Aproximando a distância média dos veículos do CVRP a partir de relações geoespaciais

Germano B. dos Santos[1]

[1]Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPeD-Lab)
Universidade Federal de Viçosa - Campus Florestal (UFV – CAF)

germano.santos@ufv.br

1. Problema

Nesta seção será apresentado o *Capacitated Vehicle Routing Problem*(CVPR) e sua formalização, assim como, o Desafio Loggibud e a abordagem escolhida para o estudo.

O CVRP é um problema de roteamento de veículos, esse último pode ser definido como: dada uma demanda de transporte e uma frota de veículos, deve-se determinar conjuntos de rotas que satisfaçam todas as demandas ou algumas demandas com a frota de veículos com um custo mínimo; em outras palavras seria definir a rota r_i do veículo v_i que atenda uma fração da demanda em um custo mínimo. A seguir existe a apresentação formal do problema em que segue a formulação *two-index*.

Minimize $$\sum_{k \in K} \sum_{(i,j) \in E} c_{ij} x_{ij}^k \ (1)$$

Tal que $$\sum_{k \in K} \sum_{i \in V, i \neq j} x_{ij}^k = 1, \ \forall j \in V - \{0\} \ (2)$$

$$\sum_{j \in V - \{0\}} x_{0j}^k = 1, \ \forall k \in K \ (3)$$

$$\sum_{i \in V, i \neq j} x_{ij}^k - \sum_{i \in V} x_{ji}^k = 0, \ \forall j \in V, \forall k \in K \ (4)$$

$$\sum_{i \in V} \sum_{j \in V - \{0\}, i \neq j} q_{ij} x_{ij}^k \leq Q, \ \forall k \in K \ (5)$$

$$\sum_{k \in K} \sum_{(i,j) \in S, i \neq j} x_{ij}^k \leq |S| - 1, \ S \subseteq V - \{0\} \ (6)$$

$$x_{ij}^k \in \{0, 1\}, \ \forall k \in K, \ \forall (i,j) \in E \ (7)$$

Dado que
$G(V, E) =$ Um grafo que representa a localização e a rota do veículo
$V = \{0, 1, \ldots, n\} =$ Conjunto de vértices. O vértice 0 representa o depósito.
$E =$ Conjunto de arestas que conectam cada vértice $e_{ij} = (i, j)$
$K = \{1, 2, \ldots, |K|\} =$ Conjunto de Veículos
$d_i \geq 0 =$ Demanda do cliente i
$Q \geq 0 =$ Capacidade máxima do veículo
$c_{ij} =$ Custos da viagem entre clientes i e j

$$x_{ij}^k = \begin{cases} 1, \text{se o veículo } k \text{ foi atribuído a rota } e_{ij} \\ 0, \text{caso contrário} \end{cases}$$

Na definição acima, temos que a primeira restrição significa que o objetivo é a minimização da soma dos custos das viagens. A segunda restrição define que um cliente não deve ser visitado por mais de um veículo. A restrição três especifica que todos

os veículos devem começar as rotas pelo depósito. A quarta restrição especifica que o número de veículos para uma direção é o mesmo número de veículos para a direção contrária, principal definição de fluxo. A quinta restrição define que a capacidade que o veículo deve entregar deve ser menor ou igual à capacidade máxima do mesmo. A sexta restrição corresponde a eliminação de *subtours*.

Essa definição do CVRP possui um número exponencial de restrições em relação ao número de clientes, o que torna esse problema intratável em alguns casos. Logo, é necessário resolvê-lo por meio de heurísticas, disponibilizados por *solvers* públicos como OR-TOOLS ou outros métodos como *Machine Learning* e *Redes Neurais Profundas*.

1.1. Loggibud Dataset

A Loggi é uma empresa de logística que tem como país-sede o Brasil. Ela fornece serviços de entrega por todo o Brasil, desde a coleta do produto, ou seja, o *first mile* até a entrega ao cliente *last-mile*. Sabe-se que o problema da última milha é recorrente em grandes empresas de logística, essa etapa contribui em até 53% do custo total da entrega [1]. Dessa forma, é importante o resolver esse problema de forma ótima para que as empresas de logísticas tenham um melhor aproveitamento do seu fluxo de caixa.

No entanto, não existe ainda uma base de dados de larga escala e considera o Brasil como sua área de estudo. Portanto, a Loggi criou o Loggibud Dataset que consiste em instâncias do problema de roteamento de veículos capacitado em 3 estados brasileiros: Pará, Distrito Federal e Rio de Janeiro. Cada instância possui cerca de 7 mil a 32 mil entregas. Ainda outras bases de dados consideram a distância euclidiana interferindo na aproximação da realidade da solução de instâncias do CVRP gerando perdas para modelos em produção. Assim, essa base considera a distância Haversine que considera a curvatura da Terra.

1.2. Abordagem do problema

Desse modo, o presente estudo utiliza da base Loggibud para aproximar a distância média percorrida por todos os veículos, utilizando *Machine Learning*. Essa abordagem é importante para planejamento de gastos da empresa, como, por exemplo, para tomada rápida de decisões sobre qual veículo alocar para as rotas específicas de um dia de entregas.

2. Métodos

Nesta seção serão apresentados os métodos utilizados para a modelagem do problema de otimização em um problema de *Machine Learning*. O projeto de ML possui a arquitetura como mostra a Figura 1.

O fluxo da arquitetura segue a seguinte estrutura: o cliente envia uma instância do problema formatado como esperado por meio de uma aplicação Web, detalhes de sua implementação será apresentado na seção 2.3. Essa aplicação irá predizer a distância com base em um modelo de ML apresentado na seção 2.2 e mostrará o resultado para o cliente. A construção do modelo possui duas etapas: a engenharia de *features* e o treinamento e seleção do modelo. A primeira etapa será abordada na seção 2.1 e a segunda etapa na seção 2.2

[1] https://fareye.com/resources/blogs/last-mile-delivery-problem-and-solution

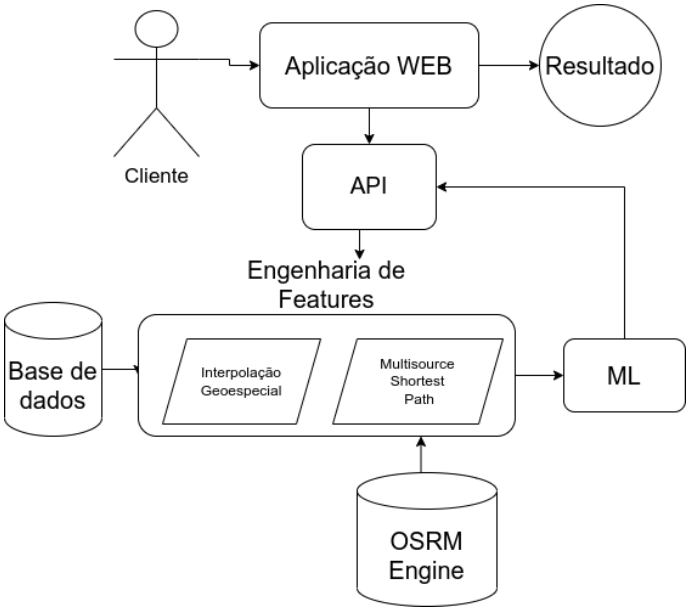


Figura 1. Arquitetura do projeto de ML

2.1. Engenharia de Features

É necessário realizar a engenharia de *features*, pois os dados do Loggibud não são estruturados em uma tabela, como um modelo de ML requer. Portanto, essa seção visa definir quais atributos foram criados tal que fosse possível o treinamento de um modelo de ML para aproximar a média da distância percorrida por todos os veículos no CVRP.

Os dados podem ser definidos como uma tupla: $\langle \text{nome}, \text{regiao}, \text{origem}, \text{capacidade}, \text{entregas} \rangle$, no qual o *nome* é o nome da instância, a *regiao* define qual o estado brasileiro que está sendo considerado, a *origem* é a localização do depósito, ou seja, possui latitude e longitude, a *capacidade* é a capacidade do veículo e o atributo *entregas* é uma lista de entregas, a qual possui diversas localizações com sua demanda.

A apresentação dos dados, pode ser vista de forma estruturada, porém, em um arquivo JSON, o que não pode ser uma entrada para qualquer modelo de ML. Dessa forma, foi necessário transformar essa base em outra base de dados tabular. Assim, cada instância de um estado tornou-se uma linha com os seus respectivos atributos, visto que é preciso resolver cada instância para se obter a distância média percorrida por todos os veículos. Portanto, a base para entrada do modelo pode ser definida como: $\langle \text{id}, \text{atributo 1}, \text{atributo 2}, \text{atributo 3}, \text{media da distância} \rangle$.

Para essa engenharia é necessário entender quais atributos podem influenciar essa aproximação. Supõe-se que a estrutura das ruas e estradas influenciam diretamente na velocidade do veículo ao entregar uma determinada demanda, assim como a característica social da área alvo de entrega de um veículo. Para isso foi preciso regionalizar as entregas para capturar essas informações de cada instância; esse processo será apresentado nas seções 2.1.1 e 2.1.2. Além da estrutura das regiões é necessário existir algum atributo relacionado com a distância percorrida dos veículos, processo que será apresentado na seção 2.1.3. Por fim, como o problema de CVRP possui a restrição da capacidade do

veículo, então supõe-se que algum atributo relacionado à demanda seja importante para a criação de um modelo de *Machine Learning* que tem como objetivo a aproximar a média das distâncias percorridas pelos veículos no CVRP.

2.1.1. Regionalização das Entregas

Neste trabalho será considerado apenas o estado de Pará, por termos de simplificação. Pará pode ser dividido em municípios, subdistritos e bairros, porém, possui o problema de ter uma grande granularidade, sendo necessário diminuir o tamanho da área do mesmo, visto que assim é possível calcular métricas como distância entre regiões de uma maneira mais realista. Para tanto, foi utilizado a indexação espacial H3, para regionalizar as entregas com célula de resolução igual a 6, que possui área igual à $36km^2$, assim como mostra a Figura 2. Note que essa regionalização é feita considerando as entregas para permitir que outros atributos que serão explicados sejam posteriormente agregados à base.

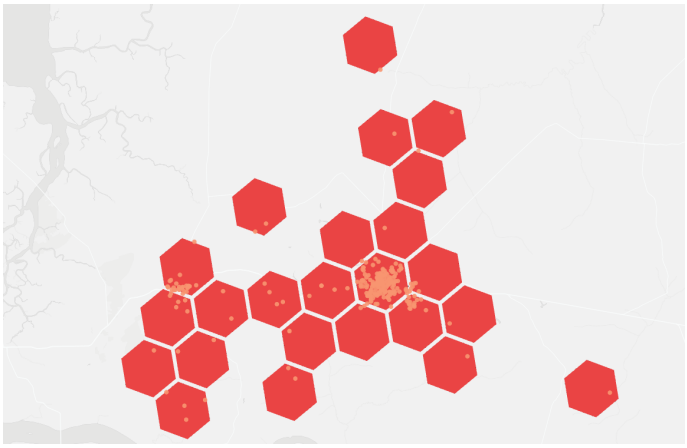


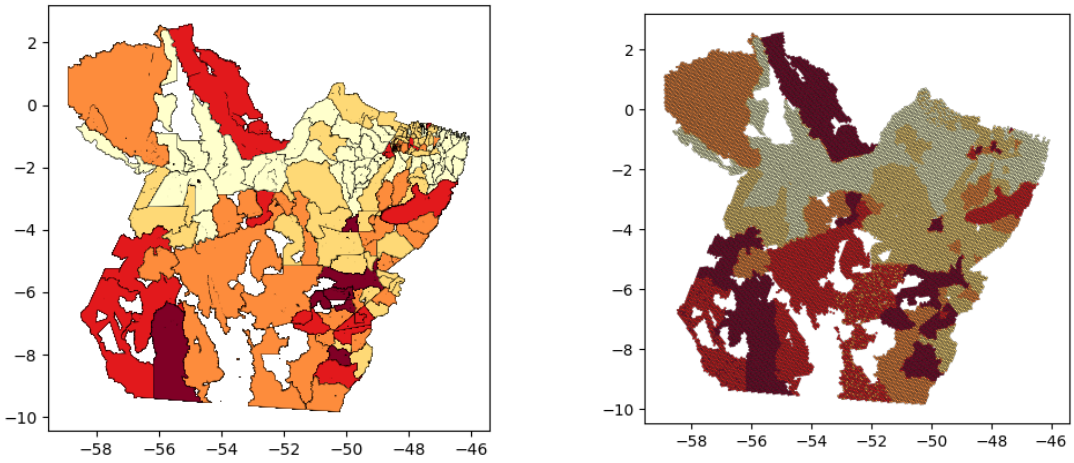
Figura 2. Indexação Espacial H3

2.1.2. Interpolação dos dados do CENSO 2010

Para atribuir características sociais às regiões definidas previamente, foi utilizado os dados do CENSO de 2010. Essa técnica permite que de forma latente seja adicionado atributos que influenciem o modelo a realizar essa tarefa considerando a estrutura da região. O intuito de agregar os dados do CENSO 2010 na base de dados foi adicionar a percepção da estrutura da rua, porém como não é possível obter esse dado facilmente, optou-se por utilizar pesquisas quantitativas de várias características sociais para suprir essa percepção. Assim, é possível considerar ainda outros atributos como demografia étnica e renda, que podem agregar o modelo positivamente.

No entanto, o CENSO 2010 é regionalizado por bairros, diferente da regionalização feita na seção anterior. Para resolver esse problema foi realizada a mesma indexação geoespacial utilizando o H3 com células de resolução igual a 6, para que existisse a junção entre as bases de dados e fosse possível agregar os atributos do CENSO às respectivas regiões.

Ainda assim, existe outro problema com essa abordagem. Agora, todas as regiões que pertencem a um bairro têm as mesmas características sociais, desconsiderando o fato de influência de regiões fronteiriças que podem ter valores diferentes. Esse problema também é conhecido como *Modifiable Areal Unit Problem* que significa que a mudança de escala pode impactar nas estatísticas geoespaciais. Portanto, para contornar esse problema foi utilizado a biblioteca *tobler*[2], que tem como princípio a interpolação de atributos, transformando atributos relacionados a uma unidade espacial em diferente níveis de agregações, como mostra a Figura 3.



(a) CENSO Pará **(b)** H3 Pará

Figura 3. Mudança do nível de granularidade

Note que a representação dos atributos em regiões fronteiriças se torna mais homogênea, o que era o esperado com essa solução.

Então são adicionadas à base de treino estatísticas descritivas dos atributos do CENSO para regiões que tenham pelo menos uma entrega, ou seja, cada característica do CENSO terá a média, mediana, variância, desvio padrão, máximo e mínimo considerando as regiões com pelo menos uma entrega em cada instância.

2.1.3. Métricas das Distâncias por Região

Como o objetivo do modelo é prever a distância média percorrida pelos veículos, é necessário agregar atributos que tenham correlação com a distância percorrida. Assim, foi calculado a matriz de Floyd-Warshall, matriz de distância mínima de todas as regiões para todas as outras regiões, através do OSRM Engine. Essa ferramenta é uma API que calcula as distâncias mínimas considerando o grafo das rodovias de algum estado, ou região, a partir de dados do *OpenStreetMap*. Assim, para calcular a matriz foi utilizado os centroides de cada região de entrega que tinha pelo menos uma entrega considerando apenas uma instância. Após o cálculo da matriz, a distância total também foi agregada em estatísticas descritivas: média, mediana, variância, desvio padrão, máximo e mínimo.

[2]pysal.org/tobler

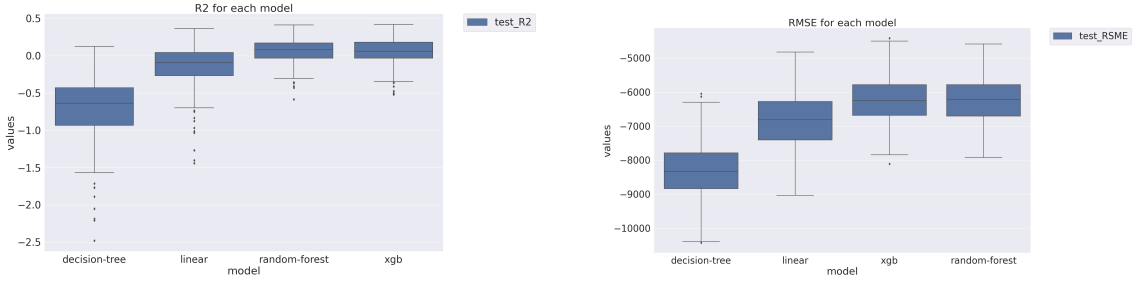
2.1.4. Métricas das Demandas

Além dos atributos de distância e de características sociais, foi adicionado ao modelo, estatísticas descritivas das demandas de cada entrega, ou seja, a média, mediana, variância, desvio padrão, máximo e mínimo do tamanho da entrega. Esse atributo se torna importante pelo fato de mapear uma restrição do problema formal para a base de dados tabular.

2.2. Modelo

Nesta seção será apresentado sobre a seleção do modelo e sobre a base de dados completa. A base de dados[3] é composta por 620 linhas e 163 colunas, sendo que uma é a variável alvo da média das distâncias percorrida pelos veículos. Note que para se obter essa média, é necessário resolver o problema CVRP; para isso foi utilizado o solucionador OR-TOOLS desenvolvido pela Google. Portanto, temos nessa base de dados, 620 instâncias de CVRP totalizando cerca de 19 milhões de entregas. Apesar de ser uma base com pouco volume de dados, essa base possui um custo computacional alto para sua construção.

Para o modelo, portanto, testou-se 4 modelos: Random Forest, Decision Tree, XGBoost e Regressão Linear, sendo que o XGBoost apresentou os melhores resultados para a validação cruzada com 10 *folds*, como mostra a Figura 4.



(a) R^2 **(b)** RMSE

Figura 4. Métricas para a Seleção do Modelo

Note que as métricas têm valores negativos, pois foram utilizadas como função de perda para a validação cruzada, assim quanto mais próximo de 0 melhor o modelo para quais quer métricas. Assim, o modelo escolhido para a implantação conforme as métricas mostradas foi o XGBoost.

A partir disso, foi feito o ajuste fino dos parâmetros do modelo com a intenção de melhorar o máximo possível. Assim, a melhor configuração dos parâmetros pode ser vista na tabela

Após o treinamento o modelo foi salvo em um arquivo *pickle*.

2.3. Implantação

Além do processo de engenharia de atributos, esse trabalho também contém o processo da implantação do modelo em uma aplicação web. Assim, para essa tarefa foi desenvolvida uma API utilizando o Flask e o Vue como framework web para a criação da interface

[3]https://www.kaggle.com/datasets/gegen07/cvrp-with-mean-distance

Parâmetro	Valor
subsample	0.75
n_estimators	600
min_child_weight	10
max_depth	50
learning_rate	0.01
lambda	1
gamma	2
colsample_bytree	0.3
alpha	0

Tabela 1. Parâmetros do XGBoost

amigável para o usuário enviar a instância para o modelo aproximar a média da distância percorrida pelos veículos. A Figura 5 mostra a interface web projetada para o envio de arquivos no formato esperado para a tarefa de perdição.

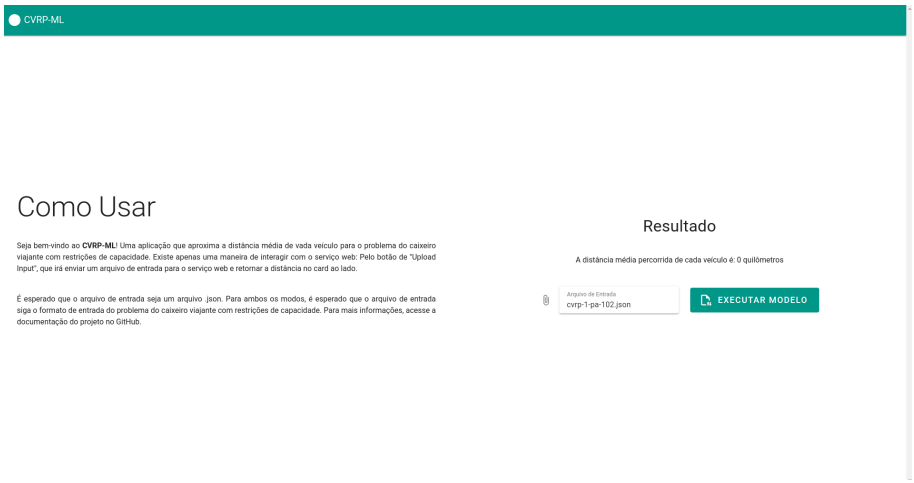


Figura 5. Aplicação Web

3. Resultados

Nesta seção serão apresentados os resultados do modelo, bem como, a importância das features da base de dados treinada.

3.1. Modelo

O modelo foi treinado utilizando uma divisão de 20% para teste e 80% para treino, sendo que na validação cruzada, utilizou-se somente os dados de treino. As métricas utilizadas para medir a qualidade da regressão foram o R^2 e o Mean Absolute Percentage Error (MAPE), para a base de treino o R^2 foi igual a 91.17%, e para a base de teste foi igual a 19.19%. O MAPE no treino foi igual a 2.14% e na base de teste 7.67%. Observando os resultados, é possível concluir que a base de dados de teste possui poucos registros para avaliação do R^2, pois o MAPE possui valores baixos para a base de teste. Logo, faz-se necessário resolver mais instâncias e agregar a base de treino para ter uma melhor interpretação do modelo.

Analisando as importâncias dos atributos na Figura 6, nota-se que os principais atributos são relacionados às distâncias mínimas das regiões. Além disso, existe uma importância relativa para o atributo mediana do número de indígenas, pois o estado do Pará possui população indígena principalmente na região norte do estado. Além disso, o desvio padrão das demandas também se mostra importante para o problema abordado nesse trabalho.

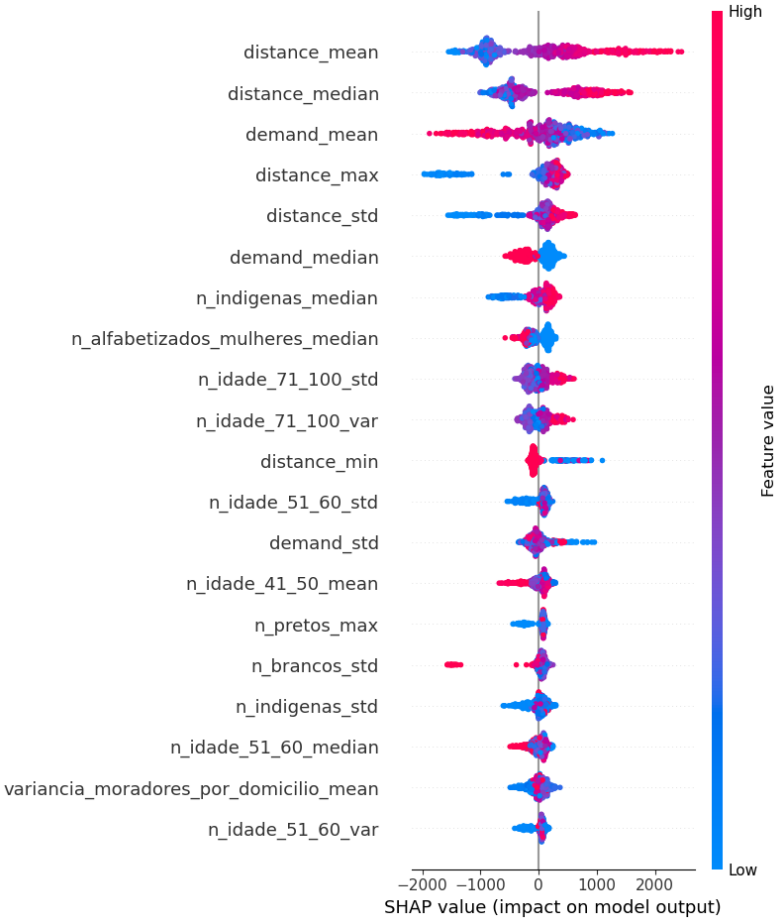


Figura 6. Importância dos Atributos

4. Conclusão e Trabalhos Futuros

Por fim, mostra-se importante a criação de atributos e o entendimento do problema para a predição ser feita corretamente e de forma otimizada.

Como trabalhos futuros, pretende-se adicionar *baselines* para comparar o modelo proposto com outras modelagens já existentes na literatura. Além disso, é preciso aumentar a base de dados, para ser possível uma melhor interpretação do desempenho do modelo nesse tipo de tarefa de regressão para problemas de otimização.