Combinatron

O impacto da COVID-19 no mercado de trabalho brasileiro mudou a vida de diversas pessoas. O índice de desemprego cresceu nos últimos meses e o cenário de incerteza pode piorar essa estatística. Existem setores que vão sofrer mais impacto do que outros, nesse momento é importante rearranjar os funcionários para minimizar a evasão do mercado de trabalho.

O Combinatron é um sistema que rastreia as demissões e conecta os candidatos para novas oportunidades de trabalho que melhor se adequam ao seu perfil. Utilizando fontes de dados públicas como o SINE o sistema consegue identificar as características do candidato e utilizando algoritmos de aprendizado de máquina faz o *match* do usuário com a vaga que melhor se encaixa ao seu perfil.

O sistema foca em dois perfis de usuários: a empresa em busca de profissionais que melhor se adequam aos seus requisitos; e o candidato que buscam se manter ou adentrar o mercado de trabalho. As jornadas dos dois tipos de usuários podem ser observadas nas imagens a seguir:

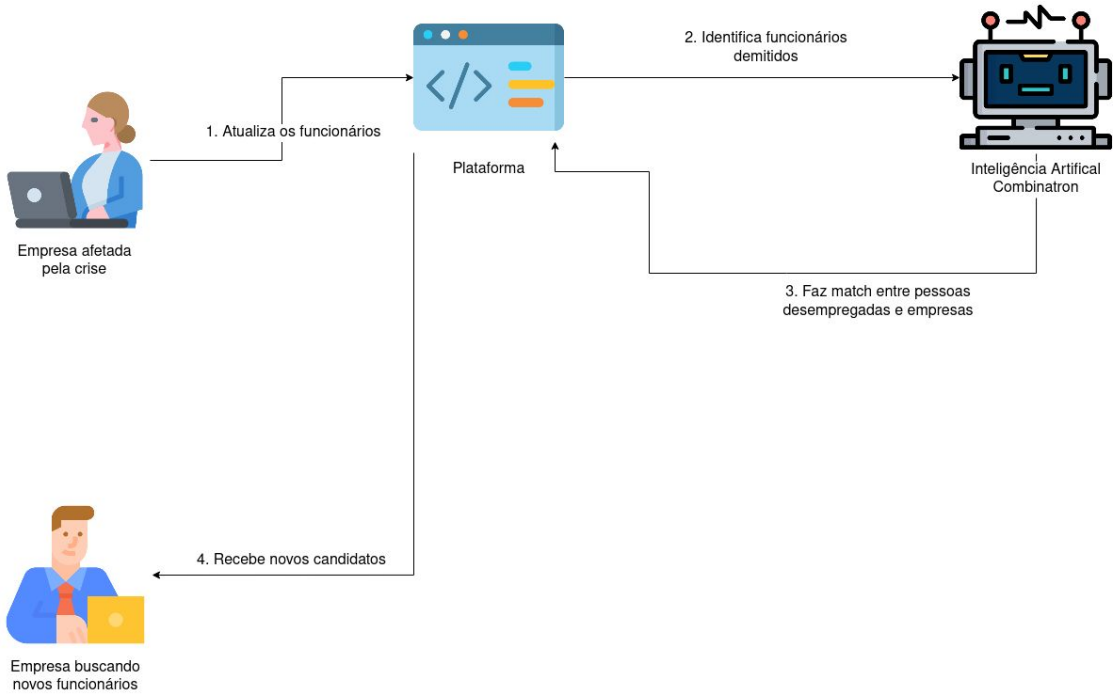

Figura 1: Jornada das empresas

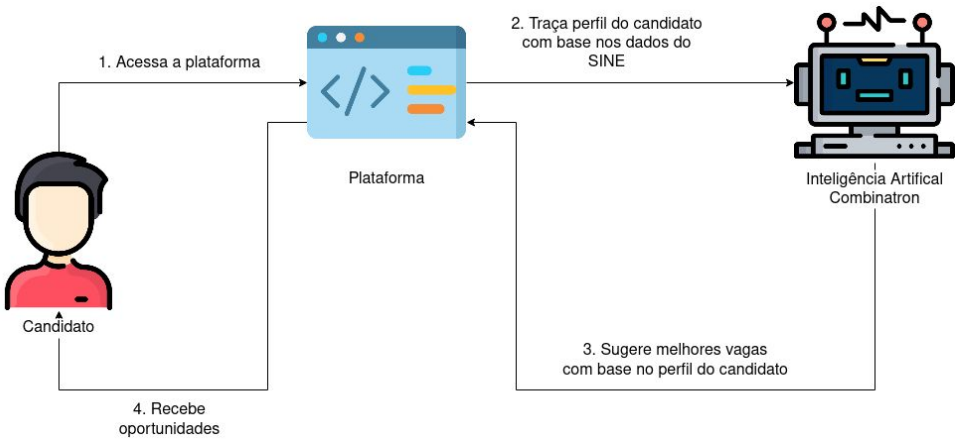

Figura 2: Jornada dos candidatos

Resultados

Matching

O matching do Combinatron utiliza por baixo um sistema de recomendação. O mecanismo de recomendação funciona usando uma lógica similar aos sistemas de recomendação do Netflix, e Youtube. Nesse modelo, o match é feito quando o perfil do trabalhador é similar a de outros trabalhadores que ingressaram na vaga, ou em vagas similares, garantindo que serão recuperadas sugestões de posições mais adequadas para o perfil do trabalhador.

De maneira similar, o empregador receberá os trabalhadores mais ‘adequados’ para sua posição, com base em escolhas anteriores e similaridades entre sua posição e outras. Além disso, é utilizada uma lógica de localização para aumentar a recomendação de trabalhos mais próximos, de forma a reduzir o deslocamento do trabalhador.

Dados utilizados

Para realização do desafio, foram utilizados os dados do sistema SINE, tanto de trabalhadores quanto para posições. O número de posições, como esperado, é muito menor que o número de trabalhadores disponíveis, por essa razão, com objetivo de balancear o dataset, utilizamos técnicas de Data Augmentation com Mocking e reprodução dos dados do SINE.

Apesar desses dados utilizados, sugerimos que seja usada a API do Linkedin para melhor matching dos dados, a API do Linkedin é acessível mediante autenticação e cadastro de app. Esse cadastro não foi possível no período da hackathon, entretanto é o mais indicado para quando a aplicação for colocada em produção.

Ferramentas

A solução foi desenvolvida com ferramentas de Data Science em linguagem Python:

- Pandas - exploração dos dados e feature engineering, quando o volume era pequeno
- Dask - manipulação dos dados em alto desempenho, quando o volume aumentou
- Surprise - biblioteca de Machine Learning com diversos mecanismos de recomendação. Utilizada para fazer benchmark do mecanismo com menor “erro”, e seleção do modelo
- Faker - biblioteca de geração de dados ‘mockados’, foi usada em conjunto com técnicas manuais para geração dos dados de posições.

Metodologia

Após a coleta dos dados de trabalhadores e posições, usamos um conjunto de regras para etiquetar automaticamente o match entre eles, entre elas:

- Aumenta os pontos caso a escolaridade exigida na posição seja atendida pelo trabalhador.
- Aumenta os pontos caso a graduação da posição esteja contida nas graduações do trabalhador;
- Aumenta os pontos caso o trabalhador esteja na mesma cidade da posição.
- Aumenta os pontos caso alguma pretensão de tipo de posição (ex: operador de caixa, auxiliar de escritório) do trabalhador seja igual ao tipo de posição da vaga.

Após esse cruzamento, unimos os dados em uma tabela única, com o id do trabalhador, o id da posição e a pontuação de match entre eles, esses dados foram usados para alimentar o modelo de Machine Learning.

Para o mecanismo de recomendação foi utilizado o algoritmo KNNBasic da biblioteca Surprise. Esse mecanismo utiliza a técnica de filtro colaborativo (https://surprise.readthedocs.io/en/stable/knn_inspired.html#surprise.prediction_algorithms.knns.KNNBasic), associado a uma métrica de MSD (Mean Squared Difference - https://surprise.readthedocs.io/en/stable/similarities.html#surprise.similarities.msd). Com isso, é possível obter o matching de novos usuários e novas posições em tempo real, sem necessidade de re-treino do modelo de Machine Learning, utilizando os que já existem no sistema, mesmo que não hajam muitos já cadastrados.

Entendemos que isso é estratégico para o SINE, pois dessa forma o time pode iniciar imediatamente o matching de novas posições e trabalhadores.

Código fonte

O código fonte do resultado apresentado pode ser encontrado no repositório: https://github.com/equipepontozip/coronathon.

Proposta de arquitetura

A seguir pode ser observado um diagrama com a proposta de arquitetura para a solução completa do sistema Combinatron:

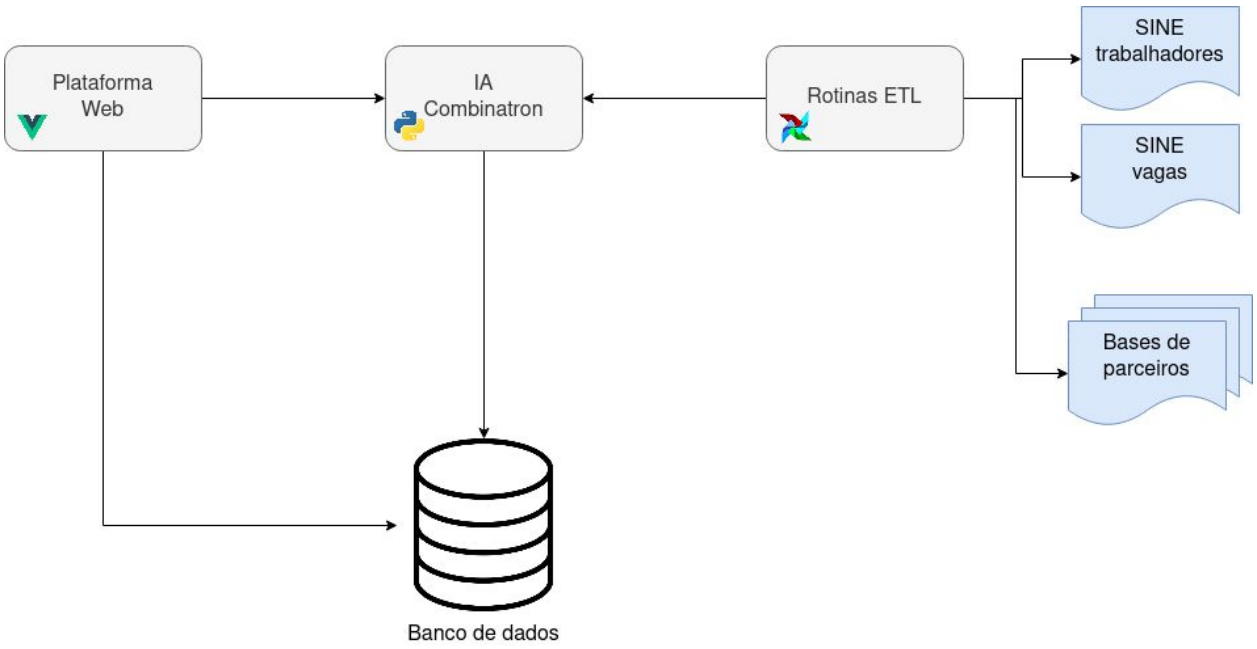

Figura 3: Proposta de arquitetura Combinatron

Plataforma Web

Interface do sistema com o usuário final, responsável por apresentar os resultados e servir como ponto de entrada dos dados das empresas e candidatos. O webapp será construído utilizando a tecnologia Vue.js(https://vuejs.org/) framework javascript open-source amplamente utilizado.

IA Combinatron

Módulo responsável por fazer o *match* dos candidatos com as vagas disponíveis. Implementado em Python utilizando as técnicas apresentadas na seção de resultados. O modelo de aprendizado de máquina resultante está disponível através de uma API construída em Flask(https://flask.palletsprojects.com/).

Rotinas ETL

Utilizando rotinas implementadas em Python e gerenciadas pelo Airflow(https://airflow.apache.org/) esse módulo é encarregado de extrair, transformar e retreinar o modelo de aprendizado de máquina com base nos novos dados disponibilizados nas fontes utilizadas.

Fontes de dados

Serão utilizadas as fontes de dados do SINE como apresentado na seção de resultados, mas buscando maximizar a performance do modelo de aprendizado de máquina é importante fornecer fontes adicionais que podem ser obtidas através de parcerias, como por exemplo, as informações dos perfis públicos e vagas registradas na plataforma Linkedin.