2018

Application of Artificial Intelligent Predictive Modeling for Completion Optimization and Refracture Candidacy

Daniel Yingling

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>

Recommended Citation

Yingling, Daniel, "Application of Artificial Intelligent Predictive Modeling for Completion Optimization and Refracture Candidacy" (2018). *Graduate Theses, Dissertations, and Problem Reports*. 7009.
<https://researchrepository.wvu.edu/etd/7009>

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

Application of Artificial Intelligent Predictive Modeling for Completion Optimization and Refracture Candidacy

Daniel Yingling

**Thesis submitted
to the Statler College
at West Virginia University
in partial fulfillment of the requirements for the degree of
Master of Science in
Petroleum and Natural Gas Engineering**

Ali Takkiri-Borujeni, Ph.D., Chair

Ebrahim Fathi, Ph.D.

Samuel Ameri, M.S.

Petroleum and Natural Gas Engineering

Morgantown, West Virginia

2018

Keywords: Petroleum Engineering, Artificial Intelligence, Predictive Modeling

Copyright 2018 Yingling

ABSTRACT

Application of Artificial Intelligent Predictive Modeling for Completion Optimization and Refracture Candidacy

Daniel Yingling

Unconventional reservoirs are full of uncertainty when dealing with conventional methods of modeling and analysis. The objective of this work is to use a trained artificial intelligent (AI) model to compare actual production data to AI predicted possible well production. Data-driven models based on AI are efficient tools for optimizing the production, stimulation, and completion design of wells and can be very beneficial when determining the success or failure of wells based on production. An AI model for production predictions require both native and design parameters, which include well characteristics, completion design, and stimulation design parameters.

Data from over 100 Marcellus Shale wells are used to train and test an AI model for production predictions. Feature selection algorithms are used to determine the most influential input parameters pre and post modeling for both increased model accuracy and quality assurance. Post modeling, Monte-Carlo simulation and Type Curves are used to assess the performance of each well based on the AI generated well production values. AI model generation is a very useful tool for predicting production performance of existing wells, which can be used to optimize design characteristics and reservoir production. Generating AI predictive models in fields with low amount of cases to train and test the artificial neural network require very delicate and careful considerations in order to maximize the effect and accuracy of the predictive model.

This study will be able to give an underlying method of applying these artificially intelligent solutions to a complex petroleum engineering problem. The ability to correctly apply these techniques will allow for the optimization of completion and stimulation designs within complex, unconventional reservoir.

Acknowledgments

I would like to express my appreciation for my committee chair, Dr. Ali Takbiri-Borujeni, for his guidance and knowledge during my research. He continually aided and drove my learning in a positive way to develop not only my thesis, but knowledge of coding skills for data mining and data science applications. I would also like to thank my committee members, Dr. Ebrahim Fathi and Professor Samuel Ameri for their support and knowledgeable insight throughout my graduate career. Finally, I would like to thank CNX Resources for providing the data that this study was performed on and permission for publication of the work.

Table of Contents

ABSTRACT.....	ii
Acknowledgments.....	iii
Table of Figures	v
List of Tables	vi
Introduction	1
Theory (Machine Learning)	2
Methodology	3
<i>Data Preprocessing</i>	3
Data Mining & Production Data Analysis	3
Data Correlation and Outlier Identification	5
Data Cleansing and Repairing	9
Feature Selection.....	10
Data Modification (Data Scaling).....	11
Principle Component Analysis (PCA)	11
<i>Machine Learning</i>	12
Data Partitioning	13
Neural Network Training, Calibration, and Testing	13
<i>Data Postprocessing</i>	14
Type Curve Generation	14
Monte Carlo Simulation and Re-Fracture Candidacy	15
Results	16
Feature Selection (Pre and Post)	16
Model Training	18
Type Curve Generation	23
Production Profiles (Monte-Carlo Analysis)	30
Re-Fracture Candidacy	32
Limitations and Conclusions	34
References	35
Appendix A – Feature Selection	36
120 D.....	36
150 D.....	37
Appendix B – Model Training Results	38

120D.....	38
150 D.....	40
90 Day Normalized	42
Appendix C – Type Curve Results	44
Original Models	44
Normalized Model	46
Appendix D – Production Profiles.....	49
90D Model	49
120D.....	51
150D.....	53
Appendix E – Refracture Candidacy	55

Table of Figures

Figure 1 - Backpropagation Neural Network Structure, (Minnaar, 2015).....	2
Figure 2 - Example of Outlier Identification for Max Measured Depth, Avg. Treating Rate, and Bottom Perforation Depth.....	6
Figure 3 - Correlation Plot of Measured Depth and Lateral Length.....	7
Figure 4 - Input Data Cross-plots with Outliers Removed	8
Figure 5 - Heat-Map of Data Correlation Factors Between Input Parameters	9
Figure 6 - 90 Day Feature Importance Scores and Rankings. Pre and Post Modeling.....	16
Figure 7 - 90 Day Normalized Feature Importance Results	18
Figure 8 - Input Data Training Distribution Compared to Full Dataset	19
Figure 9 - Input Data Testing Distribution Compared to Full Dataset	19
Figure 10 - Model Training Loss Profile	20
Figure 11 - Model Training Results Cross plot	21
Figure 12 - Model Testing Results Cross-Plot.....	21
Figure 13 - Model Output Vs. Field Output Trend Confirmation	22
Figure 14 – 90D Model Type Curve Results, Proppant per Foot	24
Figure 15 – 120D Model Type Curve Results, Proppant per Foot	25
Figure 16 – 150D Model Type Curve Results, Proppant per Foot	26
Figure 17- 90 Day Boundness and Proppant/Ft Vs. Normalized BOE.....	28
Figure 18 - 90 Day Stage Spacing and Proppant/Ft Vs. Normalized BOE	29
Figure 19 - Well Monte-Carlo Generated Production Profile, greater than the 90th Percentile ..	30
Figure 20 - Well Monte-Carlo Generated Production Profile, greater than the 50th Percentile ..	31
Figure 21 - Well Monte-Carlo Generated Production Profile, less than the 50th Percentile.....	31
Figure 22 - Well Monte-Carlo Generated Production Profile, less than the 10th Percentile.....	32
Figure 23 - 120 Day Feature Scores Graph, Pre and Post Modeling.....	36
Figure 24 - 150 Day Feature Scores Graph, Pre and Post Modeling.....	37
Figure 25 - 120 Day Model Training Loss Curve.....	38

Figure 26 - 120 Day Model Training Results Cross-Plot	38
Figure 27 - 120 Day Model Testing Results Cross-Plot	39
Figure 28 - 120 Day Model Output Vs. Field Output Trend Confirmation.....	39
Figure 29 - 150 Day Model Training Loss Curve.....	40
Figure 30 - 150 Day Model Training Results Cross-Plot	41
Figure 31 - 150 Day Model Testing Results Cross-Plot	41
Figure 32 - 150 Day Model Output Vs. Field Output Trend Confirmation.....	41
Figure 33 - 90 Day Normalized Training Loss Profile	42
Figure 34 - 90 Day Normalized Model Training Accuracy Results	42
Figure 35 - 90 Day Normalized Model Training Accuracy Results	43
Figure 36 - 90 Day Normalized Model Output Trend Analysis	43
Figure 37 - 90/120/150 Day ATP Type Curve	44
Figure 38 - 90/120/150 Day ATR Type Curve	44
Figure 39 - 90/120/150 Day Boundness Type Curve	44
Figure 40 - 90/120/150 Day Cluster Spacing Type Curve	45
Figure 41 - - 90/120/150 Day ATP Flush Amount Curve	45
Figure 42 - 90/120/150 Day Proppant per Foot Type Curve	45
Figure 43 - 90/120/150 Day Total Shots Type Curve.....	46
Figure 44 - 90 Day Normalized Type Curve, Shots/Ft.....	46
Figure 45 - 90 Day Normalized Type Curve, Fluid/Ft	47
Figure 46 - 90 Day Normalized Type Curve, Cluster Spacing.....	47
Figure 47 - 90 Day Normalized Type Curve, ATR	48
Figure 48 - 90D Production Profile Example, Bad Production	49
Figure 49 - 90D Production Profile Example, Poor Production	49
Figure 50 - 90D Production Profile Example, Avg Production.....	50
Figure 51 - 90D Production Profile Example, Good Production.....	50
Figure 52 - 120D Production Profile Example, Good Production.....	51
Figure 53 - 120D Production Profile Example, Avg Production.....	51
Figure 54 - 120D Production Profile Example, Poor Production	52
Figure 55 - 120D Production Profile Example, Bad Production	52
Figure 56 - 150D Production Profile Example, Good Production.....	53
Figure 57 - 150D Production Profile Example, Avg Production.....	53
Figure 58 - 150D Production Profile Example, Poor Production	54
Figure 59 - 150D Production Profile Example, Bad Production	54

List of Tables

Table 1- Example of Daily Production Data Structure	4
Table 2 - Example of Well Completions and Stimulation Data Format	5
Table 3 - Example of Condensed Completions and Stimulation Data Averaged for Each Well ...	5
Table 4 - Well Completions and Stimulation Data Scaled Formatting	11
Table 5 - 90 Day Feature Importance Scores and Rankings. Pre and Post Modeling	16
Table 6 - 90 Day Normalized Feature Importance Rankings and Scores	17
Table 7 - Model Training and Testing Accuracy Scores	22

Table 8 - Normalized Model Training and Testing Accuracy Scores	23
Table 9 - Monte-Carlo Well Profile Results and Refracture Candidacy	33
Table 10 - 120 Day Feature Importance Scores and Ranking Table, Pre and Post Modeling.....	36
Table 11 - 150 Day Feature Importance Scores and Ranking Table, Pre and Post Modeling.....	37
Table 12 - 120 Day Refracture Candidacy Results.....	55
Table 13 - 150 Day Refracture Candidacy Results.....	56

Introduction

The development of production from unconventional resources took the oil and gas industry by storm within the past few years, generating a massive increase of natural gas production in the northeastern United States, specifically in the Marcellus Shale formation in the Appalachian basin. With this new spike in oil and gas production from this previously thought underwhelming formation type, many problems arose to successfully understand what is causing the oil and gas production. A general technique to determine the optimal design and reservoir production plan was to create reservoir model using conventional oil and gas reservoir techniques. This process involved the use of the major fundamental equations and theories but altering them in a way to force it to work for shale reservoirs. This adaptation of the original fundamental equations introduced many forms of user bias and uncertainty to the reservoir models. In order to counter this issue, a new way of thinking and analyzing field data from unconventional reservoirs was needed. Without a new way to develop accurate and reliable models to generate optimal drilling, completion, and stimulation design scenarios, shale formation production will continue to be generated using these conventional methods based on copious amounts of uncertainty, and user bias.

Around the same time as the discovery of Marcellus Shale and the economic success potential of unconventional reservoirs, a new emerging idea was in development in the technology sector. The use of a system of interconnected artificial neurons to replicate the systematic approach of problem solving, similar to a human brain, led to the development of Machine Learning or Artificial Intelligence (AI) algorithms. Machine learning is the foundation behind this new type of reservoir analysis and production modeling. Machine learning lets the unknowns about unconventional reservoirs not interfere with the analysis by reducing all major assumptions when it comes to modeling. Using artificial intelligence to model unconventional reservoirs requires substantial amounts of real field data to be trained and calibrated correctly. This innovative technology is not biased because only field data is used when generating the AI model, whereas conventional techniques impose many assumptions to generate models. Applications of machine learning are being used in other areas of the petroleum industry for example, using artificial intelligence to optimize well hydraulics in real time during the well drilling process (Wang & Salehi, 2015).

Theory (Machine Learning)

The development of the neural network training process for machine learning applications is revolutionizing the world by offering an easy solution to once previous complex scenarios. Applications of machine learning range from self-driving cars to voice recognition programs in a standard smart phone. Neural networks are modeled after the neurons in a human brain, which takes input data to derive a solution. Due to this method's simplistic form, it can provide an accurate solution to complex scenarios with a fraction of the computational power. The neural network technique used in this study is a backpropagation training process. In order to successfully train a backpropagation model, three controllable factors can be adjusted to affect the learning rate of the process. The initial learning rate, momentum, and early stopping criteria can be used to determine the speed of the machine learning process. The initial learning rates dictate the amount each hidden neuron can be adjusted, while the momentum controls the amount of iterations an error adjustment is allowed. The early stopping criteria can be applied to let the model stop training if a specific value or iteration number is met. No general rules are defined for setting these values, but typically the momentum will be set to approximately 0.5, which is half the maximum limit for the training process. (Choy, Lee, & Lo, 2003) Seen below in Figure 1 is an example of a backpropagation neural network training scenario.

Figure 1 - Backpropagation Neural Network Structure, (Minnaar, 2015)

Neural network training involved the process of identifying patterns within the input data to derive a solution. This process requires multiple iterations of training for the neural network to accurately assign a weight to certain input parameters and their significance to the output result. The ability to identify an accurate solution without computing complex equations allows for the computational power and testing of the models more accurate in these highly complex situations.

Neural network method can be used in modeling complex production estimation problems using records of past response.

Methodology

Creating predictive AI models requires a comprehensive knowledge about how data science is performed. In this section, all the steps taken to create a machine learning based predictive model for completion and stimulation optimization for the Marcellus shale reservoir. The major sections of this study are condensed into 3 main categories; Data Preprocessing, Neural Network Training and Testing, and Data Postprocessing. Each category is then divided up in to smaller sub groups to create a linear workflow of the study.

Data Preprocessing

Data preprocessing was the first step in this study, which included the use of data mining techniques, production data analysis, data modification and cleansing, and Feature Selection/Elimination, as well as a study of Principle Component Analysis and its benefits. This process and all of its different sections are used to take the initial raw field data and turn it into an acceptable format for the machine learning algorithms to train and test. The process of generating the correct format of data allows for the training and testing process of machine learning to be more accurate and reliable. Understanding the data and how it needs to be applied to obtain the maximum amount of information is the most important factor in the preprocessing section.

Data Mining & Production Data Analysis

The first step in the data analytics process is to extract as much information from the raw data as possible. In this study; Well, Production, Completion, and Stimulation data was provided using excel worksheets. The production data was recorded on a daily basis for each well, which included the lateral length as well as the Gas, Water, and Condensate rates. Table 1 shows the original production data in its daily formatting.

WELL NAME	PLAY	PAD UNIT	Lateral Length	Active Days	Gas	Condensate	Water
Well #1	MARCELLUS	Pad #1	2437.7	0	20.56	0	0
Well #1	MARCELLUS	Pad #1	2437.7	0	97.654	0	0
Well #1	MARCELLUS	Pad #1	2437.7	1	1184.452944	0	0
Well #1	MARCELLUS	Pad #1	2437.7	2	2203.836271	0	0
Well #1	MARCELLUS	Pad #1	2437.7	3	4475.109554	0	252.2591482
Well #1	MARCELLUS	Pad #1	2437.7	4	2718.449668	0	96.36535435
Well #1	MARCELLUS	Pad #1	2437.7	5	3884.139402	0	148.135169
Well #1	MARCELLUS	Pad #1	2437.7	6	2120.316814	0	84.51853196
Well #1	MARCELLUS	Pad #1	2437.7	7	4419.00142	0	155.6345004
Well #1	MARCELLUS	Pad #1	2437.7	8	3688.305668	0	75.26267818
Well #1	MARCELLUS	Pad #1	2437.7	9	1162.324694	0	22.94158526
Well #1	MARCELLUS	Pad #1	2437.7	10	786.058934	0	0
Well #1	MARCELLUS	Pad #1	2437.7	11	1752.724235	0	0

Table 1- Example of Daily Production Data Structure

The production data being in a daily format leads to the presence of days where the well is shut in and not producing, or a major spike is present that doesn't fit the average trend of the data. In this scenario, that day would be removed due to the lack of accurate or available production data available, which can be seen in above in Table 1 highlighted in red.

The data was loaded into a python code that was generated to create the data in a format of cumulative production for a specified active producing day range. In this study, 30,60,90,120, and 150 Day cumulative production was generated for each well to show how each parameter affects production over different time intervals.

The next step of the data mining process was to analyze the completion and stimulation data. This process was much more tedious due to the different formatting styles and naming conventions used from the different service companies used to perform the fracture stimulation jobs. A list of common attributes for each master sheet of completion and stimulation data was created to extract common attributes from each service company's field data (Table 2). Once the common attributes were identified, a table of attributes for all of the available wells was created to be combined with the production data to form the master dataset for each well available. This process however reduced significantly the number of available wells in the final data set, from 181 wells with production data to 100 wells with available production, completion, and stimulation

data. The final format for the data is on a per well basis with a variety of completion, stimulation, and production attributes (Table 3).

Well	Pad Stage #	Stage	Mid Perf TVD	Clusters	Total Shots	Phasing	Diam	Depth	Top Perf	Bot Perf	Interval	Break Down	ATP	ISIP
Well #1	1	1	8861.71	4.54	81.78	68.15	0.64	29.76	11828.48	12060.19	231.71	7932.60	8396.02	6601.42
Well #1	3	2	8285.63	5.32	63.86	63.86	0.60	27.88	10762.28	11012.39	250.11	7353.30	8307.98	5670.63
Well #1	5	3	9458.82	6.08	72.98	72.98	0.68	31.87	11931.47	12225.80	294.33	8380.00	9792.07	5852.62
Well #1	7	4	9254.68	5.95	71.41	71.41	0.67	31.18	11316.03	11604.05	288.02	8062.14	9289.19	6024.59
Well #1	9	5	6608.37	4.25	50.99	50.99	0.48	22.27	7823.64	8029.30	205.66	5772.97	6427.35	4350.34
Well #1	11	6	8660.75	5.57	66.81	66.81	0.62	29.17	9914.55	10170.65	256.10	7585.12	8695.27	6186.57
Well #1	13	7	7573.35	4.87	58.39	58.39	0.54	25.50	8371.36	8606.87	235.51	6815.24	7473.11	5584.15
Well #1	15	8	7210.22	4.63	55.60	55.60	0.52	24.28	7691.15	7915.40	224.25	6097.32	7192.61	4751.83
Well #1	16	9	9579.62	3.71	66.78	74.20	0.69	32.40	9981.55	10179.43	197.87	8951.37	9671.14	5888.01
Well #2	2	1	8139.27	4.18	75.18	62.65	0.58	27.36	10316.36	10529.37	213.01	9013.24	8260.39	4992.16
Well #2	4	2	6609.81	4.24	50.88	50.88	0.47	22.22	8122.75	8319.50	196.75	5727.82	6742.95	4346.33
Well #2	6	3	9313.42	5.98	71.73	71.73	0.67	31.32	11090.02	11379.35	289.33	8376.10	9477.21	6195.39
Well #2	8	4	6268.55	4.03	48.31	48.31	0.45	21.09	7225.01	7419.85	194.84	5998.84	6311.22	4898.26
Well #2	10	5	7964.03	5.11	61.37	61.37	0.57	26.80	8870.28	9117.82	247.53	7106.87	7849.47	5629.85
Well #2	12	6	6309.70	4.05	48.64	48.64	0.45	21.24	6785.58	6981.77	196.19	5755.18	6171.07	3674.10
Well #2	14	7	8399.05	4.33	52.01	65.02	0.61	28.39	8752.30	9004.78	252.48	6542.83	8247.34	5267.42

Table 2 - Example of Well Completions and Stimulation Data Format

Well Name	100 Mesh	1_ISIP	ATP	ATR	Bot_Dpth	Brkdwn_Press	Clean_Vol	Clusters
Well #1	57430.43052	4501.059671	9516.526161	105.3096949	12785.5936	7504.556143	5985.553686	4.563204976
Well #2	115893.0677	3741.413275	8096.293665	87.6495635	11504.08154	6208.802192	4711.893765	3.806609528
Well #3	57884.13597	4539.824212	9791.185863	108.4915934	13719.54132	7207.834378	5608.333001	4.565212186
Well #4	47421.72129	3782.369451	8060.665254	91.54176506	11324.90096	5875.772904	4514.537457	4.733678332
Well #5	109719.0958	3577.087246	7428.617226	83.9135911	10503.66848	5371.687139	4118.785623	4.362320186
Well #6	56342.19271	4604.504538	9623.486228	105.4958361	14005.24032	6836.152289	5624.326576	5.623052664
Well #7	60519.8366	0	9898.713452	112.2158168	12069.02417	8804.39617	11897.71749	5.965904186
Well #8	43489.46695	0	7084.31879	75.87789405	8375.672196	6487.04023	6603.325701	4.287620726
Well #9	224459.2475	4313.2727	7924.418724	83.4431134	10968.43562	6822.232175	9234.728878	4.798484124
Well #10	200265.4978	3633.359771	6914.719171	71.41005124	9515.426616	6121.567727	8545.851576	4.261505714
Well #11	237256.9927	4604.717168	8285.812528	86.18181322	10548.01067	6854.475668	9024.642609	4.931850613

Table 3 - Example of Condensed Completions and Stimulation Data Averaged for Each Well

Data Correlation and Outlier Identification

Once the data is compiled into a per well format, an initial analysis of the available data can allow identification and classification. The initial phase of the data analysis was to identify any data that would be considered an outlier along with any parameters that are highly correlated with each other. Data outliers are a very serious problem to have within the dataset, due to the distortion it adds to the normalization process. An example of this can be seen in Figure 2 below. Another method for outlier identification revolves around Z-Score analysis of the input data, which is a test of the statistical significance of the input data in order to identify if it lies within the

acceptable range of data values. Values deemed unacceptable can be marked as outliers to provide an initial list of cases for each data input parameter. This technique paired with the traditional crossplot method allows for a very thorough examination of the input data to successfully identify and remove outliers from the dataset. When creating a model based on real field data, outliers within the dataset is a very common issue. Some possible explanations for these data outliers are caused by machine or human error. Machine recordings from the field could result in the presence of outliers, while data recording from the field could cause inaccurate measurements as well. These types of scenarios are the major causes of data outliers, which can be successfully dealt with during the data pre-processing stage if the data analyst has expertise in the petroleum engineering field and understands the acceptable values for each data input parameter.

Figure 2 - Example of Outlier Identification for Max Measured Depth, Avg. Treating Rate, and Bottom Perforation Depth

Identification of highly correlated parameters within the dataset is an important part of the initial feature selection and elimination process. Highly correlated data provides multiple complexities to the model training process. Inclusion of highly correlated parameters can result in lower model accuracy due to the lack of variation within the input dataset. Highly correlated data has the effect of not adding new viable data to the model training process (Kuhn & Johnson, 2016). For example, if two inputs, parameter A and B, are 100% correlated and included in the model training process, parameter A will be fully utilized for the training process, but parameter B will not contribute any new viable information to the training due to the lack of variation from parameter A. This however from a Petroleum Engineering standpoint must be examined closely however due to the known significance of the different input parameters. For this study, the final goal is to optimize completion practices for a completion and stimulation operation for a Marcellus shale gas well, so removal of key completion parameters due to correlation can hinder the accuracy

of study. Addition of highly correlated parameters can add significance to an input parameter, which can be beneficial if the multiple inputs contribution trends are known. An example of this practice would be the inclusion of Lateral Length and Measured Depth of a well into a model training process. Both of these parameters can be highly correlated with each other, but it is a known practice that increasing the formation exposure increases production in an unconventional reservoir. Examples of data correlation can be seen below in Figure 3 through Figure 5.

Figure 3 - Correlation Plot of Measured Depth and Lateral Length

Figure 4 - Input Data Cross-plots with Outliers Removed

Figure 5 - Heat-Map of Data Correlation Factors Between Input Parameters

Data Cleansing and Repairing

When creating a dataset based on field recorded data, there is a high possibility of the presence of data outliers and incorrect information. The presence of these outliers can highly skew the resolution of the data attributed due to its inconsistencies with the common attribute values. Once the data has been converted from its raw format and into the final structured per well format, Data cleansing and removal of outliers can be performed. This process of removing outlier data

must be performed with extreme caution, due to the already limited available data. In this scenario, for the completion and stimulation parameters, it was determined that the data wouldn't be eliminated, but rather replaced with the pad average for the specific attribute. This technique was deemed appropriate for this study due to the common design nature of service companies for a specific gas production pad. The production data in its final form is correct without outliers due to the cleaning process during the production data analysis.

Feature Selection

When training and testing a neural network for an AI predictive model, it is vitally important to understand the impact each input parameter has on the output value for a variety of reasons. If the input parameter has a direct correlation to the output value, it could potentially be a very good predictor for the model. Another reason for feature selection is to select a variety of diverse types of input parameters, whether they are static and cannot be changed, for instance well and reservoir characteristics, or completion and stimulation design parameters that can be altered after a well is completed through refracturing. Selecting parameters that are important to the completion and stimulation is vital as well, for example the parameters such as proppant per foot, fluid amount, stage length, etc. are all vital parameters when dealing with a hydraulic fracture operation (Belyadi, Fathi, & Belyadi, 2017). This flexibility allows for the modification and optimization of design techniques to allow for increased production and well economics. In this study, two types of feature selection are applied to the data to ensure the most successful and accurate predictive model. Having an abundance of data that is highly correlated can hurt the accuracy and success of the model due to the vast number of redundant trends within the input parameters. The identification and elimination of most highly correlated data increases the success rate of the predictive model. Initially, cross-plots are created to identify any input parameters with high correlation to each other. These cross-plots can be used to generate heat maps of highly correlated input parameters. This process must be handled carefully, due to the significance of each input parameter in a real-life field development strategy. Examples of these input data cross-plots and heatmaps can be seen above in Figure 4 and Figure 5 respectively. After the initial feature selection and elimination, a logistical regression style of feature selection can be applied to the input parameters influence on the desired target output value. For this study, an algorithm was used in Python called L1-based feature selection. This algorithm is part of Python's SkLearn

package in the Feature Selection library. This method identified any pattern of influence an input parameter has on the target output.

Data Modification (Data Scaling)

When the dataset is compiled via the data mining and cleaning processes, the final step is to scale the data. Scaling the data is a technique that reduced the dimensionality of the inputs and output data. The technique used in this study was to scale the data, so the max and min of each parameter was 0 and 1 respectively. A data preprocessing library within Python called MinMaxScalar converts the input and output data into scaled, dimensionless values. This technique allows for the neural networks to train and calibrate without the reliance of overlying similarities between the input and output parameters. The MinMaxScalar operation performs the following equation to each data parameter.

Equation 1 - MinMaxScalar (Data Scaling)

$$X_{scaled} = \frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})}$$

Once the data is scaled between 0-1 for each parameter, feature importance and elimination can be performed. An example of this process on the available completions data shows how the data is transformed into its dimensionless values for feature selection and machine learning.

Well Name	100 Mesh	1_ISIP	ATP	ATR	Bot_Dpth	Brkdwn_Press	Clean_Vol	Clusters
Well #1	0.071946852	0.977488846	0.871920903	0.830756221	0.78334986	0.621336963	0.239977428	0.350390089
Well #2	0.373662204	0.812517498	0.395970764	0.397971023	0.555710363	0.243864261	0.076245448	0
Well #3	0.074288346	0.985907287	0.963965216	0.908732911	0.94925028	0.53489743	0.191484821	0.351319657
Well #4	0.020293671	0.821411894	0.384030925	0.493354641	0.523881885	0.146847799	0.05087483	0.42933872
Well #5	0.341799425	0.776831044	0.172218177	0.30641601	0.37800347	0	0	0.257357492
Well #6	0.066330649	0.999953824	0.907765499	0.835317862	1	0.426620822	0.193540833	0.841220595
Well #7	0.087890732	0	1	1	0.65606311	1	1	1
Well #8	0	0	0.056836442	0.109490479	0	0.324919205	0.319393475	0.222763112
Well #9	0.933953096	0.936707412	0.338371812	0.294886323	0.460561693	0.422565683	0.657666546	0.459351202
Well #10	0.809093424	0.78905167	0	0	0.202458589	0.218451544	0.569109748	0.210668879
Well #11	1	1	0.459482569	0.362001834	0.38588013	0.4319587	0.630659462	0.521115116

Table 4 - Well Completions and Stimulation Data Scaled Formatting

Principle Component Analysis (PCA)

Principle component analysis (PCA) is the practice of reducing the dimensionality of input parameters to identify the true impact each parameter has on a specified number of principle components. A principle component is a dimensionless parameter that is used to train a more

reliable predictive model (Kuhn & Johnson, 2016). PCA is used to generate a specified number of principle components which are made from a combination of each individual input parameter. The effect each input parameter has on a principle component can be used to determine the feature importance of the field input parameters. PCA was determined to be an unsuccessful model training method due to the lack of data cases available within the trial dataset. Principle component analysis was used however to cross check the feature importance and selection for the final predictive model. In order to create a list of most influential parameters, a combination of all input parameters was used to generate a singular value decomposition and the weights of each parameter would determine the most influential parameters (Kuhn & Johnson, 2016). This case can be related to training a predictive model, where the model looks at a combination of all input parameters and create an output results from the trending data. The fundamentals of the two scenarios rely on the underlying principle that different input parameters effect the results in separate ways, making certain parameters more important than others.

Machine Learning

The process of machine learning to create a predictive model utilizes a key component in the data science field called artificial neural networks. Artificial neural networks are similar to the structure of a human brain, with different groupings of neurons that are used for problem solving and everyday tasks.

In the neural networks model each neuron is linked to its neighbors with varying coefficients of connectivity. These coefficients represent the strengths of these connections. By adjusting the strength of these connections neurons can then be grouped into layers and therefore learning is accomplished. The input layer consists of neurons that receive input from the external environment. The output layer consists of neurons that communicate the output of the system to the user or external to the environment. There are usually a number of layers between these two layers. These hidden layers act as a black box to link the relationship between input and output. For this project due to the non-linear relationship between the input and output variables, the hidden layer extracts higher level features that enable the generalization of outputs.

Artificial neural networks can be applied to the petroleum engineering field by understanding and solving the unknowns of unconventional modeling (Mohaghegh, 2017). This

study utilizes artificial neural networks to create a predictive model for completion and stimulation design optimization with limited amounts of available data.

Data Partitioning

Once the input data has been properly configured, scaled, and selected, the next step of the process is to separate the data into a training dataset and a blind test dataset. This process allows for the overall increased accuracy of the model due to testing on model unseen data as a true test of accuracy (Mohaghegh, 2017). This process begins by analyzing the amount of available training cases present in the dataset. Determining the proper ratio of training to blind data is all based on the complexity and amount of the input and output data. In this scenario, the dataset was split into the following configuration; 85% Training data and 15% Blind Testing data. This ratio was selected due to the limited amount of cases available, keeping the maximum amount of data in the models training process. Multiple techniques can be used for data partitioning, traditional randomly selected data subset generation and a technique that involves Latin Hypercube Sampling (LHS). Latin hypercube sampling can be used to select values from every point of the full dataset to accurately represent the data distribution. The random partitioning can also be used to accurately represent the initial data distribution, but this process is completed by creating actual vs. split distributions and re-rolling the sampling until the distributions are similar. LHS does this process inherently but can be difficult to perform when dealing with datasets with sparse number of samples. The random data partitioning technique was used for this study, but LHS is being implemented in future studies.

Neural Network Training, Calibration, and Testing

Once the data has been separated into the training and blind dataset, model training and testing can begin. In this study, a Python package/module called SKlearn/MLPRegressor was used as the structure of the artificial neural network. The module required the data to be in a specific format, with each well's individual parameters to be set as the inputs, with the corresponding production value at the users preferred production period. This study focused on production data at the 3, 4, and 5-month production period for maximum model efficiency. The 1 and 2-month data was determined to be too inconsistent in terms of production data and were therefore eliminated from the study. The limited amount of data for this study also prohibited the use of data past the 5-month mark due to the reduction of cases with production above 5 months. The training process begins by specifying the initial learning rate and momentum of the neural network, along

with the early stopping characteristics of the module. The initial learning rate value determines how quick the model takes to change its values based on the training loss values. The momentum is the speed at which the model moves through the data and availability to move past local minimums within the training loss curve. The tolerance value is the value at which the model will stop itself if a training loss doesn't improve by the set amount over multiple iterations. For this study, the model was trained using a learning rate of 0.015, momentum value of 0.85, and a tolerance value of 0.00001.

Data Postprocessing

Once the AI predictive model is trained and tested, it can be used to analyze how changing input data can alter the production of a well. Multiple techniques involve the use of simulation data to create multiple scenarios for each individual well. These processes involve the use of Monte-Carlo simulation on data that can be altered in completion or stimulation designs. The input data within the data frame are split into two categories, dynamic and static data, with dynamic data being data that can be parameters that can be altered after the well has been drilled. Static data are parameters that cannot be changed once the well has been drilled, such as reservoir, or well design characteristics. Identifying what causes changes in production is a vital part of increasing well production and economics.

Type Curve Generation

Similar to the well quality analysis, type curves are a tool that shows how trends in the data respond to the change of other input parameters. Type curves can be used to take an in depth look at how well production is affected by the changing of multiple well parameters. Well type curves can also be used as a general accuracy check of the AI generated predictive model. Similar trends in the input data results for the well quality analysis and the type curves can be a model confirmation check. Along with the confirmation aspect, the type curve's that are generated can give an in depth look at how changing the dynamic well input parameters effect the production of the sample wells. Similar to the production profile generation, only in the type curve scenario, it compares only two input parameters against the wells production.

Monte Carlo Simulation and Re-Fracture Candidacy

Once the AI trained model has been trained and tested, analysis of the data and wells can be performed by creating the production potential profiles for each individual well within the field. This process is started by performing Monte-Carlo simulation on the input data that is classified as dynamic. The dynamic input data is the data that can be altered after a well is drilled, which include the completion and stimulation design parameters. These input parameters are the focus of the present study, which can be used to identify and optimize the success of a wells completion/stimulation design. Once the Monte-Carlo distributions are generated for each well, the data is put into the AI model to predict the production profile for each well. This profile is then cross matched with the actual production of the well, which can be used to determine the wells success compared to its overall potential. Re-fracturing a well involves identifying the most beneficial candidates based on the production potential of each well compared to its actual production. Re-fracturing a well is used to increase the overall production each selected well. The wells candidacy selection is based on the rank of each well compared to the other wells in the field by using the Monte-Carlo simulated production data.

Results

Feature Selection (Pre and Post)

Using a combination of the input data correlation plots and L1 based feature importance values, final feature selection was performed to create list of input data to import for model training the testing. Once the model was trained and tested, output values for every well was calculated and used to re-rank the features as a model confirmation (Table 5 and Figure 6).

	L1 - FI Pre	L1 - FI Post
MaxMD	1.0000	1.0000
Lateral Length	0.9277	0.9594
Stages	0.7673	0.7644
Flush Actual	0.6293	0.5916
ATP	0.6011	0.5978
AVG Treating Rate	0.5285	0.4311
BoundnessFinal	0.4530	0.4921
Proppant/Ft	0.1216	0.1053
Total Shots	0.0582	0.1007
Cluster Spacing	0.0200	0.0463

Table 5 - 90 Day Feature Importance Scores and Rankings. Pre and Post Modeling

Figure 6 - 90 Day Feature Importance Scores and Rankings. Pre and Post Modeling

The results of the pre and post model feature importance rankings can be seen for each of the 120 and 150 day predictive models in Appendix A: Figure 23 - Figure 24, and Table 10 - Table 11. The model confirmation comes from the order of feature importance. Generally, the ranking of the features should be in similar order for both pre and post model training. Examination of this feature importance suggests that for model training, the length of the wellbore allows reacts very positively to the production output, while the specifics of the completion job, Total Shots and Cluster Spacing, are less impactful for model training. These features may not be as impactful for model training but are vital to include due to the predictive model's ability to optimize only parameters included within the model. This study proved to be accurate for all three predictive models. In addition to the original three models, a fourth model was generated by using data that was normalized based on lateral length of the well. The Original KPI for this study is shown below in Table 6 and Figure 7.

	L1 - FI Pre
Proppant_Ft	100.00%
ATR	24.95%
Stage_Spc	4.91%
ATP	4.08%
Cluster_Spc	1.43%
Boundness	0.1012%
Fluid_Ft	0.0023%
Shots_Ft	0.0018%

Table 6 - 90 Day Normalized Feature Importance Rankings and Scores

Figure 7 - 90 Day Normalized Feature Importance Results

Model Training

To begin the model training process, data partitioning was performed on the dataset to provide a training and testing datasets for model creation and confirmation. The data partitioning was performed by random sampling but with monitored results. Sampling was performed until the distribution of training and testing data samples matched the distribution of the entire dataset. This was performed in order to accurately sample the data, without focusing on only one specific section of the dataset. For this study, with limited number of cases to select from, this process was extremely important in order to achieve an accurate predictive model. Seen below in Figure 8 and Figure 9 are respectively the results of the training and testing data partitioning distributions. Achieving a similar distribution was generated by multiple iterations of random sampling.

Figure 8 - Input Data Training Distribution Compared to Full Dataset

Figure 9 - Input Data Testing Distribution Compared to Full Dataset

After the data partitioning was performed with similar sample distributions, the neural network training process was performed to create the predictive model. The training dataset was imported into the MLPRegressor package in Python to begin the model training, with the testing dataset being used as the model confirmation. This process was “looped” in the code to run until a training and testing accuracy (R^2) value of 80% was achieved. Once these criteria were met, model calibration was performed by changing the initial learning rate and model momentum until the optimal parameters were determined. Once the model was calibrated correctly, the model was reinitialized into the code loop until an accuracy of 90% was achieved for both Training and Testing datasets. For the 90 Day model shown in Figure 11 to Figure 12, a 92% accuracy for both the Training and Testing datasets. Shown in Figure 10 is the Training loss value compared to the number of iterations the neural network performed. The number of iterations for each model exceeded the 1000 mark due to the minimal training loss limit, which was set to achieve a very thorough model training process due to minimal computational limits.

Figure 10 - Model Training Loss Profile

Figure 11 - Model Training Results Cross plot

Figure 12 - Model Testing Results Cross-Plot

Another model confirmation was the generation of the model output trend graph which can be seen below in Figure 13. This plot shows the actual field and model predicted output for each data sample within the entire dataset. The trend of these output values should be very similar if the model has achieved a high accuracy, which in this study, has been achieved. Other model confirmation techniques and post-processing scenarios will be discussed later in this report.

Figure 13 - Model Output Vs. Field Output Trend Confirmation

The results of this study were split into 3 main sections, which include the 90, 120, and 150 day training and model generation for each time period. The model training was performed the same for the 120 and 150 day model as the 90 day model, with similar model training and testing results. The results of these training and testing accuracies can be seen in Table 7 below as well as the images for each individual model in Appendix B: Figure 25 through Figure 32.

	90 Day Model	120 Day Model	150 Day Model
Training Accuracy	92%	94%	91%
Testing Accuracy	92%	92%	90%

Table 7 - Model Training and Testing Accuracy Scores

The results of the Normalized model that was created was generated in a similar manner, but the training and testing accuracies resulted in lower scores. The table of model accuracies can be

seen below in Table 8, along with every graph generated during the model training can be seen in Figure 33 through Figure 36. Multiple combinations of adjustments to the predictive model was made to create the most accurate predictive model possible.

	90 Day Normalized
Training Accuracy	83%
Testing Accuracy	77%

Table 8 - Normalized Model Training and Testing Accuracy Scores

Type Curve Generation

Once the model training, calibration, and testing is complete, another form of model confirmation and post-processing can be performed on the predictive models. Type curves are a tool to identify how certain parameters affect each other compared to an output value. For this scenario, specific completion and design parameters were compared to lateral length and BOE production. The use of type curves is to determine the influential and optimal completion scenarios based on the predictive model's output. Seen below in Figure 14 through Figure 16 are type-curves for Proppant per foot. Seen below in the figures are Proppant per Foot compared to Lateral Length. The 90 Day model shows that initially within a well that increasing the Proppant per Foot and Lateral Length will increase production, but the 150 Day model shows that production begins to reduce once a Lateral Length of 8000 ft is reached for high Proppant concentrations and around 4500 ft for lower proppant concentrations. Other examples of type curves can be seen in the Appendix C: Figure 37 to Figure 43.

Figure 14 – 90D Model Type Curve Results, Proppant per Foot

Figure 15 – 120D Model Type Curve Results, Proppant per Foot

Figure 16 – 150D Model Type Curve Results, Proppant per Foot

Similar to the original models, the normalized model also generated type curves to identify specific trends in the completion and optimization process. Seen below in Figure 17 and Figure 18 are type curves that represent the relationship between boundness and Stage Spacing Vs. proppant per foot and Normalized BOE. In the boundness type curve, we can identify initially

that wells that are unbound can produce more hydrocarbons compared to wells that are bounded by other horizontal wells. This results shows that the more confined a well it within the formation, the less hydrocarbons it will produce. In a conventional reservoir, this is a known fact due to the drainage area of each well, but for unconventional reservoirs, specifically Marcellus Shale, this communication between wells doesn't typically occur unless there are intersecting wells within the formation. In Figure 18, Stage Spacing is compared to Proppant per foot and normalized BOE. This graph identifies that wells with lower stage spacing (more stages per lateral), the increase of production. For the study presented, there was a limit to the smallest amount of stage spacing available, so the model couldn't be generated to predict when or if the production would decrease with smaller stage spacing. These results provide an in depth look at the characteristics of the completion and stimulation designs when compared to the normalized production values and proppant per foot. The remainder of types curves generated can be seen in the appendix in Figure 44 through Figure 47.

Figure 17- 90 Day Boundness and Proppant/Ft Vs. Normalized BOE

Figure 18 - 90 Day Stage Spacing and Proppant/Ft Vs. Normalized BOE

Production Profiles (Monte-Carlo Analysis)

Once the predictive models were created, analysis of the field data continued with well production potential based on Monte-Carlo simulated production profiles. The first step of this process was to simulate the dynamic data for each well for 1000 iterations. This data was then input into the predictive models to create the production profiles. The actual field production was compared to the P10, P50, and P90 values of the production profile to determine the wells production-based success. In Figure 19 through Figure 22 below, different examples of wells production quality can be seen. Respectively, the graphs below are examples of great, good, poor, and very poor wells based on the simulated production profile.

Figure 19 - Well Monte-Carlo Generated Production Profile, greater than the 90th Percentile

Figure 20 - Well Monte-Carlo Generated Production Profile, greater than the 50th Percentile

Figure 21 - Well Monte-Carlo Generated Production Profile, less than the 50th Percentile

Figure 22 - Well Monte-Carlo Generated Production Profile, less than the 10th Percentile

Re-Fracture Candidacy

After the Monte-Carlo simulated production profiles were developed, ranking of each well can be calculated using the field production and the different percentile values. The P50 and P90 values were subtracted from the actual field production to develop a P50 and P90 score, these values were sorted from min to max create a well rank for each. The P50 and P90 scores and ranks were recorded in Table 9 below. Once the rank for P50 and P90 were determined, the average of the P50 and P90 ranks are used to determine the overall well rank. The refracture candidacy ranking can be used to select specific wells to be re-stimulated and completed to increase its overall production and increase well economics.

Well #	Field BOE	P10	P50	P90	P50_Score	P90_Score	P50_Rank	P90_Rank	Overall Avg	Overall Rank
88	38478.02518	71387.50988	149925.1699	234606.3073	-111447.1448	-196128.2821	1	1	1	1
15	102940.5055	94849.20603	189216.2134	282935.3451	-86275.70788	-179994.8395	2	2	2	2
2	93121.1976	88389.09677	168101.9893	269652.19	-74980.79166	-176530.9924	5	3	4	3
25	93640.03204	83232.74946	167066.1585	267866.2113	-73426.12648	-174226.1793	6	4	5	4
92	17244.62288	24016.01137	99100.86693	170144.1816	-81856.24406	-152899.5587	3	7	5	5
52	27708.07454	32766.91572	103806.969	175915.0285	-76098.89446	-148206.954	4	8	6	6
6	88299.45744	75377.95178	157481.1176	241338.3732	-69181.66017	-153038.9158	7	6	6.5	7
49	111931.0857	89570.82014	175208.5904	266213.7811	-63277.50472	-154282.6954	11	5	8	8
69	46844.16815	37929.79416	114157.0141	194884.0401	-67312.84591	-148039.8719	8	9	8.5	9
30	28752.11608	18620.43797	93432.88024	167281.1384	-64680.76416	-138529.0224	9	11	10	10
27	32778.02237	20972.02889	96671.49779	170327.4046	-63893.47542	-137549.3822	10	12	11	11
56	46329.20219	36512.68756	106428.9276	182512.1144	-60099.72544	-136182.9123	15	13	14	12
89	28280.68976	16579.2212	90586.5003	163387.3163	-62305.81054	-135106.6265	12	16	14	13
4	36180.02876	24957.44457	97503.38899	169245.9918	-61323.36023	-133065.9631	13	21	17	14
60	46455.86612	30461.75268	105207.3482	180269.8838	-58751.48204	-133814.0177	19	17	18	15
84	39190.33021	31653.05176	99391.34818	171310.7722	-60201.01797	-132120.442	14	22	18	16
81	89067.03815	70942.73085	146430.4563	224681.1933	-57363.41811	-135614.1552	22	15	18.5	17
26	45180.86287	30452.07956	104235.4389	178811.618	-59054.57604	-133630.7551	18	19	18.5	18
70	48767.79884	37858.75753	107995.0959	179433.2408	-59227.2971	-130665.442	17	24	20.5	19
54	40011.46458	31505.74685	99838.06031	169086.9938	-59826.59573	-129075.5292	16	28	22	20
57	35416.56027	18763.39893	91314.30291	168858.7518	-55897.74264	-133442.1916	25	20	22.5	21
83	28255.64711	14930.90309	86875.0056	158615.8979	-58619.35849	-130360.2508	20	25	22.5	22
47	141395.2299	99292.92376	188075.0526	287789.2081	-46679.82264	-146393.9782	40	10	25	23
82	40690.6239	27103.3127	98623.98506	169025.9654	-57933.36116	-128335.3415	21	30	25.5	24
86	49327.60527	32262.07386	105442.8859	178051.8296	-56115.2806	-128724.2243	24	29	26.5	25
46	62719.62161	47120.69167	118990.6533	189728.0738	-56271.03169	-127008.4522	23	33	28	26
5	32958.84829	15340.39346	87477.78129	160555.2495	-54518.93301	-127596.4012	26	32	29	27
12	118572.7149	81453.78096	164865.7564	252330.8036	-46293.04154	-133758.0887	42	18	30	28
73	131651.5993	93624.80969	176367.1326	267777.2301	-44715.53338	-136125.6308	47	14	30.5	29

Table 9 - Monte-Carlo Well Profile Results and Refracture Candidacy

Limitations and Conclusions

The applications of artificial intelligence and machine learning are processes that solely rely on available data that is useful to the study at hand. For the process of machine learning for completion optimization, the availability of well data dictates the application of the study. This study was performed on a real field within the Marcellus shale reservoir through a series of unconventional, horizontal wells. The initial count of wells with partial data surpassed 200 wells, but once the data was condensed into a usable format with sufficient production ranges, the final well count was 93. Having this number of wells significantly increased the difficulty of the machine learning process due to the high probability of error. Model training and calibration increases accuracy with number of clean samples available, but extra precautions need to be taken to perform the operations with limited data. Limited data cases also reduce the total number of model inputs allowed to be used. This study used a total of 10 well, completions, and stimulation parameters to train the model over the 93 cases. In addition to the lack of input parameters allowed to be used, data processing techniques are hindered by the sparse number of samples within the dataset. PCA analysis provides a very unique set of input data to train AI models, but require more usable data to be applicable. The predictive model results from the normalized model generated seemed to provide a more in depth looks at the quality of completion and stimulation designs due to the reduction of the wells dimensionality for a more even comparison. The results from the normalized model seemed to provide a more accurate representation of what was expected for certain parameters, and provided informative insight on other, more uncertain parameters.

In conclusion, using an AI predictive model is a very effective and efficient solution to the complex nature of unconventional wells optimization. Sufficient data needs to be available for a potential reservoir in order for more successful and accurate models. This study was performed on a field with a small amount of available data for analysis but was successful due to the careful and meticulous care taken during the fundamental processes. AI predictive models are very beneficial to the oil and gas industry and can be used to solve the complex questions of unconventional shale formation production.

References

- Belyadi, H., Fathi, E., & Belyadi, F. (2017). *Hydraulic Fracturing in Unconventional Reservoirs*. Cambridge: Elsevier Inc.
- Choy, K., Lee, W., & Lo, V. (2003). Design of an Intelligent Supplier Relationship Management System; A Hybrid Case Based Neural Network Approach. *Expert Systems with Applications*, 225-237.
- Kuhn, M., & Johnson, K. (2016). *Applied Predictive Modeling*. New York: Springer Science+Business Media LLC.
- Minnaar, A. (2015, September 6). *Implementating the DistBelief Deep Neural Network Training Framework with Akka*. Retrieved from alexminnaar.com:
<http://alexminnaar.com/implementing-the-distbelief-deep-neural-network-training-framework-with-akka.html>
- Mohaghegh, S. D. (2017). *Shale Analytics: Data-Driven Analytics in Unconventional Resources*. Cham: Springer International Publishing.
- Wang, Y., & Salehi, S. (2015). Drilling Hydraulics Optimization Using Neural Networks. *Society of Petroleum Engineers* (p. 13). The Woodlands: Society of Petroleum Engineers.

Appendix A – Feature Selection

120 D

	L1 - FI Pre	L1 - FI Post
MaxMD	1.0000	1.0000
Lateral Length	0.9520	0.9758
Stages	0.7499	0.8050
Flush Actual	0.6453	0.5943
AVG Treating Rate	0.5812	0.5199
ATP	0.5148	0.5314
BoundnessFinal	0.5114	0.4536
Proppant/Ft	0.1682	0.1332
Cluster Spacing	0.0905	0.1956
Total Shots	0.0866	0.1727

Table 10 - 120 Day Feature Importance Scores and Ranking Table, Pre and Post Modeling

Figure 23 - 120 Day Feature Scores Graph, Pre and Post Modeling

150 D

	L1 - FI Pre	L1 - FI Post
MaxMD	1.0000	1.0000
Lateral Length	0.9626	0.9480
Stages	0.8664	0.8452
Flush Actual	0.6978	0.6273
ATP	0.6610	0.7238
AVG Treating Rate	0.5756	0.4584
BoundnessFinal	0.5289	0.4841
Cluster Spacing	0.2395	0.3040
Proppant/Ft	0.1538	0.1010
Total Shots	0.1111	0.1924

Table 11 - 150 Day Feature Importance Scores and Ranking Table, Pre and Post Modeling

Figure 24 - 150 Day Feature Scores Graph, Pre and Post Modeling

Appendix B – Model Training Results

120D

Figure 25 - 120 Day Model Training Loss Curve

Figure 26 - 120 Day Model Training Results Cross-Plot

Figure 27 - 120 Day Model Testing Results Cross-Plot

Figure 28 - 120 Day Model Output Vs. Field Output Trend Confirmation

150 D

Figure 29 - 150 Day Model Training Loss Curve

Figure 30 - 150 Day Model Training Results Cross-Plot

Figure 31 - 150 Day Model Testing Results Cross-Plot

Figure 32- 150 Day Model Output Vs. Field Output Trend Confirmation

90 Day Normalized

Figure 33 - 90 Day Normalized Training Loss Profile

Figure 34 - 90 Day Normalized Model Training Accuracy Results

Figure 35 - 90 Day Normalized Model Training Accuracy Results

Figure 36 - 90 Day Normalized Model Output Trend Analysis

Appendix C – Type Curve Results

Original Models

Figure 37 - 90/120/150 Day ATP Type Curve

Figure 38 - 90/120/150 Day ATR Type Curve

Figure 39 - 90/120/150 Day Boundness Type Curve

Figure 40 - 90/120/150 Day Cluster Spacing Type Curve

Figure 41 - 90/120/150 Day ATP Flush Amount Curve

Figure 42 - 90/120/150 Day Proppant per Foot Type Curve

Figure 43 - 90/120/150 Day Total Shots Type Curve

Normalized Model

Figure 44 - 90 Day Normalized Type Curve, Shots/Ft

Figure 45 - 90 Day Normalized Type Curve, Fluid/Ft

Figure 46 - 90 Day Normalized Type Curve, Cluster Spacing

Figure 47 - 90 Day Normalized Type Curve, ATR

Appendix D – Production Profiles

90D Model

Figure 48 - 90D Production Profile Example, Bad Production

Figure 49 - 90D Production Profile Example, Poor Production

Figure 50 - 90D Production Profile Example, Avg Production

Figure 51 - 90D Production Profile Example, Good Production

120D

Figure 52 - 120D Production Profile Example, Good Production

Figure 53 - 120D Production Profile Example, Avg Production

Figure 54 - 120D Production Profile Example, Poor Production

Figure 55 - 120D Production Profile Example, Bad Production

150D

Figure 56 - 150D Production Profile Example, Good Production

Figure 57 - 150D Production Profile Example, Avg Production

Figure 58 - 150D Production Profile Example, Poor Production

Figure 59 - 150D Production Profile Example, Bad Production

Appendix E – Refracture Candidacy

Well #	Field BOE	P10	P50	P90	P50 Score	P90 Score	P50_Rank	P90_Rank	Overall Avg	Overall Rank
72	21669.01817	55501.92346	171559.8642	308521.2204	-149890.846	-286852.2022	1	1	1	1
11	37050.17933	55668.91468	167569.9139	316773.0909	-130519.7345	-279722.9116	4	2	3	2
54	51003.26233	57743.37319	185430.4133	328334.1974	-134427.151	-277330.935	3	3	3	3
47	54902.83483	76726.22771	191849.9384	326610.0438	-136947.1035	-271707.2089	2	4	3	4
41	46286.00367	52467.59795	175947.4015	317434.5472	-129661.3979	-271148.5436	6	5	5.5	5
30	35210.0515	55955.39689	165593.407	303039.2147	-130383.3555	-267829.1632	5	6	5.5	6
58	63376.5315	60969.17398	185590.5501	330139.475	-122214.0186	-266762.9435	8	7	7.5	7
87	34848.73767	54712.39314	160253.4389	295263.4103	-125404.7012	-260414.6726	7	10	8.5	8
61	43621.445	45910.93926	164708.4552	305540.3444	-121087.0102	-261918.8994	9	9	9	9
92	48937.47133	49908.49251	166548.2182	312410.6055	-117610.7469	-263473.1342	11	8	9.5	10
80	61329.39417	55018.2764	174639.5084	319710.0974	-113310.1142	-258380.7032	12	11	11.5	11
31	42604.615	47839.48404	163270.9988	294252.8833	-120666.3838	-251648.2683	10	17	13.5	12
39	39325.79717	45796.06963	149768.4106	293922.7794	-110442.6134	-254596.9823	15	15	15	13
84	56308.502	55246.73456	163389.826	313926.3025	-107081.324	-257617.8005	18	13	15.5	14
40	46804.63083	48381.37745	158199.8837	297713.7983	-111395.2529	-250909.1675	13	18	15.5	15
25	45929.02033	52181.05038	152040.8192	303970.1758	-106111.7989	-258041.1554	20	12	16	16
19	52962.252	42486.95892	163745.5292	302803.6597	-110783.2772	-249841.4077	14	20	17	17
88	59218.75917	52139.04132	164563.3763	315346.9941	-105344.6171	-256128.2349	23	14	18.5	18
59	47665.09167	41616.68573	153493.9512	301033.6212	-105828.8595	-253368.5296	21	16	18.5	19
7	62370.6785	52306.0283	172748.4494	310861.6513	-110377.7709	-248490.9728	16	22	19	20
62	51235.1695	53570.85515	156775.5205	301684.6119	-105540.351	-250449.4424	22	19	20.5	21
79	47321.54583	45796.10249	153520.3524	295766.4638	-106198.8066	-248444.918	19	23	21	22
68	51450.60783	37567.30061	156245.5674	300928.5858	-104794.9596	-249477.978	25	21	23	23
73	72453.7535	60665.37074	179582.8775	316292.0606	-107129.124	-243838.3071	17	29	23	24
23	54233.91317	47037.87881	158546.302	301082.6552	-104312.3888	-246848.7421	26	26	26	25
86	60993.05733	50766.22	166000.7682	304925.3406	-105007.7108	-243932.2832	24	28	26	26
13	63521.25033	53591.8293	165686.4651	310698.5307	-102165.2147	-247177.2804	28	25	26.5	27

Table 12 - 120 Day Refracture Candidacy Results

Well #	Field BOE	P10	P50	P90	P50_Score	P90_Score	P50_Rank	P90_Rank	Overall Avg	Overall Rank
85	72125.47483	79639.84036	210119.9234	345282.0192	-137994.4486	-273156.5444	1	1	1	1
53	41280.16633	47461.5156	137899.3426	238522.976	-96619.17624	-197242.8097	2	6	4	2
32	93040.17333	70420.82011	180613.5343	290798.8346	-87573.361	-197758.6613	4	5	4.5	3
66	25643.74283	14835.22036	113987.1948	221619.6282	-88343.45198	-195975.8853	3	7	5	4
19	94832.75817	59398.13926	170976.6639	292889.4304	-76143.90573	-198056.6722	10	4	7	5
49	74535.868	54677.59962	158256.9686	259295.6766	-83721.10059	-184759.8086	6	12	9	6
9	40442.61933	27459.47717	127444.5132	222973.9474	-87001.89391	-182531.3281	5	14	9.5	7
23	123226.5138	65691.95511	189664.7103	330226.6345	-66438.19654	-207000.1207	19	2	10.5	8
72	70274.5495	36381.00961	138260.8201	256910.0177	-67986.27056	-186635.4682	14	10	12	9
7	65674.851	45363.35619	143853.2486	244431.4139	-78178.39762	-178756.5629	8	17	12.5	10
39	44732.25817	31142.01239	125229.0657	220349.4344	-80496.80758	-175617.1762	7	22	14.5	11
76	88261.83717	61411.00499	155208.4568	267145.5773	-66946.61963	-178883.7401	16	16	16	12
60	69256.0055	49097.89673	138606.5982	246592.4821	-69350.59267	-177336.4766	13	19	16	13
24	115836.0243	72188.09782	181740.0792	299593.8992	-65904.05486	-183757.8749	20	13	16.5	14
22	78621.144	57379.07011	156224.9684	252147.0332	-77603.82441	-173525.8892	9	24	16.5	15
16	93942.55717	58402.7457	163408.9235	269419.5612	-69466.36629	-175477.004	12	23	17.5	16
18	137339.5107	65676.02358	194272.0185	338834.413	-56932.50784	-201494.9023	33	3	18	17
45	97412.50883	60907.78124	162648.2355	276605.5003	-65235.72669	-179192.9915	22	15	18.5	18
86	51169.81167	28653.49409	123642.6073	222896.8321	-72472.79564	-171727.0204	11	26	18.5	19
44	131242.1738	66430.68398	184457.1688	316655.9522	-53214.99503	-185413.7784	36	11	23.5	20
89	57843.40367	10607.7349	118382.3134	235006.4866	-60538.90972	-177163.0829	28	20	24	21
51	73917.66967	27725.68718	138358.9601	246551.424	-64441.29047	-172633.7543	24	25	24.5	22
1	78606.36167	41061.30253	143619.3411	248474.1934	-65012.97944	-169867.8317	23	28	25.5	23
91	78464.56983	42455.79546	143877.9757	246188.1115	-65413.40584	-167723.5417	21	30	25.5	24
65	166112.5431	88710.02765	214953.1893	353327.4747	-48840.64622	-187214.9316	43	9	26	25

Table 13 - 150 Day Refracture Candidacy Results