energies

MDPI

Article

Machine Learning-Based Production Prediction Model and Its Application in Duvernay Formation

Zekun Guo *, Hongjun Wang, Xiangwen Kong, Li Shen and Yuepeng Jia

The Research Institute of Petroleum Exploration & Development CNPC, Beijing 100083, China; whj@petrochina.com.cn (H.W.); kongxwen@petrochina.com.cn (X.K.); shenli19980622@163.com (L.S.); jyuep1996@163.com (Y.J.)

* Correspondence: gzk@petrochina.com.cn; Tel.: +86-186-0081-1053

Abstract: The production of a single gas well is influenced by many geological and completion factors. The aim of this paper is to build a production prediction model based on machine learning technique and identify the most important factor for production. Firstly, around 159 horizontal wells were collected, targeting the Duvernay Formation with detailed geological and completion records. Secondly, the key factors were selected using grey relation analysis and Pearson correlation. Then, three statistical models were built through multiple linear regression (MLR), support vector regression (SVR), gaussian process regression (GPR). The model inputs include fluid volume, proppant amount, cluster counts, stage counts, total horizontal lateral length, gas saturation, total organic carbon content, condensate-gas ratio. The model performance was assessed by root mean squared errors (RMSE) and R-squared value. Finally, sensitivity analysis was applied based on best performance model. The analysis shows following conclusions: (1) GPR model shows the best performance with the highest R-squared value and the lowest RMSE. In the testing set, the model shows a R-squared of 0.8 with a RMSE of 280.54 $\times$ 10^4 m^3 in the prediction of cumulative gas production within 1st 6 producing months and gives a R-squared of 0.83 with a RMSE of 1884.3 t in the prediction of cumulative oil production within 1st 6 producing months (2) Sensitivity analysis based on GPR model indicates that condensate-gas ratio, fluid volume, and total organic carbon content are the most important features to cumulative oil production within 1st 6 producing months. Fluid volume, Stages, and total organic carbon content are the most significant factors to cumulative gas production within 1st 6 producing months. The analysis progress and results developed in this study will assist companies to build prediction models and figure out which factors control well performance.

Keywords: machine learning; sensitivity analysis; production prediction; grey relation analysis

Citation: Guo, Z.; Wang, H.; Kong, X.; Shen, L.; Jia, Y. Machine Learning-Based Production Prediction Model and Its Application in Duvernay Formation. *Energies* **2021**, *14*, 5509. https://doi.org/10.3390/en14175509

Academic Editor: Galih Bangga

Received: 2 August 2021
Accepted: 31 August 2021
Published: 3 September 2021

1. Introduction

There are enormous shale resources distributed worldwide and it requires advanced exploration and development strategies to get economic production. Due to the application of horizontal wells and hydraulic fracturing technologies, the shale reservoirs have achieved an economic production, which plays an important role in world's gas supply. For shale reservoir, the production performance is influenced by many factors such as geology, drilling, completion. A production model that contains a comprehensive set of variables is required for production prediction and optimization.

Decline curve analysis (DCA) has been wildly adopted for production forecast and it is a fast method by matching production rate-time history data. Traditional DCA models [1] is designed for conventional reservoir and assume a boundary-dominated flow, which is not applicable for shale wells with multiple flow regimes. Many DCA models have been specifically designed for shale gas reservoirs [2–7]. However, DCA models still have some drawbacks as it requires a long production history to get a reliable result, which means they are mainly applied to on-production wells [8]. In addition, it is impossible to accommodate

additional geological and completion information to improve prediction accuracy, which means it is difficult to optimize production.

Reservoir numerical simulation is another approach to forecast production performance of unconventional reservoirs. It is a physically driven model and can provide accurate results when the data is complete and precise. However, it is difficult and expensive to gather the necessary data such as the complex hydraulic fracture distribution and properties. Moreover, the gas flow system in shale formation is poorly understood [9] and simulation techniques requires large amounts of computational time [8].

The emerging machine learning technique has provided a potential method for production modeling as a result of advanced computing powers and access to large data set. Liao et al. [10] used random forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM) to build a stacking model and identified that stimulated length, total stage count, pumped proppant per stage, pumped fluid per length and injection rate are the most important factors for Wapiti-Montney tight gas formation. However, the model did not include reservoir parameters such as total organic carbon content (TOC). Hirschmiller et al. [11] used recursive feature elimination to select features and used random forest to predict and optimize well performance. Sheikhi et al. [12] used linear regression, Random forest, Gradient Boost, XGBoost, Bagging, ExtraTrees and neural network to build models and applied Individual Conditional Expectation and Partial Dependency to assess completion performance, but the input variables selected for the model were limited to completion information. Shelley et al. [13] built a feed forward neural network model to estimate Wolfcamp production and evaluate the economics for completion designs. Han et al. [14] used deep neural network and exploratory data analysis to build robust model for production prediction and provide some insights about shale reservoirs. However, the input variables were limited to completion factors. Wang et al. [15] used random forest, adaptive boosting, support vector machine and neural network to estimate the well performance. It was concluded that random forest has the best performance and these models were useful for designing hydraulic fracture treatments. Liang et al. [16] used multi-objective random forest to predict dynamic production data. However, there remains a challenge to choose the right method for production prediction. In general, multiple linear regression can only describe linear relationship. Regression tree requires a lot of data to perform well. The construction of neural network is time-consuming and tedious [15]. Compared with neural work, regression tree and multiple linear regression, support vector regression and gaussian process regression have a better performance for small data sets [17,18].

The originality of this paper is that geological and completion data are coupled to more accurately describe the reservoir property. In addition, feature selection is used to provide a variety of geologic and completion parameters specifically related to production.

In this study, a workflow based on data driven approach was proposed for production modeling. A large data set (including geological and completion factors) in Duvernay formation was used to illustrate the application of workflow. Grey relation analysis and Pearson Correlations were applied to screen the geological and engineering parameters related to shale gas productivity. Gaussian process regression, support vector regression, and multiple linear regression were applied for production predicting. A tornado plot was used to identify the most important factors for production performance.

2. Methodology

Figure 1 is a flowchart illustrating the procedures of the study, Pearson correlation degree and grey relation degree is calculated to select features. Gaussian process regression, support vector regression, and multiple linear regression were applied to model cumulative production. After training and testing the model, the model with the best performance is selected to applied sensitivity analysis.

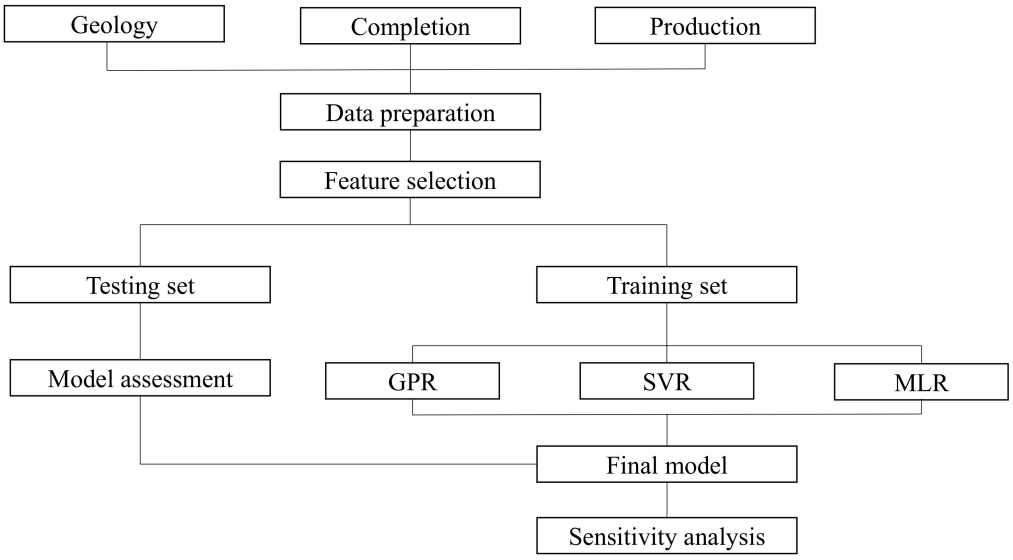

Figure 1. Flowchart of the research progress.

2.1. Pearson Correlations

The Pearson correlation degree (PCD) is one of the commonly used relation measurement standards, it can measure the monotonic relationship between different variables [19]. The Pearson-correlation formula is as follows:

$$\rho = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} \quad (1)$$

where $Cov(x,y)$ denotes covariance between x and y, $Var(x)$ is the variance of x, and $Var(y)$ represents the variance of y.

PCD varies from -1 to 1. The value of -1 or 1 shows that there is a strong linear relationship between two variables, whereas 0 means no linear relationship. Normally, the relationship strength of variables is estimated on the basis of following value ranges [20], i.e., 0.8–1.0: very strong correlation; 0.6–0.8, strong correlation; 0.4–0.6, moderate correlation; 0.2–0.4, weak correlation; 0.0–0.2 very weak correlation.

2.2. Grey Relation Analysis

The grey relation analysis, which was originally proposed by Deng et al. [21], provides a multi-factor analysis method which describes the posture relationship among factors. In a sample data, the grey relation degree (GRD) represents the change trend (speed, size, direction), A high GRD between two factors means a strong correlation. On the contrary, a small value means a weak correlation [22,23]. The advantage of grey correlation analysis is that it requires less data and calculation compared to other multi-factor analysis methods [24] (random forest, regression, etc.). The basic steps of grey relation analysis are as follows:

2.2.1. Determine the Reference Series and Comparison Series

Reference series is a data series which reflect the system behavior. Comparison series is a data series composed of variables affecting system behavior. $X_{0'}$ represents reference

series, $X_{i'}(i = 1, 2, \ldots m)$ denote comparison series. The whole data series are represented as matrix (2):

$$(X_{0'}, X_{1'}, \ldots, X_{m'}) = \begin{bmatrix} x_{0'}(1) & x_{1'}(1) & \ldots & x_{m'}(1) \\ x_{0'}(2) & x_{1'}(2) & \ldots & x_{m'}(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_{0'}(n) & x_{1'}(n) & \ldots & x_{m'}(n) \end{bmatrix}_{n \times (m+1)} \quad (2)$$

where $X_{i\prime} = (x_{i'}(1), x_{i'}(2), \ldots, x_{i'}(n))^T$, I = 0,1,2, … , m; n is the length of variable series.

2.2.2. Dimensionless Processing

In general, each variable series has a specific physical meaning and dimensions or order of magnitude vary from variable to variable. Therefore, it's important to process raw data before using, the following formula is used for dimensionless:

$$x_i(n) = \frac{x_{i'}(n) - \min(X_{i'})}{\max(X_{i'}) - \min(X_{i'})} \quad (3)$$

After dimensionless processing, the data matrix can be represented as matrix (4):

$$(X_0, X_1, \ldots, X_m) = \begin{bmatrix} x_0(1) & x_1(1) & \ldots & x_m(1) \\ x_0(2) & x_1(2) & \ldots & x_m(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_0(n) & x_1(n) & \ldots & x_m(n) \end{bmatrix}_{n \times (m+1)} \quad (4)$$

2.2.3. Calculate Grey Relation Degree

Compute the absolute value between reference series and comparison series and get absolute difference matrix (5), then transform matrix (5) into correlation coefficient matrix (7) using formula (6), finally, use Equation (8) to calculate grey relation degree.

$$\begin{bmatrix} \Delta_{01}(1) & \Delta_{02}(1) & \ldots & \Delta_{0m}(1) \\ \Delta_{01}(2) & \Delta_{02}(2) & \ldots & \Delta_{0m}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{01}(n) & \Delta_{02}(n) & \ldots & \Delta_{0m}(n) \end{bmatrix}_{n \times m} \quad (5)$$

$$\xi_{0i}(k) = \frac{\Delta_{min} + \rho\Delta_{max}}{\Delta_{0i}(k) + \rho\Delta_{max}} \quad (6)$$

$$\begin{bmatrix} \xi_{01}(1) & \xi_{02}(1) & \ldots & \xi_{0m}(1) \\ \xi_{01}(2) & \xi_{02}(2) & \ldots & \xi_{0m}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{01}(n) & \xi_{02}(n) & \ldots & \xi_{0m}(n) \end{bmatrix}_{n \times m} \quad (7)$$

$$r_{0i} = \frac{1}{N} \sum_{k=1}^{N} \xi_{0i}(k) \quad (8)$$

2.3. Multiple Linear Regression (MLR)

Multiple linear regression describes the relationship between several independent variables, *X* and a dependent variable, *Y*. *Y* is often called response variable and independent variables X are named predictor variables. Equation (9) is the universal form of a multiple linear regression model.

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \quad (9)$$

where p is the number of predictor variables, ϵ is the noise terms, it is assumed that noise terms are uncorrelated and have independent and identical normal distributions with mean zero and constant variance [19,25,26]. The noise terms represent some variables that not include but may have contribution to Y.

2.4. Support Vector Regression (SVR)

Support vector machine (SVM) is a supervised learning method, which was originally identified by Cortes et al. [27]. SVM has been comprehensively used as a robust technique for classification and SVR is a regression technique used from SVM. SVR is regarded as a nonparametric method as it relies on kernel functions. Kernel function can project the data to a higher dimensional feature space, hence avoiding the non-linearity in lower space. The aim of SVR is to find a function $f(x)$ that diverge from response y by a value no more than ϵ, and in the meantime the function $f(x)$ as even as possible (Figure 2). More details about SVR are described in Al-Azani et al. [28], Da Silva et al. [29], El-Sebakhy et al. [30], Li et al. [31] and Schuetter et al. [26].

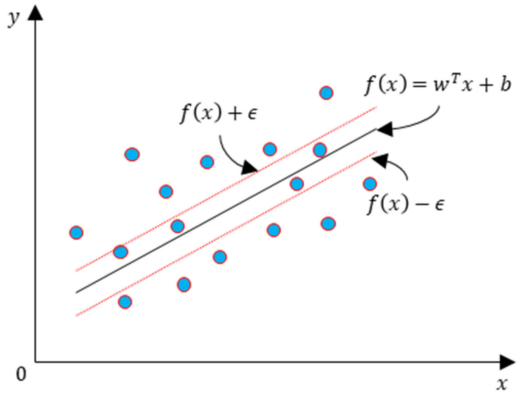

Figure 2. Support vector machine regression diagram.

In SVR, the choice of kernel function directly determines the performance of the model, the main kernel functions used in this study are shown as follows:

Linear:

$$G(x_j, x_k) = x_j' x_k \tag{10}$$

Gaussian:

$$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right) \tag{11}$$

Polynomial:

$$G(x_j, x_k) = \left(1 + x_j' x_k\right)^q \tag{12}$$

2.5. Gaussian Process Regression (GPR)

Gaussian progress regression (GPR) is a non-parametric model. A Gaussian process (GP) refers to a collection of random variables, any finite number of random variables in this collection has a joint gaussian distribution [32,33]. GP has quite good adaptability for dealing with some complex problems such as small samples, high dimensionality, and nonlinearity. GP is determined by its mean function and covariance function, and it inherits many properties of the Gaussian distribution [34,35]. A finite dimensional subset of a Gaussian process obeys a Gaussian distribution.

Given a set of data:

$$D = \{(x_i, y_i), \mathrm{i} = 1, 2, \ldots, \mathrm{n}\}, x_i \in R^d, y_i \in R \tag{13}$$

Mean function is defined as:

$$m(x) = E[f(x)] \tag{14}$$

Covariance function is defined as:

$$k(x, x') = E\left[(f(x) - m(x))(f(x') - m(x'))\right] \tag{15}$$

where: $x, x' \in R^d$, the GP can be defined as:

$$f(x) \sim GP\left[m(x), k(x, x')\right] \tag{16}$$

As for regression question, the model can be defined as:

$$y = f(x) + \varepsilon \tag{17}$$

Assuming the noise $\varepsilon \sim$ N $(0, \sigma_y^2)$, then the prior probability distribution of the observation value y is obtained:

$$y \sim N\left[0, K(X, X) + \sigma_y^2 I_n\right] \tag{18}$$

y and $f(x^*)$ have a joint Gaussian distribution:

$$\begin{bmatrix} y \\ f(x^*) \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X, X) + \sigma_y^2 I_n & K(X, x^*) \\ K(x^*, X) & K(x^*, x^*) \end{bmatrix}\right) \tag{19}$$

where: $K(X, X) = K_n = K_{ij}$, it is a $n * n$ positive definite matrix, the matrix elements $K_{ij} = K(x_i, x_j)$ describe the correlation between x_i and x_j, $K(X, x^*) = K(x^*, X)^{-1}$ is used to measure the correlation between data X and x^*, $K(x^*, x^*)$ is used to describe the correlation between data x^*.

Then the posterior probability distribution of $f(x^*)$ can be expressed as:

$$P(f(x^*)|x^*, X, y) \sim N(\mu^*, \Sigma^*) \tag{20}$$

where:

$$\mu^* = K(X, x^*)\left[K(X, X) + \sigma_y^2 I_n\right]^{-1} y \tag{21}$$

$$\Sigma^* = K(x^*, x^*) - K(X, x^*)\left[K(X, X) + \sigma_y^2 I_n\right]^{-1} K(x^*, X) \tag{22}$$

Σ^* and μ^* represent the covariance and mean of $f(x^*)$.

Kernel function controls the accuracy of GP as GP is parameterized by a mean function and a kernel function. The main kernel functions used in this study are shown as follows:

Squared Exponential Kernel:

$$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2}\frac{(x_i - x_j)^T (x_i - x_j)}{\sigma_l^2}\right] \tag{23}$$

where σ_l is the characteristic length scale and σ_l is the signal deviation.

Exponential Kernel

$$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right) \tag{24}$$

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{25}$$

where σ_l is the characteristic length scale and r is the Euclidean distance between x_i and x_j.

Mater 5/2:

$$k(x_i, x_j) = \sigma_f^2 \left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right) \exp\left(-\frac{\sqrt{5}r}{\sigma_l}\right) \tag{26}$$

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{27}$$

where r is the Euclidean distance between x_i and x_j.

Rational Quadratic Kernel:

$$k(x_i, x_j|\theta) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha} \tag{28}$$

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{29}$$

where σ_l is the characteristic length scale, r is the Euclidean distance between x_i and x_j and α is a positive-valued scale-mixture parameter.

2.6. Goodness-of-Fit Metrics

There are many measures for evaluating the quality of a model. In this study, two techniques are used: root mean squared errors (RMSE) and R-squared value (Note that the calculation of R-squared is not always as easy as squaring the Pearson correlation, as R-squared can be smaller than 0. However, in linear regression, the square of the Pearson correlation is equal to R-squared). R-squared value indicates the part of variability in the response which is interpreted by the model, it is defined as the ratio of explained sum of squares (ESS) to total sum of squares (TSS):

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \tag{30}$$

RMSE can be interpreted as the average distance between the residuals and zero.

$$RMSE(X, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2} \tag{31}$$

3. Problem Description

The Duvernay Formation, an Upper Devonian source rock, lies within Western Canada Basin in Alberta, Canada. It is an organic-rich mudstones surrounded by Leduc reef complexes [36]. Because of high formation pressure coefficient (1.7–2.0) and significant degree of undersaturation, Duvernay formation is a tight liquid-rich reservoir. The phases of hydrocarbon liquids vary from wet gas in southwest part of the reservoir to black oil in the northeast portion due to various thermal maturity [37–40]. The depth of Duvernay is 3000–4150 m deep, and the buried depth rises from southwest to northeast. The study area is located in west shale basin, with an area of around 737.92 km^2. The basic reservoir parameters in study area are described in Table 1.

Table 1. Duvernay reservoir characteristics.

Reservoir Characteristics	Value Range	Units
Depth	3000–4150	m
Thickness	20–56	m
Total organic Carbon Content (TOC)	2–6	%
Porosity	1–8	%
Vitrinite Reflectance (Ro)	1.1–1.6	%
Clay minerals	10.9–56	%
Brittle minerals	44–89.1	%

The difference of well performance can be attributed to two factors, geological factors and completion factors. Usually, geological factors include contents of brittle minerals, formation pressure, porosity, permeability, reservoir thickness, TOC, thermal maturity, natural fractures, gas content, condensate gas ratio, etc. Completion factors include total horizontal lateral length, total fluid amount, total proppant amount, number of stages, number of clusters, sand contents, cluster distance, stage distance, etc.

Currently, there are 159 horizontal wells in the Duvernay Formation in the study area, the whole wells have a production time of more than one year, 70 wells have a production time of more than 5 years. In the study area, PNP (Plug & perf) segmented completion technology is adopted, high-viscosity fracturing fluid and small particle size quartz sand proppant are used in the hydraulic fracturing process, the clusters for each stage vary from 4 to 10 and the pump rate per stage varies from 10 m^3/min to 14 m^3/min. In the past 8 years, the stage spacing has decreased from 90 m to 50 m and the proppant intensity has increased from 1 t/m to 4 t/m. In this study, 11 variables contained geological information and completion parameters are collected from 159 shale gas wells in Duvernay Formation according the data availability and statistical efficiency. As the hydrocarbon phase in study area is gas condensate, the responses are divided into condensate production and gas production. A list of the whole variables is illustrated in Table 2.

Table 2. List of variables in the study area.

Type	Variable	Description
Response	M6COND	Cumulative oil production within 1st 6 producing months (t)
	M6GAS	Cumulative gas production within 1st 6 producing months (10^4 m^3)
Predictor	Fluid	Total fluid amount for hydraulic fracturing (m^3)
	PROP	Total proppant amount (t)
	Clusters	Number of frac clusters
	Stages	Number of frac stages
	Lateral Length	Total horizontal lateral length (m)
	Thickness	Formation total thickness (m)
	POR	Porosity (%)
	PERM	Permeability (mD)
	Sg	Gas saturation (%)
	TOC	Total organic Carbon Content (%)
	CGR	Condensate gas ratio (t/10^4 m^3)

4. Feature Selection

In grey relation analysis, two responses M6COND and M6GAS are defined as the reference series, and the comparison series are defined as 11 geological and completion predictors. Normally, the smaller the grey relation degree (GRD), the higher the difference is between the comparison series and the reference series. Table 3 shows GRD and rank between responses and predictors, and the influencing parameters with GRD bigger than 0.69 are selected. The geological factors include CGR, TOC, Sg, the completion factors contain Fluid, PROP, Clusters, Stages, Lateral length.

Table 3. The correlation between responses and predictors.

Rank	The Correlation between M6GAS and Input Feature		The Correlation between M6COND and Input Feature	
	Indicator	Coefficient of Association	Indicator	Coefficient of Association
1	PROP	0.791	Stages	0.780
2	Stages	0.791	PROP	0.776
3	Clusters	0.767	Fluid	0.763
4	Lateral Length	0.744	Lateral Length	0.759
5	Fluid	0.735	Clusters	0.749
6	CGR	0.697	TOC	0.708
7	TOC	0.694	Sg	0.690
8	Sg	0.691	CGR	0.690
9	Thickness	0.656	Thickness	0.679
10	POR	0.640	POR	0.621
11	PERM	0.628	PERM	0.615

Figure 3 illustrates Pearson correlation degree (PCD) between responses and inputs, PCD bigger than 0.2 are selected. The result is the same as grey relation analysis, PCD between Thickness, POR, PERM and two responses are all less than 0.2, which means there is a weak correlation between them. Therefore, Thickness, POR and PERM are eliminated from the dataset.

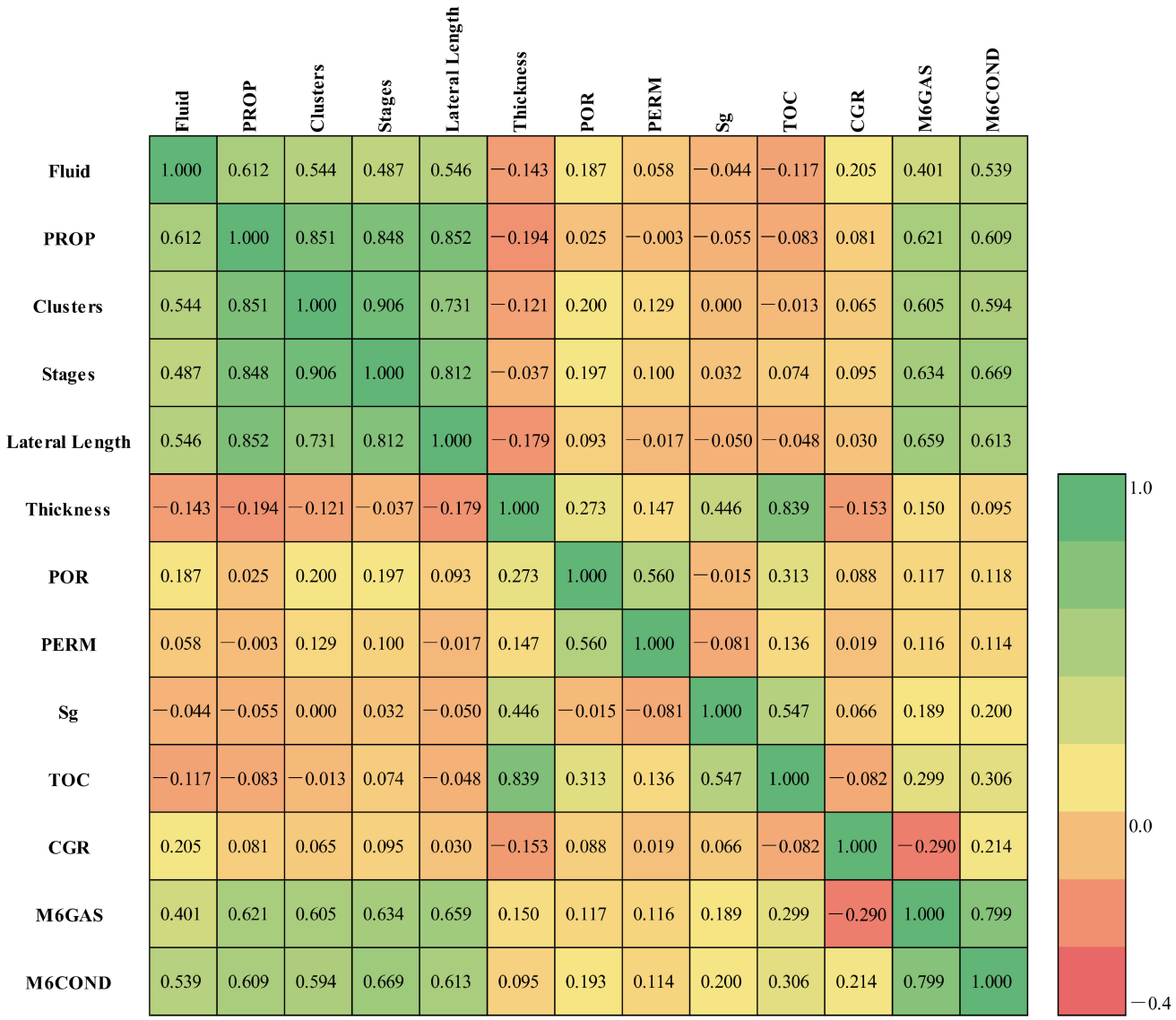

Figure 3. Correlation matrix (Green: positive correlation, Red: negative correlation, the darker the color, the stronger the correlation).

Finally, combined with Pearson correlations and grey relation analysis method, predictors selected as key factors to build statistical models are as follows: Fluid, PROP, Clusters, Stages, Lateral Length, Sg, TOC, CGR. The histogram plots of all selected variables are demonstrated in Figure 4. Table 4 shows the results of Kolmogorov-Smirnov (K-S) test. The results illustrate that the significance of all variables (except for Lateral Length) is less than 0.05, meaning that Fluid, PROP, Clusters, Stages, Sg, TOC, CGR are not normally distributed. Therefore, a normal score method (a method which can return a normal distribution dataset) is applied to transform those non-normal distribution variables.

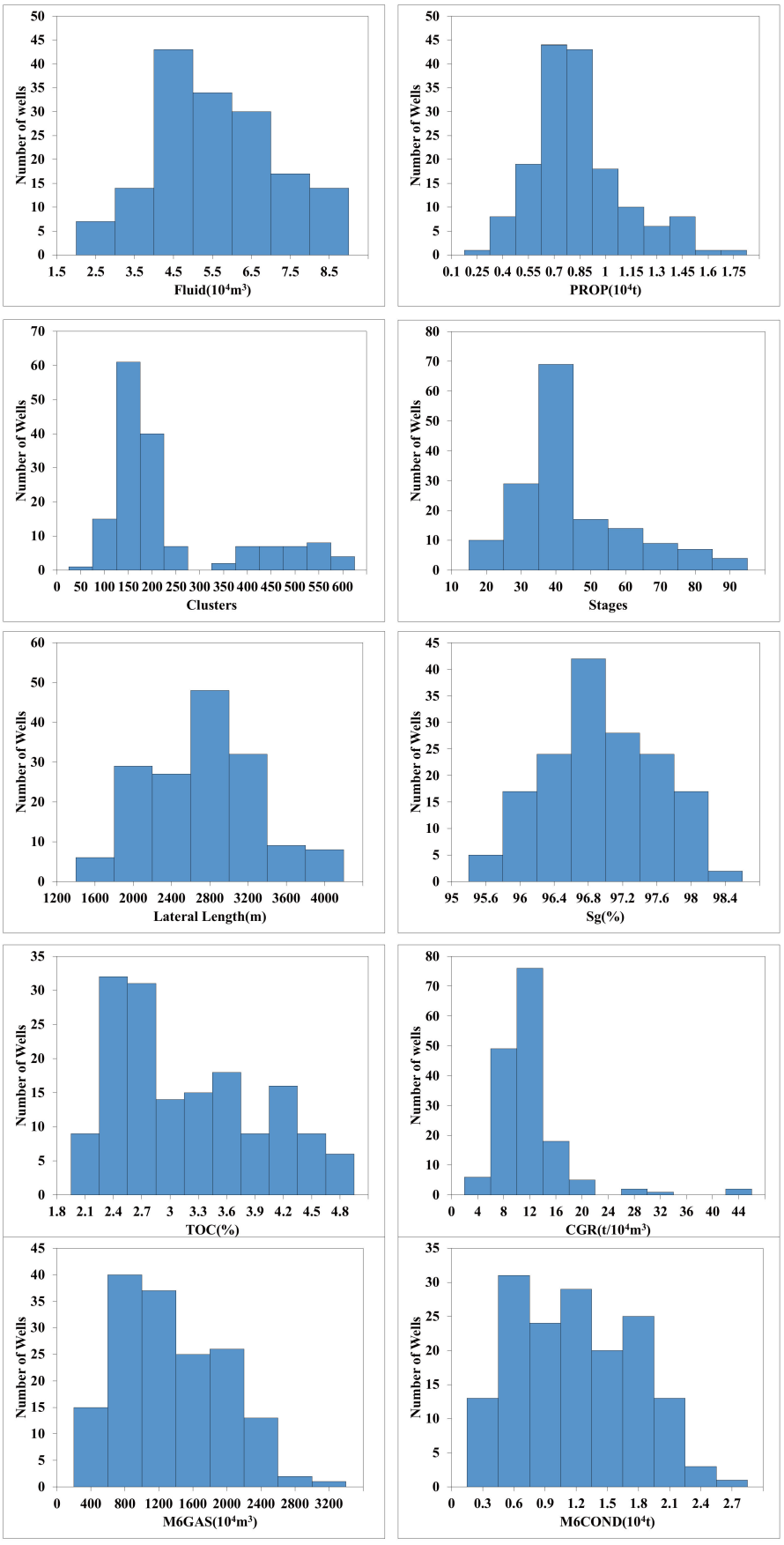

Figure 4. Histogram plots of all variables.

Table 4. The results of Kolmogorov-Smirnov Test.

	Fluid	PROP	Clusters	Stages	Lateral Length	Sg	TOC	CGR
Statistic	0.089	0.131	0.290	0.195	0.056	0.075	0.139	0.182
Sig.	0.004	0.000	0.000	0.000	0.200	0.030	0.000	0.000

5. Model Development

In this study, Matlab 2019 was used to build the model. To avoid overfitting, the total data was separated into a training set, and a testing set. A total of 129 samples were used as the training set and the remaining 30 samples were used to test the effectiveness of three machine learning models. Training set are performed a 4-fold cross-validation. In this method, the dataset is partitioned into 4 disjoint sets or folds. For each fold, the model is trained using the out-of-fold data and model performance is assessed using in-fold data. Finally, calculate the average test error over all folds. This method performs an excellent estimate of the predictive accuracy and makes full use of all the training data, it works quite well for small data sets. The most difficult part for GPR model and SVR model is parameter optimization. To solve this problem, a Bayesian optimization was applied. Bayesian optimization internally maintains a Gaussian process model of the objective function, and uses objective function evaluations to train the model. After optimization, the hyperparameters of SVR and GPR were demonstrated in Tables 5 and 6:

Table 5. Hyperparameters for SVR model.

	Kernel Function	Kernel Scale	Box Constraint	Epsilon-Insensitive Band
M6GAS	Gaussian	2.4	747.96	74.8
M6COND	Gaussian	2.8	7620	762.07

Table 6. Hyperparameters for GPR model.

	Kernel Function	Kernel Scale	Sigma
M6GAS	Exponential	5346.3319	0.56
M6COND	Exponential	3653.12	0.0001

Tables 7 and 8 illustrate the model performance for the prediction of M6GAS and M6COND respectively. The model performance for training set and testing set is compared in Figure 5. As for M6GAS prediction, the root mean squared errors (RMSE) of the testing set are found to be 280.54 $\times$ 10^4 m^3 for GPR, 366.25 $\times$ 10^4 m^3 for SVR, 377.72 $\times$ 10^4 m^3 for MLR. GPR model shows the highest R^2 and the lowest RMSE in both training set and testing set. Especially in testing set, GPR model shows R^2 at 0.8 and RMSE at 280.54 $\times$ 10^4 m^3, it means that the model is able to explain 80% of the gas production variance in the study area and on average there is around 280.54 $\times$ 10^4 m^3 uncertainty in the prediction of gas production within 1st 6 producing months for each well. It is found that GPR model shows the best performance for forecasting the gas production. It should be noted that SVR method performs a comparable prediction accuracy in training set compared with GPR. However, the differences of RMSE and R^2 between training set and testing set are very large, indicating that SVR undergoes overfitting problems.

Table 7. Model results for M6GAS.

	Training			Testing		
	GPR	SVR	MLR	GPR	SVR	MLR
R^2	0.81	0.78	0.66	0.8	0.67	0.65
RMSE (10^4 m^3)	277.73	296.04	371.27	280.54	366.25	377.72

Table 8. Model results for M6COND.

	Training			Testing		
	GPR	**SVR**	**MLR**	**GPR**	**SVR**	**MLR**
R^2	0.78	0.73	0.59	0.83	0.8	0.67
RMSE (t)	2246.2	2477.3	3936.9	1884.3	1903.7	2544.5

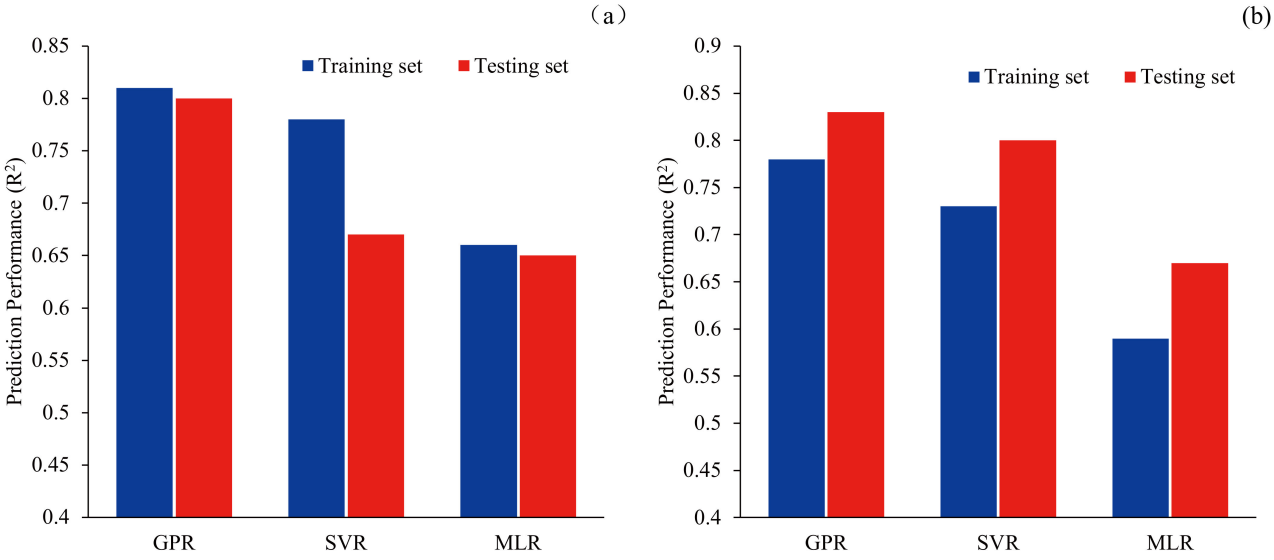

Figure 5. Comparison of prediction performance of the GPR, SVR, and MLR algorithms: (**a**)prediction performance about M6GAS (**b**) prediction performance about M6COND.

As for M6COND prediction, the result is the same as M6GAS prediction, the root mean squared errors (RMSE) of the testing set are found to be 1884.3 t for GPR, 1903.7 t for SVR, 2544.5 t for MLR. GPR model demonstrates the highest R^2 at 0.83 and the lowest RMSE at 1884.3 t in the testing set, it also means that on average there is around 1884.3 t uncertainty in the prediction of M6COND for each well and the model can illustrate 83% of the oil production variance in Duvernay formation. In addition, as for GPR model, the differences of RMSE and R^2 between training set and testing set are relatively small, indicating that GPR doesn't undergo overfitting or underfitting problems.

Considering that the R-squared value of most machine learning based model varied from 0.5–0.8 (Luo et al. [41], Kong et al. [9], Wang et al. [15]), the GPR models built in this study are able to provide relative high accuracy for production prediction. Compared with MLR and SVR, GPR has better performance for the prediction of M6COND and M6GAS. Therefore, GPR method is selected to apply sensitivity analysis.

6. Sensitivity Analysis

Figure 6 is a Tornado Plot that describes the sensitivity of 8 geological and completion factors on M6GAS. In the process of sensitivity analysis, the factors are set based on the statistical distribution of input features used in machine learning model, which is shown in Figure 6. The base case used the average value of each input feature. The sensitivity values are calculated by setting each interested factor to its maximum and minimum values one-factor-at-a-time as shown in Table 9, while the other factors are set to their base values. The red bars in Figure 6 equal to a sensitivity measure (compared with a base example) when a factor is designed to its maximum value and the green bars in Figure 6 equal to a sensitivity measure (compared with a base example) when a factor is designed to its minimum values. It can be observed from Figure 6 that Fluid, Stages and TOC are regarded as the most significant factors. It is followed by CGR, Clusters and Lateral Length. Sg and

PROP shows the least effect on M6GAS. It can be inferred that these least important factors could have an unimportant important on M6GAS.

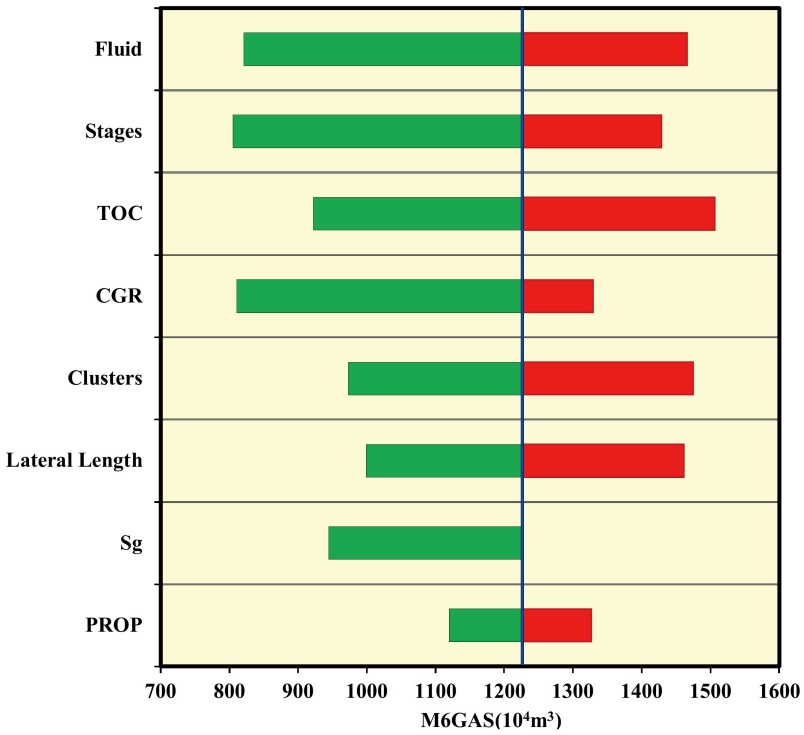

Figure 6. Tornado plot illustrating the most and the least important factors for M6GAS. Each red bar equals to a sensitivity measure (compared with a base example) when a factor is designed to its upper bound. The green bars indicate the lower-bound case. The blue line in the middle means the base case. Factors are arranged by descending order from top to bottom.

Table 9. Summary of geological and engineering factors for sensitivity analysis.

Type	Factor	Unit	Range	Base Case
Geology	Sg		95%, 96%, 97%, 98%	97%
	TOC		2%, 2.5%, 3%, 3.5%, 4%	3%
	CGR	t/10^4 m^3	2, 6, 10, 14, 18	10
Completion	Fluid	m^3	20,000, 30,000, 40,000, 50,000, 60,000	40,000
	PROP	t	2000, 3500, 5000, 6500, 8000	5000
	Clusters		80, 140, 200, 260, 320	200
	Stages		20, 35, 50, 65, 80	50
	Lateral Length	m	1600, 2000, 2400, 2800, 3200	2400

Figure 7 is a Tornado Plot that indicates the sensitivity of 8 geological and completion factors on M6COND. It is clear that CGR, Fluid and TOC are the most important factors. The following factors are Stages, Clusters, and Sg. The effect of Sg and PROP is marginal. In the study area, the production of condensate leads to higher economic efficiency, so it is significant to drill wells in high CGR areas as CGR plays the most important role in condensate production. Furthermore, a more progressive completion treatment (such more fluid volume) is required to get more production.

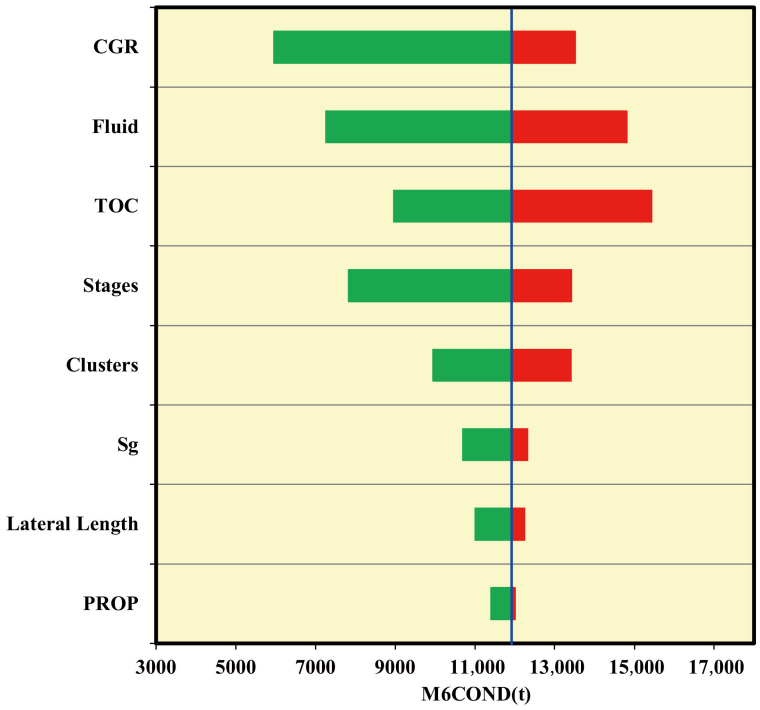

Figure 7. Tornado plot illustrating the most and the least important factors for M6COND. Each red bar equals to a sensitivity measure (compared with a base example) when a factor is designed to its upper bound. The green bars indicate the lower-bound case. The blue line in the middle means the base case. Factors are arranged by descending order from top to bottom.

It is significant to note that this model may provide some suggestions for the study area in Duvernay Formation, it is not applicable to apply this model in a completely different reservoir. Moreover, the model doesn't take parent and child well interference into consideration, which may limit the model prediction accuracy for new offset wells.

7. Conclusions

In this study, grey relation analysis and Pearson Correlation analysis were applied to demonstrate how to select significant geologic and completion factors. Then, MLR, SVR, and GPR were used to predict cumulative production with selected factors. At last, the best performance model was used to conduct sensitivity analysis.

Based on data-driven methodology presented in this paper, it is concluded that:

Based on grey relation analysis and Pearson Correlation analysis, 8 parameters are selected as input in predictive model, they are Fluid, PROP, Clusters, Stages, Lateral Length, Sg, TOC, CGR.

GPR model has the best performance among three predictive models, it results in highest R-squared and lowest RMSE. In the testing set, the model shows a R-squared of 0.8 with a RMSE of 280.54 $\times$ 10^4 m^3 in predicting M6GAS and gives an R-squared of 0.83 with a RMSE of 1884.3 t in predicting M6COND.

Using GPR model, sensitivity analysis indicates that Fluid, Stages and TOC are the most important features for M6GAS. As for M6COND, CGR, Fluid, and TOC are the most important features.

The approach includes feature selection, model development, and sensitivity analysis. The results and technique may provide some advice in making development decisions in Duvernay formation. The method may apply to different shale reservoirs

Author Contributions: Data curation, L.S. and Y.J.; supervision, H.W.; validation, X.K.; writing–review and editing, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Major Project of China, grant number 2016ZX05029-005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Cov(x,y)	Covariance between x and y
Var(x)	The variance of x
GRD	grey relation degree
PCD	Pearson correlation degree
MLR	Multiple linear regression
SVR	Support Vector Regression
GPR	Gaussian progress regression
DCA	Decline curve analysis
RMSE	Root mean squared errors
M6COND	Cumulative oil production within 1st 6 producing months
M6GAS	Cumulative gas production within 1st 6 producing months
PROP	Total proppant amount
POR	Porosity (%)
PERM	Permeability (mD)
Sg	Gas saturation (%)
TOC	Total organic Carbon Content (%)
CGR	Condensate gas ratio ($t/10^4$ m^3)

References

1. Arps, J.J. Analysis of Decline Curves. *Trans. AIME* **1945**, *160*, 228–247. [CrossRef]
2. Clark, A.J.; Lake, L.W.; Patzek, T.W. Production Forecasting with Logistic Growth Models. In Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, CO, USA, 30 October–2 November 2011.
3. Clarkson, C.R.; Qanbari, F.; Williams-kocacs, J.D. Innovative use of rate-transient analysis methods to obtain hydraulic-fracture properties for low-permeability reservoirs exhibiting multiphase flow. *Lead. Edge* **2014**, *33*, 1108–1122. [CrossRef]
4. Duong, A.N. Rate-Decline Analysis for Fracture-Dominated Shale Reservoirs. *SPE Reserv. Eval. Eng.* **2011**, *14*, 377–387. [CrossRef]
5. Ilk, D.; Rushing, J.A.; Perego, A.D.; Blasingame, T.A. Exponential vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves. In Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, CO, USA, 21–24 September 2008.
6. Paryani, M.; Ahmadi, M.; Awoleke, O.; Hanks, C. Using Improved Decline Curve Models for Production Forecasts in Unconventional Reservoirs. In Proceedings of the SPE Eastern Regional Meeting, Canton, OH, USA, 2016.
7. Seshadri, J.N.; Mattar, L. Comparison of Power Law and Modified Hyperbolic Decline Methods. In Proceedings of the Canadian Unconventional Resources and International Petroleum Conference, Calgary, AB, Canada, 19–21 October 2010.
8. Wang, S.; Qin, C.; Feng, Q.; Javadpour, F.; Rui, Z. A framework for predicting the production performance of unconventional resources using deep learning. *Appl. Energy* **2021**, *295*, 117016. [CrossRef]
9. Kong, B.; Chen, Z.; Chen, S.; Qin, T. Machine learning-assisted production data analysis in liquid-rich Duvernay Formation. *J. Pet. Sci. Eng.* **2021**, *200*, 108377. [CrossRef]
10. Liao, L.; Li, G.; Zhang, H.; Feng, J.; Zeng, Y.; Ke, K.; Wang, Z. Well Completion Optimization in Canada Tight Gas Fields Using Ensemble Machine Learning. In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, United Arab Emirates, 9–12 November 2020.
11. Hirschmiller, J.; Biryukov, A.; Groulx, B.; Emmerson, B.; Quinell, S. The Importance of Integrating Subsurface Disciplines with Machine Learning when Predicting and Optimizing Well Performance–Case Study from the Spirit River Formation. In Proceedings of the SPE Annual Technical Conference and Exhibition, Calgary, AB, Canada, 30 September–2 October 2019.
12. Sheikhi Garjan, Y.; Ghaneezabadi, M. Machine Learning Interpretability Application to Optimize Well Completion in Montney. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, USA, 28 September–2 October 2020.
13. Shelley, R.; Oduba, O.; Melcher, H. Machine Learning and Artificial Intelligence Provides Wolfcamp Completion Design Insight. In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, 4–6 May 2021. Online.

14. Han, D.; Kwon, S.; Kim, J.; Jin, W.; Son, H. Comprehensive Analysis for Production Prediction of Hydraulic Fractured Shale Reservoirs Using Proxy Model Based on Deep Neural Network. In Proceedings of the SPE Annual Technical Conference and Exhibition, 26–29 October 2020. Online.
15. Wang, S.; Chen, S. Engineering. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *J. Pet. Sci. Eng.* **2019**, *174*, 682–695. [CrossRef]
16. Xue, L.; Liu, Y.; Xiong, Y.; Liu, Y.; Cui, X.; Lei, G. A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *J. Pet. Sci. Eng.* **2021**, *196*, 107801. [CrossRef]
17. He, Z.K.; Liu, G.B.; Zhao, X.J.; Wang, M.H. Overview of Gaussian process regreession. *Control. Decis.* **2013**, *28*, 1121–1129.
18. Ding, S.F.; Qi, B.J.; Tan, H.Y. An Overview on Theory and Algorithm of Support Vector Machines. *J. Univ. Electron. Sci. Technol. China* **2011**, *40*, 2–10.
19. Zhou, Q.; Dilmore, R.; Kleitre, A.; Dilmore, R. Evaluating Gas Production Performances in Marcellus Using Data Mining Technologies. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, Denver, CO, USA, 25–27 August 2014.
20. Niu, W.; Lu, J.; Sun, Y. A Production Prediction Method for Shale Gas Wells Based on Multiple Regression. *Energies* **2021**, *14*, 1461. [CrossRef]
21. Deng, J. Introduction to grey system theory. *J. Grey Syst.* **1989**, *1*, 1–24.
22. Feng, Y.; Ji, B.; Gao, P.; Li, Y. An Improved Grey Relation Analysis Method and Its Application in Dynamic Description for a Polymer Flooding Pilot of Xingshugang Field, Daqing. In Proceedings of the North Africa Technical Conference and Exhibition, Cairo, Egypt, 14–17 February 2010.
23. Yue, X.; Bing, H.; Xing, L.; Bowen, C. Research on Main Control Factors Influencing Fracturing Effect of Jiaoshiba Area Based on Grey Relational Analysis. In Proceedings of the ARMA-CUPB Geothermal International Conference, Beijing, China, 5–8 August 2019.
24. Ma, K.; Jiang, H.; Li, J.; Zhang, R.; Zhang, L.; Fang, W.; Shen, K.; Dong, R. A Novel Early Warning System of Oil Production Based on Machine Learning. In Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, United Arab Emirates, 11–14 November 2019.
25. Lolon, E.; Hamidieh, K.; Weijers, L.; Mayerhofer, M.; Melcher, H.; Oduba, O. Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History. In Proceedings of the SPE Hydraulic Fracturing Technology Conference, The Woodlands, TX, USA, 9–11 February 2016.
26. Scheuetter, J.; Mishra, S.; Zhong, M.; LaFollette, R. Data Analytics for Production Optimization in Unconventional Reservoirs. In Proceedings of the SPE/AAPG/SEG Unconventional Resources Technology Conference, San Antonio, TX, USA, 20–22 July 2015.
27. Cortes, C.; Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
28. Al-Azani, K.; Elkatatny, S.; Abdulraheem, A.; Mahmoud, M.; Ali, A. Prediction of Cutting Concentration in Horizontal and Deviated Wells Using Support Vector Machine. In Proceedings of the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 23–26 April 2018.
29. Da Silva, L.M.; Avansi, G.D.; Schiozer, D.J. Support Vector Regression for Petroleum Reservoir Production Forecast Considering Geostatistical Realizations. *SPE Reserv. Eval. Eng.* **2020**, *23*, 1343–1357. [CrossRef]
30. El-Sebakhy, E.A.; Sheltami, T.; Al-Bokhitan, S.Y.; Shaaban, Y.; Raharja, P.D.; Khaeruzzaman, Y. Support Vector Machines Framework for Predicting the PVT Properties of Crude Oil Systems. In Proceedings of the SPE Middle East Oil and Gas Show and Conference, Manama, Bahrain, 11–14 March 2007.
31. Li, X.; Miskimins, J.L.; Sutton, R.P.; Hoffman, B.T. Multiphase Flow Pattern Recognition in Horizontal and Upward Gas-Liquid Flow Using Support Vector Machine Models. In Proceedings of the SPE Annual Technical Conference and Exhibition, Amsterdam, The Netherlands, 27–29 October 2014.
32. Habu, I.; Ioki, T.; Okuyama, E. Study on Prediction Method of Propulsive Performance by Means of Gaussian Process Regression for Onboard Monitoring Data. In Proceedings of the 30th International Ocean and Polar Engineering Conference, Shanghai, China, 11–16 October 2020.
33. Iwamoto, H. Generation of nuclear data using Gaussian process regression. *J. Nucl. Sci. Technol.* **2020**, *57*, 932–938. [CrossRef]
34. Yu, H.; Wang, Z.; Rezaee, R.; Zhang, Y.; Xiao, L.; Luo, X.; Wang, X.; Zhang, L. The Gaussian Process Regression for TOC Estimation Using Wireline Logs in Shale Gas Reservoirs. In Proceedings of the International Petroleum Technology Conference, Bangkok, Thailand, 14–16 November 2016.
35. Almasov, A.; Onur, M. Life-Cycle Optimization of the Carbon Dioxide Huff-n-Puff Process in an Unconventional Oil Reservoir Using Least-SquaresSupport Vector and Gaussian Process Regression Proxies. *SPE J.* **2021**, *26*, 1914–1945. [CrossRef]
36. Sharma, G.; Galvis-portilla, H. Integrated Workflow for Analysis of the Adsorbed Phase Contribution in Liquid Rich Wells in the Unconventional Duvernay Formation. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, Canada, 28 September–2 October 2020.
37. White, A.; Prefontaine, N.; Thomas, F.B. Pseudo Formation Volume Factor: A Consistent and Continuous Volumetric Assessment Across Multiphase Reservoirs, a Wet Gas to Black Oil Duvernay Formation Example. In Proceedings of the SPE Annual Technical Conference and Exhibition, Calgary, AB, Canada, 30 September–2 October 2019.

38. Wüst, R.A.; Ziarani, A.S.; Cui, A.X. Interbedded Carbonate and Calcareous Shales of the Devonian Duvernay Formation of Alberta, Canada: Implications for Completion Due to High Variability of Geomechanical Properties. In Proceedings of the SPE Canada Unconventional Resources Conference, Calgary, AB, Canada, 28 September–2 October 2020.
39. Kong, X.; Wang, H.; Yu, W.; Wang, P.; Miao, J.; Fiallos-Torres, M. Compositional Simulation of Geological and Engineering Controls on Gas Huff-n-Puff in Duvernay Shale Volatile Oil Reservoirs, Canada. *Energies* **2021**, *14*, 2070. [CrossRef]
40. Tian, H.; Zou, C.; Liu, S.; Hong, F.; Hao, J. Reservoir porosity measurement uncertainty and its influence on shale gas resource-assessment. *Acta Geol. Sin.* **2019**, *94*, 233–242. [CrossRef]
41. Luo, G.; Tian, Y.; Bychina, M.; Ehlig-Economides, C. Production-Strategy Insights Using Machine Learning: Application for Bakken Shale. *SPE Reserv. Eval. Eng.* **2019**, *22*, 800–816. [CrossRef]