Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar

Paulo J. L. Adeodato[1,2], Maílson M. Santos Filho[1], Rodrigo L. Rodrigues[1]

[1]Centro de Informática da Universidade Federal de Pernambuco

[2]NeuroTech Tecnologia da Informação S.A.

pjla,mmsf,rlr{@cin.ufpe.br},paulo@neurotech.com.br

Abstract. *This paper presents a data mining solution for assessing the quality of private Brazilian secondary schools based on the official school survey and students tests. These two data sets have been transformed to the school grain for embedding data and expert´s knowledge and been integrated in a single data set made compatible with the AI techniques applied. Logistic regression produced a propensity score, decision tree produced the sequential decision making a human would follow and rules were induced for supporting the decision based on the score. The AUC_ROC and Max_KS metrics assessed the propensity score performance and coverage, confidence and lift assessed the quality of the rules induced. The results showed that this domain-driven data mining approach was very successful in modeling the problem and validating public policies.*

Resumo. *Este artigo traz uma solução de mineração de dados para avaliar a qualidade da educação secundária privada no Brasil, a partir das bases do ENEM e do Censo Escolar. Essas bases foram transformadas para o grão da decisão embutindo o conhecimento dos dados e especialistas, integradas e os formatos compatibilizados com cada técnica de IA. Regressão logística gerou o escore de propensão ao sucesso, árvore de decisão expôs a decisão seqüencial humana ideal e a indução gerou as regras para apoiar a decisão baseada no escore. As métricas AUC_ROC e Max_KS avaliaram o desempenho do escore de propensão e, cobertura, confiança e lift, mediram a qualidade das regras. Os resultados mostraram que essa abordagem (domain-driven data mining) teve muito sucesso na resolução do problema e na validação de políticas públicas.*

1. Introdução

No Brasil, o INEP faz o Censo Escolar das escolas da educação básica e o Exame Nacional do Ensino Médio (ENEM) que avalia o desempenho do estudante ao fim da educação básica. Em 2009, o ENEM passou a ser precisa fonte de informações sobre a educação secundária do país. Além do conhecimento técnico do aluno, o ENEM captura informações sócio-econômico-culturais sobre o seu perfil (INEP, 2012). A integração das informações do Censo Escolar com as do ENEM possibilitam ao Governo Federal definir e validar as políticas públicas para a educação nacional (TRAVITZKI, 2013). O ENEM também é usado para gerar um *ranking* anual das escolas a partir do desempenho médio dos seus alunos nas provas das diversas áreas do conhecimento.

Porém, pouco conhecimento sistemático é extraído dessas bases, pelo seu grande volume de dados (>10Gb/ano) e por estarem em grãos distintos (escolas, alunos e professores das escolas). A mineração de dados é um processo computacional (inteligência artificial, estatística e banco de dados) capaz de extrair conhecimento dos dados (FAYYAD, 1996) para apoiar decisões. Este artigo aplica a metodologia CRISP-DM (***CR****oss* ***I****ndustry* ***S****tandard* ***P****rocess for* ***D****ata* ***M****ining*) (WIRTH, 2000) para extrair conhecimento sobre a educação do ensino médio do Brasil, desses dados.

2. Entendimento dos dados

Para este trabalho, foram utilizados os Microdados do ENEM 2011 que contém informações sócio-econômico-culturais e de desempenho dos alunos ao final do ensino secundário, e o Censo Escolar 2011 que detalha as condições das escolas secundárias, da infra-estrutura ao corpo docente. Este trabalho focará apenas as escolas privadas.

Somente as escolas privadas foram selecionadas e as bases do ENEM e do Censo foram integradas pelo código único de identificação da escola. Anomalias que pudessem tanto desviar o foco quanto deteriorar a qualidade foram eliminadas. Os alunos que não realizaram todas as provas e os não estavam concluindo o ensino médio foram descartados, bem como os atributos somente de escolas públicas (*e.g.* EJA). No grão escola, o trabalho preservou apenas as escolas privadas com 15 ou mais alunos para haver massa estatisticamente válida para as análises.

Para sistematizar este estudo, o objetivo foi definido como "binário". Nesse processo há dois pontos polêmicos: 1) que métrica seria utilizada para avaliar a qualidade da escola e 2) que limiar seria adotado como critério para definir o que seria uma escola "boa". Mantendo o pragmatismo e isenção, este trabalho considerou a média aritmética das notas dos alunos como medida de qualidade das escolas, similar ao feito no *ranking* das escolas. A nota de cada aluno é a média aritmética das suas notas em cada prova.

Em seguida, definiu-se o quartil superior como limiar de binarização da nota para caracterizar o objetivo (escola *forte* ou *fraca*). Essa abordagem estatística evita polêmica e usa a robustez das separatrizes (quartis *etc*.) em relação a valores extremos (*outliers*), como tem sido feito (SOUSA *et al*., 2008) em outros domínios de aplicação. Assim, escola forte é aquela cuja média esteja no quartil superior das médias das escolas.

A diferença na granularidade dos atributos aumenta a complexidade do trabalho. 1) Como atribuir uma *renda familiar* à escola a partir da *renda familiar* dos seus alunos? 2) Como atribuir uma *formação docente* à escola a partir da *formação docente* dos seus alunos? Esse aspecto dificulta equipes sem profissionais de mineração de dados desenvolverem projetos, tanto pelo grande volume de dados quanto pela necessidade de uso de inteligência artificial para embutir conhecimento dos especialistas em educação nas transformações dos atributos para mudança de grão, no processo *Domain-Driven Data Mining* (D^3M) (CAO, 2008). Para evitar explosão combinatória, este trabalho adotou a média e a moda para conjugar no grão escola os atributos numéricos e categóricos, respectivamente. Assim, as respostas às perguntas acima passam a ser 1) a média da *renda familiar* dos alunos considerados e 2) a moda da *formação docente* dos professores considerados em cada escola.

Em geral, dados ausentes foram preenchidos pela média nos atributos numéricos e, nos categóricos, foi criada a categoria *dado ausente*, exceto em casos dependentes do domínio, em que havia relação de ordem, quando era preenchido por uma categoria extrema (*e.g*. educação dos pais ausente foi codificado como analfabeto). Nos numéricos, os *outliers* foram substituídos pelo extremo de 3 desvios-padrão.

3. Extração do conhecimento e resultados

A extração do conhecimento consistiu de três funcionalidades básicas voltadas para a construção de um Sistema de Suporte à Decisão (DSS - *Decision Support System*): um estimador de propensão à escola ser forte, uma árvore de decisão para explicar como

seria a seqüência decisória ideal na visão de um especialista humano e uma base de regras para explicar as decisões e identificar nichos de alta relevância no domínio.

Regressão logística foi a técnica escolhida para gerar a pontuação de propensão pelo seu sucesso em problemas de classificação binária, em diversos domínios [REF]. Além da pontuação, a técnica identifica e quantifica os principais atributos que têm influência sobre a variável-alvo binária cujo comportamento pode ser expresso pela equação abaixo, com base no conjunto de variáveis independentes ou explicativas (atributos):

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

onde π é a probabilidade de um evento e β é o coeficiente que dá a influência do atributo. A técnica *forwardstepwise* selecionou os atributos de maior discriminação no modelo pelo critério de máxima verossimilhança (*LikelihoodRatio* – LR) (Hair 2009). A Tabela 1 mostra o parâmetro β que mede a influência do atributo e seu *p-value*, para 10 dos atributos mais influentes (dentre 21) ao nível de significância 5% para o modelo.

TABELA I

DEZ VARIÁVEIS MAIS INFLUENTES NO MODELO DE REGRESSÃO LOGÍSTICA PELO MÉTODO *FORWARDSTEPWISE*

Variável	**Coeficiente β**	***p-value***	***Natureza***	***Fonte***
Escola situada na região Sudeste	1,11	0,000	Categórica	Escola
Q75 (Quantidade de banheiros em casa)	0,83	0,000	Moda	Aluno
Escola tem pelo menos 1 doutor	0,62	0,000	Transformada	Professor
Escola situada na região Sul	0,48	0,033	Categórica	Escola
Escola tem laboratório de ciências	0,36	0,003	Binária	Escola
Etnia predominante	-0,47	0,000	Moda	Professor
Escola tem laboratório de informática	-0,59	0,000	Binária	Escola
Escola tem TV	-0,83	0,026	Binária	Escola
Escola situada na região Norte	-0,96	0,013	Categórica	Escola
Q25 (Motivação para o Enem)	-1,41	0,033	Moda	Aluno

Como será visto mais adiante, dos atributos mais influentes, apenas a quantidade de banheiros aparece explicitada pelas técnicas de regras. Porém, a maioria dos atributos embute aspectos econômico-financeiros. É importante destacar que a escola ter laboratório de ciências contribui para o sucesso enquanto ter laboratório de informática contribui para o insucesso, assim com ter TV. Computadores usados para interação em redes sociais assim como assistir a televisão podem ser distratores da aprendizagem.

Árvore de Decisão é uma estrutura hierárquica para decisão seqüencial em forma de árvore invertida da raiz para as folhas (SAFAVIAN, 1991) que particiona o espaço de entrada para maximizar o ganho de informação em relação ao objetivo. As regras da árvore explicitam o conhecimento embutido nos dados de forma humanamente compreensível por regras do tipo *se-então*. A Tabela II ilustra parte de um ramo da árvore, indicando a concentração de escolas fortes e a sua representatividade percentual. No topo de árvore (à esquerda) ficam os atributos mais importantes. Nas regras de três condições, a concentração de escolas fortes varia desde 2,2% no pior extremo a 91,3% no melhor. Novamente, a influência das características econômico-financeiras é intensa, na renda familiar diretamente ou no ProUni ou banheiros da casa, indiretamente.

TABELA III

RAMO DA ÁRVORE DE DECISÃO INDICANDO CONFIANÇA E COBERTURA DE CADA REGRA COMBINADA COM AS ANTERIORES

POPULAÇÃO	CONDIÇÃO NÍVEL-1	CONDIÇÃO NÍVEL-2	CONDIÇÃO NÍVEL-3
Raiz	Optou ProUni=Não	Renda Familiar>30 S.M.	No. Alunos > 46
25%, 100%	63,8%, 27,2%	81,9%, 11,7%	**91,3%,** 6,7%
			22 < No. Alunos <=46
			75%, 3,7%
			No. Alunos <=22
			51,7%, 1,3%
	Optou ProUni=Sim	Quant. Banheiros >= 3	
	10,6%, 72,8%	32,7%,8,4%	
		Quant. Banheiros = 2	
		13,4%, 25,3%	
		Quant. Banheiros = 1	Renda Familiar < 2 S.M.
		4,0%, 39,1%	**2,2%,** 25,5%

A indução de regras também gera regras do tipo *se-então*, mas sem particionar o espaço de entrada nem ponderar a força da regra pela sua massa, minerando "pepitas de conhecimento". As melhores regras medidas pelo *lift* (razão das concentrações no grupo / população) (Tabela III), mostram o ganho de qualidade regra de 2 condições.

TABELA IIIII

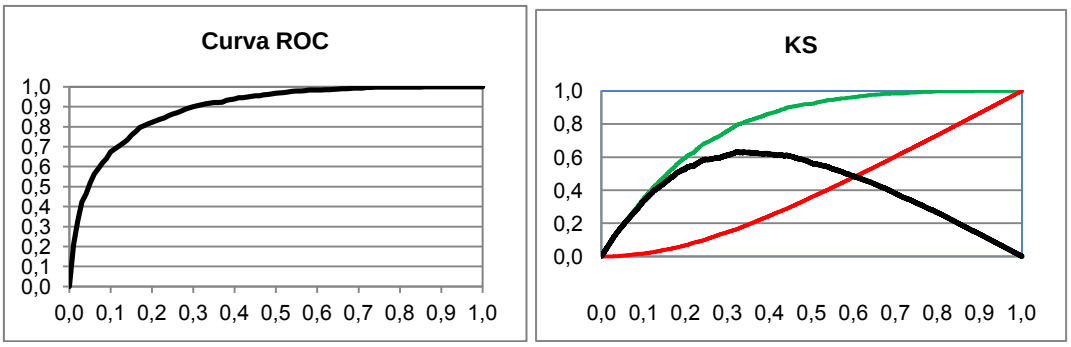
REGRAS INDUZIDAS COM 1 E 2 CONDIÇÕES, AVALIADAS PELAS MÉTRICAS: COBERTURA, CONFIANÇA E *LIFT*.

Condição-1	Condição-2	Cobertura	Confiança	Lift
Escolaridade do pai >= Superior	No. de funcionários da escola > 75	18,3%	62,1%	2,48
Escolaridade do pai >= Superior		36,7%	51,6%	2,06
No. de funcionários da escola > 75		32,5%	41,7%	1,66
No. de funcionários da escola < 75		66,3%	11,6%	0,46
Escolaridade da mãe <= 4a série		4,2%	3,5%	0,14
Escolaridade da mãe <= 4a série	No. de funcionários da escola < 75	3,8%	1,6%	0,07

4. Avaliação de desempenho

O escore de propensão teve seu dempesenho avaliado pelos testes Kolmogorov-Smirnov (KS) e a Curva ROC. A área sob a Curva ROC (AUC_ROC) e a máxima distância do KS (Máx_KS) foram as métricas de desempenho usadas numa amostra estatisticamente independente do modelo. Ambas as métricas, para o classificador ideal, têm valor igual a um (1). Neste estudo foi AUC_ROC=0,897 e Máx_KS=0,632.

Figura I – Curva ROC à esquerda e curva KS2 à direita



5. Conclusão

Este artigo analisou e extraiu conhecimento das bases de dados do ENEM e do Censo Escolar de 2011 sobre a qualidade do ensino médio privado brasileiro. Mesmo em bases tão grandes e em diversos grãos, a metodologia CRISP-DM com a visão de *Domain-Driven Data Mining* permitiram 1) caracterizar a "boa" escola como objetivo e 2) extrair o conhecimento voltado para 3 funcionalidades de um sistema de suporte à decisão, em relação ao objetivo definido.

A regressão logística produziu um classificador capaz de gerar uma pontuação de propensão ao sucesso da escola, a partir das suas características e daquelas dos seus docentes e discentes e famílias. A pontuação tem alto poder discriminante, de qualidade medida pelas AUC_ROC=0,897 e Máx_KS2=0,632. Os coeficientes mostraram que o fator econômico é relevante e aparece indiretamente na região da escola. Também, mostraram que ter infra-estrutura de laboratório pode ser bom ou ruim; precisa ser verificado se a efetividade depende da sua alocação, forma de uso e controle.

A árvore de decisão extraiu o conhecimento explicitando como o especialista humano decidiria de forma seqüencial utilizando regras. Outras técnicas geraram regras similares que corroboraram os resultados e deixaram claro que os principais fatores que influenciam a boa qualidade das escolas está nos aspectos econômico-financeiros, seja direta (renda familiar) ou indiretamente (No. banheiros da casa, opção do ProUni) ou em aspectos culturais (nível de educação da mãe ou do pai) da família.

O artigo apresentou uma solução de mineração de dados para avaliar e prever a qualidade das escolas de ensino médio, desenvolvida visando à sua futura implantação em um Sistema de Suporte à Decisão para operação e navegação em tempo real. Estudo análogo foi desenvolvido para escolas públicas e confirmou as políticas públicas da merenda escolar e da licenciatura.

6. Referências

Portal INEP. Disponível em: <http://portal.inep.gov.br/>. Acesso em:18 / 11 / 2013.

Travitzki, R. *ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar*. Tese doutorado, USP, São Paulo, 2013.

Wirth, R.; J. Hipp, *CRISP-DM: Towards a standard process model for data mining*. In Proc. of the Fourth Int. Conf. on PADD, 2000.

Sousa, M.U.R.S. Silva, K.P. Adeodato, P.J.L. *Data Mining Applied To The Processes Celerity Of Pernambuco's State Court Of Accounts*. In: Proc. CONTECSI 2008.

Cao, L.: Introduction to Domain Driven Data Mining. In: Cao, L., et al. (eds.) *Data Mining for Business Applications*, pp. 3–10 (2008)

Hair Jr.; J.F. *et al. Análise multivariada de dados*. 5ª ed. Porto Alegre: Bookman, 2005.

Safavian, S.R.; Landgrebe, D. A survey of Decision Tree Classifier Methodology. *IEEE Trans. Systems, Man and Cybernetics*, vol. 21, 660-674, May/June 1991.

Censo Escolar. Disponível em: <http://portal.inep.gov.br/basica-censo>. Acesso em: 18 de Nov. 2013.

ENEM. Disponível em: <http://portal.inep.gov.br/web/enem/sobre-o-enem>. Acesso em: 18 Nov. 2013.