Nanodegree Engenheiro de Machine Learning

Rebeca Andrade Baldomir

Abril 2018

1 Proposta do Projeto

Esse projeto irá resolver o problema de reserva de novos usuários do Airbnb (Airbnb New User Bookings), procurando prever em qual país esses novos usuários irão procurar reservas e, assim, o Airbnb pode proporcionar um conteúdo mais personalizado pro seus clientes e prever melhor a demanda.

O conjunto de dados está disponível no Kaggle e consiste em uma lista de usuários dos EUA, dados demográficos, registros de sessões da *web* e dados estatísticos. Como países de destino possíveis estão: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' e 'other', sendo 'NDF' utilizado caso nenhuma reserva tenha sido feita.

Os dados são divididos em 5 arquivos:

1. train_users.csv: conjunto de treino dos usuários:
   (a) id: user id
   (b) date_account_created: datad a criação da conta
   (c) timestamp_first_active: data da primeira atividade, podendo ser antes da criação da conta.
   (d) date_first_booking: data da primeira reserva
   (e) gender
   (f) age
   (g) signup_method
   (h) signup_flow: página de onde o usuário veio se inscrever
   (i) language: idioma de preferência
   (j) affiliate_channel: tipo de marketing
   (k) affiliate_provider: onde está o marketing, ex: google, craigslist, outro
   (l) first_affiliate_tracked: tipo de marketing que o usuário interagiu antes de se cadastrar
   (m) signup_app
   (n) first_device_type

(o) first_browser

(p) country_destination: variável a ser predita

2. test_users.csv: conjunto de teste dos usuários
3. sessions.csv: dados da sessões web dos usuários
   (a) user_id
   (b) action
   (c) action_type
   (d) action_detail
   (e) device_type
   (f) secs_elapsed
4. countries.csv: estatística dos países de destino
5. age_gender_bkts.csv: estatística dos usuários em questão de idade, gênero e país de destino.

Esse é um problema multiclasse e será resolvido utilizando algoritmos de classificação, como árvores de decisão, gaussian naive bayes, k-nearest neighbors e logistic regression.

A métrica de avaliação desses modelos que serão propostos é NCDG, normalized discounted cumulative gain, pois é como o kaggle classifica as submissões feitas.

O benchmark a ser utilizado como referência está no leaderboard do kaggle que está entre 0.881 e 0.882 entre os 10 primeiros participantes.

A solução será feita com a exploração e visualização dos dados, aplicação dos algoritmos, comparação entre os resultados dos algoritmos, adoção do modelo de referência, a avaliação e validação dos modelo Exploração dos dados