2022 Chevron Rice Datathon:
Predicting Rate of Penetration (ROP) for new well plans

Description

The process of drilling new wells, especially offshore, is extremely challenging and costly. After reaching the seabed more than 3,000 feet under water, rigs in the Gulf of Mexico must drill through an additional 20,000 feet of rock. In these extreme environments, where temperatures and pressures far exceed regular drilling conditions, specialized equipment and teams are required.

These operations can involve hundreds of people and equipment with very high daily drilling costs. Reducing the time it takes to drill by even a few hours per well can result in significant savings for the company and provide a significant competitive advantage as more and more wells are drilled.

Your task is to use historic drilling data to build a model that can predict the rate of penetration conditional on controllable drilling parameters and/or parameters that are known before the drilling process begins. This model could then be used as part of the "Drilling Roadmap" development process, where the WOB and RPM (Among other variables) are determined before drilling begins.

Data

DISCLAIMER: The data provided for this challenge can be used only to develop a model for the challenge and should be discarded after the end of the challenge.

The challenge involves several files for your project. The files required to build and test your predictive model will be made available at the start of the Datathon. However, 30 minutes before the judging period starts, we will also release a scoring file that contains additional data for which you will run your final model and provide corresponding predictions.

The following files will be available on the Datathon website at the start of the Datathon :

- **training.csv**: contains raw data used to build, train and test your model. The dataset is listed by well segment and contains information regarding drilling parameters such as weight on bit (WOB) and RPM, drill bit information and a classification of the formation.
- **data_dictionary.xlsx**: provides description of the variables in training.csv
- **scoring_format.csv**: shows the format of the scoring file (scoring.csv, released 30 minutes before the Datathon ends). This is a sample of rows from training.csv, but the real scoring.csv file will comprise completely new rows of data.
- **submission_format.csv**: gives the required format of the model predictions file obtained by applying your model to the scoring.csv file. Note: the segment_id column in this file corresponds to the segment_id in the scoring.csv file below and ***the rows must be submitted in this same order***!

The following file will be available on the Datathon website 30 minutes before the Datathon judging begins:

- **scoring.csv**: used to generate model predictions required for us to score your model accuracy

Project Submission

Please submit your Datathon project to devpost for judging. At minimum, the submission should include a link to a github repo that contains:

- your project code
- a summary of your methodology and findings
- submission_format.csv: the file that is the output of running your model on the scoring.csv dataset

Again, due to the specificity of the submission file format please follow the formatting provided in submission_format.csv. All submission files are required to have exactly the same column names as this key, and ***the rows must be in the same order as they are in the submission key***! Thus, we highly recommend you use a copy of this file to submit in order to ensure your submission is valid and corresponds to the correct order for the wells. The file must be a comma-separated csv file!

Evaluation Criteria

Projects will be evaluated holistically based on model accuracy the criteria outlined by Rice Datathon:

- **Technical Difficulty**: We are looking for technically advanced solutions to difficult problems that make use of a diverse set of modeling and data science techniques. That being said, if you can solve a challenging problem with a simple solution, we will be very impressed!
- **Analysis & Exploration**: We are looking for projects that take time to analyze and explore the nuances of whatever data they are working with.
- **Creativity**: We are looking for original ideas or new angles on existing ideas.”
- **Predictive Accuracy**: predictive accuracy will be measured by the Root Mean Squared Error (RMSE) in ROP predictions for the well segments in the scoring.csv file:

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2}}$$

  where $\boldsymbol{y_i}$ is the actual ROP corresponding to the $\boldsymbol{i}$-th row in scoring.csv, and $\boldsymbol{\hat{y}_i}$ is the corresponding model prediction.

Areas that can be looked at include, but not limited to feature generation, feature selection, and model selection/building.

Rules

1. Contestants must respect the privacy of the data and remove it from their computers upon completion of the competition.
2. Contestants' solution must be a model that is repeatable, adjusting model results manually (including “arbitrary” factors/constants) to tune model predictions is not permitted.
3. Contestants' prediction models can use only that data from the training.csv dataset provided to build the model.