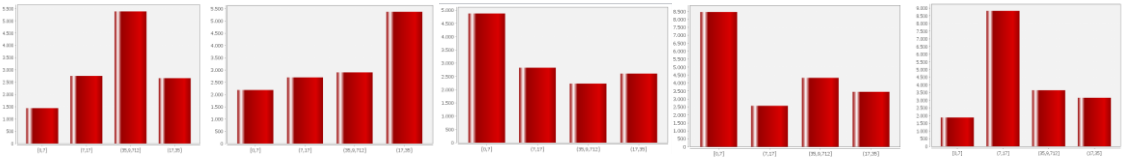


- Among the other clusters, with the range [8-185) of electronic listenings, cluster0 have most people who listen to electronic.

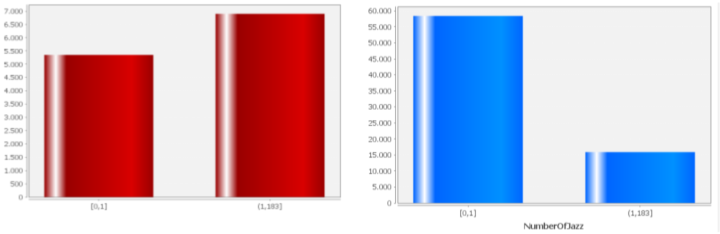




- If it is compared with other clusters, it can be seen that most friendship people are in Cluster 0.

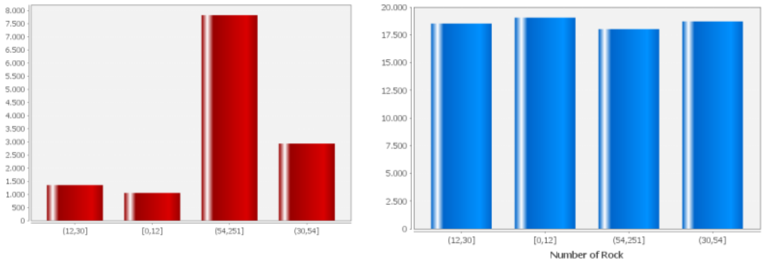


- For the users who listen to jazz, more than half of these people are in the Cluster0.

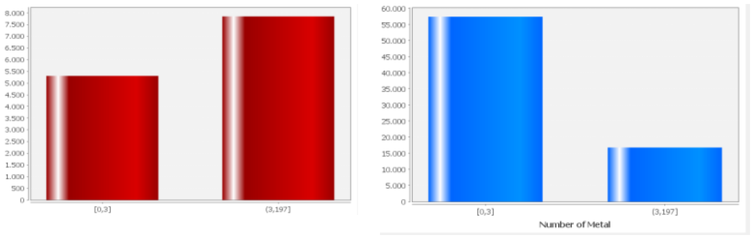


Cluster 1:

- 18% of data
- When the number of rock listenings in Cluster1 is compared to other cluster and all data set, the more than half of users tend to be member of cluster1. The range [54-251) of rock can be assign rocker users.

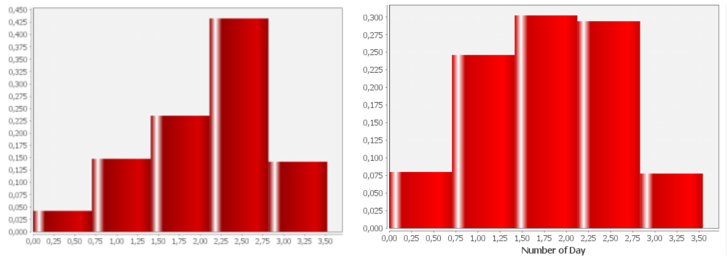


- Also the rocker user tend to listen to metal, because the mostly of metal listenings are occurred in Cluster1.



Cluster 2:

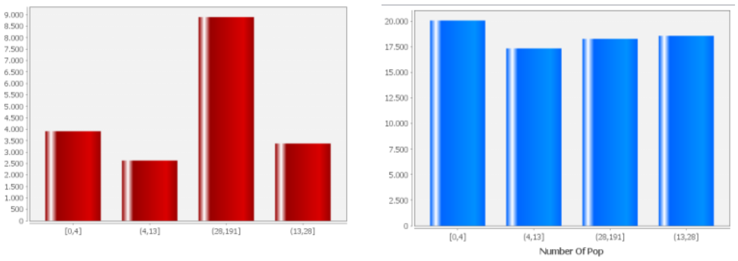
- 17% of data
- The oldest users are in Cluster 2. The variable "number of day" is assign to the number of passed day by user. (Cluster2/All data set)



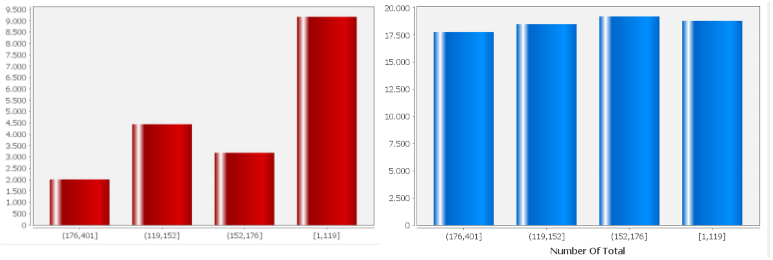
- Hip-hop listenings of cluster2 have the biggest range value. When it is compared to other genre listenings, portion of biggest range of hip-hop listenings win.

Cluster 3:

- 25% of data
- Except of Pop, other genres have lowest range value. The more than half of Pop listeners are in Cluster 3.



- Total listening range [1-119) is the lowest one when it is compared to other clusters.



Cluster 4:

- 24% of data
- It cannot be said that one genre is dominate other one for cluster 4. The range of all genres has lower range values.

Profiling the Clusters:

Clusters	Total Listening	Users	Listening index	Leverage	Tenure
4	21%	24%	0,89	0,87	187,26
1	20%	18%	1,16	1,11	214,678
3	20%	25%	0,81	0,8	166,838
0	19%	16%	1,14	1,19	158,979
2	19%	17%	1,12	1,12	303,09

- According to above table, the best users who are in the LastFM for a long time are in Cluster 2 and 1 because of the tenures. They have high-value and low-risk so retention strategies can be applied on these users. Cluster 1 is the most profitable users since tenure is high and also avgListening variable of this cluster is higher than cluster 2.
- Cluster 0 has index close to cluster 1 so cross-selling can be applied.
- Clusters 4 and 3 have the lowest index. Up-selling strategies can be used to convert the users to 0.