Competição
Airbnb New User Bookings

Solução para NDGC = 0.885xx
(aprox. #149/1463 - LB privado)

Competição Airbnb

- desafio: a partir de dados do usuário e de ações de navegação no site, predizer para qual país o usuário residente nos EUA fará o seu primeiro "booking".
- países destino (classes): 'NDF', 'US', 'other', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL', 'DE', 'AU'
- métrica: NDCG
- dados fornecidos:

train.csv; test.csv	dados do usuário
age_gender_bkts.csv	dados da pirâmide etária por país
sessions.csv	dados de navegação de um subconjunto de usuários
countries.csv	dados do países de destino e língua-mãe

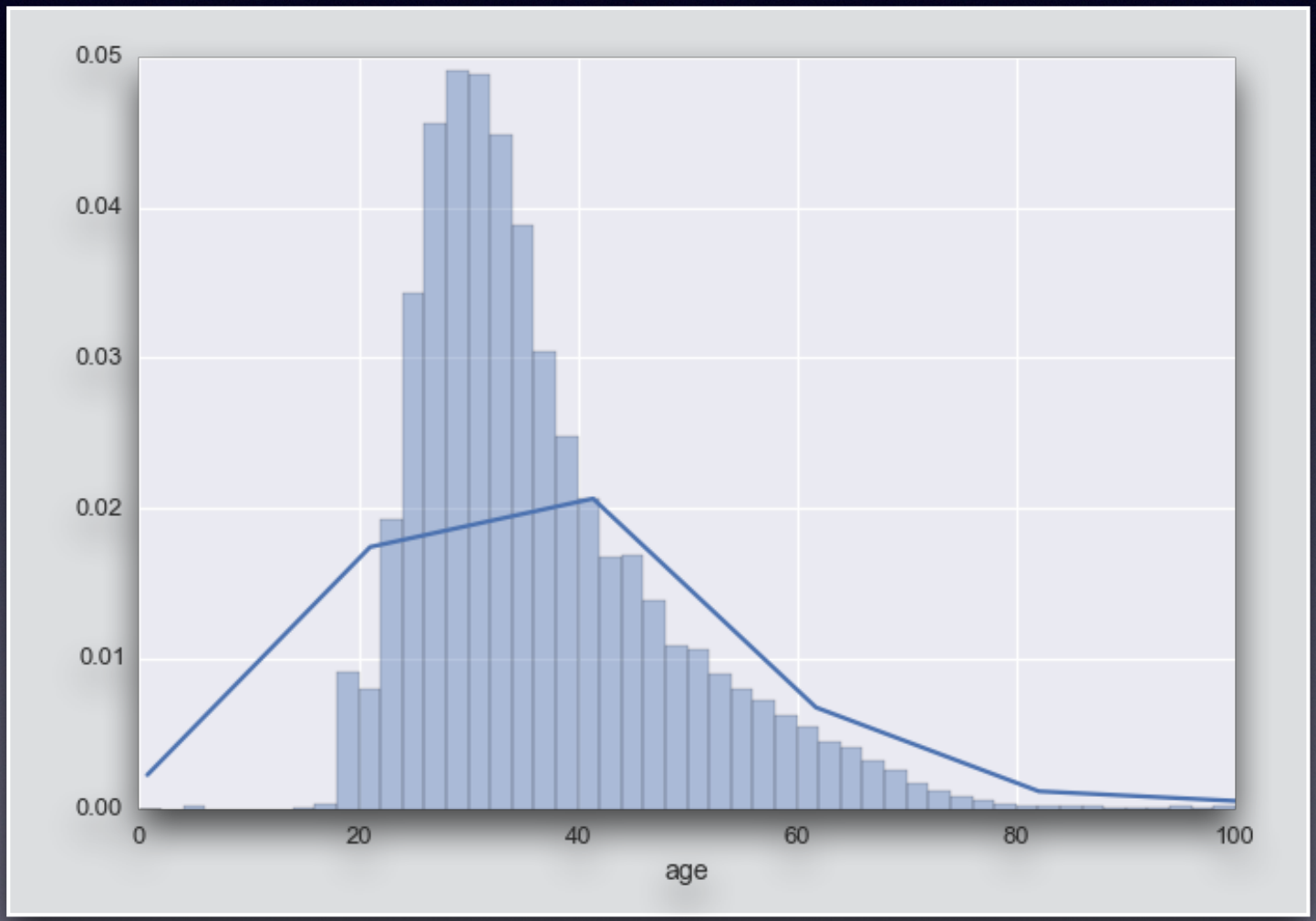
Intuições para feature extraction

- Principal caracterização de usuarios:
  - ➡ cadastro e cookies compõem dados demográficos (*)
  - ➡ maneira como navega revela os dados psicográficos e comportamentais (**)
- Outras fontes:
  - ➡ momento do primeiro acesso
  - ➡ pessoas preferem destinos com predominância de faixas na mesma idade, e sexo oposto (ou não).
  - ➡ 'booking' para fora dos EUA requer mais planejamento, e associado a proximidade dos feriados
  - ➡ quantidade de dispositivos distintos indica propensão a viagens mais caras
  - ➡ duração acumulada de todas as sessões de navegação, e sua quantidade total diz se o usuário "só está olhando" (label NDF) ou tem intenção de viajar e fazer um booking.

(*) dados demográficos
(**) dados psicográficos e comportamentais

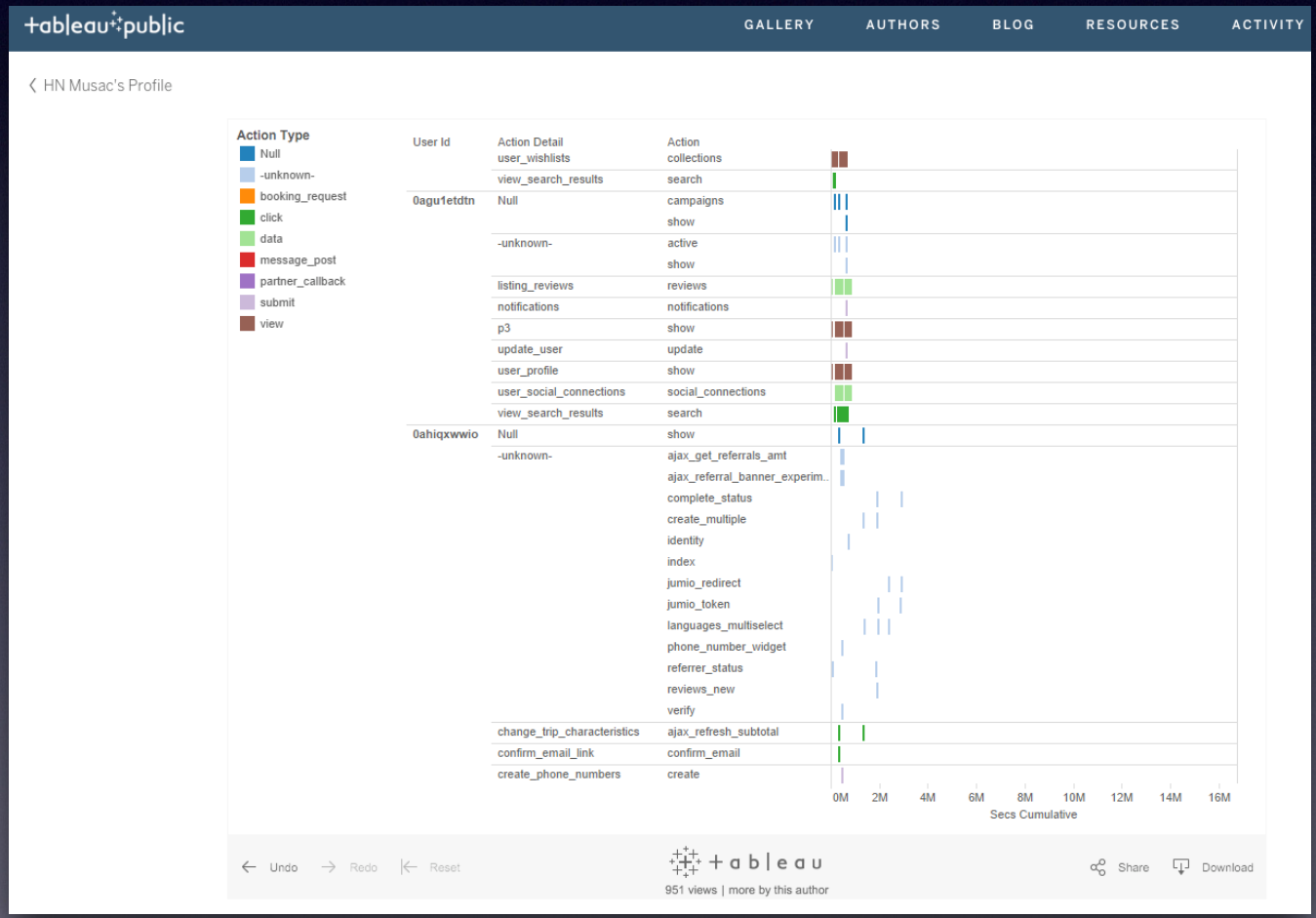
cadastro e cookies...

... dizem quem voce é



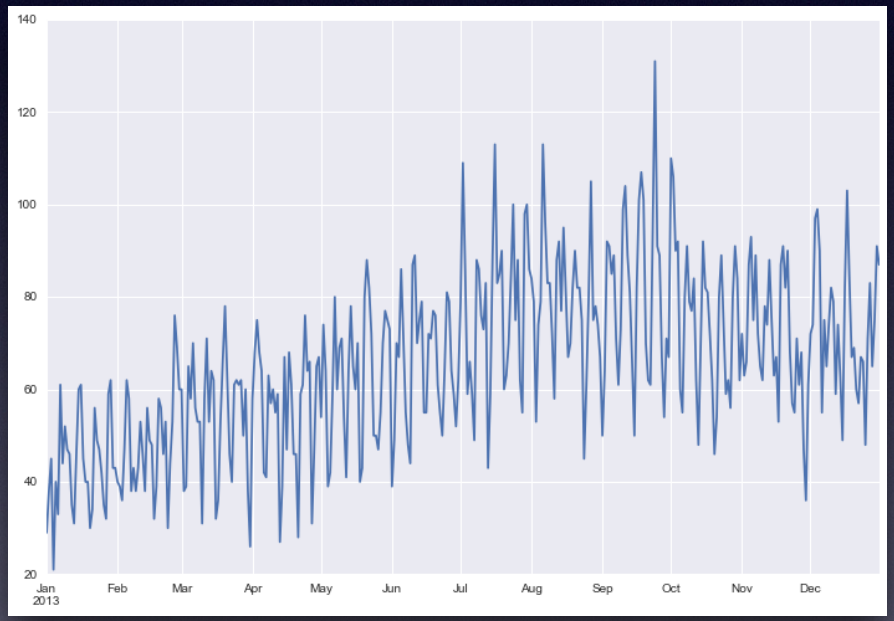
ações de navegação...

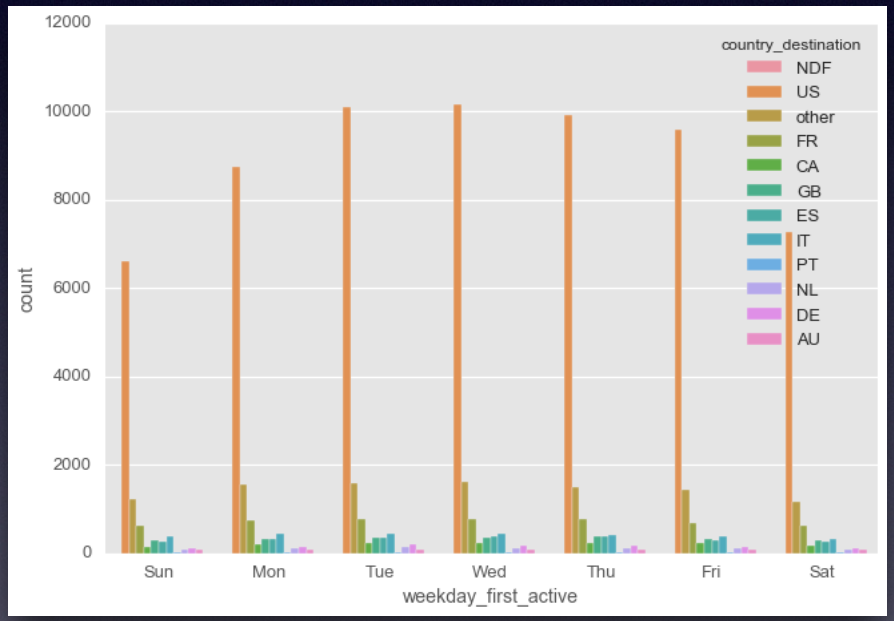
... revelam o que voce quer



outras análises

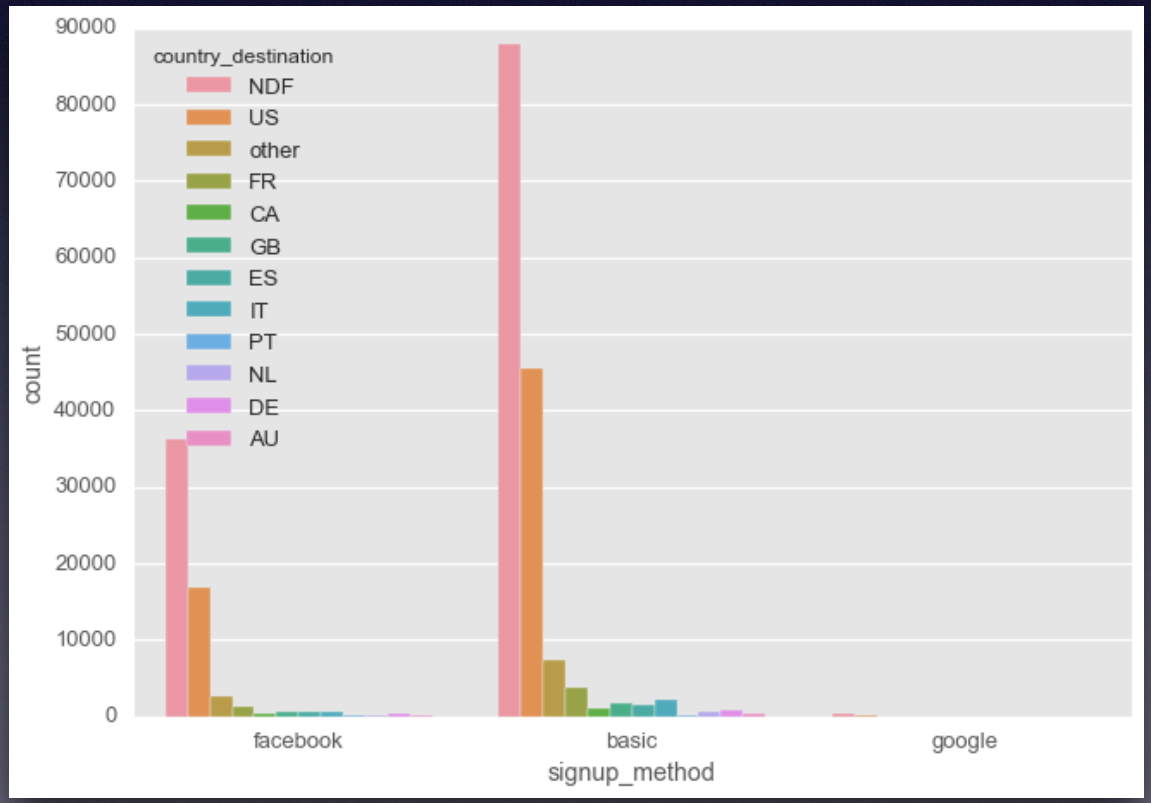
existe sazonalidade





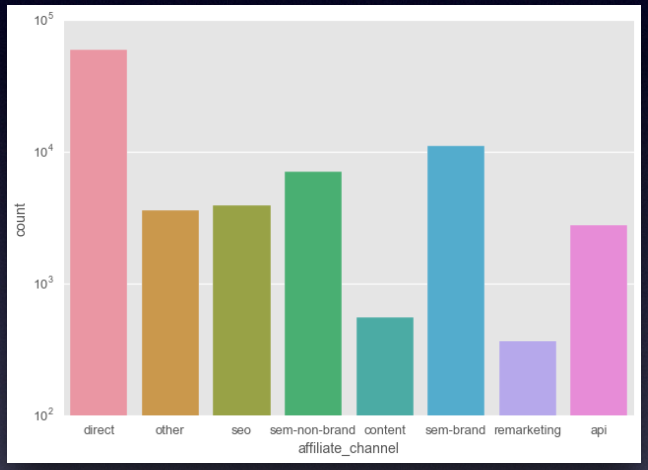
outras análises

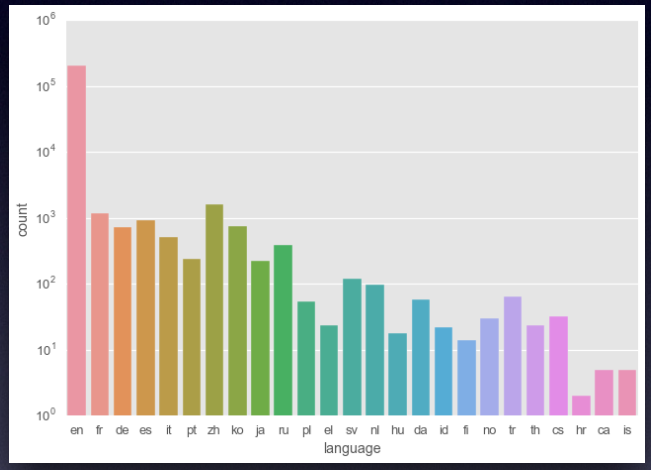
algumas features “talvez” não sejam informativas…

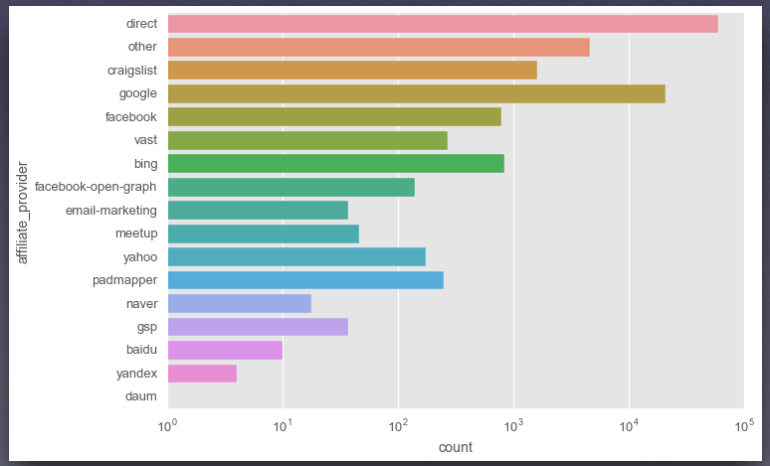


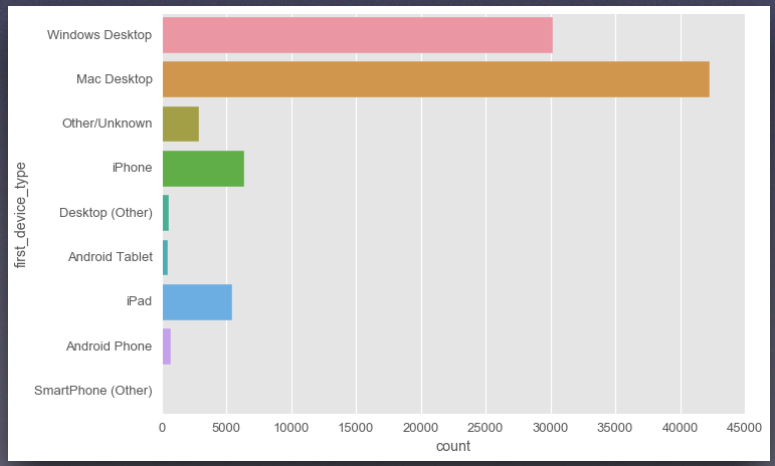
outras análises

for i = 1 to n : {visualizar; aplicar; medir desempenho}









Features...

- dados do usuario:
  - ➡ sexo, idade
  - ➡ browser, signup, linguagem, hardware,…
  - ➡ marketing: canal afiliado, provedor,…
  - ➡ da primeira atividade: dia da semana, dia do mês e hora do dia
- dias até o proximo feriado nos EUA
- ranking de países, baseado em sexo e idade
- unigramas, bigramas das ações de navegação:
  - ➡ unigramas: contagem de ações isoladas
  - ➡ bigramas: contagem das ações consecutivas
- quantidade total de sessões (liminar de inatividade = 30 mins)
- número de dispositivos unicos

Roteiro simplificado

- extração de features
- seleção univariada
  testes com 500, 1000 e 1500 features
- duas validações cruzadas com 2 folds
  trimestre anterior aos dados de teste, e mesmo trimestre dos dados de teste, mas no ano anterior
- classificador único XGBC

coisas que não fiz

- porque pioraram o score:
  - ➡ PCA
- porque não tinha hardware suficiente:
  - ➡ cadeias de Markov de ordem n>2
- porque não tive tempo:
  - ➡ ensembling e stacking
  - ➡ redes neurais