IMPACTO CIENTÍFICO
EDITORA

MODELING THE COST OF CONSTRUCTING AN OFFSHORE DRILLING RIG USING RANDOM FOREST

MODELANDO O CUSTO DE CONSTRUÇÃO DE UMA SONDA DE PERFURAÇÃO MARÍTIMA ATRAVÉS DE FLORESTA ALEATÓRIA

MODELADO DEL COSTO DE CONSTRUCCIÓN DE UNA PLATAFORMA DE PERFORACIÓN MARINA UTILIZANDO RANDOM FOREST

https://doi.org/10.56238/edimpacto2025.089-004

Ricardo de Melo e Silva Accioly[1], Fernanda da Serra Costa[2]

ABSTRACT

Offshore drilling rigs are vital equipment for oilfield exploration and development; therefore, accurate estimates of their construction costs are crucial for planning rig construction projects. This paper explores the development of a random forest model to forecast offshore drilling rig construction costs. The model aims to provide accurate and reliable estimates of cost drivers based on a robust dataset that includes historical construction costs and rig design characteristics. The prediction-based ensemble learning approach used in the random forest algorithm effectively captures complex relationships and interactions in the data, improving forecast accuracy compared to traditional regression methods. The model's construction, validation, and significant findings will be detailed, highlighting its ability to minimize estimation errors and support decision-making in project budgeting. The objectives of this research encompass resource management, cost control, and strategic planning in rig construction projects.

Keywords: Random Forest. Offshore Drilling Rig. Cost Drivers. Random Forest.

RESUMO

As sondas de perfuração marítimas são equipamentos vitais para a exploração e o desenvolvimento de campos de petróleo, consequentemente uma boa estimativa de seus custos de construção é fundamental para o planejamento de projetos de construção de sondas. Este artigo explora o desenvolvimento de um modelo de floresta aleatória (random forest) para prever os custos de construção de sondas de perfuração marítimas. O modelo busca fornecer estimativas precisas e confiáveis dos direcionadores de custos, com base em um conjunto de dados robusto que inclui custos históricos de construção e características de projeto das sondas. A abordagem de aprendizado por combinação de previsões, utilizada no algoritmo de floresta aleatória, permite capturar de forma eficaz relacionamentos e interações complexas nos dados, melhorando a precisão da previsão em comparação com os métodos de regressão tradicionais. Serão detalhadas a construção, a validação e as descobertas significativas do modelo, destacando sua capacidade de minimizar erros de estimativa e de apoiar a tomada de decisões na elaboração do orçamento do projeto. Os

[1] Dr. in Production Engineering. Universidade do Estado do Rio de Janeiro (UERJ).
E-mail: raccioly@ime.uerj.br
Orcid: https://orcid.org/0000-0001-6513-3443

[2] Dr. PhD in Civil Engineering. Universidade do Estado do Rio de Janeiro (UERJ). E-mail: fcosta@ime.uerj.br
Orcid: https://orcid.org/0000-0002-4414-7755

objetivos desta pesquisa abrangem a gestão de recursos, o controle de custos e o planejamento estratégico em projetos de construção de sondas.

Palavras-chave: Random Forest. Sonda de Perfuração Marítima. Direcionadores de Custos. Floresta Aleatória.

RESUMEN

Las plataformas de perforación offshore son equipos vitales para la exploración y el desarrollo de yacimientos petrolíferos; por lo tanto, la estimación precisa de sus costos de construcción es crucial para la planificación de proyectos de construcción de plataformas. Este artículo explora el desarrollo de un modelo de bosque aleatorio para pronosticar los costos de construcción de plataformas de perforación offshore. El modelo busca proporcionar estimaciones precisas y confiables de los factores de costo con base en un conjunto de datos robusto que incluye costos históricos de construcción y características de diseño de las plataformas. El enfoque de aprendizaje por conjuntos basado en predicciones, utilizado en el algoritmo de bosque aleatorio, captura eficazmente relaciones e interacciones complejas en los datos, mejorando la precisión del pronóstico en comparación con los métodos de regresión tradicionales. Se detallarán la construcción, validación y hallazgos significativos del modelo, destacando su capacidad para minimizar errores de estimación y respaldar la toma de decisiones en la presupuestación de proyectos. Los objetivos de esta investigación abarcan la gestión de recursos, el control de costos y la planificación estratégica en proyectos de construcción de plataformas.

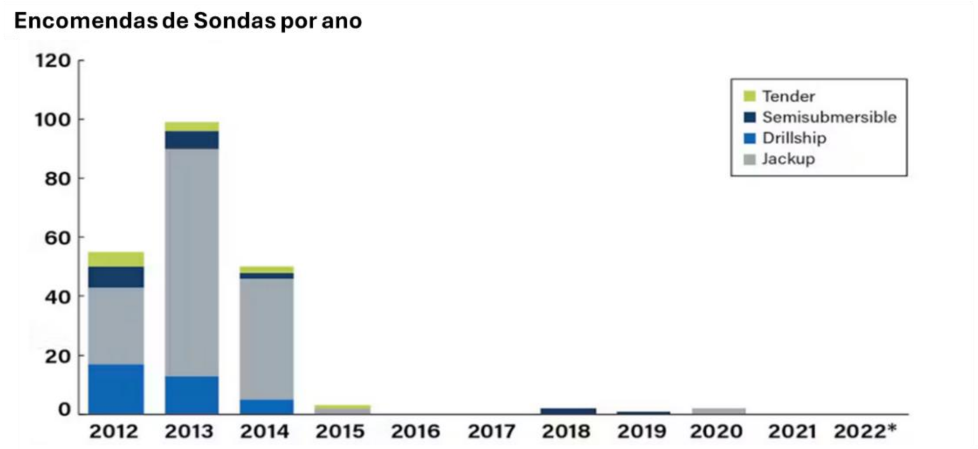
Palabras clave: Bosque Aleatorio. Plataforma de Perforación Offshore. Factores de Costo. Bosque Aleatorio.

1 INTRODUCTION

At the beginning of the 2010s, the market for the construction of platforms was extremely heated, with many new constructions, some of them speculative, which indicated that there were no contracts signed to finance them. As presented by Boman (2008), at that time, 140 mobile offshore drilling units (MODU) were under construction, with others ordered and in planning. In Figure 1, we see the rig construction orders from 2012 to 2022 (SMITH, 2022), where it can be seen that when oil prices plummeted in 2014, the rig construction market was heavily affected. This effect persists in the following years.

Figure 1

Orders for rigs between 2012 and 2022



Source: Smith (2022).

According to Offshore Magazine (2023), the construction of a new drillship requires an investment of almost 1 billion dollars. While there are currently no prospects for a new construction cycle, it is essential to assess the cost factors that influence the CAPEX of a new MODU. Kaiser and Snyder (2012) and Kaiser et al. (2013) extensively studied the offshore drilling industry and rig construction in the Gulf of Mexico and pointed out several factors that influence rig construction costs in their 2012 and 2013 studies, respectively. The model developed by them (multifactorial linear model) to estimate construction costs, due to its characteristics, limited the type of relationship between the response variable and the explanatory variables to a linear pattern.

Breiman et al. (1984) created tree models, which are powerful for classification and regression tasks. The method divides the data into subsets based on explanatory variable values, creating a tree-like structure that represents decisions that lead to outcomes. However, models based on a single tree are generally not competitive for prediction accuracy and model stability. In addition, the number of predictions of a tree is finite and is determined by the number of terminal nodes. Finally, these trees suffer from a selection bias: predictors with more distinct values are favored over more granular predictors (HASTIE et al., 2009).

Subsequently, Breiman (1996) improved the initial development through techniques of combination of predictions (ensemble), using what he called "bagging". "Bagging" is an abbreviation for "bootstrap aggregation." In 2001, he improved his methodology through predictions with random forests, which corrected the problem of correlation between the trees generated by the bootstrap samples, further improving the method.

This type of modeling has several interesting features, such as, it accepts different types of variables (sparse, asymmetric, continuous, categorical, etc.) without the need for pre-processing; it does not require the user to specify how the explanatory variables relate to the answer, as a linear regression model does; it can effectively deal with missing data and implicitly lead to the selection of variables; desirable aspects for many real-life modeling problems.

2 THEORETICAL FRAMEWORK

2.1 TREE-BASED MODELS

In the 1990s, combination techniques, or "ensemble" (methods that combine predictions from multiple models), began to emerge. *Bagging*, an abbreviation for "bootstrap aggregation", was originally proposed by Leo Breiman and was one of the first combination techniques developed (BREIMAN, 1996). Bagging is a general approach that uses bootstrapping in conjunction with any regression (or classification) model to generate a set of predictions.

Bootstrap aggregation, or *bagging*, is a general-purpose procedure to reduce the variance of a statistical learning method. It is particularly useful and often used in the context of trees. Recall that given a set of n independent observations Z1, Z2, ..., Zn, each with variance , the variance of the mean of $\sigma^2 n\sigma^2/n$ observations is given by . In other words, averaging a set of observations reduces variance. In general, this is not practical because we do not have access to various training sets. On the other hand, we can use bootstrap, obtaining new "samples" through resampling with replenishment of the (single) training dataset. In this approach, we generate B distinct training datasets. We then train our method on the b-th set of *bootstrap* training in order to get a prediction at point x.$\hat{f}^{*b}(x)$

The average of all predictions is:

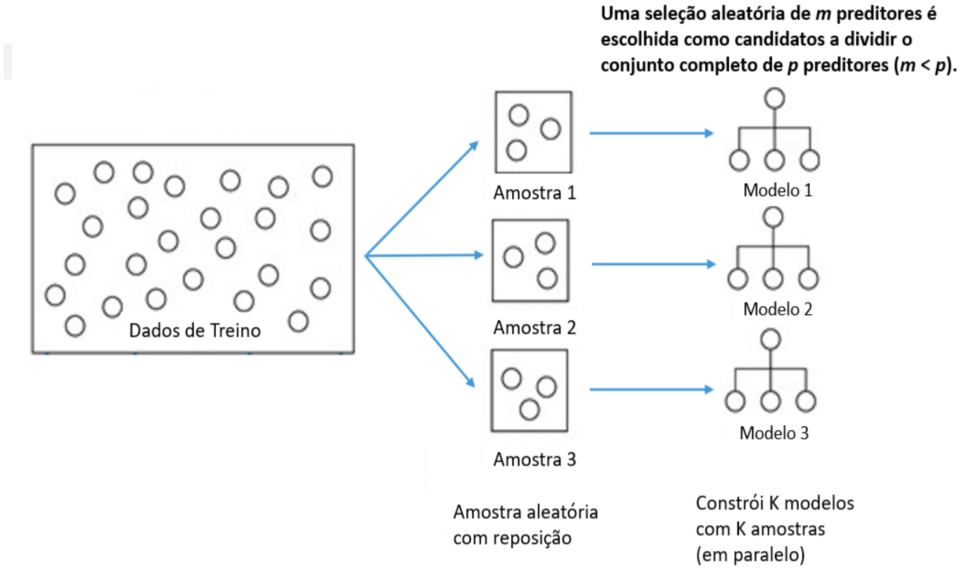
$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \qquad (1)$$

The random forest provides an improvement over *bagging* through a small adjustment that decorrelates the trees. This reduces variance when averaging the trees. In this method, when constructing the trees, each time a division in a tree is considered, a random selection of *m* predictors is chosen as candidates to divide the full set of *p* predictors (*m* < *p*). Division is only allowed for one of these *predictors* . A new selection of *m* predictors is performed in each model.

In the context of regressions, Breiman (2001) recommends that *m* be equal to one-third of the number of predictors. Figures 2 and 3 illustrate the concepts mentioned above.

Figure 2

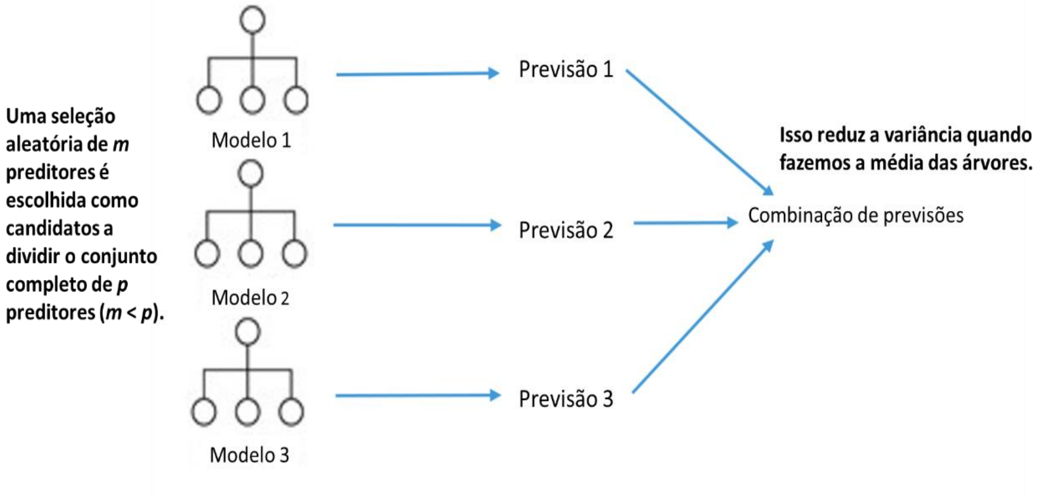
Representation of the random forest methodology


Uma seleção aleatória de *m* preditores é escolhida como candidatos a dividir o conjunto completo de *p* preditores (*m* < *p*).
Dados de Treino
Amostra 1
Amostra 2
Amostra 3
Modelo 1
Modelo 2
Modelo 3
Amostra aleatória com reposição
Constrói K modelos com K amostras (em paralelo)

Source: The authors (2025).

Figure 3

Representation of the concept of ensemble of predictions


Uma seleção aleatória de *m* preditores é escolhida como candidatos a dividir o conjunto completo de *p* preditores (*m* < *p*).
Modelo 1
Modelo 2
Modelo 3
Previsão 1
Previsão 2
Previsão 3
Isso reduz a variância quando fazemos a média das árvores.
Combinação de previsões

Source: The authors (2025).

2.2 INTERPRETATION OF COMPLEX MODELS

An important issue when working with forecasts is to interpret them, in order to understand and justify them. Combination techniques, or "ensemble" (methods that combine predictions from many models), have increased the predictive capacity of models by reducing prediction errors. However, in the case of trees, by combining the responses of several trees, the ease of interpretation of a single tree was lost. This problem occurs in all models that use the prediction combination, but also in other models that are commonly referred to as black-boxes, such as neural networks and SVM. To improve this issue, some techniques were introduced, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), according to Molnar (2019). Here, we will use LIME, developed by Ribeiro et al. (2016).

LIME is an approach based on the idea that complex models can be approximated locally by simple models. To interpret the prediction of a given point:

Multiple disturbed versions of this observation are generated (for example, by slightly changing the values of the variables).

The complex model is used to predict these versions.

A linear (or simple) model is adjusted locally based on these predictions.

The coefficients of this explanatory model indicate the importance of the variables in the local prediction.

3 METHODOLOGY

The random forest methodology requires tuning a set of parameters to define the behavior of the prediction model. The main parameters of this model are: the number of trees, the number of variables considered in each division, the minimum size of observations in each terminal node (leaf), the division rule and the method of calculating the importance of the variables. In the problem in question, the division rule is variance, a common method in regression problems. On the other hand, permutation will be used to calculate the importance, because the computational cost in this problem is small, but allows for more robust and reliable evaluations (BREIMAN, 2001 and STROBL et al., 2007).

The other parameters will be defined through cross-validation, using a grid of values for each of them.

The database contains 101 offshore drilling rigs, delivered and under construction between 2001 and 2010, and was collected in 2011 through consultations with rig operators, publications from the International Association of Drilling Contractors (IADC), Offshore Magazine and other sources. From the database with 40 explanatory variables (features) related to these probes, a random forest model was built to predict and evaluate the importance of each trait. These explanatory variables act as cost drivers and are essential for predicting actual construction costs. Table 1 presents the explanatory variables, with the dependent variable being the CAPEX in 2011 MMUS$

Table 1

Database features

Variable Name	Description
Rig Type	Type of Probe (Semi-submersible or Drillship)
Rig Water Depth	Water Depth (feet)
DP	Dynamic Positioning Probe (Y/N)
DP Class	Dynamic Positioning Class
Design Water Depth	Design water depth (feet)
Drilling Depth	Drilling depth (feet)
BOP WP Max	Maximum Pressure Supported by BOP (psi)
Draft	Submerged Length (feet)
Dual Activity	Dual Activity (Y/N)
EWT Capable	EWT Drilling Capacity (Y/N)
Harsh Environment	Ability to operate in harsh environment (Y/N)
North Sea Capable	North Sea Operating Capability (S/N)

Zero Discharge	Zero discharge (S/N)
Thruster Assist	With the presence of thrusters (S/N)
Top Drive	With presence of topdrive (Y/N)
Quarters Capacity	Accommodation capacity
Variable Load	Variable load (mt)
Hull Newbuild	New Hull (S/N)
AnoTerm	Year of completion of Construction
AnoPed	Year of probe order
CAPEX 2011	2011 Capex in MMUS$
Spec Build	Speculative construction (Y/N)
Year Hull Built	Year of hull construction
Year In Service	Years in operation
Norwegian SUT	Norway Declaration of Conformity (Y/N)
Hull Length	Hull length (feet)
Hull Breadth	Hull width (feet)
Hull Depth	Hull Depth (feet)
Bulk Cement	Cement capacity (cubic feet)
Bulk Mud	Mud capacity (cubic feet)
Liquid Mud	Mud Capacity (bbl)
Drill Water	Industrial Water Capacity (bbl)
Potable Water	Drinking water capacity (bbl)
BOP Qty	Number of BOPs
Engine Qty	Number of machines
Generator Qty	Number of Generators
Derrick Or Mast	Towers or Mast (T/M)
Derrick Capacity	Turret Load Capacity (lbs)
Mudpump Qty	Quantity of Mud Pumps
Tensioner Qty	Riser Tensioners quantity
Tensioner Capacity	Riser Tensioner Capacity (lbs)

Source: The authors (2025).

To facilitate the interpretation of the forecast results, the LIME methodology was used (MOLNAR, 2019).

4 APPLICATION

For the application of the random forest model, a program was developed in the R language, version 4.5.1, using the ranger package (WRIGHT and ZIEGLER, 2017), version 0.17.0. To interpret the results, the lime library was used (HVITFELDT et al., 2022), version 0.5.3. The programming environment was RStudio 2025.09.1.

The database contained data from 101 probes, semi-submersible probes and drillships, with 40 technical specifications (features), which were separated into two sets: 75% used for training and the rest for model testing. To define the best set of parameters, a grid of values was defined for the number of trees (num.trees in the ranger), the number of variables considered in each division (mtry), the minimum size of observations in each terminal node (min.node.size) and the division rule (*splitrule*: variance only). The grid contained 70 different configurations. Table 2 presents the selected values.

Table 2

Random forest model parameter grid

Parameter	Values
num.trees (# trees)	100, 200, 300, 400, 500
mtry (# max. explanatory var. in the division)	2, 4, 6, 8, 10, 12, 14
min.node.size (# min. of node observations)	1, 5

Source: The authors (2025).

The choice of the best parameters was performed by means of cross-validation in the training set, with 5 data envelopes. Table 3 shows the results.

Table 3

Best (five) results of the cross-validation

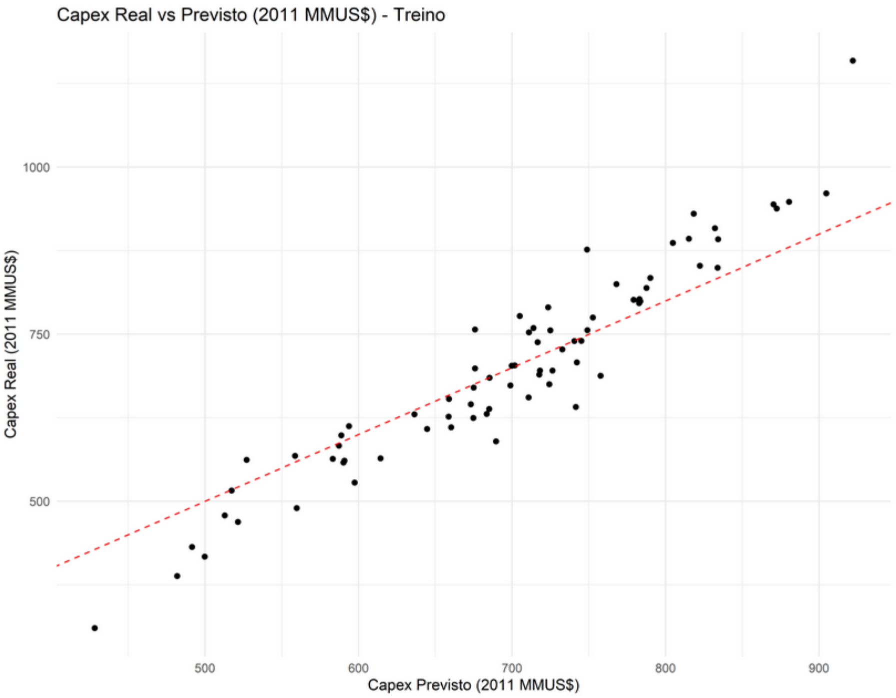
mtry	min.node.size	splitrule	num.trees	Average RMSE
2	1	Variance	100	105,2259
2	1	Variance	200	106,0781
2	1	Variance	300	106,3491
2	1	Variance	400	106,3901
2	1	Variance	500	106,4608

Source: The authors (2025).

Figure 4 shows the comparison between the predicted values and the actual values (training set), showing a good performance of the model.

Figure 4

Comparison between real value and forecast with random forest - Training



Source: The authors (2025).

Figure 5 shows the comparison between the predicted values and the actual values (test set), showing a good performance of the model.

Figure 5

Comparison between real value and prediction with random forest – Test

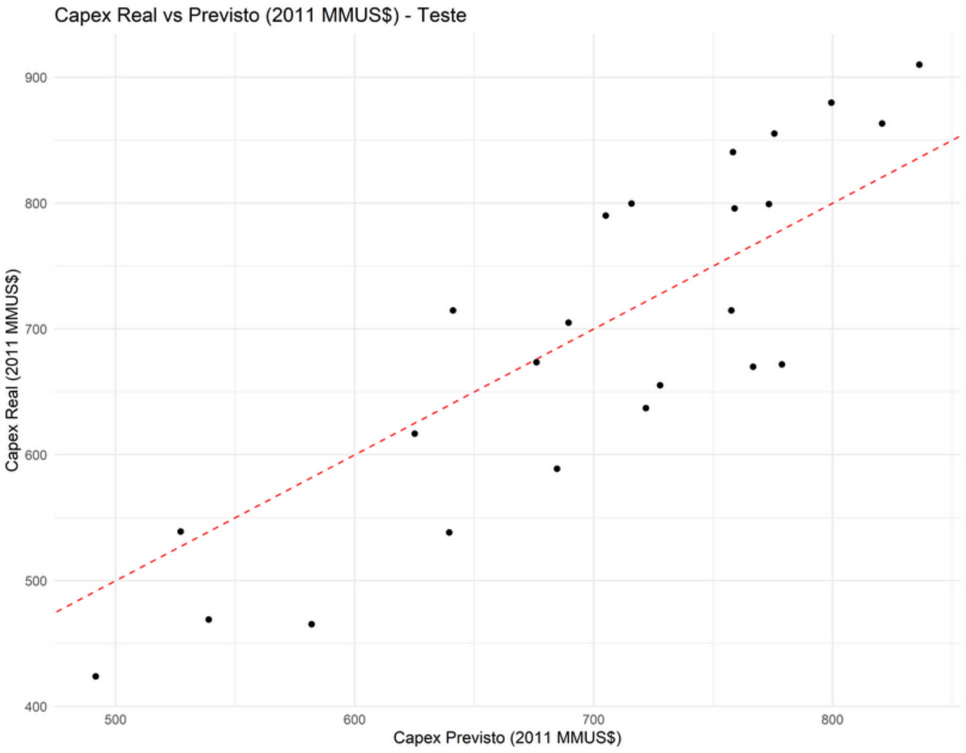


Table 4 shows the comparative statistics of the results of the adjustments of the training and test sets.

Table 4

Metric by dataset (training and testing)

Ensemble	RMSE	MAE	MAP	R2
Training	58,14	44.44	6.70	0,91
Test	72,63	64,89	9,94	0,76

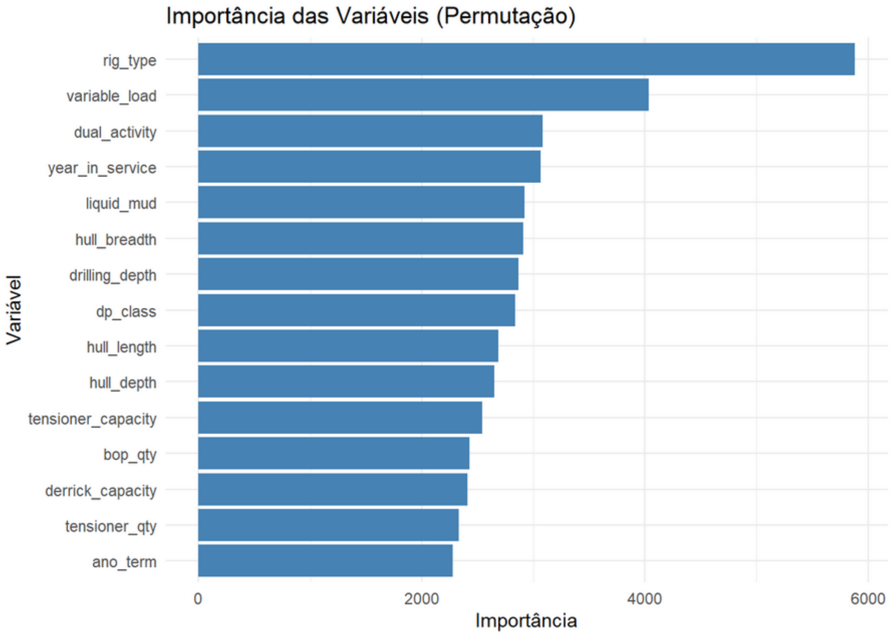
Source: The authors (2025).

The statistics indicate a good performance of the model both in the training and testing phases. The most important result is the test result, in which, observing the MAPE, we have a predicted error of around 10%, which can be considered adequate in this context.

The importance of the variables can be seen in Figure 6. Only the 15 most important variables were presented. We can observe that the type of probe was the most important variable. Drillships have a higher construction cost than semi-submersibles, as we can see in Kaiser et al. (2013). The drillship has a higher variable load capacity (variable_load) than semi-submersible rigs.

Figure 6

Importance of the model variables



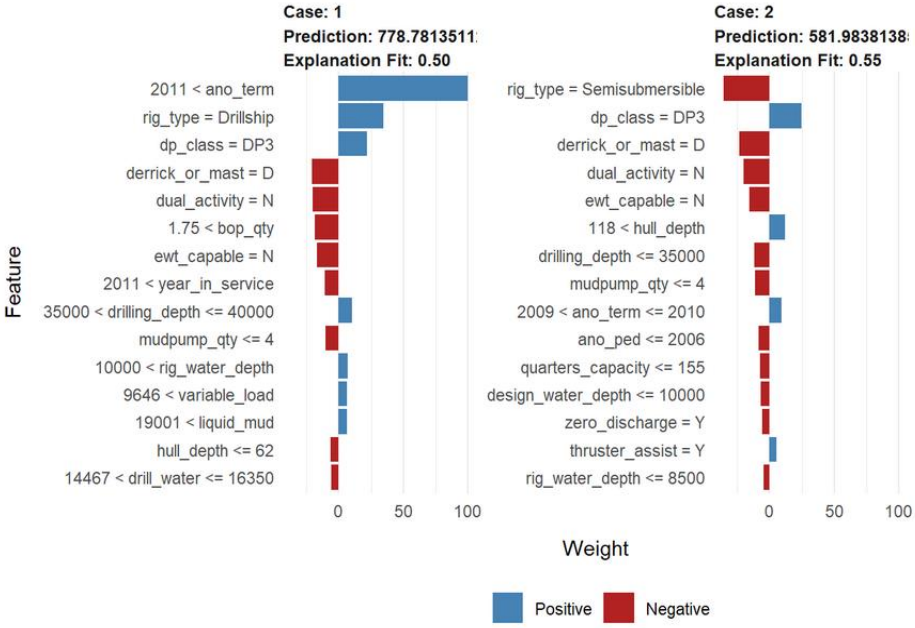
Source: The authors (2025).

In Figure 7 we see the LIME result in the 1st and 2nd observations of the test set. In it, we can see the impact of the type of rig: increase in value when it is a drillship and loss of value when it is a semi-submersible rig. The type of dynamic positioning class was positive in both observations. Another interesting point is that the two probes lost value because they did not have dual activity (dual_activity). Rigs with dual activity have two towers, which allows speeding up certain operations, that is, saving time in the construction of wells.

In these two examples, it was possible to explore the factors that affected the forecasts, which is very important to understand and justify them. This analysis also allows us to identify the sources of deviation from the actual values.

Figure 7

Evaluation of the results of the 1st and 2nd observations of the test set



Source: The authors (2025).

4 CONCLUSIONS

The random forest represents a powerful methodology in the construction of prediction models, as it works with several types of variables and allows complex (nonlinear) relationships between them. The identification of the importance of the variables is essential to select the most relevant ones, which would facilitate the obtaining of a more parsimonious model.

The existence of techniques for interpreting complex models solves the difficulty of understanding the results of these models, allowing the identification of what led to the achievement of a certain prediction. In this work, we used LIME, but SHAP (MOLNAR, 2019) has been expanded in use.

As for the results, the model allowed us to identify that the type of rig is the factor with the greatest impact on the construction cost. Secondly, the variable load capacity of the probe. Thirdly, the dual activity, i.e. the ability to carry out simultaneous operations. This type of information is important in the process of selecting the technical characteristics of a project.

REFERENCES

Boman, K. (2008). Rig construction boom continues. Offshore Magazine. https://www.offshore-mag.com/business-briefs/equipment-engineering/article/16761659/rig-construction-boom-continues

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1986). Classification and regression trees. Wadsworth and Brooks/Cole Advanced Books & Software.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hvitfeldt, E., Pedersen, T. L., & Benesty, M. (2022). lime: Local interpretable model-agnostic explanations. https://doi.org/10.32614/CRAN.package.lime

Kaiser, M. J., Snyder, B. F., & Pulsipher, A. G. (2013). Offshore drilling industry and rig construction market in the Gulf of Mexico (OCS Study BOEM 2013-0112). Louisiana State University Center for Energy Studies Coastal Marine Institute.

Kaiser, M. J., & Snyder, B. (2012). Reviewing rig construction cost factors. Offshore Magazine. https://www.offshore-mag.com/business-briefs/equipment-engineering/article/16760123/reviewing-rig-construction-cost-factors

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. Lulu.com.

Offshore Magazine. (2023). Report: No resurgence seen for newbuild rig construction. https://www.offshore-mag.com/rigs/article/14298424/westwood-global-energy-group-report-no-resurgence-seen-for-newbuild-rig-construction

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. arXiv:1602.04938. https://doi.org/10.48550/arXiv.1602.04938

Smith, J. (2022). Rig construction market remains quiet but with room for long-term possibilities. Offshore Magazine, 82(6).

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77(1), 1–17.