OTC-29243-MS

An Intelligent Data Driven Approach for Production Prediction

Christine Ikram Noshi, Texas A&M University; Marco Risk Eissa, Cairo University; Ramez Maher Abdalla, The American University in Cairo

This paper was prepared for presentation at the Offshore Technology Conference held in Houston, Texas, USA, 6 – 9 May 2019.

Abstract

The objective of this work is to further explore the potential application of Machine Learning algorithms in production prediction and ultimate recovery. Intelligent Machine Learning Approaches such as Gradient Boosted Trees (GBT), Adaboost, and Support Vector Regressor (SVR) are applied to detect the most important features contributing to cumulative production prediction within the first 12 producing months. The models are applied on a data set composed of 5 wells in the Volve field in the North Sea. The collected data was then filtered and used to structure and train the different Regression algorithms and fine tune the appropriate hyperparameters. The different models were All models were evaluated by measuring the Mean Absolute Error (MAE). The generalization and precision performance of the proposed models are established by comparing the forecasted outcome after cross validation with field data. The optimized model can predict production response with high accuracy. The data-fitting process comprises of splitting the data into training using 70% of the data set, 15% validation, and 15% testing. Constructing a regression model on the training set and validating it with the test set. Recurrent application of a "cross-validation" process produces important information concerning the robustness of any regression-modeling method. Six parameters were considered as input factors to build the model. Factors affecting production prediction included on stream hours, average choke size, bore oil volume, bore gas volume, bore water volume, average wellhead pressure were used as input. The outcome showed that the developed model provided better prediction compared to analytical models with a 11.71% MAE prediction for SVR. This novel data mining application could be trained on any dataset to help predict future production performance at any conditions in any given scenario.

Introduction

Data-driven modeling have become rather the buzzword in the past decade, particularly, in the context of oil and gas reservoir production analysis. The definitive goal was to advance data-driven insights for developing and optimizing the behavior and production optimization of unconventional plays. Many statistical and machine learning methods have been proposed to visualize, summarize and draw conclusions from the data. Each of them has its own application background and set of assumptions. They were developed to work extremely well in specific scenarios but can be terrible choices in others. Lately, Data Mining and

Machine Learning methods have been proved to be the effective approaches for production analysis in unconventional plays (Noshi et al. 2018), especially, when the underlying relationships among variables are highly complex, and non-linear. Production forecasting in unconventional reservoirs is a complex problem because of the complicated interrelationship between the geology, heterogenous lithology, stimulation techniques, etc and production optimization problems (i.e., analysis of production, well completion, and/or formation evaluation data) have been an issue since the dawn of the petroleum era. Conventionally, to infer important reservoir properties and performance-related parameters, historical flow rate and pressure data from producing wells are analyzed using techniques such as straight-line (regime-flow) analysis, analytical models, and numerical-simulation approaches (Clarkson 2013). In addition, techniques such as empirical decline-curve models are commonly used to fit trends to well-production data. However, most of these traditional methods (specifically type-curve analysis and regime-flow methods) are formulated for vertical wells in conventional reservoirs, which means that they are generally predisposed to boundary-dominated flow and only provide average information about bulk reservoir properties (Anderson et al. 2010). In addition, they do not consider the complex physics of fluid storage and nanoscale fluid flow in the formation. A number of dual-porosity analytical models have therefore been proposed to characterize production from multistage hydraulically fractured shale wells (Bello and Wattenbarger 2010). However, these analytical models are often developed in Laplace space corresponding to specific boundary conditions, which means that numerical-inversion schemes must be used to express the solutions in the time domain. In addition, they typically require several matrix, fracture, and well input parameters, which are often expensive or impossible to infer accurately. Overall, many of these existing methods tend to require detailed input data and/or lack generality to other reservoir settings, in addition to being time consuming to implement, which limits their applicability to problems involving massive data sets. While Data Mining techniques are recently used to forecast the oil and gas productions in unconventional shale or tight reservoirs, the performance of such models has not been impressive (Shelley et al. 2008). One of the chief causes is that oil and gas productions are greatly affected not only by the characteristics of the reservoir such as permeability, porosity, oil saturation, and pressure, but by the horizontal well stimulation characteristics and pressures bottom-hole. It is difficult to collect all the data for a horizontal well. Another reason is that data sets are mostly obtained from oil and gas fields suffering from a large uncertainty. Therefore, a more comprehensive data mining framework involving both descriptive and predictive data mining techniques is needed to copy with the complexity and uncertainty of the production prediction problems. Practically, these data-driven techniques deliver a simple framework to real time design and optimization, since the corresponding mechanistic models including physics-reliant simulators would be more time-consuming to develop, implement. and interpret.

In this work, we used a data set from the Volve field wells in the North Sea as a case study to illustrate the implementation of several popular analytic methods in the Data Mining domain. Our method encompassed the implementation of Gradient Boosted Trees (GBT), Support Vector Regressor (SVR)and Adaptive Boosting (AdaBoost) algorithms for predicting the future performance of the oil wells, based on historical production data of the previous years. The production decline was captured during the algorithms' training development and was implemented to the production data during the prediction phase. The method used was time series analysis and it captured the trend, value fluctuations, decline rate, and correlation with the past to produce fast and accurate forecasts. The application of the trained data on the observed data is extremely fast, thereby rendering the method amenable to real-time prediction on the basis of well performance. This in turn has the potential to affect decision-making work flows involving massive data sets collected from producing wells within a field, which helps to mitigate uncertainty in reservoir management and decision making through continuous analysis. Our dialogue will accentuate an analytical workflow that can be simply implemented by various engineers collaborating with data scientists.

Field Description

Volve is a decommissioned field located in the North Sea and was revealed in 1993. The Field is situated 200 kilometers west of Stavanger at the southern terminal of the Norwegian division as shown **i**n Figure 1 and five **k**lometers north of the Sleipner Øst field with water depths of 80 meters in the block 15/9. Drilling started in May 2007, came into production the following year and ended in 2016 after 8.5 years in operation, more than twice as long as originally planned. New wells were being drilled up until 2012-13, which contributed to the increased recovery rate and extended the life of the field. However remaining resources were very limited and with the decrease in oil price over recent years, new wells were no longer profitable. All possibilities to extend the life of the field were explored, which yielded very good results.

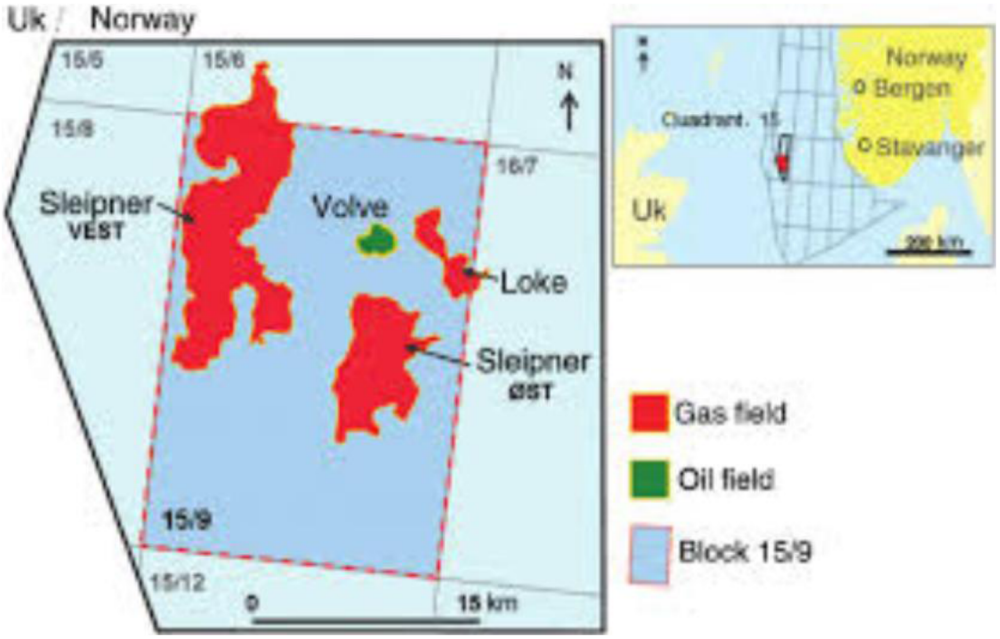


Figure 1—Volve Field Location

The reservoir is located at 2750 – 3120 m depth. The produced oil was from sandstone in the Hugin Formation of the Jurassic age. The reservoir in the Volve field consists of Jurassic sandstones at 2750 – 3120 m TVD ss, the field was produced with water injection as pressure support. It has not been possible to determine the age of the sandstones in the wells 15/9-9, -15 and -17. They could probably represent deposits of Early - Middle Jurassic age. In the Sleipner Øst area, the Hugin Formation is therefore with certainty only identified in the 15/9-11, -13 and -19 SR wells. The 15/9-11 well has also limited value for such analysis, because the Hugin Formation was not cored in this well, well data information is thereby very limited and scattered. Only two wells have been cored in this formation (15/919 SR and 15/9-13). In one of them (15/9-19 SR), the formation is heavily faulted.

Volve reached a recovery rate of 54%, at plateau Volve produced some 56,000 barrels/day and produced an accumulation of 63 million bbls. The field delivered about 9.5 million barrels of oil beyond what was expected in the development and operation plan. Permanent plugging of the wells and other process activities were carried out in December.

Methodology

The scope of this work was to focus on predicting the future production rates (bbl) from the publicly reported Volve Field production data set. The data set is composed of 7 wells (5 producers and 2 injectors). The

producers namely are (15/9-F-1 C, 15/9-F-11, 15/9-F-12, 15/9-F-14, 15/9-F-15 D) Figure 2, while, injectors are (15/9-F-4 and 15/9-F-5).

Exploratory Data Analysis (EDA) and Data Preparation

Several data quality challenges were encountered and resolved as part of the EDA. This included handling missing data, spuriously low or high values (outlier peaks) and undesirable distributions. Data pre-processing and cleaning included de-noising, removing outliers, and treating the incomplete and inconsistent data in the set of variables. To account for the missing data or erroneous data, several treatments were applied based on modelling judgement. These included clamping certain parameters to minimize the impact of spurious extreme values on model training, scaling, imputing values for missing data and applying a range of data transformations to improve predictor model performance Moreover, a total injection feature (BORE_WI_VOL) was created and added to the dataframe.

A good understanding of the dataset is critical to the success of any statistical modelling project. A quick data summary using a 5-number summary, uni-variate, and bi-variate plots as shown in Figure 3 were performed along with a Pearson correlation matrix displayed in Figure 4. Using the entire data set, the correlations generated suggested that:

- Average downhole pressure is strongly correlated with average downhole temperature and average DP pressure
- The average well head pressure is strongly correlate with DP choke size which is expected since the choke size controls the well head pressure
- The produced gas volume is strongly correlated with the produced oil volume which conforms with the theory that the produces gas is released from the oil stream
- the linear correlation between produced oil volume and other parameters is weak so regular multiple linear regression model won't present a good prediction method. The exploratory data analysis revealed no apparent relation between the initially selected features and the target (Oil Volume Produced)

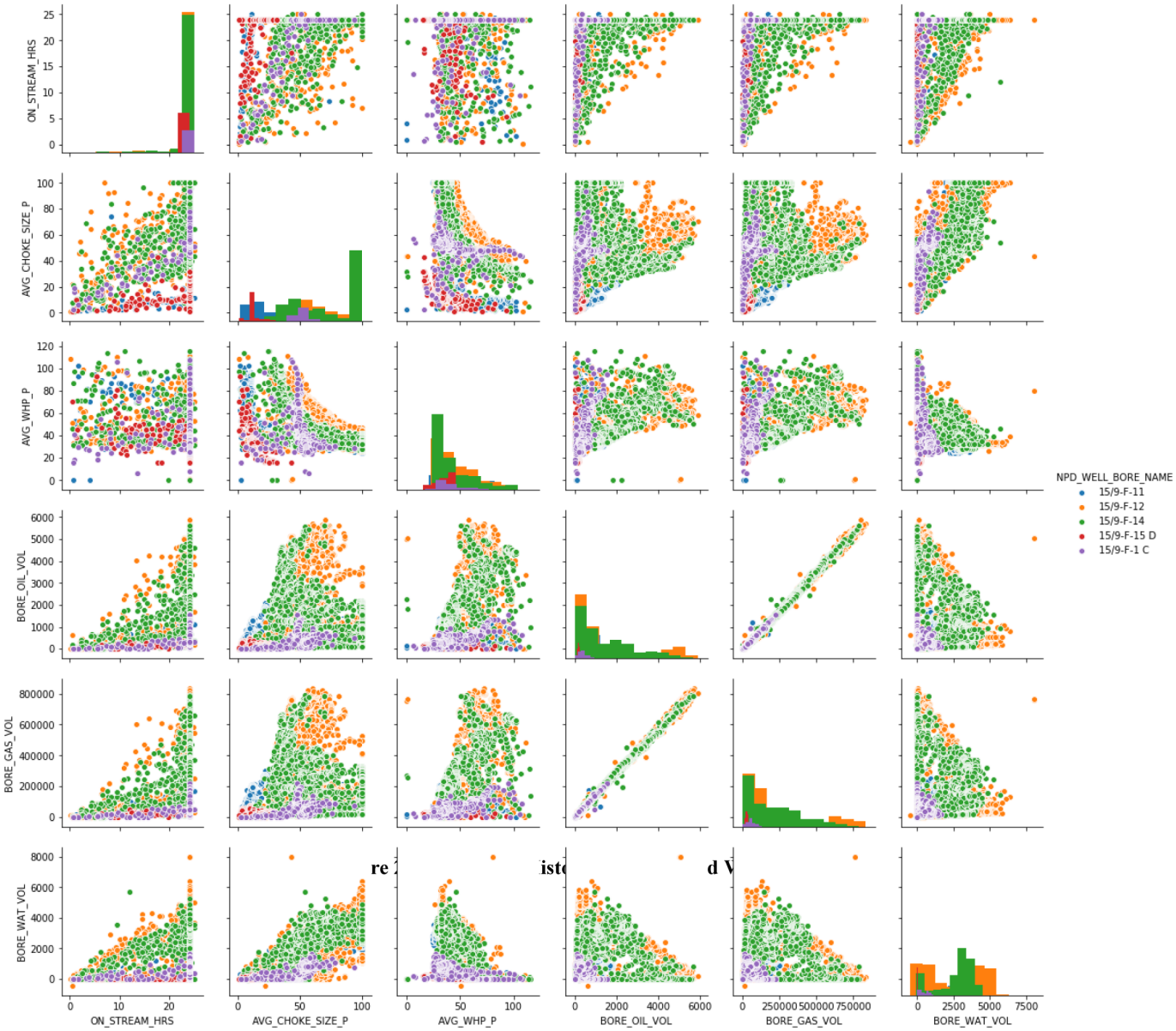


Figure 3—Uni-variate and Bi-variate plots of the selected features colour-separated by well.

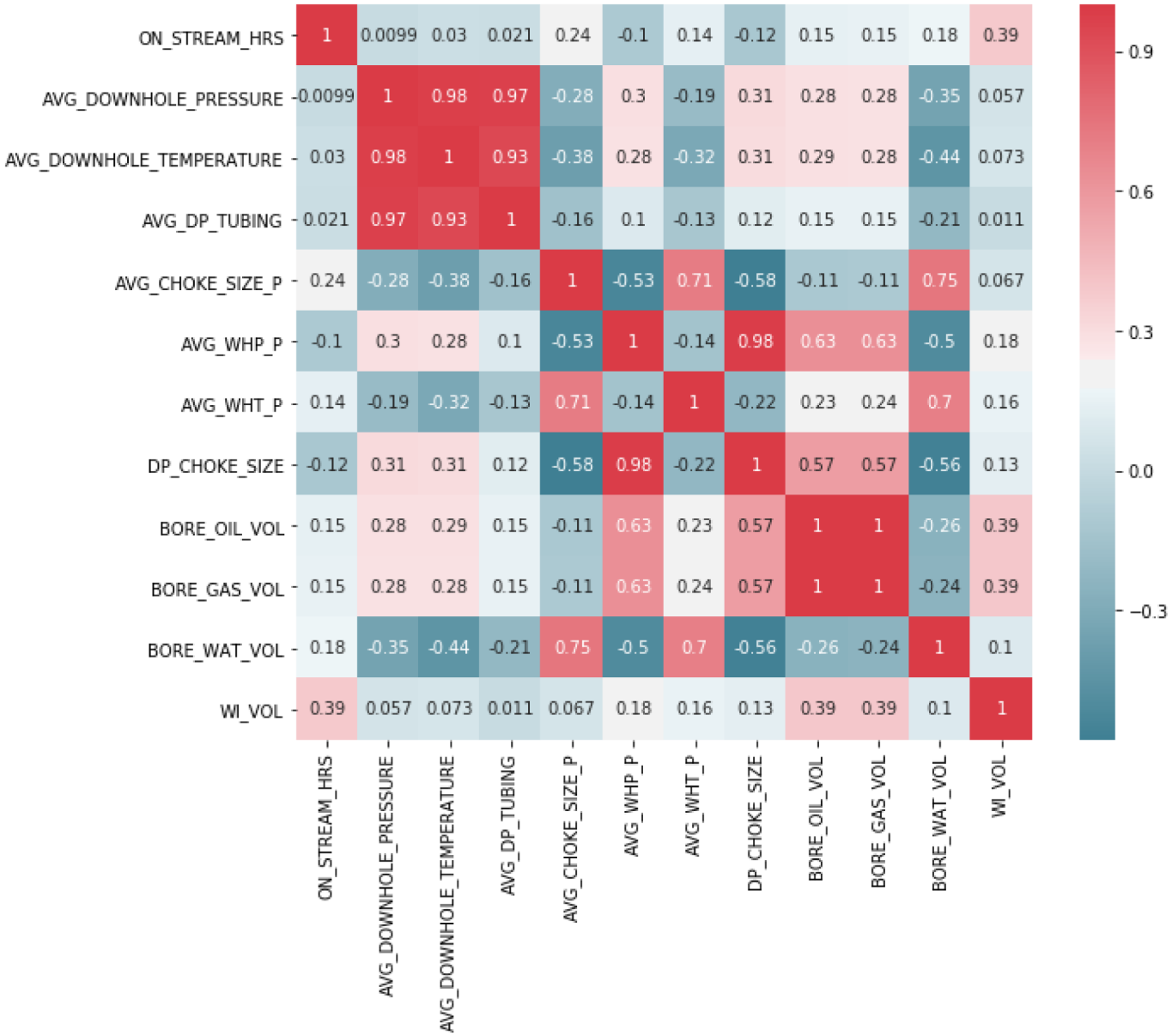


Figure 4—Pearson Correlation Matrix for Features.

Finally lag plots as shown in Figure 5 were generated to check if the produced oil volume can be successfully modelled using time series. The linear trend displays that the data is powerfully non-random and additionally advocates that an autoregressive model might be suitable.

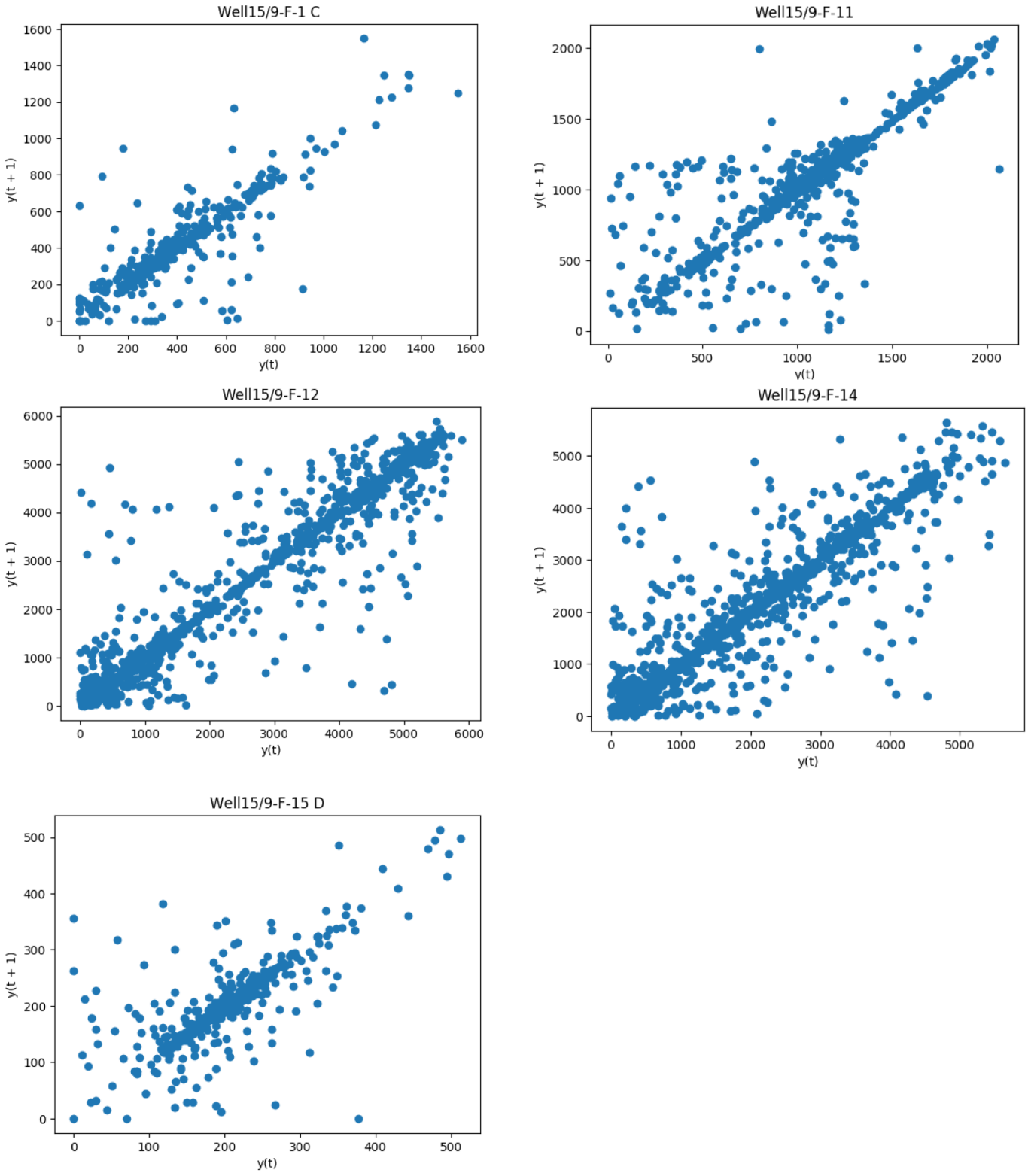


Figure 5—Lag Plots for the 5 Producing Wells.

Feature Selection

Features are the building blocks of data which should be carefully considered depending on domain knowledge and field expertise. Since initially there was a large number of attributes, the faulty attribute signatures could lead to inaccurate results hindering Machine Learning predictive algorithms. The initial feature selection was based on the theory of oil well production and what factors were most likely to affect

it. The data set obtained, contained daily and monthly production data per well. Supplementary to the well IDs, the dataset also comprised of 23 input features. The entirety of features was associated to the operational characteristics of the wells, including the time period of drilling the well, pressure information, production, and water injection information, and operator type. The target response of primary interest was BORE_OIL_VOL, which measures the daily produced oil volume in barrels (BBL).

A list of all features in the original dataset is displayed in Table—1 as follows:

Table 1—List of features in the study data set.

Type	Variable	Description
-	**ID**	**Well-Identification number**
Response	BORE_OIL_VOL	the daily produced oil volume (bbl)
	DATEPRD	Production Date (in Days)
	WELL_BORE_CODE	The ID of the well within the field
	NPD_WELL_BORE_CODE	A well official identification code that is identifying a well or wellbore on the Norwegian Continental Shelf
	NPD_WELL_BORE_NAME	Name of the wellbore in Norwegian Continental Shelf
	NPD_FIELD_CODE	Field code in Norwegian Continental Shelf
	NPD_FIELD_NAME	Field name in Norwegian Continental Shelf
	NPD_FACILITY_CODE	Code for Facility used
	NPD_FACILITY_NAME	Facility owner name
	ON_STREAM_HRS	Flow period for well (in hours)
	AVG_DOWNHOLE_PRESSURE	Average downhole pressure
Predictor	AVG_DOWNHOLE_TEMPERATURE	Average downhole Temperature
	AVG_DP_TUBING	Average production tubing size (in mm)
	AVG_ANNULUS_PRESS	Average Annulus Pressure
	AVG_CHOKE_SIZE_P	Average Choke Size
	AVG_CHOKE_UOM	Average Choke Unit of Measurement (%)
	AVG_WHP_P	Average Well head Pressure
	AVG_WHT_P	Average Well head Temperature
	DP_CHOKE_SIZE	production choke size (in mm)
	BORE_OIL_VOL	Produced Oil volume / Day
	BORE_GAS_VOL	Produced Gas volume / Day
	BORE_WAT_VOL	Produced Water volume / Day
	BORE_WI_VOL	Injected Water Volume / Day
	FLOW_KIND	whether a production or injection well
	WELL_TYPE	Same as previous but encoded (OP & WI)

After preliminary analysis and using petroleum production engineering intuition the features were downsized to 12 features:

- Duration in Days (On Stream Hours)
- Average Downhole Pressures
- Average Downhole Temperature
- Average Tubing diameter

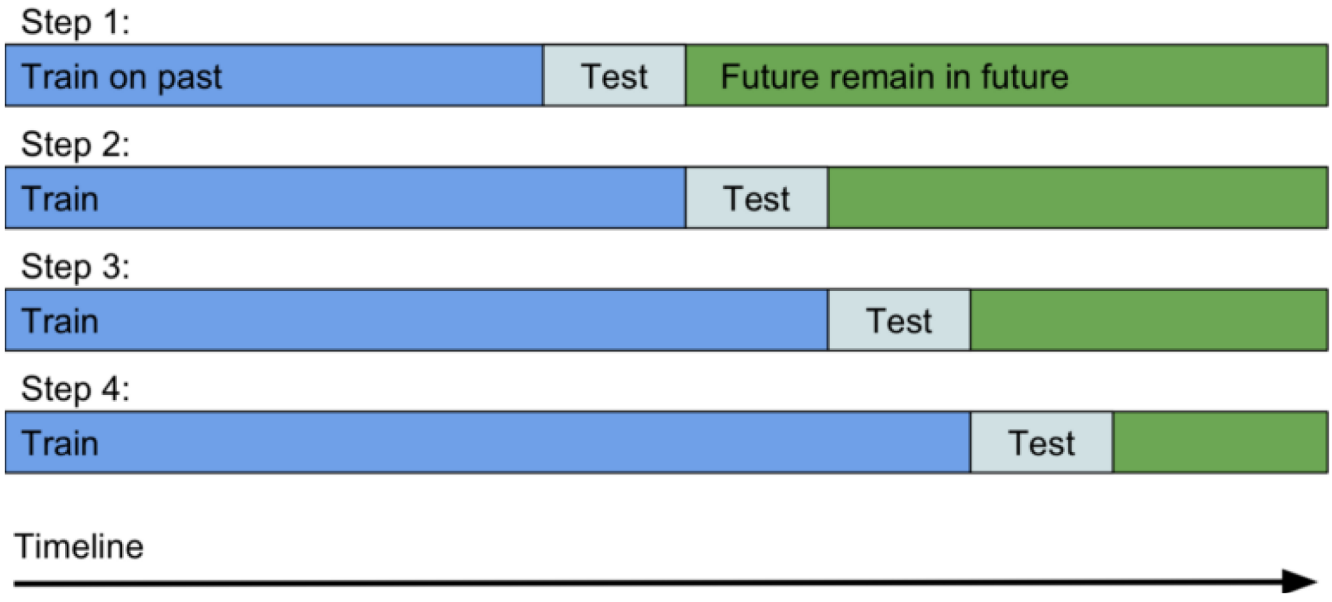
- Average Annulus Pressure
- Average Choke size
- Average Well Head Pressure
- Average Well Head Temperature
- The influence of Injector Wells
- The previous time step produced oil/gas/water volume (3 features)

Further selection, from twelve input features down to six key predictor features, was done through the application of a range of Data Science techniques and domain knowledge. The final feature selection was done through running different permutations of the initially selected features and measuring how they reflect on the final model accuracy. The set of features that gave the best model accuracy were:

- Duration in Days (On Stream Hours)
- Average Choke size
- Average Well Head Pressure (dropping DP_choke_size since the choke size controls WHP, but WHP is the influencing factor in the oil production equations)
- The previous time step produced oil/gas/water volume (3 features)

Data Splitting

For time series data, the traditional training/testing data split or cross validation are not applicable since the purpose of the model was to predict future behaviour. An alternate approach is to split the data sequentially at different points in the life of the well allowing for model validation on different segments of the well production life (see figure below).



Predictive Modeling Algorithms

Moving past investigative data analysis, a mutual goal in the oil and gas industry is to build predictive models. For instance, in the Volve dataset, the intention may be to forecast the production rates for a new well with a certain set of operational features. SVR (Noshi et al. 2018b; Noshi and Schubert 2018), GBT

and AdaBoost (Noshi et al. 2018a) algorithms were selected due to their robustness. A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely associated than observations far apart. Furthermore, time series models will usually take advantage of the natural one-way collation of time so that values for a certain period will be articulated as deriving in some manner from previous values, rather than from future values. Considering the previous definition of time series modelling, supervised machine learning algorithms do not directly accommodate for the sequential relation between values of time series. The introduction of time lagged features allows machine learning algorithms to capture the time component relation of features and target values.

A two-time step lag was used in modelling this dataset. Higher lags were tested and provided an improvement of 1% which was traded off to keep the model less complex and less computationally expensive.

Predictive Modeling Results

The three selected models were assessed by their ability to predict the future trend correctly and have a low mean absolute error (MAE).

The SVR technique is found to forecast the production spikes and follows the shifts and trends extremely well. There are some differences (slight offsets) between the predicted versus actual that could potentially be improved/reduced through further model refinement (e.g. model calibration, weighting of predictor features, optimized time lags). Six input features including on stream hours, average choke size, bore oil volume, bore gas volume, bore water volume, average wellhead pressure, were used as input. These features should be known to the operator before initiating prediction. For this purpose, SVR model was tested on five different segments of the life of the well, two out of the five wells exhibited high MAE, this can be attributed to high fluctuations which in return gives the model high error results. However, with regards to the remaining three wells, when the production performance is stable, the model predicts with an accuracy of 90%.

GBT and Adaboost were implemented on the data set but the prediction outcome was poor. Tree regressors did not work in this problem as the method involved predicting the mean value at the leaf giving a constant value when the prediction rate is below a threshold. Only one learning rate gave unsatisfying results Figure 6.

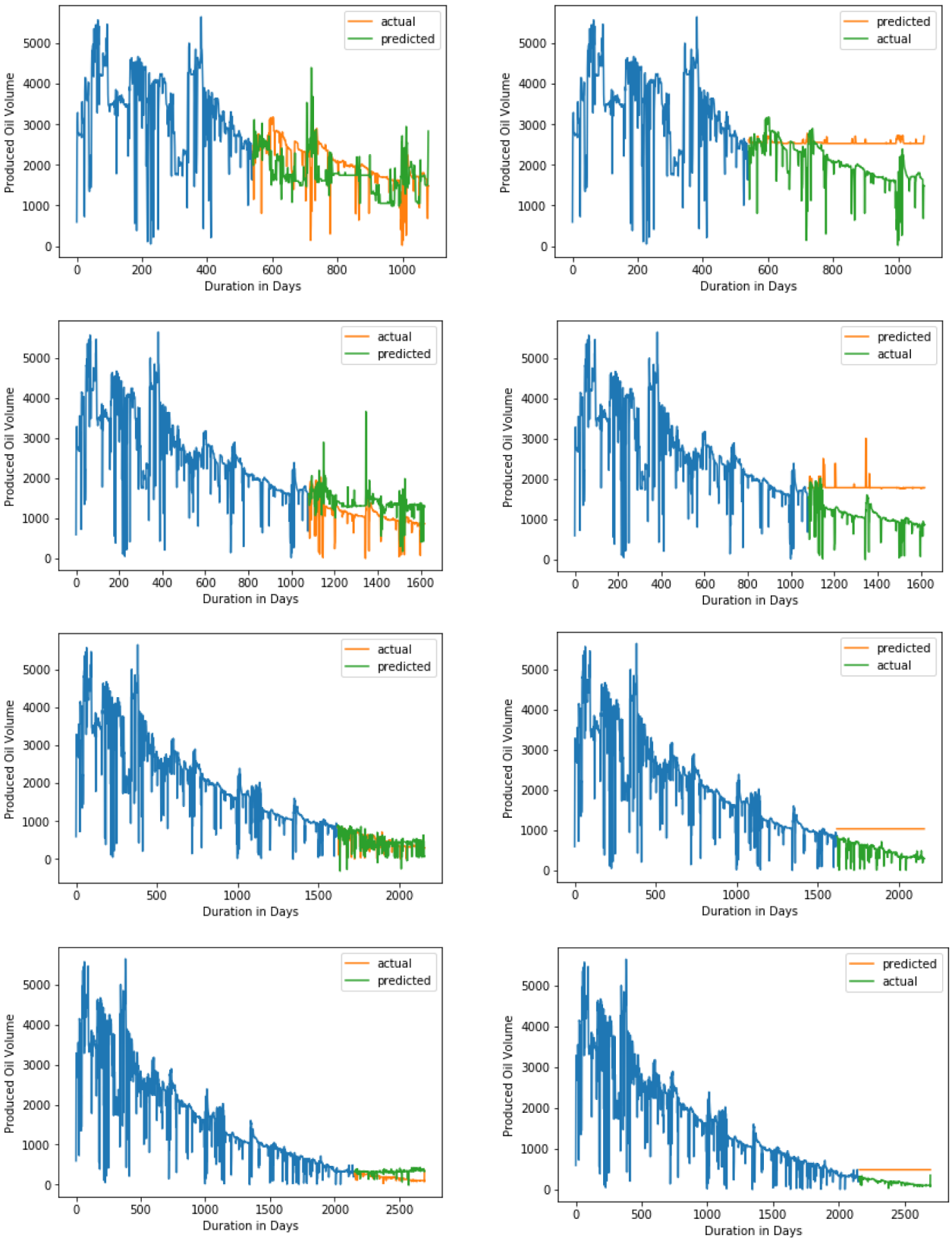


Figure 6—On the Left GBT Algorithm and on the Right AdaBoost Algoritm, predicting future rates on 4 intervals of well 15/9-F-14.

Out of the three models, only SVR showed good results as shown for the 5 wells in Table — 2. With a MAE scoring 90.9% of predictions accuracy at the best case, the feasibility of building a predictor model

using Machine Learning approaches was deemed to have been demonstrated for this application. Since the model takes only 6 variables which are relatively straightforward measurements, predictions are provided instantaneously due to the simplicity of the time series model. The plots of the SVR model predictions for the 5 wells are shown in Figures 7 to 11. Table — 3 shows a MAE comparison for the all the models for well 15/9-F-14

Table 2—*The SVR MAE comparing the metrics comparing the 5 wells.

Metric	Well 15/9-F-1 C	Well 15/9-F-11	Well 15/9-F-12	Well 15/9-F-15 D	Well 15/9-F-14
MSE ((bbl/ft)2)	70.3	57.2	62.6	9.11	10.3

*Note: the MAE error in the tables above is for the most stable region of the well production life (usually the last of the four splits)

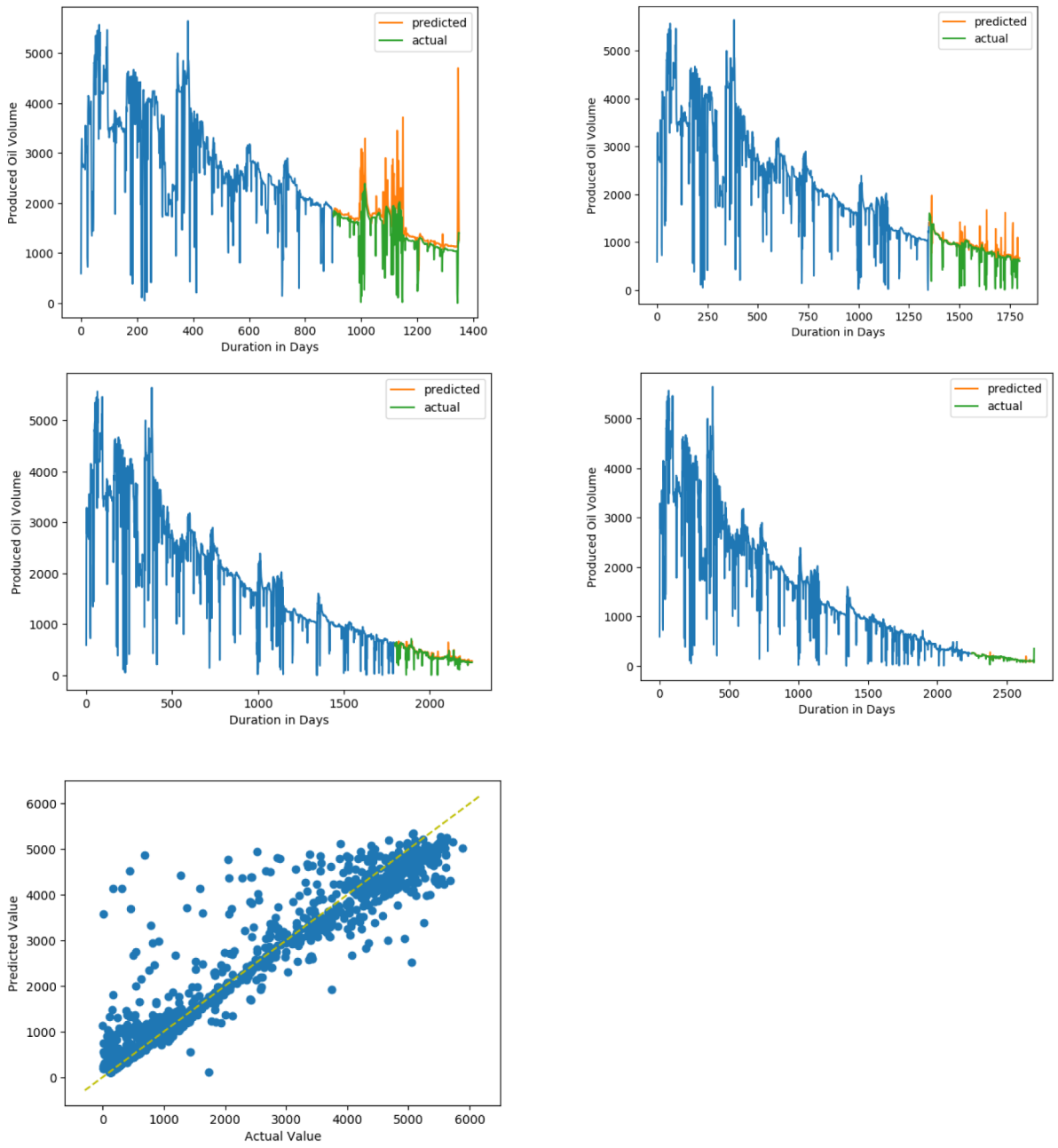


Figure 7—SVR predictions on 4 intervals of well 15/9-F-14.

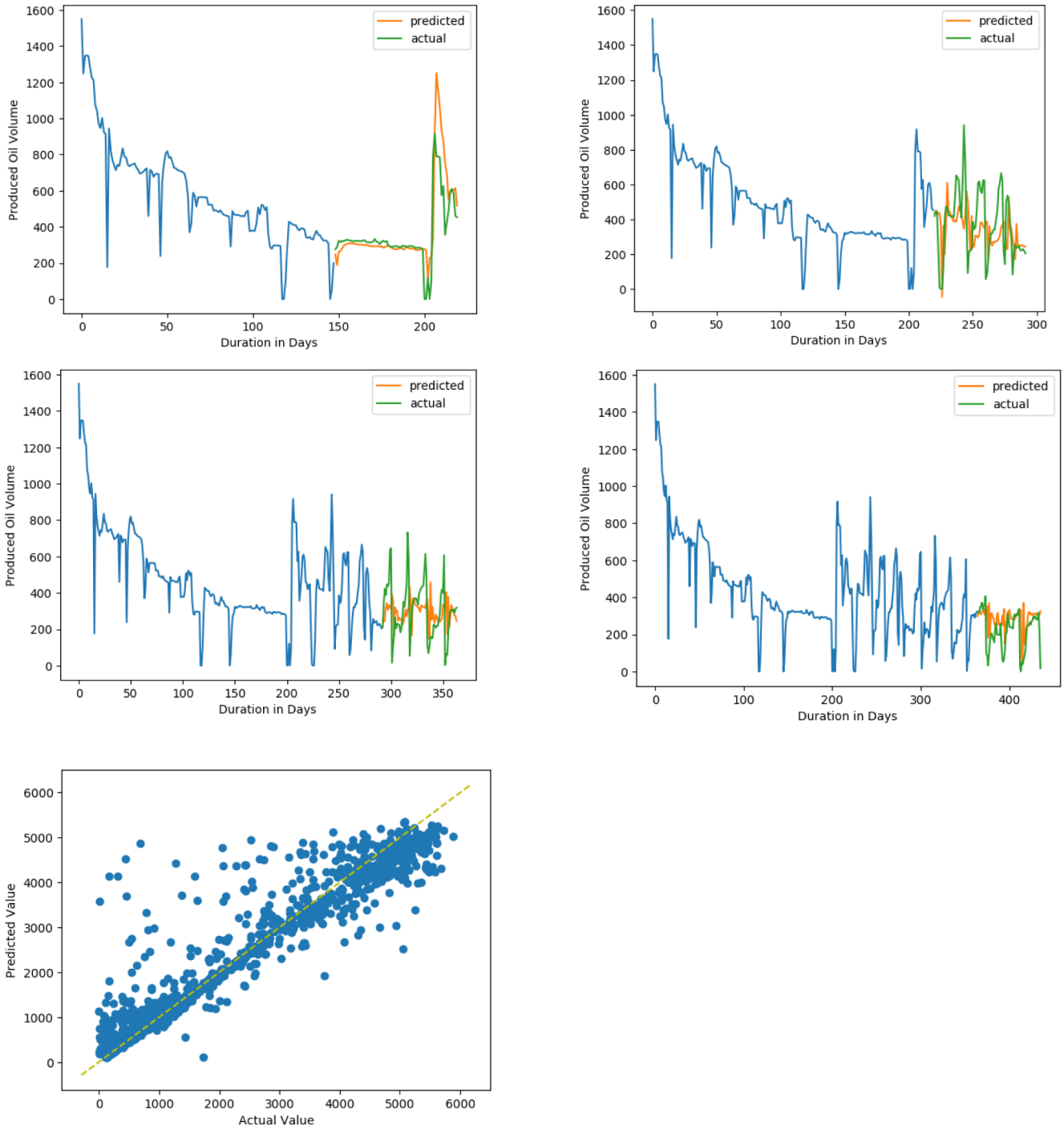


Figure 8—SVR predictions on 4 intervals of well 15/9-F-1 C.

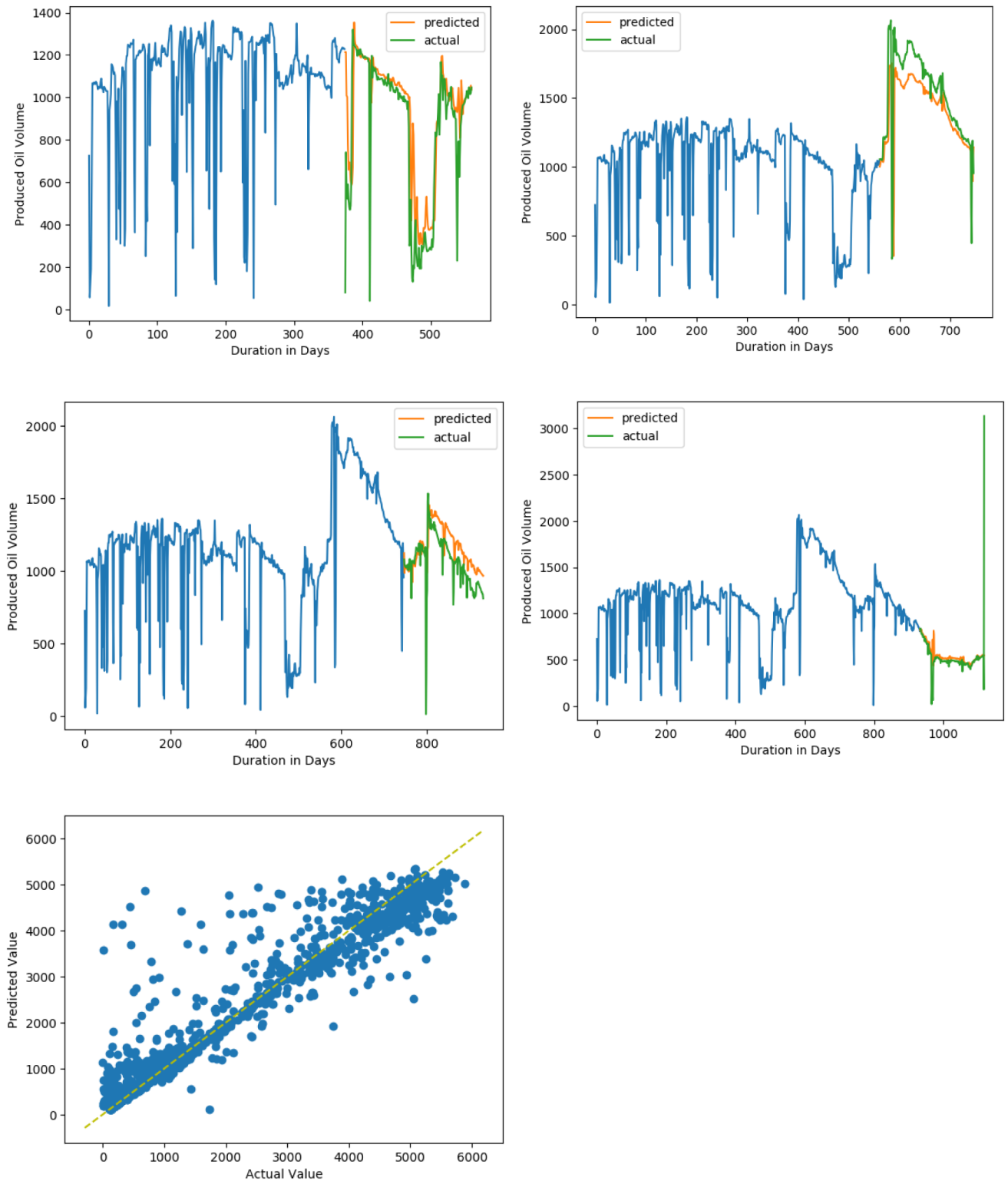


Figure 9—SVR predictions on 4 intervals of well 15/9-F-11.

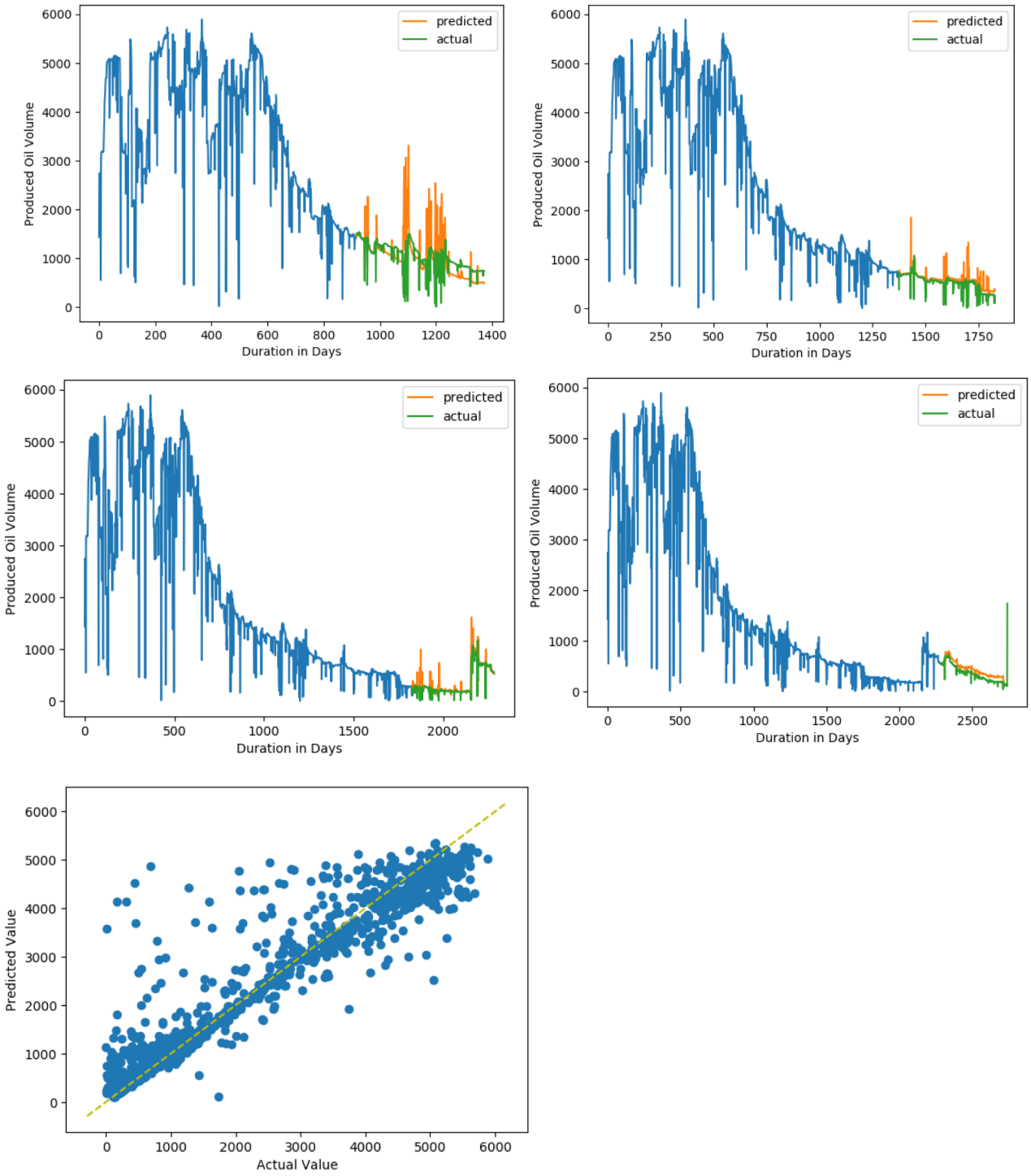


Figure 10—SVR predictions on 4 intervals of well 15/9-F-12.

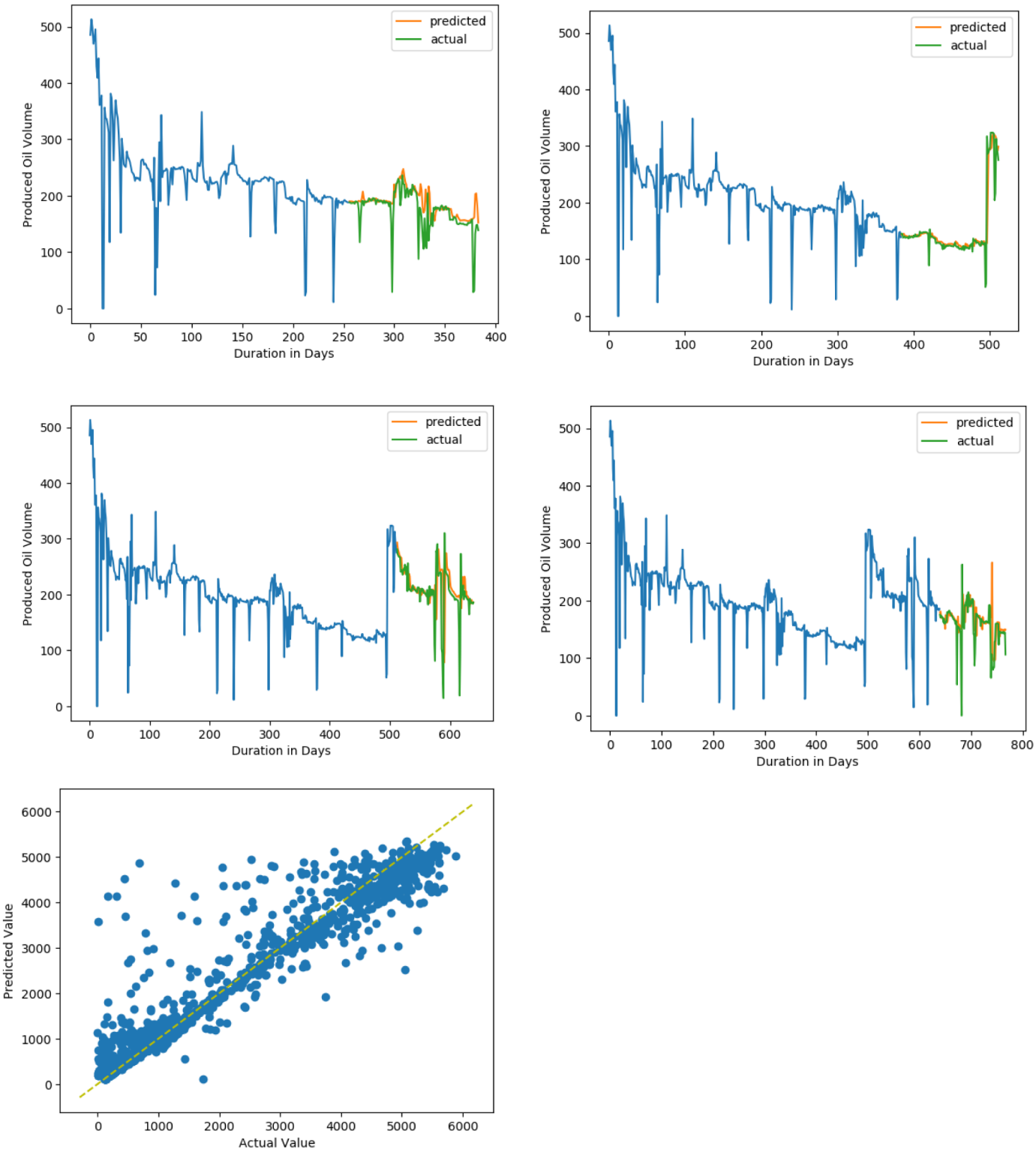


Figure 11—SVR predictions on 4 intervals of well 15/9-F-15 D.

Table 3—*MAE comparison from the four models for well 15/9-F-14.

	AdaBoost	SVR	GBT
MAE	300	12	151

*Note: the MAE error in the tables above is for the most stable region of the well production life (usually the last of the four splits)

Testing Model Generalization. To further validate the generalization capability of this model, the model was trained on well '15/9-F-14' was tested on well '15/9-F-12' as shown in Figure 12. The prediction results captured the trend significantly well, mimicking its spikes and following the decline pattern smoothly, considering its representation over the entire life of the well.

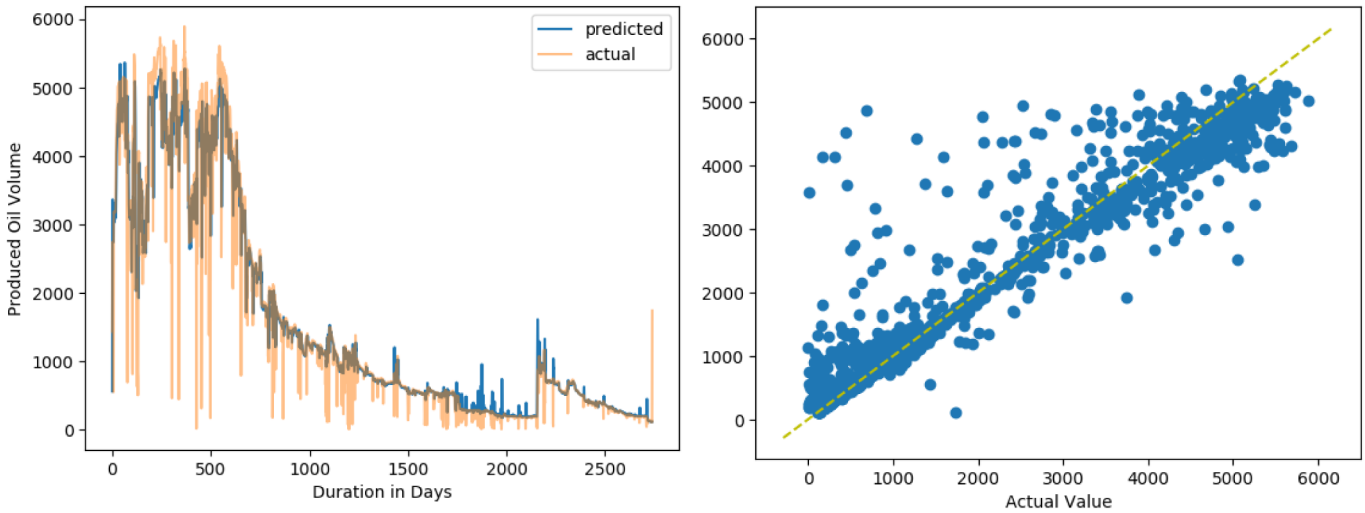


Figure 12—Model Generalization, SVR trained on well 15/9-F-14 and tested on well 15/9-F-12.

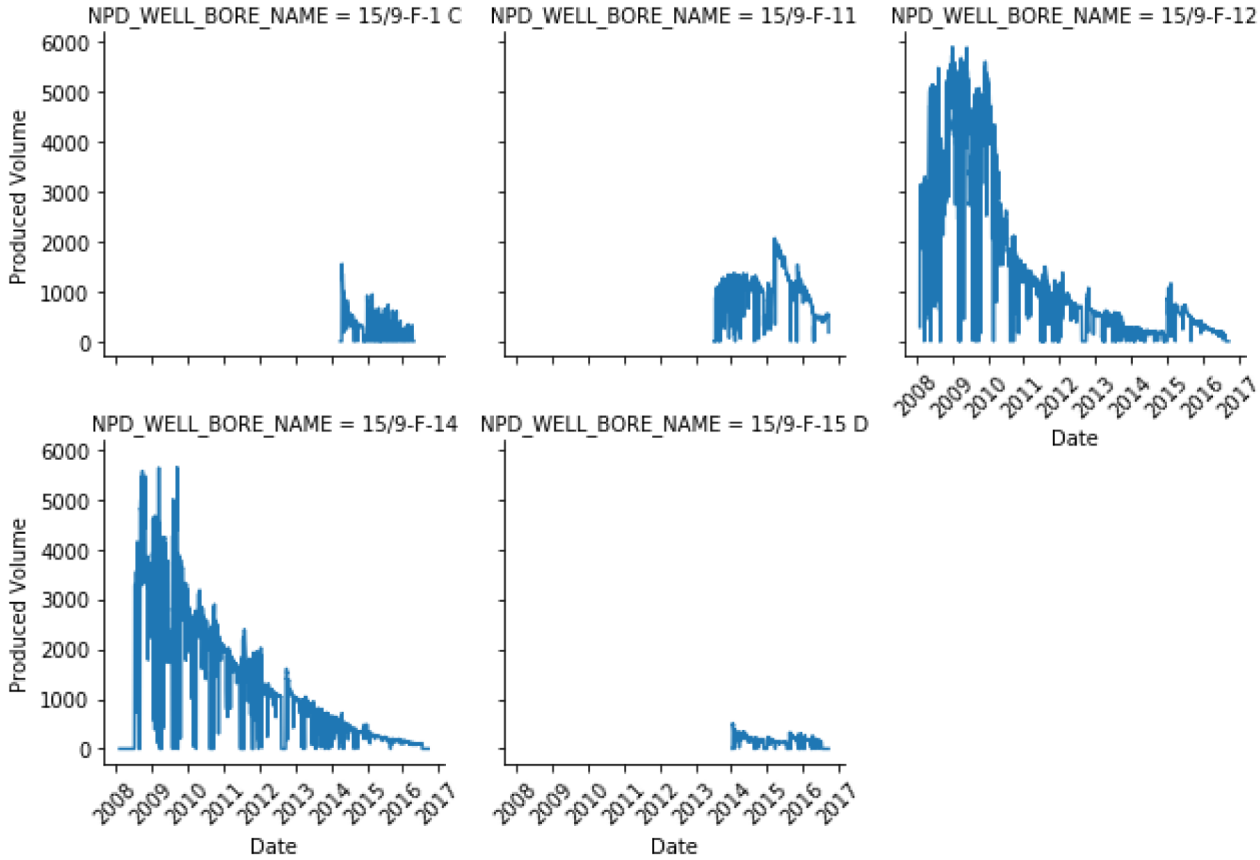


Figure 13—Zero Data observations for the produced volume of the 5 wells versus time.

Conclusions

The Machine Learning methods rely on the volume and the granularity of the data to develop the capability of predicting the production of an existing or new well based on past and offset production. The quality of prediction is highly dependent on the input data quality. If the input is noisy and inconsistent, then the prediction can be unreliable.

- Time series analysis is a simple workflow and can work in conjunction with other decline curve methods to provide more validation to forecasts from those techniques.
- The approach in this paper does not presuppose any physics of the unconventional plays and bases the future on the past history.
- The inputs used in training should be available for new wells, which enhances the generality of the model.

  Training data on well was tested on a different well and the prediction results were shown to capture the trend significantly well, mimicking its spikes and following the decline pattern smoothly.
- SVR model was tested on five different segments of the life of each of the 5 producing wells from the Volve field data set.
- When the production performance is stable, the model predicts with an accuracy of 90%.
- GBT and Adaboost were implemented on the data set but the prediction outcome was poor. Tree regressors did not work in this problem, perhaps further model tuning could enhance their performances
- More testing on the SVR model is needed to asses it's predictive capabilities as machine learning regression time series models are difficult to evaluate.

References

Clarkson, C. R. 2013. Production Data Analysis of Unconventional Gas Wells: Review of Theory and Best Practices. *Int. J. Coal Geol*. 109–110: 101–146. https://doi.org/10.1016/j.coal.2013.01.002.

Anderson, D. M., Nobakht, M., Moghadam, S. et al. 2010. Analysis of Production Data From Fractured Shale Gas Wells. Presented at the SPE Unconventional Gas Conference, Pittsburgh, Pennsylvania, 23–25 February. SPE-131787-MS. https://doi.org/10.2118/131787-MS.

Arps, J. J. 1945. Analysis of Decline Curves. *Trans. AIME* **160** (1): 228–247. SPE-945228-G. https://doi.org/10.2118/945228-G.

Bello, R. O. and Wattenbarger, R. A. 2010. Multi-Stage Hydraulically Fractured Horizontal Shale Gas Well Rate Transient Analysis. Presented at the North Africa Technical Conference and Exhibition, Cairo, 14–17 February. SPE-126754-MS. https://doi.org/10.2118/126754-MS.

Gupta, S., Fuehrer, F., & Jeyachandra, B. C., 2014, Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis. Society of Petroleum Engineers. doi: 10.2118/171588-MS.

Noshi, C. and Schubert, J.J. 2018. The Role of Machine Learning in Drilling Operations; A Review. Presented at the SPE/AAPG Eastern Regional Meeting, 7–11 October, Pittsburgh, Pennsylvania, USA. SPE-191823-18ERM-MS. https://doi.org/10.2118/191823-18ERM-MS.

Noshi, C. I., Assem, A. I., Schubert, J. J. 2018. The Role of Big Data Analytics in Exploration and Production: A Review of Benefits and Applications. Presented at the SPE International Heavy Oil Conference and Exhibition, 10–12 December, Kuwait City, Kuwait. SPE-193776-MS. https://doi.org/10.2118/193776-MS.

Noshi, C. I., S. F. Noynaert, Schubert, J.J. 2018a. Casing Failure Data Analytics: A Novel Data Mining Approach in Predicting Casing Failures for Improved Drilling Performance and Production Optimization. Presented at the SPE Annual Technical Conference and Exhibition, 24–26 September, Dallas, Texas, USA. SPE-191570-MS. https://doi.org/10.2118/191570-MS.

Noshi, C. I., S. F. Noynaert, Schubert, J.J. 2018b. Failure Predictive Analytics Using Data Mining: How to Predict Unforeseen Casing Failures? Presented at the Abu Dhabi International Petroleum Exhibition & Conference, 12–15 November, Abu Dhabi, UAE. SPE-193194-MS. https://doi.org/10.2118/193194-MS.

Schuetter, J., S. Mishra, M. Zhong and R. LaFollette, 2015. Data Analytics for Production Optimization in Unconventional Reservoirs. SPE/AAPG/SEG Unconventional Resources Technology Conference.

Shelley, B., Grieser, B., Johnson, B.J., Fielder, E.O., Heinze, J.R., and Werline J.R. 2008. Data Analysis of Barnett Shale Completions. *SPE J*. **13** (3): 366–374. SPE-100674-PA. http://dx.doi.org/10.2118/100674-PA.

Zhong, M., J. Schuetter, S. Mishra and R. Lafollete, 2015. Do data mining methods matter? A "Wolfcamp" shale case study. SPE Hydraulic Fracturing Technology Conference.