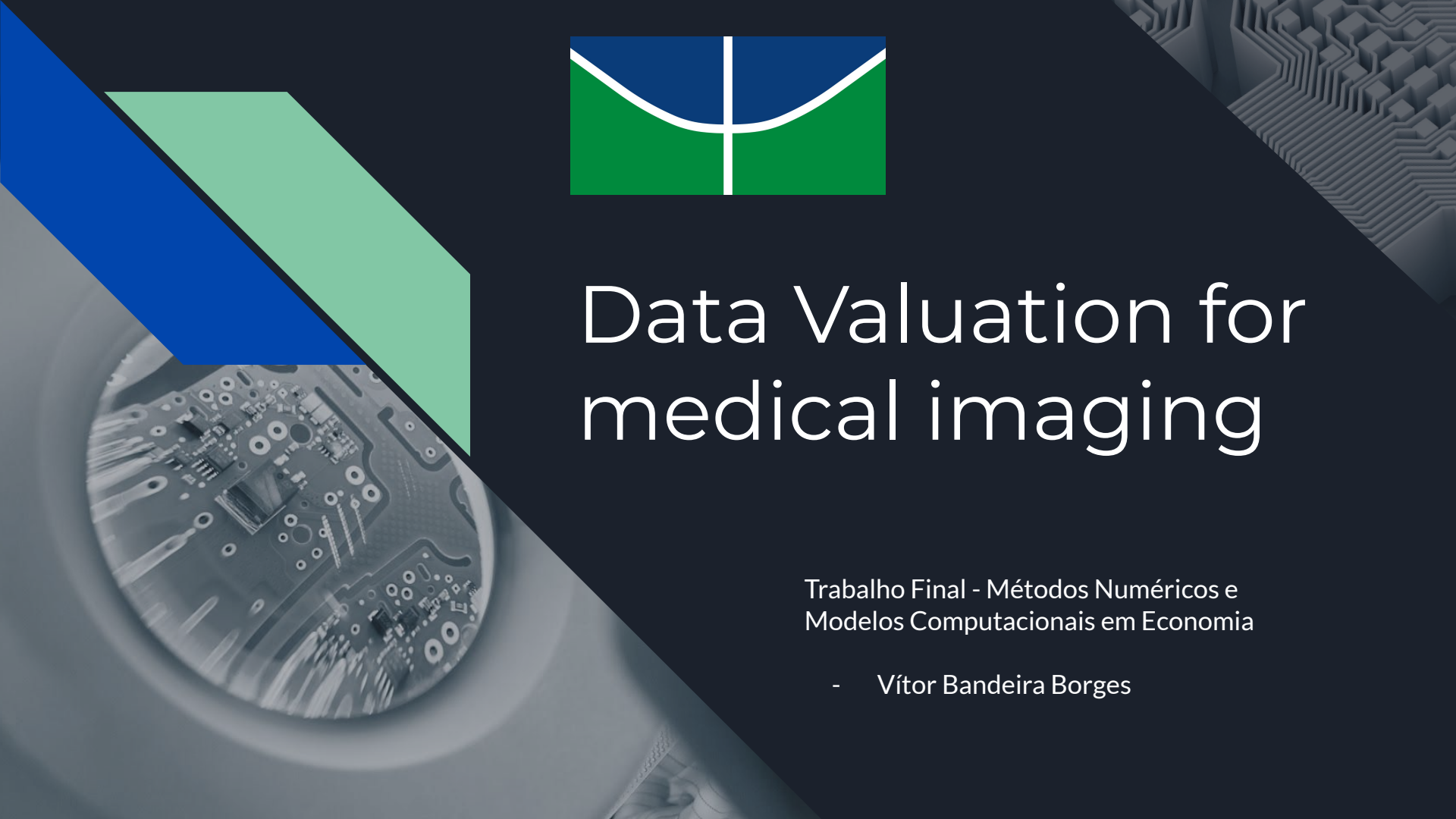


Data Valuation for medical imaging

Trabalho Final - Métodos Numéricos e Modelos Computacionais em Economia

- Vítor Bandeira Borges

Slide 1 - Referências e Motivações

01

O artigo de **Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A. Dunnmon, James Zou & Daniel L. Rubin** utiliza o valor de Shapley da teoria dos jogos para ranquear os dados de treino de uma rede neural convolucional de acordo com a sua importância.

02

O objetivo principal é avaliar se uma base de dados de baixa qualidade pode comprometer a acurácia do modelo. No artigo foram usados imagens de Raio-X do pulmão para previsão de pneumonia.

03

Escolhi este artigo com o objetivo de botar em prática o conhecimento adquirido sobre Redes Neurais Convolucionais, e por ser um experimento que eu poderia replicar partindo de pesquisas prévias.

Slide 2 - Teoria dos Jogos

O valor de Shapley é um conceito da teoria do jogos que mensura a contribuição individual de cada agente para o ‘payoff’ de um jogo cooperativo. A definição formal deste estimador é a seguinte:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

Se considerarmos um modelo de rede neural como um jogo cooperativo em que cada ponto de dado de treinamento é responsável por uma parte da eficácia do modelo, podemos mensurar o valor de Shapley para cada ponto de dado como uma métrica da qualidade deste dado.

Slide 3 - Base de Dados

O exercício empírico foi o treinamento de uma rede neural convolucional para a previsão do movimento dos preços de ações baseado no padrão de ‘candlesticks’ dos períodos anteriores, e subsequente estimação dos seus valores de Shapley. Um exemplo de dado da amostra:

Slide 3 - Base de Dados

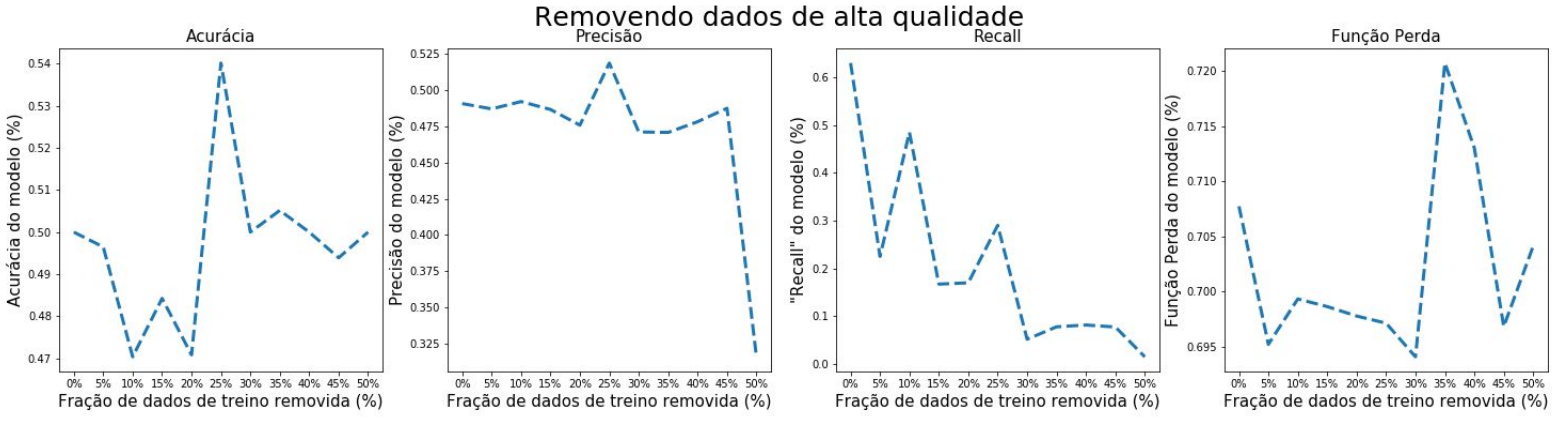
A linguagem de programação utilizada foi Python 3. Foram gerados um total de 1000 imagens utilizando a biblioteca ‘mpl_finance.candlestick2_ohlc’ para desenhar os gráficos e ‘alpha_vantage.timeseries’ para as variações de preço. Se o preço de uma ação subiu após a obervação daquele padrão de ‘candles’ a imagem foi classificada como ‘up’, e se o contrário aconteceu, como ‘down’.

Esta base foi separada em 700 para treino e 300 para teste. O modelo binário foi treinado usando o pacote ‘keras’ do ‘TensorFlow’, e os valores de Shapley para os dados de treinamento foram calculados com a biblioteca ‘shap’.

```
In [ ]: from alpha_vantage.timeseries import TimeSeries
        from mpl_finance import candlestick2_ohlc
        from tensorflow import keras
        import shap
```

Slide 4 -

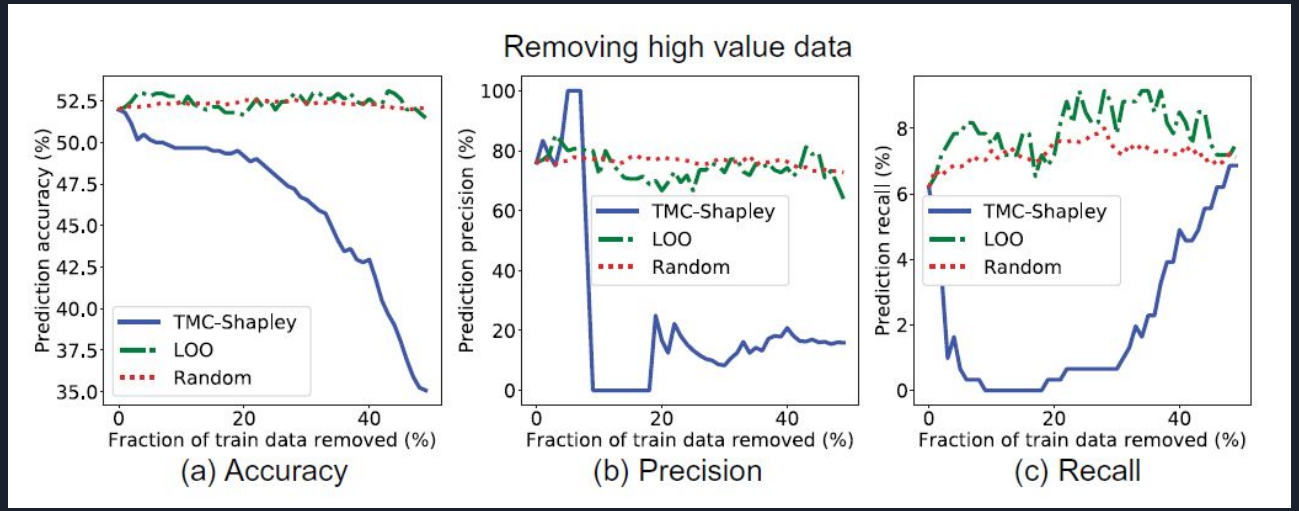
Após serem estimados os valores de Shapley para cada ponto de dado, foram treinados novos modelos removendo iterativamente os x% melhores dados para observar que efeito isto teria na precisão, acurácia, função perda e recall da rede neural.



O efeito que mais chamou atenção neste caso foi a redução progressiva do ‘recall’ quando o modelo foi perdendo seus dados de alta qualidade. O resto se parece com variações aleatórias.

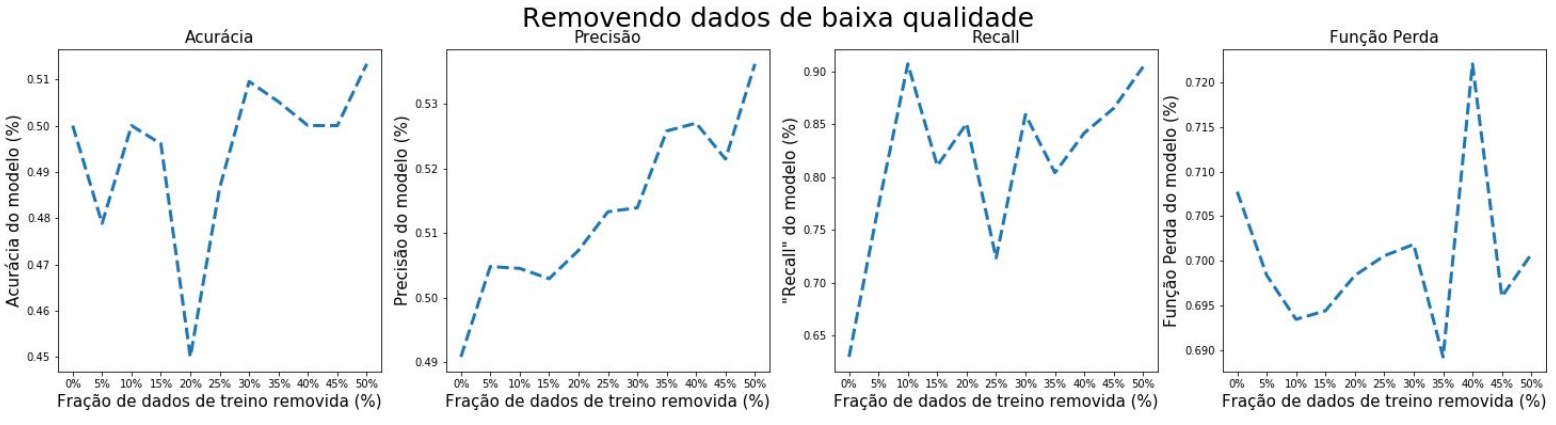
Slide 4 - Comparação de Resultado

No artigo, o principal resultado foi a redução de acurácia com a remoção dos dados de alta qualidade, mostrando que realmente os dados com alto valor Shapley eram os que mais causavam a previsão.



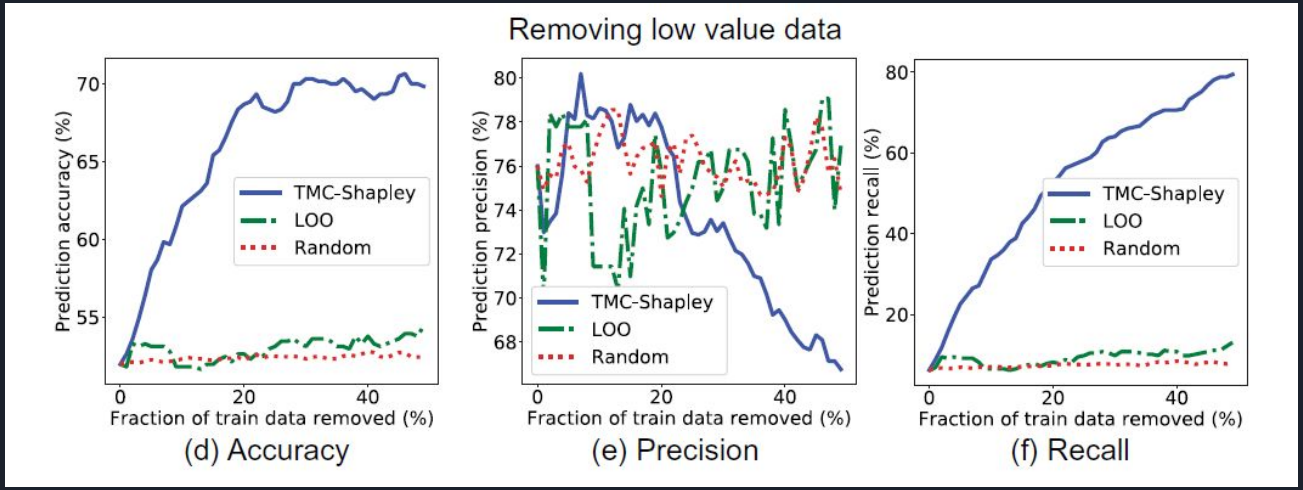
Slide 4 -

Quando removemos os dados de baixa qualidade da base de treinamento, continuamos observando o efeito dos valores de Shapley no recall, porém agora a correlação com a precisão se torna bem mais expressiva. Essas correlações serão estudadas mais profundamente nos resultados econométricos.



Slide 4 - Comparação de Resultado

Os resultados do artigo são bem mais expressivos que do experimento realizado por mim, no caso da remoção de valores ‘ruins’ os pesquisadores conseguiram elevar a acurácia do modelo até mais que 70%.



Slide 5 - Discussões

Os pesquisadores atribuíram essa melhora expressiva na acurácia à perda de dados classificados erroneamente na sua base. Os dados de imagens de raio-X podem muitas vezes conter baixa definição e problemas de mal-classificação.

No caso do experimento realizado, não temos este problema pois as variações nos preços das ações são observados com quase absoluta certeza, basta verificar se o preço era mais alto ou mais baixo após aquele padrão para obter a classificação correta. Por isso provavelmente não observamos uma melhora expressiva na acurácia, e sim uma melhora na precisão/recall.

Slide 5 - Discussões

Para mensurar isto os pesquisadores fizeram uma reavaliação da sua base de dados, utilizando a opinião de três radiologistas, e observaram que realmente os pontos de dados com valores de Shapley mais extremos tinham em sua maioria algum problema de classificação.

Com finalidade de somente reproduzir este experimento, foi solicitado por mim para que dois colegas com experiência na área de 'Day Trade' e análise gráfica fizessem uma avaliação destes valores extremos encontrados. O resultado foi bastante interessante:

Colega 1	15 maiores valores de Shapley	15 menores valores de Shapley	Total
Acertos	6	7	13
Erros	9	8	17
Proporção de Acertos	40%	46.667%	**43.333%**

Colega 2	15 maiores valores de Shapley	15 menores valores de Shapley	Total
Acertos	6	10	16
Erros	9	5	14
Proporção de Acertos	40%	66.667%	**53.333%**

Slide 5 - Discussões

O resultado se torna mais interessante quando comparamos estes resultados com as previsões de uma pessoa que havia acabado de descobrir o que eram gráficos de ‘candlesticks’:

Leigo	15 maiores valores de Shapley	15 menores valores de Shapley	Total
Acertos	8	6	14
Erros	7	9	16
Proporção de Acertos	53%	40.000%	**46.667%**

O leigo teve um poder de previsão melhor que um dos colegas com experiência no assunto, e melhor ainda que alguns dos modelos complexos de redes neurais convolucionais. Este resultado levanta hipóteses sobre a eficácia deste tipo de estratégia de investimento, porém está fora do escopo desta apresentação.

Slide 6 - Resultados Econométricos

01 Nesta seção nós iremos testar se os resultados encontrados na remoção dos valores de Shapley da base de dados são somente barulho, ou se realmente possuem alguma dependência linear.

02 Para isto vamos estimar modelos de regressão linear simples e avaliar não só o p-valor da hipótese de existência de correlação, como o R-quadrado estimado.

Slide 6 - Resultados Econométricos

Resultado das remoções de dados de alta qualidade

	Accuracy	Precision	Recall	Loss
p-value	0.477	0.085	0.003	0.0456
R-squared	0.058	0.293	0.646	0.063
Correlation	0.24	-0.541	-0.803	0.251

Resultado das remoções de dados de baixa qualidade

	Accuracy	Precision	Recall	Loss
p-value	0.234	0	0.07	0.728
R-squared	0.153	0.896	0.32	0.014
Correlation	0.391	0.9465	0.565	0.118

Os resultados das regressões confirmam as hipóteses iniciais de que havia dependência entre a remoção dos dados e a variação no recall e na precisão dos modelos. O alto R-quadrado sugere que esta correlação não é somente barulho e que os valores são bem relacionados.

As demais regressões podem até ter indicado existência de correlação entre os efeitos, porém seu R-quadrado e módulo da correlação não se mostraram tão expressivos. Por isso é mais seguro assumir que os valores encontrados são pura aleatoriedade no sorteio das amostras, e que os valores de Shapley não as influenciaram.

Agradecimentos

- Gostaria de agradecer primeiramente aos meus colegas que se fizeram dispostos à avaliar imagens de ‘candles’ para me ajudar neste trabalho, também ao meu pai que fez um esforço à mais para primeiro entender o que eram aqueles retângulos na tela, e depois classficá-los. (kkkkk)

- Também aos colegas de turma que comentaram nos exercícios que foram resolvidos ao longo do semestre, e sempre estiveram presentes no grupo de Whatsapp.

- E por último ao professor Cajueiro que me permitiu participar deste curso mesmo estando ainda na graduação, e sempre disponibilizou os melhores conteúdos e exercícios que auxiliaram no nosso aprendizado.