II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Modelo de previsão da qualidade do petróleo produzido no Brasil

Rhenan Gomes dos Santos Queiroz[1], Cássio Antonio Andrade [2], Claudio Marcio Cassela Inacio Junior[3], Roseli A. F. Romero[4]
ICMC-USP

1 Introdução

Com o advento da exploração do pré-sal, o Brasil passou a figurar entre os principais produtores de petróleo no cenário mundial [1]. Tal descoberta foi severamente importante, uma vez que esses poços possuem o potencial de produção dez vezes maior que os poços em pós-sal. Atualmente, o Brasil produz cerca de 2 milhões de barris por dia (Mbpd), representando mais do que 50% da produção total nacional [1]. Ainda assim, alguns poços podem aumentar a sua produção em mais 5 Mbpd nos próximos anos e devido a isso, estudos na área tem sido conduzidos de diversas formas como, por exemplo algumas técnicas de visualização de informação que vêm sendo publicadas e desenvolvidas nos últimos anos [3–5]. Tais fatores permitem análises profundas no que tange a aquisição, pré-processamento, transformação, análises, classificação e previsão de dados. Entre essas técnicas destacam-se os algoritmos de Machine Learning (ML) que permitem um treinamento de determinados atributos desejados nos datasets a fim de classificar características a partir de atributos que possuem relações entre si.

O objetivo deste trabalho é abordar algumas técnicas de exploração de dados e algoritmos de Machine Learning na base de dados de produção dos poços brasileiros publicados pela ANP. Os principais resultados obtidos demonstram uma diferenciação no subconjunto de atributos selecionados a partir de duas técnicas de seleção de variáveis: o Mutual Information Score (MIS) e o Sequential Feature Selection (SFS). Além disso, a modelagem da base de dados usando os métodos Random Forest e XGBoosting obtiveram uma alta performance quanto a predição da variável de Grau API, sendo o XGBoosting o algoritmo que obteve o melhor resultado de acordo com as métricas de erro médio absoluto (MAE) e erro médio quadrático (MSE).

[1] rhenan.queiroz@usp.br
[2] andrade.cassio@usp.br
[3] claudio.inacio@usp.br
[4] rafrance@icmc.usp.br

2 Métodos e Experimentos

2.1 Dataset, Exploração e Pré-processamento

O dataset nomeado por "Dados de Produção por Poço (pós-2018)" é fornecido pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), responsável pela regulação do mercado de petróleo brasileiro. Como obtido pelo metadados disponível via website [2], a base de dados possui informações detalhadas sobre os volumes produzidos mensalmente em cada poço. A temporalidade dos dados é mensal e disponibilizada no formato "csv".

No estudo foram avaliados os dados de produção de janeiro de 2021 a abril de 2022. A fim de se obter uma análise mais abrangente, optou-se por trabalhar com uma granularidade maior e, portanto, foi realizada uma sumarização dos dados com base nos atributos *bacia*, *periodo*, *tipo_producao* e *grau_api*, permitindo-se o manuseio de uma base de dados mais compacta. Após isso, foi efetuado o tratamento dos valores ausentes através da inserção por meio de interpolação linear e optou-se por eliminar apenas as instâncias cujos valores da variável resposta, Grau API, fossem ausentes. A base final para modelagem resultou em 42 variáveis.

2.2 Experimentos

Para a seleção de variáveis, duas técnicas diferentes foram utilizadas para predição do valor do Grau API: o Mutual Information Score (MIS) e o Sequential Feature Selection (SFS) [6]. Aplicando-se o MIS foi possível verificar que com 21 variáveis, o valor acumulado de informação dos dados fosse responsável por cerca de 90% da informação que explica a variável resposta. Foram também selecionadas 21 variáveis a fim de realizar uma compararação com o método SFS, no qual definiu-se também um estimador linear e direção 'backward'.

No que tange a modelagem de previsão, foram utilizados os métodos XGBoosting que obtém a árvore de decisão mais adequada, segundo o método dos gradientes e o método Random Forest, que é uma coleção de árvrores de decisão.

Primeiramente, ao se comparar as duas técnicas de seleção de variáveis, observa-se pela Tabela 1 que o subconjunto de variáveis selecionadas por cada método foram divergentes: enquanto o SFS selecionou como variáveis importantes algumas bacias específicas, o MIS selecionou mais variáveis numéricas referentes aos dados de produção do petróleo. É importante destacar que o SFS estabelece uma relação linear para a remoção das variáveis, enquanto o MIS utiliza uma métrica de similaridade que mede a dependência entre duas variáveis.

Para definição dos parâmetros do modelo Random Forest, o método *RandomizedSearchCV* foi aplicado considerando-se 100 iterações e uma validação cruzada em 5 folds, que buscasse minimizar o erro quadrático médio. No que diz respeito ao modelo XGBoosting, o método *RandomizedSearchCV* foi também aplicado com as mesmas configurações. Na Tabela 2 verifica-se a melhor combinação de parâmetros para cada modelo combinado com cada método de seleção de variáveis.

Após treinar os modelos com as configurações obtidas na Tabela 2, foram realizadas as previsões do Grau API para as observações reservadas no conjunto de teste. A Tabela 3 mostra a avaliação dos erros dos valores previstos por meio das métricas erro médio absoluto (MAE) e erro médio quadrático (MSE).

Tabela 1: Feature Selection

MIS	SFS
frac_dest_leves_volume_md	condensado_bbl_dia_sum
frac_dest_medios_volume_md	vol_gas_mm3dia_sum
frac_dest_pesados_volume_md	frac_dest_medios_volume_md
pcs_gp_kjm3_md	frac_dest_pesados_volume_md
metano_md	metano_md
etano_md	etano_md
co2_md	propano_md
propano_md	isopentano_md
nitrogenio_md	hexanos_md
dens_glp_gas_md	heptanos_md
butano_md	undecanos_md
npentano_md	dens_glp_gas_md
isopentano_md	pcs_gp_kjm3_md
hexanos_md	bacia_Alagoas
heptanos_md	bacia_Campos
octanos_md	bacia_Espírito Santo
petroleo_bbldia_sum	bacia_Potiguar
nonanos_md	bacia_Santos
oxigenio_md	bacia_Sergipe
vol_gas_mm3dia_sum	bacia_Solimões
agua_bbl_dia_sum	tipo_extracao_terra

Tabela 2: Melhores hiperparâmetros

		MIS	SFS
Random Forest	max_features	7	7
	n_estimators	108	102
XGBoosting	max_depth	5	15
	learning_rate	0.3	0.2
	subsample	0.9	0.9
	colsample_bytree	0.9	0.9
	colsample_bylevel	0.9	0.8
	n_estimators	500	1000

Tabela 3: Modelagem de Classificação do Grau API

Modelo	Método Features	Método Parâmetros	MAE	MSE
XGBoosting	MIS	Random Search	0.08	0.15
	SFS		0.14	0.27
Random Forest	MIS	Random Search	0.19	0.39
	SFS		0.30	0.94

3 Conclusões

A escolha do método de seleção de variáveis é de grande importância, pois a suposição de relações lineares para as covariáveis e variável resposta pode implicar em uma pior predição, mesmo usando os mesmos algoritmos que em outra seleção. No caso, a relação obtida com o MIS foi superior ao da SFS. Tanto o Random Forest quanto o XGBoosting resultaram num bom ajuste e desempenho na tarefa de previsão da qualidade do petróleo brasileiro a partir dos dados de produção. Ainda assim, o modelo XGBoosting obteve um melhor resultado em comparação com o Random Forest quando avaliado pelo MAE e MSE nos dados de teste. Portanto, para uma análise de regressão utilizando o conjunto de dados de produção de petróleo brasileiro, recomenda-se a utilização do algoritmo XGBoosting em conjunto com a técnica de seleção de variáveis MIS.

Referências

[1] Petersohn, E. Pre-Salt Super Play: Leading Brazil into the World's Top 5 Oil Suppliers. *AAPG Latin America And Caribbean Region Geoscience Technology Workshop*. (2019)

[2] Agencia Nacional do Petroleo, Gas Natural e Biocombustiveis Dados Abertos. (2022), https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos, acesso em 2022-05-28

[3] Leal, A. B, Moura, T. R. da Silva. Data analytics applied to the analysis of petroleum production in Brazil. *Brazilian Applied Science Review*. (2021)

[4] Liu, W., Liu, W. & Gu, J. Petroleum production forecasting based on machine learning. *Proceedings Of The 2019 3rd International Conference On Advances In Image Processing*. pp. 124-128 (2019)

[5] Noshi, C., Eissa, M. & Abdalla, R. An Intelligent Data Driven Approach for Production Prediction. (2019,5), https://doi.org/10.4043/29243-MS, D041S048R007

[6] Scikit Learn - Feature Selection. (2022), https://scikit-learn.org/stable/modules/feature_selection.html, Last accessed on 2022-07-25