geoconvention
Calgary • Canada • May 7-11 2018

Data science tools for petroleum exploration and production

Matteo Niccoli[1] *Thomas Speidel*
1 MyCarta

Summary

Our aim is to demonstrate practical aspects of data science by illustrating the analysis of publicly available oil and gas production data. Two will be the areas of focus: exploratory data analysis (EDA) and predictive modeling. In both cases, we will make abundant use of sound visualization techniques while following the principles of reproducible research. Every aspect of the analysis will be available as a series of notebooks shared on a GitHub repository.

Methods

Key aspects we plan to cover are:

Exploratory Data Analysis

- Distribution visualization (such as boxplots, kernel density, histograms, novel plots)
- Summary Statistics (such as measures of central tendency, spread)
- Univariate screening (such as t-test/Wilcoxon, critical r, chi-squared tests)
- Correlations measures and correlation matrix plots (such as Spearman correlations)
- Redundancy analysis (as a data reduction technique)
- Variable Clustering (as a data reduction technique)

Predictive & Inferential Analysis

- Linear models
- Non-linear models
- Machine Learning models
- Feature selection
- Model interpretation

Computational Tools

We will demonstrate each technique in detail using either Python or R. Each technique will utilize one of the two popular languages. Where possible, the authors will strive to have both languages.

Examples

Figure 1 showcases a Python tool used to facilitate the culling (screening) of independent variables based on a critical r test. The critical r is the value of correlation coefficient at which one can rule out chance as an explanation for the relationship. The tool first generates a correlation matrix, and after comparing the correlation coefficient between the dependent variable and all the independent variables one at a time, it highlights the ones that do not pass the critical r test. In this case (data from Hunt, 2013) the variables position, pressure, X5, and X6 fail the critical r test. How is this knowledge useful? On the one hand, since there is no reasonable physical explanation for the relationship between X5, X6 and production, these variables should be rejected as predictors of production. On the other hand, as pointed

out by Hunt (2014), prior knowledge of production suggests it is often a multivariate problem, and that pressure and position should matter in predicting it, in spite of the test failure.

Figure 2 is an example of response surface plot (data from Doublet, 2001) for the permeability ratio (Kmax/k90) based on a non-linear ordinal regression model (Harrell, 2015) that adjusts for helium porosity, photoelectric capture cross-section log, neutron porosity, depth and rock type. The plot is generated by predicting the median permeability ratio for each rock type while holding neutron porosity and depth fixed at the medians (0.085 and 7046, respectively). Each panel represents a different rock type. The color scale on the right represents the predicted median permeability ratio. For this model, one can visually assess the relationship both helium porosity and photoelectric capture cross-section log have on permeability for different rock types. For instance, overall, rock type 4 has the highest permeability (green shades), while rock type 3 one of the lowest. The effect of photoelectric capture cross-section log is over and above that of rock type.

References

Doublet, L.E. (2001) An Integrated Geologic and Engineering Characterization of the North Robertson (Clear Fork) Unit, Gaines County, Texas. Petroleum Engineering PhD thesis, Texas A&M University.

Harrell, F.E. (2015) Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. Second Edition. Springer-Verlag New York, Inc. New York, USA

Hunt, L. (2013) Many correlation coefficients, null hypotheses, and high value. CSEG Recorder, 38 (10)

Kalkomey, C. (1997) Potential risks when using seismic attributes as predictors of reservoir properties. The Leading Edge 16 (3)

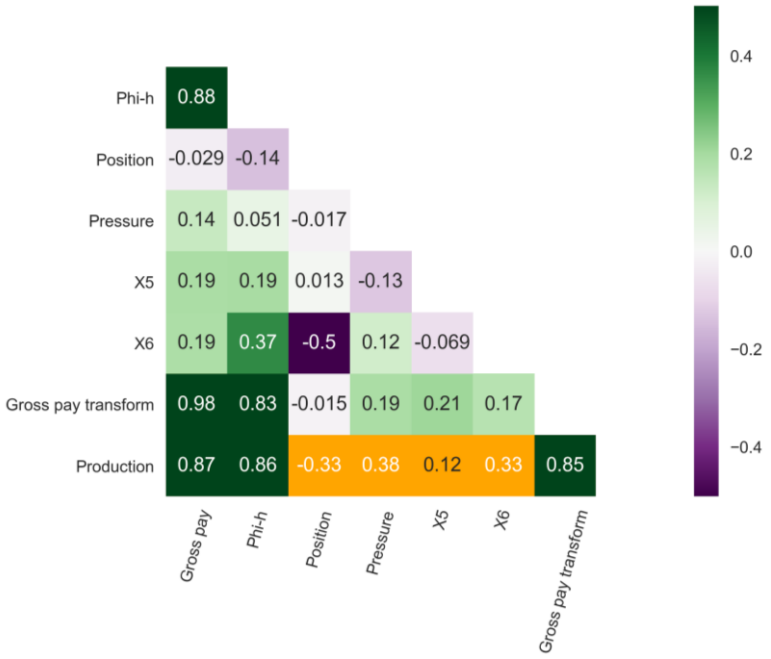

Figure 1. Output plot of the Python tool used for screening independent variables gross pay, Phi-h, ranked position (facies) within the reservoir, pressure draw-down, and a gross pay-transform. The dependent variable is production. The critical r test tells us that with 21 wells (19 degrees of freedom) and a confidence level of 95% we need a value of correlation coefficient of at least 0.43 to rule out chance as an explanation for a specific correlation.

Figure 2: Surface plot based on a non-linear regression model showing the simultaneous effect, on the dependent variable permeability ratio (Kmax/k90), of changing two continuous variables (helium porosity and photoelectric capture cross-section log) while holding (marginalizing) all

other variables fixed at their median. The non-linear model utilizes restricted cubic splines with 4 knots for helium porosity, 3 knots for photoelectric capture cross-section log, 3 knots for neutron porosity and 3 knots for depth.