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

CC BY

Estudo Comparativo de Algoritmos de Aprendizado de Máquina para Previsão de Parâmetros Técnicos e Financeiros em Projetos de Mineração

Emerson Rocha Pereira[1], Thomas Peron[2]
ICMC-USP

1 Introdução

O processo de avaliação financeira de projetos de mineração é essencial para empresas que pretendem se manter no mercado a longo prazo e requer a previsão de parâmetros para a tomada de decisões eficazes. Este estudo compara diferentes algoritmos de aprendizado de máquina para prever variáveis críticas como custos e capacidade industrial. A análise visa identificar os algoritmos mais eficazes para previsão de cada parâmetro, proporcionando direcionamentos estratégicos para a otimização de projetos de mineração.

2 Materiais e Métodos

Os dados utilizados neste estudo são provenientes de informações técnicas e financeiras de 817 projetos de mineração que produzem ou produzirão cobre (primario ou secundário), com base em relatórios públicos de análise econômica (*Economic Assessment*). Utilizou-se a premissa de que sempre se conhecem características básicas de um novo projeto como tonelagem de minério disponível, concentrações de metais, status de atividade, região geográfica e tipo de mina, com o objetivo de testar e comparar cinco algoritmos de aprendizado de máquina, utilizando os dados inicialmente apenas pré-processados e, posteriormente, com a aplicação de PCA, para prever os parâmetros técnicos e financeiros: custo inicial, capacidade industrial, vida útil do empreendimento, taxa de recuperação dos metais, custos de sustentação, custos de mineração e custos de processamento, que serão utilizados posteriormente como *input* em modelos de avaliação financeira de projetos dessa natureza. O processo de análise envolveu as seguintes etapas:

1. Leitura e Pré-processamento dos Dados:
    (a) Separação de Variáveis: As variáveis categóricas foram separadas das numéricas.

[1]emersonrocha@usp.br
[2]thomas.peron@usp.br

(b) Transformação Logarítmica: Variáveis com ordens de grandeza muito grandes como recursos e reservas e capacidade de produção foram transformadas logaritmicamente para reduzir a variabilidade e melhorar a performance dos modelos.

(c) Normalização: As variáveis numéricas foram normalizadas para garantir que todas tenham a mesma escala, o que é essencial para algoritmos que dependem da distância entre os dados.

(d) Codificação: As variáveis categóricas foram codificadas para serem utilizadas nos modelos.

(e) Tratamento de Valores Nulos: Utilizou-se *K-Nearest Neighbors* (KNN) *Imputer* para preencher valores nulos, garantindo que os dados estivessem completos para a análise.

2. Treinamento e Avaliação dos Modelos:

   (a) Foram utilizados os algoritmos *Linear Regression*, *Random Forest*, *Support Vector Regresso*r, *KNN* e *Gradient Boosting Regressor*.

   (b) Os modelos foram treinados e utilizados para prever cada uma das variáveis alvo.

   (c) O desempenho dos modelos foi comparado utilizando as métricas *Mean Squared Error* (MSE) e R^2 *Score*. O MSE mede a média dos quadrados dos erros, enquanto o R^2 *Score* indica a proporção da variância dos dados que é explicada pelo modelo.

3. Análise de Componentes Principais (PCA):

   (a) O PCA foi aplicado para reduzir a dimensionalidade dos dados, mantendo a maior parte da variância. Isso ajuda a simplificar os modelos e a reduzir o risco de *overfitting*.

   (b) O processo de Treinamento e Avaliação dos Modelos foi aplicado novamente, a partir do PCA.

4. Comparação dos Modelos:

   (a) Inicialmente, os modelos foram treinados com os dados pré-processados.

   (b) Em seguida, os mesmos modelos foram aplicados utilizando PCA, para comparar os resultados e entender as diferenças e impactos.

   (c) Aplicar os algoritmos diretamente aos dados permite avaliar a robustez dos modelos em condições mais próximas dos dados reais, sem transformações. Isso ajuda a entender como os modelos lidam com a complexidade e variabilidade dos dados brutos. Por outro lado, aplicar a Análise de Componentes Principais (PCA) antes dos algoritmos reduz a dimensionalidade dos dados, simplificando os modelos e diminuindo o risco de *overfitting*. Comparar ambas as abordagens é importante para determinar se o pré-processamento melhora significativamente a precisão ou se os modelos são robustos o suficiente para lidar com dados não processados, fornecendo embasamento para a otimização de projetos de mineração.

Algoritmos Utilizados

- *Linear Regression*: Modelo linear que busca a melhor linha reta que descreve a relação entre variáveis independentes e a variável dependente. É de fácil interpretação, porém pode não capturar relações complexas nos dados.
- *Random Forest*: Modelo de *ensemble* que utiliza múltiplas árvores de decisão para aprimorar a precisão e mitigar o *overfitting*. As previsões são obtidas pela média das previsões de todas as árvores.
- *Support Vector Regressor* (SVR): Modelo que busca um hiperplano em um espaço de alta dimensionalidade, sendo eficaz na captura de relações não lineares.
- *K-Nearest Neighbors* (KNN): Modelo baseado nas instâncias mais próximas, que realiza previsões com base nos k vizinhos mais próximos dos dados de entrada. É simples e intuitivo, mas pode ser computacionalmente oneroso para grandes conjuntos de dados.
- *Gradient Boosting Regressor*: Modelo de *ensemble* que constrói árvores de decisão de forma sequencial, onde cada nova árvore corrige os erros das anteriores. É poderoso e capaz de capturar relações complexas nos dados, mas pode ser suscetível ao *overfitting* se não for devidamente regulado.

3 Resultados

Comparar as abordagens foi relevante para determinar se a aplicação do PCA melhora a precisão ou se os modelos são robustos o suficiente para lidar com os dados apenas pré-processados. A aplicação do PCA ajudou a reduzir a dimensionalidade dos dados, simplificando os modelos e diminuindo o risco de *overfitting*. No entanto, aplicar os algoritmos diretamente aos dados permitiu avaliar a robustez dos modelos em condições mais próximas dos dados reais.

Os resultados indicam que, embora a aplicação do PCA possa melhorar a precisão dos modelos, especialmente para algoritmos como *Linear Regression* e *Support Vector Regressor*, modelos como *Random Forest* e *Gradient Boosting Regressor* demonstraram ser mais robustos, mantendo um bom desempenho mesmo sem o PCA, como apresentado nas tabelas 1 e 2. Isso sugere que esses modelos são capazes de lidar com a complexidade e variabilidade dos dados apenas pré-processados de maneira eficaz.

Tabela 1: Resultados com aplicação de PCA

Parâmetro	Melhor Modelo	MSE	R^2 Score
Custo Inicial	Gradient Boosting	0.0817	0.9337
Capacidade Industrial	Gradient Boosting	0.1492	0.8578
Vida Útil	Gradient Boosting	0.2949	0.7253
Taxa de Recuperação	Random Forest	0.0775	0.8609
Custos de Sustentação	Random Forest	0.1619	0.8581
Custos de Mineração e Processamento	Random Forest	0.1287	0.7801

Tabela 2: Resultados sem aplicação de PCA

Parâmetro	Melhor Modelo	MSE	R^2 Score
Custo Inicial	Random Forest	0.0121	0.9879
Capacidade Industrial	Random Forest	0.0156	0.9839
Vida Útil	Random Forest	0.0337	0.9658
Taxa de Recuperação	Random Forest	0.0268	0.9579
Custos de Sustentação	Random Forest	0.0201	0.9805
Custos de Mineração e Processamento	Random Forest	0.0216	0.9651

4 Conclusões

Este estudo evidenciou que diferentes algoritmos de aprendizado de máquina apresentam eficácia variada na previsão de parâmetros técnicos e financeiros em projetos de mineração. O *Gradient Boosting Regressor* e o *Random Forest* destacaram-se pela sua eficiência na maioria dos casos. A seleção do algoritmo deve levar em conta o parâmetro específico a ser previsto e a natureza dos dados disponíveis. Os resultados iniciais obtidos podem contribuir significativamente para a otimização desses projetos, promovendo decisões mais informadas e eficientes. A aplicação do PCA mostrou-se vantajosa para aprimorar a precisão de alguns modelos, especialmente para algoritmos como *Linear Regression* e *Support Vector Regressor*. Entretanto, modelos como *Random Forest* e *Gradient Boosting Regressor* demonstraram robustez suficiente para lidar com a complexidade e variabilidade dos dados apenas pré-processados. A utilização de técnicas de aprendizado de máquina pode proporcionar uma vantagem competitiva substancial para empresas de mineração, permitindo previsões mais precisas de parâmetros críticos e, consequentemente, uma melhor tomada de decisão estratégica.

Referências

[1] H. Wasserbacher and M. Spindler *Machine Learning for Financial Forecasting, Planning and Analysis: Recent Developments and Pitfalls*, Digital Finance, 4, 63–88, 2022. DOI: 10.1007/s42521-021-00046-2.

[2] J. C. Munizaga-Rosas and K. Flores *Advanced Analytics for Valuation of Mine Prospects and Mining Projects. In: Advanced Analytics in Mining Engineering: Leverage Advanced Analytics in Mining Industry to Make Better Business Decisions*. Cham: Springer International Publishing, 2022. p. 95-145.

[3] N. Foo, H. Bloch, and R. Salim *The Optimisation Rule for Investment in Mining Projects*, Resources Policy, 55, 123-132, 2018.