



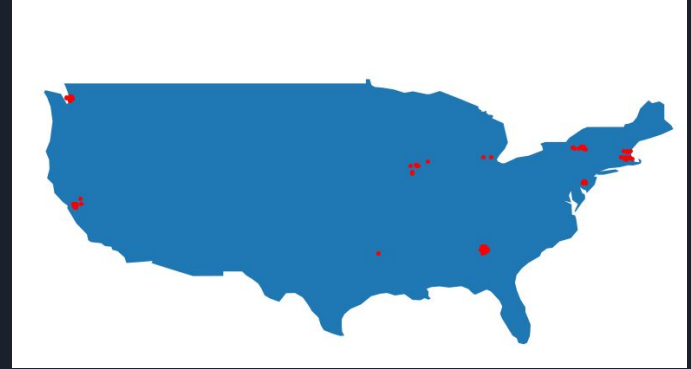
Objetivo do desafio

- Selecionar 3 Zip Codes para a instalação de novos laboratórios de uma rede de medicina diagnóstica dos EUA, por meio da análise de informações sobre os laboratórios atuais e dados demográficos, econômicos e de seguro de saúde do país.

Exploratory Data Analysis

Plot dos locais dos laboratórios atuais no mapa dos EUA

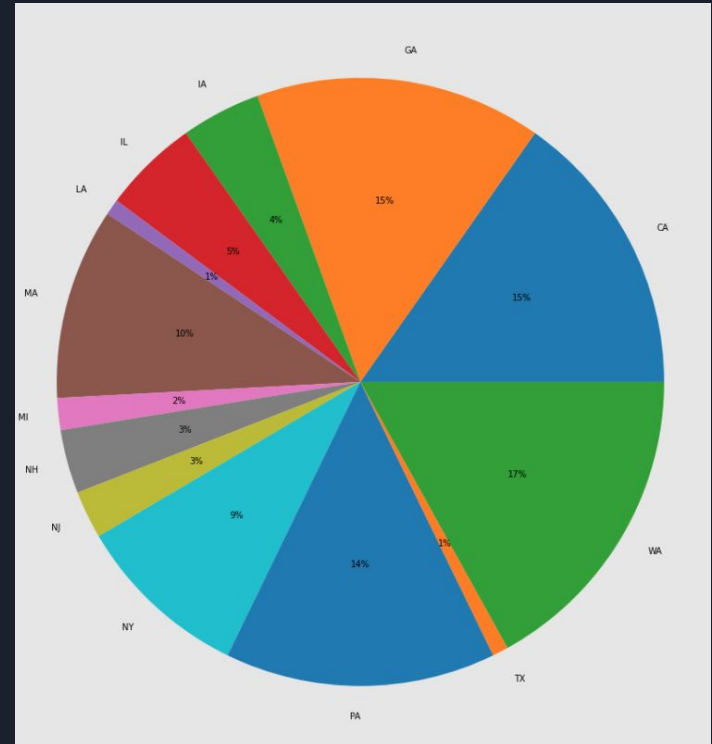
- Este plot inicial tem como objetivo, expressar visualmente o estado atual dessa rede de laboratórios e onde concentram suas instalações no país.



Exploratory Data Analysis

Instalações de laboratórios por estado em termos percentuais

- Para expressar quantitativamente as instalações de laboratórios por estado o gráfico ao lado foi plotado.
- É possível notar que apenas 3 estados são responsáveis pela maior parte da localização das instalações atuais.
- Para analisar melhor os resultados das instalações em cada estado, será feita uma estimativa de lucro dos laboratórios baseada na receita das transações e custos de realização dos exames.



Exploratory Data Analysis

Estimativa de Lucro por laboratório

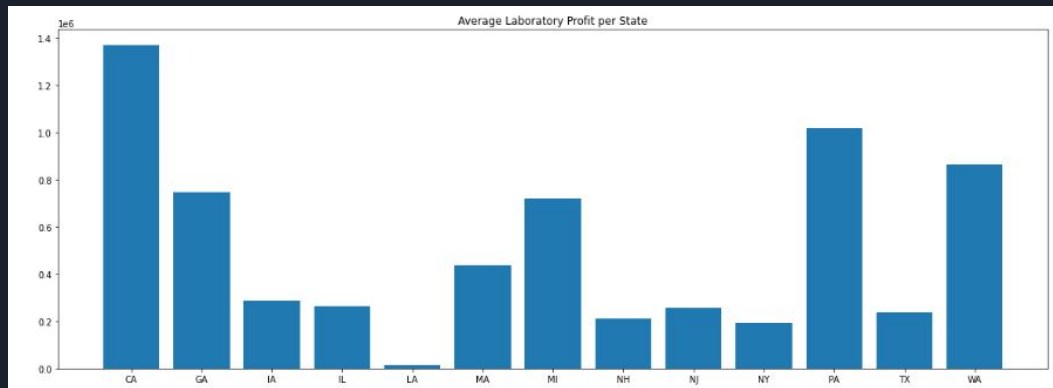
- Para estimar o lucro de cada laboratório foram usados os dados de transação realizada e de custo por exame, os dados foram adicionados a uma tabela local para agregar custo e receita por laboratório. Essa informação será usada para estimar o lucro de uma instalação de laboratório em um local arbitrário usando métodos estatísticos.

	Patient Id	Gender	Date of birth	Date of service	Service Id	Lab Id	Coditem	Testing Cost	exam_cost
0	10210830256-BIO003	F	01/08/1976 00:00:00	2019-01-07	571904533475-38	L133	70003237	9.0	1.78
1	10210830256-BIO003	F	01/08/1976 00:00:00	2019-01-07	571904533475-38	L133	70000638	13.0	2.46
2	10210830256-BIO003	F	01/08/1976 00:00:00	2019-01-07	571904533475-38	L133	70001597	49.0	2.11
3	10210830256-BIO003	F	01/08/1976 00:00:00	2019-01-07	571904533475-38	L133	70000103	11.0	0.80
4	10210830256-BIO003	F	01/08/1976 00:00:00	2019-01-07	571904533475-38	L133	70000224	10.0	1.02
...
2355236	7664157546-1	M	06/03/1971 00:00:00	2021-02-12	7664157546-1-1	L697	70004038	10.0	1.37
2355237	7664157546-1	M	06/03/1971 00:00:00	2021-02-12	7664157546-1-1	L697	70004134	10.0	0.95
2355238	7664157546-1	M	06/03/1971 00:00:00	2021-02-12	7664157546-1-1	L697	70003056	9.0	1.12
2355239	7664157546-1	M	06/03/1971 00:00:00	2021-02-12	7664157546-1-1	L697	70004185	13.0	2.39
2355240	7664157546-1	M	06/03/1971 00:00:00	2021-02-12	7664157546-1-1	L697	70000392	8.0	1.43

Exploratory Data Analysis

Média de lucro de um laboratório por estado

- O gráfico abaixo expressa a média do lucro de um laboratório por estado, essa análise preliminar permite visualizar quais localidades são mais propícias para instalação de um novo laboratório, sem considerar os dados demográficos.





Exploratory Data Analysis

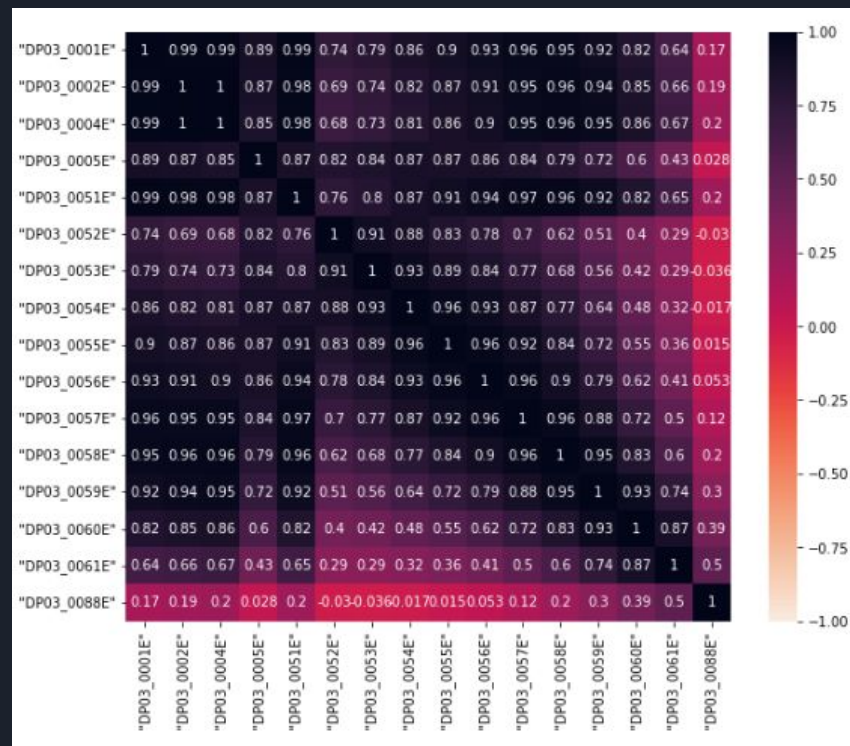
Indicadores econômicos

- Houve uma seleção inicial de indicadores: dados estratificados por estimativa da renda média por ZCTA5, a população maior de 16 anos empregada e desempregada e a renda per capita em dólares.
- Esses itens foram selecionados em detrimento dos outros devido a forma geral em que eles podem descrever o poder socioeconômico da região e a força laboral de forma simplificada.
- Considerar essas características advém da intuição de que a medicina diagnóstica, por ter um caráter preventivo, pode ser procurada por pessoas que dispõem de recursos para tratar da saúde fora de um regime de urgência.

Exploratory Data Analysis

Indicadores econômicos

- O plot ao lado, da correlação de Pearson das variáveis selecionadas, é usado para identificar relações lineares entre as variáveis observadas. Essa análise influenciará a escolha de um modelo preditivo (linear ou não linear) e a seleção de novos marcadores, combinando as variáveis para serem expressas linearmente.






Exploratory Data Analysis

Informações de seguro de saúde

- Quanto às informações disponíveis de seguro de saúde, foram selecionados marcadores, estratificados por idade e pela população não institucionalizada assegurada ou não por algum seguro de saúde.
- Como o seguro de saúde cobre testes diagnósticos, esses marcadores podem indicar uma população inclinada ou não a procurar serviços de medicina diagnóstica, uma vez que poderiam usufruir dos serviços sem despesas extras.



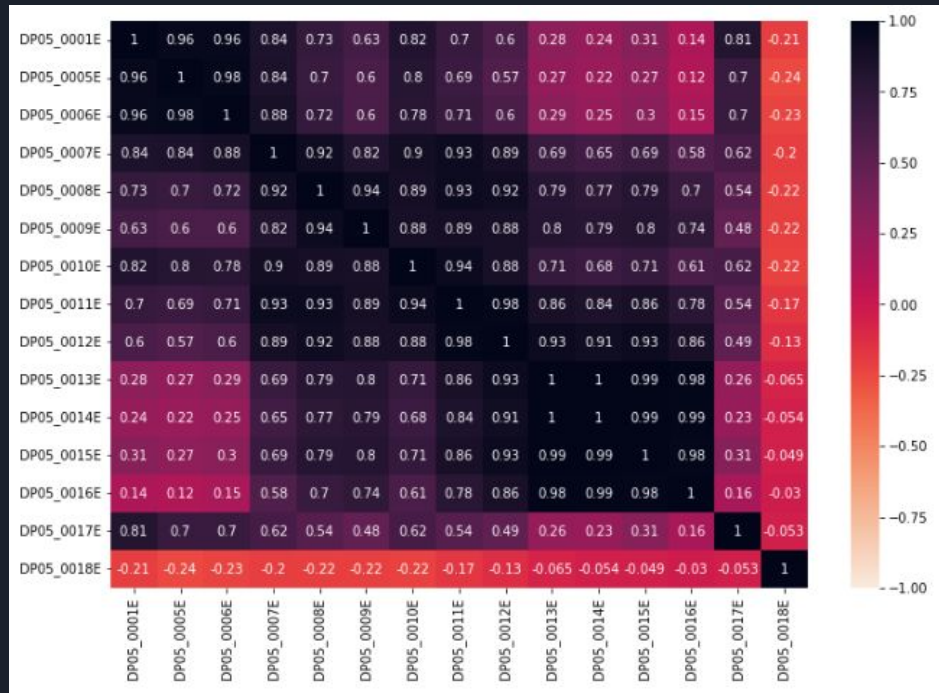
Exploratory Data Analysis


Indicadores demográficos

- Para os indicadores demográficos foram selecionados os indicadores, estratificados por idade da população total de cada região e a população total.
- Esses dados são importantes para considerar, em magnitude, a possível demanda por serviços de saúde e quais faixas etárias procuram mais serviços de medicina diagnóstica.

Exploratory Data Analysis

Indicadores demográficos





Solução proposta - Parte 1

Abordando o problema como uma regressão

- Parte da solução proposta para este problema é usar os marcadores sociais, demográficos e econômicos para estimar o lucro de um laboratório instalado em uma ZCTA.
- Com o lucro aproximado dos laboratórios existentes, usando as informações de custo e de transações, e sua localidade expressa na forma de código postal, é possível inferir o lucro de um novo laboratório e utilizar essa nova marcação como critério inicial para a seleção de localidades.
- Para cada região de laboratório a estimativa de lucro foi usada como um dado anotado para o treinamento dos modelos estatísticos de predição.



Solução proposta - Parte 1

Feature Engineering

- Os dados tabulados foram convertidos em informação numérica e foram aplicadas duas técnicas importantes para melhorar a predição, principalmente de modelos lineares: Standardization e Feature Crossing.
- Standardization escala cada uma das features selecionadas para uma distribuição normal com média igual a zero e desvio padrão igual a um. Esse processo é importante para diversos algoritmos como Ridge Regression, Lasso, Deep Learning, que aprendem pesos a serem aplicados aos dados, pois diminui significativamente a variação de cada feature fazendo com que os pesos também sejam escalados com pouca variação e convergindo mais rápido para um valor ótimo.
- Feature Crossing é usado para combinar features lineares para que os modelos estatísticos lineares possam capturar a relação entre elas.

Solução proposta - Parte 1

Feature Engineering

Add Polynomial features

```
In [40]: 1 from sklearn.preprocessing import PolynomialFeatures
          2
          3 poly = PolynomialFeatures(2)
          4 poly.fit(X)
          5 X = poly.transform(X)
```

```
In [41]: 1 X.shape
```

```
Out[41]: (116, 2556)
```

Standardize Features

```
In [42]: 1 from sklearn.preprocessing import StandardScaler
          2
          3 scaler = StandardScaler()
          4 scaler.fit(X)
          5 X = scaler.transform(X)
          6 X
```

```
Out[42]: array([[ 0.          ,  2.84455043,  2.73149481, ...,  8.7470147 ,
                  9.09742163,  8.09214288],
                [ 0.          ,  0.39461191,  0.41111767, ..., -0.29456561,
                 -0.30621026, -0.35032714],
                [ 0.          ,  0.51969241,  0.502884  , ..., -0.22937385,
                 -0.2142859 , -0.27992208],
                ...,
                [ 0.          , -1.61100321, -1.30753455, ..., -0.36992302,
                 -0.3237438 , -0.35386545],
                [ 0.          , -0.25324092,  0.04274141, ..., -0.32031449,
                 -0.25260883, -0.2654078 ],
                [ 0.          ,  0.27950935,  0.50616137, ...,  0.40547225,
                 0.11008087, -0.14330014]])
```



Solução proposta - Parte 1

Feature Selection

- Para estimar quais features são mais importantes para a solução do problema foi usado uma técnica de eliminação de features que não contribuíram com a solução.
- Esse processo é importante devido ao número reduzido de dados anotados (laboratórios já instalados e sua estimativa de lucro) , pois, ao reduzir o número de variáveis do problema, menos informação precisará ser aprendida pelo modelo escolhido. Assim é possível generalizar melhor pra novos casos, evitando o fenômeno de overfitting, quando os dados se adequam apenas aos dados em que foram treinados e não se generalizam para novos pontos porque o modelo é complexo demais.



Solução proposta - Parte 1

Feature Selection

Feature Selection

```
In [43]: 1 from sklearn.feature_selection import RFECV
          2
          3 # Number of samples ~= Number of parameters * 10
          4 MIN_FEATURES = 10
```



Solução proposta - Parte 1

Metrics Selection

- Duas métricas foram selecionadas para comparar as diferentes abordagens para resolver o problema de regressão: RMSE e R2 Score
- RMSE (Root Mean Squared Error) : Esse valor metrifica a distância entre o valor predito pelo modelo e o valor real. Por ser obtido como a medida do erro ao quadrado, erros grotescos são penalizados por essa avaliação. Isso é importante para esta análise devida a quantidade limitada de dados anotados que pode não expressar corretamente a distribuição de todos os dados da inferência. No geral, quanto menor o RMSE melhor o modelo.
- R2 Score: Essa métrica representa o coeficiente de quão bem os valores se comparam com os valores originais. Então um modelo com R2 Score de 0.5 pode explicar aproximadamente metade das entradas utilizadas.



Solução proposta - Parte 1

Model Selection

- Para a seleção do modelo, foram utilizados diversos paradigmas para regressão como modelos lineares, conexionistas, bayesianos e de árvore de decisão.
- Os algoritmos foram comparados pela performance nas métricas na média do conjunto de teste dos dados dividido em 5 grupos mutuamente exclusivos. Essa técnica é chamada de validação cruzada (Cross Validation) e é utilizada principalmente em conjunto de dados limitados para averiguar melhor a capacidade do modelo de generalizar os resultados para dados novos.

Solução proposta - Parte 1

Model Selection

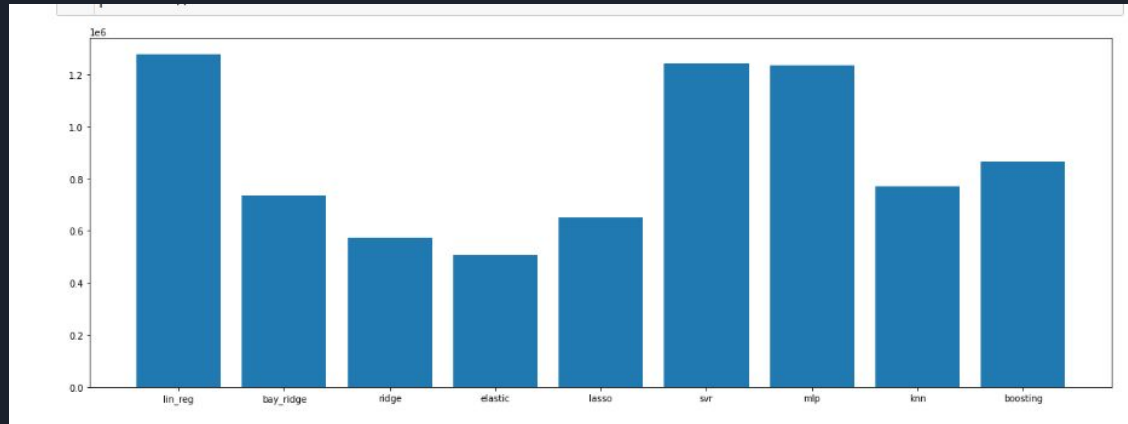
Model Selection

```
] : 1 from sklearn.model_selection import KFold
    2 from sklearn.linear_model import LinearRegression, BayesianRidge, Ridge, ElasticNet, Lasso
    3 from sklearn.svm import LinearSVR
    4 from sklearn.neural_network import MLPRegressor
    5 from sklearn.neighbors import KNeighborsRegressor
    6 from sklearn.ensemble import GradientBoostingRegressor, StackingRegressor
    7
    8 models = {
    9     'lin_reg':LinearRegression,
   10     'bay_ridge':BayesianRidge,
   11     'ridge':Ridge,
   12     'elastic':ElasticNet,
   13     'lasso':Lasso,
   14     'svr':LinearSVR,
   15     'mlp':MLPRegressor,
   16     'knn':KNeighborsRegressor,
   17     'boosting':GradientBoostingRegressor
   18 }
   19
   20 n_splits = 5
   21
   22 kf = KFold(n_splits=n_splits)
```

Solução proposta - Parte 1

Model Selection

- O desempenho dos modelos lineares, considerando RMSE, foi o melhor. O método de ElasticNet que combina Ridge e Lasso regression obteve o melhor resultado.



Solução proposta - Parte 1

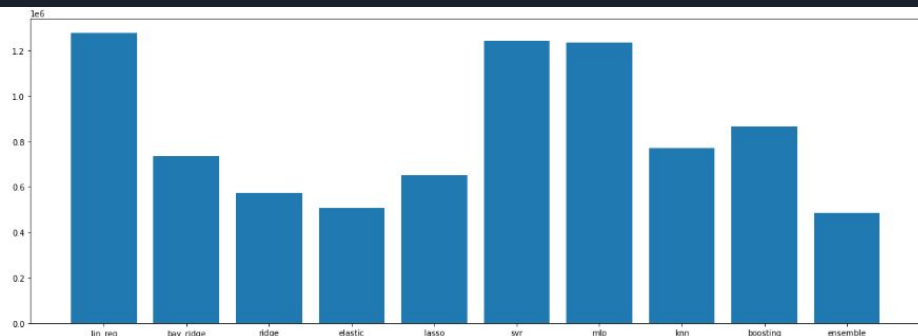
Model Selection

- A performance foi drasticamente aumentada utilizando o método de Ensembling para combinar as predições dos melhores modelos de cada paradigma em uma predição final, obtendo resultados muito competitivos.

Ensembling best methods from distinct paradigms

```
1 model_name = 'ensemble'
2 estimators = [
3     ('knn', KNeighborsRegressor()),
4     ('boosting', GradientBoostingRegressor()),
5     ('elastic', ElasticNet()),
6     ('bay_ridge', BayesianRidge()),
7 ]
```

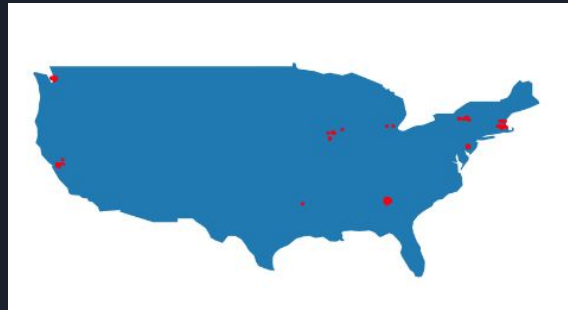
```
{'model_name': 'ensemble',
 'selector': RFECV(cv=5, estimator=ElasticNet(), min_features_to_select=10, step=30),
 'rmse': 484277.385718098,
 'mae': 309461.08104919066,
 'r2': 0.7181914509396776}
```



Solução proposta - Parte 1

Inferência

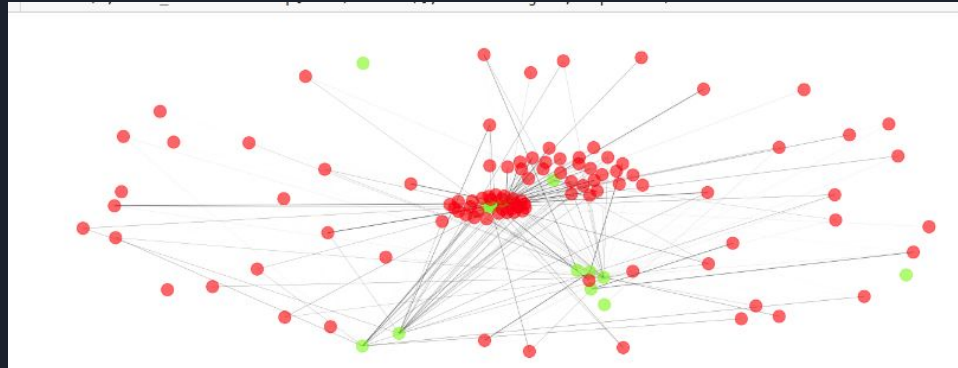
- Ao lado, o mapa com pontos verdes marcam as regiões divididas por ZCTA5, que apresentaram as 100 maiores estimativas de lucros na instalação de um laboratório da rede de medicina diagnóstica.
- É possível notar a correlação visual entre as regiões indicadas em verde e em vermelho, principalmente em estados como Califórnia e a baía de Nova York. Isso pode indicar que a predição está assertiva em relação aos estados mais lucrativos.



Solução proposta - Parte 2

O problema de recursos iniciais para expansão como uma otimização de grafos

- Como exposto na descrição do problema, é importante considerar questões como o abastecimento das novas instalações, para tanto foram selecionados os 15 pontos mais lucrativos de acordo com as estimativas e o problema foi remodelado para selecionar os nós mais relevantes de um grafo.
- Latitude e longitude foram transformadas para coordenadas cartesianas e utilizadas para calcular o peso das arestas entre os candidatos (verde) e os laboratórios (vermelho), quanto maior o peso, mais perto de um laboratório.



Solução proposta - Parte 2

O problema de recursos iniciais para expansão como uma otimização de grafos

- Utilizando um método de PageRank para calcular a importância de cada nó na rede, medido pela conectividade e força das arestas, foram selecionados 3 nós representando os candidatos mais importantes para a rede, ou seja, mais próximos de mais laboratórios. Mitigando assim o problema de desabastecimento.



60629 - Chicago, IL

47906 - West Lafayette, IN

11219 - Brooklyn, NY



Limitações

- A abordagem selecionada foi baseada numa estimativa de lucro enviesada, que não considera fatores como os diferentes custos para cada estado de acordo com as leis vigentes.
- Ademais devido ao limite na quantidade de exemplos rotulados, a estimativa pode apresentar alta variância para inferência uma vez que é possível que mesmos divididos em 5 grupos de validação cruzada os exemplos tenham distribuição diferente do resto dos dados.