Nanodegree Engenheiro de Machine Learning

Proposta de Projeto Final

Rafael Novello
11 de janeiro de 2018

Proposta

Esta proposta consiste em classificar em uma base histórica de um site de leilões os lances realizados de forma automatizada (feita por robôs) e aqueles feitos pelos usuários tradicionais (humanos).

Os dados usados foram fornecidos pelo Facebook na plataforma Kaggle na forma de um desafio.

Histórico do Assunto

O leilão é uma forma de negociação, venda ou compra, de bens ou serviços muito utilizada. A dinâmica consiste, de forma muito resumida, na oferta de um item ou lote ao qual é estabelecido um preço mínimo inicial. Os interessados no item fazem suas ofertas (lances) por um determinado período e adquire o item aquele que fizer o maior lance até o fim do período do leilão.

Uma modalidade de leilão muito popular no Brasil é o Pregão Eletrônico que é realizado pelo governo em processos de licitação (aquisição de produtos ou serviços pelo governo). O pregão tem uma dinâmica inversa ao leilão. Neste, o governo anuncia publicamente a necessidade de compra ou contratação e os interessados competem entre si para alcançar o menor preço.

Segundo dados no Ministério do Planejamento, em 2013 o pregão eletrônico proporcionou uma economia de 9,1 bilhões de reais nas compra públicas do governo federal [1].

Descrição do Problema

Como descrito na seção anterior, o objetivo do leilão, assim como, e principalmente, do pregão eletrônico é permitir a competição igualitária entre os interessados. Porém, o uso de robôs (softwares que automatizam tarefas, neste caso o envio do lance ao leilão ou pregão) vem prejudicando sites de leilão e o pregão eletrônico do governo brasileiro, como podemos ver em várias matérias jornalísticas [2], [3], [4]

Identificar lances feitos por robôs é um grande desafio para plataformas de leilão, que precisam garantir a igualdade entre os “competidores”. Apenas assim as plataformas serão capazes de evitar que todos os ganhadores serão sistemas automatizados.

Para a identificação dos lances feitos por robôs e por operadores humanos vamos usar uma base de lances históricos feitos em uma plataforma de leilão online. Nesta base os operadores foram categorizados em robôs e humanos e desta forma podemos identificar todos os lances feitos por cada categoria de operador. Esta será nossa entrada.

O resultado esperado é a identificação dos lances feitos por cada uma das categorias de operadores através de uma variável binária, onde o valor 1 representa um lance feito por robô e o valor 0 representa um lance feito por humano. Esta será nossa saída.

Para obter o resultado esperado será usado um modelo de aprendizagem supervisionada, um classificador binário, que será melhor detalhado na seção Descrição da Solução.

Conjuntos de Dados e Entradas

Os dados disponibilizados na plataforma Kaggle para este desafio [5] estão organizados da seguinte forma:

Para o conjunto de dados do licitante (participante do leilão)

- bidder_id - Identificador exclusivo de um licitante.
- payment_account - conta de pagamento associado a um lance. Estes são obscuros para proteger a privacidade.
- endereço - endereço de correspondência de um licitante. Estes são obscuros para proteger a privacidade.
- Resultado - Etiqueta de um licitante que indica se é ou não um robô. O valor 1.0 indica um robô, onde o valor 0.0 indica humano.

Para o conjunto de dados de licitação

- bid_id - ID exclusivo para este lance
- bidder_id - Identificador exclusivo de um licitante (o mesmo que o bidder_id usado no train.csv e test.csv)
- Leilão - Identificador exclusivo de um leilão
- mercadoria - A categoria da campanha do site de leilões, o que significa que o licitante pode chegar a este site por meio da busca de "bens domésticos", mas acabou por licitar "bens esportivos" - e isso leva a que esse campo seja "bens domésticos". Este campo categórico pode ser um termo de pesquisa ou propaganda online.
- dispositivo - modelo de telefone de um visitante
- Hora - Tempo em que a oferta é feita (transformada para proteger a privacidade).
- país - O país ao qual o IP pertence
- IP - endereço IP de um licitante (ofuscado para proteger a privacidade).
- url - url do qual o licitante foi encaminhado (ofuscado para proteger a privacidade).

A base de lances contém 7.656.334 registros que são divididos em 3.071.224 no conjunto de treinamento (que deve ser dividido em treinamento e validação) e 4.585.110 no conjunto de testes. O conjunto de testes não contém a variável alvo, que deve ser inserida pelo resultado do modelo treinado para submissão na plataforma Kaggle e avaliação.

Na base de treinamento a proporção de lances feitos por humanos é de 87% e 13% feitos por robôs, o que mostra que o dataset está desbalanceado e por isso deve-se tomar algumas precauções que serão discutidas na seção Métricas de Avaliação.

Descrição da Solução

O objetivo do projeto é classificar, com o máximo de precisão, os lances realizados por robôs e por humanos. Para isso será usada a abordagem de aprendizado supervisionado usando classificadores pois é fornecida a base de treinamento com a identificação dos lances feitos por humanos e robôs.

Será necessária a análise das features mais relevantes no conjunto de treinamento, para esta análise pode-se utilizar PCA por exemplo. Com a informação de quais as features mais relevantes, será possível testar modelos de classificadores como SVC, Logistic Regression, Random Forest ou Redes Neurais.

Modelo de referência

Por se tratar de um desafio na plataforma Kaggle, é provável que a melhor referência para comparação seja o ranking associado ao desafio. Neste ranking [6] o melhor modelo teve uma taxa de acertos de 94,3% e as melhores 50 submissões ficaram acima de 92,3% de acerto.

Para apoiar durante o processo de treinamento e escolha do melhor modelo, pode-se utilizar o algoritmo Dummy Classifier da biblioteca SciKit-Learn [7] como baseline de comparação.

Métricas de Avaliação

Devido a natureza desbalanceada dos dados e a importância de não descartar erroneamente um lance feito por um usuário real da plataforma, a métrica de avaliação usada será Precision por lidar bem com conjuntos desbalanceados e por permitir comparar objetivamente os resultados de cada modelo testado. Em conjunto, será usado a matriz de confusão para identificar falsos positivos e falsos negativos na classificação.

Design do projeto

O projeto deve iniciar com a análise dos dados de treinamento, verificando se existe ausência de dados para alguma feature e fazendo algumas visualizações como para a distribuição entre os lances feitos por robôs e por humanos. O objetivo é saber se o dataset é balanceado, isso quer dizer, se a quantidade de lances feitos por robôs e humanos é a mesmo ou se esta distribuição está balanceada.

Em seguida pode-se verificar a correlação entre as features do modelo e a variável alvo. Neste processo uma análise de PCA pode ajudar a reduzir a dimensionalidade (número de features) mantendo apenas features que melhor identificam as categorias de lance. Outra opção é a utilização de algoritmos de árvore de decisão para avaliar as features mais relevantes para a classificação.

Na sequência será necessário processar os dados, transformando features categóricas em one-hot-encoding e ajustando a escala de features contínuas por exemplo. Este processo é necessário para a maioria dos modelos de machine learning.

Após o pré-processamento é possível testar modelos como SVC, Logistic Regression, Randon Forest e Redes Neurais com um subset dos dados de treinamento ou com todo o dataset, a depender da performance do treinamento. Para esta fase pode-se usar os parâmetros padrão de cada modelo, pois o objetivo é descobrir apenas qual deles apresenta melhor taxa de acertos nas classificações.

Ao descobrir o melhor modelo, a última etapa será a otimização do mesmo. Neste processo é possível usar a classe GridSearch da biblioteca Scikit-Learn para encontrar o melhor conjunto de parâmetros para o modelo escolhido.

Referências

- https://pt.wikipedia.org/wiki/Leil%C3%A3o
- https://pt.wikipedia.org/wiki/Preg%C3%A3o_eletr%C3%B4nico
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
- https://en.wikipedia.org/wiki/Precision_and_recall

Citações

1. http://agenciabrasil.ebc.com.br/economia/noticia/2014-02/com-pregao-eletronico-governo-economizou-r-91-bilhoes-em-2013
2. https://istoe.com.br/139247_GOLPE+NO+PREGAO+ELETRONICO/
3. https://idag.jusbrasil.com.br/noticias/2611942/pregao-eletronico-robos-ganham-licitacoes
4. http://g1.globo.com/tecnologia/blog/seguranca-digital/post/robos-participam-de-pregoes-compram-ingressos-e-atuam-na-bolsa.html
5. https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/data
6. https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/leaderboard
7. http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html