CORONATHON
Equipe 28

Participantes: Allan Rogge, Érico Encarnação, Ivan Ormenesse, Jonas Lima, Maurício Trujillo, Rodrigo Figueira

PROJETO: PROPOSTA DE API PÚBLICA BASEADA EM MODELAGEM ESTATÍSTICA PARA MELHORIA DE ASSERTIVIDADE DO SINE

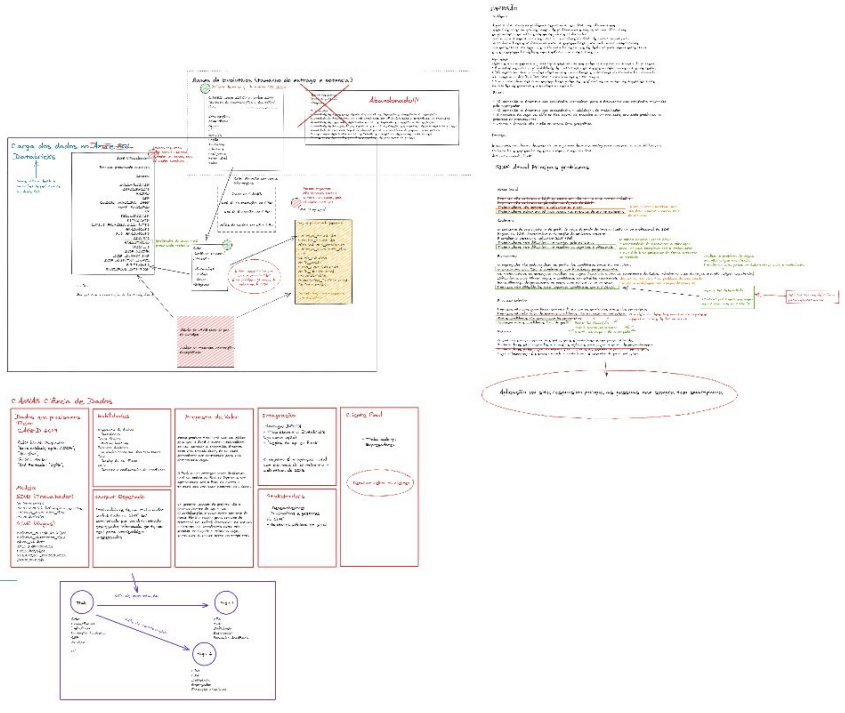
Resumo do Projeto

Nosso produto final será uma API pública pela qual o SINE e, eventualmente, outros stakeholders possam acessar a propensão, descrita como uma probabilidade de um trabalhador ser contratado para uma determinada vaga. Nossa missão é reduzindo o atrito e agilizar o processo de contratação para trabalhadores e empregadores, melhorando a visualização de vagas no front-end e facilitando encontrar pessoas e posições, inclusive aquelas portadoras de deficiência.

Etapa de Imersão

Iniciamos o entendimento do problema discutindo detalhadamente os principais problemas apresentados pelo SINE e, a partir da explicação das bases apresentadas nas primeiras etapas do evento, escolhemos alguns problemas para resolver.

Documentamos todo o processo por meio do Excalidraw gerando tanto uma imagem em PNG como um arquivo JSON que detalha não apenas o processo de Imersão, mas também a ideação, prototipação e a solução proposta.



Além disso, fizemos algumas análises descritivas com o propósito de levantar e validar algumas hipóteses que se mostraram valiosas para a etapa de ideação.

Etapa de Ideação

Na Etapa de ideação abandonamos algumas ideias que haviam surgido anteriormente e desenhamos a solução a ser desenvolvida nas etapas futuras.

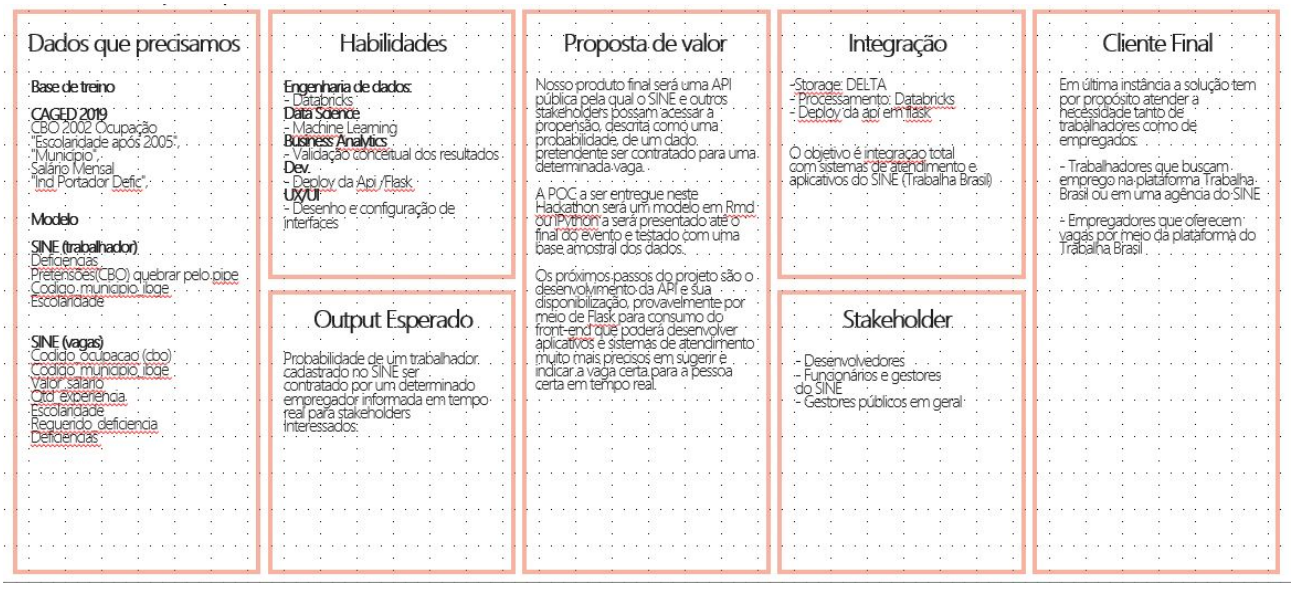
Ideias adotadas

- Modelo preditivo por área geográfica e ocupação para a predição à partir da base do CAGED
- Utilização do Databricks versão comunidade para lidar com o volume de informações
- A entrega final seria uma API que consulta um algoritmo de priorização de vagas e entrega resultado para trabalhadores, empregadores e stakeholders

Ideias abandonadas

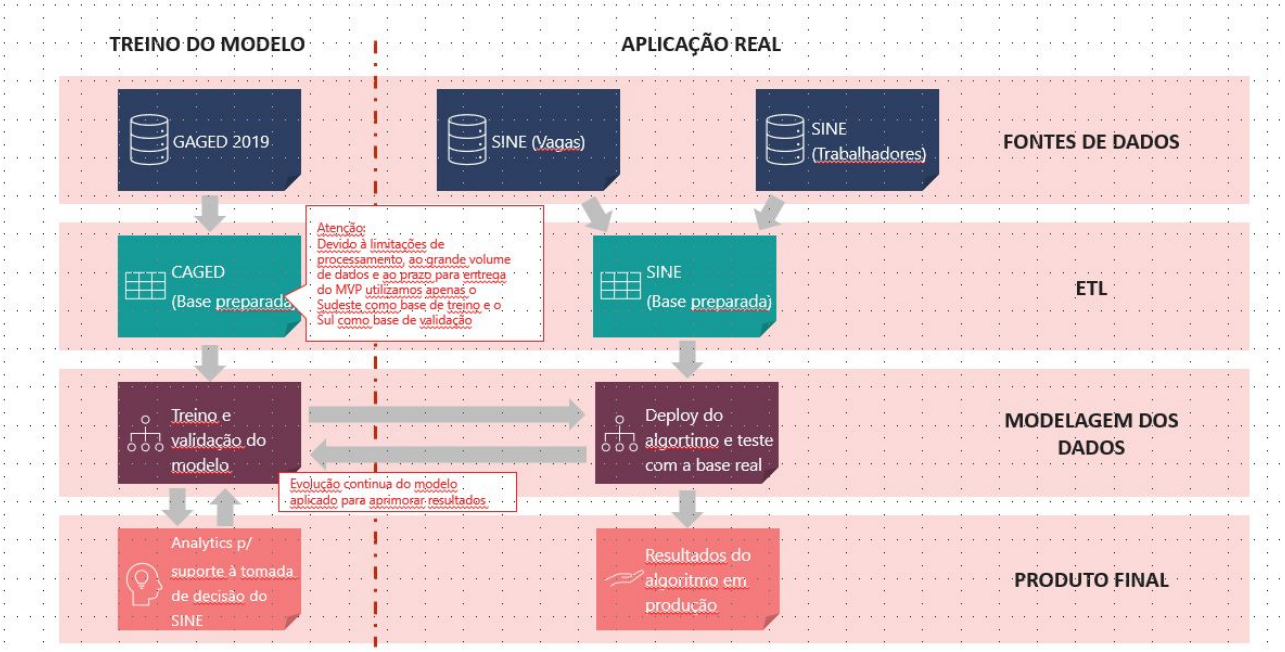
- Informações sobre o seguro desemprego, por serem dados agregados não contribuem para construção do modelo
- Base da RAIS: entendemos que o CAGED de 2019 possuia as informações
- necessárias para a modelagem
- BASE CAGED 2020: Poucas informações e dados fortemente impactados pela Covid-19
- Utilização das bases do SINE tanto de vagas como de trabalhadores para treinar um modelo de machine learning, pois, devido à presença de informações sensíveis, o volume de dados disponibilizados é insuficiente

Além disso, produzimos o seguinte canvas com o detalhamento de toda a solução:

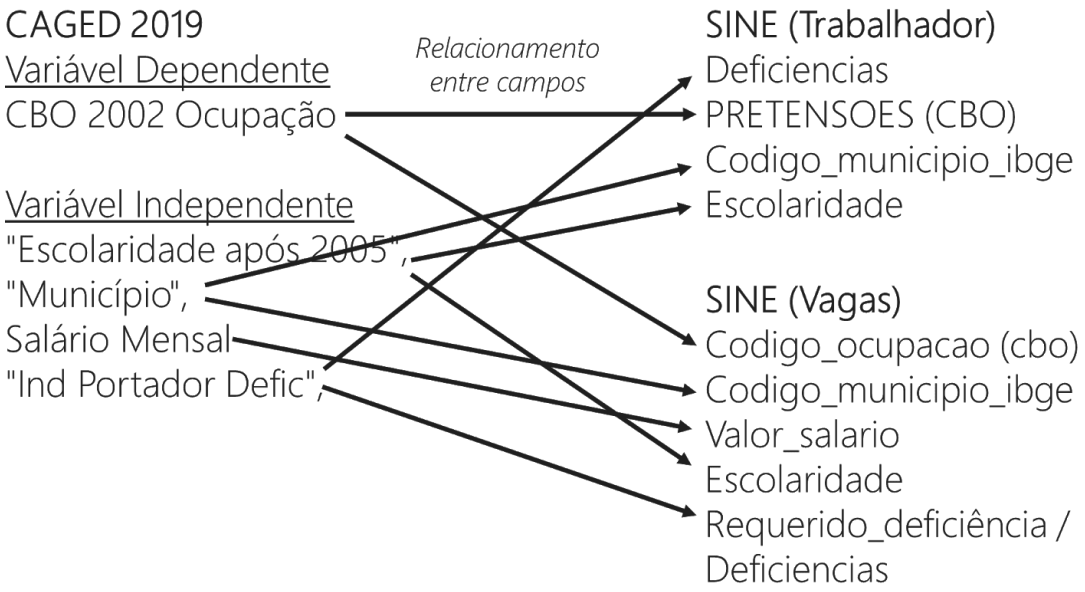


Etapa de Prototipação

Partimos então para a etapa de prototipação, desenhando como a solução funcionaria e especificando como a API se conectaria ao front-end do SINE.



Também descrevemos um de-para das informações disponíveis tanto nas bases do SINE com nas bases do CAGED para facilitar o processo de tratamento e modelagem dos dados. Haja visto que o volume de dados nos obrigou a filtrar as informações por UF e também para os 30 CBOs mais representados na base:



O algoritmos será construído com base nas informações da base do CAGED e será treinado com um grande volume de informações para que, com base nas informações também disponíveis na base do SINE possam alimentar o modelo em tempo real e responder a probabilidade de contração para uma dada relação colaborado-vaga.

O modelo será treinado com base nos dados de São Paulo da base da CAGED, validado com base nos dados do Sul e aplicado aos dados disponibilizados pelo SINE. Devido à necessidade de reduzir as informações disponibilizadas no modelo, não foi possível utilizar a variável município no MVP.

O modelo

Aplicou-se o modelo de Regressão Logística para cada CBO da base de demitidos e admitidos do CAJED, para a região de São Paulo. Foram selecionados 30 CBOs para um primeiro experimento, devido capacidade de servidor e espaço para treinamento do modelo.

Foram gerados 30 modelos, um para cada CBO, com o objetivo de descobrir a probabilidade que cada trabalhador possui em relação ao CBO que ele se identifica. Com isso, foi calculada a acurácia de cada modelo, escolhendo como critério a probabilidade de 0,2 para o sucesso da ocorrência.

As acurácias oscilaram e conseguiram alcançar poder de predição de até 95%. Deseja-se melhorar o ajuste do modelo a fim de encontrar os melhores pontos de cortes para cada CBO e consequentemente melhorar o poder de predição.

Além disso, deseja-se aplicar o modelo para a base nacional do CAGED e para a base de indivíduos que procuram vagas de emprego do SENE

Etapa de Solução

Com base no detalhamento do modelo e conceito de integração e retro alimentação das bases atacamos a frente de disponibilização das informações via API. Para tanto idealizamos um processo, ainda à ser desenvolvido, via Flask para envio e recebimento de dados.

Nesta abordagem, o Flask irá realizar um túnel de API para conectar a base de dados ao nosso modelo e ao front-end do Trabalha Brasil.

Ao realizar o cadastro, será feito um post em direção a base, para inserir os dados via JSON na base de dados.

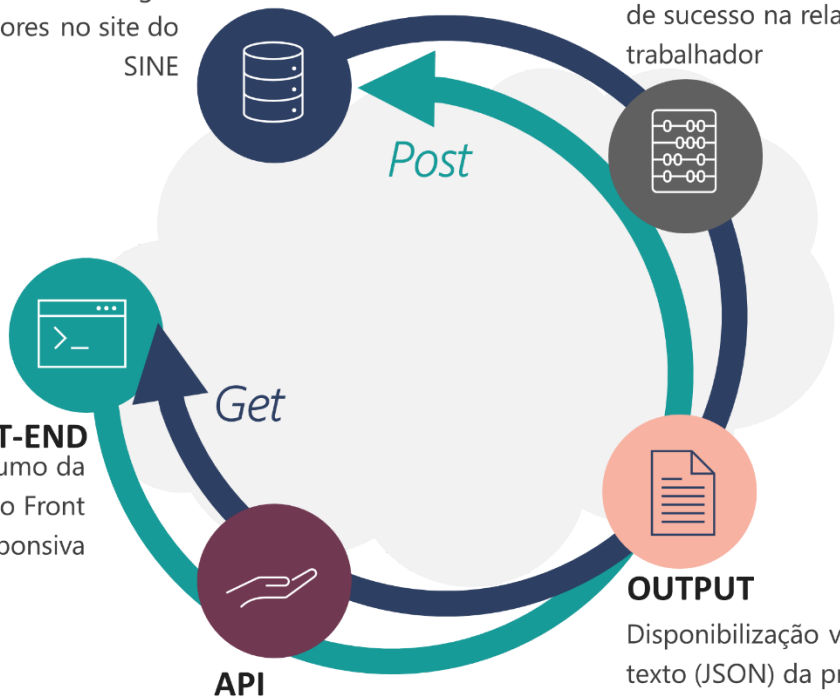
Ao solicitar uma procura de emprego, um método ‘get’ consome a informação do modelo e outro ‘get’ exibe a informação na tela.

BASES SINE

Bases com cadastro de vagas e trabalhadores no site do SINE

ALGORITMO

Cria o score de probabilidade de sucesso na relação vaga-trabalhador



FRONT-END

Consumo da informação pelo Front via página responsiva

OUTPUT

Disponibilização via arquivo de texto (JSON) da priorização

API

O Flask disponibiliza uma rota de conexão com o front-end do Trabalha Brasil