Semantix

Documentação Final

Projeto Data Science

BIG DATA & ANALYTICS

Documentação Final

Projeto Data Science

BIG DATA & ANALYTICS

Semantix S.A.

www.semantix.com.br

Versões do documento

Data	Versão	Autor	Descrição
17/05/2018	1.0	Renata Lima da Silva	Criação do documento

Informações de contato com a Semantix

Nome: Stefany Villas Boas de Araújo
Telefone: (11) 972.324.119
E-mail: stefany.araujo@semantix.com.br

Nome: Eduardo Miranda
Telefone: (11) 946.118.635
E-mail:eduardo.miranda@semantix.com.br

Nome: Renata Lima da Silva
Telefone: (11) 991.705.115
E-mail:renata.lima@semantix.com.br

Nome: Aruã de Mello Sousa
Telefone: (12) 981.039.482
E-mail:arua.sousa@semantix.com.br

Nome: Rodolfo Uchida
Telefone: (11) 985.523.425
E-mail:arua.sousa@semantix.com.br

Índice

1. Introdução

Este documento faz parte do conjunto de documentações referentes ao Projeto **Data Science**, elaborado e desenvolvido pela **Semantix** para o cliente **Elo**. Nele consta o fluxo de trabalho, premissas, metodologias e técnicas abordadas.

O escopo do projeto envolvia um estudo pontual sobre os clientes da **Elo.** Este estudo buscou responder algumas perguntas sobre o padrão de consumo e ciclo de vida dos clientes.

O projeto foi dividido em quatro diferentes fases: *Feature Engineering*, Clusterização, Previsão de *Churn* e Previsão de Faturamento.

2. Feature Engineering

Durante essa fase foram feitos os entendimentos das bases de dados, levantamento dos dados mais importantes e a criação das variáveis que foram usadas nos modelos.

Nesse momento desenvolvemos o script: **projeto_ds_2017_primeira_ramo_bin.py** que cria a tabela de mesmo nome. Esta contém as transações de todos os clientes, criação das variáveis importantes e os relacionamentos com outras tabelas que armazenam dados importantes como: data da primeira transação, ramos e subsetores, motivos de negadas.

Também foi gerada a tabela **projeto_ds_primeira_ramo_bin.py** que além dos dados de 2017 contém o histórico de 2016 e o conceito de **lag**.

O conceito de *lag* foi criado como parâmetro para comparação entre clientes. Que seria a diferença, em semanas, entre a data da transação e a data da primeira transação do cartão.

3. Clusterização

3.1. Justificativa e Metodologia

Para uma análise inicial do comportamento dos clientes usamos o método *K-means* para clusterizar empiricamente os clientes da **Elo**.

As premissas foram:

- Clientes que ativaram o cartão entre 01/01/2016 e 28/02/2017.

- Divididos em dois grandes grupos: Crédito e Débito.
- Analisamos uma janela de 27 semanas a partir do nascimento do cliente.

A partir destas segmentações extraímos três parâmetros principais:

- O valor total transacionado no período;
- A quantidade de transações no período;
- O número de semanas onde houveram transações.

Para uma melhor performance do algoritmo foi realizada uma transformação nas variáveis: valor transacionado e quantidade de transações. Foram testadas duas transformações, primeiramente uma transformação logarítmica e depois, com melhor resultado, a normalização dos dados.

Foram testadas combinações de K={2, 3, 4, 5, 6, 7, 8}, usando a soma dos erros quadrados como métrica.

Definimos um limiar para a escolha do número de centróides ideal (K=4), tanto para Crédito como para Débito.

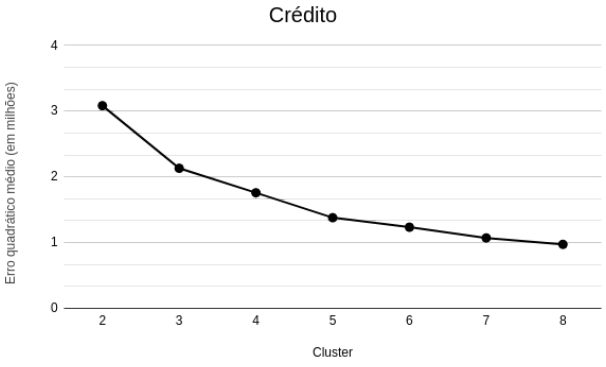


Imagem 1 - Gráfico de erros para clusterização Crédito

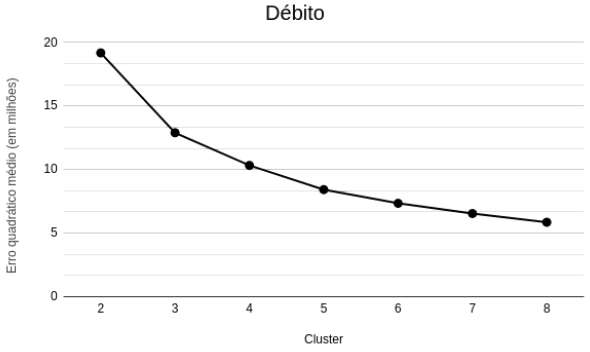


Imagem 2 - Gráfico de erros para clusterização Débito

Como resultado tivemos os seguintes centróides:

Label	Número de Transações	Valor de transações	Número de semanas
A	86,29	$ 7.593,80	19,45
B	49,25	$ 3.169,19	16,68
C	21,05	$ 2.163,11	9,38
D	5,88	$ 628,36	2,87

Tabela 1 - Centróide de Crédito

Label	Número de Transações	Valor de transações	Número de semanas
A	136,15	$ 4.742,69	22,87
B	60,42	$ 3.446,74	17,34
C	26,62	$ 1.406,34	10,99
D	6,40	$ 376,52	3,27

Tabela 2 - Centróide de Débito

A seguinte distribuição entre os clusters:

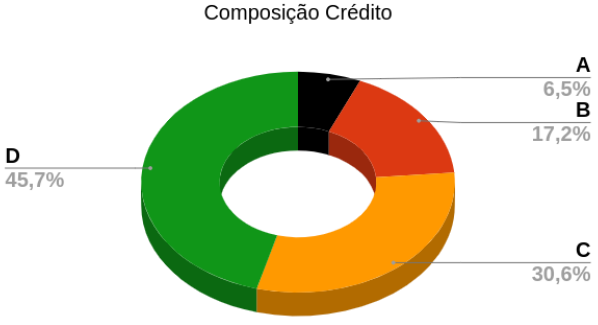


Imagem 3 - Composição Crédito

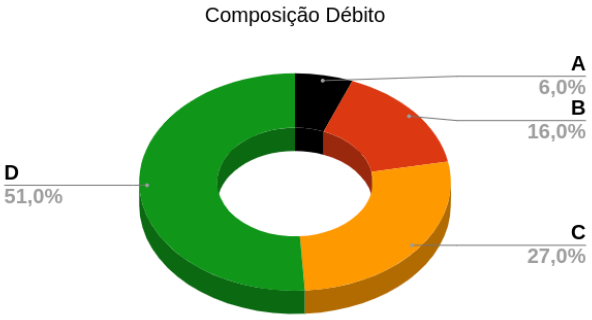


Imagem 4 - Composição Débito

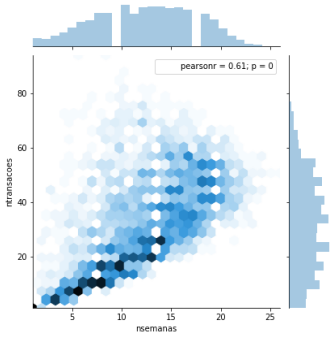
3.2. Visualização dos *clusters*

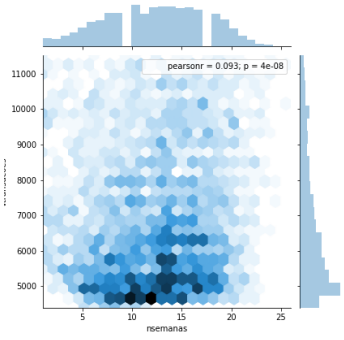
Abaixo estão as visualizações das amostras dos clusters, combinados entre número de transações, valor de transações e número de semanas. Estes gráficos foram de grande importância para interpretar as características de cada cluster. Abaixo estão os clusters A, B, C, e D respectivamente, para crédito e débito.

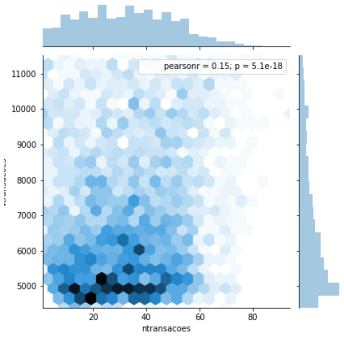
Estas visualizações ajudam a **Elo** entender melhor o seu cliente e seu perfil de consumo. Com essa clusterização a **Elo** é capaz de focalizar melhor suas ações.

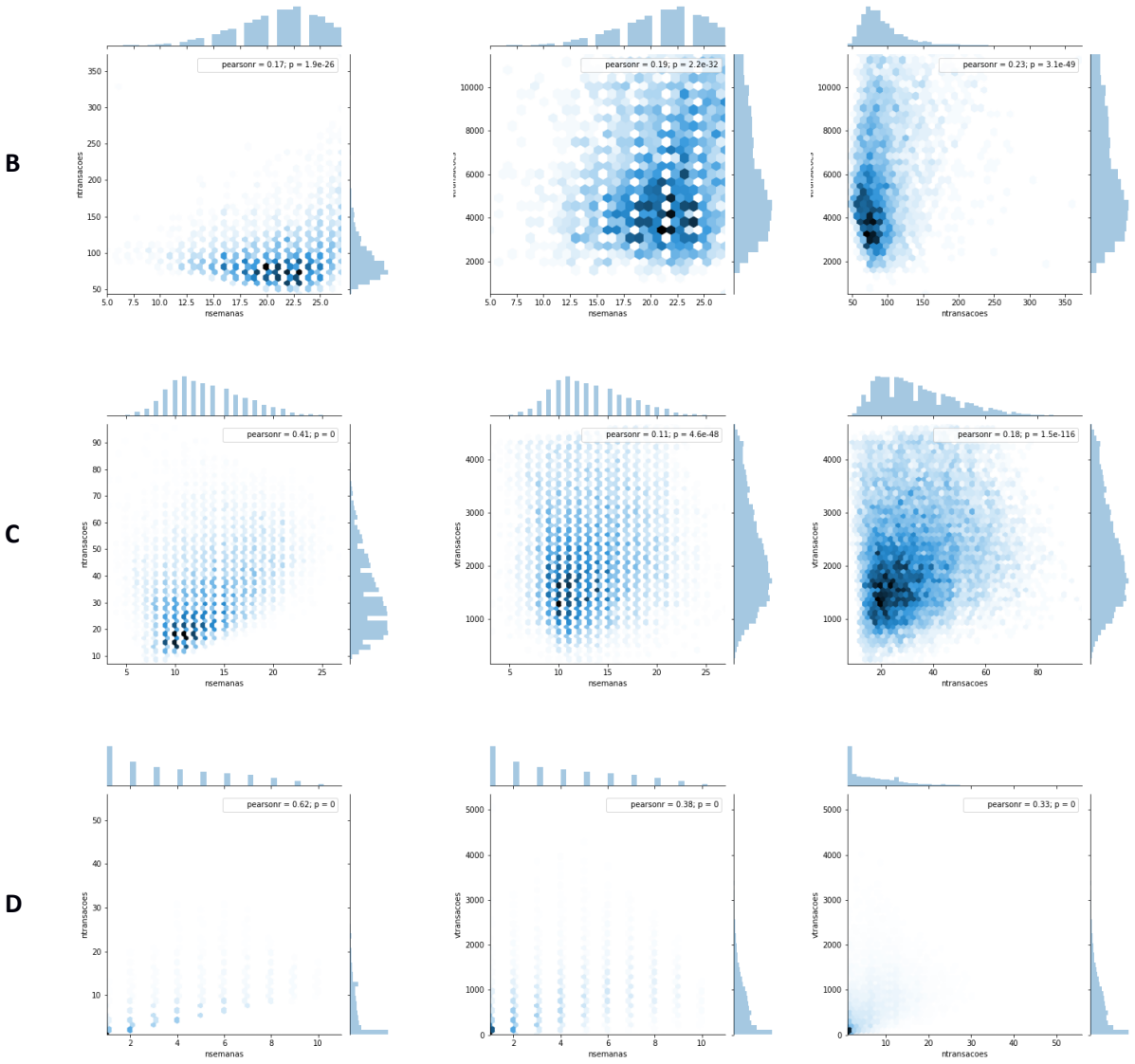
#Semanas X #Transações | **#Semanas X $Transações** | **#Transações X $Transações**

A

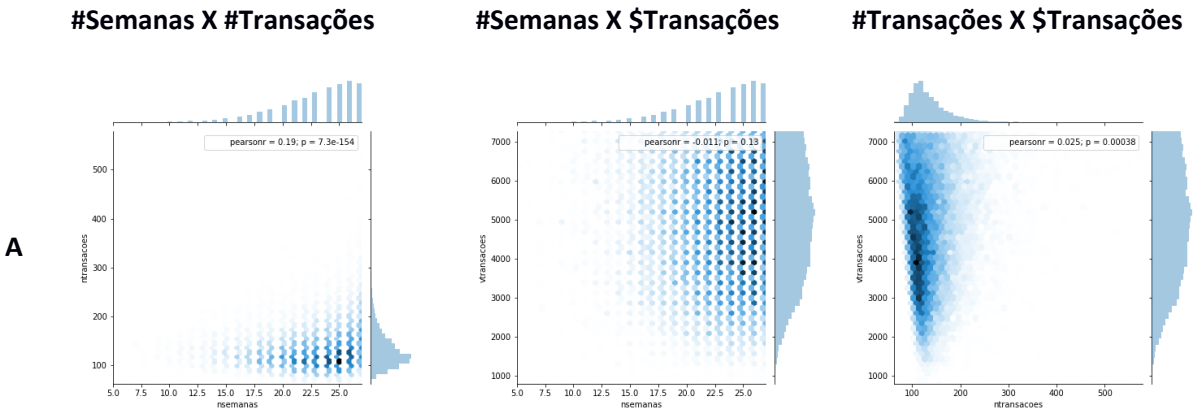


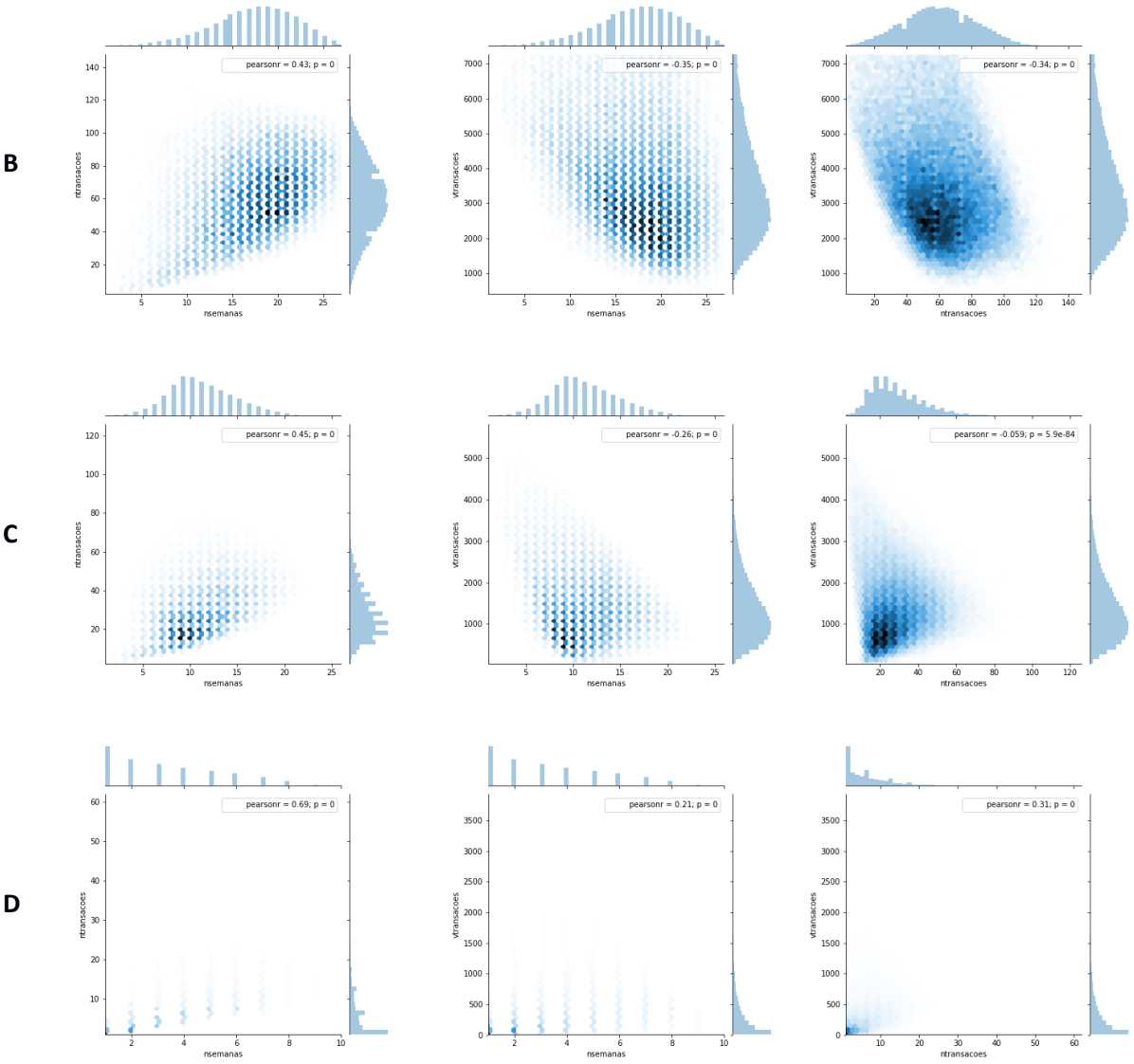






Visualização 1 - *Clusters* Crédito





Visualização 2 - *Clusters* Débito

3.3. Comparações entre *clusters*

Após a clusterização dos clientes foram feitas análises comparativas a fim de estabelecer as principais diferenças entres os tipos de *clusters*.

Comparamos o perfil de consumo em: Ramos de Atividade, Subsetor, Rede de Estabelecimento comercial.

Para a escolha usamos diferentes abordagens:

- Ramos com as maiores diferenças entre os *clusters*.

- O ponto negativo dessa análise é que ela pode apresentar **Ramos** que não são relevantes para o faturamento **Elo**.

- Top 50 Ramos mais relevantes para a **Elo**.
    - A escolha desses 50 **Ramos** foi feita a partir do faturamento total no período analisado.

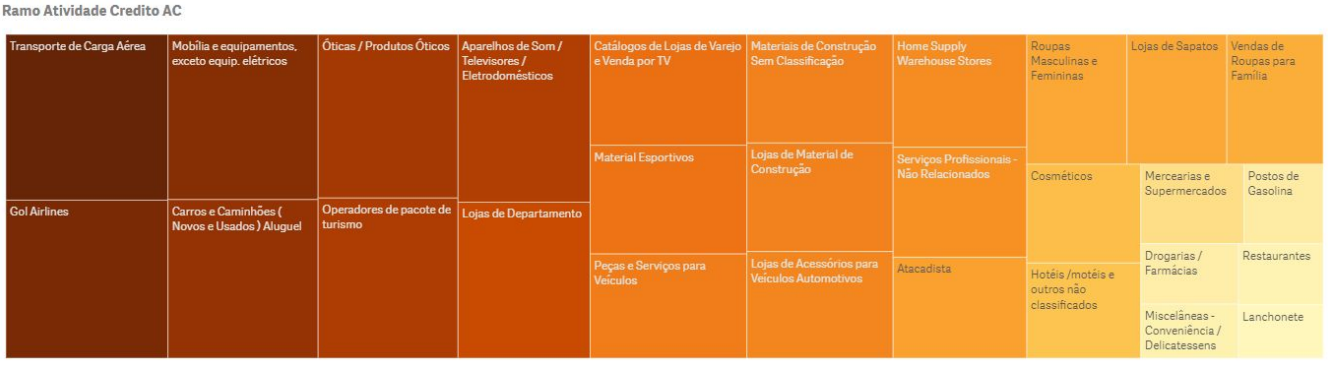
- Como **Subsetor** possui menos de 50 categorias todas foram usadas para o comparativo.

- Para o campo Rede de Estabelecimento Comercial usamos o campo **Marca**.

Para executar essas comparações foi desenvolvido o *script* **comparacao_cluster_total.py**. É possível fazer a comparação apenas entres os *clusters* ou também usando a segmentação por tipo de cartão (Básico, Mais, Grafite, Nanquim). Sempre lembrando que a análise está segmentada entre Crédito e Débito.

Essas comparações ajudam a **Elo** entender padrões de consumo e como influenciar um cliente de um *cluster* transitar para outro.

Segue um exemplo das comparações apresentadas.



Comparação 1 - Ramo Atividade Crédito *Cluster* A e C

4. Previsão de *Churn*

4.1. Premissas e Variáveis

A intenção desse modelo é obter uma medida para a previsão do comportamento do cliente no futuro e conseguir discernir dessa previsão os clientes que são churn.

Para o estudo não houve uma segmentação da base, uma janela de tempo de observação de três meses e analisados todos os clientes presentes .

Com o propósito de descrever o comportamento da distribuição do faturamento semanal desse cliente na janela analisada.

Com os dados transacionais de cartões de crédito e débito modelamos a probabilidade do cliente ser *Churn* no próximo mês. A partir disso criamos as variáveis com as seguintes características:

- Emissor do cartão.
- Família Cartão.
- Tipo Cartão.
- *Kurtosis* da distribuição do faturamento semanal.
- *Skewness* da distribuição do faturamento semanal.
- Soma das transações no período.

As variáveis de maior importância foram:

- 1ª *Skewness*.
- 2ª Soma das transações.
- 3ª *Kurtosis*.

A saída deste algoritmo foi usada como variável para o modelo de Previsão de Faturamento.

4.2. Análise de performance do modelo

Os algoritmos foram validados até Agosto de 2017. O conjunto treino e teste foram divididos em 70% e 30% do total dos dados. Foram aplicados os algoritmos *Random Forest* e *Gradient Boosting Tree*, o algoritmo final com melhor performance foi o *Gradient Boosting Tree*.

- **Crédito**

	Ativos	**Churn**
Ativos	**206510**	16320
Churn	35083	**25040**

  - **Precision**: 0.80
  - **Recall**: 0.82
  - **Accuracy**: 0.82

- **Débito**

	Ativos	**Churn**

Ativos	**1725819**	53339
Churn	596678	**65350**

- **Precision**: 0.72
- **Recall**: 0.76
- **Accuracy**: 0.76

5. Previsão de Faturamento

5.1. Premissas e Variáveis

A partir de dados transacionais de cartões de crédito e débito modelamos o faturamento previsto para os próximos dois meses de cada cliente. O período utilizado para criar as variáveis transacionais do modelo foi de três meses, a partir disso criamos as variáveis com as seguintes características:

- Valor total transacionado
- Quantidade total transacionada
- Gastos em cada subsetor
- Top 100 estabelecimentos por faturamento: gasto em cada estabelecimento
- Tipo, emissor e família do cartão
- Estado (UF)
- Transações Negadas
- Probabilidade de Churn: resultado do modelo de churn
- Grupo no qual o cliente pertence no bimestre atual

Ao total foram criadas 198 variáveis

O faturamento bimestral alvo foi dividido em três faixas de valores, tais faixas foram determinadas de acordo com os quantis da distribuição. Os gráficos de faturamento são relativos ao período de Jullho-Agosto de 2017:

- **Crédito:**

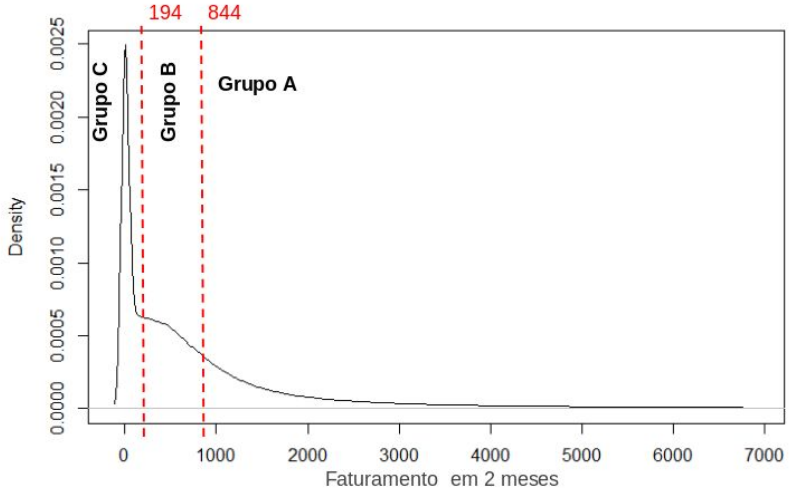


Imagem 5 - Faturamento Crédito

- **Débito:**

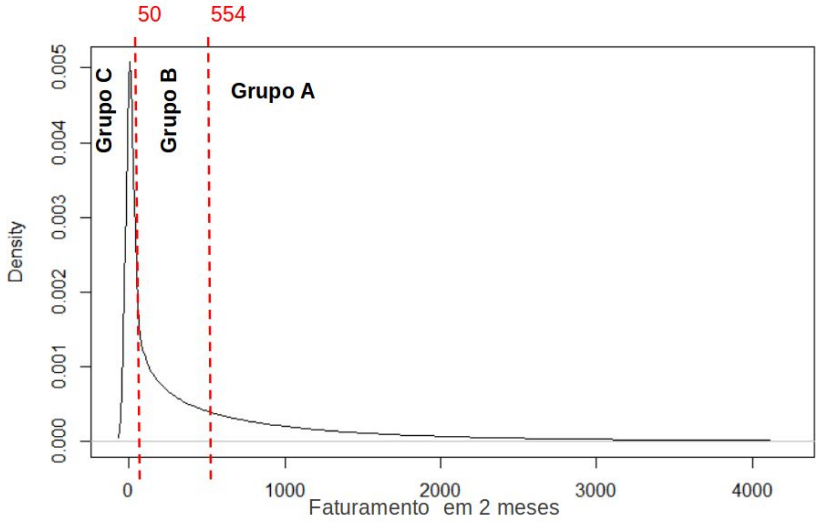


Imagem 6 - Faturamento Débito

5.2. Análise de Performance dos Modelos

Os algoritmos foram validados até Agosto de 2017, a matriz de atributos total continha respectivamente 1,6 e 14,7 milhões de cartões crédito e débito. O conjunto treino e teste foram divididos em 70% e 30% do total dos dados. Foram aplicados os algoritmos *Random Forest*, *Gradient Boosting Tree* e um modelo Multinomial, o algoritmo com maior **Medida F1** balanceada por classe foi o *Random Forest*.

A seguir apresentamos as medidas de performance do Random Forest para o crédito e para o débito.

- **Crédito**

➔ Matriz de confusão

Predito / Real	C	B	A
C	**110732**	4378	15298
B	43793	**99657**	26715
A	16885	39038	**113427**

➔ Medidas de precisão de recall

Grupo	Precisão	Recall	Medida F1
C	0.65	0.65	0.65
B	0.55	0.58	0.56
A	0.73	0.67	0.70

- **Débito**

➔ Matriz de confusão

Predito / Real	C	B	A
C	**1023641**	288592	163925
B	506307	**622364**	342285
A	148051	287753	**1036711**

➔ Medidas de precisão de recall

Grupo	Precisão	Recall	Medida F1
C	0.61	0.69	0.65
B	0.52	0.42	0.47
A	0.67	0.70	0.69

5.3. Principais resultados

As variáveis mais importantes do modelo foram as variáveis transacionais auto-regressivas e de totais, tal resultado é esperado pois variáveis de gasto em determinado estabelecimento ou subsetor contém informações mais específicas e portanto de menor alcance.

Abaixo se encontra lista de variáveis mais importantes com o respectivo peso para a queda da entropia nos dados:

- Cartão de **Crédito**

Ranking	Variável	Peso
1	Valor transacionado no mês 3	0.137
2	Valor total transacionado	0.128
3	Quantil no bimestre atual	0.120
4	Contagem de transações no mês 3	0.110
5	Valor transacionado no mês 2	0.0757
6	Contagem total de transações	0.0707
7	Contagem de transações no mês 2	0.0500
8	**Probabilidade de Churn**	**0.0470**
9	Valor transacionado no mês 1	0.0446
10	**Gasto total em varejo alimentício generalista**	**0.0303**
11	**Gasto total em drogarias e farmácias**	**0.0213**
12	**Gasto total em alimentação**	**0.0212**
13	Contagem de transações no mês 1	0.0202
14	**Gasto total em postos de combustível**	**0.0181**
15	**Gasto total em vestuário**	**0.0167**
16	Cartão Básico	0.0154
17	**Gasto total em varejo alimentício especializado**	**0.0113**
18	**Gasto total em outros varejos**	**0.00865**
19	Cartão Grafite	0.00771
20	**Gasto total em lojas de departamentos**	**0.00641**

- Cartão de **Débito**

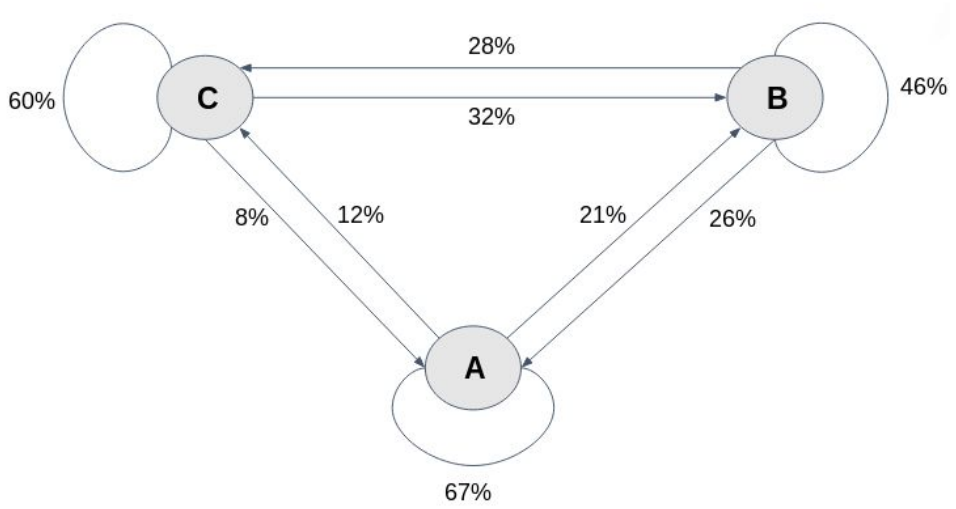
Ranking	Variável	Peso
1	Valor transacionado no mês 3	0.161
2	Valor total transacionado	0.117
3	Contagem de transações no mês 3	0.115
4	Quantil no bimestre atual	0.0929
5	**Probabilidade de churn**	**0.0873**
6	Contagem total de transações	0.0655
7	Valor transacionado no mês 2	0.0560
8	Contagem de transações no mês 2	0.0483
9	Valor transacionado no mês 1	0.0364
10	**Gasto total em varejo alimentício generalista**	**0.0350**
11	Contagem de transações no mês 1	0.0267
12	**Gasto total em alimentação**	**0.0233**
13	**Gasto total em drogarias e farmácias**	**0.0209**
14	**Gasto total em varejo alimentício especializado**	**0.0203**
15	**Gasto total em postos de combustível**	**0.0194**
16	**Gasto total em outros varejos**	**0.0121**
17	**Gasto total em vestuário**	**0.0103**
18	**Gasto total em lojas de departamento**	**0.00530**
19	**Gasto total no Pagseguro**	**0.00504**
20	Número de estados visitados	0.00467

As variáveis destacadas estão dentro do grupo onde é possível agir em forma de parcerias, desta maneira uma ação em tais subsetores produz resultados mais impactantes no faturamento bimestral do cliente em comparação com o restante dos subsetores.

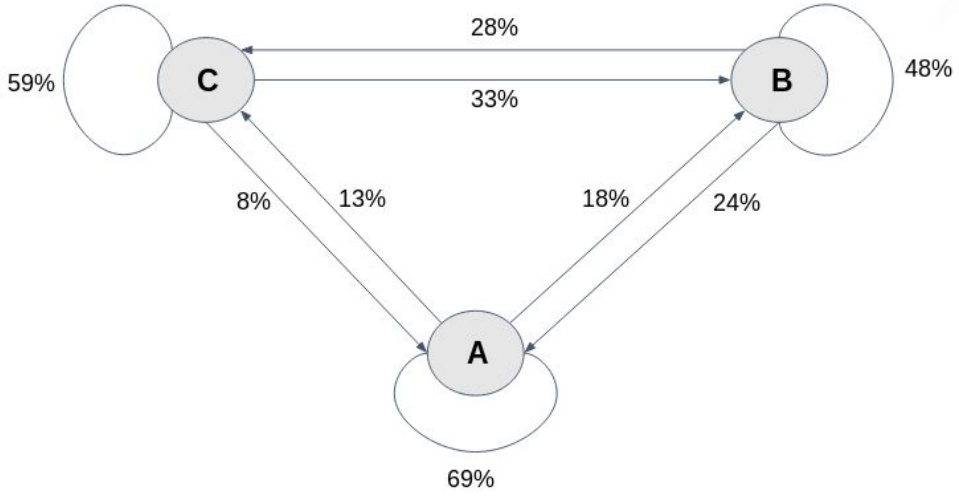
O impacto de cada subsetor no faturamento final do cliente foi mensurado através de probabilidades de transição entre grupos de faturamento dado que o cliente teve gasto em um determinado subsetor.

Foram utilizados os períodos de Junho, Julho e Agosto de 2017 para o cálculo das probabilidades de transição de estados do cliente. A dinâmica média de transição de um cliente de um grupo X em um bimestre específico para um grupo Y no bimestre consequente é dado abaixo:

Comportamento geral dos clientes no **débito**:



Comportamento geral dos clientes no **crédito**:



Notamos que os clientes tendem a permanecer no mesmo grupo no quadrimestre, a maior taxa de transição de estados está entre o grupo B e C no qual temos muitos clientes que oscilam entre ativos e inativos.

A partir do comportamento médio de transição mensuramos o quantidade de vezes que a probabilidade de transição altera dado que o cliente realizou compra em determinado subsetor ou estabelecimento. Denominamos a medida que mensura tal grandeza de lift.

Separamos as transições em categorias de acordo com a legenda abaixo:

Legenda	
	Cliente se mantém no mesmo grupo
	Cliente muda para um grupo superior
	Cliente muda para um grupo inferior

A tabela abaixo contém os dados de *lifts* de cartão de débito segregados por subsetores com maior impacto preditivo no faturamento do cliente:

Transição	Varejo Alimentício Generalista	Drogarias e Farmácias	Alimentação	Postos de Combustível	Vestuário	Lojas de departamento
A -> A	1,02	1,06	1,04	1,06	1,03	1,05
B -> B	1,01	1,04	1,01	0,98	1,03	1,04
C -> C	0,99	0,9	0,9	0,92	0,89	0,89
Média	1,006	1	0,983	0,986	0,983	0,993
B -> A	1,01	1,09	1,10	1,12	1,02	1,06
C -> B	1,02	1,15	1,13	1,07	1,17	1,18
C -> A	0,95	1,09	1,16	1,20	1,06	1,06
Média	0,99	1,11	1,13	1,13	1,08	1,10
A -> C	0,91	0,78	0,85	0,85	0,87	0,84
A -> B	0,97	0,9	0,92	0,89	0,95	0,92
B -> C	0,95	0,84	0,88	0,91	0,92	0,86
Média	0,943	0,84	0,883	0,883	0,913	0,873

Concluímos a partir da tabela acima:

- ➔ Subsetores com maior impacto na estabilização de clientes (amarelo):
  - ◆ Não foram obtidos valores significativos de lift que indiquem desvios do padrão médio do cliente
- ➔ Subsetores com maior impacto em "alavancar" clientes (verde):
  - ◆ Alimentação e Postos de Combustível
- ➔ Subsetores com maior impacto em reduzir queda de gasto no cliente (vermelho) :
  - ◆ Drogarias e Farmácias, Alimentação e Postos de Combustível

- Cartão de **Crédito**

Transição	Varejo Alimentício Generalista	Drogarias e Farmácias	Alimentação	Postos de Combustível	Vestuário	Lojas de departamento
A -> A	1,06	1,1	1,1	1,08	1,06	1,05
B -> B	1,04	1,04	1,02	1,01	1,05	1,03

C -> C	0,9	0,85	0,9	0,89	0,86	0,85
Média	1	0,996	1,006	0,993	0,99	0,976
B -> A	1,05	1,12	1,12	1,13	1,01	1
C -> B	1,17	1,23	1,13	1,15	1,22	1,24
C -> A	0,94	1,1	1,09	1,09	1,04	1,03
Média	1,0533	1,15	1,113	1,123	1,09	1,09
A -> C	0,79	0,72	0,77	0,77	0,8	0,85
A -> B	0,91	0,8	0,78	0,83	0,91	0,89
B -> C	0,87	0,82	0,85	0,86	0,89	0,93
Média	0,856	0,78	0,8	0,82	0,866	0,89

Concluímos a partir da tabela acima:

- Subsetores com maior impacto na estabilização de clientes (amarelo):
  - Não foram obtidos valores significativos de *lift* que indiquem desvios do padrão médio do cliente
- Subsetores com maior impacto em "alavancar" clientes (verde):
  - **Drogarias e Farmácias** e **Postos de Combustível**
- Subsetores com maior impacto em reduzir queda de gasto no cliente (vermelho) :
  - **Drogarias e Farmácias** e **Alimentação**

Observa-se que **Drogarias e Farmácias** obteve um impacto satisfatório nas análises acima, desta maneira comparamos os *lifts* de duas maiores redes de **Drogarias** no estado de São Paulo:

- Cartão de **Débito**:

Transição	Droga Raia	Drogaria Sâo Paulo
A -> A	1,16	1.15
B -> B	1	0,99
C -> C	1	0,99
Média	1,0533	0,99
B -> A	1,3	1,3
C -> B	1,22	1,22
C -> A	1,41	1.47
Média	1,31	1,26
A -> C	0,63	0,64
A -> B	0,69	0,7
B -> C	0,73	0,74

Média	0,6833	0,6933

- Cartão de **Crédito**:

Transição	Droga Raia	Drogaria São Paulo
A -> A	1,19	1,18
B -> B	0,95	0,97
C -> C	0,83	0,81
Média	0,83	0,81
B -> A	1,31	1,27
C -> B	1,15	1,22
C -> A	1,47	1,36
Média	1,47	1,36
A -> C	0,61	0,63
A -> B	0,54	0,59
B -> C	0,82	0,81
Média	0,82	0,81

A partir das tabelas acima conclui-se que há diferença significativa entre as duas redes no quesito "alavancar" clientes, onde a **Droga Raia** tem melhor *lift* ambos no crédito e débito.

Por fim comparamos as medidas de *lift* para alguns estabelecimentos nas duas modalidades de cartões:

- **Crédito**

Transição	Pão de Açúcar	Netflix	Lojas Americanas
A -> A	1,15	1,05	1,11
B -> B	1	1,05	1,08
C -> C	0,89	1,05	0,83
Média	1,013333333	1,05	1,006666667
B -> A	1,18	1,09	1,06
C -> B	1,15	0,9	1,3
C -> A	1,09	0,99	0,93
Média	1,14	0,9933333333	1,096666667
A -> C	0,68	0,84	0,7
A -> B	0,65	0,92	0,78
B -> C	0,85	0,83	0,81
Média	0,7266666667	0,8633333333	0,7633333333

Observa-se que o **Pão de Açúcar** é o melhor em alavancar e reduzir queda de gasto de clientes entre os três estabelecimentos analisados. O **Netflix** apresenta apenas relevância em reduzir queda de gasto de clientes, portanto clientes que realizam gastos no **Netflix** tendem a manter a mesma faixa de faturamento em todos os bimestres.

- **Débito**

Transição	Pão de Açúcar	MC Donalds	Riachuelo	Droga Raia
A -> A	1,1	1,1	1,08	1,17
B -> B	1,01	1,03	1,03	1,01
C -> C	0,89	0,87	0,84	0,81
Média	1	1	0,9833333333	0,9966666667
B -> A	1,16	1,16	1,07	1,27
C -> B	1,14	1,18	1,23	1,23
C -> A	1,2	1,16	1,19	1,37
Média	1,166666667	1,166666667	1,163333333	1,29
A -> C	0,73	0,73	0,75	0,6
A -> B	0,81	0,81	0,87	0,69
B -> C	0,82	0,78	0,87	0,73
Média	0,7866666667	0,7733333333	0,83	0,6733333333

Na tabela acima conclui-se que o **Pão de Açúcar**, **MCDonald's** e **Riachuelo** produzem comportamento similar em todos os quesitos, já a **Droga Raia** é superior em alavancar e reduzir queda de gasto em clientes.