DATA ANALYTICS AND MACHINE LEARNING WORKFLOWS FOR OPTIMIZATION OF UNCONVENTIONAL ASSETS. CASE STUDY: NEUQUÉN BASIN, VACA MUERTA PLAY

Jesús Ochoa[1], Lean Elizabeth Gardland[2], Knut Utne Hollund[3], José Julián Salazar[4,5,6], Michael Pyrcz[4,5], Haoyuan Zhang[7]

1: Technology, Digital and Innovation, Equinor US, Houston, USA, jocho@equinor.com

2: Exploration and Production International, Equinor ASA, Oslo, Norway, lega@equinor.com

3: Technology Digital and Innovation, Equinor ASA, Oslo, Norway, kuho@equinor.com

4: Hildebrand Department of Petroleum and Geosystems Engineering, Cockrell School of Engineering, The University of Texas at Austin, USA, jsalazarn@austin.utexas.edu

5: Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, USA, mpyrcz@austin.utexas.edu

6: Facultad de Ingeniería en Ciencias de la Tierra, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5 Via Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador.

7: Technology Digital and Innovation, Equinor ASA, Oslo, Norway, hzan@equinor.com

Keywords: multivariate analytics, machine learning, integration, petroleum system, produced volumes

RESUMEN

El objetivo de este trabajo es ilustrar cómo los métodos avanzados de análisis de múltiples variables y aprendizaje automático son útiles para la mejora económica de la productividad de pozos en el intervalo Cocina de la Formación Vaca Muerta, Cuenca Neuquina. Las metodologías basadas en ciencia de datos brindan la oportunidad de aumentar la eficiencia y extraer más detalles de extensas bases de datos.

Todas las decisiones de campo y optimización se toman en presencia de incertidumbre. Debido a las incertidumbres inherentes al campo físico y la escasez de puntos de control, proponemos métodos basados en datos para evaluar el impacto de diferentes parámetros en la producción. El flujo de trabajo se compone de análisis de datos junto con métodos de aprendizaje automático supervisados y no supervisados.

Primero: Se implementa una fase de análisis espacial, mejorando el rigor estadístico de los parámetros de entrada para la predicción usando aprendizaje automático al tiempo que se respeta el contexto espacial de los datos del subsuelo. El flujo de trabajo incluye identificación de anomalías espaciales, detección, valores atípicos de datos espaciales, problemas de calidad de datos y modelado de tendencias óptimas.

Siguiente: Un análisis de múltiples variables en la zona de perforación de interés. El modelo usa datos solo de este intervalo con el objetivo de determinar las características de alto impacto y en qué rangos de valor estas características gobiernan la productividad del pozo. Cuantificar y clasificar la importancia de todas las características geológicas, geofísicas y de ingeniería disponibles permite

realizar pruebas iterativas rápidas de diferentes diseños de producción. Los modelos basados en datos se calibran con modelos basados en la física (Cruz, ***et al*** 2021 y presentación de colegas de Equinor en CONEXPLO22 por Arief ***et al***).
Finalmente: Una predicción usando aprendizaje automático (ML) para pozos futuros vinculada a un modelo económico que proporciona parámetros comerciales como: tiempo de recobro de inversión para cada pozo y cobertura de gastos en un corto plazo. El modelo de pronóstico de aprendizaje automático utiliza todo el conjunto de datos de producción de la Formación Vaca Muerta del Capítulo IV como análogo a la predicción de producción. El objetivo de este paso final es modelar la optimización económica produciendo un pronóstico para cada pozo para un plan de desarrollo, brindando la capacidad de evaluar múltiples escenarios y cientos de iteraciones. Consulte la Figura 1: Flujo de trabajo completo basado en datos.

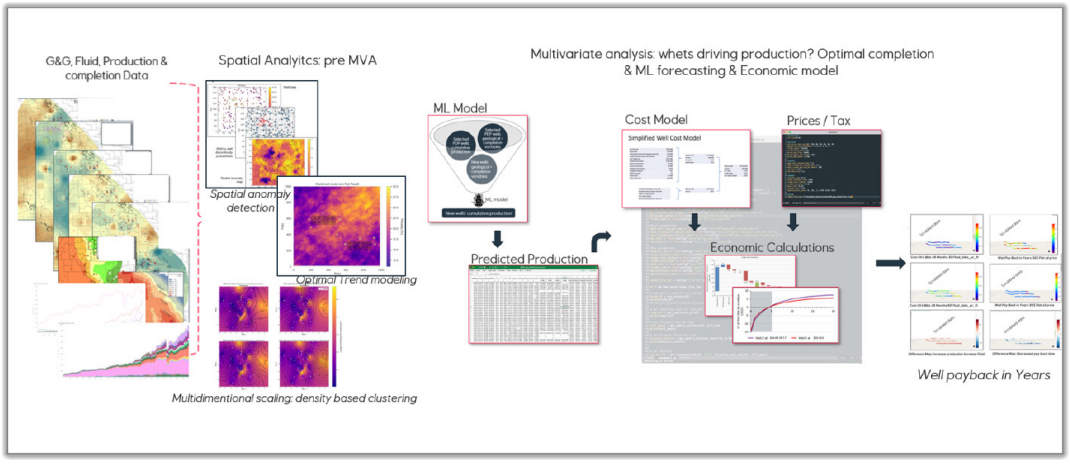


Figure 1. Data analytics workflows for optimization of unconventional asset. Novel spatial analytics techniques, handling the G&G input data -Clustering Techniques-Multivariate analysis: What's driving production? Testing completion scenarios: Machine learning multi scenario forecasting and well payback model.

INTRODUCTION

We propose a data driven, end-to-end workflow that integrates all the data. Geological and geophysical data, with spatial analytics applied to honor the spatial context of the data. This is combined with engineering, production, financial and cost data as a means of extending the current modeling capabilities and enhancing the current decision workflow practices.

The workflow facilitates understanding influential drivers on well productivity, understood by using Multivariate analysis and the impact of geospatial variations, what happens when we move away from a known well bore? This is understood using semi-variograms in 2D modeling (Optimal trend model) and semi-supervised Machine learning clustering techniques to generate facies maps, for a consistent reservoir quality comparison across the basin of all G&G properties combined to define facies classes. (Figure 5).

The aim of this is gaining insight into well completion optimization. and knowledge for well spacing considerations- We can also gain knowledge on what impacts payback time (time

to 100% return on investment). Multivariate analysis gives insight, to the impact of engineering and geological variables over time, e.g., what was significant in the first years and what drives production in the later years of field development (Figure 10), assisting with optimization of shortest well payback time and development sequences.

Our focus is data-driven characterization of the subsurface. We build practical automated workflows for subsurface data analysis and predictive machine learning.

Data science capabilities allow our physics-based models to be augmented, with spatial analytics, non-supervised, supervised machine learning techniques, and multivariate analysis.

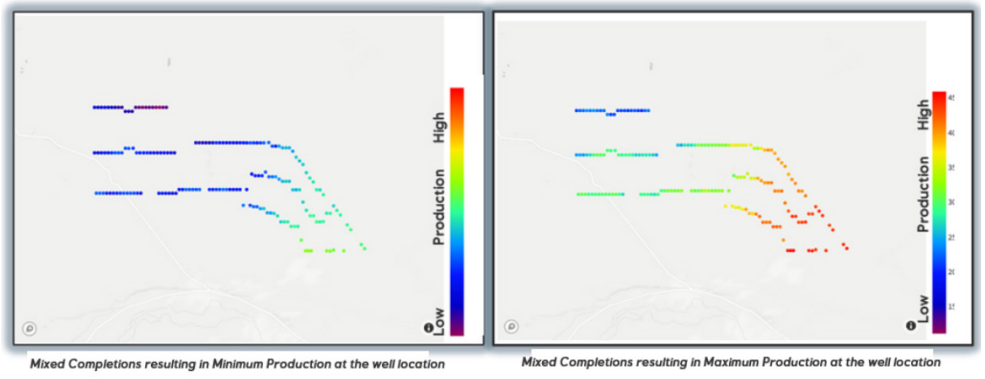
Unconventional dynamic systems pose the challenge of nonlinearity of properties and high dimensionality within the data sets, the many properties also have noise and are stochastic of nature. Machine learning is a good technique for overcoming these challenges as it supports prediction and inference in multivariate and complex data sets. Our data driven models are calibrated to physics-based models. In the paper, Cruz, et. al., 2021 physics-based, and data-driven models were seen as complementary to each other's short comings. The final physics-based models are better equipped, compared to data-driven models, to predict behaviors in situations where large amounts of spatially distributed data are not available. Physics-based models can be used to answer more detailed and granular completions questions i.e., what is the optimal clusters spacing and number of perforations per cluster? (Arief et. al., 2022) However, physics-based models are generally time and computationally intensive and, depending on the degree of reservoir heterogeneity in the studied basin/reservoir, may only be applicable to a small region near the studied well or pad. Spatially aware data-driven models allow for insight away from the small region of detailed data.

Combining the physics-based and data-driven modeling provide an opportunity to bridge the data and physical understanding gaps to improve accuracy of production forecasts and completions optimization. Digital driven automation of forecasting models tackles the issue of bias in forecasting to avoid misallocation of capital. Data-driven forecasting generates prediction updates at speed.

The vehicle for automated workflow is a scripting / coding solution. When you automate workflows there are many gains:

1. All data types can be analyzed and integrated, all geological and geophysical properties (geomechanical, geochemical) not just STOOIP inputs. Completion parameters, cost data, and economic data. If all this information is taken out of context, we could mislead decision making / field optimization.
2. Size of data set is not an obstacle; 100-1,000 wells can be analyzed. Scaling information input means we can learn from all (currently approximately 1,200) horizontal producers in the Vaca Muerta Formation. Legacy software does not allow for this level of data integration.

3. Automation leads to efficiency: multiple scenarios can be run fast & repeatable, varying completion parameters, costs & oil price, outputting forecast iterations fast and testing for optimization. An obvious gain is a higher resolution of information, a forecast for every well in the development plan versus one type-curve per area. Geoscientists and engineers can focus simulation time on scenarios that add value.
4. Multiple scenarios facilitate an uncertainty centric workflow, representing a span of possible outcomes rather than a few single scenarios. Figure 2 demonstrate the degree of flexibility, in evaluating optimal completions scenarios. For the license area in figure 2, thirteen different completion scenarios where tested, and a predicted forecast generated for every well in the development plan, these scenarios can be compared, figure 2 also shows the resulting 12 months cumulative production if the worst performing completion for each well is chosen (image on the left) and the resulting 12 months cumulative production if the best performing completion is selected for each well from the 13-scenario combination (Figure on the Right).



13 completion scenarios run: images show the min production from mixed completions and the max production at each well location from all completions. Actually production volumes scrubbed

Figure 2.Illustrates the flexibility of the automated workflow13 completion scenarios where run on the reservoir interval of the study area, resulting in a production forecast for every well in the development plan , the image shows the 12 months cum production result from a mix of the 13 completion in the lowest/ minimum 12 month cum production attained, image on the left, and the completion scenarios for each well that resulted in the maximum 12 month cum production prediction image on the right.

The workflow is broken into three major phases

1. Spatial analytics:

This step improves that statistical rigor of predictive ML models:

- 1A: Optimal Trend modeling workflow:
  Removes the disconnect of geological and geophysical (G&G) parameters variability from engineering to better honor the spatial context, autocorrelation and variance of G&G

properties. 2D models of the geological properties are generated using GeoModeling principles such as semi-variograms and sequential gaussian simulation.

- 1B: KNN (k-Nearest Neighbors algorithm) a clustering analysis for the interval of interest across the whole basin. Consistent reservoir quality comparison of all G&G properties combined to define facies classes.
- 1C: Fair test train split of spatial data in ML prediction models:
  Model input data is spatially related, the workflow offers spatial aware data sets ready for predictive machine learning problems (Salazar, *et al* 2022).

2. Multi Variate Analysis: (MVA):

Understand bivariate and advanced multivariate correlations and collinearity. Ranking and properties significance of G&G and completion properties on well productivity at time intervals: e.g., 1 year & 3 years. The results of the MVA models are then calibrated to physics-based models (Cruz *et al*, 2021; Arief *et al*., 2022).

3. Auto ML forecasting model & Value Model:

Automated planned well production forecasting for every well in the development plan, with uncertainty, combining cost and economic models for single well payback & short-term measures of value outputs: breakeven, IRR & NPV. The automatic workflow allows for testing multiple completion scenario for well optimization all the way to economics. Completion scenarios are chosen using the results of the MVA analysis and subject matter expert input, this allows to output 9-year production profile; 3 to 5 years predictions on cumulative production or 5 to 8 years decline curve analysis.

DATA INPUT

Equinor's inhouse regional database was used, petrophysical logs and basin analysis models.

Production and completion data are from the Chapter IV public database.

The Production and completion data was downloaded from the Argentinian Energy Department (AED) (Repositorio oficial del Equipo de datos del Ministerio de Energia y Mineria de la Republica Argentina). The github address is the following: https://github.com/datosminem/produccion-de-petroleo-y-gas-por-pozo.

Production wells from 2009 to present day with greater than 500m lateral length where used, approximately 730 wells. Equinor Petrophysical database included approximately 130 vertical pilot wells across the basin.

One target zone was analyzed in the MVA and Machine learning forecast and value model.

If proprietary data was used the model's accuracy would improve, exact target zones of the production wells would be understood and insightful features such as number of clusters/ clusters spacing and well spacing could be calculated and added to the models.

The MVA model had petrophysical data from an additional approximately 100 horizontal wells. The 2D modeling, optimal trend model, was generated for each petrophysical feature as input to the MVA model and Machine learning forecast and value model: G&G information sampled at the production well location from the model.

Table 1 shows the input properties to the data driven models. Of a total of 16 initial G&G parameters, 7 were rejected in the final models test and train phase, due to collinearity and/or data quality.

G&G Properties MVA Regional Model	G&G Properties ML forecast Model	Completion Properties	Production properties
TOC	TOC	Number of stages	Monthly Volumes
Youngs Modulus	Youngs Modulus	Lateral lenght	
Top of zone	Top of zone	Propannt Pumped	
Base of zone	Base of zone	Fluid pumped	
Porosity	Porosity		
Thickness	Thickness		
Saturation	Saturation		
Pore pressure	Pore pressure		
Pressure gradient	Pressure gradient		
GOR			
Sonic Potenial Log Measurments			
Sonic log Measurments			
Denisty Log Measurments			
Deep Resitivity Log Measurments			
Medium Resitivity Log Measurments			
Shallow Resitivity Log Measurments			

Table 1: G&G Properties used in Regional MVA model & ML Forecast & Value model Completion properties used in both models along with Actual monthly production volumes.

PHASE 1: SPATIAL ANALYTICS

Several spatial analytics workflows that can be stand alone have been generated. The workflows used included an initial optimal trend modeling, KNN Clustering, fair train-test split of spatial data for ML models.

Phase 1A - Optimal trend model

The objective for the development of the optimal trend model is to address uncertainty of properties at unknown well locations and produce robust 2D models of G&G properties. Geological data present challenges for mapping; data paucity is dominant; presence of trends and spatial correlation need to be honored. Many off-the-shelf techniques assume the data are independent and identically distributed, but real geological phenomena present trends. Neglecting

the spatial continuity and inherent data paucity can produce unreliable uncertainty models and impair decision making. The optimal trend model is an innovative approach that combines data analytics, geostatistics, and optimization techniques to provide a workflow to analyze 2D datasets and produce models that are reliable to sample planed production well locations to feed G&G data into the Auto ML forecast prediction model (Salazar *et al*, 2022).

Spatial model includes multiple realizations to access uncertainty and the primary feature can be modeled and co-simulated with features that hold a strong correlation, this can include data from seismic surfaces, modelled surface results from basin models, or surfaces derived from well data.

Methodology

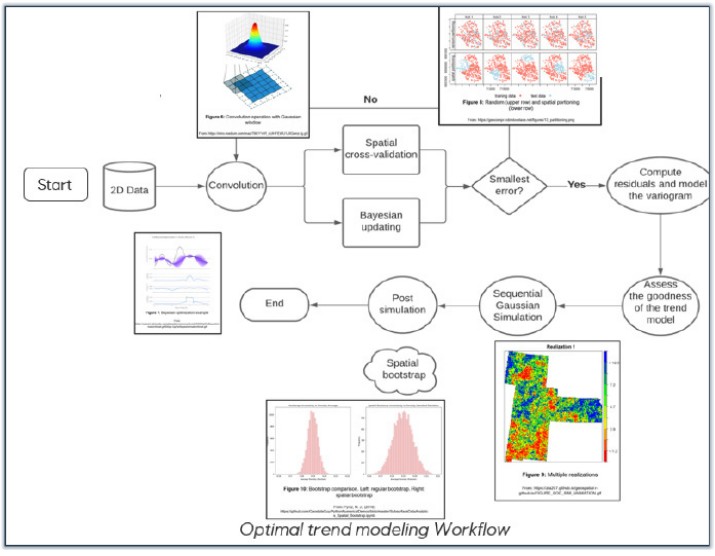


Figure 3. Optimal Trend model workflow combines data analytics, geostatistics and optimization techniques that provides a standalone workflow for 2D modeling from Salazar *et al*, 2022.

Step 1: Outlier Detection

The workflow uses several algorithms to identify outliers for removal. Mahalanobis distance is used and a confidence interval to advise data points that could be outliers based on their location and a confidence interval, next an algorithm that classifies the samples three standard deviations away from the mean as outliers. Then the isolation forest algorithm which categorizes an instance as an outlier by measuring its predisposition to being isolated. Another algorithm used is the elliptic envelope that assumes the data follows a Gaussian distribution and measures the distance of the samples to the central samples to estimate their outlier degree. Finally, the last algorithm is the local outlier factor that uses the distances from each sample to its k-nearest neighbors to

compute a local density and classifies as outliers' samples that have a smaller neighbor density than their k-nearest neighbors.

Step 2: Plotting the experimental semi variograms

This indicates the presence of trends in the data sets, the results are generation of directional semi-variograms and variogram maps, this will depict the trend for the feature being modeled and/or co-simulated.

Optimization is in place using a Gaussian window and an evolutionary algorithm (Salazar, *et al* 2022), finding the optimal dimensions of the Gaussian window equivalent to the grid cells.

Step 3: Identify the direction, azimuth of maximum spatial continuity using a Bayesian optimization.

Step 4: Model the 2D variogram using evolutionary algorithms for optimization of the nugget effect, so the nugget effect and the variance contributions add up, to evaluate the goodness of the semi-variogram model fit.

Step 5: Perform sequential Gaussian simulation: (SGS)

SGS overcome limitations that exist with kriging, Kriged maps are deterministic and therefore incompatible with uncertainty analysis. (Jensen *et al.*, 2000; Journal *et al.*, 2000). SGS create stochastic realizations in reservoir modeling that reproduce the global distribution.

Step 6: Co-simulation

Understanding feature correlation allows for modeling of the subsurface with analysis of more than one feature. The cosimulation method used collocated cokriging to prioritize the reproduction of the primary feature's histogram and variogram while maintaining the Pearson's correlation coefficient with the collocated secondary feature. For cosimulation, the algorithm requires the realizations from the secondary feature, the variance reduction factor using Bayesian optimization is applied.

Results: Output diagnostics

The optimal trend workflow outputs are solving two of the major issues we encounter when working with geological or geophysical properties; first we either underestimate or overestimate

the property values and second the resulting uncertainties are too wide since the trends are still present in the data, violating the stationary assumption of the sequential gaussian simulation approach. The workflow automates the semi-variogram definition, model the trend in the data so we can remove it and correctly apply a sequential gaussian simulation on data that is now stationary. Furthermore, the uncertainty is within range and the realizations are now close to the input histogram. 2D uncertainty maps of the realizations are calculated for P10, P50, and P90 but more realizations can also be added as needed. Finally, a local probability of exceedance map is calculated to better visualize the risk and uncertainty for the property under analysis.

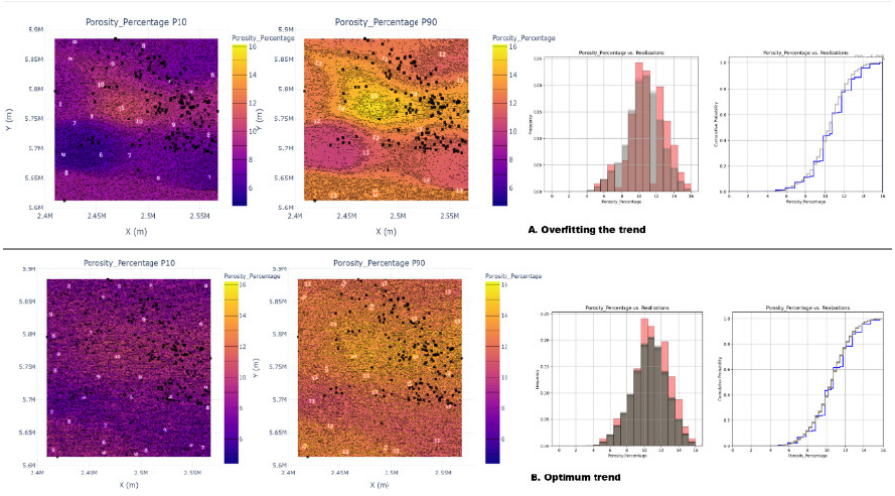


Figure 4.A & B Outputs from the optimal trend model of Porosity. A. Maps with trend overfitting the property (realizations fail to reproduce the histogram, porosity values of 8% and 12%, we will be inaccurate and imprecise at other predrill locations) and B. Maps without the trend after its removal (realizations close to the input histogram, porosity values of 8% and 12% are present now, bullseyes are removed). The variogram model is implemented with diagnostics for the goodness of the model fit to the input data. This model has been run on all properties. G&G inputs for planned development wells and production prediction are sampled from these resulted modeled surfaces considering the trend of the data.

Phase 1B: KNN clustering analysis for the whole basin or the interval of interest

Objective

The objective of non-supervised machine learning technique of KNN clustering is to integrate all G&G properties data into a consistent comparable basis. This allows for regional reservoir quality analysis. The cluster results and, facies map, can be used in conjunction with the features of significance & transform plots, understanding influential drivers on well production, and the values at which they are significant, in the MVA part of the workflow, to understand the gross regional spatial variations of significant features.

Methodology

KNN is a supervised classification algorithm that gathers and groups data into K number of clusters.

The algorithm is used to classify different objects into groups in such a way that the similarity between two objects is maximal if they belong to the same group and minimal otherwise.

Regional geological data consisting of 17 G&G (i.e., geological, geophysical, petrophysical and geomechanical) properties of the Cocina interval underwent a cluster analysis which combines the 17 G&G properties into facies classes. The clustering results are four defined facies classes that characterize the different areas. Two of the classes (1 and 2) had the best petrophysical and property parameters for the Cocina interval (Figure 5).3

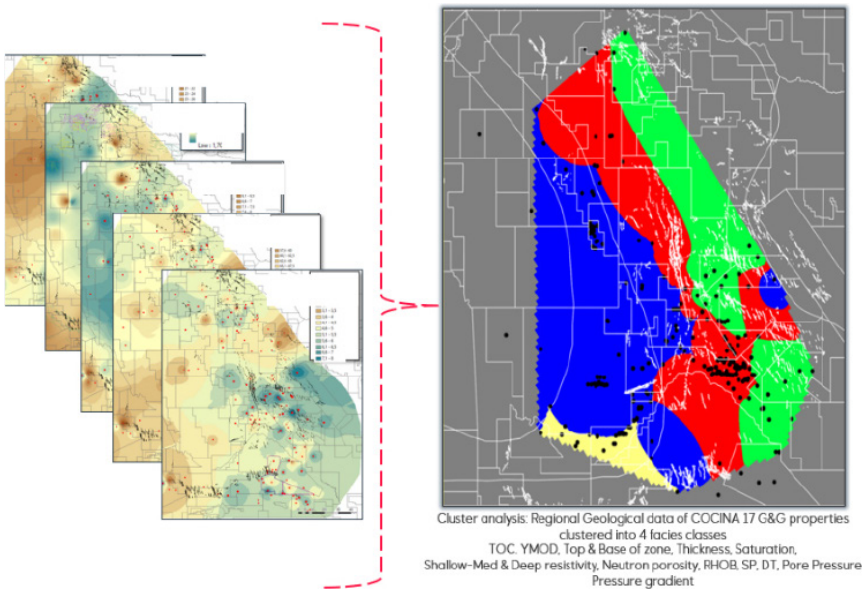


Figure 5. Petrophysical and properties clustering analysis for the Cocina interval. Classes 1 and 2 collected the best geological parameters for the interval. Map and cluster view can be prepared and analyzed

Phase 1C: Fair train-test split of spatial data in Machine Learning

Objective

Mitigation of spatial autocorrelation for improved prediction accuracy in Machine learning models. Overlooking the spatial autocorrelation prevalent in our data from the optimal trend model can result in over optimistic models.

Method

The fair train-test split workflow is a novel cross-validation method for spatial predictive machine learning modeling that provides fair test splits with spatial prediction difficulty distributions (Figure 6).

Step1

The workflow applies the semi variogram model of the target to compute the simple kriging variance as a proxy of spatial estimation difficulty based on the spatial data configuration.

Step 2

The workflow employs a modified rejection sampling to generate a test set with similar prediction difficulty as the planned real-world use of the model...

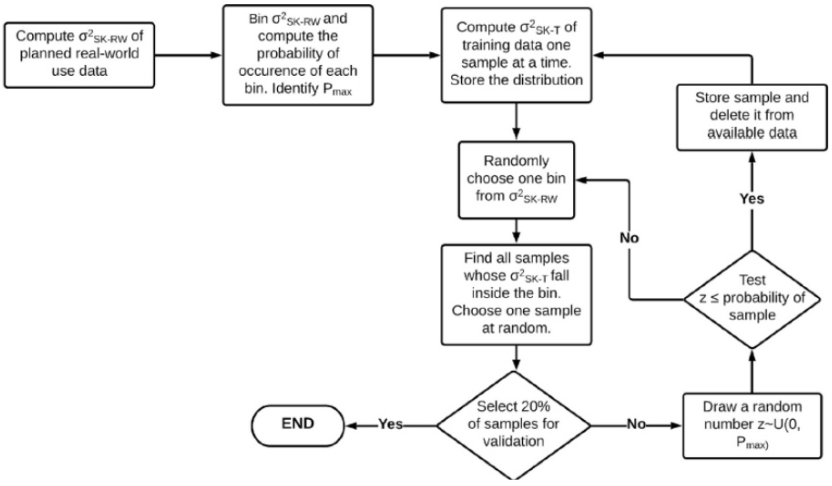


Figure 6. Fair train-test split of spatial data from Salazar *et al.*, 2022.

Results

Figure 7 shows the spatial configuration results or two cross validation methods, the spatial fair train-test split and a validation set approach for the Cocina interval. The spatial train test set choses different test wells to honor the spatial variance in the data set. The workflow outputs are training, and test sets ready for model fit and assessment with the Auto Machine learning forecast model.

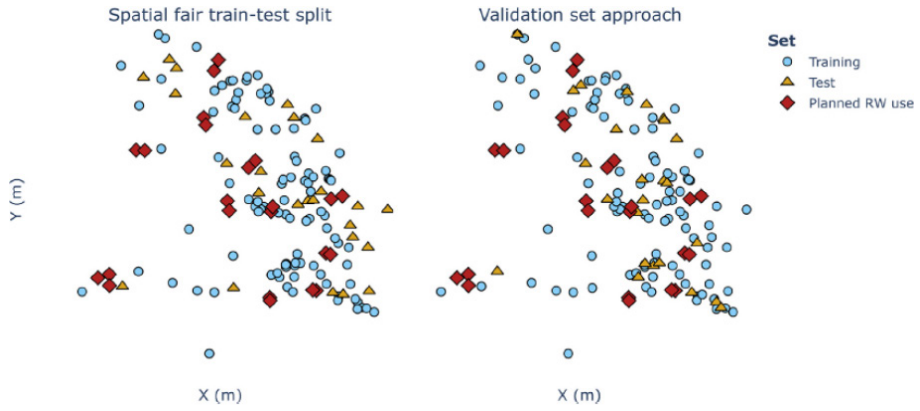


Figure 7. Regional spatial configuration results across the Vaca Muerta Play of realization for two cross-validation methods in the Cocina interval. Left: Spatial fair train-test split. Right: Validation set approach. (RW: real world).

PHASE 2: MULTIVARIATE ANALYTICS (MVA)

Objective

The data-driven model consisted of a multivariate analytics method that included a nonlinear regression model and guided transforms was used to determine how multiple independent predictor variables related to a single dependent response variable, in this case cumulative hydrocarbon production. This gives us insight into what is driving production with the output of feature significance plots. (Figure 10) The model will output feature significance plots through time, so we gain insight into what drives initial production and what drives later production volumes. The objective of analyzing and explaining the effect of multiple independent predictor variables (G&G and completion features) on the response variable: actual production, is facilitated by transforming the independent variables (Duolao Wang *et al* 2004).

The transforms give insight into the relationship of these variables. This assists in describing the relationships and uncovers non-linear relationships (Duolao Wang *et al* 2004). The transform algorithm is a decerning algorithm that gives more detailed information on the value ranges at which significant features are impactful, in this case on actual production.

These results can be compared with the clustered facies classes to understand regional productivity potential. This gives insight on the effect of spatial variation on actual production.

The transform plots can identify optimal engineering designs within the data. This information can be passed to the engineer to answer more granular completions questions i.e., what is the optimal clusters spacing and number of perforations per cluster? (Arief *et al*, 2022).

The objective is to complement physics-based models as spatially aware data driven models allow for insight away from the small region of detailed data.

Methods

We applied a multivariate analytics method to evaluate the predict twelve months of cumulative oil production. This data analytics method allows integration of statistical tools with geological, geophysical, and engineering data for predictive modeling and quantitative analysis. The MVA method uses an algorithm that includes a proprietary nonlinear regression (NLR) model, from a commercial software (Enverus), determining how multiple independent predictor variables relate to a single dependent response variable.

The Enverus proprietary transform algorithm was applied to the variables. This algorithm operates by applying a standard normalization to the response variable by subtracting its mean and dividing by the standard deviation and by normalizing the predictor variables to be zero mean by subtracting the mean of their respective distributions. The model in this study also used a transforming response variable (Y) to better match the combined system of transformed predictor variables, $\theta(Y)=\sum_{(i=1)}^{p}[\varnothing_i(X_i)+\varepsilon]$ where θ is a function of the response (Y), and $\varnothing_i$ are functions of the predictors (X_i), i=1, …, p.

In this MVA workflow, the data is reviewed using an outlier analysis, where outliers are identified and removed using a non-parametric technique based on distribution smoothing and rejection threshold arguments (Figure 8). In this study, a full univariate analysis was done for each outlier candidate. Then, variables are analyzed for redundancy using a multicollinearity analysis that is based on a specified correlation threshold. In cases where variables are highly correlated, only the one with the lesser overall aggregate correlation to the others is retained. This is followed by analyzing the variables using a correlation table (Figure 9) where predictor and response variables are evaluated interactively. Then, standard and rank correlations are calculated for every variable pair with a customized color to denote the correlation threshold.

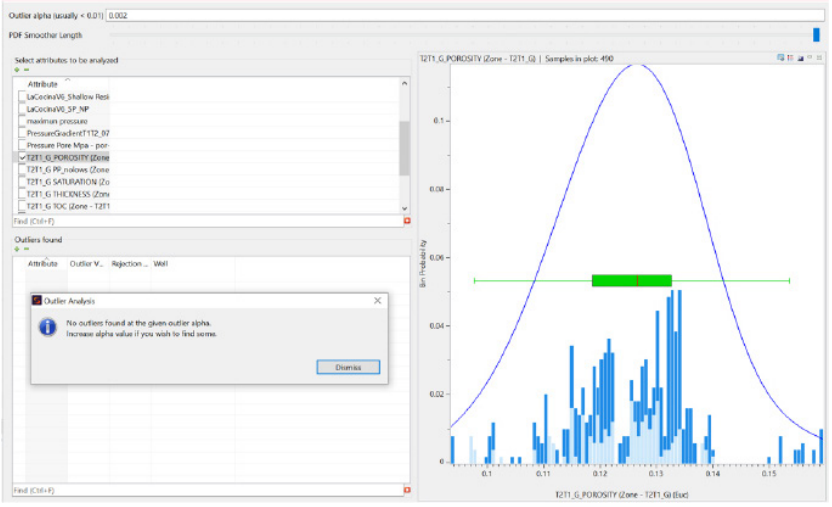


Figure 8. Data outlier analysis and collinearity workflow for maximum multiple correlation among predictor variables.

Correlation Table 1 | Data Correlations (144)

	T2T1_G T...	T2T1_G ...	T2T1_G T...	ClustersT...	T2T1_G_...	T2T1_G ...	T2T1_G T...	T2T1_G ...	T2T1_G L...	T2T1_G ...	T2T1_G ...	Pressure...	Pressure ...	Water inj...	Total San...	Sand pu...	Sand pu...	Horsepo...	Well -
T2T1_G TOC (Zone - T2T1_G)	1.0	0.135	0.429	-0.569	0.306	0.429	0.007	0.746	-0.132	0.332	0.512	-0.043	-0.087	-0.177	-0.067	0.077	-0.095	0.061	0.073
T2T1_G YMOD (Zone - T2T1_G)	0.135	1.0	-0.087	0.219	-0.061	-0.08	0.353	0.029	0.149	-0.099	-0.241	-0.209	0.492	0.133	0.134	-0.048	0.14	0.177	0.034
T2T1_G TopZone (Zone - T2T1_G)	0.429	-0.087	1.0	-0.76	0.176	1.0	0.03	0.681	-0.249	0.843	0.742	-0.237	0.008	-0.142	-0.097	0.127	-0.144	0.113	-0.145
ClustersT2T1_G (Zone - T2T1_G)	-0.569	0.219	-0.76	1.0	0.025	-0.763	-0.188	-0.785	0.171	-0.696	-0.684	0.062	0.377	0.103	0.004	-0.036	0.021	-0.01	0.05
T2T1_G POROSITY (Zone - T2T1_G)	0.306	-0.061	0.176	0.025	1.0	0.161	-0.808	0.021	-0.369	-0.147	0.106	0.163	0.184	-0.316	-0.296	0.225	-0.366	0.224	-0.116
T2T1_G BaseZone (Zone - T2T1_G)	0.429	-0.08	1.0	-0.763	0.161	1.0	0.048	0.686	-0.242	0.848	0.741	-0.243	0.009	-0.136	-0.091	0.124	-0.138	0.109	-0.142
T2T1_G THICKNESS (Zone - T2T1_G)	0.007	0.353	0.03	-0.188	-0.808	0.048	1.0	0.272	0.339	0.302	-0.064	-0.356	0.014	0.298	0.315	-0.137	0.342	-0.186	0.142
T2T1_G SATURATION (Zone - T2T1_G)	0.746	0.029	0.681	-0.785	0.021	0.686	0.272	1.0	-0.178	0.723	0.544	-0.167	-0.041	-0.083	0.056	0.169	-0.03	0.207	-0.059
T2T1_G LATERAL_LENGHT (Zone - T2T1_G)	-0.132	0.149	-0.249	0.171	-0.369	-0.242	0.339	-0.178	1.0	-0.109	-0.186	0.053	0.025	0.712	0.633	-0.299	0.696	-0.13	0.541
T2T1_G PP_orig (Zone - T2T1_G)	0.332	-0.099	0.843	-0.696	-0.147	0.848	0.302	0.723	-0.109	1.0	0.705	-0.284	0.002	-0.01	0.056	0.114	-0.004	0.069	-0.132
T2T1_G PP_nolows (Zone - T2T1_G)	0.512	-0.241	0.742	-0.684	0.106	0.741	-0.064	0.544	-0.186	0.705	1.0	-0.002	-0.289	-0.126	-0.051	-0.016	-0.037	-0.102	0.089
PressureGradientT1T2_07	-0.043	-0.209	-0.237	0.062	0.163	-0.243	-0.356	-0.167	0.053	-0.284	-0.002	1.0	-0.095	0.101	0.137	-0.021	0.13	-0.006	0.161
Pressure Pore Mpa - por-sp-depth-temp-cgr	-0.087	0.492	0.008	0.377	0.184	0.009	0.014	-0.041	0.025	0.002	-0.289	-0.095	1.0	0.038	-0.01	0.211	-0.108	0.458	-0.223
Water injected	-0.177	0.133	-0.142	0.103	-0.316	-0.136	0.298	-0.083	0.712	-0.01	-0.126	0.101	0.038	1.0	0.943	-0.12	0.886	0.089	0.678
Total Sand Pumped	-0.067	0.134	-0.097	0.004	-0.296	-0.091	0.315	0.056	0.633	0.056	-0.051	0.137	-0.01	0.943	1.0	-0.009	0.884	0.086	0.683
Sand pumped - Imported	0.077	-0.048	0.127	-0.036	0.225	0.124	-0.137	0.169	-0.299	0.114	-0.016	-0.021	0.211	-0.12	-0.009	1.0	-0.475	0.357	-0.314
Sand pumped - National	-0.095	0.14	-0.144	0.021	-0.366	-0.138	0.342	-0.03	0.696	-0.004	-0.037	0.13	-0.108	0.886	0.884	-0.475	1.0	-0.091	0.748
Horsepower fracture equiptment	0.061	0.177	0.113	-0.01	0.224	0.109	-0.186	0.207	-0.13	0.069	-0.102	-0.006	0.458	0.089	0.086	0.357	-0.091	1.0	-0.272
Well - NumberOfStages_copy-4 Nearest Point...	0.073	0.034	-0.145	0.05	-0.116	-0.142	0.142	-0.059	0.541	-0.132	0.089	0.161	-0.223	0.678	0.683	-0.314	0.748	-0.272	1.0
LaCocinaV6_SP_NP	0.082	0.533	-0.064	0.357	0.345	-0.064	-0.048	-0.043	0.019	-0.198	-0.416	-0.094	0.743	-0.06	-0.157	0.126	-0.197	0.345	-0.185
LaCocinaV6_DT_NP	-0.623	-0.424	-0.482	0.651	0.125	-0.49	-0.488	-0.69	-0.017	-0.52	-0.315	0.38	0.055	-0.008	-0.082	-0.023	-0.061	-0.133	0.042
LaCocinaV6_CGR_norm_NP	0.589	0.139	0.664	-0.466	0.423	0.661	-0.139	0.627	-0.259	0.485	0.354	-0.214	0.388	-0.205	-0.198	0.237	-0.285	0.466	-0.232
LaCocinaV6_RHOB_norm_NP	-0.086	-0.272	0.032	-0.101	-0.257	0.032	-0.004	0.01	-0.059	0.173	0.435	0.227	-0.374	0.014	0.102	-0.055	0.115	-0.129	0.129
LaCocinaV6_Deep-Resistivity_NP	0.497	-0.159	0.497	-0.799	-0.355	0.505	0.433	0.625	0.039	0.569	0.602	-0.097	-0.508	0.075	0.173	-0.063	0.182	-0.167	0.11
LaCocinaV6_Neutron Porosity_NP	0.383	0.001	0.509	-0.641	-0.062	0.516	0.401	0.57	-0.083	0.439	0.332	-0.164	0.016	-0.107	-0.05	-0.03	-0.03	-0.047	-0.077
LaCocinaV6_Medium Resistivity_NP	0.688	0.169	0.701	-0.845	-0.069	0.708	0.406	0.852	-0.099	0.678	0.525	-0.308	-0.178	-0.073	0.007	0.011	0.001	0.045	-0.029
LaCocinaV6_Shallow Resistivity_NP	0.786	0.244	0.468	-0.617	0.144	0.471	0.138	0.864	-0.195	0.51	0.464	-0.03	-0.024	-0.118	0.02	0.146	-0.05	0.256	-0.055
Well - CocinaV6_maxT-4 Nearest Points	0.69	0.08	0.789	-0.849	0.148	0.792	0.164	0.868	-0.189	0.674	0.606	-0.112	-0.046	-0.116	0.007	0.152	-0.065	0.25	-0.085
LaCocinaV6_vitrinite_reflectance_top-4 NP	-0.74	-0.072	-0.59	0.788	-0.044	-0.594	-0.206	-0.874	0.202	-0.536	-0.595	0.059	0.234	0.119	-0.031	-0.049	-0.004	-0.046	-0.039
Horizontal length	-0.132	0.149	-0.249	0.171	-0.369	-0.242	0.339	-0.178	1.0	-0.109	-0.186	0.053	0.025	0.711	0.632	-0.299	0.696	-0.131	0.541
maximun pressure	-0.104	0.11	0.045	-0.004	-0.227	0.047	0.139	0.036	0.385	0.11	-0.05	-0.025	0.138	0.422	0.43	-0.085	0.418	0.32	0.171
CUMOil_365days	0.461	0.015	0.56	-0.661	0.091	0.562	0.106	0.598	0.194	0.46	0.403	0.046	-0.037	0.141	0.199	0.063	0.146	0.284	0.034

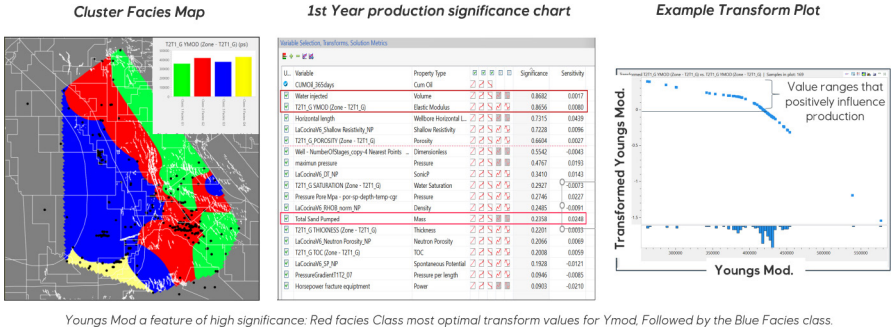
Figure 9. Data correlation table. Seventeen variables selected after the collinearity analysis to model and predict the last 365 days of cumulative oil production.

In this study, several variables were rejected (e.g., deep resistivity, medium resistivity, VSH,) from the total of more than 30 variables including geological, geophysical, and engineering variables listed in table 1.

A model parameterization was then prepared including sample sizing and seed number assignments to support repeatability. The solutions were not biased to the range of the input data and there was enough data in the horizontals and vertical tracks within the interval of interest to include and avoid biasing.

After the parameterization was completed, we performed a model correlation where the predicted response from the model was compared to the actual response from the input data, showing standard and rank correlations. The feature significance results for the 1st year of actual production is illustrated in Figure 10.

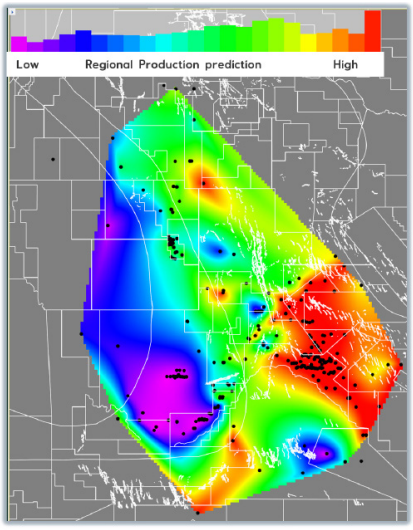
Finally, the model was validated using N-fold and leave-one-out cross-validation options that provided two methods of validation for the predictive model. The same amount of error was achieved in each method that conveys a well-conditioned and consistent model.



Youngs Mod a feature of high significance: Red facies Class most optimal transform values for Ymod, Followed by the Blue Facies class.

Figure 10. Shows the features of significance and their sensitivty with the correlation coefficents for the Non-linear regression in the MVA, for the first years of actual production (Response varaiable). This is shown with the Facies class map from the cluster analysis and an example transform plot: of youngs modulus, showing the actual values at which the feature positivily.

A nonlinear predictor-response relationship was analyzed to determine the optimal transformation that should be applied to each predictor variable. Once the functional relationships were established for each variable, the transformed predictor variables re-expressed the predictor-response relationships as continuous functions that minimize error between the actual and predicted response variable values.



Response to transformed derived « optimal Completion» showing the regional variation due to variance in geological features significant to production

Figure 10 b. Regional production prediction: Regional response to transformed derived «optimal Completion» showing the regional variation due to variance in geological features significant to production.

Results

The key results of the MVA for insight to well optimization by understanding drivers of productivity away from detailed well data showed that volume of water injected and Young's modulus are the most significant features during the 1st year of production. Both of these properties are very close in significance. This is followed by horizontal length, resistivity, and porosity. Proppant is seen to be less significant in early production.

After 3 to 5 years of production the most significant features are similar, Young's modulus, volume of water injected, porosity and TOC increased in significance for later production. Also, proppant pumped increased in significance for the 3-5 years production compared to the first year.

A set of transformed responses were generated for all the input features. Two-fold information is gained from this. First in figure 10 a: an example of the transform plot for Young's modulus is given, we gain more detailed information on this feature, 3.5 - 4 Gpa are the optimal value ranges for Young's modulus having a positive impact on production.

This transform plots can be combined with the Facies Class map from the KNN cluster workflow to see which facies classes have the optimal high significant feature and at which values

they positively impact production. We see in the Young's modulus example that the Red facies class is the most optimal for production.

Secondly, we can look at the best values from the transform plots for the engineering parameters, water injected, proppant pumped, number of stages & lateral length. We can gain an optimal well design from the best values in each transform. We can then predict production with this "optimal well design" and see regionally where uplift in production can be gained, where in the basin responds the most positivity, Figure 10 b. The transform plots for proppant values where the total amount of sand pumped are impacting production are within a of very close range with a limited impact on production, when compared to the total water injected volumes or horizontal length.

The "optimal well design" from transform plots is used to generate completion scenarios for the auto ML forecast and value model at a specific development plan area and passed to the engineers, so physics-based models are run to understanding the granularity of why increased fluid is highly significant to production, Cruz, *et al* 2021 & Arief *et al*, 2022. The physics-based models can also run scenarios that extrapolate beyond the data.

Simulations were performed to account for the oil uplift when using the different values from each Transform in the model. The results indicate that a 49% oil uplift can be achieved for the water injected volume versus less than 20% for the sand pumped into Formation (Cruz, *et al*, 2021).

Phase 3: Auto ML forecasting model & Value Model

Objective

The objective was to build an automated machine learning model for forecasting prediction of planned development wells and to use it to investigate what an economically optimal well completion design could be.

Here we wanted to integrate all the G&G, production, and completion data on a basin scale to forecast the production of hypothetical well locations in a development plan for an area of interest. A workflow was established to swiftly test varying stimulation designs evaluating the economics of the chosen design given the regional varying geological input.

The final goal was then to bring the production forecast to commercial value with the addition of a scripted cost and economy model for economic scenario testing and optimization and calculation of economic metrics like well NPV, break-even and payback in years. We would then be able to test for the economically optimal well design throughout any license area under varying price scenarios in just minutes.

Figure 13 shows the full workflow. Dashboards of all the results, predicted production forecasts

and economic metrics were created to facilitate discussions with all disciplines (i.e., geologists, engineers and economists).

Method

The Auto ML forecast model uses an XG Boost algorithm. XGBoost is a decision tree-based ensemble Machine Learning algorithm. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient and flexible. Figure 11 shows the ML forecast model workflow. The workflow follows the typical machine learning steps: 1) Feature engineering and feature selection, 2) Model training, 3) model performance evaluation and finally 4) prediction.

Feature engineering is the first step of the workflow. This allows for understanding of correlation and collinearity between the input G&G and engineering parameters. This is important so the algorithm can understand the influence of independent variables on the production and predict based on learning from the selected features data set.

This step is performed using multiple methods. Principal component analysis: Pearson's correlation coefficient is used for covariance and correlation coefficient to account for bivariate variance /covariance in linear relationship instances. The rank correlation coefficients where then used to relax the linear assumptions within the data and removing the sensitivity to dispersion of the data and then partial correlation coefficient used for understanding correlations due to the multivariate nonlinear nature of the data set. Finally, mutual information method was used where we can quantify the amount of information each feature holds, assuming independence and by comparing to all features. The results from these methods along with subject matter expertise were used to select the features to input into the model.

The next step is a classification and clustering of the input production wells of approximately 760 horizontal wells across the basin, the algorithm is searching for similar wells or unbiased analogue wells based upon multiple G&G features as input to the predicted wells production. This output seen in Figure 12 and allows the engineer to see and evaluate the existing production wells that contribute to the predicted forecasts of a given well in the development plan.

The model is tested and trained on the selected features and the models performance is evaluated, the main outputs of this are the mean squared error and R2 result of the prediction time steps, 6 months, 1 year, 2 years and 3 years predictions. Finally, a predicted forecast is generated for every well in the development plan of the study area. The first 3 years are prediction with decline curve analysis using an Arps equation pushing the forecast out to nine years. Prediction of EUR, 30 years production, was tested with multiple B factors, but the result was deemed too uncertain. The uncertainty in the nine-year forecast was +/-10%.

Forecast predictions were generated for 13 different completion scenarios with input from the MVA and engineers.

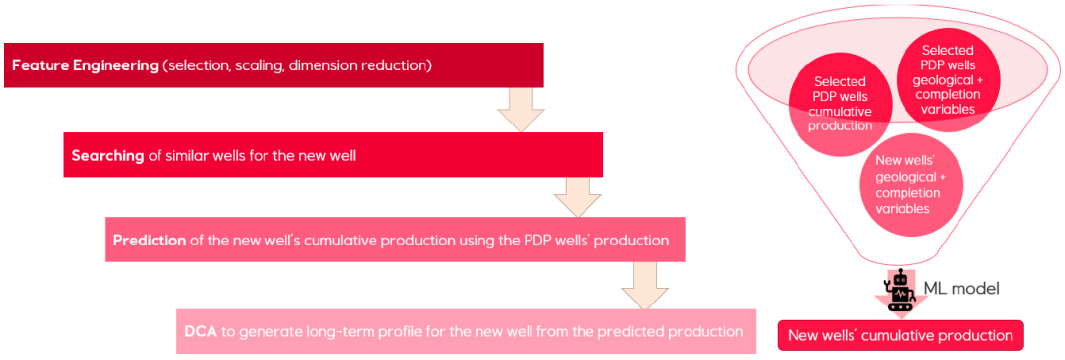


Figure 11. Auto ML forecast model workflow.

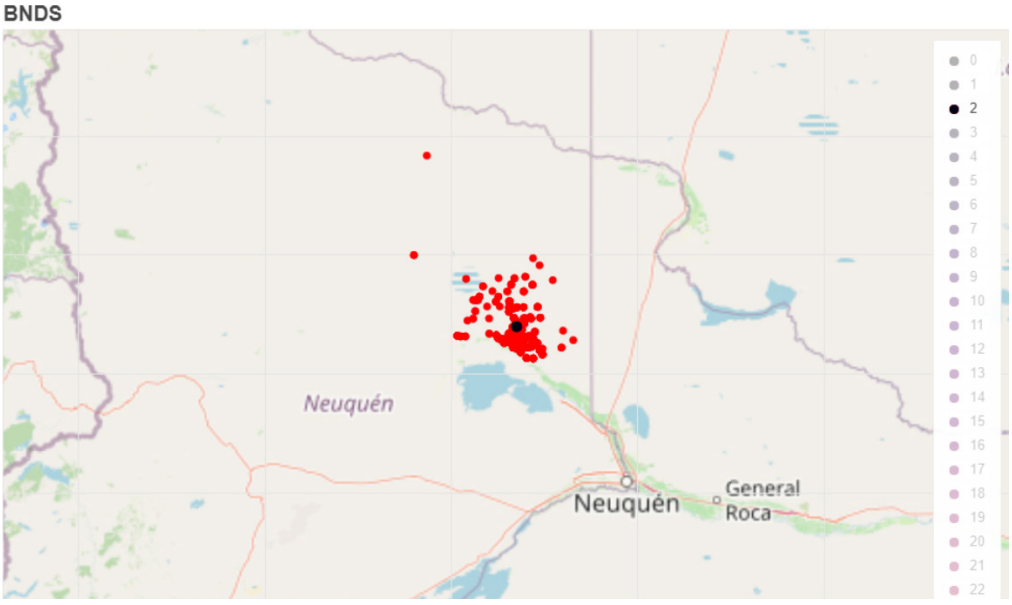


Figure 12. Planned well 2 shown in Black with contrbuiting wells to well 2 forecast prediction in red. Clasification step: unbiased selection of similar Wells.

To bring the forecast to value, a scripted economic and cost model was developed. This was a collaboration between the software engineer, the drilling and well engineer and the economic analyst. A joint model of cost and single well economy from the analysis was used to design the code and perform QA/QC of the model outputs. The code replicated results from the analysis software without any discrepancies. The code calculates the well payback time within a band of uncertainty and other short-term measures of value, IRR, NPV and break-even oil price. The production forecast outputs were linked to the cost and economy model and a dashboard integrating all outputs from the Auto ML forecast and value model was generated.

In the dashboard completion scenarios could be compared, well production forecast, NPV, payback in years and IRR could be visualized to gain insight in the most economically viable optimal completion. Figure 13 shows the full workflow.

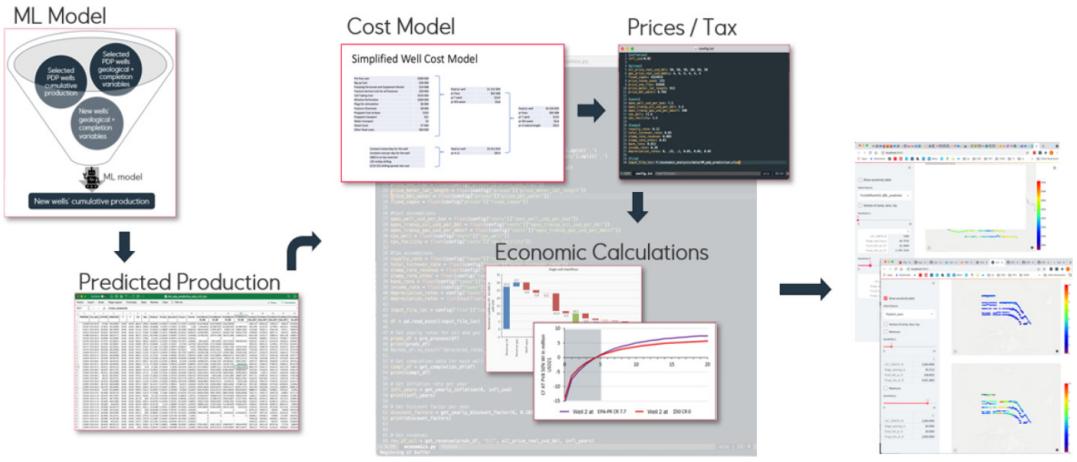


Figure 13. Full Auto ML forecast and value model.

Results

The following results are generalized to show major findings of the economic optimization of wells at the study area and does not intent to present any true value assessments of any asset.

An auto ML forecast model was trained and tested in the area of interest to predict a production forecast for each hypothetical well in a development plan. Thirteen different forecast scenarios per well where run each with a different completion design. Stage spacing, fluid intensity, proppant loading and lateral length where varied. Results from the MVA showed a high significance of water intensity in the first and third year of production, with proppant having a lesser effect on production. This was supported in the auto ML forecast results and the full value model workflow gave insight to the economy of well completion optimization.

Figure 14 a shows the production prediction on scenarios 1 & 2. Here the number of stages, fluid pumped, and lateral length are kept equal, and the only difference is proppant intensity. Scenario 2 has 2.5 times the proppant per stage than Scenario 1. The difference map shows an universal uplift in production of 10-20 kbbls of oil. However, the well payback maps and the difference map, reveal that the increased proppant erodes value by increasing the well payback time by 1-2 years. Hence, a higher proppant intensity and the resulting minor increased production appear not to be profitable.

Figure 14 b shows the production prediction of increased fluid intensity (scenarios 3 & 4). Both scenarios have the same number of stages and lateral length. Scenario 4 has 2,5 times higher fluid intensity and 1,3 times the proppant intensity than Scenario 3. The difference map shows a significant universal uplift in production by increasing fluid intensity, the central and western areas gain significantly in increased production with increased fluid pumped. Up to 100 KBbls oil increase is seen in most wells, with a minimum of 30 kbbls. The production increase leads to a decrease in well payback time by 1 to 2,5 years.

The auto ML forecast model allows for new iterations to be run fast; based on the findings we can easily change the completion design and rerun or test for different oil price scenarios.

Figure 15 a & b are dummy scenario plots from the Auto ML forecast dashboard for illustrative purposes to demonstrate the additional information the model outputs. The results are hypothetical. Figure 15a shows the additional metrics derived from the cost and economic model. The well location dot color reflect IRR (red indicates a high IRR and green indicates a lower IRR). For each well we display production, break-even and NPV as function of time. Breakeven is penalized due to the limiting the forecast to 9 years. A low break even at 9 years will in the life of the well get more robust.

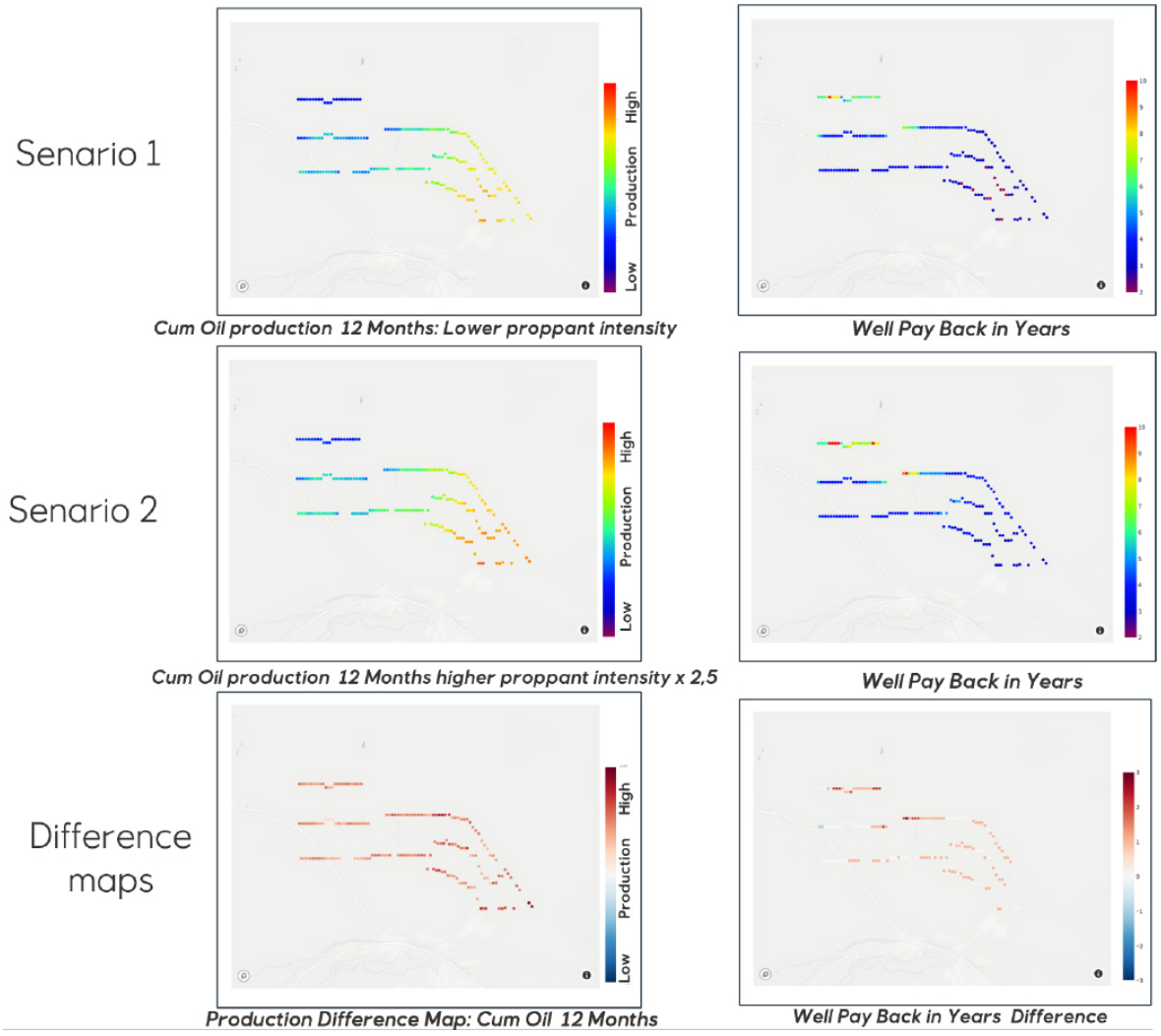


Figure 14. A: Production and well payback maps for scenario 1 & 2 and difference maps. Upper and middle left picture show the 12 months cumulative production of all hypothetical development wells for two different completion scenarios (Scenario 1 & 2), whereas the lower left shows the production difference between the two scenarios. The right figures show the well payback in years for the same two scenarios and the difference. Scenarios 1 and 2 represent two completion designs with different proppant intensities.

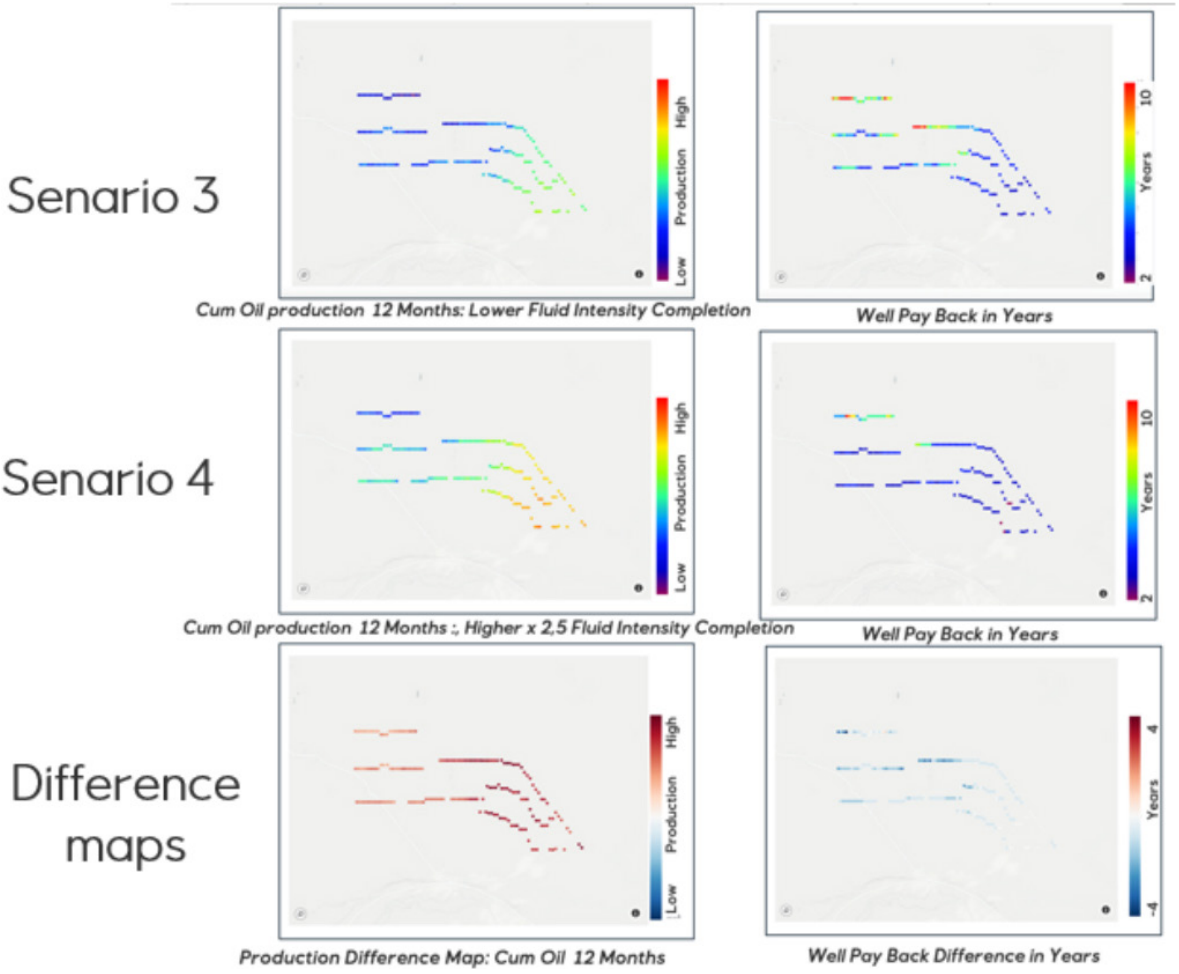


Figure 14. B: Production and well payback maps for scenario 3 & 4 and difference maps. Scenarios 3 and 4 represent two completion designs with different fluid intensities.

Figure 15 b shows a shapely plot (https://en.wikipedia.org/wiki/Shapley_value). This plot shows the weighted average of the features (G&G & Engineering variables) contribution significance to production. Information on each well and which features are most significant for the resulting production forecast.

The variation in production in Figure 14 a and b and shapely values in Figure 15 b reflects the variation in the underlying G&G properties data. The dashboards where all 13 completion scenario forecasts and well payback time, IRR, NPV and Break-even are compared and analyzed facilitates multidisciplinary discussion between geologists, engineers and economists and decision makers.

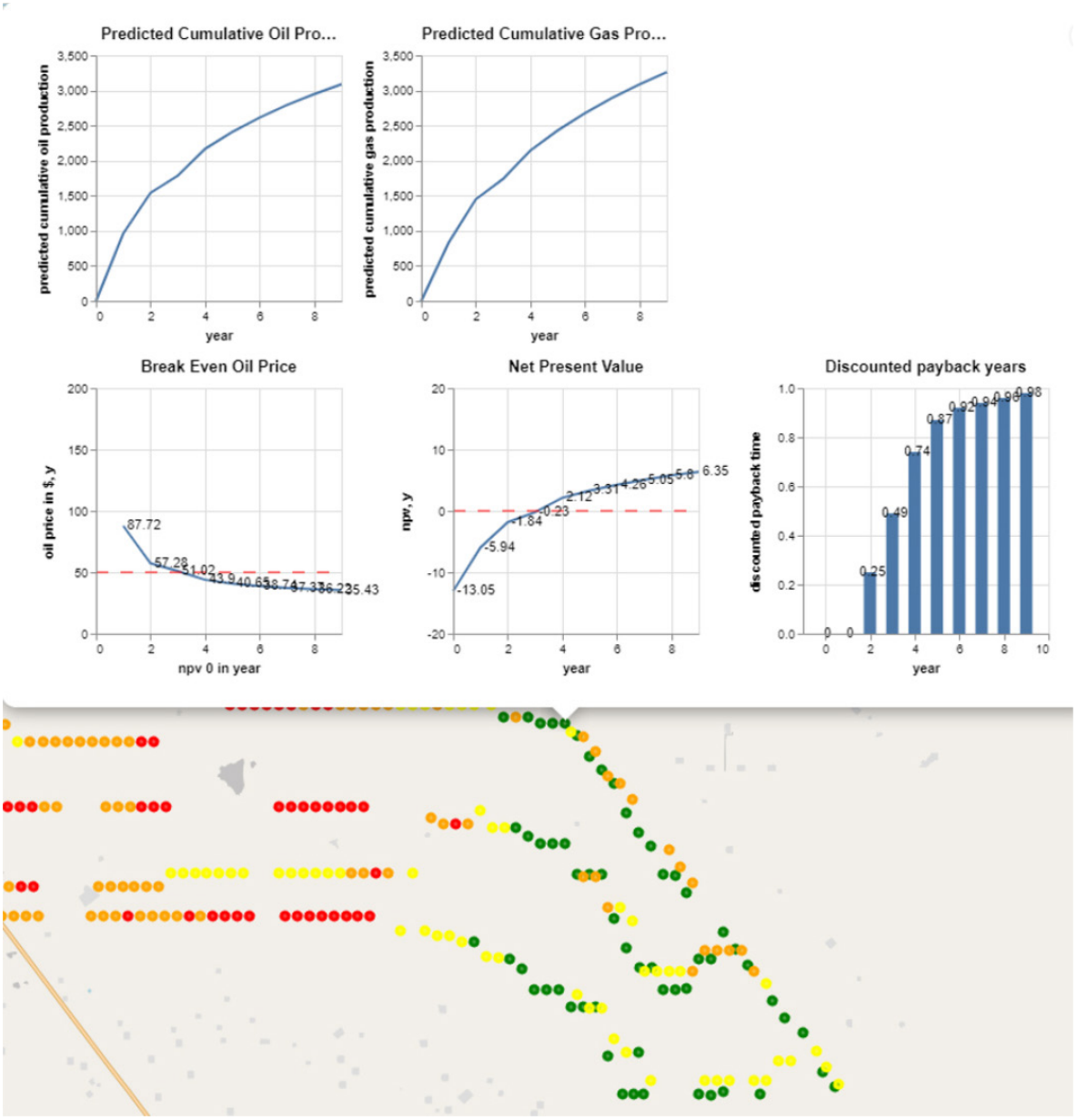


Figure 15. Figure 15 a Shows the financial information in the Auto ML forecast output dashboard. Color of dots in this picture reflects IRR (red is high IRR and green is low). The graphs on the top show predicted cumulative oil and gas production. The lower row of graphs shows Break-Even, NPV and well payback in years with uncertainty. The numbers are hypothetical and for illustration purposes only.

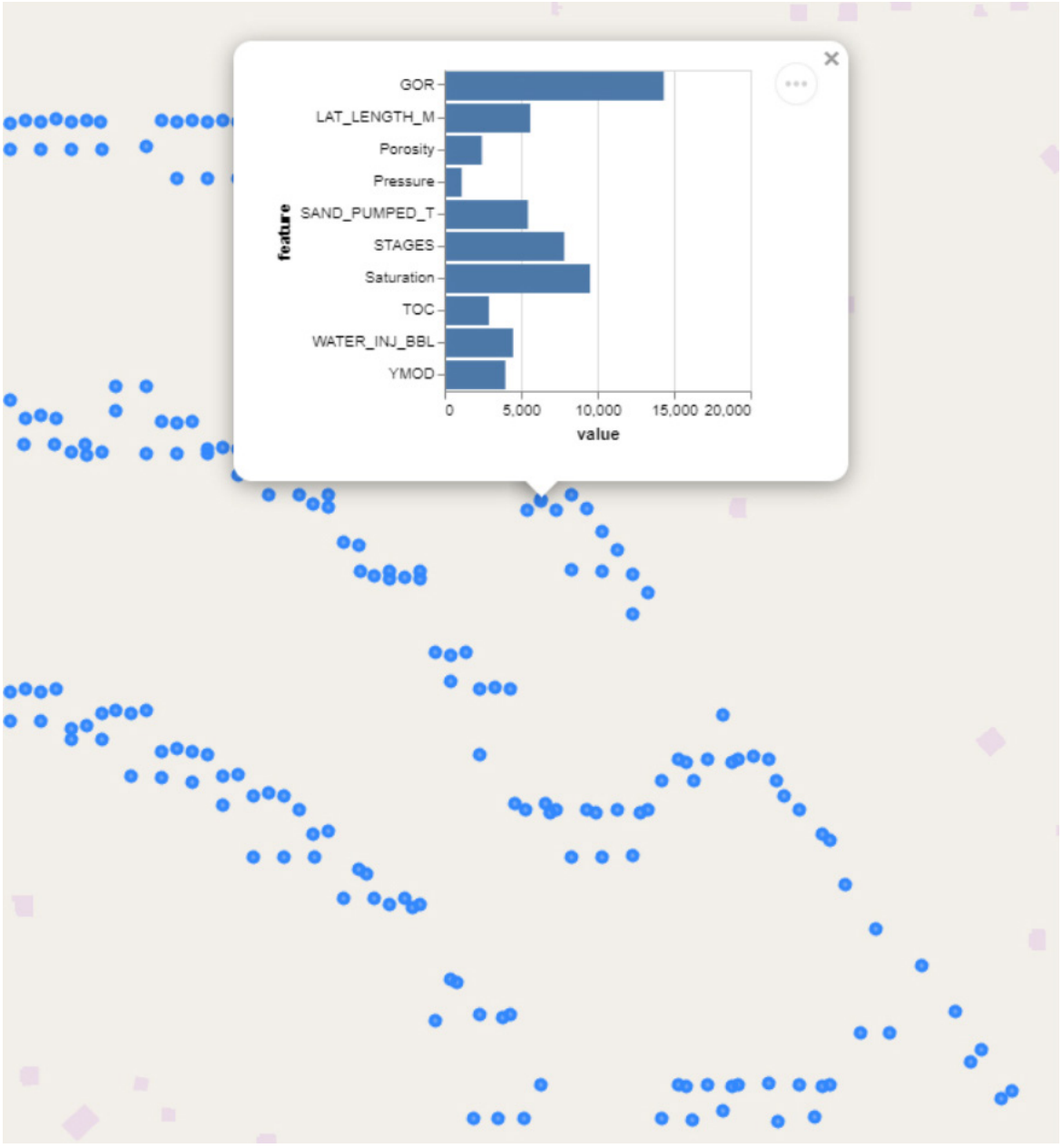


Figure 15 b: Blue dots represent the well locations in the development plan, chart is a Shapley plot showing the significance of the G&G and engineering features that contribute to the production (Not real Values: Illustrative only).

CONCLUSIONS

The regional MVA results showed that the most significant geological feature that drives production is the Young's Modulus. The most significant engineering feature was volume of water pumped. Increased / high proppant volume was seen to have a much lower significance in

driving production. Lateral length and total injected fluid per unit length are identified as primary production drivers in both data-driven and physics-based models (Cruz *et al.* 2021, Arief *et al.* 2022).

The facies class map from KNN clustering shows the "red" facies as the most favorable for geological features of high significance to production.

The auto ML forecast and value model in the study area concurred with the MVA showing a significant uplift in production between the low and high-water intensity completion. This proved to be economically viable as the increased production with increased water volume decreased payback of all the wells in the development plan, the majority up to 2 years faster payback time.

The auto ML forecast and value model also showed that higher proppant volumes had only a marginal uplift in production, yet this eroded the value of the wells as it increased payback time by 1- 2 years.

Physics based models can be used to answer more detailed and granular completions questions i.e., what is the optimal clusters spacing and number of perforations per cluster? (Arief *et al*, 2022).

However, physics-based models are generally time and computationally intensive and, depending on the degree of reservoir heterogeneity in the studied basin/reservoir, and may only be applicable to a small region near the studied well or pad. Spatially aware data-driven models give insight on parameters influencing production from the small region of detailed data. Detailed location and completion data for production wells would allow for calculation of the importance of well and clusters spacing, well interference and parent/child effects are not addressed in this paper but can easily implemented as features into the models.

ACKNOWLEDGMENTS

Equinor would like to acknowledge the collaboration with the University of Texas on the spatial analytics workflows. The authors wish to acknowledge the greater Equinor Global analytics team, Knut Utne Hollund & Haoyuan Zhang. The authors also wish to acknowledge the help from Equinor colleagues, Morten Fejerskov, Mogens Ramm & Raghavendra Kulkarni in the writing of this technical paper.

REFERENCES

Cruz, L., and Ochoa, J. 2021. Physics-based and Data-driven models to predict production drivers in the Vaca Muerta Formation. Unconventional Resources Technology Conference, Houston – Texas.

Duolao Wang and Michael Murphy Estimating optimal transformations for multiple regression using the ACE algorithm Journal of Data Science 2(2004), 329-346

Jensen, J., Lake, L.W., Corbett, P.W.M., Goggin, D.,

2000. Statistics for Petroleum Engineers and Geoscientists, Second. Elsevier Science, Amsterdam.

Salazar, J., Garland, L., Ochoa, J., Pyrcz, M., 2022. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. J. Petrol. Sci. Eng. 209, 109885. https://doi.org/10.1016/j.petrol.2021.109885

Salazar, J., Ochoa, J., Garland, L., Pyrcz, M., 2022. Spatial data analytics – assisted subsurface modeling: a Duvernay case study. J. Petrol. Sci. Eng. In press, 215 part B.