Organização do Desafio Técnico do PD

Preparando a base BI:

- Na parte de montar BD's, posso preparar um conjunto de dados, uma versão do BD para alimentar uma rede neural, que ele esteja com boa formatação para uma rede neural (Introduction to DL for Tabular Data, fast.ai);
- Fazer uma cópia do BD original, fazendo a imputação de dados faltantes e validar a qualidade desse novo BD;
- Pegar dados de quantos usuários no Brasil de celular usam iOS ou Android, e deixar proporcional para avaliar a performance em termos de eficiência ou interface do aplicativo, em relação a ambas as plataformas, por que pode haver uma diferença de gosto estético e dinâmica do aplicativo de S.O. pra S.O., podem ter públicos diferentes, e se normalizar ambas e o tráfego em uma for muito diferente do outro, é possível que haja alguma coisa para mudar aí, que valha a pena investigar mais a fundo em relação aos aplicativos.

Sobre o pipeline:

- Por ser um Data Pipeline, o processo deveria ser eficiente, logo pode ser conveniente no final botar um wrapper sobre as funções mais custosas para otimizar elas.

Predições:

- Qual faculdade nos próximos meses vai ter mais assinantes?
- Podia pensar qual a tendência futura de uso deu aparelhos na plataforma, para fazer análises de projeto e alocação de equipe.
- Qual faculdade nos próximos meses vai ter mais visualizações/acessos por mês?
- Qual a tendência de assunto a aumentar nos próximos meses?
- Seria possível deduzir o estado e cidade dos usuários para imputar os dados faltantes pela universidade, data de registro ou tipo de assinatura? (usando métodos de clusterização, NN etc)

Outras análises possíveis a partir da nova base (precisaria validar):

- Os alunos de quais universidades tendem a assinar o pacote premium, e outra, quais assuntos os alunos que assinam o pacote tendem a consumir?
- Das 10 principais faculdades/universidades, quais os assuntos mais acessados, seguidos?
- Quais os assuntos mais consumidos?
- Ver a correlação entre faculdade x acessos semanais, faculdade x registro de usuário, faculdade x assinatura, estado x acessos mensais, estado x assinaturas; fazer uma visualização no Bokeh com o mapa do brasil talvez, verde claro pro

escuro ou azul claro pro escuro.

- Tenho que ver as características dos usuários que não botaram esses dados, e pensar como poderia fazer pra botarem.
- Estudantes do tipo X seguem matérias Y, qual a correlação mais forte nessas categorias?
- W pessoas da universidade R seguem matérias Y, qual a correlação mais forte nessa associação?
- O banco de dados vai de 2012 até 2017, pois o último iOS e Android são, respectivamente, o Android Oreo e o iOS 11. Seria possível tirar alguma relação dos aparelhos mais frequentes e o tipo de registro no site (Facebook/Google/E-mail)?
- Dos usuários cadastrados num intervalo de tempo X, quais as principais características Y dos que mais entraram no site? (tenho 60000 cadastros e 283941 entradas)
- Os dados de State e City têm muitos NaN, por que?

quais as principais características de um estudante que quase não acessa o site, e quais as características de um estudante que segue menos que o valor de 25% da distribuição de acompanhamento de matéria

Sugestões:

- Dar sugestões para análises futuras, como usar Markov Logic Networks para extração de conhecimento e padrões.

alunos de qual universidade vêm consumido mais em termos de acesso à plataforma nos últimos 2012-2017?