Retenção de Alunos no Ensino Superior Privado usando Machine Learning

Autoria

Francisco Coimbra Carneiro Pereira - f.coimbra.pereira@gmail.com
Mestr e Dout em Admin de Empresas/IAG-A Esc de Negócios da PUC-Rio - IAG/PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro

Jorge Brantes Ferreira - jorgebf@gmail.com
Mestr e Dout em Admin de Empresas/IAG-A Esc de Negócios da PUC-Rio - IAG/PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro

Andrea Ribeiro Carvalho de Castro - andrearc.castro@gmail.com
Mestr e Dout em Admin de Empresas/IAG-A Esc de Negócios da PUC-Rio - IAG/PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro

Cristiane Junqueira Giovannini - mestrekis@gmail.com
Mestr e Dout em Admin de Empresas/IAG-A Esc de Negócios da PUC-Rio - IAG/PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro

Bacharelado em Administração/IBMEC-RJ - Faculdades IBMEC Rio de Janeiro

Fernanda Leão Ramos - leaoramos@gmail.com
Mestr e Dout em Admin de Empresas/IAG-A Esc de Negócios da PUC-Rio - IAG/PUC-Rio - Pontifícia Universidade Católica do Rio de Janeiro

Resumo

O ensino superior privado perde em média mais de 20% da base de alunos a cada semestre. Esta evasão de alunos representa um desafio para a gestão das instituições de ensino. Para combater este problema, diversos tratamentos são usados. A identificação e a diferenciação dos alunos são os primeiros passos necessários para aplicar uma estratégia de marketing de relacionamento personalizada para a retenção de clientes. Sendo assim, este trabalho apresenta uma metodologia quantitativa para classificação de risco de evasão de alunos ativos. Foram utilizados dados históricos reais de alunos que evadiram ou se formaram em uma instituição de ensino superior, para a geração de modelos preditivos por algoritmos de machine learning. Estes modelos foram calculados e comparados e, posteriormente, usados na classificação dos alunos ativos. Em sequência, foi estimado o lifetime value destes estudantes para servir de base para a definição de estratégias de retenção.

Retenção de Alunos no Ensino Superior Privado usando Machine Learning

RESUMO

O ensino superior privado perde em média mais de 20% da base de alunos a cada semestre. Esta evasão de alunos representa um desafio para a gestão das instituições de ensino. Para combater este problema, diversos tratamentos são usados. A identificação e a diferenciação dos alunos são os primeiros passos necessários para aplicar uma estratégia de marketing de relacionamento personalizada para a retenção de clientes. Sendo assim, este trabalho apresenta uma metodologia quantitativa para classificação de risco de evasão de alunos ativos. Foram utilizados dados históricos reais de alunos que evadiram ou se formaram em uma instituição de ensino superior, para a geração de modelos preditivos por algoritmos de *machine learning.* Estes modelos foram calculados e comparados e, posteriormente, usados na classificação dos alunos ativos. Em sequência, foi estimado o *lifetime value* destes estudantes para servir de base para a definição de estratégias de retenção.

Palavras-chave: Modelos preditivos, classificação, evasão de alunos, *lifetime value, machine learning.*

INTRODUÇÃO

Devido a expansão da oferta nas instituições de ensino superior, houve um aumento do número de alunos matriculados de 3,8 milhões em 2003, para 7,3 milhões de alunos em 2013 (INEP, 2015). Este fato ocorreu, principalmente, nas instituições de ensino privado posicionadas para atender às classes C e D. No entanto, mesmo com o crescimento da demanda, a taxa de evasão manteve-se acima de 20% por semestre ao longo do período. Esse é um problema que ocorre principalmente no ensino privado, onde o preço é um item importante na decisão do aluno de permanecer na instituição.

Com relação à evasão, este tema aparece na literatura dividido em três tipos: evasão do curso, da instituição e do sistema (LOBO, 2012). Na primeira, o aluno muda de curso, mas permanece na Instituição de Ensino Superior (IES), na segunda, muda de IES e na terceira, abandona o ensino o superior. O interesse da IES é concentrado nas duas últimas pois representam uma perda para a instituição. No caso da evasão por mudança de curso, o aluno continua na instituição, apenas consumindo um serviço diferente.

De acordo com os conceitos de *Customer Relationship Management* (CRM), quanto menor a taxa de evasão de alunos, mais valiosa é a base de clientes. Isto se deve pela obtenção de um maior *lifetime value* para a instituição. Desta forma, é importante compreender as causas das evasões para que as instituições possam agir de forma preventiva na manutenção dos alunos em suas bases. Embora haja estudos no Brasil que examinam os motivos da evasão de alunos no ensino superior (ADACHI 2009, LOBO 2012, PRIM e FÁVERO 2013, GERBA 2014), estas pesquisas tratam o tema de forma qualitativa ou quantitativa, porém sem muito aprofundamento estatístico devido as restrições de acesso a dados. A importância deste estudo deve-se ao acesso exclusivo aos dados reais de uma IES. Sendo assim, esse trabalho tem como objetivo propor um modelo, a partir de dados acadêmicos e financeiros, capaz de estimar o risco de evasão de alunos de uma IES privada do Rio de Janeiro posicionada para a classe C. Para compor a amostra de alunos, foram selecionados os indivíduos matriculados em cursos de graduação da IES a partir de 2010. Para isso, o estudo se propõe a, por meio de técnicas de *machine learning* e de informações disponíveis nos bancos de dados da IES, desenvolver um modelo preditivo para a evasão de alunos. Este modelo permitirá que a IES tenha uma ferramenta para atuar de forma preventiva na retenção de seus alunos. Para a amostra de alunos

ativos (etapa de *scoring*), os modelos estimados apresentaram como resultados a probabilidade de evasão desses alunos.

REVISÃO DE LITERATURA

CRM (*Customer Relationship Management*)

O objetivo primordial do CRM é elevar o valor da base de clientes da empresa por meio da fidelidade de seus clientes. Esta fidelidade pode ser alcançada, principalmente, por boas experiências de relacionamento, ocasionando uma maior satisfação dos clientes. A manutenção deles na base por um prazo longo está associada ao conceito de lealdade. A consequência de uma boa gestão de relacionamento com o cliente é a lealdade deste com a instituição. Há evidências empíricas em estudos acadêmicos (PEPPERS & ROGERS GROUP, 2000) da relação entre o nível de utilização de ferramentas de CRM com a lealdade de longo prazo do cliente com a empresa. O foco deste trabalho está na retenção dos alunos na base.

Marketing e retenção no setor de educação superior

Vincent Tinto apresentou em 1975 um modelo para compreender a evasão no ensino superior que se tornou referência na área. Este modelo foi revisitado pelo autor em 1987 e 1993. Para analisar a evasão, Tinto recorre a elementos de psicologia e baseia-se na teoria do suicídio de Durkheim (1961) – para ele, o suicídio é mais favorável de ocorrer em indivíduos não muito integrados à sociedade. Tinto trata o campo acadêmico como um sistema social com seus próprios valores e estruturas, compara o abandono da instituição de ensino com o suicídio para Durkheim. Para Tinto, é preciso distinguir os tipos de evasão, como decisões espontâneas ou falha no desempenho acadêmico, pois para cada um deles, há envolvimentos de diferentes tipos de estudante e padrões de relacionamento com a instituição.

No modelo de Tinto, há dois conceitos muito importantes, a integração e o comprometimento. Nesse modelo, os fatores antecedentes ao ingresso na instituição, que podem ser o histórico familiar, atributos individuais ou a educação escolar, moldam o comprometimento do aluno com suas próprias metas e à instituição. O grau de comprometimento com suas metas influencia a integração ao sistema acadêmico, enquanto o comprometimento com a instituição influencia a integração ao sistema social. De forma eventual, caso um dos dois graus de comprometimento se mostrar muito baixo, o aluno tende a abandonar o curso. O comprometimento do aluno com seus objetivos e com a instituição está associado diretamente à lealdade.

Lifetime Value e *Customer Equity*

Lifetime value é uma medida usada para obter o valor de uma base de clientes (BERGER; NASR, 1998; JAIN; SINGH, 2002). O seu cálculo consiste, usualmente, em traçar o lucro por cliente – totalidade das receitas menos totalidade dos custos diretamente vinculados ao atendimento daquele cliente – num determinado intervalo de tempo e num determinado tempo de relacionamento da empresa com o cliente. Os lucros futuros são descontados a valor presente por uma taxa de juros que reflita o custo de oportunidade da empresa de forma análoga à apuração do valor presente líquido (VPL). O *lifetime value* de um aluno na graduação é a margem de contribuição desse aluno pelo prazo esperado até sua formatura, trazida a valor presente, multiplicada pela probabilidade de conclusão. O somatório dos *lifetime values* corresponde o *customer equity*, que representa o valor da base de clientes para a empresa. A fórmula do *lifetime value* será mostrada na seção de Metodologia.

Machine Learning

O objetivo de técnicas de *machine learning* é aprender continuamente com novos dados para melhorar a previsão de uma determinada variável. Usando estas técnicas, o pesquisador irá incluir variáveis que são suspeitas de causar impacto na variável alvo para gerar o modelo. No caso do problema de evasão de alunos no setor de educação, encontra-se nas técnicas de *machine learning* uma oportunidade eficaz de aplicação. Para este trabalho, o foco da sua aplicação foi na obtenção da diferenciação de clientes (por nível de risco de evasão). Os dados dos alunos, tais como, desempenho acadêmico, informações demográficas e financeiras podem indicar padrões associados ao abandono da instituição.

METODOLOGIA

Este estudo foi feito com dados reais obtidos de uma IES situada no Rio de Janeiro. Foi adotada uma abordagem quantitativa, aproveitando-se do acesso aos dados dos sistemas financeiro e acadêmico da instituição. A pesquisa foi elaborada em três etapas. Na primeira, foi aplicada uma abordagem quantitativa de análise, utilizando modelos de *machine learning* de classificação em uma base de dados secundários para encontrar modelos que expliquem a evasão de alunos. Na segunda, uma amostra de alunos ativos foi submetida aos modelos encontrados para classificação da probabilidade de evasão destes alunos. Para finalizar, foi calculado o *lifetime value* desses alunos e o *customer equity* da base. A amostra contém os alunos da IES que ingressaram de 2010 até 2016.

Para compor o trabalho, a amostra foi dividida em três grupos de alunos: aqueles (i) que concluíram seu curso, (ii) que evadiram e (iii) que estavam ativos no instante da coleta dos dados. Os grupos (i) e (ii) compreendem a amostra para a qual os modelos de *machine learning* foram calculados na primeira etapa, enquanto o grupo (iii) com os alunos não classificados foram utilizados na segunda etapa (*scoring* da base). Adotou-se o software WEKA (HALL, 2009; AHER; LOBO, 2011; SHARMA; JAIN, 2013) para desenvolver modelos de *machine learning*, empregando cinco tipos de algoritmos para classificação. Após a construção dos modelos na primeira etapa, estes foram rodados para classificação do grupo (iii), de alunos ativos, baseado em seus atributos. Na terceira etapa, foi calculado o *lifetime value* de cada aluno ativo estimado conforme as probabilidades geradas na segunda etapa. Uma limitação desta metodologia é que os resultados dos modelos não são capazes de indicar quando o aluno evadirá, informando somente a probabilidade de evasão.

Amostra e Procedimentos de Coleta de Dados

Foram considerados os alunos de 15 cursos de bacharelado, licenciatura e tecnólogos, com durações de 4 a 10 semestres, totalizando a amostra inicial de 4040 alunos. Foi feito um corte a partir do primeiro semestre de 2010 devido à dificuldade de obter dados anteriores a este período no sistema da IES. A base final usada nos modelos contém informações extraídas diretamente da base de dados da instituição. A lista de atributos foi: matrícula, forma de ingresso na instituição, indicação se aluno possui FIES, distância entre a residência e universidade, curso, gênero, turno, tipo de colégio, número de anos entre a entrada na universidade e a saída na escola, idade, médias das notas no primeiro e segundo semestres, indicador de aumento de média do primeiro para o segundo semestre, descontos médios obtidos no primeiro e segundo semestres, redução de descontos, percentuais de boletos atrasados, em negociação e com isenções e status (evade ou conclui – a variável alvo).

Baseado nos tipos de status encontrados no sistema acadêmico da IES, foram classificados os alunos como concluinte ou evadido. Foi definido no critério de seleção da amostra que seriam considerados apenas os alunos que cursaram ao menos dois períodos na IES antes de evadir. Desta forma, a amostra final foi composta por 4.078 alunos. A adoção

deste critério deve-se a pouca informação produzida por um aluno que cursou apenas um semestre e isto pode prejudicar a capacidade de aprendizagem e predição de algoritmos de *machine learning*. É importante destacar que é observado em estudos anteriores que a evasão ocorre mais no primeiro ano do aluno na instituição.

Para criar e treinar modelos, usa-se uma base de dados de alunos com a classe conhecida para que os algoritmos calculem e modelem as associações entre os atributos e a classe. Para fazer a previsão, usa-se o modelo criado para calcular a classe de cada cliente em uma case na qual tal classe é desconhecida, usando os mesmos atributos utilizados na base que originou o modelo. Para cada instância, é gerado o *score* (classificação) como resultado do modelo. O critério adotado para a amostra de alunos ativos, foi selecionar aqueles que ingressaram na instituição em 2015.1 e 2015.2. Foram selecionados 1.502 alunos que foram classificados pelos modelos, para cada instância, sendo calculadas suas probabilidades de evasão e seus *lifetime values.*

Machine Learning

Este trabalho utilizou o WEKA, o software mais utilizado em processos de *machine learning* para criar, testar e comparar modelos preditivos (HALL, 2009). O estudo trabalha na classificação pelo reconhecimento de um aluno como integrante de um dos grupos definidos a priori, "evade" ou "concluiu" (classificação supervisionada). O trabalho foca em dois algoritmos para classificação, regressão logística (WILSON; LORENZ, 2015; PENG; LEE; INGERSOLL, 2002) e árvore de decisão. Além desses dois, outros algoritmos foram rodados e comparados. São eles: J48 – outro tipo de árvore de decisão, conhecido como C4.5 (QUINLAN, 1996; KOTSIANTIS, 2007), *K-Nearest Neighbors* (ALTMAN, 1992), *Support Vector Machines* (CORTES & VAPNIK; FRADKIN & MUCHNIK, 2006) e *Naive Bayes* (HAND; YU, 2001).

Lifetime Value

Para chegar ao *lifetime value* do aluno, aplicou-se a equação a seguir (MAHISHI, 2014): $LV_x = p_x \cdot \sum_{i=1}^{n} \frac{M_x . mgC}{(1+r)^i}$ onde **LV** representa o *lifetime value* do aluno **x**; **p** a probabilidade estimada de o aluno **x** concluir o curso; **n** o número de meses até a sua formatura, **M** a mensalidade média do aluno **x**; **mgC** a margem de contribuição média da IES e **r** a taxa de desconto do custo de oportunidade para a IES. Para chegar a **n**, foi calculado o número de semestres até a conclusão do curso multiplicado por 6, obtendo-se o número de meses restantes esperado do aluno na IES, não considerando atrasos na formação. Foi considerado também a mensalidade média paga pelo aluno, utilizando a mensalidade média do último semestre disponível, 2016.1. A margem de contribuição (58%) e taxa de desconto (15% a.a. nominal) usadas foram obtidas pela direção da instituição. Este estudo escolheu por usar a margem de contribuição para o custo variável de servir o aluno. O custo docente é o mais importante dentre os custos variáveis. Uma possível deficiência desse critério é que o custo docente varia em função do número de turmas e não diretamente do número de alunos. No entanto, o número de turmas é função direta do número de alunos. As probabilidades **p** de conclusão foram extraídas dos modelos de *machine learning* adotados. Ao final, foram somados o *lifetime value* dos alunos para obter o *customer equity value* da base de alunos ativos.

RESULTADOS

São apresentados para comparação os resultados dos diversos algoritmos gerados pelo WEKA (Tabela 1).

Tabela 1 – Resumo dos resultados de desempenho dos modelos

	Logit	REPTree	J48	IBk	SMO	N.Bayes	Média
Acurácia	**76.88%**	**76.83%**	**76.24%**	**70.47%**	**76.66%**	**73.79%**	**75.14%**
Sensibilidade	53.10%	52.44%	56.52%	50.84%	48.73%	38.38%	50.00%
Especificidade	89.13%	89.39%	86.39%	80.58%	91.04%	92.01%	88.09%
Precisão	71.54%	71.78%	68.13%	57.40%	73.68%	71.22%	68.96%
Evade/total	25.22%	24.83%	28.19%	30.10%	22.48%	18.32%	24.86%

A regressão logística (logit) e a árvore de decisão (reptree) alcançaram os índices mais altos de acurácias dentre os algoritmos testados. Os métodos *support vector machines* (SMO) e J48 (árvore de decisão J48) também tiveram bom desempenho, muito próximos dos dois primeiros. Já o algoritmo SMO, em especial, apresentou o melhor desempenho na capacidade de prever a classe de evasão, com precisão de 73,7%, opondo-se com o modelo o algoritmo *k-nearest neighbors* (IBk) que apresentou o pior desempenho, com precisão de 57,4%. As classes são binárias (0 – negativo que representa conclui e 1 – positivo que representa evade). A precisão é a taxa de acertos quando a previsão é positiva. A sensibilidade conhecida também como *recall* ou taxa de verdadeiro positivo, apura o percentual de previsões de positivos, quando o real é positivo. A especificidade ou taxa de verdadeiro negativo mensura o percentual de previsões de negativo quando o real é negativo. As taxas altas de especificidade mostram um risco baixo de incorrer no erro tipo I (alarme falso), no entanto o risco de cometer o erro tipo II (não identificação) é elevado. Pelo resultado da métrica sensibilidade, o modelo não apresentou bom desempenho, pois em torno de 50% de alunos que evadiram não foram corretamente classificados na previsão.

O método da regressão logística, originado no início do século XIX (CRAMER, 2002), é um dos mais usados para classificação binária (WILSON; LORENZ, 2015). Este método desperta o interesse dos pesquisadores pela sua capacidade de identificar as chances de ocorrência de um evento binário (a classe) pela influência de um atributo. Estas chances calculadas são nomeadas de *odds ratio*. Para ilustrar como elas são calculadas, será apresentado um exemplo a seguir. De cada 10 alunos evadidos, 7 são homens, logo a probabilidade de evasão de um homem é de 0,7 e sua complementar é de 0,3. A *Odds ratio* é a razão 0,7/0,3 = 2,33. Pode-se expressar, neste exemplo, que homens têm 2,33 vezes mais chances de evadirem do que mulheres. Quando uma *odds ratio* de um determinado atributo está entre 0 e 1, isto significará que a presença deste atributo está associada a uma menor chance de ocorrência da variável dependente. Contudo, quando a razão é igual ou próxima de 1, isto indicará que o atributo não tem influência sobre a classe. Já para valores maiores do que 1, isto significará que o atributo aumenta as chances de ocorrência da classe e, quanto maior o valor, maior a chance. A interpretação dos resultados da *odds ratio* estimada dos atributos deve ser cautelosa. Os valores não têm significado real, sendo assim, eles somente podem ser utilizados para explicar a amostra analisada. Mesmo assim, de posse dos resultados apurados, é possível inferir algumas informações, ainda que com algum grau de incerteza.

A árvore de decisão, o outro algoritmo usado no estudo, classifica os dados com base nas regras (nós) que os dividem em ramos (setas que saem dos nós e representam os resultados). Essas divisões são também chamadas de *splits*. A partir de cada ramo, outras regras podem ser aplicadas, criando nós que geram novos *splits* e assim, sucessivamente. O primeiro nó é denominado raiz e o último, que não tem mais subdivisões, é chamado de folha. Já os nós intermediários, recebem o nome de nós de decisão. Uma vez construída a estrutura da árvore, novos dados não classificados seguem o caminho dos ramos conforme as regras definidas em cada nó. A classe do dado é identificada quando este chega na folha. Neste estudo, o tamanho total da árvore foi calculado em 302 nós e foram encontradas folhas desde do nível 3 até 11. Cabe ressaltar que, embora este método seja capaz de explicar bem a amostra na qual foi

construído, ele pode não ter um bom poder preditivo, já que alguns dos nós criados não se relacionam com fenômenos reais, sendo apenas consequências de particularidades da amostra em que foram baseados. As variáveis mais determinantes para a classificação da instância aparecem nos primeiros *splits* conforme as regras do algoritmo, sendo assim, é importante concentrar as análises nestes níveis iniciais. O nó raiz representa a variável média das notas do segundo período e o primeiro *split* dividiu o nó em dois ramos, média maior ou menor do que 3,9 (numa escala de 0 a 10). Os nós seguintes, para os dois ramos, representam o tipo de ingresso na instituição. No terceiro nível, temos os nós que dividem os alunos nos cursos.

Após rodados os modelos, estes foram utilizados para prever o *score* da base de alunos ativos. A tabela 2 compara as probabilidades de evasão e *customer equity* com base nos diversos métodos.

Tabela 2 – Probabilidade de evasão média e *customer equity* por modelo

	Logit	REPTree	J48	IBk	SMO	N.Bayes	Média
Prob.Evasão Média	41.4%	47.8%	51.0%	44.4%	30.5%	31.0%	**41.0%**
Customer Equity (R$ M)	7.14	6.35	4.86	5.65	6.97	6.80	**6.29**

DISCUSSÃO

No resultado apresentado pela Regressão Logística, a sensibilidade ficou em torno de 50%, isto indica que quase metade dos alunos que evadiram não foram corretamente classificados. Desta forma, fica difícil identificar o aluno com risco de evasão somente utilizando os atributos deste estudo. Este obstáculo pode estar associado ao fato de que determinados atributos presentes dão indícios fortes de evasão (Exemplo: Um CR muito baixo, porém, a sua ausência não garante que o aluno irá concluir o curso). É possível que outros motivos não capturados pelos atributos usados no estudo levem um aluno a evadir, apesar do aluno não apresentar CR baixo. Outro resultado significativo da Regressão Logística, foi obtido do atributo "melhorou" (*odds ratios*: 0,62) que representa o aumento da média das notas no primeiro para o segundo período. Este *odds ratio* indica que há uma chance menor de evasão e pode estar associado a um maior nível de comprometimento do aluno ou maior nível de integração acadêmica, conforme observado por Tinto (1975). Por outro lado, a redução do desconto médio (*odds ratios*:1,59) tem um efeito contrário de mesma magnitude: o fim de um benefício econômico pode motivar a saída de alunos da instituição. Vale destacar que o nível de desconto no segundo período é muito pouco relevante (*odds ratio* próxima de 1) quando comparado ao desconto do primeiro período, porém este desconto inicial tem relação direta com a evasão (*odds ratio*: 3,97). Há uma correlação positiva entre alunos que entram com descontos maiores e evadem. Ao indicar que os maiores níveis de desconto na entrada elevam as chances de evasão, o modelo alerta para um possível erro no processo seletivo de alunos. Alguns descontos agressivos são dados para aumentar a captação de alunos e, possivelmente, a instituição pode estar captando alguns alunos interessados apenas na mensalidade barata e menos interessados pelo curso ou instituição. Os alunos do FIES também aparecem com *odds ratio* maior que 1, indicando maior chance de evasão, porém pode ser um viés da amostra: isto porque, somente a partir de 2013, o FIES tornou-se importante para a instituição. Desta forma, não há tantos alunos formados no período, logo há uma tendência em identificá-los no perfil de abandono. A forma de ingresso na IES também pode ter uma boa influência: alunos transferidos de outras instituições tem menor chance (*odds ratio*: 0,47) de abandonarem o curso. Isto pode estar relacionado à uma maior motivação do aluno em fazer o mesmo curso nesta instituição. Já com o aluno de transferência interna (*odds ratio*: 2,39), ocorre o contrário, a chance de evasão se eleva. Neste caso, ele troca de curso, porém permanece na IES. É possível que esse aluno tenha mais dúvida quanto a sua trajetória acadêmica, mesmo permanecendo na instituição, que

poderia demonstrar um certo vínculo com a IES. Os alunos do PROUNI aparecem com maior chance de evasão, indicando serem mais propensos a evadir. Foi observado que alunos vindos de escolas públicas aparecem com chance mais baixa de evasão. Os atributos "distância em km da residência a instituição" e o "número de anos desde que saiu da escola" parecem não ter relação com a classe investigada. Eles apresentaram valores de *odds ratio* muito próximos a 1, indicando que eles são insignificantes para explicar a classe nesta amostra.

Pelos resultados apresentados pela Árvore de Decisão, observa-se também que a média do segundo período, a forma de ingresso e o curso são atributos importantes na determinação da classe do aluno, corroborando os resultados encontrados no modelo construído via regressão logística. A partir do *split* dos alunos com média menor do que 3,9, observa-se que a quantidade de nós a partir desse *split* é bem menor. De maneira geral, 83% dos alunos que obtiveram uma média menor do que 3,9 no segundo período evadiram. Analisando o lado esquerdo da árvore, o nó subsequente à média do segundo semestre é a forma de ingresso, onde o modelo retorna evasão para todas as modalidades (são todos folhas) exceto para o vestibular (nó de decisão) – caso essa seja a modalidade de entrada, ele vai para um terceiro *split*, de acordo com o curso. Do lado direito da árvore, após o *split* inicial da nota do segundo período maior ou igual a 3,9, novamente aparece no segundo nível o critério de ingresso, porém diferentemente do lado esquerdo, deste lado as formas de ingresso não são folhas, mas sim nós de decisão, dependentes dos cursos. As probabilidades de evasão aqui variam de 50% (ENEM) a 14% (portador de diploma). Foi observado que alunos que ingressaram, independente do curso, via transferência externa ou já são portadores de diploma tem chance mais baixa de evasão.

Em relação ao *lifetime value* e *customer equity*, normalmente, quanto maior a taxa de conclusão média (probabilidade complementar a da evasão) indicada pelo modelo, maior seu *lifetime value*. Contudo, ao comparar os resultados dos diversos modelos apresentados, observa-se que as taxas de conclusão são diferentes, porém com valores próximos. Pela previsão da regressão logística, em torno de 60% dos alunos da amostra concluirão seu curso, 10% a menos que nos modelos de *naive bayes* e *support vector machine*. No entanto o *lifetime value* do primeiro é o mais alto de todos. A diferença está no valor dos alunos classificados como futuros concluintes. A regressão logística classificou, na média, mais alunos valiosos como concluintes do que os demais modelos. A média dos resultados dos modelos indica 41% de probabilidade de evasão. O *customer equity* da amostra indica um valor de aproximadamente R$ 6,3 milhões. Todavia, esse resultado pode ser otimista, pois considerando os resultados analisados observou-se que os modelos foram sistematicamente conservadores em suas capacidades de prever a evasão: muitos casos de alunos que evadiram não foram previstos.

CONCLUSÕES

As previsões apresentadas neste trabalho permitiram obter um *ranking* dos alunos em risco de evasão e seus *lifetime values*. De posse destas informações, a IES pode elaborar uma estratégia para decidir onde priorizar seus esforços de retenção. Vale ressaltar que a robustez dos resultados está diretamente associada ao tamanho da amostra. Considerando as deficiências encontradas na obtenção de dados históricos nos sistemas da IES e do próprio tamanho desta instituição, o estudo selecionou o maior número factível de alunos e seus atributos mais relevantes para desenvolver os modelos. É importante que os modelos sejam continuamente aprimorados pela inclusão de informações a cada semestre. É esperado, neste tipo de metodologia, que os modelos fiquem mais robustos pela continua realimentação.

De todos os atributos estudados, a média de notas no segundo período foi considerada a variável principal do desempenho acadêmico, um dos fatores mais fundamentais para determinar a futura evasão do aluno. Este resultado era esperado e corrobora a teoria do

abandono estudantil, de Tinto (1993), que tem na ausência de comprometimento acadêmico uma das principais causas para a evasão tanto quanto a ausência de integração social.

A forma de ingresso na instituição também é um fator relevante na evasão. Alunos que ingressaram pelo ENEM possuem uma maior chance de evasão do que as outras formas. É possível que este achado também tenha repercussão na teoria da integração de Tinto. Pelo ENEM, o aluno não está vinculado a nenhuma instituição em particular, ele tentará ingressar numa IES de acordo com a sua nota. Ao prestar o vestibular específico para a IES, ele já demonstra seu interesse na instituição, podendo indicar uma intenção de integração social na instituição. Outro achado que reforça a teoria de Tinto é que alunos vindos de transferência externa tem chance bem mais baixa de evasão. Nesta forma de entrada, o aluno está mudando de instituição, mantendo o curso, o que demonstra uma atitude de comprometimento acadêmico e interesse pela instituição. De maneira semelhante, os alunos que já possuem diploma também demonstram terem um grau de comprometimento acadêmico maior do que o observado nos alunos oriundos do ENEM.

Em outro exemplo de aderência dos resultados à teoria de Tinto, a transferência interna aparece com chance alta de evasão: neste caso o aluno muda o curso, mas opta por se manter na instituição demonstrando que há integração social do aluno com a IES, ainda que não tenha havido a devida integração acadêmica – razão para o abandono do primeiro curso. Segundo Tinto, uma excessiva integração social com baixa integração acadêmica também favorece a evasão, caso em que um aluno prioriza atividades sociais em detrimento dos estudos. Enquadram-se neste caso alunos que, embora não se identifiquem com o curso, tenham amigos e uma interação social intensa na instituição e podem ter dificuldades em abandoná-la num primeiro momento. Assim, tentam outro curso na mesma instituição, resultando em uma evasão posteriormente.

Outro fato com possível repercussão gerencial são os descontos dados na mensalidade. Na regressão logística, foi observado que quanto maior o desconto dado no semestre inicial, maior a propensão à evasão. É possível que na ânsia de captar mais alunos, a IES pode estar dando descontos atrativos aos alunos sem que estes estejam realmente interessados em seus cursos. Por outro lado, há também um aumento de chance de evasão quando há uma redução do nível de desconto do primeiro para o segundo período. Isto pode indicar que seria recomendável a manutenção do desconto num mesmo patamar. Porquanto, a sua redução implica em efeito negativo na probabilidade de retenção.

LIMITAÇÕES E PESQUISAS FUTURAS

Uma limitação deste estudo deve-se ao número de instituições estudadas, no caso apenas uma. Para esse tipo de análise, seria interessante obter dados de várias IES para que os fatores específicos de cada uma delas pudessem ser, devidamente, comparados. A impossibilidade de se prever o momento em que a evasão ocorrerá é uma outra limitação do estudo decorrente desta metodologia.

Uma oportunidade de novos estudos seria aplicar esta metodologia de *machine learning* para marketing de relacionamento em cursos de ensino a distância (EaD). É possível neste ambiente virtual identificar os caminhos seguidos pelos alunos, possibilitando sua medição e análise.

A metodologia de *machine learning* aqui utilizada representa uma ferramenta absolutamente escalável que, na prática, funciona melhor quanto mais instâncias existem, e que tem como *output* um identificador único para cada aluno baseado nos seus próprios dados. Portanto, este segmento de ensino superior representa uma oportunidade excelente para aplicação de metodologias de *machine learning*, não apenas para classificação de evasão, mas também podendo ser exploradas em outras questões do negócio.

REFERÊNCIAS

ADACHI, A. A. C. T. **Evasão e evadidos nos cursos de graduação da Universidade Federal de Minas Gerais**. 30 jan. 2009. 214 f. Dissertação – Universidade Federal de Minas Gerais. Belo Horizonte/MG, 30 jan. 2009.

AHER, S. B.; LOBO, L. M. R. J**. Data mining in educational system using Weka**. IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT), p. 20-25, 2011.

ALTMAN, N. S. **An introduction to kernel and nearest-neighbor nonparametric regression**. The American Statistician, 46.3, 175-185, 1992.

BERGER, Paul D.; NASR, N. I. **Customer lifetime value: Marketing models and applications**. Journal of interactive marketing, v. 12, n. 1, p. 17-30, 1998.

CORTES, C., VAPNIK, V. **Support-vector networks**. Machine learning, 20.3, p. 273-297, 1995.

CRAMER, J. S. **The origins of logistic regression**. Tinbergen Institute Discussion Paper, 119/4, 2002.

DURKHEIM, E. **Suicide**. J. Spaulding & G. Simpson, trans. Glencoe: The Free Press, 1961.

FRADKIN, D.; MUCHNIK, I. **Support vector machines for classification**. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 70, p. 13-20, 2006.

GERBA, R. T. **Análise da evasão de alunos nos cursos de Licenciatura: Estudo de caso no Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina.** 17 set. 2014. 157 f. Dissertação – Universidade Federal de Santa Catarina. Florienópolis/SC, 17 set. 2014.

HALL, M., et al. **The WEKA data mining software: an update**. ACM SIGKDD explorations newsletter, 11.1, p. 10-18, 2009.

HAND, D J.; YU, K**. Idiot's Bayes – not so stupid after all?** International statistical review, 69.3, p. 385-398, 2001.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior**. Brasília: INEP, 2016. Disponível em: http://portal.inep.gov.br/web/censo-da-educacao-superior. Acesso em: 18 nov. 2015.

JAIN, D.; SINGH, S. S. **Customer lifetime value research in marketing**: A review and future directions. Journal of interactive marketing, 16.2, p. 34-46, 2002.

KOTSIANTIS, S. B. **Supervised machine learning: A review of classification techniques**. Informatica, 31.3, p. 249-269, 2007.

LEE, W.; STOLFO, S. J.; MOK, K. W. **A data mining framework for building intrusion detection models**. Security and Privacy. Proceedings of the 1999 IEEE Symposium on. IEEE, p. 120-132, 1999.

LOBO, M. B. C. M. **Panorama da evasão no ensino superior brasileiro:** aspectos gerais das causas e soluções. ABMES Cadernos. Brasília, set./dez, 2012.

MAHISHI, A. **Customer Lifetime Value** – Not Just a Marketing Metric. Tata Consultancy Services, 2014. Disponível em: http://www.tcs.com/SiteCollectionDocuments/White-Papers/Customer-Lifetime-Value-Not-Just-a-Marketing-Metric-1214-1.pdf. Acesso em: 1 mar. 2017.

PENG, C. Y. J.; LEE, K. L.; INGERSOLL, G. M. **An introduction to logistic regression analysis and reporting**. The journal of educational research, 96.1, p. 3-14, 2002.

PEPPERS AND ROGERS GROUP. **Roper Starch Worldwide survey**. Setembro, 2000.

PEREIRA FILHO, E. **Compromisso com o graduar-se, com a instituição e com o curso**: estrutura fatorial e relação com a evasão. 89 f. Dissertação (Mestrado em Educação) – Universidade Estadual de Campinas, São Paulo, 2012.

PRIM, A. L.; FÁVERO, J. D. **Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau**. Revista E-Tech: Tecnologias para Competitividade Industrial, Florianópolis, n. Especial Educação, p. 53-72, 2013/2.
QUINLAN, J. R**. Improved use of continuous attributes in C4.5**. Journal of artificial intelligence research, 4, p. 77-90, 1996.
SHARMA, T. C.; JAIN, M. **WEKA approach for comparative study of classification algorithm**. International Journal of Advanced Research in Computer and Communication Engineering, 2.4, p. 1925-1931, 2013.
TINTO, V. **Dropout from higher education: a theoretical synthesis of recent research**. Review of Educational Research, 45, 89-125, 1975.
TINTO, V. **Leaving college**: Rethinking the causes and cures of student attrition. The University of Chicago Press, Chicago, 1987.
TINTO, V. **Leaving college**: Rethinking the causes and cures of student attrition. 2 ed.
WILSON, J. R.; LORENZ, K. A. **Short History of the Logistic Regression Model**. In: Modeling Binary Correlated Responses using SAS, SPSS and R, p. 17-23, Springer International Publishing, 2015.