DESENVOLVIMENTO DE MODELOS RFM COM AJUDA DE PLATAFORMAS DE APRENDIZADO DE MÁQUINA NA NUVEM

Leandro da Silva Carvalho

lcarvalho@ufrj.br

Programa de Engenharia Civil da COPPE/UFRJ, Universidade Federal do Rio de Janeiro

Cidade Universitária – CT Bloco B Sala 100, 21945-970, Rio de Janeiro, RJ, Brasil

Abstract. *O presente trabalho tem por objetivo aplicar técnicas de mineração de dados com a ajuda de uma plataforma de aprendizado de máquina na nuvem para a criação de modelos RFM. Tais modelos têm se mostrado uma ferramenta muito útil na adoção de estratégias de marketing direto para empresas. Embora tais modelos possam ser gerados de várias maneiras e por diversas plataformas existentes, o estudo teve como enfoque avaliar de forma prática a construção destes modelos a partir do ambiente em nuvem da Microsoft, denominado Azure Machine Learning (que é um dos mais recentes no mercado). A avaliação foi feita a partir de um conjunto de dados de um site de comércio eletrônico voltado para o setor de alimentação. Esta análise levou em consideração a criação de modelos RFM a partir da segmentação dos dados através da técnica de classificação não supervisionada k-means. Para cada modelo RFM gerado foi apresentado um conjunto de estratégias de marketing. Ao final, é apresentada uma conclusão sobre o uso da plataforma e os resultados obtidos durante o estudo.*

Keywords: *Modelos RFM, Aprendizado de Máquina, Computação em Nuvem, K-means.*

1 INTRODUÇÃO

Como a concorrência entre as empresas se torna mais intensa a cada dia, a maioria das empresas se concentram em objetivos de curto prazo, como aumentar as vendas e os volumes de transação, em vez de metas de longo prazo, como a melhoria da confiança dos consumidores e manutenção dos clientes fiéis (Hong et al, 2011).

Por isso, o presente trabalho tem por objetivo aplicar técnicas de mineração de dados com a ajuda de uma plataforma de aprendizado de máquina na nuvem para a criação de modelos RFM.

Tais modelos têm se mostrado uma ferramenta muito útil na adoção de estratégias de marketing direto, e permitem que os tomadores de decisão possam desenvolver estratégias de marketing únicas, ao examinar os registros dos históricos de compras feitas pelos clientes e seus respectivos gastos (Hughes, 1994).

Entretanto, embora as empresas tenham a capacidade para coletar estes dados, muitas destas companhias não sabem como promover o acesso a estas informações.

Por isso, um grande esforço aplicado à inteligência computacional tem sido feito para auxiliar na busca por técnicas que ajudem na extração de conhecimentos em grandes bases de dados. Estas técnicas estão relacionadas ao que a literatura denomina por mineração de dados ou aprendizado de máquina.

Contudo, tais técnicas estão praticamente disponíveis aos pesquisadores e aos cientistas de dados, o que dificulta a adoção pelas companhias. Em especial, pelas dificuldades impostas quanto às necessidades de vultosos investimentos na compra de equipamentos, softwares e contratação de equipe especializada.

Por isso, o desenvolvimento de aplicações de mineração de dados de forma mais eficiente e confiável é um desafio crítico para muitas empresas (Liu et al, 2011).

Como respostas a estas dificuldades, grandes empresas fornecedoras de software passaram a disponibilizar estes recursos de aprendizado de máquina a partir de um modelo de negócio que utiliza uma plataforma como serviço, o que permite uma contratação de acordo com a necessidade e a capacidade de investimento de cada companhia.

Isto só é possível graças aos recursos existentes na computação em nuvem, o que torna as soluções mais robustas, integradas e escaláveis.

Já que com este modelo de computação em nuvem, segundo Buyya et al (2009), as empresas e os usuários são capazes de acessar aplicativos de qualquer lugar do mundo sob demanda, transformando o desenvolvimento de software para milhões de clientes como um serviço, em vez de executar em seus computadores individuais.

Com isso, estas empresas não só conseguem oferecer um ambiente de desenvolvimento flexível, mas também garantem uma utilização eficaz e eficiente dos recursos computacionais para produzir resultados mais precisos, completos e oportunos (Delen et al, 2013).

Neste contexto, o trabalho proposto tem por objetivo analisar a criação de modelos RFM, através da segmentação de um conjunto de dados com o uso do algoritmo K-means, a partir do ambiente de aprendizado de máquina em nuvem da Microsoft, denominado Azure Machine Learning – Azure.microsoft.com (2015).

Neste estudo, o conjunto de dados avaliado é referente a um site de comércio eletrônico voltado para o setor de alimentação.

2 REVISÃO DO MODELO RFM E DO ALGORITMO K-MEANS

2.1 Modelo RFM

O modelo RFM é um método de classificação de registros de um banco de dados de clientes que permite saber quem são os compradores mais recentes, os compradores mais frequentes, e os mais rentáveis, ou seja, é uma forma que permite a criação de perfis através do comportamento de cada consumidor (Hughes, 1994).

Assim, com a adoção desses modelos, os tomadores de decisão podem efetivamente identificar os clientes mais valiosos para então desenvolver estratégias de marketing mais eficazes (Wei et al., 2010).

Isto é muito importante, porque é com a análise de informações, reconhecimentos e personalizações é que são atingidos os benefícios fundamentais para o desenvolvimento de um boa estratégia de marketing direto (Kahan, 1998).

Especificamente, o R de Recente denota a duração do período de tempo desde a última compra; já o F de Frequência significa o número de compras dentro de um determinado período; e o M de Monetário representa a quantidade de dinheiro gasto durante um determinado prazo (Wang, 2009).

Um outro benefício do uso do modelos RFM é que eles possuem uma fácil compreensão pelas equipes de marketing, o que facilita a sua implementação.

Uma maneira comum de adotar o modelo RFM é classificar os dados do cliente através de cada dimensão R, F e M; e, em seguida, dividir os segmentos em cinco partes iguais. Onde os dados de R são classificados com base nas datas de compra por ordem crescente, enquanto que F e M são classificados em ordem decrescente.

Os 20% dos registros iniciais do topo do ordenamento é codificado como 5, o próximo segmento de 20% é codificado como 4 e assim por diante. Deste modo, todos os clientes podem ser apresentados por rótulos como 555, 554,..., 111; o que dá um total de 125 possibilidades de classificação (5 x 5 x 5) (Wei et al., 2013).

Esta abordagem tradicional, rotulando todos os clientes com atributos de 1 a 5, que produz um total de até 125 grupos distintos pode gerar um problema, pois se os recursos são escassos, que tipo de cliente deverá ser trabalhado?

Neste caso, uma outra abordagem (Wu et al., 2009) pode ser realizada sem que as classificações do conjunto de dados dos atributos R, F e M sejam feitas em rótulos de 1 a 5.

Neste caso, o valor de R é reorganizado em ordem crescente, ou seja, da data de compra mais nova para a mais antiga, com a atribuição de um valor que representa a diferença da data de compra mais recente até a data de compra mais antiga, mas dentro de um determinado período. Já os valores de F e M são atribuídos, respectivamente, pelo número total de compras realizadas e pelo o gasto médio por cliente dentro do período da análise.

Neste estudo, a abordagem adotada será esta última apresentada e não a tradicional, de modo que R foi calculado de acordo com o intervalo de tempo em semanas (e não em dias), F com o seu valor de frequência absoluto e M com o valor do ticket médio.

2.2 Algoritmo K-Means

O algoritmo K-means – introduzido inicialmente por McQueen (1967) – é um método de agrupamento não hierárquico que tem por objetivo particionar um conjunto de dados em K grupos.

Sua execução é realizada através de um processo iterativo que trabalha baseado na minimização do erro de uma função de critério. Neste método, um determinado dado só pode pertencer a um único cluster por vez.

A principal vantagem do uso do algoritmo K-means é que ele é um método simples e de baixo custo computacional, o que permite ser executado de forma eficiente em grandes volumes de dados.

Entretanto, este método não traz nenhuma garantia quanto ao melhor resultado a ser obtido, já que a quantidade de centros a serem trabalhados pelo modelo deve ser definida previamente.

De qualquer forma, uma vez determinado os melhores valores para a parametrização inicial, o algoritmo K-means é sem dúvida um método muito eficiente na tarefa de agrupar dados devido a sua comprovada escalabilidade e eficiência.

3 APRESENTAÇÃO DA PLATAFORMA AZURE MACHINE LEARNING

O ambiente Azure Machine Learning (Azure ML) é um serviço em nuvem, desenvolvido para a criação de soluções de aprendizado de máquina em um modelo de plataforma de serviço.

Uma das principais vantagens deste serviço é a capacidade de integrar facilmente o desenvolvimento em um padrão de fluxo de trabalho repetitivo para criar soluções de análises preditivas. Isto o torna acessível tanto para um novato como para um cientista de dados experiente (Barnes, 2015).

A utilização destas plataformas em nuvem evitam a necessidade de instalação e execução dos aplicativos no computador do usuário, além das tarefas de manutenção e suporte (Zorrilla et al, 2013).

Seu uso é feito a partir de um navegador de Internet que se conecta com o ambiente de desenvolvimento, onde é possível criar modelos de análise e predição de acordo com a necessidade de cada empresa.

Um ponto interessante desta plataforma fica por conta da capacidade de criação de *Web Services* para acesso aos modelos gerados. Esta é uma grande vantagem, pois facilita a integração com outros sistemas, já que os *Web Services* são implementadas por um conjunto de tecnologias que fornecem os mecanismos para comunicação, descrição e descoberta de serviços (Devi et al., 2012).

Quanto aos custos, a cobrança da plataforma é feita pelo uso dos recursos computacionais a partir da quantidade de horas utilizadas e volume de processamento efetuado.

Para este estudo, foi usado o ambiente Microsoft Azure Machine Learning Studio, Studio.azureml.net (2015), que não tem qualquer tipo de cobrança ou necessidade de criação de uma conta Azure da Microsoft.

Este ambiente, por ser de graça (e não se sabe até quando), possui algumas limitações de execuções, como o volume de dados a ser processado e quantidade de predições que podem ser realizadas.

De qualquer modo, para a realização deste artigo, o estudo conseguiu utilizar os recursos deste ambiente muito bem.

4 ESTUDO DE CASO

4.1 Apresentação do conjunto de dados

A base de dados em questão possui um total de 228.188 mil registros transacionais de pedidos, e um total de 63.931 clientes cadastrados. Estes são dados históricos de mais de 10 anos de operação de um site de comércio eletrônico voltado para o setor de alimentação.

Porém, neste estudo, serão avaliados somente os registros referentes aos pedidos realizados entre o período de abril de 2015 a junho de 2015. Desta forma, serão analisadas as informações de um total de 9.952 pedidos e de 3.874 clientes.

Para a criação dos modelos o estudo definiu os respectivos valores de R, F e M da seguinte forma:

R – definido pelo registro de compra mais recente no banco de dados para cada cliente. Neste estudo, este valor foi calculado em semanas e não em dias. Assim, um cliente que fez uma compra em até 7 dias a partir do dia 30 de junho de 2015 ficou com um valor igual a 1. Já os clientes que fizeram compras entre o 8º e 14º dia, ficaram com valor 2 e assim por diante;

F – calculado pela quantidade de compras realizadas pelo cliente dentro do período de análise do estudo;

M – medido pelo valor do ticket médio de cada cliente.

A partir daí, o estudo passou a analisar o conjunto de dados através destas três perspectivas, ou seja, considerando apenas estes três atributos.

4.2 Estatísticas Básicas

A tabela 1 descreve as estatísticas básicas do conjunto de dados analisado.

Como pode ser visto, os valores mínimo e máximo do atributo R são 1 e 13 respectivamente. Para o atributo F, os valores são 1 e 66 respectivamente. E para o atributo M, os valores apresentados são, R$ 16,80 e R$ 372,50 respectivamente.

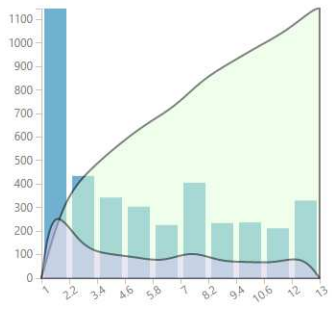
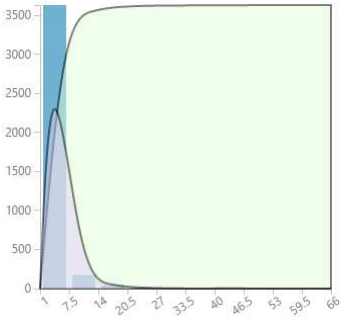
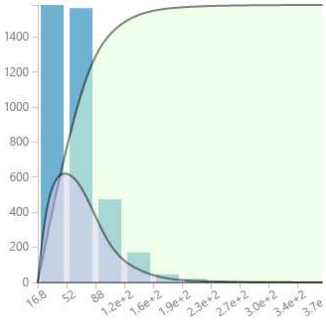
Tabela 1. Estatísticas básicas do conjunto de dados

	R (Recente)	F (Frequente)	M (Monetário)
Mínimo	1	1	16,80
Máximo	13	66	372,50
Média	5,464119773	2,568921012	67,42256582

Quanto aos histogramas de cada variável, eles podem ser visualizados nas imagens contidas da tabela 2.

Vale ressaltar que estes histogramas foram produzidos a partir do próprio ambiente Azure ML, e que cada histograma pode ser gerado com suas respectivas curvas de distribuição acumulativa (a curva esverdeada do gráfico) e de probabilidade de densidade (a curva de coloração rosa claro do gráfico).

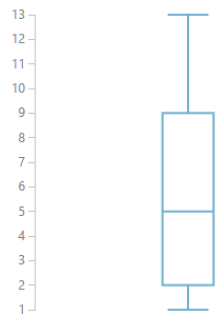
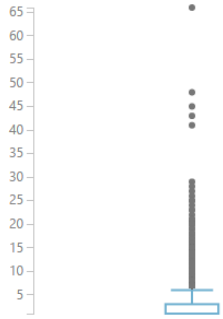
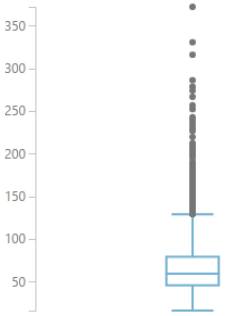
Tabela 2. Histograma das variáveis RFM

R (Recente)	F (Frequente)	M (Monetário)
		

Podemos notar que o atributo R é o mais balanceado entre todos e que o atributo F é o menos balanceado, o que pode ter um forte impacto no momento da criação dos modelos.

Em seguida, conforme pode ser visto na tabela 3, foram analisados os gráficos de box-plot de cada variável. Nele, é possível perceber que há uma grande quantidade de outliers tanto nas variáveis F e M, mas não na variável R.

Tabela 3. Gráficos box-plot das variáveis RFM

R (Recente)	F (Frequente)	M (Monetário)
		

4.3 Desenvolvimento do estudo

A plataforma Azure ML só possui o algoritmo K-means para a resolução de problemas de classificação não supervisionada. Sua execução é realizada a partir da variação de 4 parâmetros, conforme podem ser vistos na Figura 1:

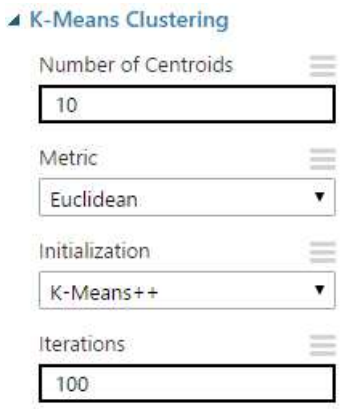


Figura 1. Parametrização do algoritmo K-means na plataforma Azure ML.

A descrição de cada parâmetro e as suas respectivas parametrizações são descritas a seguir:

- Número de Centroides (K): a quantidade grupos que o algoritmo irá trabalhar;
- Métrica de Distância: usada para calcular a distância entre os clusters. Opções de parametrização:
  - Euclidiana – a distância Euclidiana normalmente é usada como uma medida de dispersão de cluster do K-means. Essa métrica é preferencial porque minimiza a distância média entre pontos e os centroides;
  - Cosseno – como alternativa, é possível utilizar a função cosseno para medir a similaridade dos clusters. A similaridade por cosseno é útil em casos onde não se importa com o comprimento de um vetor, somente com seu ângulo.
- Método de Inicialização: escolha do algoritmo de inicialização dos centroides. Opções de parametrização:
  - First N – também chamado de método Forgy, onde um número inicial de pontos de dados é escolhido no conjunto de dados e usado como o meio inicial;
  - Random – também chamado de método de partição aleatória, onde o algoritmo coloca aleatoriamente um ponto de dados em um cluster e, em seguida, calcula a média inicial para ser o centroide dos pontos atribuídos aleatoriamente do cluster;
  - K-Means++ – que aprimora os meios de K usando um método melhor para escolher os centros de cluster iniciais, desenvolvido por Arthur et al (2007);
  - K-Means++ rápido – que é uma variante do algoritmo K-means++, porém mais otimizado;
  - Evenly – onde os centroides estão localizados equidistantes uns dos outros no espaço de D-Dimensional de N pontos de dados.
- Iterações: número de iterações a ser usado.

Como os resultados gerados pelo algoritmo sofrem influência direta com relação aos parâmetros de entrada – em especial pelo valor da variável utilizada na definição inicial da quantidade de clusters – o estudo optou por analisar somente a variação do atributo K para a geração dos modelos.

Nos demais parâmetros, foram utilizados sempre os mesmos valores, definidos por: métrica de distância – Euclideana; método de inicialização – K-means++; Iterações – 100.

Deste modo, diversos modelos foram criados dentro do ambiente utilizado com o objetivo de se encontrar a melhor distribuição de registros entre os clusters.

Mas para isso, o estudo estabeleceu como critério a busca por um modelo que pudesse gerar a maior quantidade de clusters possível, de modo que o menor cluster encontrado não tivesse menos do que 5% do total de registros.

Sendo assim, para cada modelo gerado, o seu resultado era analisado para saber se atendia ao critério de parada estabelecido pelo estudo. Caso contrário, um novo valor para o parâmetro K era atribuído e um novo modelo era gerado e analisado.

Como ponto de partida para a definição inicial do valor de K, o estudo fez uma rápida análise de segmentação com o uso de redes neurais SOM (Kohonen, 1995).

Neste caso, como o ambiente Azure ML não possui este tipo de rede, o estudo optou por utilizar o software de mineração de dados WEKA, que não é executado na nuvem e nem como plataforma de serviço, mas sim localmente em uma estação de trabalho.

O resultado desta análise obteve uma resposta indicando um total de 4 grupos. Assim, o estudo iniciou suas análises sobre a criação de modelos RFM na plataforma Azure ML a partir da criação de 4 grupos através do algoritmo K-means.

Dado o início do processo de desenvolvimento de modelos, a partir da variação do parâmetro de entrada K, e o critério de parada estabelecido anteriormente, o melhor modelo encontrado foi o que gerou um total de 10 clusters, conforme pode ser visto nas figuras 2, 3 e 4.

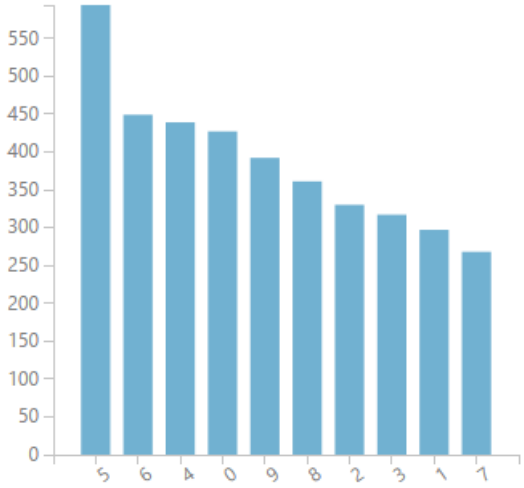


Figura 2. Histograma da distribuição dos registros entre os clusters.

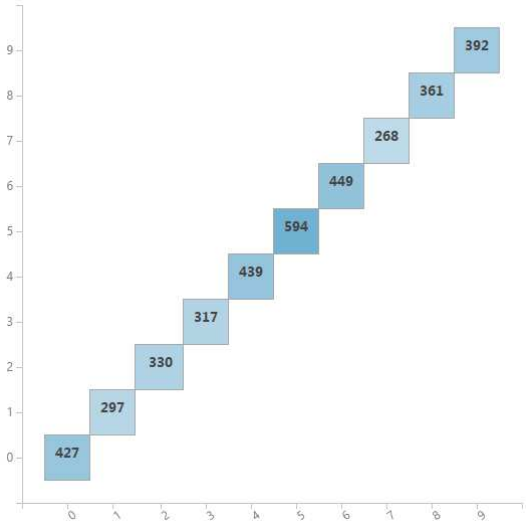


Figura 3. Gráfico com a quantidade de registros em cada cluster.

Conforme observado, o cluster 5 possui um total de 594 registros, o que representa 15% do conjunto de dados, e o cluster 7, possui um total de 268 registros, com 6,9% do total de dados.

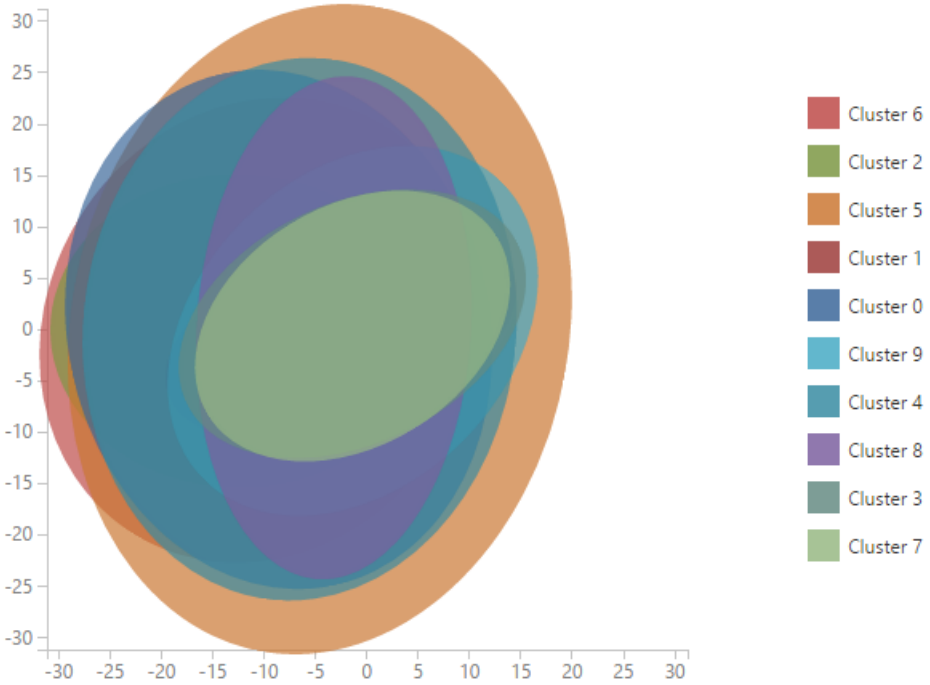


Figura 4. Visualização gráfica dos clusters a partir do ambiente Azure ML.

Na tabela 4 é possível observar os resultados obtidos em todos os clusters e o valor RFM de agrupamento (positivo ou negativo) de cada grupo.

Aqui vale explicar o que foi considerado no momento de definir os atributos positivos e negativos de agrupamento para cada modelo RFM gerado:

• No caso do atributo R, um cliente possui mais importância quanto mais recente for a sua compra, ou seja, quanto menor o valor de R, maior relevância possui este cliente. Assim, todos os clusters que possuem média menor do que a média global foram considerados positivos. Já os clusters que possuem média maior do que a média global foram considerados negativos;

• Quanto aos atributos F e M, um cliente possui maior valor quanto mais frequente ele for e quanto maior for o seu ticket médio. Sendo assim, todos os clusters com valores maiores do que a média global, foram classificados como positivos. Já os clusters com valores menores, foram classificados como negativos.

Tabela 4. Clusters gerados e seus respectivos modelos de agrupamento RFM

Cluster	Tamanho	Média R	Média F	Média M	Grupo RMF +	Grupo RFM-
0	427	8,55	1,44	72,13	M	RF
1	297	2,43	1,00	59,86	R	FM
2	330	12,52	1,05	67,57	M	RF
3	317	1,36	6,18	45,61	RF	M
4	439	6,49	1,84	67,81	M	RF
5	594	4,51	2,01	65,81	R	FM
6	449	10,47	1,25	70,62	M	RF
7	268	1,34	4,36	63,96	RF	M
8	361	3,15	3,09	78,11	RFM	-
9	392	1,31	4,81	76,41	RFM	-

Podemos observar que temos cluster com valores positivos para: M, R, RF e RFM; e valores negativos para: RF, FM e M.

Estes grupos podem ser descritos da seguinte forma:

• Clusters 0, 2, 4 e 6 – M (Positivo) e RF (negativos) – clientes rentáveis, ou seja, clientes que possuem um alto valor no ticket médio, mas que não são recentes e nem frequentes.

Assim, podemos considerar, pelo alto valor da média de R e do baixo valor da média de F, que este é o típico caso do cliente perdido (por não ser recente), e que não teve uma experiência de compra muito satisfatória (por não ser frequente).

Nestes casos, como estes clientes possuem um valor de ticket médio alto, valeria a pena, em uma futura estratégia de marketing, abordar estes clientes para entender os motivos que os

levaram a não comprar novamente, em especial com os clientes dos clusters 0 e 6 que possuem valores de ticket médio mais altos do que os dos clusters 1 e 4.

Outro dado importante é que os clientes do cluster 4 possuem um média no atributo F maior do que a dos clusters 0, 2 e 6. Isto pode indicar que estes clientes estão mais propensos a fazer uma nova compra, principalmente se comparado aos clientes do cluster 2, que possui o menor valor médio para o atributo F.

Resumindo, os clientes destes clusters são os clientes perdidos, mas potencialmente rentáveis, ou seja, clientes com um valor de ticket médio alto e que não compram mais, ou que migraram para um concorrente, ou que deixaram de consumir alimentos fora de casa, ou que tiveram uma experiência de compra ruim.

• Clusters 1 e 5 – R (Positivo) e FM (negativos) – estes são os clientes mais recentes, mas que não possuem uma taxa de frequência alta e que não são tão rentáveis.

O que diferencia o cluster 1 do cluster 5 é que o primeiro é basicamente formado por novos clientes (frequência de compra igual a 1), enquanto que o segundo é formado por pessoas que já compraram mais de uma vez (frequência de compra maior do que 1).

Neste caso, fica como sugestão para futuras estratégias de marketing a elaboração de campanhas que permitam o aumento do ticket médio destes clientes, como, por exemplo: campanhas de *Up-selling* (quando se expõe aos clientes produtos *Premium*, ou seja, que são mais caros do que os tradicionais) e/ou campanhas de *Cross-selling* (para fazer com que os clientes gastem mais adicionando novos produtos ao carrinho de compras).

Outro ponto importante é manter a fidelização destes clientes para fazer com que eles se tornem compradores recorrentes. Diante disso, se faz necessário garantir aos novos clientes uma boa experiência de compra e um ótimo atendimento no pós venda.

• Clusters 3 e 7 – RF (Positivos) e M (negativo) – formado pelos clientes fiéis, aqueles clientes que compram com frequência e que continuam ativos, mas que não são rentáveis, ou seja, compram pouco e podem estar associados aos clientes que só querem promoções e descontos.

Neste caso, vale um esforço para tentar migrar estes clientes para um nível de monetização maior. Assim, estratégias de *Up-selling* e *Cross-selling*, conforme descritas anteriormente, se tornam essenciais para a conversão destes consumidores em clientes rentáveis.

Inclusive, é possível notar que no cluster 7 os clientes já estão bem próximos de atingir a média do atributo M. Deste modo, uma simples adição de mais um produto ao carrinho (feito através de um sistema de recomendação) já seria, quem sabe, o suficiente para alcançar este objetivo.

• Clusters 8 e 9 – RFM (Positivos) – estes são os melhores consumidores, ou seja, aqueles que são fiéis, rentáveis e que possuem um alto valor para empresa. São os clientes VIPs, aqueles que a empresa deve tratar com o maior carinho e com a máxima atenção.

Para estes clientes vale a pena oferecer um serviço de *Concierge* (assistente pessoal) para resolver qualquer eventualidade, além de serviços exclusivos como uma entrega expressa e um atendimento especializado, prioritário e direto.

5 CONCLUSÕES

Com este estudo foi possível compreender o funcionamento de uma plataforma de aprendizado de máquina na nuvem para a criação de modelos RFM através do uso de uma técnica de classificação não supervisionada.

Além disso, foi possível identificar as variáveis que impactam na criação dos modelos RFM e que tipos de estratégias de marketing poderão ser adotadas em futuras campanhas.

No que diz respeito à análise, ela demonstra que, embora o modelo de classificação tenha produzido 10 clusters, é possível notar que apenas 4 modelos RFM distintos foram criados, ou seja, com os mesmos valores de agrupamento RFM (positivo ou negativo) nos clusters gerados.

Se comparado ao modelo tradicional propostos Hughes (1994), que pressupõe a criação de 125 grupos RFM por análise, a criação de modelos RFM a partir de técnicas de mineração de dados pode ser encarada com uma grande vantagem, já que ela reduz o número de grupos e informa com mais precisão o entendimento de cada grupo

Ainda assim, fica claro que mesmo que um cluster possua os mesmos valores de agrupamento RFM (positivo ou negativo) de um outro, eles podem possuir características distintas de valores nos atributos RFM (quando analisados somente os atributos de forma isolada, e não o agrupamento).

Isto permite que clusters com os mesmos agrupamentos RFM possam ser trabalhados de forma distinta, o que dá mais liberdade aos estrategistas de marketing das empresas.

No que diz respeito à plataforma, é possível dizer que ela é bastante prática, visto que não há qualquer necessidade de configuração de hardware e nem a instalação de software, o que permite o foco naquilo que é mais importante, a análise dos dados.

Além disso, ela possui uma abordagem orientada a serviços, que representam um modelo padrão para compartilhamento de recursos em sistemas distribuídos e oferece uma abordagem genérica para a integração com diversos sistemas (Olejnik et al, 2009).

Por ser um ambiente novo, ainda há muito para se desenvolver, já que a quantidade de algoritmos é limitada e há poucos recursos para visualização dos dados diretamente no ambiente, em especial à visualização da formação sobre cada cluster.

Como uma primeira análise, vale a pena ressaltar que talvez este tipo de plataforma seja mais bem indicada para os métodos que não precisem de muita interação com um especialista do negócio, ou seja, modelos que possam ser automatizados, como por exemplo, sistemas de recomendação.

Isto porque na análise de classificação não supervisionada os modelos gerados sempre precisam ser avaliados por um conhecedor da área fim, e neste ponto, é possível dizer que a ferramenta não possui os recursos mais apropriados para análise.

De todo modo, é possível dizer sem dúvida alguma que a ferramenta é capaz de classificar um conjunto de dados. Mas é preciso melhorar este ponto específico sobre a forma de visualização dos resultados obtidos.

O ponto positivo do uso deste ambiente é a vantagem de não precisar ficar se preocupando em levar a estação de trabalho e os dados para todo lugar, uma vez que todas as informações ficam armazenadas no própria ambiente, ou seja, na nuvem.

Outra vantagem é que, por utilizar um modelo de plataforma como serviço, ela facilita o desenvolvimento e implantação de aplicativos sem os custos e a complexidade de comprar e gerenciar as camadas de hardware e software subjacentes (Marston et al, 2011).

Porém, esta vantagem traz um risco, pois não há como garantir que a segurança dos dados estará protegida fora do ambiente da empresa.

Afinal, segundo Demirkan et al (2013), a manipulação de dados fora das instalações do local da empresa aumenta o número de potenciais riscos de segurança. Além disso, na computação em nuvem, os dados são fisicamente localizados em um determinado país e está sujeita a regras e regulamentações locais.

Entretanto, para proteger os dados nos ambientes de nuvens, técnicas tradicionais de segurança como criptografia de dados e autorização tem fornecido uma base sólida para tais objetivos, mas ainda assim são passíveis de falhas (Sun et al, 2011).

De qualquer forma, a análise do estudo foi bastante positiva, indicando um campo de estudo bastante promissor para novas pesquisas e aplicação das técnicas de mineração de dados pelas empresas.

REFERENCES

Arthur, D., & Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding. *In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Azure.microsoft.com, 2015. *Machine Learning - Predictive analytics | Microsoft Azure*. [online] Available at: http://azure.microsoft.com/en-us/services/machine-learning/ [Accessed 17 Jul. 2015].

Barnes, J., 2015. *Microsoft Azure Essentials: Azure Machine Learning*. Redmond: Microsoft Press, p.25.

Buyya, R., Yeo, C., Venugopal, S., Broberg, J. and Brandic, I., 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), pp.599-616.

Delen, D. and Demirkan, H., 2013. Data, information and analytics as services. *Decision Support Systems*, 55(1), pp.359-363.

Demirkan, H. and Delen, D, (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, 55(1), pp.412-421.

Devi, B., Devi, Y., Rani, B. and Rao, R., 2012. Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce. *Procedia Engineering*, 30, pp.20-27.

Hong, I. and Cho, H., 2011. The impact of consumer trust on attitudinal loyalty and purchase intentions in B2C e-marketplaces: Intermediary trust vs. seller trust. *International Journal of Information Management*, 31(5), pp.469-479.

Hughes, A., 1994. *Strategic database marketing*. Chicago, Ill.: Probus Pub. Co.

Kahan, R., 1998. Using database marketing techniques to enhance your one-to-one marketing initiatives. *Journal of Consumer Marketing*, 15(5), pp.491-493.

Kohonen, T., 1995. *Self-organizing maps*. Berlin: Springer.

Liu, H. and Xu, J., 2011. Web Service Framework Research of Data Mining in E-business. *Procedia Engineering*, 15, pp.1968-1972.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California, pp.281-297.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J. and Ghalsasi, A., 2011. Cloud computing — The business perspective. *Decision Support Systems*, 51(1), pp.176-189.

Olejnik, R., Fortiş, T. and Toursel, B., 2009. Webservices oriented data mining in knowledge architecture. *Future Generation Computer Systems*, 25(4), pp.436-443.

Studio.azureml.net, 2015. *Microsoft Azure Machine Learning Studio*. [online] Available at: https://studio.azureml.net [Accessed 17 Jul. 2015].

Sun, D., Chang, G., Sun, L. and Wang, X., 2011. Surveying and Analyzing Security, Privacy and Trust Issues in Cloud Computing Environments. *Procedia Engineering*, 15, pp.2852-2856.

Wang, C., 2009. Robust segmentation for the service industry using kernel induced fuzzy clustering techniques. *Industrial Engineering and Engineering Management, 2009. IEEM 2009. IEEE International Conference on*, pp.2197-2201.

Wei, J., Lin, S. and Wu, H., 2010. A review of the application of RFM model. *African Journal of Business Management*, 4(19), pp.4199-4206.

Wei, J., Lee, M., Chen, H. and Wu, H., 2013. Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, 40(18), pp.7513-7518.

Wu, H., Chang, E. and Lo, C., 2009. Applying RFM Model and K-Means Method in Customer Value Analysis of an Outfitter. *Advanced Concurrent Engineering*, pp.665-672.

Zorrilla, M. and García-Saiz, D., 2013. A service oriented architecture to provide data mining services for non-expert data miners. *Decision Support Systems*, 55(1), pp.399-411.