Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Pedro Buzzi Filho
22/01/2018

Proposta

Histórico do assunto

"Airbnb New User Bookings" é um desafio do Kaggle que propõe aos competidores o objetivo de prever onde será a primeira experiência de viagem de um usuário. O Airbnb permite que os usuários façam reservas em mais de 34.000 cidades em 190 países. Sendo assim, um bom modelo preditivo pode auxiliar o Airbnb a sugerir destinos mais atrativos e acertivos aos novos usuários, gerando com isso mais reservas.

Este tipo de trabalho tornou-se atrativo para mim, por eu ser um desenvolvedor de software e ter acesso frequente a bases de dados e informações como estas. Sendo assim, com este projeto pretendo me aperfeiçoar na resolução desse tipo de problema.

Descrição do problema

A competição do Kaggle, junto ao Airbnb, forneceram dados sobre usuários, destinos escolhidos na sua primeira reserva, dados sobre a sessão e sobre os países. O problema está em utilizar estes dados para prever onde novos usuários têm mais chance fazer suas primeiras reservas. Com uma solução como esta, o Airbnb pode dar sugestões mais eficazes aos seus novos usuários, ganhando com isso mais adesão por parte deles.

Conjuntos de dados e entradas

Os dados estão disponíveis na descrição da competição do Kaggle, porém estes dados ainda precisarão ser preparados antes de serem usados para construção de um modelo preditivo. Abaixo seguem os dados fornecidos:

- train_users: Os dados dos usuários, sugeridos para treinamento.
- test_users: Os dados dos usuários, sugeridos para teste.
- countries: estatísticas dos países de destino.
- sessions: dados sobre a sessão e o dispositivo que usuários utilizaram.
- age-gender-bkts: dados gerais sobre a faixa etária dos usuários, gênero e país de destino.

Ao total, nos dados de treinamento, existem 11 países que foram destinos de novos usuários e "NDF" que significa que usuário não fez nenhuma reserva. Sendo assim, o modelo proposto irá prever qual dessas 12 opções os novos usuários tem mais chance de escolherem.

Descrição da solução

Para este problema, deve ser construído um modelo preditivo que seja treinado com os dados dos usuários uma vez e apartir disso, possa fazer a predição dos destinos mais prováveis para os novos usuários. O resultado deverá mostrar a porcentagem de chance do usuário viajar para cada país.

Serão testados e avaliados três modelos após a preparação dos dados:

- Gradient Boosting Classifier
- RandomForestClassifier
- XGBoost

Modelo de referência (benchmark)

Os competidores do Kaggle vêm utilizando fortemente o XGBoost e obtendo ótimos resultados, porém acredito que sempre possamos melhorar o resultado com uma boa exploração e preparação dos dados. Pretendo encontrar uma nova "teoria" na fase de exploração de dados e então aplicar os modelos mais utilizados neste problema:

- Gradient Boosting Classifier
- RandomForestClassifier
- XGBoost

Métricas de avaliação

Como métrica de avaliação será usado o "NDCG scorer", um script conhecido no Kaggle para fazer avaliação de desempenho dos modelos. O código desta métrica está disponível no próprio site do Kaggle, compartilhado por um usuário (davidgasquez) no link: https://www.kaggle.com/davidgasquez/ndcg-scorer/code.

Design do projeto

Exploração dos dados

Primeiramente serão explorados os dados para um melhor compreendimento das características e definição das abordagens. Nesta etapa também pode ser aplicada Estatística Inferencial para validar alguma teoria sobre os dados, e serão considerados os dados das sessões e dos países para criar teorias. Dependendo do resultado das teorias, alguns atributos poderão ser desconsiderados ou até mesmo criados na próxima etapa.

Preparação dos dados

Após a exploração e visualização dos dados, será feita a preparação dos mesmos, removendo atributos desnecessários, preenchendo dados faltantes e gerando novos atributos tratados se necessário.

Machine Learning

Nesta etapa serão construídos os modelos preditivos utilizando três diferentes algoritmos:

- Gradient Boosting Classifier
- RandomForestClassifier
- XGBoost

Os modelos deverão ser melhorados a partir de testes com diferentes parâmetros e sempre validados com a métrica escolhida. Por fim, eles deverão ser comparados e se possível, criado um modelo de Ensemble.

Conclusão e resultados

Este é o passo final onde será avaliado o melhor modelo e tiradas as conclusões sobre o projeto. Nesta etapa também será concluído o relatório e preparado para envio juntamente com o Notebook.