Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Helton Souza Lima

abril de 2019

Proposta

Histórico do assunto

O Programa Bolsa Família (PBF) é o maior programa de distribuição de renda do Brasil [1], através de um benefício em dinheiro transferido diretamente do governo federal para famílias dentro da linha da pobreza e extrema pobreza, para garantir um alívio mais imediato à pobreza, complementando a renda dessas famílias e condicionando à participação nos serviços de saúde e educação. De acordo com o artigo da Dra. Daniela Dias Kuhn [2], o programa foi efetivo na melhoria dos índices de desenvolvimento humano no Estado do Rio Grande do Sul. Podemos citar outro estudo, realizado em Minas Gerais [3] que aponta a mesma conclusão no âmbito deste estado.

Por outro lado, é recorrente a veiculação de notícias [4] referentes a fraudes nos benefícios do Programa Bolsa Família. Essas fraudes acarretam saques de valores superiores ao necessário para o atingimento do objetivo do programa e precisam ser eliminadas, pois acarretam um custo desnecessário ao governo, chegando ao patamar de bilhões [5] de reais.

A empresa em que trabalho é a DATAPREV [6], empresa de processamento de dados do governo federal. Uma atividade recorrente de nossa empresa é o levantamento e cruzamento de informações entre bases de dados para verificar o correto cumprimento de políticas públicas através de sistemas informatizados. O trabalho com os dados do Bolsa Família permitirá a investigação de situações semelhantes ao dia-a-dia de nossa empresa, e a experiência será útil dentro de um contexto semelhante ao problema abordado neste trabalho.

Descrição do problema

O público-alvo do PBF são as pessoas que estão dentro da faixa da pobreza ou pobreza extrema. Entende-se que os volumes financeiros disponibilizados para o programa é proporcional à quantidade de pessoas dentro das faixas sociais que são alvo do programa, de forma que, a partir de dados de informações sociais e econômicas, como a população total, esperança de vida ao nascer, taxa de analfabetismo, percentual de crianças na escola, taxa de frequência, renda per capita, percentual de distribuição de renda, proporção de pobres, etc, é possível predizer o volume financeiro a ser utilizado para o PBF. Em suma, acredita-se que municípios com índices mais baixos (que compõem o IDH) deveriam ter um maior recebimento de recursos do PBF ao passo que municípios com índices mais altos deveriam receber menos recursos do PBF, resguardadas a quantidade de pessoas residentes nesses municípios.

De posse desses dados granularizados a nível de município brasileiro, propõe-se a utilização de modelos de machine learning que serão treinados utilizando-se os dados de parte desses municípios e poderão predizer o volume financeiro de outra parte desses municípios. Em um momento inicial, a análise dos dados poderá apontar a correlação entre os indicadores sociais e o volume financeiro do PBF associado com cada município. Em seguida, é possível que alguns municípios apontem discrepância nessas correlações e possam ser apontados como municípios onde há maior incidência de fraudes. Uma das respostas que se deseja responder é: Será que existem municípios com alto IDH mas que, mesmo assim, recebem muitos recursos do PBF, em comparação com outros municípios semelhantes?

Conjuntos de dados e entradas

Os dados utilizados foram obtidos de duas fontes. A primeira fonte são os dados relacionados ao Índice de Desenvolvimento Humano Municipal (IDHM), disponibilizado pelo site Atlas do Desenvolvimento Humano no Brasil [7] ou no site da Kaggle [8]. Os dados do IDHM são disponibilizados para cada um dos 5565 municípios brasileiros, abrangendo índices que podem ser agrupados em 3 dimensões: índices de longevidade, índices de acesso ao conhecimento e índices de renda. Os dados utilizados são de janeiro de 2010, ou seja, 7 anos após o ano de lançamento do PBF, que pode ser considerado como suficiente para o programa ter atingido uma maturidade em sua operacionalização e gestão.

A segunda fonte são os dados relacionados à quantidade de famílias beneficiárias e o total de pagamentos disponibilizados pelo PBF para cada município brasileiro. Os dados são disponibilizados pelo Ministério da Cidadania [9].

Como ambos os conjuntos de dados são listados para cada município, é possível associarmos para um município os dados do IDHM e os dados de valores repassados do PBF. Desta forma, será possível treinar modelos de machine learning que poderão verificar a correlação entre as informações sociais dos municípios e os respectivos pagamentos do PBF. Parte dos dados servirão para treinamento do modelo

e parte dos dados serão para o teste e validação.

Descrição da solução

Considerando a grande quantidade de variáveis que compõem a base e o cálculo do IDHM, inicialmente propõe-se a avaliação das variáveis que podem ser retiradas da base do IDHM sem prejuízos para a predição dos algoritmos. Em seguida, propõe-se realizar um split dos dados para separar entre conjunto de treinamento, validação e testes. Em seguida, a utilização de pelo menos 3 algoritmos a serem avaliados e comparados quanto à sua acurácia. Por fim, a verificação dentre os casos de teste, aqueles que tiveram maior discrepância em relação à predição dos algoritmos. Essa discrepância pode estar associada às fraudes, que aumentam os valores dos recursos repassados.

Uma questão a ser abordada é: se dentro do meu conjunto de treinamento houver casos que contenham fraudes, então o modelo será treinado já considerando as possíveis fraudes. Sendo assim, vai diminuir o seu poder de identificar os municípios com maior probabilidade de fraude.

Modelo de referência (benchmark)

Dentro do conjunto de dados já existem os valores que se deseja predizer, de forma que será possível calcular a acurácia dos algoritmos escolhidos. Esse trabalho, nesta fase inicial de análise, é útil no sentido de avaliar uma metodologia que pode ser aplicada a dados mais recentes em busca de identificar lugares onde há maior incidência de fraudes. Serão buscadas notícias sobre as fraudes já conhecidas e publicadas e, de posse dessas informações, avaliar os dados dos municípios afetados no ano de 2010. Notícias posteriores a esse ano, mas não tão distantes, serão as mais relevantes, por exemplo, entre os anos de 2011 e 2014.

Não encontramos, até o momento, algum trabalho que realizou trabalho semelhante para que possamos realizar uma comparação direta. Sendo assim, o modelo que vamos usar como de referência será a utilização do modelo de Regressão Logística para realizar as predições.

Métricas de avaliação

Neste momento inicial de projeto da solução, já que o valor a ser previsto é um valor contínuo, podemos antecipar que a métrica que deverá ser utilizada é o Root Mean Squared Erros (RMSE), pois é uma métrica que avalia a distância entre o valor previsto e o valor real. Essa métrica poderá ser calculada pelo próprio scikit-learn comparando os valores previstos e os valores reais.

Design do projeto

- Unir os dois datasets em apenas um dataset
  - Através do campo "codmun6" do arquivo de IDHM e do campo "ibge" do arquivo do PBF
  - Este é o identificador do município, em ambos os campos, utilizado pelo IBGE
- Realizar análise inicial do arquivo
  - Realizar análise das variáveis. Quais são categóricas e quais são contínuas e avaliar seus padrões (média, mediana, moda, quartis, variância, desvio padrão, percentual)
  - Avaliar o Kernel inicial automático gerado no site do Kaggle [10]. Embora precisaremos excluir os dados anteriores a 2010 (anos de 2000 e 1991)
  - Verificar as correlações entre as variáveis usando as bibliotecas de análise de dados
  - Por exemplo: Scatter Plot, Correlation, Stacked Column Chart, T-test
  - Avaliar variáveis que podem ser excluídas para o aprendizado dos modelos
  - Identificar possíveis outliers e avaliar se a remoção será pertinente
  - Avaliar a necessidade de transformação ou criação de variáveis
- Escolher algoritmos de machine learning a serem avaliados neste trabalho
  - A analisar pelos dados que estão sendo avaliados e as restrições e objetivos das predições
  - Verificar a possibilidade de pelo menos 3 algoritmos
  - Buscar por notícias que já possam sugerir a existência de fraudes em determinados municípios nos anos de 2010 até 2014
- Implementar o uso desses algoritmos e treiná-los
- Realizar a predição nos dados de teste.
- Avaliar o F1 score dos algoritmos e escolher o melhor
  - Cabe avaliar outra métrica
  - Avaliar os dados mais discrepantes entre o que foi previsto e o real
- Avaliar trabalhos futuros
- Redigir o projeto final

Referências

[1] Portal do Programa Bolsa Família. Ministério da Cidadania.

[2] Kuhn, Daniela Dias. Tonetto, Elci da Silva. O Programa Bolsa Família e os indicadores sociais no Rio Grande do Sul. Desenvolvimento em Questão

[3] Denubila, Lais Atanaka. Ferreira, Marco Aurelio Marques. Monteiro, Doraliza Auxiliadora Abranches. Programa Bolsa Família: Análise Da Trajetória Dos Indicadores Sociais Em Minas Gerais. Associação Nacional de Pós-Graduação e Pesquisa em Administração

[4] Busca no Google sobre fraudes no Bolsa Família

[5] "Controladoria-Geral acha R$ 1,3 bi em fraudes no Bolsa Família", Revista Exame Online, 4 de janeiro de 2018

[6]Portal da Dataprev. Empresa de Tecnologia e Informações da Previdência Social

[7] Portal do Atlas do Desenvolvimento Humano no Brasil

[8] Human Development Indexes and Census data for Brazilian municipalities. Portal Kaggle

[9] Visualizador de Dados Sociais. Um portal do Ministério da Cidadania

[10] Human Development Indexes and Census data for Brazilian municipalities. Kaggle DataSet. Setembro/2018