URTeC: 598

Eagle Ford Fluid Type Variation and Completion Optimization: A Case for Data Analytics

Fahd Siddiqui[*1], Ali Rezaei[1], Birol Dindoruk[1,2], Mohamed Y. Soliman[1]; 1. University of Houston, 2. Shell International Exploration and Production Inc.

Abstract

Prior knowledge of reservoir fluid type and properties aids in selecting and optimizing completion and surface facilities. Fluid properties prediction has an impact on in-place volumes and reservoir performance management including optimized well placement. We present a data-driven fluid variation modeling approach using machine learning. The aim is to predict the fluid type and oil API gravity for a given location and depth and optimize the completion design for the Eagle Ford shale.

Data from 9400 Eagle Ford shale wells were compiled, cleaned, and analyzed. Data was then divided into training and test sets. The test set was set aside for validation to prevent any training bias. Data visualization and statistical analysis was carried out, which revealed patterns and features within the training data. Three separate artificial neural networks (ANNs) were then constructed on those features, and a supervised learning algorithm was employed to train on the training set.

The first ANN predicts the oil API gravity based on a given coordinate: latitude, longitude and depth information. This network uses Mean Squared Error (MSE) loss function with the Root Mean Squared (RMS) regression optimizer. ANN-1 reported an error of 2.4 API which is well within process dependency of the API measurements and within the potential experimental errors. The second ANN predicts the most likely fluid type along with the probability, which can be used as a measure of confidence. ANN-2 uses the categorical cross-entropy loss function with the Adam optimizer (Kingma (2014)). Finally, ANN-3 predicts the hydrocarbon production of the first 12 months based on the well location, lateral length, depth, number of stages, proppant volume and gel volume. All three models were then validated on the test set, and a good match was obtained. Based on the data-driven models, an optimization scheme was created to maximize cash flow from the first 12 months of production based on varying the lateral length, the number of stages, proppant volume, and gel volume used. The resulting optimum parameters are then represented visually on the map of Eagle Ford, along with oil and gas production, and cash flow.

Even though the presented method was trained for Eagle Ford, data from other formations can be incorporated and re-trained, including other proxies for every additional basin, to create a general neural network predictive model on all formations; or to create smaller networks that would make accurate predictions within the specified formation. This approach will lead to a continuously improving and learning process for each additional field and play.

Introduction

Unconventional reservoirs play the central role in the increased US oil and gas production in the last decade. The chances of developing these reservoirs depend mainly on the formation properties, fluid type, and how the horizontal wells are completed. Prior knowledge of reservoir fluid type and properties aids in selecting and optimizing completion and surface facilities. Fluid properties prediction has an impact on in-place volumes and reservoir performance management including optimized well placement. Formation properties dictate the type of completion, while the fluid type plays a crucial role in selecting proper completions and surface facilities. Therefore, the two ends of successful field development are surface facility and hydraulic fracturing design that is an inevitable part of the completions in unconventional reservoirs. In this paper, we utilize a data-driven approach to optimize the two ends of production from the Eagle Ford shale. We aim to create artificial neural network (ANN) models to predict the fluid type, API, and the first year production of oil and gas in the Eagle Ford shale. We base our models on the parameters that can be selected before drilling the well. These parameters include latitude and longitude, TVD, proppant volume, etc.

Physics-based modeling and data-based approaches are standard practices for building predictive models that can be used for both forecasting the production and optimizing the completion design. The physics-based approach includes analytical or numerical solutions of the problem an often is based on several simplifications. The most commonly used analytical models to analyze formations are Decline Curve Analysis (DCA) and Rate Transient Analysis (RTA). On the other hand, the data-based approach includes correlations, cross-plots, analog fields, etc. and are heavily dependent on the availability of data. The advantage of the physics-based models is that it gives better predictions, but often requires significant number of parameters, and fails in several cases due to the uncertainties in in-situ rock and fluid properties. While the data-based approach does not require as many parameters, yet may fail in the case of extrapolation, it also requires a considerable amount of example data to train the models. These models have gained popularity in recent years as a result of a massive increase in the amount of produced data (big data). The produced data can be utilized to overcome the shortcomings of the data-based approach. Thus, it has now become feasible to train a data-based model to create reliable tools that can be used for prediction of oil and gas production. Data-based approaches can be grouped into two main classes: supervised learning and unsupervised learning. Supervised learning is the task of building a function from a set of labeled input data, whereas unsupervised learning is used to find hidden structures of unlabeled data. The most popular data analyses are least square regression, gradient descent, support vector regression, random forest, genetic algorithm, gradient boosting models, artificial neural network, self-organizing maps.

Using artificial neural network (ANN) models in the oil industry has increased in the last decade. The literature contains a wide variety of models including supervised and unsupervised models. A review of the commonly used data analytics models and their application in the oil and gas industry was presented by Mishra (2017). Usually, in these types of studies the production (6 months or 12 months) is considered as the model output. The follwoing are a few examples of using the data-driven approach in the oil and gas

industry. Shelley et al. (2012) constructed an artificial neural network model to forecast production. Gupta (2014), used neural network (NN) and time series analysis techniques for predicting the performance of shale gas wells in unconventional reservoirs. In their method, they used the production data of the previous year as the inputs of the model.

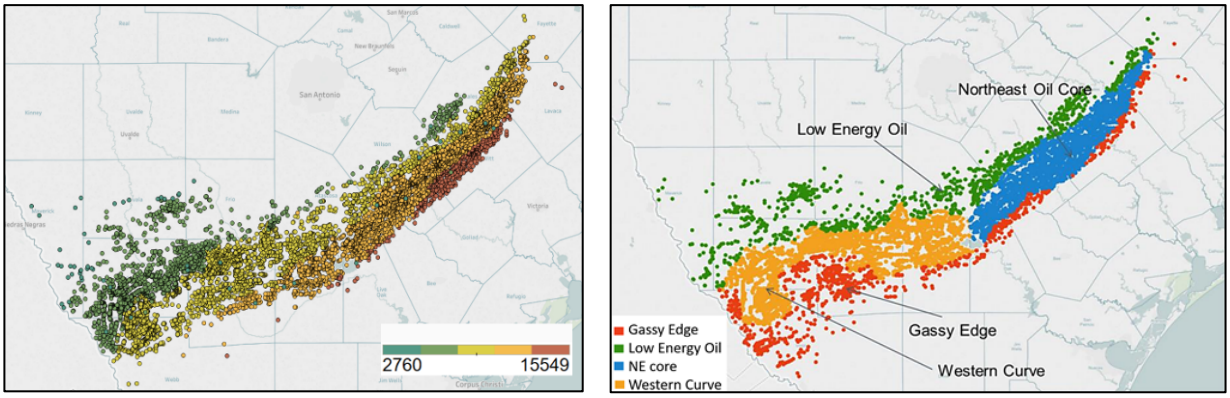
Nejad et al. (2015) used an ANN model to optimize the well completion strategy of three operators in Eagle Ford Shale. Zhong et al. (2015) compared the performance of several methods, including Ordinary Least Square Regression (OLSR), Support Vector Regression (SVR), Random Forest (RF) and Gradient Boosting Model (GBM), on the Wolfcamp shale data set. They concluded that the Random Forest performs better in terms of prediction. Bhattacharya et al. (2016) used various supervised and unsupervised techniques on Bakken and Marcellus data sets to identify the geological trends. They found that the support vector machine (SVM) works better for lithofacies classification. Mohaghegh (2016) presented a workflow for selection of refrac candidate selection using data analytics. A so-called "Shale Analytics" for using data analysis approach for shale reservoirs was introduced later by the same author (Mohaghegh et al., 2017). Repchuk et al. (2018) developed a decision forest regression model to forecast production in the Denver Basin. Cai et al. (2018) proposed a nonparametric smoothing model to analyze the well performance in Bakken shale. Luo et al. (2018) built a predictive ANN model based on geologic inputs to forecast the production from Middle Bakken shale.

In this study, we present three predictive artificial neural network (ANN) models to address some of the challenging questions that an operator may face regarding the designing of surface facilities and optimization of the stimulation in the Eagle Ford shale. We aim at using the inputs that can be obtained without drilling the well. The models that are presented in this study include different ANNs to predict the oil API, fluid type, and first year production from the prospective well. The ANNs are constructed using the data from 9400 wells in the Eagle Ford formation. Following section provides a summary of the Eagle Ford formation characteristics. After that, we describe the pre-processing and the steps we took to make the data ready for the analysis. Then, in the discussion and results section, we present the three artificial neural network models. ANN-1 is a model that is built to predict the oil API in the reservoir. ANN-2 predicts the fluid type base on the location. Also, ANN-3 forecasts the production of oil and gas based on completion and hydraulic fracture parameters. Moreover, a production optimization analysis is presented to help operators optimize the parameters that affect the production from Eagle Ford shale. Finally, the conclusions are summarized in the last section.

Characteristics of the Eagle Ford Shale

Eagle Ford is a major unconventional oil and gas play located in the southeast of Texas. This formation extends from the Mexican border to Leon County on the North of Houston. Eagle Ford covers 30 counties in Texas in an area about 50 miles by 400 miles. Eagle Ford is both oil and gas play and dips from northwest to southeast (5000 ft. to 13000 ft.). Figure 1a shows the TVD variation in Eagle Ford.

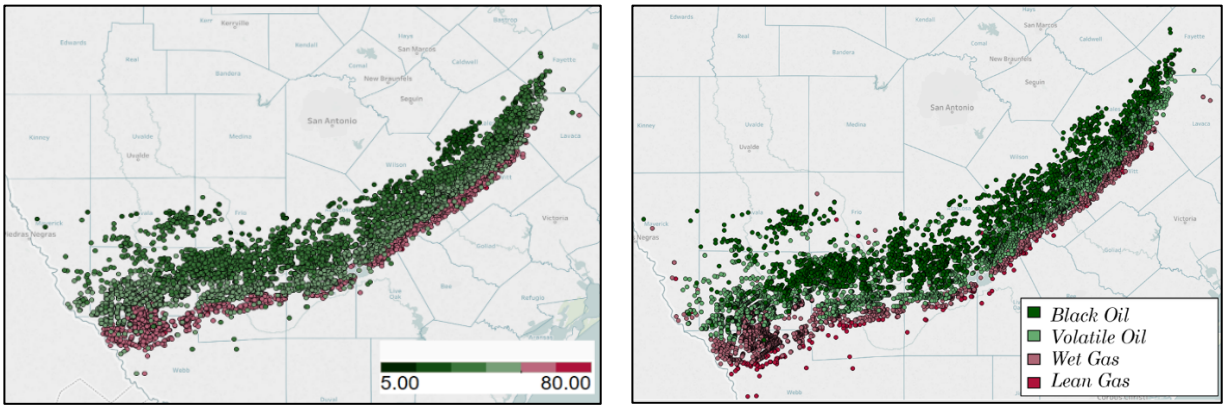
The formation has four sub plays: Northeast Oil Core, Low Energy Oil, Gassy Edge, and Western Curve (EIA, 2016). Figure 1b shows the location of these four sub plays. Production from each of the sub plays integrated with several challenges and cost issues. Operators observe varying characteristics of the produced fluid across the play and different types of fluids in different areas. It is also necessary to perform local area reservoir characteristics analysis for hydraulic fracturing, which yet results in the success of stimulation in one area and failure in another (Mullen, 2010).



(a) (b)

Figure 1. Location of Eagle Ford: (a) the TVD variation [ft], (b) sub plays (EIA (2016))

Figure 2 shows the variation of the fluid API gravity and fluid type distribution in the reservoir. As can be seen in Figure 2a, the fluid API changes from oil in the northwest (low energy oil sub play) to gassy oil and wet gas in the central region of Eagle Ford (northeast oil core and western curve sub plays). The fluid type is mostly lean gas in the southeast part of the reservoir, which is also the deepest part (gassy edge). The definition of the different fluid in Eagle Ford shale is based on the percentage of the liquid. Oil, gassy oil, wet gas, and lean gas are fluids that contain more than 90%, 50-90%, 10-50%, and less than 10% liquid respectively. About 55% of the reservoir hydrocarbon production is gassy oil. Oil and wet gas form about 43% of the production with 22% and 21% each, respectively. The least amount of production comes from lean gas with about 2%.



(a) (b)

Figure 2. Variation of well and fluid API gravity in Eagle Ford: (a) API gravity, (b) fluid type variation and distribution

Pre-Processing the Data and Methodology

Pre-processing includes data cleaning, feature selection, splitting the data to training and test sets, and normalization of the data. Data cleaning includes filtering out the missing data, unexpected non-numeric data, and non-meaningful zeros from the dataset. After this step, a range for each variable was considered and only values within that range were filtered. The ranges of the values for the selected parameters in the following ANNs are reported in Table 1. Please note that the data that we used in this study are based on a

report from 2014-2015. So, some of the values that are reported in the table have changed since then. For example, the frac stages that are currently selected are higher than the values presented in the table. One may run a similar analysis to the one presented in this paper on a more recent data to obtain a more updated results from the model.

Table 1. Selected range of input and output for training of ANN-3

Parameter		Min	Max
Input	Lateral Length (LL) [ft]	2500	7500
	Proppant Volume (PV) [lb.]	2e6	1e7
	Frac Fluid Volume (FFV) [gal]	1e6	1e7
	Frac Stage (FS)	5	25
Output	Oil [BBLD]	50	600
	Gas [MCFD]	50	2000

After the data is cleaned, a proper set of variables needs to be selected that have effects on the model output. For example, if a model is generated to predict the production, parameters such as TVD, location, and other completion elements are chosen as appropriate input variables because they have the most significant impact on the production. In this study, we selected different inputs for each model depending on the output of the model.

Splitting the Training and Test Sets

We split the data as follows: 75% (~6,600 wells) of the data was allocated as the training set (TrSet-1) and the rest as the test set (TsSet-1). TsSet-1 was set aside for verification purposes only, and it was ensured that it is not used for training or validation to avoid bias. From the TrSet-1, 5% was allocated as the validation set (ValSet-1) for use during the active training of the models. The purpose of the validation set is to avoid overfitting the data.

A total of 3453 gas well and 5492 oil well were remained after cleaning the data, out of which 2417 and 3844 wells were selected for training set TrSet-2 and TrSet-3 respectively.

Because of missing data, number of wells with complete data for ANN-3 was only ~3500 wells for gas and ~5500 wells for oil. The training and test sets were split according to the same ratio described above giving a TrSet-2 consisting of 2500, TsSet-2 with 875, and ValSet-2 of 125 gas wells. Similarly, for oil wells the split was TrSet-3 with 3920, TsSet-3 with 1375, and ValSet-3 of 205 points.

Normalization

Because the data consists of various ranges, the training process can be expedited if the data are normalized using the mean and standard deviation. Therefore, the mean and standard deviation of the training set was used to normalize both, training and test sets. Because the models will only predict based on normalized data, the mean and standard deviation of the training sets (i.e., TrSet-1, TrSet-2, and TrSet-3) are saved to normalize all future unseen input data to get the correct prediction.

Methodology

The methodology used in this manuscript are also used in typical ANN exercises and are found in the common literature, we skipped their definitions here and include them in Appendix A.

Discussion and Results

Rock and fluid properties change from a point to another point within a formation. The variation is more severe in the Eagle Ford because of the changes in the formation depth. The depth variation affects the reservoir temperature (geothermal gradient), which in turn affects the hydrocarbon maturity. Therefore, depending on the location of the well and the target depth, fluid properties will vary and appropriate design for surface facilities and well completion is required. To predict the oil API gravity and the fluid type for a given location and depth, ANN-1 and ANN-2 are created. ANN-3 is created to predict the oil and gas productions for the first 12 months, for that location, depth, and some other completion parameters which will be explained later.

ANN-1: Oil API model

The variation of oil API of Eagle Ford Shale is visually demonstrated in Figure 2a. ANN-1 was constructed as a regression type artificial neural network with the architecture shown in Figure 3a. The input layer consists of bottom hole location information such as latitude, longitude, and true vertical depth (TVD). The two hidden layers are of type Rectified Linear Unit (ReLU) with active L2 regularization, and with 512 nodes for each layer. Finally, the output layer consists of one node: oil API gravity. ANN-1 was trained on TrSet-1 using the root mean squared prop (RMSProp) optimizer, the mean square error loss function, and mean absolute error as the reporting metric.

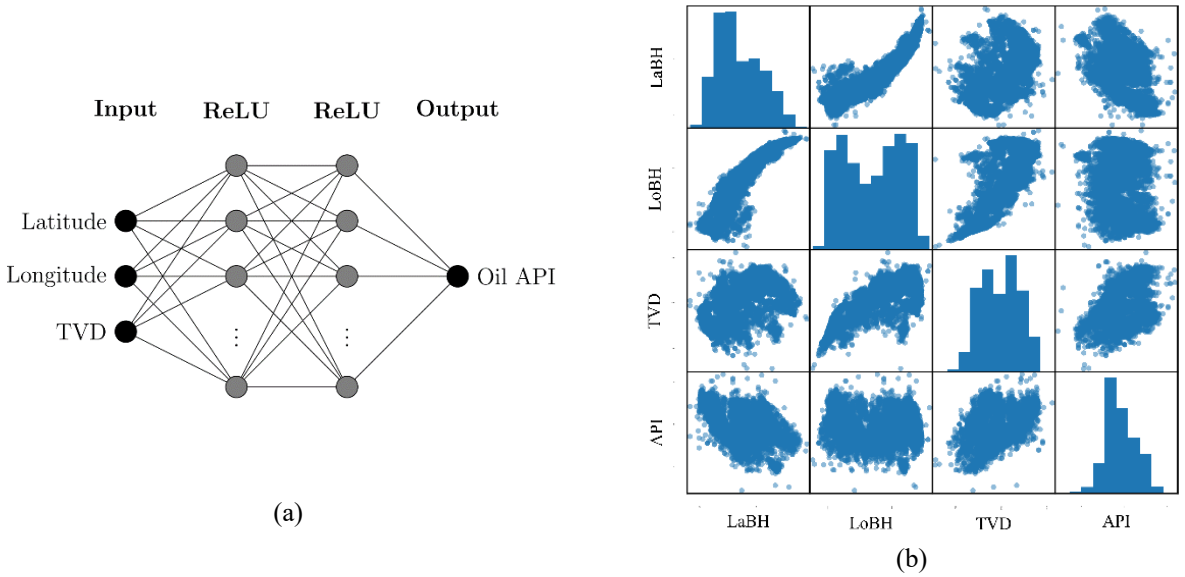


Figure 3. Model variables. (a) Oil API ANN-1 Architecture and (b) Scatter Matrix Plot: Relationship of Location Parameters to Oil API Gravity

Figure 4 shows the performance of the model on the TsSet-1 by plotting actual vs predicted oil API, with fluid types marked for enhancing the understanding of deviation of the data. A 45-degree diagonal line through the scatter plot would mean accurate prediction, and ideally, all data points should lie on that line. Data points to the right of the line indicate to over-prediction and data points to the left indicate under-prediction by the model. In this case, the mean absolute error is about ~2.7 API, which is reasonably close considering the model is based purely on location information. Understandably, the model performs poorly for lean gas wells because of suspected measurement errors in the data (oil API vs. condensate API). Therefore, it is suggested to use caution while predicting oil API on gas wells.

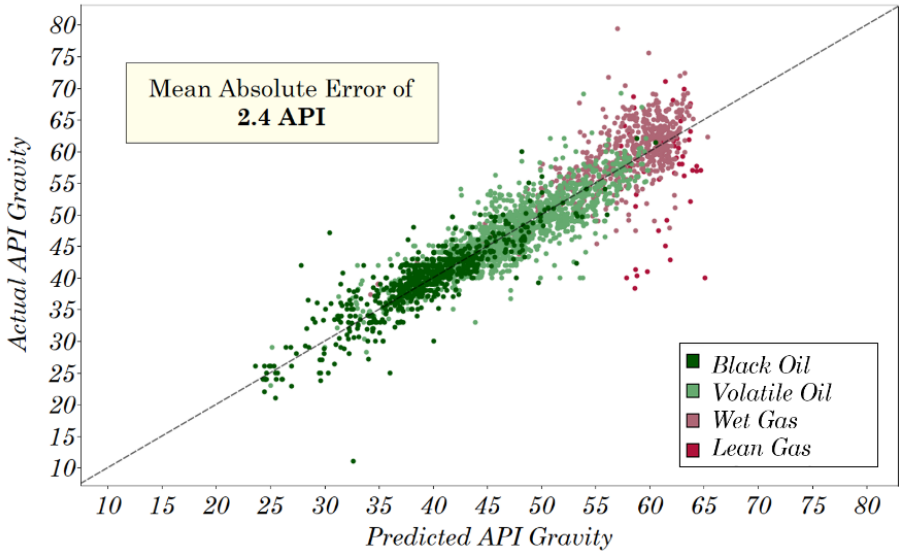


Figure 4. Actual vs Predicted Plot for Oil API with Fluid Types Marked

ANN-2: Fluid type model

Fluid properties of the Eagle Ford Shale vary from lean gas to black oil based on the location of the well and its depth (Figure 2b). The fluid type model (ANN-2) was trained to predict the fluid type probability in the Eagle Ford shale given the location information. Figure 5 shows the architecture of the artificial neural network that was used. ANN-2 takes three bottom hole location inputs: latitude, longitude, and TVD. The two hidden layers are of type ReLU with L2 regularization and consist of 512 nodes each. The final layer is of softmax type containing four output categories. The four categories of fluid type were chosen based on the criteria described by Dindoruk (2012). ANN-2 predicts the most likely fluid type at a given location and can help operators design completion and surface facilities.

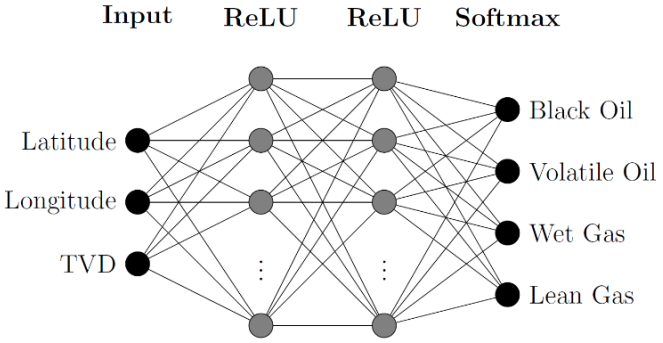


Figure 5. Fluid Type ANN-2 Architecture

Figure 6 shows the confusion matrix table for ANN-2 multiclass classifier. The rightmost column gives the sum of the number of samples predicted for each class. The bottom-most row gives the totals of the actual number of samples in the test data classes. The leading diagonal shows the True Positives for each class. Off-diagonals in each row are the False Positives, while the off-diagonals in each column are False

Negatives with respect to each class. Consider the case for Gassy Oil: the recall is 84%, and the precision is 79%. The overall accuracy of the model is 77% with a misclassification rate of 23%. Even though there is some misclassification, overall a reasonable certainty is obtained in predicting the type of fluid at a particular location in Eagle Ford.

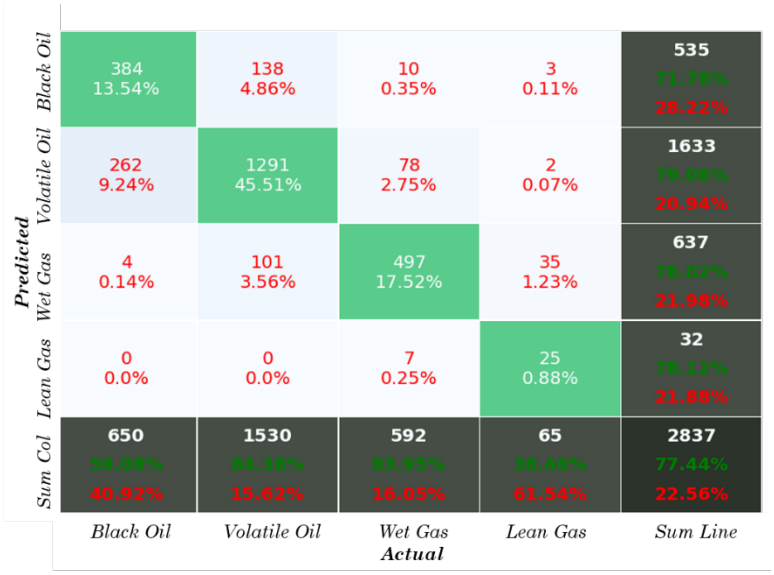


Figure 6. ANN-2 Confusion Matrix

ANN-3: Production forecast and completion optimization

A reliable model for predicting the production of hydrocarbon based on the location of well and completion variables can be constructed before drilling the well to help operators better decide on the most profitable parts of the play and optimize the well completion design for maximum profit. For this purpose, ANN-3 is built to predict the production of oil and gas from Eagle Ford based on some parameters that can be selected before drilling. These parameters are the lateral length (LL), TVD, fracture stages (FS), fracturing fluid volume (FFV), bottom hole latitude (LaBH), bottom hole longitude (LoBH), and proppant volume (PV). Figure 7 shows the model inputs and structure of ANN-3. Our aim is to predict the production of oil and gas for the first year of the well's life. Two such models are trained separately for oil and gas, each with a total of two hidden layers with 512 nodes as before. ANN-3 for gas was trained on TrSet-2, and ANN-3 for oil was trained on TrSet-3 using the RMSProp optimizer, the mean square error loss function, and mean absolute error as the reporting metric.

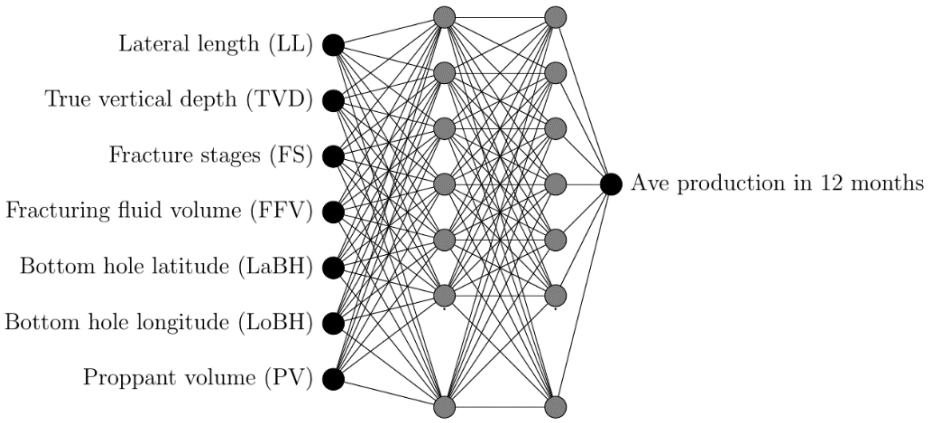


Figure 7. ANN-3 Architecture

Figure 8 shows the scatter plot of the parameters used for ANN-3. As can be seen, among all the variables only longitude and latitude are highly related to each other, which is expected. Some relationship was also observed between TVD and bottom hole longitude (LoBH). Furthermore, we interpolated the missing reported values of fracture stage (FS) based on the well's lateral length using a linear regression model. As a result, there is some relationship between these two values in the scatter plot matrix (Figure 8).

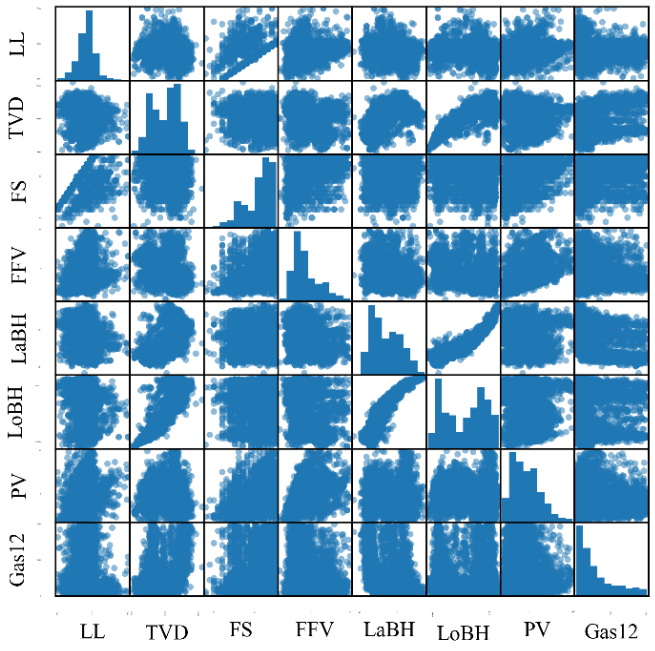
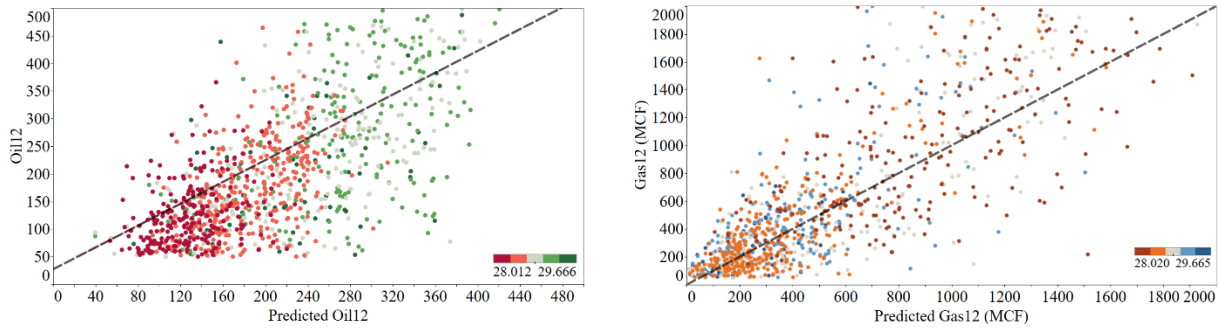


Figure 8. Scatter Matrix Plot for ANN-3 model

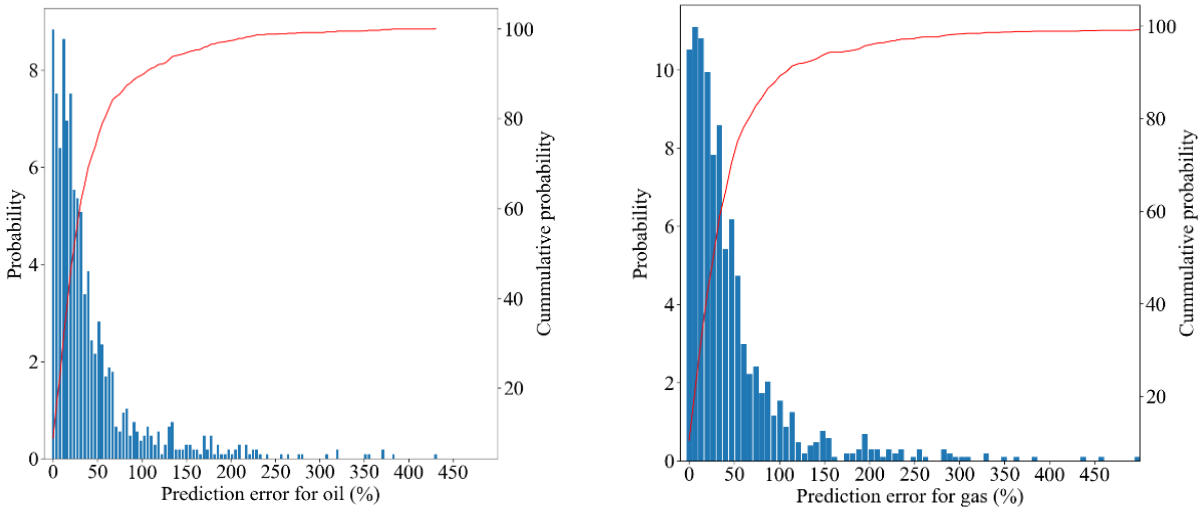
The result of model performance on the test set for oil and gas production are depicted in Figure 9. The figure is color-coded with the longitude of the wells. The model performs better for the points located on the oil bearing zone of the Eagle Ford shale. Also, both oil and gas ANN-3 models perform much better for predicting oil and gas production for oil productions of less than 240 BBLD and gas production of less than 600 MCFD. The errors associated with oil model for TrSet-3 and TsSet-3 were 29% and 32%, and for the gas model on TrSet-2 and TsSet-2, it was 30% and 33%. Figure 9 shows the oil and gas production predictions on the test sets as a function of latitude.



(a) (b)

Figure 9. Model performance as a function of Latitude, a) one-year oil production, b) one-year gas production

An analysis of the reported error by the model is in order. Figure 10a shows the error distribution and probability for oil model. Blue bars in the figure show the probability of the error associated with the model predictions and the red line shows the cumulative error. As can be seen, the probability of predicting oil production with less than 40% error using ANN-3 for both oil and gas models is about 60%. Adding more data and input features to the ANN-3 would increase the accuracy of prediction.



(a) (b)

Figure 10. Error distribution of the model, a) Oil model, b) Gas model

Production Optimization and Economic Analysis

An optimization-based on hydraulic fracturing variables, well lateral length, TVD, and the location is performed to explore the profitability of the Eagle Ford shale. The constraint parameters are drilling and stimulation expenses, and price of oil and gas. Our goal is to use ANN-3 to maximize cumulative cash flow for 12 months, subject to LL, PV, FV, and FS. We selected 2000 linearly distributed locations in Eagle Ford for the optimization analysis. The points were chosen to ensure most of the Eagle Ford is covered. For each point, we used its longitude and latitude, and created another artificial neural network model to predict the associated TVD for the point. The purpose of this model was to constrain the TVD to produce intervals only. Figure 11 shows the performance of the model on the test data, which is relatively accurate.

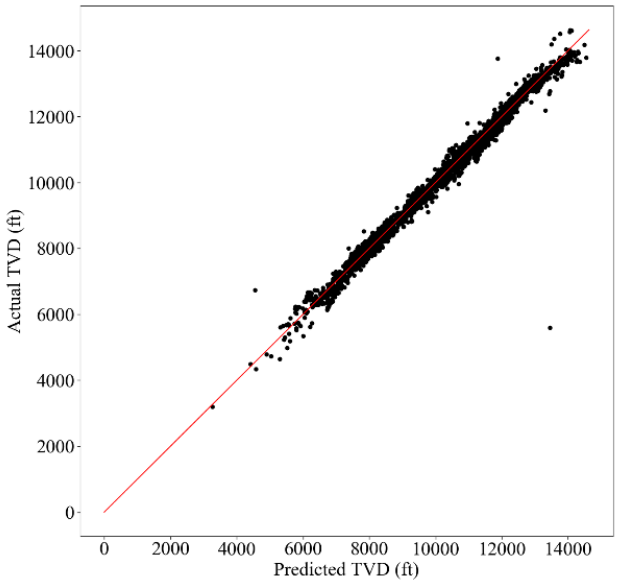


Figure 11. Predicted vs. actual value of the TVD for economics model (inputs of this model are latitude and longitude)

We selected a range for the input variables of ANN-3 and used a local sensitivity analysis approach to predict the production of oil and gas using ANN-3 for each point (prospective well). The range was chosen based on the minimum and maximum reported values for each input in our dataset. Then, we looped through all the possible combinations for each point and used ANN-3 to predict the average of first year oil and gas productions. Cumulative cash flow is then computed and the combination with maximum cash flow (undiscounted) is chosen as the optimum for each location point. The cumulative cash flow is given by:

$$\text{Cumulative cash flow} = \text{Revenue} - \text{CAPEX} - \text{OPEX} \quad (1)$$

We used 50 $/bbl and 1.8 $/Mscf as the respective oil and gas prices for calculating the revenue from production (Table B1). Revenue from oil and gas is summed and total revenue is thus computed for each point. Therefore, oil and gas production from each location are considered as input into the revenue stream of the economic model. The primary cost of a well in Eagle Ford is related to drilling, completions, and associated facilities. The typical cost for a well in different sub plays of Eagle Ford is shown in Figure B1 in Appendix B. Based on Figure B1, we assumed a typical CAPEX of well as $7.5 million and prorated this into the cost of drilling as $160 per foot drilled, $0.143 per pound of proppant, and $0.167 per gallon

of fracturing fluid (Table B2). The drilling cost in our analysis is the combination of vertical and horizontal parts of the wellbore together. For the operating expenses (OPEX) we considered the oil transport cost of 5.17 $/bbl, gas processing and transport of 1.25 $/Mscf, and water handling of 2.25 $/bbl (Table B3). Rest of the associated costs are not considered because we assume them to be equal for all location points for this basic economic study. Based on Equation 1, we then calculated the cash flow, and the combination that gave the best cash flow for each point was selected. Figure 12 shows the optimized cash flow map for the Eagle Ford. As can be seen, the North East Core sub play is determined to be the most profitable, and Gassy Edge and Western Curve are the worst profitable regions in the Eagle Ford. This conclusion is inline with the cost of wells in each sub region presented by EIA (2016) in Figure B1. Moreover, oil being the more profitable than gas, ANN-3 predicts more oil production in the North East Core than in the Gassy Edge, and more gas in the Gassy Edge. Thereby reducing the profitability, based on the first year average production from single well.

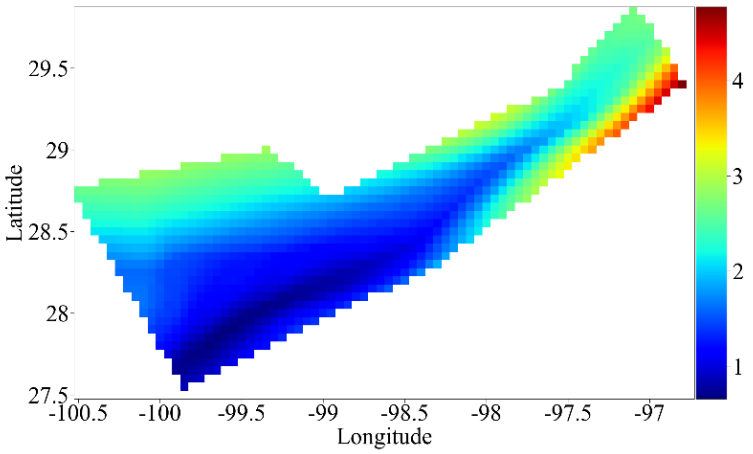


Figure 12. Optimized Cumulative Cash Flow (MM $/ 1st year)

Figure 13 shows the optimized oil and gas production from Eagle Ford shale. The figure is generated using the optimized cash flow in the play. The oil (Figure 13a) and gas (Figure 13b) productions match the reported values. The North East Core is the most profitable oil producing sub play. Also, the Gassy Edge and Western Curve are the best gas producing sub plays.

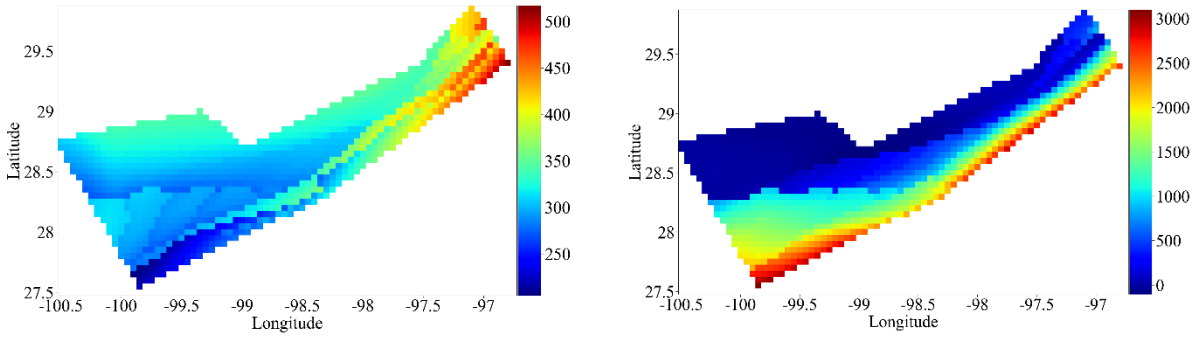


Figure 13. Optimized production from Eagle Ford (a) 12 months average oil production [STBD] (b) 12 months average gas production [MCFD]

Contour plots of the combinations that gives the best cash flow are presented in Figure 14. The optimized lateral length is shown in Figure 14a. The model predicted the maximum and the minimum value of the given lateral lengths for the gassy edge and low energy oil respectively. Also, the optimized lateral length gets longer as moving from the center of the play to the southwest. This trend indicates that the model suggests a longer lateral length for the gassy part of the play and a shorter lateral length in oil-producing sections. Combining this observation with the results of optimized fracturing stages (Figure 14b) yields a better picture of the model suggestions. By this combination, we can conclude that the model suggests the maximum possible frac stages to maximize the production in oil-producing zone. Also, the same results can be observed by combining the results of optimized lateral length and frac stages. The results are different in the gas-producing zone. In the southern part of the zone, the model suggests having longer lateral length and a smaller number of stages. While, in the northern part of the gas producing zone, longer laterals with a greater number of frac stages are suggested. Figure 14c-d show the optimized frac fluid volume and optimized proppant volume respectively. The model suggested the minimum number for the frac fluid volume in most parts of Eagle Ford. A small area on the North section of Eagle Ford is the only region with a moderate optimized frac fluid volume. Also, the model picked the lowest proppant volume for the Gassy Edge region and the maximum given number in the range for the rest of the play. This prediction is to minimize the cost of fracturing in the Gassy Edge, which may not be as profitable as oil.

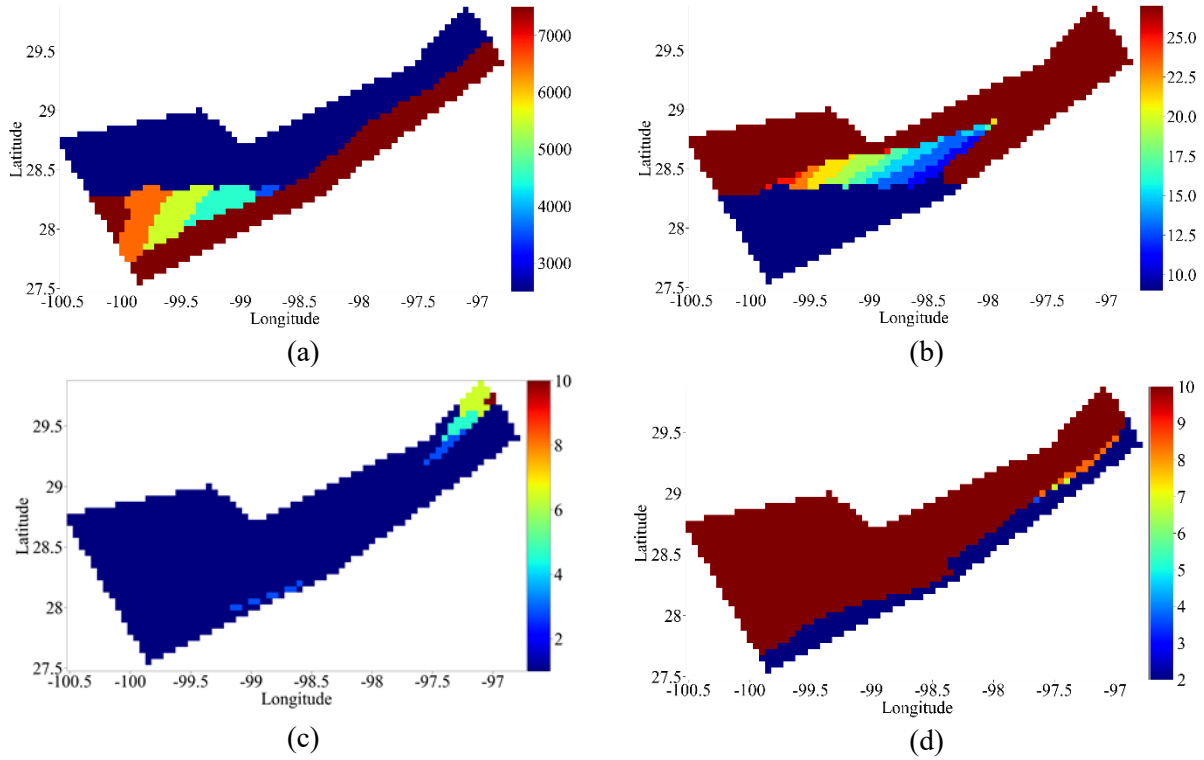


Figure 14. Optimized input variables for the maximum cash flow in Eagle Ford (a) optimized lateral length [ft] (b) optimized frac stages (c) Frac Fluid Volume (Million gal) (d) Optimized Proppant Volume (Million lbs)

Conclusions

Three artificial neural network models are built for analyzing the production data of the Eagle Ford shale. We refer to these models as ANN-1, ANN-2, and ANN-3. We used a data set of 9243 wells to build the models. The data was cleaned and separated into a training and test sets with a ratio of 3:1. ANN-1 is a model to predict the oil API of the wellbore based on the location of the well and TVD. This model has the mean absolute error of 2.4 API. ANN-2 was built to predict the most probable fluid type based on prospective drilling location. ANN-3 is the model for predicting the production of oil and gas from Eagle Ford based on the location and TVD of the well, fracture stages, lateral length, frac fluid volume, and proppant volume. The mean absolute error of the model for predicting the production on the test set was 29% for oil production, and 30% for gas production. The accuracy of the model will improve by adding more features such as geological and petrophysical properties that contribute to the production. Using ANN-3, an economic analysis was performed to optimize the fracture and drilling parameters to obtain maximum cash flow in the first year. For this purpose, 2000 locations were selected and, through a sensitivity analysis, all of the typical combinations for the model inputs were tested and the combination with the maximum cash flow in the first 12 months was selected as the optimized values. This results in best possible combination of fracturing and completion parameters that are suggested based on optimum cash flow. The results showed that the North and North East zones of Eagle Ford are the most profitable locations for drilling and production. The model can be run for optimizing a user defined location as well. The result of this data driven study can be used as an axillary tool along with physics based simulation studies for selecting optimum drilling location, fracture stages, fracture volume and proppant volume in Eagle Ford. This framework may also be used to generate data driven models for other plays.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Man_e, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Vi_egas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorow.org.

Bhattacharya, S., Carr, T.R., Pal, M. (2016). Comparison of supervised and unsupervised approaches for mudstone lithofacies classi_cation: Case studies from the bakken and mahantango-marcellus shale, usa. Journal of Natural Gas Science and Engineering;33:1119-1133.

Cai, Q., Yu, W., Liang, H.C., Liang, J.T., Wang, S., Wu, K., et al. (2018). Development of a powerful data-analysis tool using nonparametric smoothing models to identify drillsites in tight shale reservoirs with high economic potential. SPE Journal 2018;23(03):719-736.

Dindoruk, B. (2012). Development of a Correlation for the Estimation of Condensate to Gas Ratio (CGR) and Other Key Gas Properties from Density Data. Society of Petroleum Engineers. doi:10.2118/160170-MS

EIA (2016). Trends in us oil and natural gas upstream costs. US Energy Information Administration.

Gupta, S. a. (2014). Production forecasting in unconventional resources using data mining and time series analysis. SPE/CSUR Unconventional Resources Conference. Canada: Society of Petroleum Engineers.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2017).Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. p. 6626-6637.

Kellogg, R., Chessum, W., Kwong, R. (2018). Machine learning application for wellbore damage removal in the wilmington _eld. In: SPE Western Regional Meeting. Society of Petroleum Engineers.

Kingma, D.P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980;.

Luo, G., Tian, Y., Bychina, M. and Ehlig-Economides, C., 2018, September. Production Optimization Using Machine Learning in Bakken Shale. In Unconventional Resources Technology Conference, Houston, Texas, 23-25(pp. 2174-2197). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Mishra, S., Lin, L. (2017). Application of data analytics for production optimization in unconventional reservoirs: A critical review. In: Unconventional Resources Technology Conference, Austin, Texas, 24-26 July 2017. Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers; p. 1060-1065.

Mohaghegh, S. D. (2016). Fact-Based Re-Frac Candidate Selection and Design in Shale-A Case Study in Application of Data Analytics. Unconventional Resources Technology Conference (pp. 514--527). San Antonio, USA: Society of Petroleum Engineers.

Mohaghegh, S., Gaskari, R., Maysami, M. (2017). Shale analytics: Making production and operational decisions based on facts: A case study in marcellus shale. In: SPE Hydraulic Fracturing Technology Conference and Exhibition. Society of Petroleum Engineers.

Mullen, J. (2010). Petrophysical characterization of the Eagle Ford Shale in south Texas. Canadian Unconventional Resources and International Petroleum Conference. Society of Petroleum Engineers.

Nejad, A.M., Sheludko, S., Shelley, R.F., Hodgson, T., Mcfall, P.R. (2015). A case history: evaluating well completions in eagle ford shale using a data-driven approach. In: SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

Repchuk, K., Reimchen, A., Gregoris, D. (2018). Using data analytics to maximize value in the denver-julesburg basin. In: Unconventional Resources Technology Conference, Houston, Texas, 23-25 July 2018. Society of Exploration Geophysicists; p. 2848-2853.

Schuetter, J., Mishra, S., Zhong, M. and LaFollette, R., (2018). A Data-Analytics Tutorial: Building Predictive Models for Oil Production in an Unconventional Shale Reservoir. SPE Journal, 23(04), pp.1-075.

Shelley, R.F., Guliyev, N. and Nejad, A., (2012). A Novel Method to Optimize Horizontal Bakken Completions in a Factory Mode Development Program. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Tijmen, T., & Hinton, G. (2012). Lecture 6.5-RMSProp, COURSERA: Neural networks for machine learning. University of Toronto: Technical Report.

Zhong, M., Schuetter, J., Mishra, S. and Lafollette, R.F., (2015), February. Do data mining methods matter?: A Wolfcamp Shale case study. In SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

Appendix A – Definition of the terminology and Methods

The terminology that is utilized throughout the paper are defined here. These terms are standard terms used in data science and may also be found elsewhere.

Learning Rate: Learning rate is a hyper-parameter that controls the amount of correction applied to the weights of the network with respect to the loss gradient. A small learning rate ensures none of the local minima are missed however it results in a slower convergence. A higher learning rate may cause significant changes in network weights and may lead to divergence.

Over-Fitting: Over-fitting happens when model learns signal as well as noise in the training data and therefore will not generalize well on new data on which the model was not trained on.

Cost Function: A cost or a loss function is a measure of how wrong the model is in terms of its ability to estimate the relationship between the predicted value: y_i' and actual ground truth: y_i. Mathematically it can be expressed in general terms as:

$$Loss\left(y_i', y_i\right) = \sum_{i=1}^{n} f\left(y_i', y_i\right) \tag{A1}$$

An appropriate loss function maybe chosen for its desirable properties (convexity etc.) for the problem. Some examples are: mean squared error, mean absolute error, hinge loss, log-cosh loss, categorical cross-entropy and others.

Epoch: An Epoch is when the entire dataset is passed forward and backward through the neural network only once.

Layer: A layer is a collection of neurons; each neuron is connected to neurons from previous and the next layer. As an example, Figure 3a shows a neural network architecture with one input layer, two hidden layers, and one output layer.

Rectified Linear Unit (ReLU) Activation Function: The rectifier is an idea from electrical engineering of half-wave rectification. This function returns the positive part of the argument; and is also known as the ramp function. It is defined by:

$$ReLU\left(x_i\right) = \max(x_i, 0) \tag{A2}$$

Categorical Cross-Entropy Loss: Categorical cross-entropy loss is a function to measure the performance of a classification model whose output is a probability distribution. The cross-entropy loss increases if the predicted probability diverges from the actual label. This loss type penalizes (or increases in value) if the model predicts a wrong value with high confidence. In this case, the categorical cross-entropy loss refers to cross-entropy loss calculated after applying a softmax activation to ensure the loss is being computed on probability distribution of categories. For multi-class classification it is given by:

$$CE\left(y_i', y_i\right) = -\sum_{i}^{C} y_i \log\left(y_i' = Softmax\left(x_i\right)\right) \tag{A3}$$

L2 Regularization: Regularization adds a penalty term, to the cost function, as the model complexity increases. This will decrease the importance given to higher order terms and will bring the model towards less complex function. In the L2 Regularization, the squared magnitude of the coefficient is added to the loss function:

$$Loss\left(y_i', y_i\right) = Cost\left(y_i' = f\left(x_i, \beta_i\right), y_i\right) + \lambda \sum_{i=1}^{n} \beta_i^2 \tag{A4}$$

Here the first summation term represents the original loss function, and the second summation term represents the L2-regularization penalty.

Normalization/Feature Scaling: Because the range of values of raw data varies widely, the objective functions will not work properly without normalization. For example, if the classifiers rely on the distance between two points by the Euclidean distance, if one of the features has a broad range of values, the distance

will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately towards the final distance. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

In this study, we used the feature standardization method which results in a zero mean and unit variance. The formula is given as follows:

$$x_{normalized} = \frac{x - \bar{x}}{\sigma} \tag{A5}$$

Softmax Activation Function: takes an un-normalized vector and normalizes it into a probability distribution. It provides a means to compare the given scores. Specifically, the higher score gets a higher probability than a lower score. It is implemented according to the following relation:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{A6}$$

RMSProp Optimizer: RMSProp optimizer is an unpublished (Tijmen & Hinton, 2012) optimizer similar to gradient descent with moment optimizer. It combines the idea of using the sign of the gradient with the idea of adapting the step size separately for each weight. It stores the running average of the square of the past gradients.

Adaptive Moment (Adam) Estimation Optimizer: Adam optimizer (Kingma & Ba, 2015) computes the adaptive learning rates for each parameter. It stores the average of past squared gradients (like RMSProp), as well as exponentially decaying average of the past gradients. Adam prefers flat minima in error surface (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017).

One Hot Vector: One hot vector is a collection of bits used to represent categorical data. Each n^{th} category is represented by the bit vector with all low values (0) except the n^{th} value being high (1).

Confusion Matrix: Confusion matrix is a table to describe the performance of a classification model. The table lists the possible outcomes from the model against the actual expected category. In the case of a binary classification, the table presents the accuracy metrics as: True Positives (TP: model predicts 'yes' and the actual class is also 'yes'), True Negative (TN: model predicts 'no' and the actual class is also 'no'), False Positives (FP: model predicts 'yes' but the actual class is 'no'), and False Negatives (FN: mode predicts 'no' but the actual class is 'yes'). Multiclass confusion matrices present the model performance is a similar manner. The accuracy of the model is given by:

$$Accuracy = \frac{TP + TN}{total} \times 100\% \tag{A7}$$

The misclassification rate is given by subtracting the accuracy from *100%*. The TP rate or 'recall' is given by:

$$Recall = \frac{TP}{Actual\ Positive} \times 100\% \tag{A8}$$

The precision of the classifier is computed using:

$$Precision = \frac{TP}{Predicted\ Positive} \times 100\% \quad \text{(A9)}$$

Mean Squared Error Loss Function: Mean squared error loss function measures the average of the squared difference between the actual and estimated values.

$$MSE(y_i', y_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \quad \text{(A10)}$$

Mean Absolute Error Loss Function: Mean absolute error measures the average of the absolute difference between the actual and estimated values.

$$MAE(y_i', y_i) = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \quad \text{(A11)}$$

The models were created using standard procedure of TensorFlow (Abadi, et al., 2016) using Python.

Appendix B – Inputs of the economic analysis

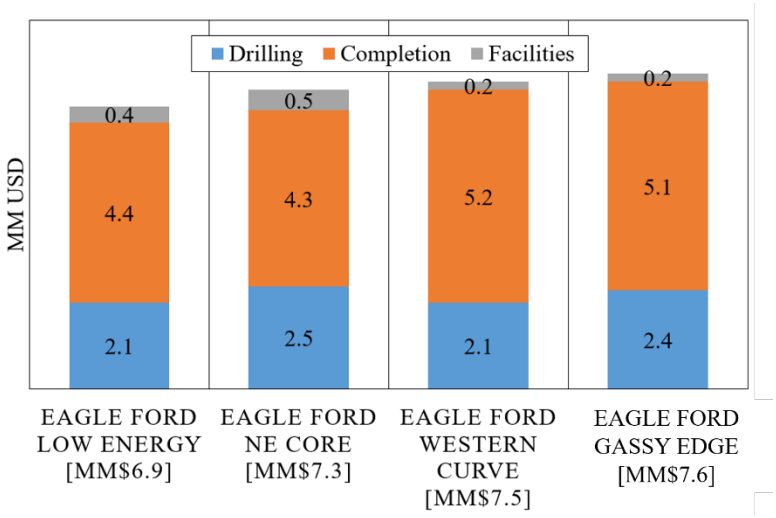


Figure B1. Typical cost of a well in Eagle Ford (source: EIA (2016))

Table B1. Assumed oil and gas prices for economic model (EIA (2016))

Fluid	Cost
Oil	50$/bbl
Gas	1.8 $/Mscf

Table B2. Estimated CAPEX (EIA (2016))

Operation	Cost
Drilling	160 $/ft (vertical + horizontal)
Frac fluid	0.167 $/gal
Proppant	0.143 $/lb

Table B3. Estimated OPEX (EIA (2016))

Operation	Cost
Oil transport	5.17 $/bbl
Gas processing and transport	1.25 $/Mscf
Water handling	2.25 $/bbl