Estimativa de Demanda Potencial de Matrículas em Ensino Superior usando Dados Públicos e Múltiplos Modelos de Regressão

Pedro Calais Guerra, Rodrigo Yuji Mizobe Nakamura, Eduardo Raul Hruschka[2]

Big Data, Brasil
{pedro.calais,rodrigo.mizobe,eduardo.hruschka}@bigdata.inf.br

Abstract. Este artigo apresenta uma proposta de aplicação de múltiplos modelos de regressão (*ensembles*) para prever a demanda potencial por vagas de ensino superior no ensino público brasileiro. Foram utilizadas variáveis socioeconômicas e educacionais disponibilizadas por MEC, INEP e IBGE para construir modelos de regressão que prevêem a quantidade atual de alunos matriculados em cada município brasileiro. Em seguida, pode-se comparar a quantidade de alunos prevista pelos modelos com a quantidade real; a diferença entre esses valores é interpretada como indicador da demanda potencial de cada município. Este trabalho em andamento (i) reforça as possibilidades de exploração de grandes volumes de dados públicos por modelos de aprendizado de máquina que geram indicadores que ajudam a aprimorar procedimentos e processos de gestão pública no Brasil, além de (ii) chamar a atenção para o fato de que métodos de aprendizado por *ensembles* são úteis também em tarefas de regressão, embora a literatura seja fortemente enviesada para tarefas de classificação, e (iii) ressaltar a utilidade de modelos de regressão para aplicações em que se está interessado na informação contida no erro da predição, e não somente na predição em si.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Palavras-chave: aprendizado de máquina, regressão, random forests, *ensembles*, economia da educação

1. INTRODUÇÃO

O uso de análise de dados para formulação de políticas públicas tem se tornado cada vez mais comum e importante como recurso que permite o aperfeiçoamento dos serviços públicos prestados aos cidadãos pelos governos das esferas municipal, estadual e federal. Possibilitados, em parte, pela crescente disponibilização de dados públicos abertos [DataRio 2014; DataViva 2014], são cada vez mais comuns, no Brasil, iniciativas que exploram dados públicos ou de redes sociais para prover serviços e promover ações como vigilância contra dengue [Gomide et al. 2011], planejamento de sistemas públicos de transporte [Cintra and Neves 2013] e combate à violência [Muggah and Diniz 2014].

Neste artigo, é proposta uma aplicação inovadora que processa dados governamentais abertos relativos a indicadores sociais, econômicos e educacionais dos 5.565 municípios brasileiros e estima, para cada um, a demanda não atendida por vagas no ensino superior público brasileiro. A ideia geral do método consiste em ajustar modelos de regressão que estimam a quantidade atual de alunos matriculados a partir de atributos socioecônomicos e características gerais dos municípios (ou região de interesse). Se o modelo ajustado tiver boa qualidade, a diferença entre a quantidade prevista

[2]Os autores agradecem a Ronaldo Prati e a Eric Yuzo pelas sugestões e contribuições na geração da visualização da demanda potencial.

pelo modelo e a quantidade real de alunos correntemente matriculados pode ser interpretada como a demanda não-atendida da região – uma consequência do fato, capturado pela regressão, de que há regiões do País com características populacionais e socioeconômicas semelhantes e que possuem uma quantidade maior de alunos matriculados. Caso o número de matriculados para um município, de acordo com o modelo, seja superior à quantidade real de alunos matriculados, existe uma demanda positiva por vagas no ensino público superior; de maneira análoga, quando o valor previsto pelo modelo é inferior à quantidade atual de alunos matriculados há um indicativo de que o município está avançado em relação a municípios similares no que se refere à oferta de vagas no ensino, ou de que a região possui características particulares que justifiquem esta oferta e que não estão capturadas pelos atributos socioeconômicos considerados.

Os resultados das análises realizadas no presente trabalho apresentam grande potencial para auxiliar a gestão pública, por meio de critérios técnicos e objetivos obtidos via técnicas de mineração de dados e *big data*, na tomada de decisões relativas à expansão da oferta de vagas no ensino público superior, em particular àquelas relacionadas à construção de novos *campi* universitários.

2. METODOLOGIA

A metodologia proposta para a estimativa da demanda potencial de alunos matriculados em instituições de ensino superior a partir de dados públicos é baseada em três passos:

(1) Inicialmente, foram obtidos conjuntos de dados publicamente disponíveis que possuem atributos dos municípios que são exógenos à variável de interesse (número de matriculados em cada município) e que, potencialmente, são correlacionados com esta variável. Em particular, foram consideradas variáveis em duas dimensões:
   —**Variáveis socioeconômicas**: correspondem a dados socioeconômicos dos municípios brasileiros, relativos à distribuição de renda, gênero, escolaridade, densidade demográfica, distribuição da população por faixa etária, entre outros. **Fonte de Dados: Censo 2010 [IBGE 2010]**.
   —**Variáveis educacionais**: representam indicadores relativos à área da Educação no Brasil, como quantidade de alunos matriculados no ensino básico, número de alunos inscritos no ENEM por ano, e número de alunos matriculados em cursos de Ensino Superior no Brasil, por curso, turno, município e tipo de instituição de ensino (privada ou pública). **Fonte de Dados: Censo do Ensino Superior 2012, MicroDados ENADE e MicroDados ENEM [INEP 2014]**.

(2) Em seguida, formula-se uma tarefa de predição via *regressão*: a partir de um conjunto de observações $((X_0, y_0), (X_1, y_1), ..., (X_N, y_N))$, para as quais X_i é o vetor de variáveis independentes obtidas na etapa anterior para o i-ésimo município e y_i é o número de alunos matriculados no Ensino Superior neste município, constrói-se uma função $\hat{f} : X \rightarrow y$ que aproxima a função real e desconhecida f [Mendes-Moreira et al. 2012]. A função $\hat{f}$ é obtida a partir de um algoritmo indutivo que examina um conjunto de n exemplos (conjunto de *treino*). Assim como em qualquer aplicação típica, procura-se pelo modelo que minimiza o erro de generalização, avaliado a partir da aplicação de $\hat{f}$ em um conjunto de exemplos não-observados durante a construção do modelo (chamado de conjunto de *teste*), por meio de métricas como R^2 e erro quadrático médio.

(3) Para cada instância (município), calcula-se o resíduo da regressão $\hat{\epsilon} = \hat{y} - y$, isto é, a diferença entre o valor predito pelo modelo e o valor real de y. Esta diferença é definida como sendo a **demanda potencial** de vagas no ensino público superior para cada município.

Foram realizados experimentos com diferentes algoritmos que instanciam a função $\hat{f}$: modelo linear, árvore de regressão, *random forest* e um modelo simples de agregação que combina o modelo linear com *random forests* a partir da média das predições de cada modelo individual. O interesse da aplicação recai em modelos que possuam duas qualidades:

(1) tenham boa qualidade na predição, isto é, uma alta capacidade de generalização;

(2) tenham baixa variância, ou seja, a previsão do valor de uma determinada instância X_i não deve variar fortemente caso uma nova amostra dos dados seja coletada.

A Tabela I exibe, para cada modelo avaliado, a qualidade da regressão e também a variância da qualidade do modelo para 1.000 execuções de validação cruzada. O melhor modelo em termos das duas métricas citadas anteriormente é aquele que combina o modelo de *random forests* e o modelo linear, e o resultado obtido – $R^2 = 0,74$ – é de boa qualidade quando comparado com modelos normalmente aceitos na área de Economia da Educação – cujos R^2 variam entre 0,3 e 0,4 [Ferraz et al. 2012; Van Klaveren 2010].

Tabela I. R^2 para diferentes modelos de regressão (média da validação cruzada com 10 pastas).

modelo	R^2	**variância** R^2
linear	0,70	0,05
árvore de regressão	0,60	0,09
random forest	0,72	0,03
combinação modelo linear + RF	**0,74**	**0,03**

As variáveis mais relevantes do modelo são (1) número de alunos inscritos no ENEM, (2) proporção de alunos de ensino básico com mãe com ensino superior, (3) população entre 19 e 30 anos, (4) renda e (5) número de alunos matriculados no ensino básico. Ressalta-se que validação cruzada é empregada para selecionar o melhor modelo; porém, para gerar o *ranking* de previsão de demanda potencial, o modelo é construído usando a estratégia *leave-one-out*, de modo que todos os dados sejam considerados para prever a demanda de cada município – com exceção do número de matriculados atual do próprio município. O critério de ordenação do *ranking* é a diferença entre o número esperado de alunos matriculados (isto é, predito pelo modelo) e o número atual.

3. ESTIMATIVA DE DEMANDA POTENCIAL POR VAGAS NOS ENSINO SUPERIOR PÚBLICO

Para exemplificar os resultados obtidos, a tabela II lista os 10 municípios com maior demanda potencial estimada de acordo com o método apresentado na Seção 2. Além disso, o mapa da Figura 1 exibe a demanda potencial por vagas no ensino superior para os 5.565 municípios brasileiros. Note que:

(1) Dentre os 10 municípios com maior demanda potencial, 5 não tem nenhum aluno matriculado, o que aponta para a necessidade da construção de novos campi para atender essas regiões.
(2) No entanto, a análise por demanda de um município deve contemplar não apenas a demanda específica de cada município, mas também a oferta e demanda dos municípios vizinhos. Municípios como Contagem (MG) e Guarulhos (SP), por exemplo, não possuem *campi* universitários públicos, mas integram regiões metropolitanas, o que diminui sua demanda real.

Tabela II. 10 municípios com maior demanda potencial de alunos de Ensino Superior Público.

posição	**município**	**# alunos atual**	**# alunos previsto**	**demanda**
1	Campo Grande (MS)	8.034	14.757	6.723
2	Ananindeua (PA)	0	4.809	4.809
3	Guarulhos (SP)	3.745	8.417	4.672
4	Contagem (MG)	0	4.202	4.202
5	Aracaju (SE)	2.644	6.761	4.117
6	Várzea Grande (MT)	0	4.107	4.107
7	Duque de Caxias (RJ)	1.381	5.249	3.868
8	Olinda (PE)	0	3.841	3.841
9	Jaboatão dos Guararapes (PE)	0	3.825	3.825
10	São Gonçalo (RJ)	2.519	6.129	3.610

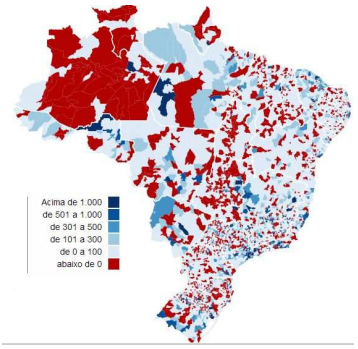

Fig. 1. Mapa de demanda potencial por vagas de ensino superior público, obtido a partir do resíduo da regressão formada por múltiplos regressores (linear + *random forest*).

4. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, foram propostos modelos de regressão construídos a partir de variáveis socioeconômicas para estimar a demanda potencial de número de matriculados no ensino superior público brasileiro. Os modelos são de boa qualidade quando comparados a modelos tipicamente reportados na literatura existente e permitem fazer inferências sobre a demanda potencial de matriculados nos municípios, a partir da comparação entre o número atual de alunos matriculados e o número predito pelo modelo.

Espera-se que este trabalho contribua para (i) reforçar as oportunidades de exploração de dados públicos por métodos de aprendizado de máquina que ajudem aprimorar a gestão pública no Brasil, (ii) chamar a atenção para o fato de que métodos de aprendizado por *ensembles* são úteis também em tarefas de regressão, embora a literatura seja fortemente enviesada para tarefas de classificação, e (iii) ressaltar a utilidade de modelos de regressão para aplicações em que se está interessado na informação contida no erro da predição, e não somente na predição em si.

Trabalhos futuros promissores incluem instanciar modelos de regressão na granularidade de curso ou área de estudo (isto é, Humanas/Exatas/Biológicas), de modo a gerar informações mais detalhadas sobre a demanda específica de cada município brasileiro. Uma direção complementar e relevante inclui, também, incorporar aos modelos a informação de proximidade e vizinhança entre as regiões, de modo a capturar a absorção de demanda de uma região por regiões vizinhas.

REFERÊNCIAS

Cintra, M. E. and Neves, O. A. A fuzzy decision tree for bus network management. *Symposium on Knowledge Discovery, Mining and Learning – KDMiLe*, 2013.

DataRio. Portal de dados abertos da Prefeitura do Rio de Janeiro. `http://data.rio.rj.gov.br/`, 2014. Acessado: 28/07/2014.

DataViva. Visualizando a economia do Brasil. `http://dataviva.info/`, 2014. Acessado: 28/07/2014.

Ferraz, C., Finan, F., and Moreira, D. B. Corrupting learning: Evidence from missing federal education funds in brazil. Working Paper 18150, National Bureau of Economic Research. June, 2012.

Gomide, J., Veloso, A., Jr., W. M., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM Web Science Conference (WebSci)*, 2011.

IBGE. Censo IBGE 2010. `http://censo2010.ibge.gov.br/en/`, 2010. Acessado: 05/08/2014.

INEP. Portal INEP. `http://portal.inep.gov.br/basica-levantamentos-acessar`, 2014. Acessado: 05/08/2014.

Mendes-Moreira, J. a., Soares, C., Jorge, A. M., and Sousa, J. F. D. Ensemble approaches for regression: A survey. *ACM Comput. Surv.* 45 (1): 10:1–10:40, Dec., 2012.

Muggah, R. and Diniz, G. Prevenindo a violência na américa latina por meio de novas tecnologias. Artigo estratégico, Instituto Igarape, 2014.

Van Klaveren, C. Lecturing Style Teaching and Student Performance. Working Papers 29, Top Institute for Evidence Based Education Research. 00, 2010.