CS229 Final Project Report

Prediction of Two Phase Flow Rate through Wellhead Chokes in Oil Wells

Negar Nazari (nazari@Stanford.edu), Talal Alshafloot (talalss@Stanford.edu)

Abstract

Wellhead assembly is an essential part of a producing oil or gas well, where it protects downstream facilities from the danger of high flow rates. An important part of this assembly is the choke that controls the flow rate of multiphase flux, in addition to protecting the hydrocarbon formation and surface equipment from probable fluctuation in pressure. Accurate prediction of flow rate through chokes is extremely helpful for assessing the reservoir performance and production forecasting. Furthermore, it is essential for establishing a controllable and stable flow in producing wells. Since flow meters are expensive and difficult in implementation for large fields, measuring the production rate of oil wells is hard. Furthermore, in fields with advanced well systems, multiple wells are connected to one manifold, and the flow rate reported from the manifold is for combined wells and not for individuals. In this work, we used machine-learning techniques to develop a reliable predictive model for predicting gross flow rate through choke. We performed extensive feature selection, data analysis, and hyper-parameter tuning to optimize the most suitable models for this purpose.

Background and Literature Review

There are several proposed correlations in the literature to predict flow rate based on simple parameters measured at the wellhead assembly [1, 2, 3]. Gilbert [4] presented the most famous correlation to calculate flow rates through chokes as:

$$Q_l = \frac{A \times P_{up}^{D} \times S^{B}}{GLR^{C}}$$

where Q_l is liquid gross flow rate, P_{up} is the upstream pressure, GLR is the gas liquid ratio, and S is the choke size as a multiple of (1/64) inches. A, B, C, and D are fitting parameters. Other authors introduced modifications to Gilbert's equation such as Baxendell [5], Ros [6] and Achong [7]. These modifications are only introduced to the parameters A, B, C and D. Nevertheless, all these version of Gilbert's equation failed to accurately estimate flow rate for plenty of wells around the world. All these correlations failed to predict accurately when applied to different set of data to the ones used to generate them. This is due to the simplicity of these correlations and their sensitivity to choke size. As a result, we need a robust model to estimate flow rate in oil wells from simple parameters measured at the wellhead assembly.

Data Processing for Feature Selection

Data was collected from three large oil fields in the Middle East. Considering the importance of different parameters, upstream pressure (P_{up}), gas liquid ratio (GLR), choke size (S), temperature (T), differential pressure ($\Delta P = P_{up} - P_{dn}$), water-cut ($WC$) [8], gas oil ratio ($GOR$), and flow regime ($Critical$ or $Subcritical$) are the initial potential features. The P_{dn}/P_{up} ratio determines the flow regime to be Critical (Ratio <= 0.5) or Subcritical (Ratio > 0.5). The gross liquid flow rate (Q_l) is the parameter to be predicted. We need to note that only the first three features were considered in the literature. Figure 1 shows different parameters against gross liquid flow rate for all fields together. The plots show that some parameters have outliers. As a result, for all features and variable to predict, we excluded data beyond three standard deviations from the mean. Data cleaning resulted in reducing data from 4677 to 4063 data points.

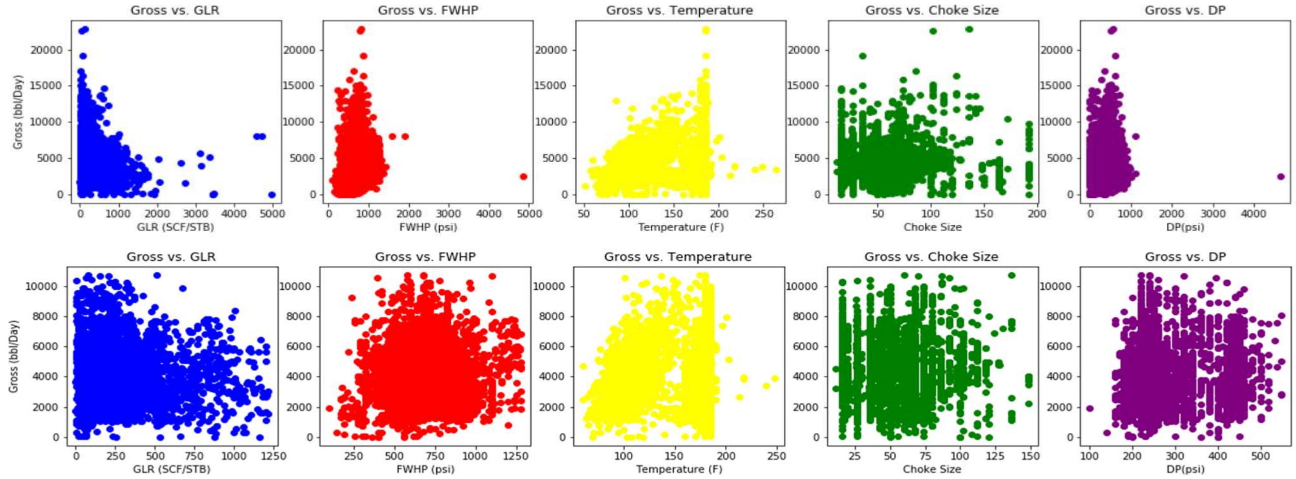

Figure 1. Gross vs. different features: before (top), and after excluding the data beyond 3σ (bottom).

Models

Figure 2 shows the poor performance of the Gilbert model to predict the gross flow rate for our entire dataset. The correlation coefficient for this model was 0.03, which shows the poor performance of this correlation to predict flow rate. This necessitate the existence of a comprehensive model to capture the maximum feature of the dataset and shows the most optimized performance for complex and thorough datasets. Therefore, we test the following models to analyze the predictability of the gross flow rate considering our introduced features.

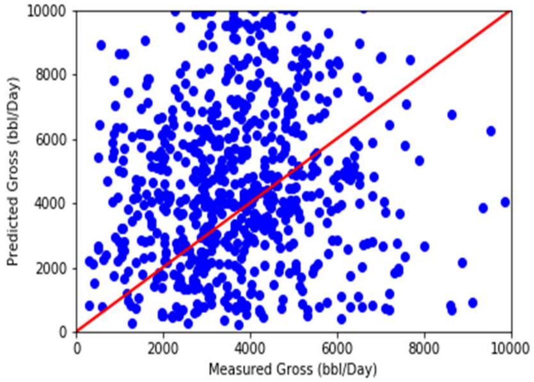

Figure 2. Predicted vs. measured gross for all three fields obtained from Gilbert correlation.

Linear Models:

- Linear/Ridge Regression: A baseline to fit a linear model with coefficients to minimize the loss function L, with and without the L_2–norm regularization, respectively. The L functions are defined as follows:

$$L = ||X\omega - \hat{y}||_2^2$$

$$L = ||X\omega - \hat{y}||_2^2 + \alpha||\omega||_2^2$$

- Bayesian Ridge Regression: It uses probability distributions rather than point estimates. The prior for the coefficient is given by a spherical Gaussian as:

$$p(\omega|\lambda) = N(\omega|0, \lambda^{-1}I_p)$$

- Polynomial Linear/Ridge Regression: It formulates the model using an nth degree polynomial to minimize L, with and without the L_2–norm regularization, respectively.

Neural Network Model:

- Multi-Layer Perceptron: The activation function for hidden layers is ReLU function. It uses backpropagation and the loss function L is defined as:

$$L = \frac{1}{2}||y - \hat{y}||_2^2 + \frac{\alpha}{2}||\omega||_2^2$$

Nearest Neighbor Model:

- K-Nearest Neighbor Regression: This model implements a learning based on the K nearest neighbors of each query point.

Ensemble Methods Models:

- Random Forrest Regression: It makes predictions by combining decisions from a sequence of base models as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots$$

- Gradient Tree Boosting: It builds an additive forward stage wise model to allow the optimization of arbitrary differentiable loss functions. It fits a regression tree on the negative gradient of the given loss function in each stage.
- Extra Tree Regression: It implements a meta- estimator $\widehat{F}(x)$ to fit a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

$$\widehat{F}(x) = arg\min_{y} E_{x,y}[L(y, F(x))]$$

Results and Discussion

We have used 10 different models from various families for this study. The total dataset size was 4677 data points, and reduced to 4323 data points after excluding the data beyond three standard deviation (3σ). After random permutation, we used 80% of the data for training and 20% for testing in the initial study. Evaluating the features reveals that the existence of different order umbers resulted in creating ill-conditioned matrices throughout the study. To solve this issue, we preprocessed the dataset by transforming the data to center by removing the mean value of each feature, and then scaling it by dividing non- constant features by their standard deviation. Furthermore, governing equations in fluid transport such as Darcy's law shows that the liquid flow rate is proportional to the pressure values linearly. Hence, we removed the upstream pressure from the features' list and fixed its exponent to be one. Figure 3 shows the initial performance of the entire model to predict the gross flow rate for the entire data set.

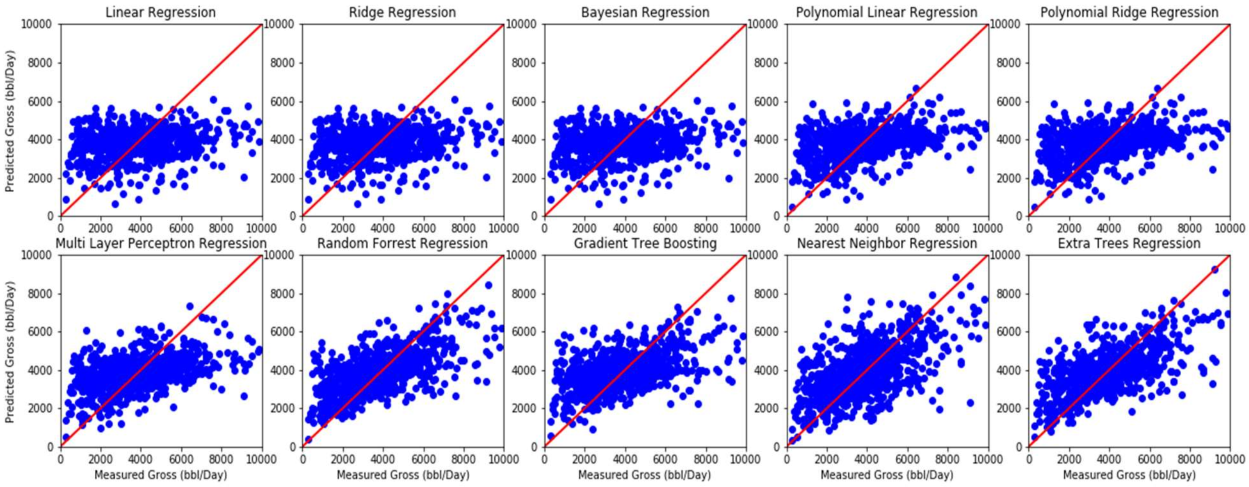

Figure 3. Predicted vs. measured gross for all three fields obtained from all models.

To make the process simple, we chose the top three models for continuing this project based on evaluating Figure 3 and the R^2-score values. Therefore, we chose extra tree regression, random forest regression and K-nearest neighbor regression as the most optimized models for continuing this study. Thereafter, we implemented the hype-parameter tuning for the best model (extra tree regression) to optimize its performance and improve the results. For this purpose, we divided the data to 80%, 10%, and 10% for training, testing, and validation, respectively. Figure 4 and Table 1 shows the performance and the statistics of the top three models for predicting the gross flow rate for the entire data set, respectively.

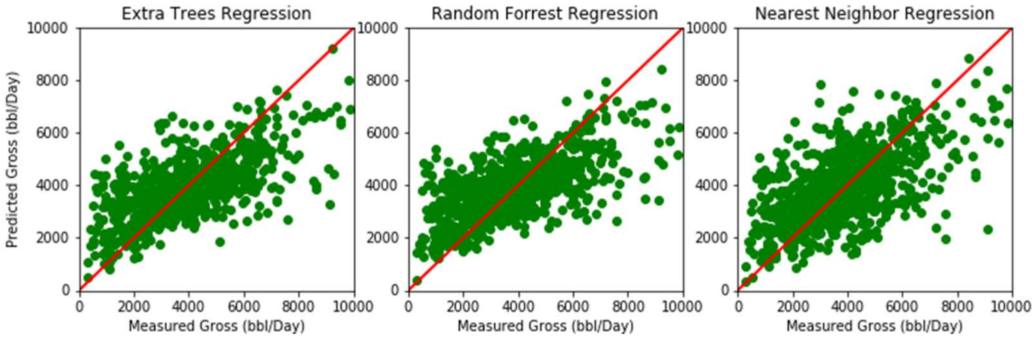

Figure 4. Predicted vs. measured gross for all three fields obtained from three top models.

Table 1. Statistics of top three models for all three fields.

Model	Training R^2-Score	Testing R^2-Score	Validation R^2-Score	Correlation Coefficient
Extra Tree Reg.	0.93	0.54	0.58	0.69
Random Forest Reg.	0.94	0.50	0.55	0.66
Nearest Neighbor Reg.	0.76	0.44	0.40	0.44

We have performed all required steps to improve our models; however, the correlation coefficient for the extra tree regression, which was our most optimized model, did not increase after 0.67. Therefore, we decided to dig into the datasets and look for other reasons causing this situation. Separate analysis of the three fields revealed that fields A, and C perform much better than field B. Figure 5 and Figure 6 shows the performance of our top three models for combination of fields A and C and field B, separately. Analyzing the statistics of the models for these two cases in Table 2 and Table 3 shows that applying the extra tree regression for fields A and C improves the correlation coefficient up to 0.78. However, this model was not able to improve the correlation coefficient of field B after 0.57. Unfortunately, due to unavailability of the well, formation, and fluids data, we were not able to run computational fluid dynamics (CFD) models to check the reliability of the data for field B. Nevertheless, we have shown that the poor performance of our models on field B can be due to other parameters that needs further investigation.

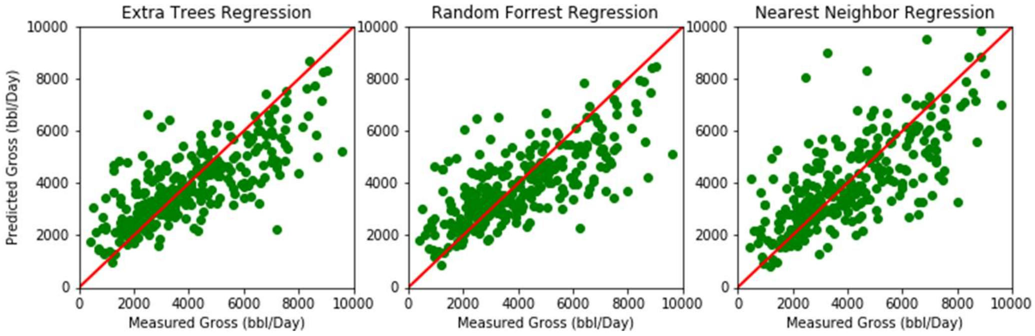

Figure 5. Predicted vs. measured gross for fields A & C obtained from three top models.

Table 2. Statistics of top three models for fields A & C.

Model	Training R^2-Score	Testing R^2-Score	Validation R^2-Score	Correlation Coefficient
Extra Tree Reg.	1.00	0.71	0.76	0.78
Random Forest Reg.	0.95	0.68	0.75	0.78
Nearest Neighbor Reg.	0.76	0.62	0.59	0.63

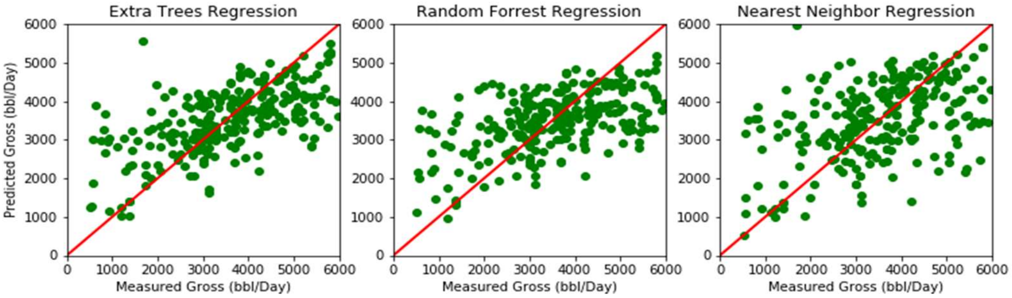

Figure 6. Predicted vs. measured gross for field B obtained from three top models.

Table 3. Statistics of top three models for field B.

Model	Training R^2-Score	Testing R^2-Score	Correlation Coefficient
Extra Tree Reg.	1.00	0.48	0.57
Random Forest Reg.	0.91	0.34	0.44
Nearest Neighbor Reg.	0.70	0.30	0.44

Conclusion

We have developed new models to predict flow rate through chokes using machine-learning method. After testing 10 different models and valuating their results, we obtained the following conclusions:

- Linear models are not capable of capturing the nonlinear behaviors, hence, they show a weak performance (R^2-score <= 0.2). The best performance in the linear models family belongs to polynomial models, as higher degree polynomials are capable of capturing different behaviors. However, due to the overfitting issue, their performance improvement is extremely limited.
- Applying the neural network models improves the R^2-score up to 0.5 for some cases, but adding more hidden layers with different activation functions may help improving the results.
- Extra tree regression as our most optimized method improved the correlation coefficient up to around 0.7 for the entire dataset. Generally, ensemble learning helps improve machine learning results by combining several models. This result in better predictive performance compared to a single model. Tuning the hyper-parameters result in capturing complex behaviors and improved the performance.
- Dataset B shows poor results, which implies that we need to add more features related to flow and formation properties to capture all different aspects of the behavior.
- The results of this study is applicable to water resources studies and reduces the flow costs significantly.

Future Work

- Due to the strange behavior of field B, we will run CFD to validate this dataset.
- A new neural network model with more complex hidden layers will be implemented
- We need to find flow and formation properties add them to the best predictive models.

GitHub Link to the Code

https://github.com/NazariStanford/CS229/blob/master/Nazari_Alshafloot_Code.py

Note on Member Contribution

Because of the limited size of the group, most of the tasks where conducted by both members while small sub-tasks were done individually.

References

[1] M. A. Al-Khalifa and M. A. l-Marhoun, "Application of Neural Network for Two-Phase Flow through Chokes," in *SPE Saudi Arabia section Annual Technical Symposium and Exhibition*, Khobar, Saudi Arabia, 2013.

[2] M. D. AlAjmi, S. A. Alarifi and A. H. Mahsoon, "Improving Multiphase Choke Performance Prediction and Well Production Test Validation Using Artificial Intelligence; A New Milestone," in *SPE Digital Energy Conference and Exhibition*, Woodlands, Texas, USA, 2015.

[3] D. W. Surbey, B. G. Kelkar and J. P. Brill, "Study of Multiphase Critical Flow," *SPR Production Engineering Through Wellhead Chokes,* pp. 142-146, 1989.

[4] W. E. Gilbert, "Flowing and Gas-Lift Well Performance," in *Drilling and Production Practice,*, New York, New York, USA, 1954.

[5] P. B. Baxendell, "Bean Performance-Lake Wells," Shell Oil Company, 1957.

[6] N. J. C. Ros, "An Analysis of Critical Simultaneous Gas/Liquid or through a Restriction and its Application to Own Metering," *Applied Scientific Research,* vol. 9, no. 1, p. 374, 1960.

[7] I. Achong, "Revised Bean Performance Formula for Lake Maracaibo Wells," internal company report, Shell Oil Co., Houston, USA, 1961.

[8] M. S. Beiranvand and M. B. Khorzoughi, "Introducing a New Correlation for Multiphase Flow Through SUrface Chokes with Newly Incorporated Parameters," *SPE Production and Operation,* pp. 422-428, 2012.