WellCast: A model to predict expected duration of offshore wells

Diego Eduardo Mercadal
Principal Data Scientist
Diego.Mercadal@valaris.com

Abstract. Drilling an offshore well is a combined effort between stakeholders, and a technically challenging activity with several sources of variability. This paper details a study of historical wells drilled by Valaris, and using statistical methods and machine learning techniques, identifies the best candidates for well duration predictors. These variables are then fit in a series of models that are trained, tested, and validated. The selected model is used to predict scenarios for each well phase duration for recent wells, and then tested against previously unseen data.

Keywords: Oil & Gas, Offshore Drilling, Statistical Modeling, Machine Learning, Predictive Analytics, Unsupervised Learning.

1) Introduction

Valaris has been drilling offshore wells using a diverse fleet in all major basins for several years, using state of the art technology and people skills. This has resulted in the world's most technologically advanced offshore drilling operations while providing safe and efficient solutions to deliver energy to the world. Combining experience with data has proven to be a promising path for the future, with endless applications within our business.

Estimating the duration of a well prior to its start isn't an easy exercise. When drilling in a remote location in an unknown area, duration is only one of the many question marks that are part of this exercise. Operators have their own financial models based on data that is not available to drilling contractors, making it even harder to predict. Additionally, unplanned events can cause dramatic changes to the schedules. In an environment in which the major part of the variables is encountered for the first time or only a few times before, the probability of having these types of events is higher.

Analyzing Valaris' historical data in wells drilled over the years helps understand how some of these variables can explain such variability in the time required to execute a well or well phase. These datapoints can be grouped in main categories as described below:

- **Location Related:** Knowing where the well is going to be drilled helps analyze past wells with same characteristics and provide a better estimate. Latitude, Longitude, Water Depth are examples of datapoints classified in this category.
- **Phase Related:** Bit diameter, starting depth and casing shoe depth, for example, can help explain variability. The time spent drilling a top hole should be much less than a deep phase in a hard formation.
- **Rig Related:** the rig, rig type, and class also play an important role. The age of the rig by the time of the project also has effect, and should be considered.
- **Client:** Every client has its own budget, plan and threshold for risk.
- **Seasonality:** Seasonality directly affects workers, equipment, weather, and sea patterns.

This work recognizes the duration of a well phase as a random variable that is influenced by the factors describes above, and attempts to find a viable regression model that could reasonably predict, the duration of a phase to be drilled. The relationships between Locations, Clients, Rigs and Seasonality were tested using machine learning techniques with the objective of finding the best candidates to work as a predictor for Phase Duration. With the best model selected some predictions are made with unseen data, represented by more recent wells that were drilled after the model was developed.

The paper has four more sections: section 2 explains the origin of the data, and what was needed to acquire and clean it in a reliable manner. This section also presents an explanation of each variable selected for the study and some discussions on how to apply the process in production. Section 3 presents all methods, assumptions, and analysis used in this work. This is the most technical section of the paper, but is necessary to deliver confidence in the results. An attempt was made to use as little technical jargon as possible to make the reading comprehensible to a more general audience. Section 4 presents the results and performance metrics of the work and simulates some scenarios. Readers more interested in the results and less on the process, can jump directly to this section. Finally, section 5 discusses the results from section 4.

2) Data

a) Data Sources

This study used four main data sources. Three of them are internal to and controlled by Valaris, while the fourth is maintained by an open-source research institute in Norway, the Peace Research Institute Oslo (PRIO). **RIMDrill** and **ROS** provided well history data through the International Association of Drilling Contractors (IADC) reports. When available, VIP data from rig sensors was used to confirm what was reported in both systems, more specifically well and bit depths. The Peace Research Institute Oslo (PRIO) is a research institute that maintains, among others, an open-source database of geographical and resources datasets. For this project the PRIO Database was used to cross reference well locations with the major oil basins. Data was aggregated and processed using **Databricks,** a data science platform hosted on Azure cloud, shown in Figure 1 below:

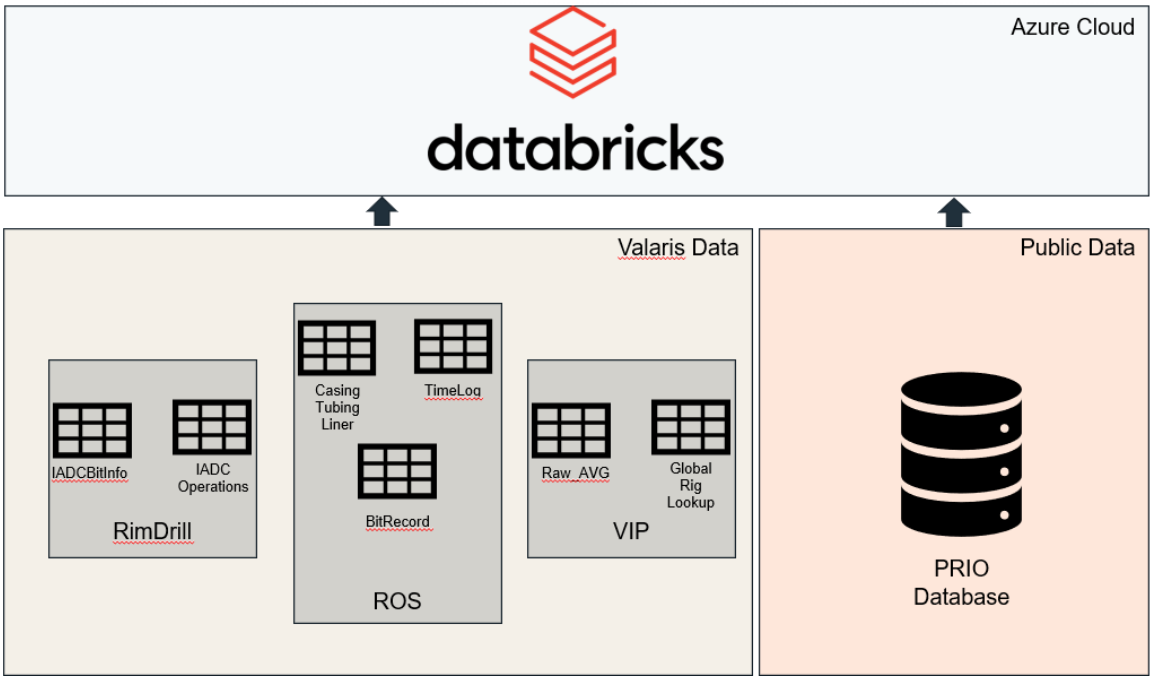

Figure 1: Data Acquisition Architecture

b) Data Cleansing

Data Quality is the backbone for every data analysis project, including this study. Most of the data came from manual inputs and where system controls were limited. The resulting errors can be classified in the following categories:

- **Typos**: Country and Client Names lead this category. Hundreds of variations for the same client were found, and these data points were all reconciled using natural language processing (NLP).
- **Latitude/Longitude**: Users recorded the wrong hemispheres by forgetting a (-) signal as part of the coordinates. Additionally, there were situations where the recorded coordinates placed the well inland; yielding no alternative other than to remove the data point.
- **Unrealistic Numbers**: Some depths informed in the reports were unrealistic, independently of the unit of measure considered. In this case, the data points we also removed.
- **Missing data**: For variables in which imputation seems reasonable, such as Client name, a K-nearest neighbor imputation was made. This means a model was created to analyze well phases with the same characteristics of the one missing the client, and the most probable client was added to the data point.

Once bad data was cleaned or eliminated, a key assumption in this process was that all remaining data was assumed to be true:

- All sensor data was taken to be accurate as it's relied upon for operations
- Transactional data from ROS and RIMDrill was assumed true (e.g., country, bit diameter, client, etc.)

Two models were trained in this study: one for cases where sensor data was not available to supplement IADC reports (Model 1), and another where it was (Model 2). The large disparity in population size between the two cases led to the development of two models.

- **Total Population**: approximately 7,500 well phases
- **Model 1**: approximately 4,751 well phases after data quality issues removed
- **Model 2**: approximately 360 well phases with supplemental sensor data

c) Variables

The main goal of the models proposed in this work was to provide a reliable estimation for the expected duration of an offshore well. To achieve that, they assume that wells are a combination of phases that have their own characteristics and therefore can also be predicted individually. This approach brings a key advantage to the analysis, allowing comparison of phases from different wells and generalizing the model. A phase in this work is defined by the Bit size reported in the IADC reports; a phase starts when a new bit size is reported for the first time and ends when a new size is reported.

A total of 4,751 observations and 25 variables were used for the study, including wells drilled by Valaris starting in December 2010 until April 30th, 2020. Data from wells starting after this date were used on Section 4 to evaluate results. Table 1 presents and describes the variables used in this project:

Table 1: Variable Definition

Variable Name	*Category*	*Data Type*	*Description*
Diameter Cluster	Calculated	Numeric	Result of a K-Means Cluster of all reported Bit diameters. This intends to organize phases in a logic and sequential manner
Formation Proxy	Calculated	Numeric	Linear Combination of Latitude, Longitude and Bit Diameter, plus a K-means cluster. This intends to group phases with the same pattern, which would be an indication of similar formation
Abnormal Phase	Calculated	Categorical	Phases that are not simply vertical. Directional, horizontal or any other special drilling fall into this category
Phase Signature	Calculated	Numeric	Linear Combination of all variables, plus a K-means cluster. This means to group phases based on similar pattern.
Rig Class	Raw - Report	Categorical	Valaris Rig Class
Rig Year Built	Raw - Report	Numeric	Rig Year Built
Rig Age	Calculated	Numeric	Age of the Rig at the time the phase started
Start Depth	Raw - Sensor	Numeric	Start Depth (feet) for the phase
End Depth	Raw - Sensor	Numeric	Final Depth (feet) for the phase
Phase Length	Calculated	Numeric	Delta(feet) between start and end depths for a phase
Outlier Phase	Calculated	Categorical	Phases that were not fully finished; for example, when a plug is set and drilling restarts after some time
Hemisphere	Calculated	Categorical	Hemisphere where the phase is being drilled
Start Month	Raw - Report	Categorical	Starting Month of the project
End Month	Raw - Report	Categorical	Ending month of the project
Start Season	Calculated	Categorical	Season by the time the phase starts
End Season	Calculated	Categorical	Season by the time the phase ends
Rig Name	Raw - Report	Categorical	Rig Name
Rig Type	Raw - Report	Categorical	Rig Type
Latitude	Raw - Report	Numeric	Well Head Latitude
Longitude	Raw - Report	Numeric	Well Head Longitude
Country	Raw - Report	Categorical	Country where the well is located
Field	Calculated	Categorical	Calculated Based on lat/long cross referenced with PRIO's database
Client	Raw - Report	Categorical	Client Name
Water Depth	Raw - Report	Numeric	Well water depth
Phase Diameter	Raw - Report	Numeric	Bit Diameter

3) Methods

a) Exploratory Analysis

The main objective of this work was to define a viable model for the duration of offshore wells. This model should be able to explain the variability of the duration in days shown on Figure 2 based on the predictors selected and presented above.

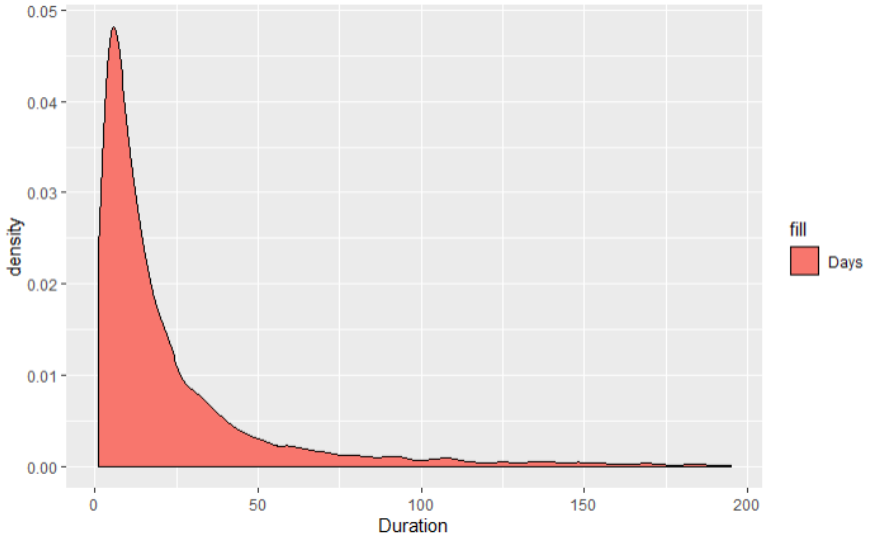

Figure 2: Variability of Valaris Well Phases (Days)

We began the process by analyzing our data, checking our distributions and effects of the predictors on the duration of the well phases. This helped understand the variables and their initial relationship with the explanatory variable and some initial patterns. Figures 3 and 4 present the variables plot for the Model 1 and Model 2 datasets, respectively.

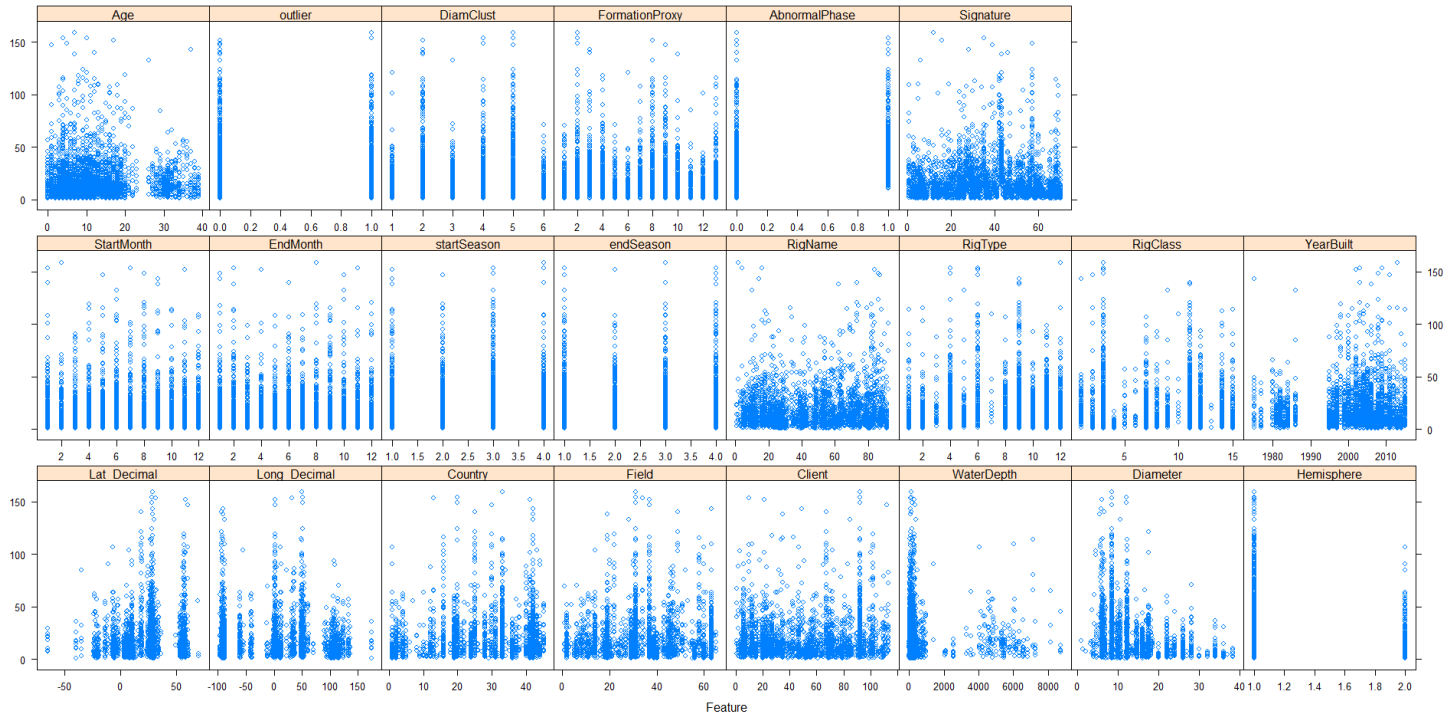

Figure 3: Variable Plot for Dataset with no Phase Depth Info (Model 1)

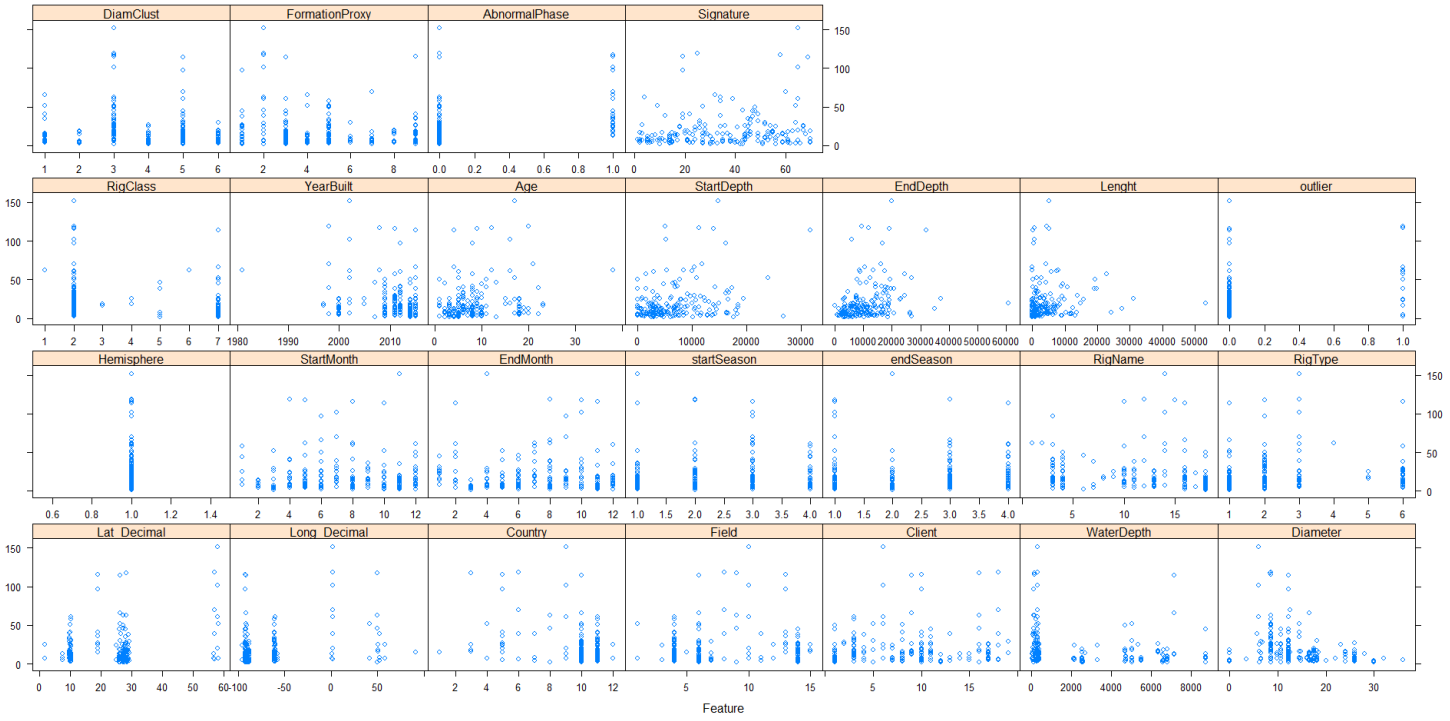

Figure 4: Variable Plot for Dataset with Phase Depth Info (Model 2)

After analyzing the relationship between the variables, the correlation between these variables was evaluated and strongly correlated pairs were removed in accordance with best practices. In this step a commonly threshold of 0.80 was selected, meaning pairs with correlation above 0.80 and below -0.80 had one of the variables removed from the data. There is no consensus in the literature regarding the threshold, and the value of 0.80 was selected based on the experience of the author with large datasets. Figure 5 presents a correlation matrix with a color code highlighting strong relationships that needed to be filtered. Values were tested at the 99.9% confidence level, meaning correlations with p-value > 0.01 were considered insignificant. Then, applying our threshold of 0.80 the variable **Year Built** is removed from the dataset.

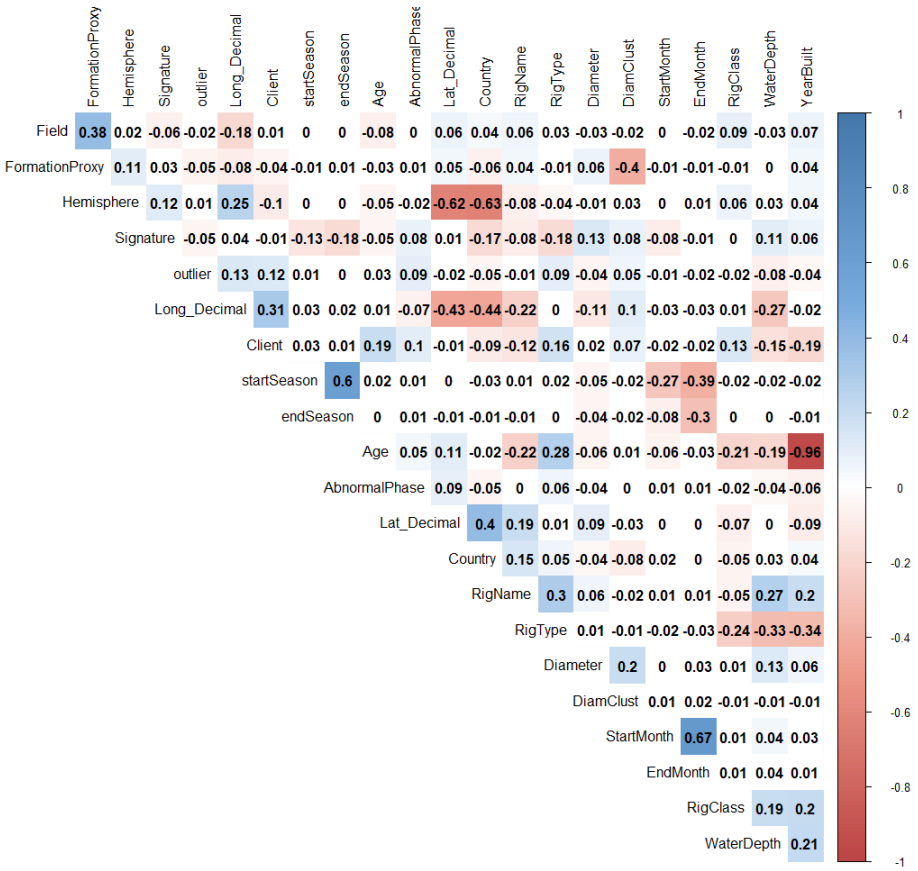

Figure 5: Correlation Matrix at 99.9% Confidence Interval

b) Pre-Process

During this phase, all needed adjustments were made to the dataset before the modeling phase began. This included applying the transformations needed, splitting it between train and test, and defining the tuning parameters to be used.

i) Transformations

The first step was to transform heavy skew data. This was intended to convert the data closer to the normal distribution and meet the assumptions for the inferences. During this step, all numeric variables were transformed using a Box-Cox transformation, which is defined by:

$$f_\varphi(x) = \frac{x^\varphi - 1}{\varphi}$$

If $\varphi \neq 0$. For $\varphi = 0$,

$$f_0(x) = \log(x)$$

From the variable plots there was considerable difference in terms of scale in the data. To fix that a Z-score normalization was applied. It consisted of subtracting the mean and dividing by the standard deviation. In such a case, each value would reflect the distance from the mean in units of standard deviation.

Assuming that after the Box-Cox transformation all variables will be closer to the normal distribution, then this normalization process would bring them all close to the standard normal distribution. The resulting distribution had a mean of 0 and a standard deviation of 1.

ii) Dummy Variables

When a predictor is categorical, such as rig name or rig class, it is common to decompose the predictor into a set of more specific variables. To use these data in models, the categories are re-encoded into smaller bits of information called “dummy variables.” Usually, each category has its own dummy variable that is a zero/one indicator for that group. By decomposing the categorical variables into binary features the total number of predictors changed from 25 to 366.

After adding these new data points, the same process described in item i) was applied. The correlation between predictors was calculated and the highly correlated variables were dropped. Appendix A details the 60 features that were dropped due to high correlation:

iii) Dimension Reduction

Dimension reduction techniques are another class of predictor transformations. These methods reduce the data by generating a smaller set of predictors that seek to capture most of the information in the original variables. In this way, fewer variables can be used that provide reasonable fidelity to the original data. For most data reduction techniques, the new predictors are functions of the original predictors. Therefore, all the original predictors are still needed to create the surrogate variables. This class of methods is often called signal extraction or feature extraction techniques.

Principal Component Analysis is a commonly used data reduction technique (Abdi and Williams 2010). This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance. The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs. Mathematically, the jth PC can be written as:

$$PC_j = (a_{j1}\ x\ Predictor\ 1) + (a_{j2}\ x\ Predictor\ 1) + \cdots + (a_{jP}\ x\ Predictor\ P)$$

P is the number of predictors. The coefficients aj1, aj2,. . . , ajP are called component weights and help to understand which predictors are most important to each PC.

Use of this technique made it possible to capture 95% of data’s variability with 205 features, meaning a reduction in 35% of the number of features. This was an important step: reducing the data dimensions reduced noise and optimized the model.

iv) Phase Signature

After reducing the dimensions, the data was grouped and labeled based on similarity using clustering techniques. When predicting the duration for a new phase, the expected range for phases with the same signature could be used to determine the prediction interval. The technique used to assign these labels to the phases was a K-Means Cluster, one of the most prolific "clustering"

algorithms. K-means stores k centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid (i.e., center) than any other centroid.

K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids, and (2) choosing centroids based on the current assignment of data points to clusters. The algorithm is as follows:

1. Initialize Cluster Centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbf{R^n}$ randomly.
2. Repeat until convergence: {

For every *i*, set

$$c^{(i)} := \arg\min ||x^{(i)} - \mu_j||^2$$

For every *j*, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

}

v) Data Splitting and Cross Validation

Data was split randomly into two datasets: train and test. In following data analysis best practices, 75% of the data was used for training the model, while 25% was used for testing the model. While tuning the model, resampling was used with cross validation and 2 rounds with 10 folds, following the steps described below:

1. Data was randomly divided in two groups: 25% reserved as testing data, and 75% as training data;
2. The training data (75%) is randomly divided in ten groups that are trained and tested
3. Steps 1 & 2 are repeated
4. Performance of the models is used to build the official model that is output.

c) Data Modeling

During this phase, a set of models were fit to the data in order of complexity.

1. An elastic net linear model that was used to confirm the significance of the selected variables
2. A pruned decision tree
3. A random forest.
4. A support vector machine
5. A cubist regression
6. A stochastic gradient boost.
7. Neural network

A description and graphical results of each model can be found in Appendix B. Performance metrics and model evaluation are discussed in Section 4 where the results are discussed. With any regression, the following assumptions need to be considered to determine a valid model:

- Normality of residuals- the residual errors are assumed to be normally distributed.
- Homogeneity of residuals variance- the residuals are assumed to have a constant variance (homoscedasticity).
- Independence of residuals error terms.

As stated in Section 2, this work used two datasets: Model 1 which excludes sensor data and the phase depth variable, and Model 2 which includes sensor data. All modeling techniques described in the next pages were applied to both datasets.

d) Prediction Interval

Section 2, item iv describes the technique for estimating prediction intervals for this project. The phases were grouped using a K-means cluster technique to select members due to its Euclidean distance. By assuming this exercise involved all wells drilled by Valaris, there is a very high probability that any new well could be classified in one of these existing clusters. Based on that, the phase to be drilled is added to one of the 300 existing clusters before the prediction. Then, using the durations of these phases a T (student) distribution is fit, and the quantiles of this distribution are used to estimate the prediction intervals of the new point. The T distribution is selected over the normal from its characteristics of a heavy tail. This means by fitting durations to this distribution the probability of a tail event is accounted for, which may be translated to a downtime with a long duration.

4) Results

This section presents the results of the work, starting with the presentation of the variable importance for the models developed. This is important as one of the objectives is to understand the relationships of the variables with the output, the Well Phase duration. It helps explain why some wells, even though similar in some aspects, have completely different expected durations.

a) Variable Importance

Variable Importance represents the statistical significance of each variable in the data with respect to its effect on the generated model. It is actually each predictor ranking based on the contribution predictors make to the model. This technique helps data scientists eliminate certain predictors that contribute little but increase processing time. Figures 6 and 7 below show the normalized variable importance for the models developed in this work:

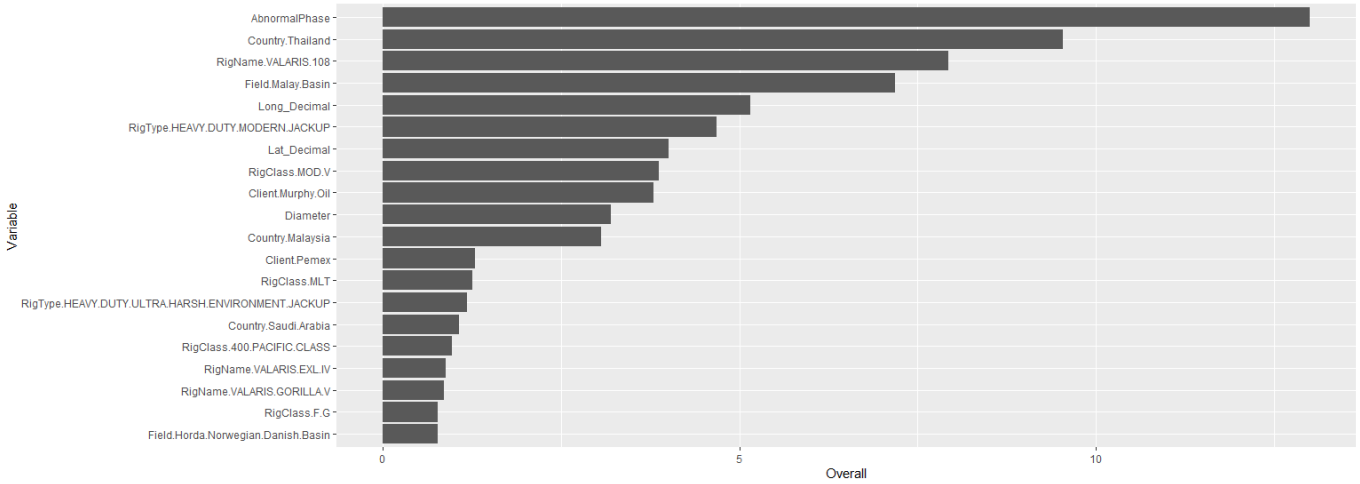

Figure 6: Variable Importance for Model 1 not using Depths

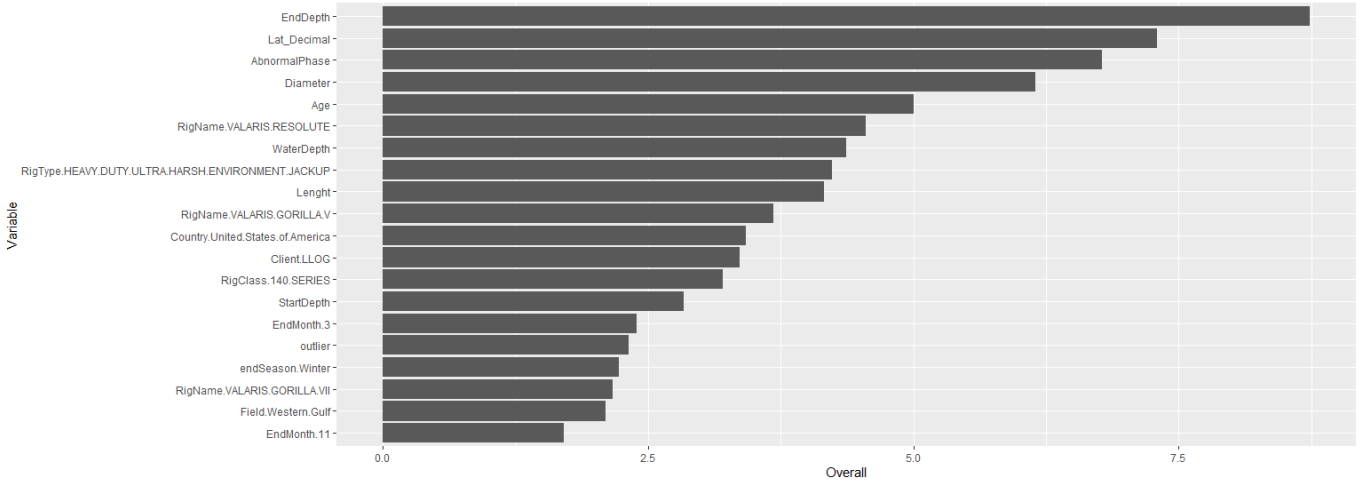

Figure 7: Variable Importance for Model using Depths

b) Performance Metrics

When working with predictive models, there are different ways to measure accuracy, each with its own nuance. For this reason, relying solely on a single metric is problematic. Visualizations of the model fit, particularly residual plots, are critical to understanding whether the model is fit for purpose. The most common method for characterizing a model's predictive capabilities is to use the root mean squared error (RMSE). This metric is a function of the model residuals, which are the observed values minus the model predictions. The mean squared error (MSE) is calculated by squaring the residuals and summing them. The RMSE is then calculated by taking the square root of the MSE so that it is in the same units as the original data. The value is usually interpreted as either how far (on average) the residuals are from zero or as the average distance between the observed values and the model predictions. Another common metric is the coefficient of determination, commonly written as R2. This value can be interpreted as the proportion of the information in the data that is explained by the model. It should be watched closely, as a high R2 may indicate overfitting. The third metric evaluated is the Mean Absolute Error (MAE). This metric has the same concept of RSME, but with the difference of being an absolute subtraction. This helps to mitigate extreme values, that end up weighing more in the RSME. Both metrics (RSME and MAE have the same unit as the target variable, which in our case is in days. Tables 2 and 3 present the metrics for the both models (with and without phase depths).

Table 2: Performance Metrics for model not using Depths (Model 1)

Model	*RMSE*	*MAE*	*R2*	*Residual*
Elastic Net	1.53	1.30	73%	Presence of Patterns
Decision Tree	1.67	1.38	57%	Residuals not normally distributed
Random Forest	1.53	1.28	73%	Skewed Residuals
Stochastic Gradient Boost	1.51	1.28	75%	Normal Random
Cubist Regression	1.53	1.27	73%	Normal Random
Support Vector Machine	1.51	1.27	74%	Normal Random
Neural Network	1.50	1.26	75%	Normal Random

Table 3: Performance Metrics for model using Depths (Model 2)

Model	*RMSE*	*MAE*	*R2*	*Residual*
Elastic Net	1.74	1.33	79%	Presence of Patterns
Decision Tree	1.75	1.43	54%	Residuals not normally distributed
Random Forest	1.63	1.38	98%	Skewed Residuals
Stochastic Gradient Boost	1.65	1.43	100%	Normal Random
Cubist Regression	1.69	1.36	96%	Normal Random
Support Vector Machine	1.68	1.29	84%	Normal Random

c) Predictions on Unseen Data

As explained on Section 2, the data used for test and training were from Valaris wells that started no later than April 30th, 2020. For this phase, data from wells starting after April 30th were collected and analyzed: a dataset of 65 phases for Model 1 and 45 phases for Model 2. The idea was to use the models selected, make predictions, and compare actuals against predictions for these well phases. This was aimed not only to verify the model performance, as demonstrated in the test phase, but also investigate the financial impact of using the tool as part of the decision-making process in pricing a Rig for a specific job. We leverage the concept of prediction intervals and present, as part of the output, some prediction scenarios that are described below:

- **Base Scenario**: The result of the prediction, or the expected value. This means that if during the drilling process everything happens as planned the duration should be close to this value
- **Contingency:** Recommended "safe" value. It is calculated based on the P75 of the selected prediction interval, meaning there is 25% chances the phase takes more than the value defined by the contingency

- **Blue Sky Scenario:** Calculated based on P20, represents the scenario in which drilling was actually much easier than planned and the phase took considerably less than the base scenario
- **Grey Sky Scenario:** Assumes a safety factor above the contingency, creating a safety net for unplanned events that can cause the phase to take considerably more than expected. It is based on the P90 of the selected phase, meaning there is a 10% chance that the phase duration exceeds this threshold.

Results of these predictions and scenarios can be found in Appendix C. For this exercise the phases were grouped by Country, Client, Rig Class and Rig Type, and show the days balance between the real phase duration and the scenarios described above.

5) Conclusion

Starting with Model 1, the technique selected for the final model was the neural network. Even though it is an expensive method, it had by far the best metrics and residuals. Using this model we are, on average, 1.26 days off the real duration of the well phases. Except for the decision tree, all models do a good job at generalizing and explaining the data variability. There is still room for improvement once R Squared is at 70-ish level for all the other models. Prediction intervals seem reasonable; they will obviously vary from phase signature to phase signature because some cluster have little historical data from which to train; this problem should be solved as more data is added to the model. Using this model on the unseen data, a bidding strategy using the **Grey Sky scenario** would have generated additional value in 84% of the phases, with an average finish **13.6 days ahead of schedule per phase**, while the **Contingency Scenario** would have generated additional value in 77% of the phases, with an average finish **5.4 days ahead of schedule per phase**.

For Model 2, the selected technique was the cubist regression. It presents the best combination between residuals and metrics when compared with the others. The R Squared values for Random Forest and Stochastic Gradient Boosting suggest they were *overfitted*; This means they do an excellent job explaining the data from which they were trained, but do poorly when presented an unseen dataset. This can be explained by the size of the dataset. Due to the lack of good depth data, this dataset has the number of columns very close to the number of measurements, making it harder to generalize. Even though more data would be very welcomed, the cubist regression does a good job and achieved very good results; the long term vision for the project should be using Model 2 as the principal model. Using this model on the unseen data, a bidding strategy using the **Grey Sky scenario** would have generated additional value in 85% of the phases, with an average finish **19.4 days ahead of schedule per phase**, while the **Contingency Scenario** would have generated additional value in 70% of the phases, with an average finish **9.7 days ahead of schedule per phase**.

From the variable importance perspective, it is clear depths help. The most important variable in Model 2 is the casing shoe depth. This implies having accurate and high quality sensor data in the near future will directly aide in producing a better model. When looking at variable importance for both models, it is apparent that location plays an important role; latitude and longitude combined with phase diameter are important in both approaches.

Data quality was really a challenge, but the future is promising. VOS will help solving several problems found when developing this work, and combined with more attention at the Rig level and VIP roll out will pay good dividends. An ideal workflow for production would have historical data clean and available in the data lake, combined with new data from VOS and VIP every new training cycle of the models.

As stated at the beginning of this section, both models are viable but there is still room for improvement. Both approaches considered operational data only; a continuation of this work could include the use of HR data to test the influence of human factor on well duration variability, for example. Some concepts presented in this work can also generate new test cases. A model for an unplanned event probability given an specific well signature, or use of the predicted phase durations to optimize supply chain times seems to be natural paths for expansion. A route involving carbon emissions can also be explored; combining the phase duration prediction with a forecast for the carbon emissions of a Rig would yield the carbon footprint of that well, which could generate a business model where part of our cost would be allocated in carbon credit tokens that can be negotiated in the free market, making Valaris the first drilling contractor drilling “carbon free” wells, and making a good case for ESG in the industry.

6) References

[1] Kuhn, M., & Johnson, K. (2016). *Applied predictive modeling*. New York: Springer.
[2] Sheather, S. J. (2009). *A modern approach to regression with R*. New York: Springer.
[3] Breiman L (2001). “Random Forests.” Machine Learning, 45, 5–32.

Appendix A: Variables dropped due to high correlation

Variables Removed from data

Field.East.Venezuela.Basin	Field.Gulf.of.Guinea	Field.Mesopotamian.Foredeep.Basin
Field.Pelagian.Basin	Field.Qatar.Arch	Field.Rub.Al.Khali.Basin
Client.Premier.Oil	Client.Saudi.Aramco	RigName.VALARIS.110
RigName.VALARIS.120	RigName.VALARIS.5002	RigName.VALARIS.5004
RigName.VALARIS.52	RigClass.6000.SERIES	YearBuilt
Country.Lithuania	Country.Brazil	Country.Bangladesh
Country.India	Country.Burma	Country.Israel
Country.Cameroon	Country.Egypt	Country.Ukraine
Country.Namibia	Country.Yemen	Country.South.Africa
Country.Canada	Country.Tanzania	Country.New.Zealand
Country.Mexico	Country.Gabon	Country.Congo..Kinshasa.
Country.Russia	Country.China	Field.Faeroes.Shetland.Orkney.Basin
Field.Browse.Basin	Country.Libya	Field.Thai.Basin
Field.Greater.Sarawak.Basin	Field.Taranaki.Basin	Field.South.African.Coastal
Field.Kutei.Basin	Field.Yinggehai.Basin	Client.PTEEP
Client.Mubdalla	Client.SOCO	Field.Red.Sea.Basin
Field.Saline.Comalcalco.Basin	Client.C..Hess	Client.Total
Client.Cobalt	RigName.VALARIS.102	RigName.THUNDER.HORSE
RigName.VALARIS.115	Client.Petrobras	RigName.VALARIS.8501
RigType.HEAVY.DUTY.ULTRA.HARSH.ENVIRONMENT	RigType.MANAGED.RIGS	RigName.SAR.202

Appendix B: Applied modeling techniques and distributions

Elastic Net Regression

The elastic net regression is a modeling technique belonging to the class of penalized linear regressions and seen as a generalization of Ridge and Lasso techniques. In short, these models allow for the exchange of bias for lower variance, which helps improve performance of the predictions. This is achieved by using a penalization parameter (lambda) that is selected through an iterative method where the function's target is to minimize the test error. Figure B-1 presents the residual analysis for the fit. Ideally, the scatter plot depicts randomness, and the histogram should follow a normal distribution.

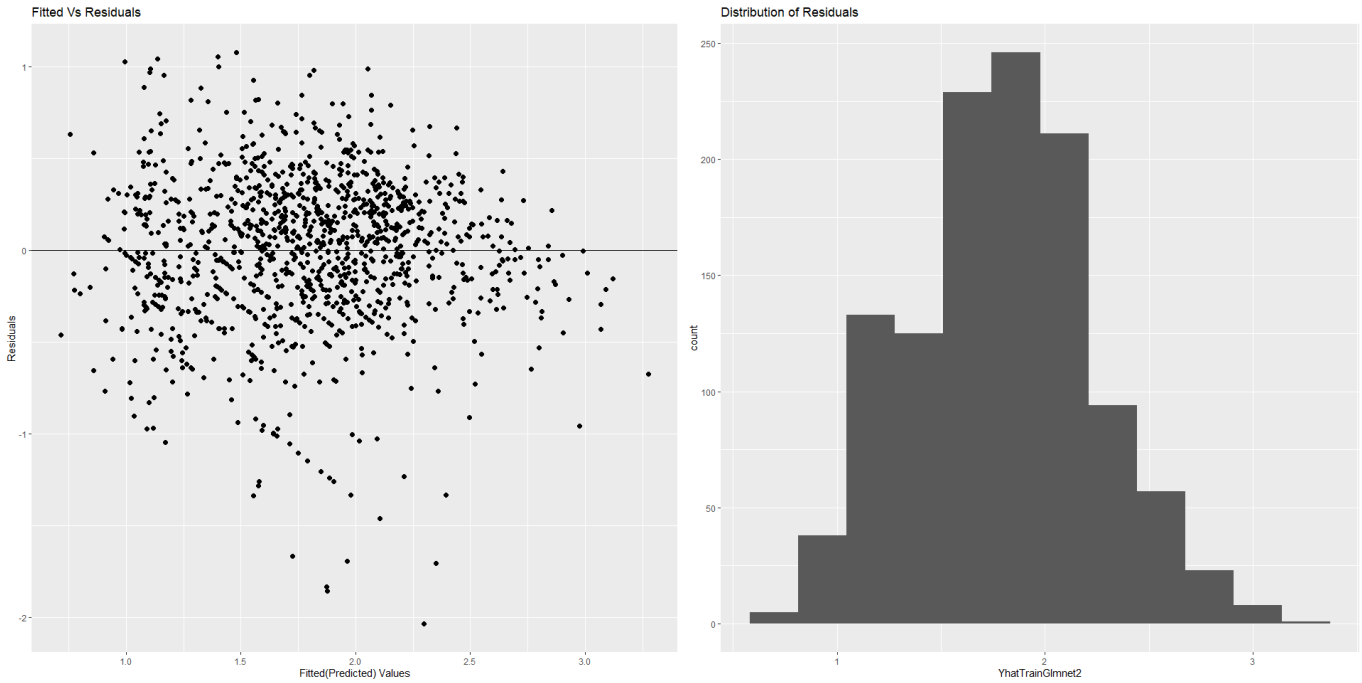

Figure B-1: Residual Analysis for the Elastic Net Model

Decision Tree

Tree-based models are a popular group of modeling techniques based on nested if-then statements for key predictors. They partition the data while producing outcomes that are easily interpretable. They are very robust models from the development perspective as they bring the following advantages:

- Transformations are not required
- It works well with skewed, continuous, and categorical data for example
- Handles missing data
- Performs feature selection as part of the algorithm.

For this work, the data used to model was transformed as described in Sub-section b. Figure B-2 presents the layout of the selected tree, while Figure B-3 presents the residual analysis.

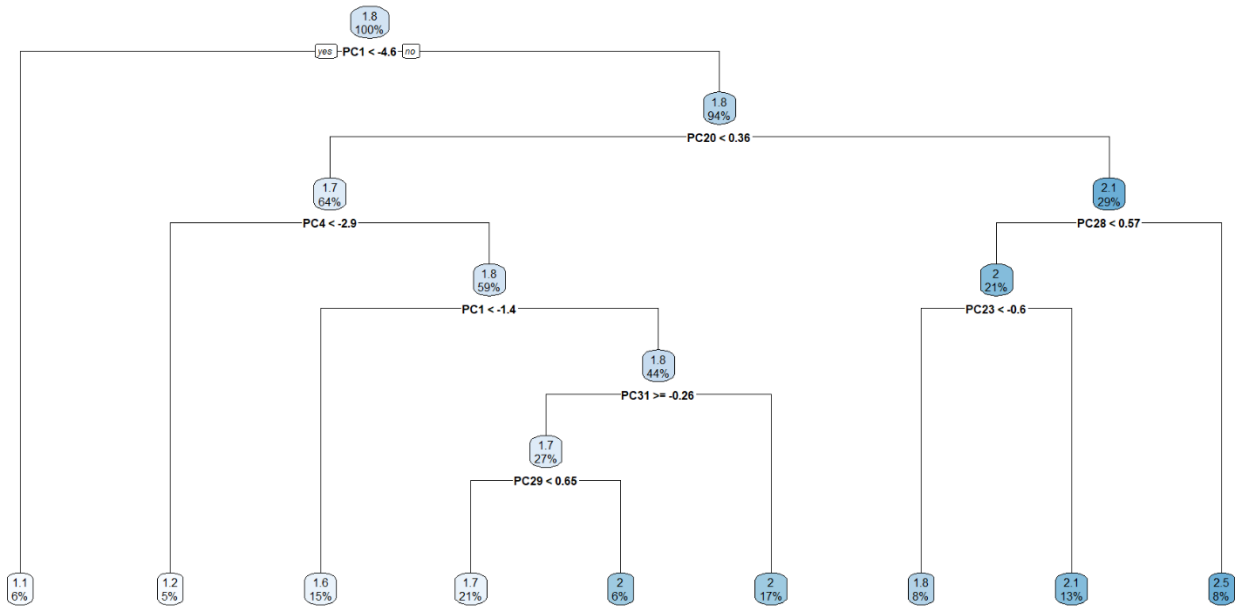

Figure B-2: Decision Tree Model for Well Phase Duration

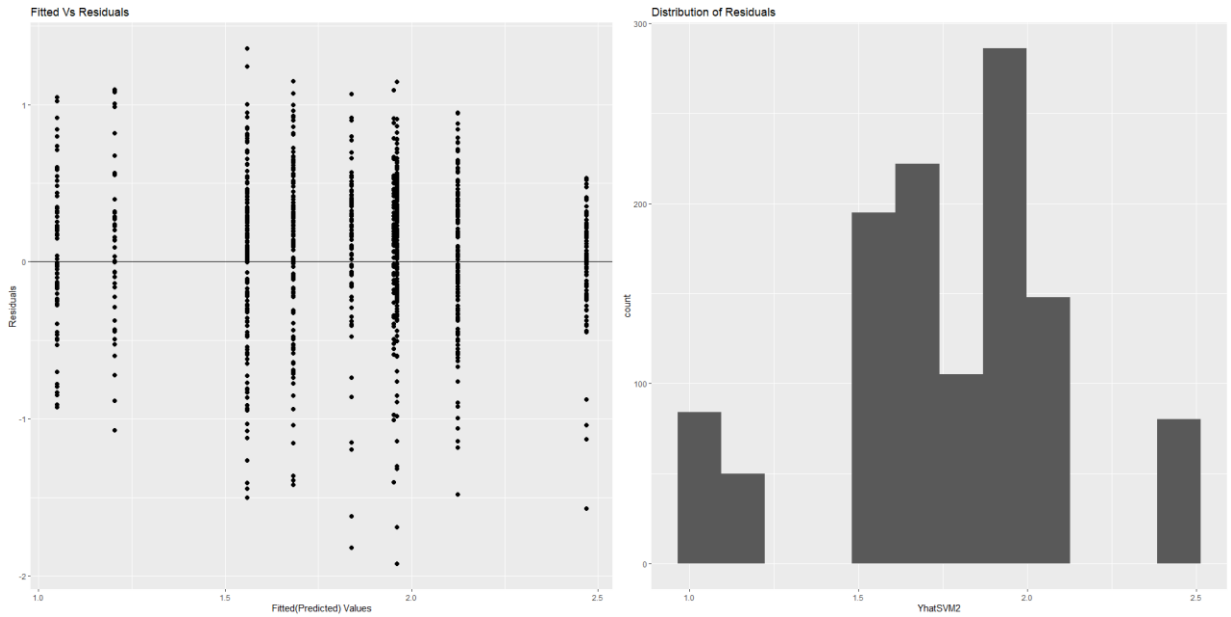

Figure B-3: Residual Analysis for the Decision Tree Model

Random Forest

Random forests can be seen as an evolution of decision trees, with the idea of splitting the data randomly and assigning it to decision trees where the top predictors are evaluated. Breiman (2001) is considered the creator of the unified random forest algorithm, a term that was used for the first time in his work with the same name. General random forest logic can be seen at right. For this work, 500 trees were used, with a minimum node size of 5. Figure B-4 presents the residual analysis for the fit.

Select the number of models to build, m

for i = 1 to m do

Generate a bootstrap sample of the original data

Train a tree model on this sample

for each split do

Randomly select k (< P) of the original predictors

Select the best predictor among the k predictors and partition the data

end

Use typical tree model stopping criteria to determine when a tree is complete (but do not prune)

end

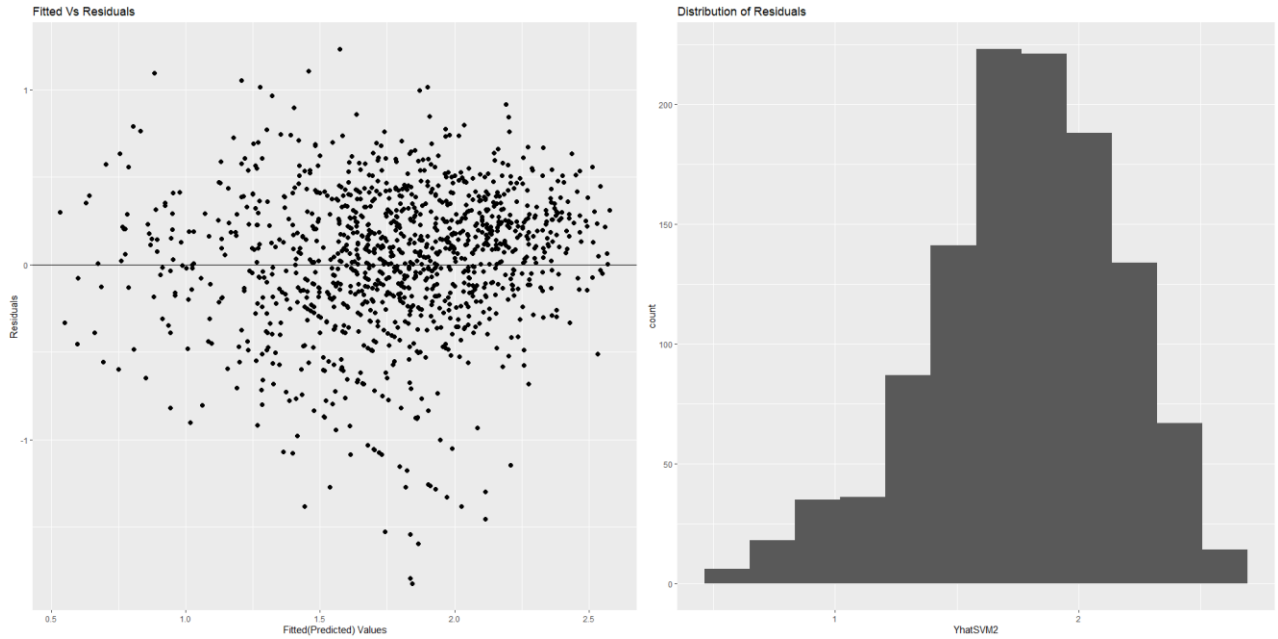

Figure B-4: Residual Analysis for the Random Forest Model

Stochastic Gradient Boost

The development of the gradient boost methodology has its origin in the AdaBoost algorithm. It then evolved to what is known today as stochastic gradient boosting machine, the boosting algorithm of choice among practitioners. It is based in a loss function such as a regression error and a weak learner, so then the algorithm works in minimizing this loss function. It is initialized with the mean response of a regression. Then, based on the residuals a new model is fit to minimize the loss function in an iterative process. A general Gradient Boosting algorithm can be seen at right, and Figure B-5 depicts the results of modeling with this method.

1 Select tree depth, D, and number of iterations, K
2 Compute the average response, $\overline{y}$, and use this as the initial predicted value for each sample
3 **for** $k = 1$ *to* K **do**
4 Compute the residual, the difference between the observed value and the *current* predicted value, for each sample
5 Fit a regression tree of depth, D, using the residuals as the response
6 Predict each sample using the regression tree fit in the previous step
7 Update the predicted value of each sample by adding the previous iteration's predicted value to the predicted value generated in the previous step
8 **end**

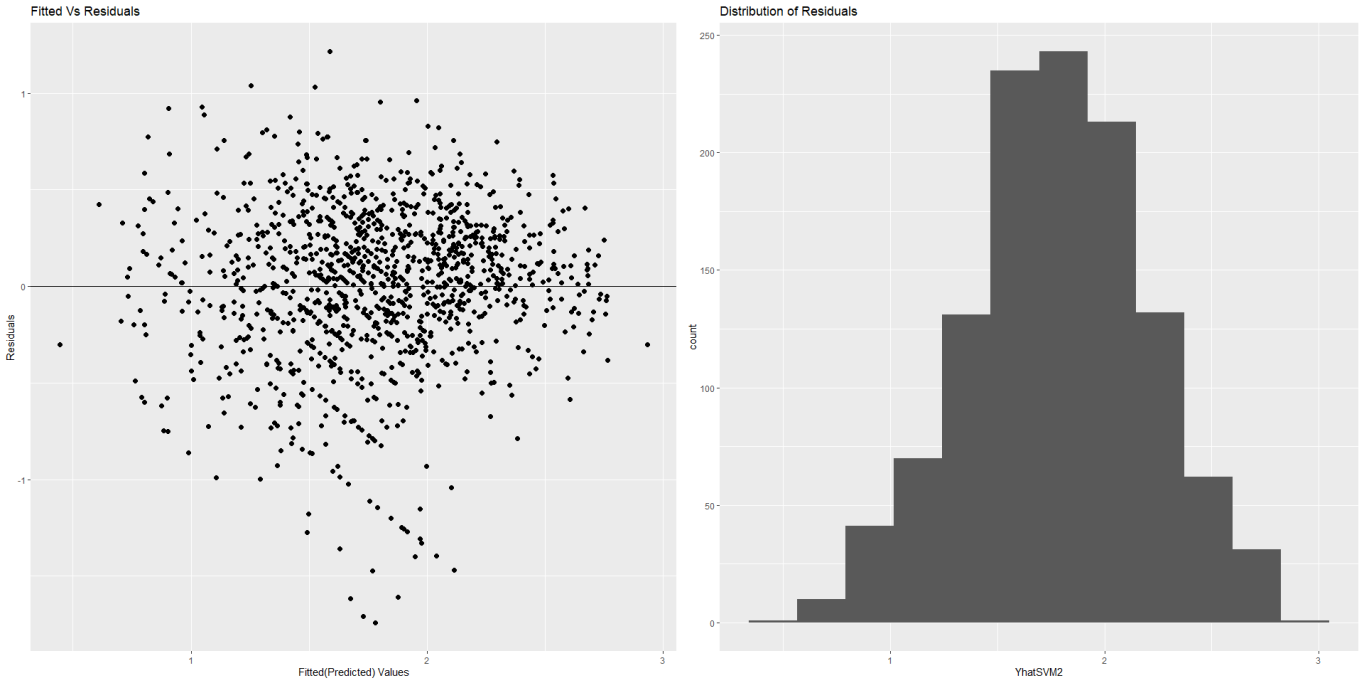

Figure B-5: Residual Analysis for the Gradient Boosting Algorithm

Cubist Regression

Cubist regression is a technique that was only available in the commercial environment. It is a combination of other rule-based techniques that had its source code released under open source license in 2011. The model tree construction is similar to the process used for pruned decision trees, with some specific differences described below:

- Methodology used for smoothing, rule creation, and pruning are different
- There is an optional boosting procedure named “committees”

Figure B-6 presents the residual analysis for the fit.

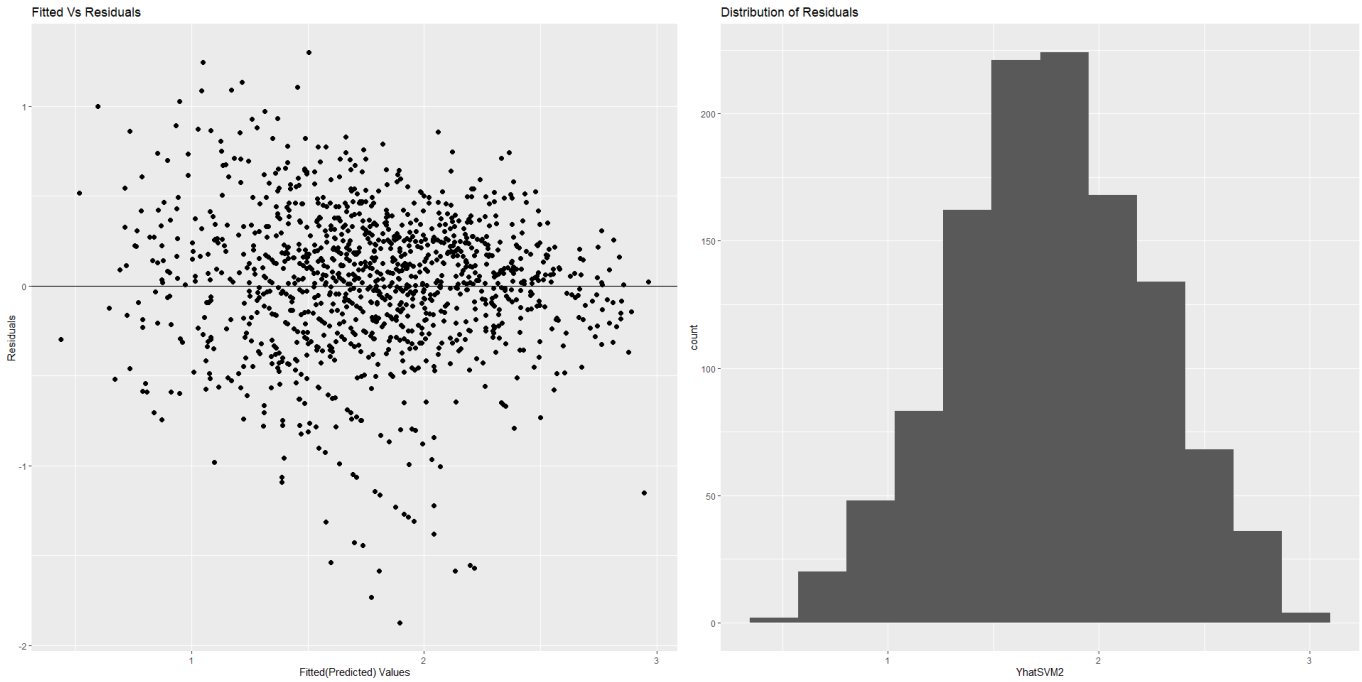

Figure B-6: Residual Analysis for the Cubist Regression Algorithm

Support Vector Machine

Support Vector Machines (SVMs) are a group of robust and powerful modeling techniques, known for their impressive accuracy and low computational cost. Initially developed as a solution for classification problems, it was adapted to become a powerful regression technique.

It constructs a hyper-plane in a dimensional space, which can be used, for example, for outlier detection, classification, regression, or other tasks. A good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Some advantages of this method are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Figure B-7 presents the residual analysis for the fit.

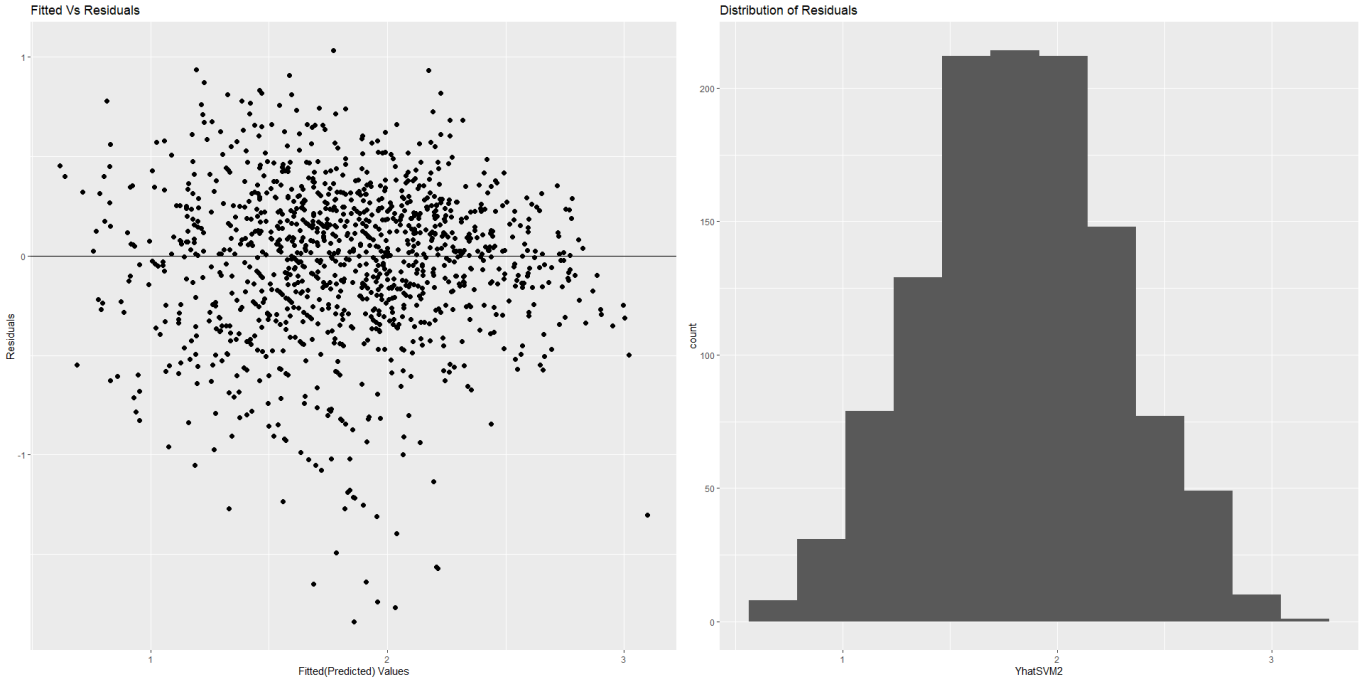

Figure B-7: Residual Analysis for the Support Vector Machine Model

Neural Network

Neural networks are powerful, nonlinear regression techniques inspired by human brain operation. They can be understood as a network of nonlinear functions called neurons that, differently from decision tree models, have no specific logic. They observe the dataset while trying to "learn" the patterns in a way that, once completed, can predict new outcomes with similar patterns. It is a compound of layers of interconnected nodes called perceptron, that feeds a signal produced by a multivariate linear regression to a non linear function. The input layer is where the patterns are received, while the output layers contain the data to be predicted. There is also a hidden layer that fine tunes the results by minimizing the loss function.

Neural networks are a computationally expensive technique that yield very accurate predictions. Figure B-8 presents the residual analysis for the fit.

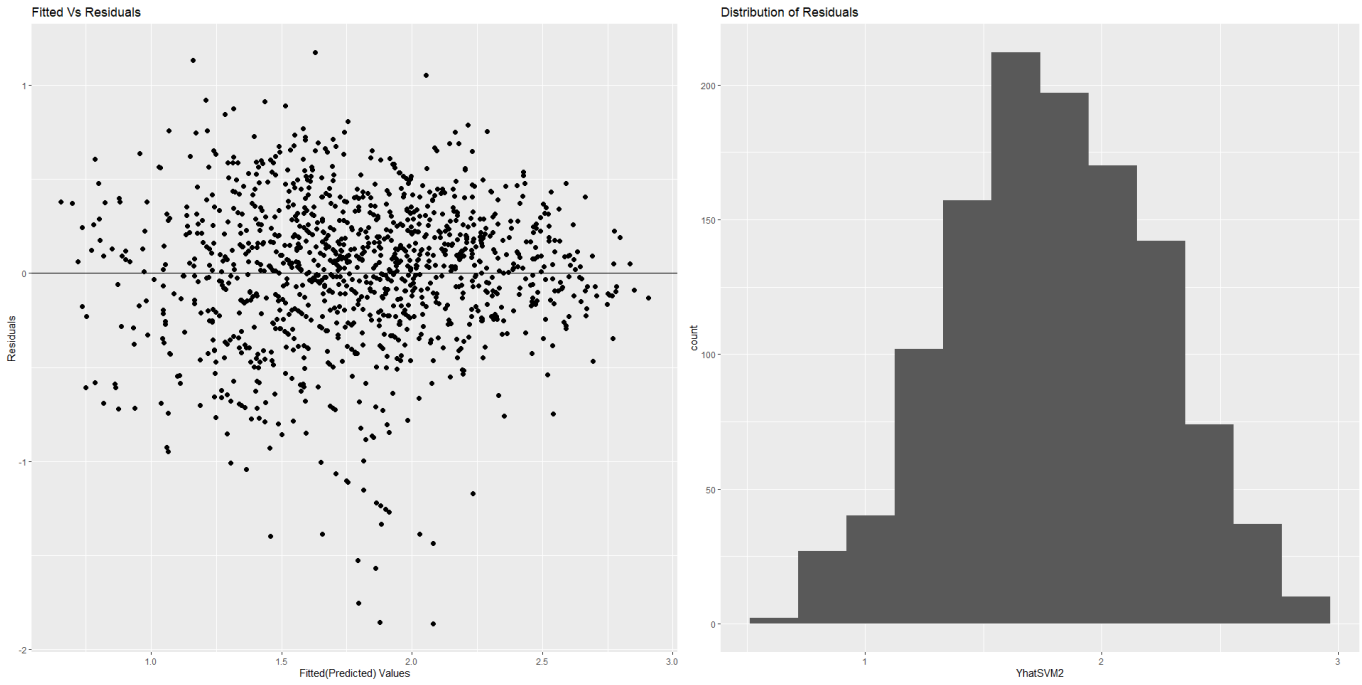

Figure B-8: Residual Analysis for the Neural Network Model

Appendix C: Predictions and Scenarios for Unseen Data

Table C-1: Model not using Depths (Model 1)

Country	Client	RigClass	RigType	Actual Duration	Base Scenario Balance	Contingency Scenario Balance	Grey Sky Scenario Balance	Sample Size
Saudi Arabia	Saudi Aramco	STANDARD DUTY LEGACY JACK-UP	F&G	16.6	13.7	18.5	21.6	3.0
Indonesia	BP	HEAVY DUTY MODERN JACKUP	MOD V	20.1	7.7	15.7	22.4	3.0
Saudi Arabia	Saudi Aramco	STANDARD DUTY LEGACY JACK-UP	MLT	16.5	3.2	8.0	11.1	2.0
Norway	Total	HEAVY DUTY ULTRA-HARSH ENVIRONMENT JACKUP	140 SERIES	26.0	2.4	52.3	79.5	3.0
United States of America	Enven	DRILL SHIP	ULTRA DEEP WATER	8.5	1.2	17.1	28.1	5.0
Thailand	Mubdalla	HEAVY DUTY MODERN JACKUP	400 PACIFIC CLASS	1.9	0.9	4.1	6.6	3.0
Qatar	NOC	HEAVY DUTY MODERN JACKUP	MOD V	14.8	-3.1	9.9	20.4	6.0
Saudi Arabia	Saudi Aramco	HEAVY DUTY MODERN JACKUP	MOD V	25.7	-3.2	0.5	2.9	3.0
Saudi Arabia	Saudi Aramco	STANDARD DUTY JACK-UP	TARZAN	25.1	-4.9	8.2	17.2	7.0
United States of America	Chevron	DRILL SHIP	ULTRA DEEP WATER	25.7	-7.3	-1.4	5.0	4.0
Saudi Arabia	Saudi Aramco	STANDARD DUTY MODERN JACK-UP	MLT	20.8	-9.5	8.7	20.0	5.0
Venezuela	EOG	HEAVY DUTY MODERN JACKUP	140 SERIES	25.5	-9.7	0.6	8.4	5.0
United Kingdom	Premier Oil	HEAVY DUTY ULTRA-HARSH ENVIRONMENT JACKUP	140 SERIES	22.6	-11.1	-6.6	-2.6	5.0
Saudi Arabia	Saudi Aramco	STANDARD DUTY LEGACY JACK-UP	140 SERIES	35.0	-13.7	-7.3	-1.5	2.0
Mexico	Fieldwood	STANDARD DUTY MODERN JACK-UP	140 SERIES	38.7	-17.5	-9.9	-5.1	5.0
Mexico	Fieldwood	HEAVY DUTY MODERN JACKUP	140 SERIES	55.9	-28.3	-17.8	-10.8	3.0
Saudi Arabia	Saudi Aramco	HEAVY DUTY MODERN JACKUP	140 SERIES	117.5	-62.2	-30.6	-10.9	1.0

Table C-2: Model using Depths (Model 2)

Country	Client	RigClass	RigType	Actual Duration	Base Scenario Balance	Contingency Scenario Balance	Grey Sky Scenario Balance	Sample Size
Norway	Total	HEAVY DUTY ULTRA-HARSH ENVIRONMENT JACKUP	140 SERIES	23	11.3	35.9	56.4	2
Saudi Arabia	Saudi Aramco	STANDARD DUTY LEGACY JACK-UP	F&G	24.6	7.7	30.5	46.8	3
Saudi Arabia	Saudi Aramco	STANDARD DUTY MODERN JACK-UP	MLT	11.3	5.2	15.9	26	3
Indonesia	BP	HEAVY DUTY MODERN JACKUP	MOD V	4	3.8	6.1	8.7	1
Venezuela	EOG	HEAVY DUTY MODERN JACKUP	140 SERIES	39.7	1.4	9.3	15.9	4
Saudi Arabia	Saudi Aramco	HEAVY DUTY MODERN JACKUP	MOD V	18	-0.1	6.6	11.2	2
Mexico	Fieldwood	HEAVY DUTY MODERN JACKUP	140 SERIES	22	-0.4	7.8	14.1	3
Qatar	NOC	HEAVY DUTY MODERN JACKUP	MOD V	30.6	-0.9	9.3	18.3	6
United States of America	Chevron	DRILL SHIP	ULTRA DEEP WATER	19.9	-4.3	7	14.9	4
Saudi Arabia	Saudi Aramco	STANDARD DUTY JACK-UP	TARZAN	44.5	-7.2	3.2	13.1	5
United Kingdom	Premier Oil	HEAVY DUTY ULTRA-HARSH ENVIRONMENT JACKUP	140 SERIES	24.9	-8.4	1.9	11.3	3
Mexico	Fieldwood	STANDARD DUTY MODERN JACK-UP	140 SERIES	49.5	-14.3	-0.5	12.4	3
Saudi Arabia	Saudi Aramco	STANDARD DUTY LEGACY JACK-UP	MLT	27	-17.1	-5.5	3.8	1