Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Rafael Roberto Dias

Métodos de Aprendizado de Máquina Aplicados ao *E-Commerce*

Curitiba
2019

Rafael Roberto Dias

Métodos de Aprendizado de Máquina Aplicados ao *E-Commerce*

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Wagner Hugo Bonat

Curitiba
2019

Métodos de Aprendizado de Máquina Aplicados ao *E-Commerce*

Machine Learning Methods Applied to textit E-Commerce

Rafael Roberto Dias[1]

[1]Especialização em Data Science & Big Data, Universidade Federal do Paraná, Curitiba, PR, Brasil

Resumo

O comércio eletrônico é uma modalidade de negócio escalável para atender milhares de clientes simult aneamente, via internet, e o desafio está em manter o relacionamento com estes consumidores de maneira personalizada. Este estudo teve como objetivo contribuir com uma parte desta personalização utilizando metodologias de aprendizado de máquina para classificar cada cliente, logo após a aprovação da sua primeira compra, em tres segmentos: Bronze, Prata e Ouro. Os classificadores supervisionados utilizados foram: método de vetores de suporte, rede neural artificial e árvore de decisão, uma vez que a empresa MadeiraMadeira já possui os clientes rotulados nos tres segmentos citados anteriormente. Realizou-se a análise descritiva das variáveis transacionais, que são as informações contidas em cada compra, extraídas diretamente do banco de dados, e escolhidas as tres principais para aplicação das metodologias. Cada modelo foi configurado, executado e avaliado comparando-se os resultados, e após identificou-se que a Árvore de Decisão é a com maior potencial para implantação em ambiente de produção pela sua acurácia e simplicidade de interpretação dos seus resultados.

Abstract

E-commerce is a scalable business modality to serve thousands of clients simultaneously via the internet, and the challenge lies in maintain the relationship with these consumers in a personalized way. This study aimed to contribute with a part of that customization using machine learning techniques to evaluate each customer shortly after their first purchase is approved, in three segments: Bronze, Silver and Gold. Supervised classification methods used were: support vector machine, artificial neural network and decision tree since the company MadeiraMadeira already has customers labeled in all three segments mentioned above. We performed a descriptive analysis of transactional variables, which are the information contained in each purchase, extracted directly from the database, and the top three were chosen for the application of the methods. Each model has been configured, executed and compared the results, and after it was identified that the decision tree is the one with the greatest potential for implamentation because of its accuracy and simplicity of interpretation of its results.

1. Introdução

1.1. Contexto

O comércio eletrônico (*e-commerce*) é a atividade que mais cresce mundialmente juntamente com a tecnologia de processamento de grandes volumes de informações, possibilitando atender milhares de clientes simultaneamente. A empresa MadeiraMadeira S/A é um e-commerce especializado na venda de itens para residencias e empresas, como eletrodomésticos, eletroeletrônicos, ferramentas, iluminação, tendo como foco principal a venda de móveis de madeira. É a única do setor que trabalha na modalidade *dropshipping*, ou seja, sem estoque físico, coletando os produtos nos fornecedores e os entregando diretamente aos clientes.

Hoje existe uma segmentação de clientes conforme comportamento de compra no decorrer de um ano móvel que é atualizado utilizando método de aprendizado de máquina, nao supervisionado, *chamado K-Means*. Existem tres grupos distintos de clientes: Ouro, Prata e Bronze. Cada um possui um comportamento de compra específico, e precisam ser abordados de maneira diferenciada pela área de Marketing,

Figura 1: Ciclo de Vida dos Clientes

conforme as melhores práticas de *marketing one to one*, ou comunicação direta.

Como a metodologia atual utiliza um ano móvel de compras dos clientes nao é possível classificar o cliente conforme a sua primeira compra. O período de um ano é necessário para entender o comportamento de cada segmento neste intervalo como a frequencia de compra, ticket médio e tempo médio desde a última compra.

O objetivo deste estudo é encontrar uma metodologia de aprendizado de máquina que possibilite classificar os novos clientes imediatamente após a aprovação da sua primeira compra, o que permitirá abordá-los conforme as estratégias desenhadas para cada segmento, utilizando os meios de comunicação diretos: e-mail, SMS, aplicativos de mensagens eletrônicas, mídias sociais, etc.

Existem práticas para entender o ciclo de compra dos clientes que são utilizadas na atualidade por várias empresas de comércio eletrônico, inclusive por outros tipos de empresas, e é chamado de Ciclo de Vida dos Clientes. Consiste basicamente em cinco etapas: Aquisição, Ativação, Manutenção, Retenção e Recuperação, conforme Figura 1. A etapa de Aquisição consiste em adquirir os meios de contatos diretos dos clientes (e-mail, SMS); Ativação é quanto o cliente realiza a primeira compra; Manutenção é onde os clientes recompram e possuem uma frequencia de compras definida; Retenção é quando o cliente já está a mais tempo sem comprar quando comparado com a sua frequencia média; Recuperação é quanto o cliente está a muito tempo sem comprar, o que indica uma propensão de que nao voltará a comprar novamente com a empresa.

A importância de ter o cliente classificado desde a primeira compra será para abordá-lo de maneira adequada logo após a sua ativação, dado que, por exemplo, os clientes do segmento Ouro tem uma frequencia de compra de tres vezes ao ano, logo a recompra destes tende a ocorrer em média a cada quatro meses. Logo os clientes do segmento Ouro entrarao na etapa de Retenção no quinto mes, diferentemente do segmento Bronze, que entrará em Retenção no décimo terceiro mes sem compras. Existem também estratégias para fazer os clientes migrarem de um segmento para outro, assim como fazer os clientes do segmento Ouro aumentarem a frequencia de compra anual, diminuindo o intervalo entre elas. Estas ações tem como objetivo maximizar o resultado da empresa através de um melhor relacionamento com sua base de clientes. A metodologia estudada neste artigo permitirá agir com este propósito de maneira antecipada.

2. Metodologia

Dado que a empresa MadeiraMadeira possui um conjunto de dados rotulados, contendo os tres segmentos e seus respectivos clientes, foram coletadas as informações transacionais da primeira compra de cada indivíduo presente na base. Após a junção das bases numa única, trataram-se os clientes sem histórico ou com dados incompletos. Neste caso empregou-se as metodologias de aprendizagem de máquina, supervisionados, para classificação dos segmentos. Escolhidos tres modelos comumente utilizados para este tipo de classificação, pela eficiencia e abundância de materiais explicativos, que são: máquina de vetores de suporte ou *support vector machine*, rede neural artificial ou *artificial neural network* e árvore de decisão. E para entendimento das variáveis presentes na primeira compra dos clientes, aplicaram-se testes estatísticos para obtenção de uma breve análise descritiva para escolha das que foram utilizadas para teste e escolha do modelo mais adequado.

2.1. Finalização: visualização, análises e recursos computacionais

Tanto a partir dos resultados intermediários na metodologia, bem como dos resultados da fusão por Borda *count*, foram combinados e produzidos visualizações em gráficos e tabelas. Após a aplicação de cada método foi necessário compatibilizar corretamente os resulta-

dos (e.g. resíduos ou *score*) com o período em que se referem.

Os principais recursos computacionais utilizados foram:

- R [20]
  - Versão 3.5.1;
  - Pacotes `benford.analysis` e `stl`;
  - IDE RStudio Desktop 1.2.1335.
- Python [21]
  - Versão 3.6.5;
  - Biblioteca `rrcf`;
  - IDE Jupyter Notebook 5.5.0.

Na próxima seção são analisados, demonstrados e argumentados os resultados obtidos.

3. Resultados

As informações sobre os segmentos foram extraídas através do uso do software estatístico R versão 3.5.1, que tem como vantagem ser *opensource*, conectando diretamente com o banco de dados relacional *MySQL*, tanto para extrair as informações dos segmentos e seus clientes, como para coletar as informações da primeira compra dos mesmos e finalizado com o cruzamento das duas

bases transformando em uma única tabela de dados. As informações transacionais coletadas das compras foram: valor ticket absoluto que é o valor dos itens somado com o valor de frete, valor do frete, quantidade de parcelas de pagamento, quantidade de itens, quantidade de sku que significa *stock keeping unit* ou unidade de manutenção de estoque, quantidade de categorias e quantidade de departamento. A proporção de clientes para cada segmento são: Bronze são 79%, Prata são 18% e Ouro são 3%, conforme distribuição representada na Figura 2.

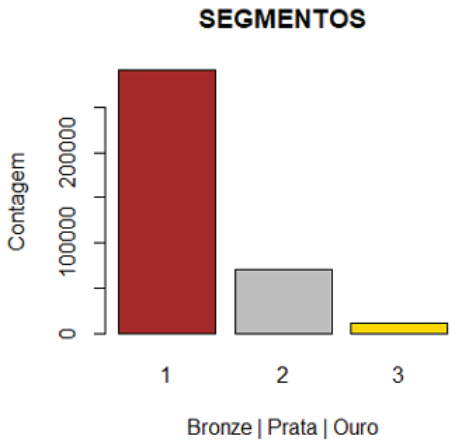

Figura 2: Distribuição dos Segmentos

Observa-se que as variáveis com diferenças marcantes são: ticket absoluto, frete, itens e categoria verificados na Tabela 1 e são possíveis candidatos para entrarem na modelagem. Quando as variáveis são analisadas com auxílio do *box-plot*, conforme Figuras 3, 4 e 5, verifica-se que as variáveis ticket absoluto, categoria e SKU tem uma distribuição distinta para cada segmento.

A análise de correlação representada na Figura 6 indica que as variáveis ticket absoluto e frete tem forte correlação positiva, o que se deve pelo fato do valor do frete pago pelos clientes estar contido no valor absoluto. As variáveis categoria, sku e departamento também apresentam forte correlação positiva.

Com as variáveis selecionadas, ticket absoluto, categoria e sku, o conjunto de dados foi separado em treino e teste, com 50% para cada partição. Escolhido o software R para aplicar as modelagens, juntamente com o pacote *caret* 3. Utilizados para otimizar os modelos as metodologias *K-fold* e

validação cruzada 2. Estas servem para avaliar a qualidade do desempenho dos classificadores selecionados para este estudo: SVM, ANN e árvore de decisão.

O primeiro método testado foi a máquina de vetores de suporte (SVM). Observado desempenho de processamento em torno de quinze minutos, configurado com *K-fold* de tamanho tres e com tres repetições. A matriz de confusão 2, representada na Tabela 2, demonstra o percentual de acerto entre os valores preditos para cada segmento em comparação aos valores rotulados do conjunto de teste.

O segundo método utilizado foi redes neurais artificiais, com processamento de sessenta minutos, igual-

Tabela 1 Análise descritiva das variáveis da primeira compra de cada segmento

			Mediana				
Segmentos	Nº de Parcelas	Ticket	Frete	Itens	Categoria	SKUs	Departamento
Bronze	4	R$ 427,20	R$ 65,80	1	1	1	1
Prata	5	R$ 859,90	R$ 103,22	2	1	2	1
Ouro	6	R$ 1.240,40	R$ 143,05	3	2	2	1

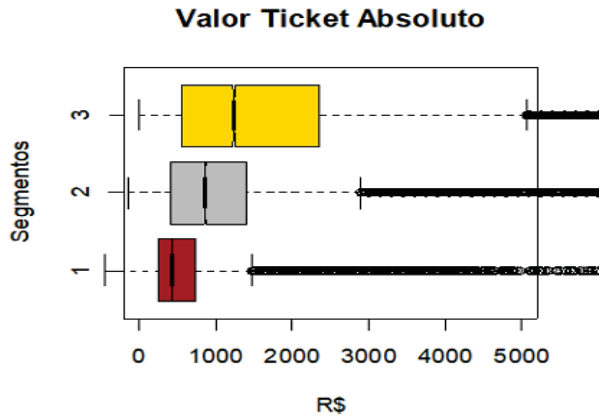

Figura 3

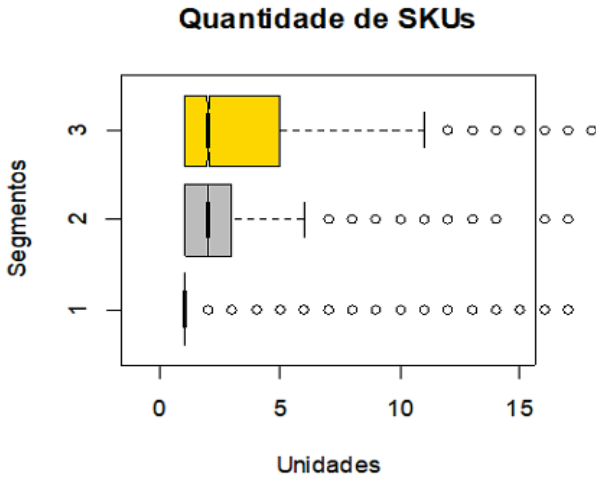

Figura 5

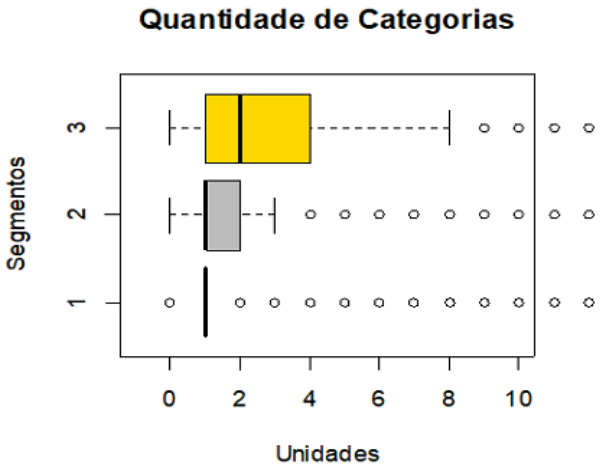

Figura 4

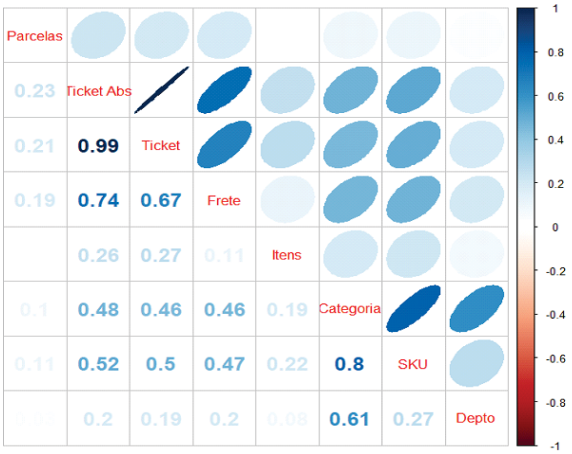

Figura 6: Correlação entre as Variáveis

mente configurado com *K-fold* de tamanho tres e com tres repetições. Tem-se a matriz de confusão demonstrada na Tabela 3, da mesma forma que a anterior.

E o terceiro método utilizado foi a de árvore de decisão, com processamento em torno de 3 minutos, configurado conforme os anteriores e com a respectiva matriz de confusão com os resultados de predição versus reais, contidos na Tabela 4.

A acurácia de cada metodologia 2 é comparada na Figura 7 para auxiliar na escolha de qual será implantada para prever os segmentos dos novos clientes.

O método de redes neurais se mostrou o mais acurado dentre os tres conforme a validação cruzada e repetições utilizadas. Através da Figura 7 verifica-se empate técnico entre as metodologias levando em consideração os intervalos de confiança.

Das regras geradas pela metodologia de árvore de decisão, os nós com maior probabilidade de acerto ao classificar os clientes estao demonstradas na Figura 8, e pode descrever que os clientes do segmento Ouro

serao os que comprarem mais de uma categoria, tiverem ter um ticket maior ou igual a R$832,90 e que comprarem mais de 3 skus.

Tabela 2 Matriz de Confusão

Support Vector Machine (SVM)			
Predição	Bronze	Prata	Ouro
Bronze	98,40%	54,50%	44,47%
Prata	1,54%	45,07%	48,46%
Ouro	0,06%	0,43%	7,07%

Tabela 3 Matriz de Confusão

Redes Neurais (ANN)			
Predição	Bronze	Prata	Ouro
Bronze	98,64%	55,15%	44,40%
Prata	1,12%	41,91%	22,82%
Ouro	0,24%	2,94%	32,78%

Tabela 4 Matriz de Confusão

Árvore de Decisão			
Predição	Bronze	Prata	Ouro
Bronze	98,55%	56,16%	46,19%
Prata	1,30%	43,34%	30,42%
Ouro	0,14%	0,50%	23,39%

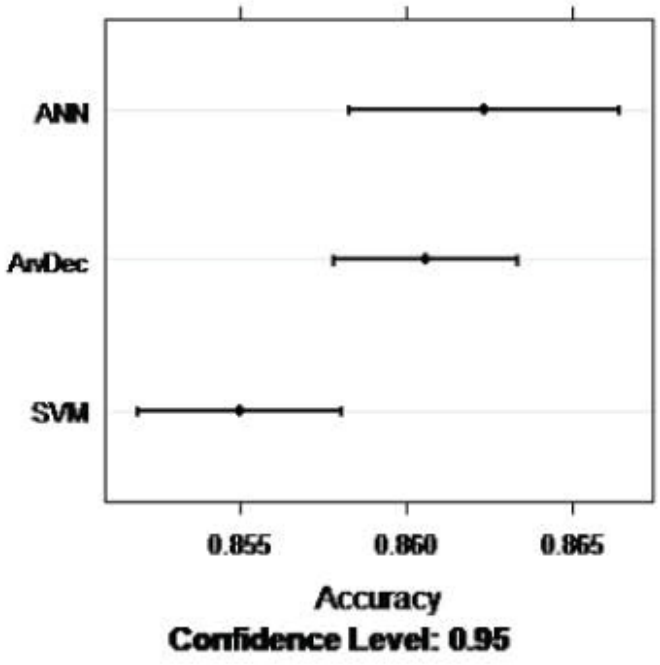

Figura 7: Comparação das Acurácias das Metodologias

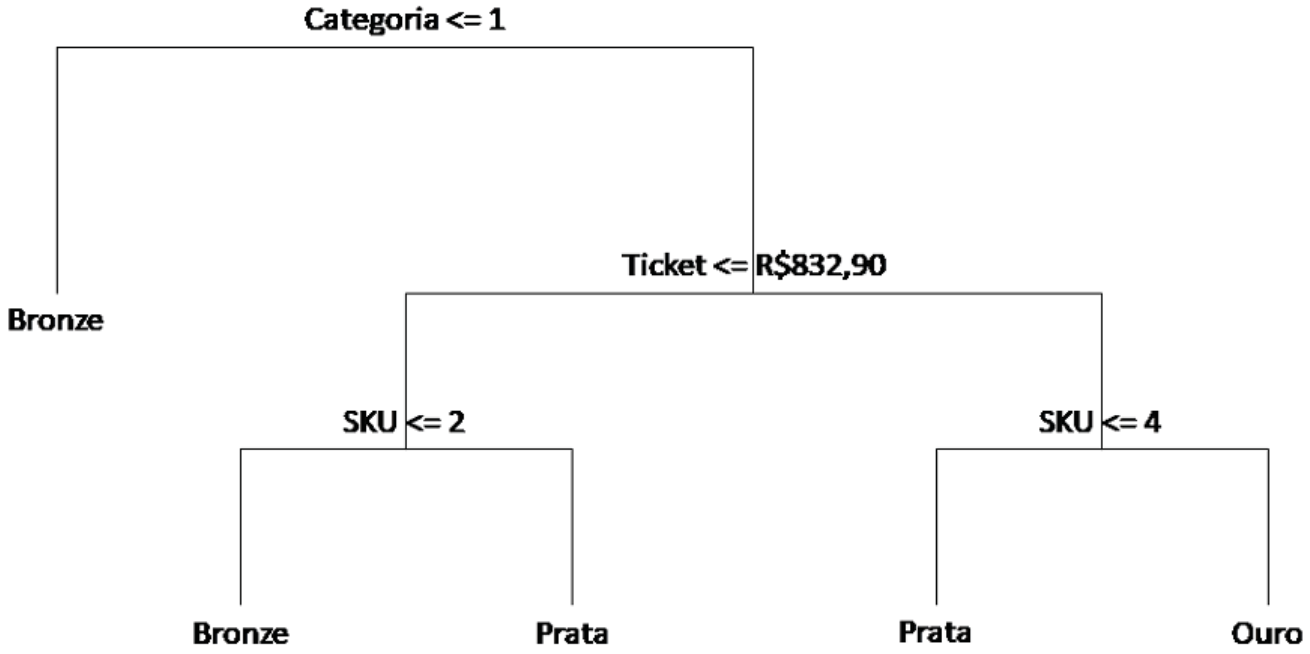

Figura 8 Regra Árvore de Decisão

4. Conclusões e Trabalhos Futuros

Os resultados demonstram que existe uma parcela importante dos clientes contidos nos segmentos Prata e Ouro que possuem o mesmo comportamento na sua primeira compra quando comparados com o segmento Bronze. Variáveis demográficas poderiam tem ajudado a diferenciar com mais acurácia os segmentos, e nao foram utilizados em decorrencia da falta destas informações na base de dados dos clientes na MadeiraMadeira. Assim mesmo, os métodos de rede neural e o de árvore de decisão demostram maior capacidade para classificar corretamente os clientes nos segmentos Prata e Ouro, diferentemente o método máquina de vetores de suporte tem baixo poder de predição para o segmento Ouro.

Comparando os métodos mais acurados, o de árvore de decisão 3 mostra-se mais viável para implantação por ter um custo computacional baixo e por ter o resultado fácil de ser interpretado, uma vez que as regras podem ser extraídas dos nós com maior probabilidade de acerto que podem ser implantadas diretamente em banco de dados utilizando funções " *if* ", " *ifelse*", " *then*"para classificar os clientes imediatamente após a aprovação da primeira compra.

Para utilizar a rede neural treinada com o conjunto de dados rotulados será necessário subir um serviço com ela em memória e enviar as informações de compras de cada cliente para que seja classificado, comumente utilizado um servidor recebendo requisições

com estas informações e enviando qual é o segmento que cada cliente pertence logo a sua primeira compra. O uso deste método neste caso seria justificável caso a sua acurácia fosse ao menos 20% superior ao de árvore de decisão.

Referencias

1) Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introdução ao Data Mining. Rio de Janeiro: Editora Moderna Ltda; 2009.
2) Max Kuhn, Kjell Johnson. Applied Predictive Modeling. New York: Springer Science+Business Media; 2013.
3) Max Kuhn. The Caret Package.
caret.r-forge.r-project.org/