Alessandre da Silva Martins

31 de maio de 2018

Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Background do domínio

Neste projeto irei analisar as empresas de saneamento quanto ao relacionamento entre a estrutura de propriedade e a eficiência operacional. Primeiro, com os dados que obtive da Sistema Nacional de Informações sobre Saneamento - SNIS iremos selecionar os 100 municípios mais populosos do Brasil em 2016, para explorar um pequeno subconjunto de dados como amostra e identificar se alguma empresa está altamente correlacionada com outra. Em seguida vamos pré-processar os dados, dimensionando cada Natureza Jurídica e remover os outliers caso exista algum. Iremos aplicar o PCA(Principal Component Analysis[1]) e um algoritmo de clustering para criar os segmentos das empresas selecionadas no conjunto de dados limpo anteriormente. Também iremos estimar a efetividade operacional de cada empresa, conforme o método DEA(Data Envelopment Analysis[2]) e analisar a relação entre os resultados do DEA e a estrutura de propriedade[3]. As empresas de cada município brasileiro do setor de saneamento básico foram selecionadas, cujos dados referentes foram coletados das demonstrações contábeis. Como resultado deste projeto, procura-se enfatizar os problemas encontrados mais significativos e analisar as estruturas de propriedades das empresas privadas e estatais do setor de saneamento básico do Brasil. Por fim, vamos comparar os segmentos encontrados com uma marcação adicional e considerar a maneira como essa informação poderia auxiliar os prefeitos dos municípios com futuras tomadas de decisões de serviço de saneamento básico de suas cidades.

A estrutura de propriedade de saneamento básico influencia diretamente o desenvolvimento de uma cidade. Pois o saneamento básico adequado nos dará

[1] "Segmentation and Clustering | Udacity." https://br.udacity.com/course/segmentation-and-clustering--ud981. Acessado em 31 mai. 2018.
[2] "Tutorial » Data Envelopment Analysis: DEAzone.com." http://deazone.com/en/resources/tutorial. Acessado em 31 mai. 2018.
[3] "Governança Corporativa e Estrutura de Propriedade no Brasil ...." http://www.pablo.prof.ufu.br/artigos/ebf3.pdf. Acessado em 31 mai. 2018.

indicativos do impacto de um município no meio ambiente, se tem condições de prevenir doenças e melhorar a saúde, a qualidade de vida da população. O saneamento básico também pode definir à produtividade do indivíduo elevando a atividade econômica do município.

Enunciação do problema

Diante dos riscos que o setor de saneamento expressa atualmente, qual a correlação entre a estrutura de propriedade das empresas de saneamento básico das principais cidades brasileiras com os níveis de eficiência através de machine learning?

Por causa de problemas de saneamento básico, os municípios brasileiros coordenaram processos de privatização do setor de saneamento, a pretexto de inscrever maior eficiência ao setor.[4] Entretanto, o processo de privatização pode repetir ao serviço público problemas e riscos típicos dos mercados financeiros, com destaque aos conflitos de agência derivados das diferentes estruturas de propriedade.[5] As empresas privadas expressam diferentes níveis de concentração acionária, com possíveis prejuízos à transparência e percepção a respeito da idoneidade dos proprietários, além de fatores relacionados à instabilidade dos mercados e falhas regulatórias. No final das contas, a população usuária do serviço experimenta as consequências de um processo inadequado de privatização, com perda à divulgada eficiência típica do setor privado. [6]

[4] "Do ownership and size affect the performance of water utilities ...." 2 abr. 2011, https://link.springer.com/article/10.1007/s10997-011-9173-6. Acessado em 31 mai. 2018.
[5] "The Structure of Corporate Ownership: Causes and Consequences." https://www.jstor.org/stable/1833178. Acessado em 31 mai. 2018.
[6] "Privatization and Regulation - unpan1.un.org, 24.07.2012." http://unpan1.un.org/intradoc/groups/public/documents/un/unpan000152.pdf. Acessado em 31 mai. 2018.

Alessandre da Silva Martins

31 de maio de 2018

Conjunto de Dados e Inputs

O conjunto de dados de entrada considerado para o projeto foi obtido do site do Sistema Nacional de Informações sobre Saneamento - SNIS[7], em Água e Esgoto no item 'Agrupamento dinâmico de indicadores e informações desagregadas por ano de referência'. O SNIS se constitui no maior e mais importante sistema de informações do setor saneamento no Brasil, apoiando-se em um banco de dados que contém informações de caráter institucional, administrativo, operacional, gerencial, econômico-financeiro, contábil e de qualidade sobre a prestação de serviços de água, de esgotos e de manejo de resíduos sólidos urbanos.[8] Essa busca permite o agrupamento dinâmico de indicadores e informações desagregadas por ano de referência. Em todas as situações de agrupamento é fornecida a totalização e, no caso dos indicadores, o resultado de cálculo do indicador médio. Permite filtrar por, Ano de referência, Abrangência, Tipo de serviço, Natureza jurídica, Região, Estado, Região metropolitana, Prestador de serviço, Familia de informações e indicadores e Informações e Indicadores propriamente ditos.

Explanação da solução

A proposta é analisar as empresas de saneamento das principais cidades brasileiras quanto ao relacionamento entre a estrutura de propriedade e a eficiência operacional usando métodos de aprendizagem não-supervisionada de machine learning. Com base nessa análise, determinar como as empresas de saneamento privada ou pública influenciam na eficiência do saneamento de um município. Pode ser que uma natureza jurídica das empresas se destaque mais que a outra. Também se propõe investigar a estrutura de propriedade das empresas de saneamento básico presentes nas principais cidades do Brasil, estimar a efetividade operacional de cada empresa, conforme o método *DEA* e analisar a relação entre os resultados do DEA e a estrutura de propriedade.

Será utilizado análise de componentes principais(PCA) para redução de dimensão e seleção daquelas features que descrevem melhor a variabilidade. Sendo assim, podemos elaborar conclusões sobre a estrutura(que não se manifestam claramente) de dados das empresas.

---

[7] "SNIS Municípios - SNIS - Série Histórica." http://app3.cidades.gov.br/serieHistorica. Acessado em 31 mai. 2018.
[8] "SNIS - Série Histórica - SINIR." http://www.sinir.gov.br/web/guest/snis-serie-historica. Acessado em 31 mai. 2018.

Posteriormente será realizado um trabalho de treinamento com aprendizagem não supervisionada para clusterização. Seguiremos transformando os pontos de volta em suas dimensões originais, recuperando os pontos de dados específicos do cluster.

Modelo de referência (benchmark)

O DEA permite, entre outros aspectos relevantes, avaliar a eficiência das DMUs que sejam referência(benchamarking) para as demais DMUs analisadas que não são eficientes. A DMU eficiente vai encapsular todas as DMUs que não são eficientes. Reparando que uma regressão não permite fazer isso. Não tem como avaliar a eficiência por meio de uma análise de regressão. Então, os resultados do índice DEA comporão a variável dependente para a análise de estatística inferencial que permitirá analisar o relacionamento entre a estrutura de propriedade e a eficiência operacional das empresas de saneamento básico das principais cidades brasileiras.

Avaliação Métrica

A estrutura de dados não é clara, caso exista. Portanto, para selecionarmos o melhor método de clusterização iremos quantificar a "eficiência" de um clustering ao calcular o coeficiente de silhueta[9] de cada ponto de dados, fornecendo um método de pontuação simples de um dado clustering. Uma silhueta define a área de contorno de cada cluster, nos mostrando que pontos se localizam dentro do cluster e quais pontos ficam em uma localização no meio de dois clusters.

Também iremos criar uma feature para auxiliar no processo de clusterização calculando o índice de eficiência *DEA* pela relação entre os insumos e produtos. Quando maior a razão entre os produtos gerados e os insumos empregados, mais eficiente será a DMU, ou seja:

$$eficiência\ DEA\ =\ \frac{y_1u_1 + y_2u_2}{x_1v_1 + x_2v_2}$$

[9] "sklearn.metrics.silhouette_score — scikit-learn 0.19.1 documentation." http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html. Acessado em 2 jun. 2018.

Onde **u** e **v** são os pesos atribuídos para **y** e **x** que são as saídas e insumos respectivamente. Atribuir pesos é algo crítico que influenciam completamente a otimização do problema, o DEA tem mecanismos para contornar isso, eles dizem quais variáveis foram importantes para determinar a eficiência ou ineficiência de cada DMU.

Design do projeto

Iremos explorar os dados por meio de visualizações e códigos para compreender a relação entre a natureza jurídica de cada DMU e determinar qual o tipo de empresa mais influencia o saneamento básico das cidades no Brasil. Também iremos observar estatística do conjunto de dados, levando em consideração a relevância de cada natureza jurídica, e selecionar alguns exemplos de pontos de dados do conjunto de dados para acompanhar durante o andamento do projeto.

De antemão, o conjunto de dados é compostos por sete natureza jurídica importantes de empresa: Administração pública direta, autarquia, empresa privada, empresa pública, organização social, sociedade de economia mista com administração privada, sociedade de economia mista com administração pública.

Iremos selecionar algumas amostras de dados de pontos e explorá-los com mais detalhes, isso vai facilitar a compreensão da análise das DMUs.

Para garantir significativamente a importância dos resultados obtidos iremos pré-processar os dados para representar as DMUs e realizar um escalonamento dos dados também como remover os dados aberrantes(outliers). Dados aberrantes podem enviesar resultados que levam em consideração os pontos de dados.

Podemos obter a rentabilidade dos acionistas pelo giro do ativo, lucro e endividamento de uma dada empresa. Essa rentabilidade serve para medir a eficiência das empresas. Porém, há áreas em que a rentabilidade aos acionistas(geração de valor aos acionistas) não é suficiente. Ou seja, há áreas em que o bom desempenho financeiro não é suficiente para saber se uma atividade está sendo eficiente diante de uma organização ou de uma sociedade. Os exemplos mais típicos disso são justamente as organizações públicas, pois não registram lucro. Essas organizações precisam mensurar a eficiência de qualquer forma. Charnes, Cooper e Rhodes em 1978 apresentaram um modelo matemático

chamado DEA-CCR[10], é a junção de 'Data Envelopment Analysis', Análise Envoltória de Dados em português, com as iniciais de seus criadores. Presume-se que todas as unidades possuem um retorno constante de escala. O primeiro modelo DEA, avalia a eficiência entre unidades e identifica quais são as unidades(beachmarking) referência para as Unidades Tomadoras de Decisão(DMU - Decision Making Units) que não são eficientes. O DEA não precisa usar dados financeiros para medir a eficiência de um conjunto de empresas.

Então, os resultados serão decorrentes do cálculo da eficiência operacional de cada empresa a partir do DEA. O DEA-CCR vai dar como resultado um indicador de no máximo 1. A empresa que tiver DEA igual a 1 é porque foi 100% eficiente, já a empresa que tiver DEA igual a ou inferior a 0,99 quer dizer que foi dada proporção eficiente.

O modelo válida a proposta inicial de Charnes, Cooper e Rhodes (1978), onde as DMUs que, para demonstrar eficiência, devem usar o mínimo de insumos, ou *inputs* possíveis para constituir o máximo de produtos, ou *outputs* possíveis.

Para o setor de saneamento básico, há que empregar como variáveis de insumos e produtos itens semelhantes à proposta de Carmo (2003)[11]. Os insumos seriam a mão-de-obra empregada, a capacidade instalada, a extensão da rede de distribuição e a extensão da rede de coleta. Os produtos seriam o volume de água faturado, o volume de esgoto faturado, a economia ativa de água e a economia ativa de esgoto. A proposta é usar os dados do Sistema Nacional de Informações sobre Saneamento. Os **outputs** selecionados são: Indicador de atendimento urbano de água (%), Indicador de atendimento urbano de esgoto (%), Indicador de esgoto tratado por água consumida (%), Indicador novas ligações de água/ligações faltantes (%), Indicador novas ligações de esgoto/ligações faltantes (%), Indicador evolução nas perdas de faturamento (%), Indicador evolução nas perdas de distribuição (%); e como **inputs**: Indicador perdas no faturamento 2016 (%), Indicador perdas na distribuição 2016 (%), Tarifa média (R$/m³), LN Investimento, LN extensão rede de água, LN extensão rede de esgotos, LN do nº de funcionários em 2016.

Há um problema chamado "problema do tamanho" em usar dados absolutos como por exemplo a extensão da rede de água ou de esgoto. Imagina compara Franca, que é uma cidade pequena e verticalizada, uns 500 metros de canos já basta para ter saneamento eficiente, com o Rio de Janeiro, que é uma cidade muito

[10] "94 2.4 Análise Envoltória de Dados – DEA O objetivo ... - Teses USP." http://www.teses.usp.br/teses/disponiveis/3/3142/tde-13122006-180402/publico/04.pdf. Acessado em 31 mai. 2018.
[11] "Avaliação da Eficiência Técnica das Empresas de Saneamento ...." https://repositorio.ufpe.br/handle/123456789/5836. Acessado em 31 mai. 2018.

populosa e horizontalizada, deve precisar de uns 12 mil km de encanamento e ainda não basta. Por isso não é bom usar dados absolutos nas análises quantitativas. A média e a mediana variam significativamente (indicando um grande desvio), portanto vamos aplicar um escalonamento não linear. Vamos reduzir o desvio aplicando o algoritmo natural. Os dados que tratamos com logaritmo natural estão indicados com as iniciais "LN". Após aplicar o algoritmo natural para o escalonamento dos dados, a distribuição para cada atributo deve parecer mais normalizado.

Vamos usar a análise de componentes principais (PCA) para elaborar conclusões sobre a estrutura subjacente de dados de DMUs. A PCA calcula as dimensões que melhor maximizam a variância das natureza jurídica envolvidas, encontrando as combinações de componentes de natureza jurídica descrevem as DMUs.

Criar um método de pontuação simples de um dado clustering para identificar o número ótimo de cluster melhor segmenta os dados.

Por fim, utilizando o modelo de clusterização escolhido, iremos calibrar para obter os melhores hiperparâmetros e documentá-lo.