Os medos dos paulistanos - Uma análise social

João Mendes, Marília Rocha, Pedro Morgado

Departamento de Estatística / IMECC - Unicamp

Introdução

Sendo o medo um sentimento frequente na vida das pessoas, esse trabalho busca analisar a relação desse sentimento com as diferentes condições socioeconômicas. A partir de um questionário aplicado na cidade de São Paulo, o objetivo do projeto é, com algumas das informações presentes no banco de dados, obter um traço latente representativo para o sentimento "medo" e prever, com técnicas de Aprendizado Supervisionado de Máquina, através do traço latente e das demais variáveis não utilizadas em seu cálculo, o sexo, a etnia e a renda familiar dos entrevistados.

Dados

O conjunto de dados foi retirado do endereço eletrônico CIS (Consórcio de Informações Sociais) e contém o resultado de um questionário aplicado com 1091 moradores da cidade de São Paulo. Ele é composto por 28 questões sobre medo, percepção de violência e características socioeconômicas, tendo um total de 86 variáveis.[1]

Metodologia

Por tratar-se de um questionário com o objetivo de analisar uma variável latente (medo), a metodologia utilizada para a redução da dimensão dos dados foi a Teoria de Resposta ao Item (TRI). Desse modo, foi possível utilizar algumas das questões relacionadas ao medo dos paulistanos para estimar um traço latente representativo do sentimento "medo". Para os itens dicotômicos foi utilizado o Modelo Logístico de dois Parâmetros (2LP), e para os politômicos de respostas graduais, o Modelo de Crédito Parcial Generalizado[2]. Com o traço latente calculado e com as outras variáveis socioeconômicas não utilizadas no cálculo deste, foi possível ajustar modelos de Aprendizado Supervisionado de Máquina para prever o sexo, a etnia e a renda familiar dos respondentes. Essas variáveis foram preditas separadamente, devido à complexibilidade de algoritmos para previsão multivariada com covariáveis categóricas. Os algoritmos utilizados para isso foram Árvore de Classificação, Floresta Aleatória, *Boosting* e Máquinas de Vetores de Suporte (MVS), comparando as acurácias obtidas.

Resultados

Com a utilização de técnicas de Teoria de Resposta ao Item, foi possível reduzir 36 variáveis relacionadas a medo, das quais 6 eram dicotômicas e 30 politômicas graduais (admitiam respostas do tipo 0 = "Nenhum medo", 1 = "Pouco medo", 2 = "Pouco medo" para a situação apresentada), em um único traço latente representativo do sentimento "medo".

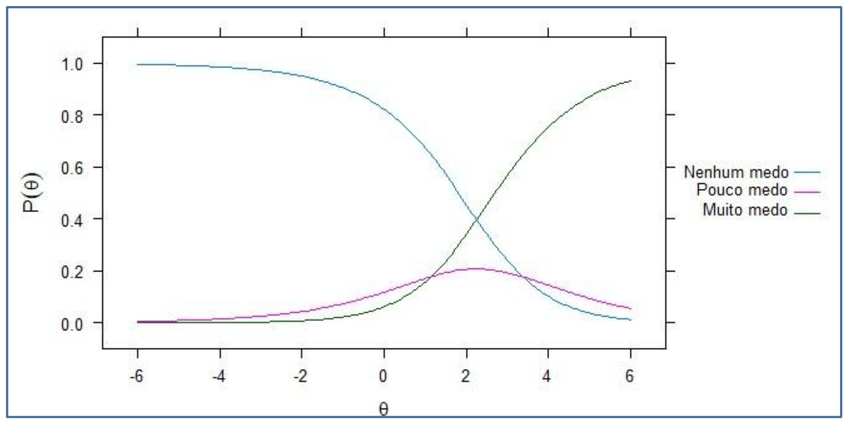

Figura 1: Linhas de traço para medo de assombração

Analisando-se a Figura 1, percebe-se que quanto maior o traço latente, maior a probabilidade de escolha da categoria "Muito medo". Além disso, como o medo de assombração é um dos únicos medos não tão condizentes com situações palpáveis, supõe-se que esse seja o motivo do deslocamento para a direita da curva referente à categoria "Pouco medo" e do decaimento leve para a curva da categoria "Nenhum medo".

Para a predição da variável sexo no banco de dados de treino, obteve-se melhor resultado, em relação à acurácia, para o método *Boosting*, sendo esta de 0.65. No entanto, o método apresentou-se computacionalmente ineficiente, devido ao tempo de execução. Além disso, a covariável de maior importância para a predição no método *Boosting* foi o traço latente, seguido pelo categoria na população economicamente ativa. O método MVS-Linear apresentou acurácia de 0.63 e demonstrou-se computacionalmente eficiente.

Para as predições das variáveis etnia e renda (dividida em faixas salariais familiares), não obteve-se boas acurácias, sendo o máximo de 0.52 por Floresta Aleatória para a variável etnia e 0.42 por MVS-Linear para a variável renda.

Conclusões

Os resultados obtidos indicam que, no geral, não se obteve boas predições para as variáveis sexo, etnia e renda dos entrevistados. As covariáveis serem todas categóricas e muitas serem nominais e politômicas limitou uma melhor análise descritiva dos dados, como o estudo das correlações entre as variáveis, e o uso de métodos mais sofisticados de Aprendizado Supervisionado de Máquina, já que muitos exigem suposições de normalidade e independência das variáveis respostas. Além disso, acredita-se que as melhores acurácias para a variável resposta sexo em relação às demais seja devido à essa ser dicotômica e não politômica.

Apesar dos baixos valores de acurácia, notou-se que para a faixa de renda, os erros de predição ocorreram entre categorias próximas.

Em relação à importância das variáveis para predição, a principal covariável para as predições tanto de sexo, quanto etnia e renda, foi o traço latente, o que faz sentido já que o mesmo resume informações presentes em 36 outras variáveis. Para a variável resposta sexo, a categoria do entrevistado na população economicamente ativa mostrou-se uma das mais importantes. Para as variáveis respostas etnia e renda, a cidade/estado de nascimento e as variáveis referentes à quantidade, em faixas (0 - Não tenho, 1 - Tenho um, 2 - Tenho dois ou mais), de pertences como eletrodomésticos, cômodos em casa, etc, mostram-se, respectivamente, com grande importância.

Referências

[1] DATAFOLHA - Instituto de Pesquisas. Medo dos Paulistanos, 2008. (Banco de dados). São Paulo, Datafolha, 2008. In: Consórcio de Informações Sociais, 2012. Disponível em: www.cis.org.br. Acesso em 25/11/2018.

[2] ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. Teoria da Resposta ao Item: Conceitos e Aplicações.