MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Análise de Viabilidade em Projetos de Exploração Mineral

Ian Lucas Ramos de Carvalho dos Santos Pinto[1], Emerson Rocha Pereira[2], Rosilanny Soares Carvalho[3], Bruna Dias Lucena[4], Aline Elí Gassenn[5], Francisco Louzada Neto[6]
ICMC-USP

1 Introdução

Empresas mineradoras, em escala global, têm enfrentado o desafio de manter sua competitividade no mercado diante do esgotamento das minas em produção atualmente e da crescente demanda por metais, especialmente o cobre, dado o avanço das tecnologias elétricas limpas e da busca por redução das emissões de carbono frente as mudanças climática. Nesse contexto, a exploração mineral - ou pesquisa mineral -advém como a fase inicial da mineração, que visa identificar novos depósitos minerais com volumes e teores economicamente viáveis, de modo a justificar a construção de uma mina para suprir as necessidades mundiais.

A maturação de projetos de mineração desde o início da pesquisa mineral até as estimativas financeiras de retorno de investimento e fluxo de caixa, é um processo longo que envolve diversas etapas. A pesquisa mineral envolve o estudo geográfico da reserva mineral e suas características, após essa pesquisa é feito uma estimativa do custo operacional da mina, do valor do minério e do tempo de operação. Essas informações possibilitam calcular a Taxa Interna de Retorno (TIR), que é uma estimativa do retorno do investimento anualmente, enquanto essa mina estiver em operação. O TIR é um indicador importante para os investidores decidirem começar as operações na mina.

Este estudo tem como objetivo desenvolver um modelo preditivo que, com base em dados da pesquisa mineral de reservas de cobre e investimentos iniciais planejados, preveja se a reserva mineral teria um retorno anual estimado menor que 15%, considerado um retorno economicamente não viável. Esses dados são coletados de diversas fontes públicas de minas cujo o retorno estimado já foi calculado. Dessa forma, o modelo simularia o retorno estimado dadas as qualidades minerais da reserva, de maneira a indicar de antemão se essa reserva seria viável economicamente

[1]ian.lucas.r.c.s@usp.br
[2]emersonrocha@usp.br
[3]rosilanny.carvalho@usp.com
[4]brunalucena@usp.br
[5]aline.gassenn@usp.br
[6]louzada@icmc.usp.br

e orientar as equipes de pesquisa mineral. Dessa forma, seriam puladas diversas etapas de estimativas financeiras necessárias para calcular o TIR, de forma a promover economia de tempo e dinheiro na analise do potencial da reserva mineral. O próximo passo imediato desse projeto seria, dado um simulador de TIR, executar uma análise contrafactual para estimar quais características a reserva mineral em questão deveria ter para que fosse economicamente viável. O trabalho foi desenvolvido na disciplina MAI5003, utilizando a abordagem de Aprendizado Baseado em Problemas (PBL).

2 Materais e Métodos

2.1 Base de Dados

A base de dados utilizada consiste em um conjunto estruturado de informações de projetos de minas de cobre primário, extraídas de publicações obrigatórias de empresas de capital aberto do setor de mineração (*Press Releases* e Relatórios Técnicos). As empresas divulgam esses dados de forma pública, eles são coletados e integrados diretamente à base. A base inclui variáveis categóricas e contínuas, como o nome da propriedade, *status* da atividade, tipo de mina, custo inicial de capital em milhões de dólares, taxa de desconto do VPL (Valor Presente Liquído) e TIR pós-impostos para o cenário base, preço de referência por tonelada, tipo de corpo geológico do minério, país/região, tonelagem de recursos e reservas de minério, e teores de cobre, chumbo, zinco, ouro e prata nas reservas e recursos.

2.2 Análise Exploratória

A base de dados foi dividida em treino e teste, com 100 amostras para treino e 94 amostras de teste. A análise exploratória iniciou-se com uma avaliação das variáveis contínuas e categóricas, para visualizar distribuições, verificar padrões e potenciais anomalias. Em primeiro lugar, foram calculadas as proporções de valores ausentes (em média 2% por variável), facilitando a identificação de atributos que requeriam imputação. A distribuição de valores foi examinada com estatísticas descritivas e gráficos, permitindo a detecção de valores extremos ou *outliers* e a verificação da variabilidade e a dispersão dos dados numéricos, especialmente custo inicial e taxas de retorno.

Para as variáveis categóricas, como GEOLOGIC_ORE_BODY_TYPE (tipo de corpo geológico), GLOBAL_REGION (regiões globais, tais como Ásia, Oriente Médio e África) e MINE_TYPE (classificação da mina como subterrânea, marinha, entre outras), a análise exploratória incluiu contagens e proporções das categorias, onde observou-se o desbalanceamento entre algumas classes, o que permitiu o posterior tratamento por agrupamento.

Observou-se que 15% das amostras de treino apresentam TIR abaixo de 15%, um indicativo de que o *target* é desbalanceado. A base é balanceada em etapas posteriores com a aplicação dos algoritmos SMOTE [2, 5] e o RandomUnderSampler [5].

2.3 Feature Enginneering

No processo de *Feature Engineering*, foram criadas *features* com potencial de aumentar o desempenho do modelo. Foram criadas variáveis compostas, como PRECIOUS_ORE_DENSITY, que é a soma da densidade de ouro e prata (g/ton). Variáveis de quantidade total também foram desenvolvidas, como GOLD_TONNAGE, SILVER_TONNAGE, e PRECIOUS_TONNAGE, ao multiplicar a densidade dos metais pela tonelagem total da mina. Outra variável informativa é INITIAL_COST_PER_TONNE, que calcula a razão entre o custo inicial e quantidade econômica (o total em toneladas de cobre mais o de ouro e prata em gramas). Esse processo de engenharia também envolveu a aplicação de transformação logarítmica para as variáveis numéricas, para extrair informação de ordens de grandeza.

2.4 Modelagem

Na modelagem, aplicou-se um *pipeline* [4] que inclui transformações de variáveis categóricas e numéricas. As categóricas foram codificadas binariamente, enquanto as numéricas foram padronizadas por meio da subtração da média e divisão pelo desvio padrão. Esse processo foi complementado com a imputação de valores ausentes pelo algoritmo KNN [4].

A seleção de variáveis foi realizada com o algoritmo Boruta [6], com base em um modelo Random Forest [4]. Identificadas as variáveis mais importantes, as numéricas são usadas em análise de componentes principais para redução dimensional. Para lidar com desbalanceamento, foram aplicadas as técnicas SMOTE, para sobre-amostragem e o RandomUnderSampler, para sub-amostragem. Esse procedimento de modelagem também foi aplicado para um modelo XGBoost [3].

Todos os modelos passaram por otimização de hipeparâmetros via otimização Bayesiana [1]. Nos dados de treino os modelos têm alto desempenho, indicando propensão à *overfitting*. Os modelos foram otimizados com ROC AUC [4] obtido via validação cruzada.

3 Resultados

O desempenho do Random Forest e XGBoost foi avaliado com estimativas de ROC AUC e curva ROC [4] por meio de *bootstraping* para dados de treino e teste. Os percentis 24%, 50% e 75% dos ROC AUC para teste, estimados para Random Forest e XGBoost, são respectivamente 0.59, 0.71, 0.81 (intervalo de confiança 0.22) e 0.58, 0.70, 0.82 (intervalo de confiança 0.24).

4 Conclusão

Até o momento foram executadas análises com curva ROC e ROC AUC para verificar a capacidade dos modelos em diferenciar as classes positivas e negativas sem um valor de corte específico para definir uma mina como rentável ou não. Esse valor será definido em passos posteriores do projeto e outras métricas e visualizações serão empregadas para a avaliação, como a matriz de confusão. No experimento relatado, o XGBoost apresenta um intervalo de confiança menor que o do Random Forest, e também possui uma mediana de ROC AUC maior. Em termos de curva ROC, as curvas são similares, contudo a do XGBoost é mais bem comportada. Até o momento,

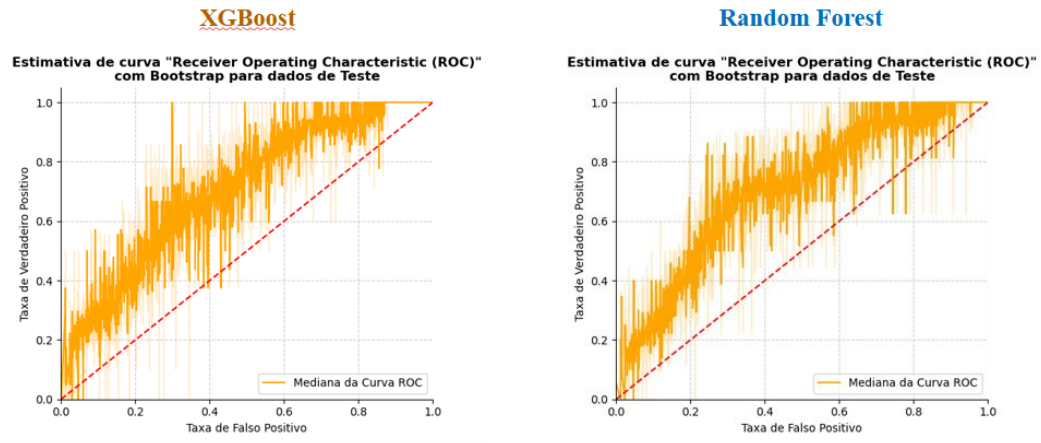

Figura 1: ROC AUC estimados com *boostraping* para dados de teste

a modelagem com XGBoost possui a melhor performance. Espera-se no futuro empregar outros modelos, técnicas de explicabilidade e eventualmente algoritmos de Monte Carlo para amostrar mais dados de treino.

Referências

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. *Optuna: A Next-generation Hyperparameter Optimization Framework*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. *SMOTE: synthetic minority over-sampling technique*. J. Artif. Int. Res., 16(1):321–357, 2002.

[3] Tianqi Chen and Carlos Guestrin. *XGBoost: A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pages 785–794, 2016.

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. *Scikit-learn: Machine learning in Python*. Journal of machine learning research, 12:2825–2830, 2011.

[5] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research, 18(17):1-5, 2017.

[6] Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11