Chevron Data Science Challenge: Predicting Production

Competition Details (PLEASE READ EVERYTHING CAREFULLY)

Description

Central to the exploration process are decisions regarding where to drill, and how to drill. Answering these questions requires an understanding of the impacts of reservoir geology and drilling & completions designs on production. In tight oil plays, such understanding is limited by the complexity of the physical processes underlying the creation of the fractured network, and the flow of fluids in that network.

While the direct application of first physical principles to unravel the effects of geology and engineering on production remains challenging, data is available to approach the inverse problem. Given drilling & completions designs and production data for a large number of wells in a specific region, can we develop a data-driven model for production as a function of geology and engineering design? That is the goal of the Chevron Data Science Challenge.

From the Wolfcamp play in the Midland Basin, public data consisting of drilling & completions design, geologic formations penetrated, and production results has been assembled for more than 2700 vertical wells. A 'training' dataset, consisting of data for a portion of these wells, will be available for contestants to construct predictive models. A second 'test' dataset, not including production values, will be distributed for model evaluation.

Data

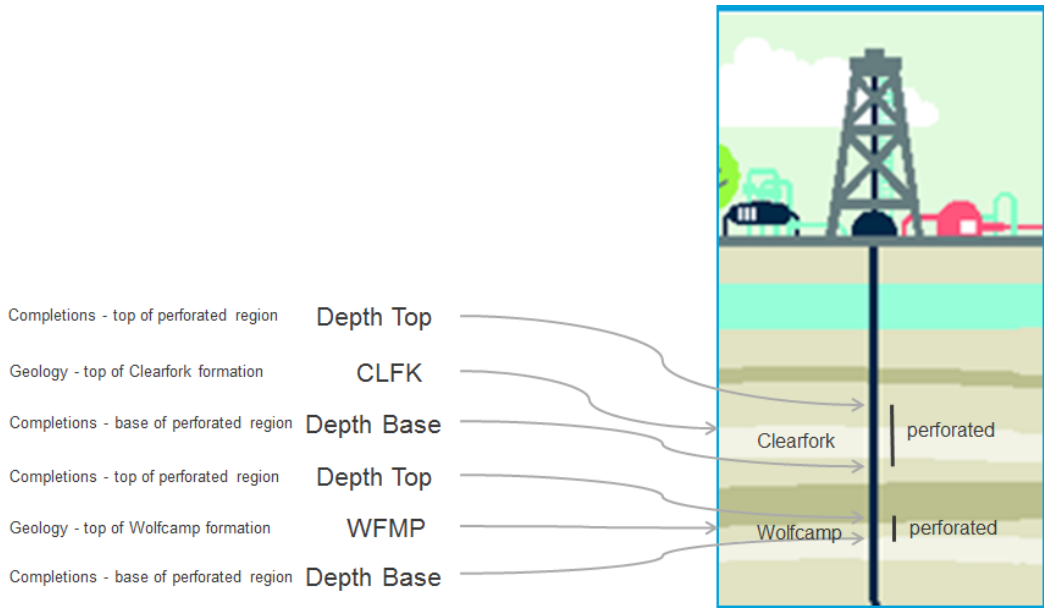
The challenge data consists of six files:

base_training.csv, geology_training.csv, completions_training.csv,

base_test.csv, geology_test.csv, and completions_test.csv

The **base** spreadsheets contain all legal, geographic, geology, drilling, and completions data except detailed formation and per stage completions data. In addition, the base_training.csv spreadsheet contains the estimated ultimate recovery for oil (EUR o) for each well. It is possible to develop models predictive of EUR only from the base_training.csv. To demonstrate the process, the benchmark described below employs only the base data. However we would encourage you to consider approaches that make use of the additional information available in the geology and completions files.

The **geology** files contain hand-picked and interpolated formation tops at each wellbore for each of the formations (potentially) penetrated. Please note that the hand-picked formation tops, although somewhat sparse, are likely more reliable than the interpolated values, and should be used when available.



For each well, the **completions** files contain data on treatments for specific perforated regions. As each well may penetrate several producing formations, each of these formations will likely be perforated and fractured. While not complete, much stage-specific data is available in the completions file. Additionally, perforated zones may be mapped to geologic formation using the formation top data in the geology files.

The **data dictionary** defines each of the features (columns) in the base, geology, and completions files.

Evaluation

Models will be scored based on their predictive accuracy as measured by the Root Mean Squared Error (RMSE) in EUR oil predictions for the test set wells. We want to ensure that the model you used to generate your submissions is not only *accurate*, but also *general*.

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

$$\text{Where } n = \text{number of Wells,}$$

$$y = \text{predicted EUR}$$

$$\hat{y} = \text{actual EUR}$$

Submission File Format:

You may only submit files in **comma, semicolon, or tab seperated file formats** (e.g. NO excel .xls or .xlsx formats). If you are working from excel, you can go to File --> save as --> choose legal file type mentioned above.

Solution file must have only 2 columns: a "**WellID**" column and a "**EUR_o..Mstb.**" column (see **benchmarkLM.csv** for an example).

Benchmark

To elucidate the modeling process, we have developed a simple model using only the base training data. To predict EUR (o), a multivariate linear regression model was designed in the statistical programming language R (benchmarkLM.csv). The code contained in the file **benchmarkLMsourcecode.R** demonstrates this process. *Of course, you are free to use any other alternative analytical tools.* Another benchmark, benchmarkAverageTrainsetEUR.csv, predicts all wells in the test set to have EUR equal to the average EUR from the training set. Not surprisingly, this benchmark did not score as well as the linear model benchmark. Hopefully, your models will score much better than either of these benchmarks!