**Raport analiza**

**Data-Science**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ver** | **Data** | **Autor** | **Descriere** |
| **0.8** | **6.12.2016** | **Octavian Bulie** | **Incarcarea, pregatirea surselor de date SQL** |
| **1.0** | **7.12.2016** | **Andrei Damian** | **Preprocesarea si modelarea datelor, raport final** |

Contents

[1 Tabelul hiperparametrilor de experiment 2](#_Toc468882931)

[2 Metodele de selectare a datelor 3](#_Toc468882932)

[3 Codul sursa de analiza a datelor 5](#_Toc468882933)

[4 Esantion date procesate 8](#_Toc468882934)

[5 Esantion date standardizate si analizate 8](#_Toc468882935)

[6 Graficele rezultate 9](#_Toc468882936)

[7 Probleme ce urmeaza a fi analizate 11](#_Toc468882937)

# Tabelul hiperparametrilor de experiment

|  |  |  |
| --- | --- | --- |
| **Indicator** | **Valoare** | **Observatii** |
| Total inregistrari analizate | 636430 | - |
| Inregistrari duplicate | 61803 | - |
| Total inregistrari dataset de modelare | 574727 | - |
| Ani de referinta | 1 | 2016 |
| Luni de referinta | 9 | Ianuarie-Septembrie |
| Locatii de referinta | 380 |  |
| Numar de clienti analizati | 18360 | s-au determinat 70 de clienti cu valori totale 0 din totalul de 18430 de clienti |
| Metode de preprocesare | 2 | Eliminarea predictorilor cu variatie aproape 0 si standardinzarea datelor |
| Parametrii noi creati | 5 | R\_LastM – luna ultimei cumparari  R\_Score – RECENT score functie de valoarea de referinta (inceput 2016)  F\_NrTr – numarul total de tranzactii  M\_TotaVal – Valoarea totala a tranzactiilor  MaxVal – Valoarea celei mai mari tranzactii |
| Metoda de analiza | 4 centroid KMeans | s-au folosit variabilele RFM pentru determinarea celor 4 clustere corespunzatoare a 4 nivele diferite de clienti  BLACK – Clienti TOP  RED – Clienti AVG-HIGH  BLUE – Clientii AVG-LOW  GREEN – Clientii LOW |

# Metodele de selectare a datelor

|  |
| --- |
| --  -- clean data view  --  CREATE VIEW [dbo].[vw\_clean]  AS  SELECT DISTINCT  dbo.[tran].TransactionElemID AS TrElem, dbo.[tran].TransactionID AS Trans, dbo.[tran].ProductID AS Prod, dbo.[tran].CustomerID AS Cust, dbo.[tran].Date AS Dt, dbo.[tran].OrderTotal AS ProdTotal,  dbo.[tran].LocationID AS Loc, dbo.cust.Age, dbo.cust.Sex, dbo.cust.State AS St, SUBSTRING(dbo.[tran].Date, 1, 4) AS Y, SUBSTRING(dbo.[tran].Date, 5, 2) AS M, SUBSTRING(dbo.[tran].Date, 7, 2) AS D  FROM dbo.[tran] INNER JOIN  dbo.cust ON dbo.[tran].CustomerID = dbo.cust.CustomerID  GO  -- END clean data view  --  -- RFM data view  --  CREATE VIEW [dbo].[vw\_rfm]  AS  SELECT Cust, Age, Sex, St, MAX(M) AS R\_LastM, COUNT(Trans) AS F\_NrTr, SUM(ProdTotal) AS M\_TotSum, MAX(ProdTotal) AS MaxProdVal  FROM dbo.vw\_clean  GROUP BY Cust, Age, Sex, St  GO  -- END RFM data view  --  -- month\_cust - monthly amount per customer  --  SELECT Cust, Age, Sex, St, Y, M, sum(ProdTotal) MonthSum  FROM vw\_clean  group by Cust, Age, Sex, St, Y, M  order by Cust  --  -- month\_cust\_loc - monthly amount per customer & location  --  SELECT Cust, Age, Sex, St, Y, M, Loc, sum(ProdTotal) MonthSum  FROM vw\_clean  group by Cust, Age, Sex, St, Y, M, Loc  order by Cust  --  -- month\_cust\_prod - monthly amount per customer & product  --  SELECT Cust, Age, Sex, St, Y, M, Prod, sum(ProdTotal) ProdSumMonth  FROM vw\_clean  group by Cust, Age, Sex, St, Y, M, Prod  order by Cust  --  -- month\_cust\_prod\_loc - monthly amount per customer & product, location  --  SELECT Cust, Age, Sex, St, Y, M, Prod, Loc, sum(ProdTotal) ProdSumMonth  FROM vw\_clean  group by Cust, Age, Sex, St, Y, M, Prod, Loc  order by Cust  ---  --- RFM\_data - recency, frequency and monerary query  ---  SELECT Cust, Age, Sex, St, max(M) as R\_LastM,count(Trans) as F\_NrTr, sum(ProdTotal) M\_TotSum, Max(ProdTotal) as MaxProdVal  FROM vw\_clean  group by Cust, Age, Sex, St  order by M\_TotSum  SELECT \* from vw\_rfm where M\_TotSum<>0 order by M\_TotSum DESC |

# Codul sursa de analiza a datelor

|  |
| --- |
| # -\*- coding: utf-8 -\*-  """  Created on Wed Dec 7 12:01:34 2016  @author: slash\_000  """  import pandas as pd  import numpy as np  from sklearn.cluster import KMeans  from sklearn import preprocessing  from bokeh.io import output\_file, show  from bokeh.plotting import figure  from bokeh.plotting import ColumnDataSource  from bokeh.models.layouts import Column  import os  if \_\_name\_\_ == '\_\_main\_\_':    df\_rfm = pd.read\_csv('rfm\_data.csv')  df\_rfm.sort\_values('M\_TotVal')    cf1 = 'R\_Score'  cf2 = 'F\_NrTr'  cf3 = 'M\_TotVal'  cfc = 'Categ'  nr\_cl = 4    cluster\_fields = [cf1,cf2,cf3]    df\_model = df\_rfm.loc[:,cluster\_fields]    np\_arr = preprocessing.scale(df\_model)  df\_model = pd.DataFrame(data = np\_arr, columns = cluster\_fields )    clf = KMeans(n\_clusters = nr\_cl, n\_jobs=-1)  clf.fit(df\_model)  labels = np.unique(clf.labels\_)    df\_model[cfc]= clf.labels\_  c = df\_model[cfc]    colors = ['red','green','blue','black','yellow']  for i in range(labels.size):  c.replace(labels[i],colors[i],inplace=True)    df\_model.to\_csv('rfm\_data\_clusters.csv')    html\_output = os.path.basename(\_\_file\_\_+'.html')  output\_file(html\_output)    p1 = figure(x\_axis\_label = cf1, y\_axis\_label = cf2, title = 'RFM Plot 1 with {} clusters'.format(nr\_cl))  cds1 = ColumnDataSource(df\_model.loc[:,[cf1,cf2,cfc]])  p1.circle(cf1,cf2,source = cds1, alpha = 0.5, size=2, color = cfc)  p2 = figure(x\_axis\_label = cf1, y\_axis\_label = cf3, title = 'RFM Plot 2 with {} clusters'.format(nr\_cl))  cds2 = ColumnDataSource(df\_model.loc[:,[cf1,cf3,cfc]])  p2.circle(cf1,cf3,source = cds2, alpha = 0.5, size=2, color = cfc)  p3 = figure(x\_axis\_label = cf2, y\_axis\_label = cf3, title = 'RFM Plot 3 with {} clusters'.format(nr\_cl))  cds3 = ColumnDataSource(df\_model.loc[:,[cf2,cf3,cfc]])  p3.circle(cf2,cf3,source = cds3, alpha = 0.5, size=2, color = cfc)    p = Column(p1,p2,p3)  show(p) |

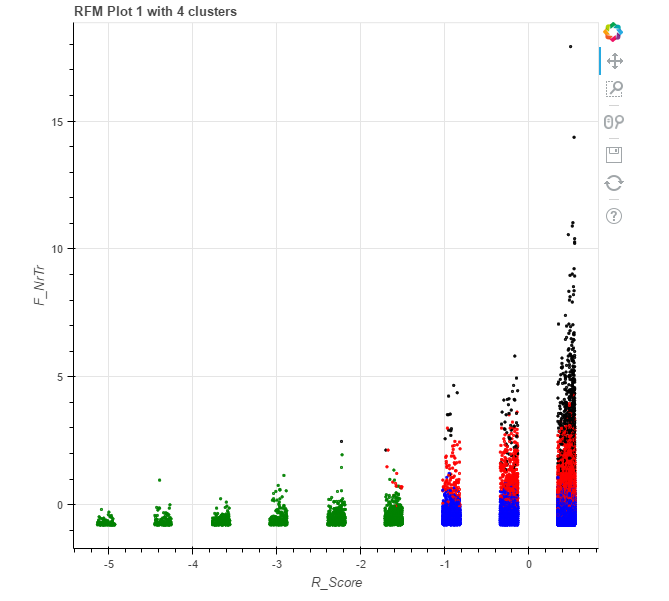
# Esantion date procesate

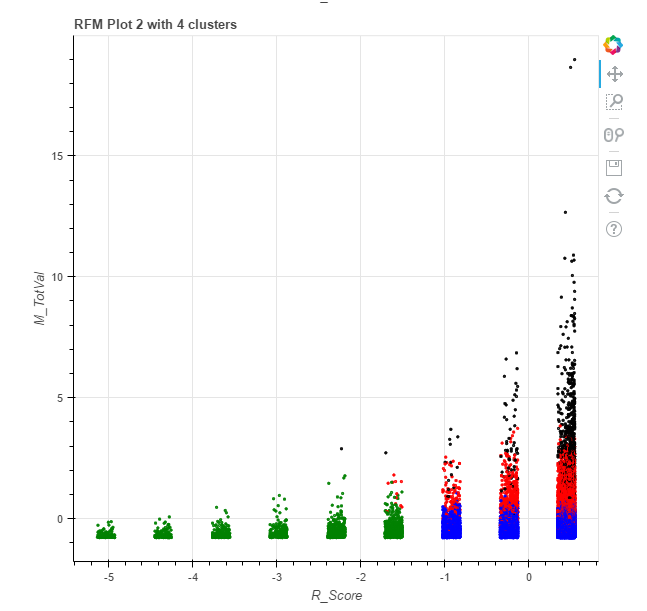
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cust | Age | Sex | St | R\_LastM | R\_Score | F\_NrTr | M\_TotVal | MaxVal |
| 4604472 | 49 | 1 | 1048648 | 9 | 930 | 430 | 8680167 | 146963 |
| 3388765 | 44 | 1 | 1053157 | 9 | 923 | 718 | 8539035 | 239992 |
| 5213133 | 28 | 1 | 1050957 | 9 | 914 | 143 | 5908915 | 1270110 |
| 4829260 | 38 | 1 | 1048648 | 9 | 928 | 266 | 5131302 | 94945 |
| 7644971 | 27 | 0 | 1048648 | 9 | 913 | 165 | 5073777 | 161658 |
| 8250238 | 80 | 0 | 1048649 | 9 | 929 | 220 | 5039696 | 143110 |
| 4476523 | 62 | 1 | 1048648 | 9 | 925 | 179 | 5020038 | 171150 |
| 4244644 | 43 | 0 | 1052914 | 9 | 926 | 247 | 4761620 | 181174 |
| 4669842 | 41 | 0 | 3157291 | 9 | 929 | 582 | 4637216 | 38798 |
| 7536272 | 24 | 1 | 3157328 | 9 | 930 | 374 | 4473691 | 84661 |
| 6202779 | 81 | 0 | 1048648 | 9 | 907 | 237 | 4368256 | 119789 |
| 4253119 | 60 | 0 | 1048648 | 9 | 930 | 181 | 4327847 | 147009 |
| 5304115 | 60 | 0 | 1048649 | 9 | 926 | 377 | 4172451 | 229167 |
| 3214779 | 84 | 1 | 1048648 | 9 | 930 | 423 | 4071119 | 73325 |

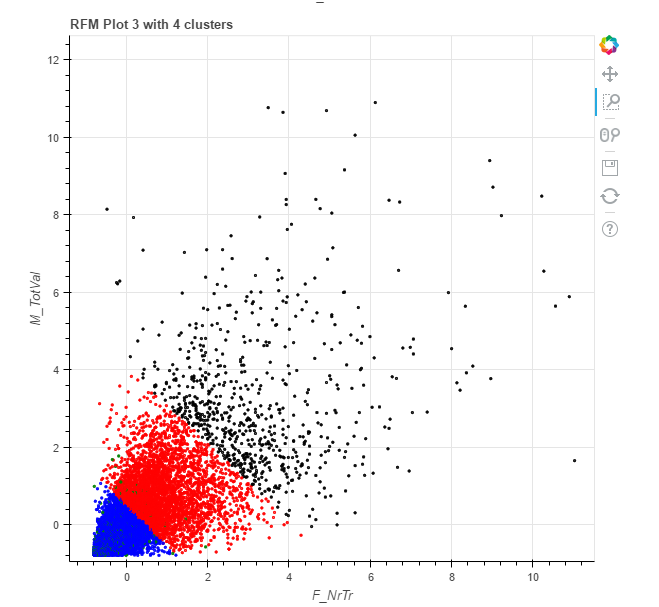
# Esantion date standardizate si analizate

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R\_Score | F\_NrTr | M\_TotVal | Categ |
| 0 | 0.545144962 | 10.39981059 | 18.98410414 | black |
| 1 | 0.497222949 | 17.9115083 | 18.6624822 | black |
| 2 | 0.435608933 | 2.914195154 | 12.66877194 | black |
| 3 | 0.531452959 | 6.122316053 | 10.89669043 | black |
| 4 | 0.428762931 | 3.488005397 | 10.76559825 | black |
| 5 | 0.53829896 | 4.922531002 | 10.68793196 | black |
| 6 | 0.510914953 | 3.853157369 | 10.64313387 | black |
| 7 | 0.517760955 | 5.626752662 | 10.05423199 | black |
| 8 | 0.53829896 | 14.36431771 | 9.770731037 | black |
| 9 | 0.545144962 | 8.939202696 | 9.39807828 | black |
| 10 | 0.38768692 | 5.365929825 | 9.157805273 | black |

# Graficele rezultate







# Probleme ce urmeaza a fi analizate

1. Pregatirea unui mediu de productie si analiza-clusterizare in timp real
2. Determinarea unui model sau a mai multor modele de predictie care sa includa si alti predictori in afara de RFM
3. Determinarea unui model de analiza si predictie la nivel de individ