Skolkovo Institute of Science and Technology

MACHINE LEARNING ON FIELD DATA FOR HYDRAULIC FRACTURING DESIGN OPTIMIZATION

Doctoral Thesis

by

VIKTOR DUPLYAKOV

DOCTORAL PROGRAM IN PETROLEUM ENGINEERING

Supervisor
Professor Andrei Osiptsov,
Project Center for Energy Transition and ESG, Skoltech, Russia

Co-supervisor
Professor Evgeny Burnaev,
Research Center in Artificial Intelligence in the Direction of
Optimization of Management Decisions to Reduce Carbon Footprint,
Skoltech, Russia

Moscow - 2023

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate: Viktor Duplyakov.

Supervisor: Prof. Andrei Osiptsov.

Co-supervisor: Prof. Evgeny Burnaev

Abstract

Over the past two decades, the number of hydraulic fracturing jobs has increased significantly, resulting in a vast amount of data available for the development of predictive models. Analysis of post-fracturing production has shown that some stages in multi-stage fractured completions can produce unevenly due to a combination of geomechanics and fracturing design factors. Therefore, there is a need to optimise the fracturing design. The solution proposed in this thesis is a data-driven fracturing optimisation model. The workflow is divided into two logical parts: the first part involves the creation of a digital database of field data from over 6000 multi-stage hydraulic fracturing jobs in 23 oil fields in Western Siberia, Russia. This database, with its large number of data points, is a rare and representative sample compared to other datasets in the literature, which typically have only tens or hundreds of points.

The database is composed of approx. 6000 data points, each represented by a vector of 92 input variables relating to the reservoir, well, and fracture design parameters. The production data is characterized by 16 parameters, including cumulative oil production. The focus of the study is on collecting data from various sources, preprocessing the data, and developing a database architecture. Machine learning techniques are used to solve the problem of production forecasting. Missing values are filled through collaborative filtering. The production forecasting is solved using the combination of Ridge Regression and CatBoost algorithms, with a predictive ability of 64% as measured by the coefficient of determination (R^2).

The second part of the study addresses the inverse problem of selecting optimal fracture design parameters to maximize production, along with a recommendation system to guide production stimulation engineers in making informed decisions. The study began with 387 parameters characterizing each well, including construction, reservoir properties, fracture design, and production. After analysis, the model was trained using 35 key parameters as input features. The model reveals the physically explainable dependencies between the target (cumulative fluid production) and design parameters such as the number of stages, proppant mass, average and final proppant concentrations, and fluid rate. To assist field engineers in analyzing previous fracturing treatments on similar wells, we developed methods using Euclidean and cosine distance to search for similar wells. These methods were also used in a workflow to determine the optimization parameters boundaries for a pilot well during the field testing of the methodology. An inverse problem of selecting the optimal fracture design parameters to maximize production was formulated as an optimization problem and solved using four different optimization methods: surrogate-based optimization, sequential least squares programming, particle swarm optimization, and differential evolution. A recommendation system was created to guide production stimulation engineers in making informed decisions about the optimized fracture design, incorporating all the methods mentioned above.

Publications

1. A. Morozov, D. Popkov, V. Duplyakov, E. Mutalova, A. Osiptsov, A. Vainshtein, E. Burnaev, E. Shel, and G. Paderin. Data-driven model for hydraulic fracturing design optimization: focus on building digital database and production forecast. *Journal of Petroleum Science and Engineering*, 194:107504, 2020a. ISSN 0920-4105. doi:https://doi.org/10.1016/j.petrol.2020.107504

2. V. Duplyakov, A. Morozov, D. Popkov, E. Shel, A. Vainshtein, E. Burnaev, A. Osiptsov, and G. Paderin. Data-driven model for hydraulic fracturing design optimization. part ii: Inverse problem. *Journal of Petroleum Science and Engineering*, 208:109303, 2022. ISSN 0920-4105. doi:https://doi.org/10.1016/j.petrol.2021.109303

3. V. Duplyakov, A. Morozov, D. Popkov, A. Vainshtein, A. Osiptsov, E. Burnaev, E. Shel, G. Paderin, P. Kabanova, I. Fayzullin, R. Uchuev, A. Mukhametov, A. Prutsakov, I. Vikhman, and M. Staritsyn. Practical aspects of hydraulic fracturing design optimization using machine learning on field data: Digital database, algorithms and planning the field tests. 09 2020. doi:10.2118/203890-MS

4. (In progress) V. Duplyakov, A. Morozov, D. Popkov, K. Pavlenko, A. Vainshtein, V. Kotezhekov, S. Kaygorodov, B. Belozerov, V. Vanovskiy, A. Osiptsov, and E. Burnaev. Data fusion of well logs, build-up test interpretations and seismic data for reservoir permeability field estimation. *Computers and Geotechnics*, 2023

Conference proceedings

1. A. Morozov, D. Popkov, V. Duplyakov, A. Osiptsov, A. Vainshtein, E. Burnaev, E. Shel, and G. Paderin. Machine learning on field data for hydraulic fracturing design optimization: Digital database and production forecast model. pages 1–5, 01 2020b. doi:10.3997/2214-4609.202032068

2. V. Duplyakov, A. Morozov, D. Popkov, A. Vainshtein, A. Osiptsov, E. Burnaev, E. Shel, G. Paderin, P. Kabanova, I. Fayzullin, R. Uchuev, A. Mukhametov, A. Prutsakov, I. Vikhman, and M. Staritsyn. Practical aspects of hydraulic fracturing design optimization using machine learning on field data: Digital database, algorithms and planning the field tests. page 24, SPE Symposium: Hydraulic Fracturing in Russia. Experience and Prospects. Moscow, Russia, 2020

3. V. Duplyakov, V. Vanovskii, D. Popkov, A. Morozov, A. Vainstein, A. Osiptsov, S. Kaygorodov, V. Kotezhekov, B. Belozerov, and E. Burnaev. Building a permeability map of an oil reservoir by combining data from logging, well testing and seismic surveys. Intelligent data analysis in oil and gas industry. Novosibirsk, Russia, 2022

Patents

1. G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. A way to select the optimal fracturing design based on intelligent analysis of field data to increase hydrocarbon production, №2775034, Russian Federation, 2022

2. (In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "well analog selection in terms of hydraulic fracturing", Russian Federation, 2023a

3. (In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "production prediction after hydraulic fracturing based on geology and hydraulic fracture parameters", Russian Federation, 2023b

4. (In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "fracturing design optimization based on well production prediction", Russian Federation, 2023c

Acknowledgments

I am grateful to the management of LLC “Gazpromneft-STC” for organizational and financial support of this work. To M.M. Khasanov, A.A. Pustovskikh, I.G. Fayzullin and A.S. Margarit for organizational support. To E.V. Shel and G.V. Paderin for scientific guidance. The help from P.K. Kabanova and A.R. Mukhametov in data gathering is greatefully appreciated.

I would like to express my deepest gratitude to my supervisor, professor Andrey Osiptsov, whose guidance, support, and encouragement were invaluable throughout my doctoral journey. To Albert Vainstein, whose mentorship have left an indelible mark on my work. Their insightful direction has shaped me into a diligent and result-oriented individual, instilling within me a profound understanding of the importance of achieving tangible outcomes. With heartfelt gratitude, I acknowledge their invaluable contributions, which have not only enhanced my research but also nurtured my personal growth. To my co-supervisor professor Evgeny Burnaev - his expertise greatly enriched my research.

I am also immensely grateful to the members of my thesis committee: Alexander Bernstein, Alexander Shapeev, Clément Fortin, Dmitry Garagash and Egor Dontsov, for their thoughtful feedback, constructive criticism, and rigorous evaluation of my work.

I would like to thank my colleagues Dmitriy and Anton for their camaraderie, support, and intellectual stimulation. Our discussions and debates challenged me to think more critically and creatively about my research, and I am grateful for the friendship that developed as a result.

I want to express my appreciation to my incredible parents and grandparents. From the very beginning, they instilled in me values of resilience, compassion, and determination, which have propelled me forward all the times.

I am forever grateful to my wife Polina, whose unwavering love and support have been a beacon of light during the most challenging moments of these times. In the face of adversity, she has been my rock, providing boundless encouragement and standing by my side.

To my aunt Olga, my mother-in-law and to all my friends: Alexander, Andreys, Enrico, Evgenii, Gabriel, Giulia, Ilya, Ivan, Kelya, Kristina, Liza, Pavel, Sergey, Seva, Susha, whose presence in my life (and countless games in Dota) have been a source of immense joy, comfort, and encouragement.

Table of Contents

1 Introduction 13

1.1 Literature review . . . 14

1.1.1 Shale fracturing . . . 15

1.1.2 Fracturing design and its optimization . . . 16

1.1.3 ML for frac design optimization . . . 17

1.1.4 Metrics of success for a fracturing job . . . 20

1.1.5 ML methods used for HF optimization . . . 21

1.1.6 Optimization of HF parameters . . . 29

1.2 Objectives, novelty and significance of the work. . . . 30

2 Data 33

2.1 Data description . . . 34

2.2 Preprocessing . . . 36

2.2.1 Handling outliers . . . 36

2.2.2 Filling missing values . . . 37

2.3 Summary . . . 38

3 Methodology 40

3.1 Forward problem. Predicting production after hydraulic fracturing . . 41

3.1.1 Target variable . . . 41

3.1.2 Feature selection . . . 43

3.1.3 Regression . . . 44

3.1.4 Feature analysis . . . 46

3.2 Inverse problem. Choosing the optimal design . . . 46

3.2.1 Selection of optimization intervals . . . 47

3.2.2 Surrogate-based optimization (SBO) . . . 52

3.3 Summary . . . 53

4 Validation and results 54

4.1 Introductory remarks . . . 54

4.2 Forward problem . . . 55

4.2.1 Filling missing values and clustering . . . 55

4.2.2 Regression . . . 56

4.2.3 Feature analysis . . . 58

4.2.4 Field test . . . 63

4.3 Inverse problem . . . 63

4.3.1 Design optimization results . . . 64

5 Conclusions and Discussions 67
5.1 Data gathering . . . 67
5.2 Forward problem . . . 68
5.3 Inverse problem . . . 71
5.3.1 Optimization interval limits . . . 71
5.3.2 Testing optimization algorithms . . . 72
5.3.3 Optimal target selection . . . 72
5.3.4 Retrospective analysis . . . 73
5.4 Future research . . . 74

Bibliography 76

Appendix 84

List of Figures

1-1 Dependence of prediction error on data set size and model complexity 23
1-2 A schematic diagram of the neuro-genetic approach 29

2-1 Distribution of the initial data 35
2-2 Distribution of the target function values 35
2-3 Distribution of wells by the field code number 36
2-4 Distribution of missing values 39

3-1 General workflow 41
3-2 Cluster procedure: steps 1 – 5, 9 48
3-3 Cluster procedure: steps 6 – 8 48
3-4 t-SNE scatter plot for cluster visualization 50

4-1 Wells clustering by DBSCAN algorithm, represented as t-SNE visualization plot 56
4-2 Comparing two models for predicting production: CatBoost (green) vs more smooth Stacked (Ridge+CatBoost) dependences 58
4-3 Regression plot for the best model on test set 58
4-4 Statistical feature importance: Sobol sensitivity scores for the entire database. 59
4-5 SHAP feature importance for refracturing operations 60
4-6 Shap feature importance for primary fracturing operations on new wells 61
4-7 Model sensitivity analysis: tornado chart of the OVAT analysis 62
4-8 Recursive feature elimination: model's score vs. number of input parameters 62
4-9 Real vs Predicted production on the real design 63

A-1 Forward model algorithm 87
A-2 Algorithms' performance on a test set (untuned): dropping NaNs method 88
A-3 Algorithms' performance on a test set (untuned): filling by well pad . 89
A-4 Algorithms' performance on a test set (untuned): matrix imputation (collaborative filtering) 90
A-5 Algorithms' performance on a test set (untuned): filling by mean values, calculated in each corresponding cluster 91
A-6 Example of offset wells selection 91

A-7 Average optimized production for all wells by different optimization algorithms . . . 92
A-8 Average recommended fluid rate for the wells . . . 92
A-9 Average recommended mean proppant concentration for the wells . . 92
A-10 Average recommended proppant masses (per stage) for the wells . . . 93
A-11 Average recommended final proppant concentration for the wells . . . 93
A-12 Average recommended pad share for the wells . . . 93
A-13 Average recommended calculated epsilon . . . 94
A-14 Production increase with optimal parameters set . . . 94
A-15 Optimized production comparison between methods . . . 94

List of Tables

1.1 Success metrics of HF job 21

2.1 Statistics of the database 36

4.1 Metrics of the prediction models on the hold-out (30%) test set . . . 58

A.1 Features used to describe a well 86

Chapter 1

Introduction

The importance of hydraulic fracturing design optimization in the petroleum industry cannot be overstated. As the demand for hydrocarbon resources continues to rise, maximizing the efficiency and productivity of hydraulic fracturing operations has become a critical endeavor. The design of effective fracture stimulation treatments is a complex task that requires a thorough understanding of reservoir characteristics, fluid dynamics, and operational parameters.

This doctoral thesis delves into the realm of hydraulic fracturing design optimization and presents a comprehensive pipeline. The objective is to reduce the workflow burden on petroleum engineers and enable them to make well-informed decisions based on proven results from previous projects.

The primary motivation behind this work stems from the desire to address the challenges faced by engineers in optimizing hydraulic fracturing designs. Traditionally, engineers have relied on trial and error approaches, simplistic analytical or time-consuming numerical models to determine optimal fracture parameters. These methods often consume substantial time and resources, with no guarantee of achieving the desired outcomes.

The approach presented in this study leverages the wealth of available data and experiences from previously completed projects to identify the most effective design parameters for a given technological environment. By capitalizing on this knowledge base, engineers can drastically reduce the time and effort required for design optimization, leading to significant cost savings and increased operational efficiency.

Furthermore, the integration of machine learning models that expedite the estimation of cumulative fluid production after hydraulic fracturing adds another layer of importance to this work. Accurate production forecasting is crucial for effective reservoir management and decision-making. By incorporating advanced machine learning techniques, engineers can achieve faster and more accurate predictions of production, enabling them to optimize production strategies and enhance the overall performance of hydraulic fracturing operations.

The outcomes of this research have far-reaching implications for the petroleum industry. Ultimately, this work contributes to the advancement of sustainable hydrocarbon extraction, ensuring the efficient utilization of resources and driving the industry towards a more environmentally responsible and economically viable future.

The goals of the literature review are to provide a comprehensive overview of existing research and methodologies, evaluate their strengths and limitations, identify gaps in the current knowledge, understand the state-of-the-art and emerging trends, examine challenges and propose potential solutions, identify key factors influencing optimization outcomes, synthesize findings to develop a theoretical framework, and establish a foundation for further research. Through a critical analysis of the literature, this review aims to present a cohesive understanding of the current knowledge in hydraulic fracturing optimization and inform subsequent research endeavors in the field.

1.1 Literature review

Hydraulic fracturing (HF) is a widely-used method for stimulating oil and gas production from wells located in hydrocarbon-rich formations Economides et al. [1989]. The process involves pumping fluid containing proppant particles at high pressures through the tubing and into the reservoir formation, creating fractures. The fractures are then filled with tightly packed proppant particles, which provide highly conductive channels for hydrocarbons to flow from the reservoir to the well and reach the surface. The technology of HF has been in commercial use since 1947 in the United States (US), and since then has advanced in complexity, with wells now being drilled directionally with horizontal segments and multi-stage fractured completion.

The evolution of hydraulic fracturing technology has been shaped by advancements in chemistry and material science, such as the creation of fracturing fluids with controlled rheology, proppants, fibers, and chemical diverters, as well as mechanical engineering innovations like ball-activated sliding sleeves for precise stimulation of specific zones. Despite being a cost-effective method, HF still has room for improvement in terms of ultimate production as studies have shown that up to 30% of fractures in a multi-stage completion are not productive He et al. [2017], Al-Shamma et al. [2014]. Research such as the analysis in Miller et al. [2011] of distributed production logs from various stages in a near-horizontal well found that almost one third of perforation clusters do not contribute to production due to various factors such as reservoir heterogeneity and geomechanics, as well as flaws in the fracturing design. Improving the design of HF jobs can be done through either continuum mechanics modeling using commercial simulators with optimization algorithms, or through data analytics techniques applied to an actual field database. The latter approach has been taken in this study.

1.1.1 Shale fracturing

The growth of hydraulic fracturing (HF) in shale oil and gas has been driven by advancements in directional drilling and multi-stage fracturing. This has led to an increase in the world oil market supply and has turned the United States into a major supplier. The shale gas technology revolution has generated a vast amount of digital field data from multi-stage fracking operations in US shale formations. This data has become a valuable resource for data science research aimed at optimizing HF design Mohaghegh et al. [2017].

The modeling of shale reservoirs is a complex issue, with the flow mechanism not yet fully understood and accurately simulated throughout the industry. Machine learning-based pattern recognition technology can be utilized to improve full-scale simulation, providing better prediction results for shale formations such as the Bakken shale. This is demonstrated in the study by Mohaghegh et al. [2011]. Another example is the full-scale simulation of the Marcellus shale, where data-driven analytics was employed instead of traditional hydrodynamic models is the work of Esmaili and Mohaghegh [2016].

The interaction between natural and artificial fractures during the fracturing process has been studied, with the use of artificial neural networks (ANN), as shown in references such as Keshavarzi et al. [2013], Burnaev and Prikhod'ko [2013], Belyaev and Burnaev [2013], Burnaev and Erofeev [2016], Guo et al. [2014]. These studies aim to predict the behavior of artificial fractures when they encounter natural fractures. A new method for controlling the gas well reservoir model in fractured reservoirs is proposed in Guo et al. [2014] through a three-level index system of reservoir properties evaluation, based on fuzzy logic theory and multilevel gray correlation.

The authors of Schuetter et al. [2015] developed a method to differentiate high-performing wells from underperforming ones using a Wolfcamp well dataset. They used Decision Tree analysis to distinguish the top 25 wells from the bottom 25 and also selected the most significant parameters that define a successful hydraulic fracturing operation.

1.1.2 Fracturing design and its optimization

In the oilfield services industry, hydraulic fracturing job evaluation and parametric analysis are typically performed using numerical simulators based on coupled solid-fluid mechanics models. There are many HF simulators that are based on models like KGD, PKN, P3D, and Planar3D, that are used to simulate the hydraulic fracture propagation process in shale formations. The more sophisticated models for shale fracturing have been presented in studies of Detournay [2016], Osiptsov [2017]. Once a robust forward model of the process is developed, an optimization problem can be formulated with a specified objective function, as outlined in Queipo et al. [2002]. One specific example of stimulation in carbonate reservoirs is acid fracturing, which has been studied in Zoveidavianpoor et al. [2012] in an Iranian field with 20 fractured wells.

A typical approach to optimization in HF design involves the use of a surrogate model output for the objective function and integrates a hydraulic fracture simulator for predicting the fracture geometry and a production model to estimate the flow rate. The computational model calculates the objective function, which can be any chosen metric for optimizing the HF design. There have been various studies that demonstrate the implementation of this optimization strategy, such as Queipo et al. [2002] and Rahman et al. [2001]. The latter involves a multi-objective optimization workflow that couples the fracture geometry module, the hydrocarbon production module, and an investment-return cash flow module.

1.1.3 ML for frac design optimization

There has been a growing body of research focusing on the application of big data analytics to the optimization of hydraulic fracturing processes. This surge in research can be attributed to the heightened interest in multistage fracturing techniques employed in shale formations.

The emergence of big data analytics offers significant potential to improve the understanding and optimization of HF operations. By leveraging large volumes of diverse data, including geological, reservoir, production, and operational data, researchers aim to uncover valuable insights and patterns that can inform more effective HF design and decision-making.

A general workflow of the data science approach to HF for horizontal wells implicate techniques that cluster similar critical time-series into Frac-Classes of frac data (surface treatment pressure, slurry pumping rates, proppant loading, volume of proppant pumped). Correlation of the average Frac-Classes with 30-day peak production is used on the second step to distinguish geographically distinct areas Anderson* et al. [2016].

Statistically representative synthetic data set is used occasionally in the fracture model to build data-driven fracture models. The performance of the data-driven models is validated by comparing the results to a numerical model, including size, number, location, phasing angle of perforations, fluid and proppant type, rock strength, porosity, and permeability on the fracture design optimization using various fracture models. Data-driven predictive models (surrogate models, see Belyaev et al. [2016], Burnaev [2019]) are generated by using ANN and Support Vector Machine (SVM) algorithms Temizel et al. [2015]. Another approach to constructing metamodels on transient data (time series) is Dynamic Mode Decomposition (DMD), which is being explored, e.g., in Chashchin et al [2020].

Affecting geomechanics parameters are Young's modulus and Poisson's ratio obtained from lab tests on core samples, that is far away from covering full log heterogeneity with missing values, hence the authors used Fuzzy Logic, Functional Networks and ANNs Abdulraheem et al. [2009].

A detailed literature review on the subject of fracturing design optimization was provided by Gao and You [2017], where the authors emphasized the necessity of bringing into the full scale shale gas systems a common integrating approach. The data-driven analytics was proposed as a trend in the HF design optimization. Authors induced game-theoretic modeling and optimization methodologies to address the multiple stakeholders.

The impact of proppant pumping schedule during the job has been investigated in Poulsen et al. [1986] by coupling fractured well simulator results and economical evaluations.

There are several approaches when different target parameters are considered as criteria to optimize. For a wide range of reasons, the proppant fraction is quite an important criteria to investigate. In Saldungaray et al. [2012] the authors made a significant step forward gaining 4 major case studies based on shale low-permeable reservoirs across the US and suggesting strategy to evaluate the realistic conductivity and impact on stimulation economics of proppant selection.

Field data, largely accumulated over the past decades, are being digitized and structured within oil companies. The market landscape in the era of declining oil prices after 2014 has stimulated shale operators to look closer at the capabilities of data science to optimize the fracturing technology Betz et al. [2015]. The issue of working with short-term data and need to find a way to turn that into long-term well performance was emphasized. Proppant loading was shown to be one of the most important variables for productivity. Increasing industry interest to artificial intelligence and to application of ML algorithms is justified by the coincidence of several points: processing power growth and amount of data available for analysis. Thousands of completions are digitalized (e.g., see Awoleke et al. [2011]), giving the grounds for the use of a wide range of big data analytics methods. One of the most recent studies Wang and Chen [2019b] investigated the relationships between the stimulation parameters and first-year oil production for a database of horizontal multistage fractured wells drilled in unconventional Montney formation in Canada. Four commonly used supervised learning approaches including Random Forest (RF), AdaBoost, SVM, and ANN Hastie et al. [2009] were evaluated to demonstrate that the RF performs the best in terms of prediction accuracy.

The state of affairs is a bit different in other parts of the world, where, though the wells are massively fractured, the data is not readily available and is not of that high quality as in the North America, which poses a known problem of "small data" analysis, where neural networks do not work, and different approaches are called for.

In Russia, there are a few attempts of using ML algorithms to process data of HF, e.g., the paper Alimkhanov et al. [2014] presents the results of developing a database of 300 wells, where fracturing was performed. Operational parameters of the treatments were not taken into account in this paper. Classification models were developed to distinguish between efficient/inefficient treatments. Job success criteria were suggested in order to evaluate the impact of geological parameters on the efficiency via classification. Regression models were proposed for predicting post-frac flow rate and water cut. A portfolio of standard algorithms was used such as decision tree, random forest, boosting, ANNs, linear regression and SVM. Limitations of linear regression model applied for water cut prediction were discussed. Recent study Makhotin et al. [2019] used gradient boosting to solve the regression problem for predicting the production rate after the simulation treatment on a data set of 270 wells. Mathematical model construction task was formulated in detail, though data sources and the details of data gathering and preprocessing were not discussed.

1.1.4 Metrics of success for a fracturing job

The ultimate optimization of a stimulation treatment is only possible if the outcome is measured. Below, various approaches to quantify the success of an HF job are presented:

- Cumulative oil production of 6 and 18 months is used by Wang et al. [2016] as a target parameter, and is predicted by a model with 18 input parameters, characterizing Bakken formation in North America.

- Predictive models for the 12 months cumulative oil production are built by Schuetter et al. [2018] using multiple input parameters characterizing well location, architecture, and completions.

- Feed-forward neural network was used by Awoleke et al. [2011] to predict average production for wells drilled in Denton and Parker Counties, Texas, of the Barnett shale based on average monthly production. The mean value was evaluated using the cumulative gas produced normalized by the production time.

- In Balen et al. [1988a], a procedure was presented to optimize the fracture treatment parameters such as fracture length, volume of proppant and fluids, pump rates, etc. Cost sensitivity study upon well and fracture parameters vs NPV as a maximization criteria is used. Longer fractures does not necessarily increase NPV, a maximum discounted well revenue is observed by Hareland et al. [1993].

- Statistically representative set of synthetic data served as an input for an ML algorithm in Temizel et al. [2015]. The study analyzed the impact of each input parameter to the simulation results like cumulative gas production for contingent resources like shale gas simulation model.

- $\Delta Q = (Q_2 - Q_1)$ was an uplift metric to seek the re-fracture candidate for 50 wells oilfield dataset using ANN to predict after the job oil production rate Q_2 based on Q_1 oil production rate before the job Yanfang and Salehi [2014].

Metrics	Source
Cumulative oil production 6/18 month just after the job	Wang et al. [2016]
12 months cumulative oil production	Schuetter et al. [2018]
Average monthly oil production after the job	Awoleke et al. [2011]
NPV	Balen et al. [1988a]
Comparison to modelling	Temizel et al. [2015]
Delta of averaged Q oil	Yanfang and Salehi [2014]
Pikes in liquid production for 1, 3 and 12 months	Pankaj et al. [2018]
Break even point (job cost equal to total revenue after the job)	Alimkhanov et al. [2014]

Table 1.1: Success metrics of HF job

- Q pikes approach is presented by implementing B1, B2 and B3 statistical moving average for one, three and twelve-month best production results consequently in Pankaj et al. [2018]. The simulation is done over 2000 dimension dataset to reap the benefit from proxy modeling treatment.

- Net present value is one of the metrics used to evaluate the success of a HF job Balen et al. [1988b]. Economical bias for HF is detailed by Balen et al. [1988a]. The proposed sequential approach of integrating upstream uncertainties to NPV creates an important tool in the identification of the crucial parameters affecting a particular job.

In Table 1.1 the list of the most common metrics for evaluation of HF job efficiency is presented.

1.1.5 ML methods used for HF optimization

ML is a broad subfield of artificial intelligence aimed to enable machines to extract patterns from data based on mathematical statistics, numerical methods, optimization, probability theory, discrete analysis, geometry, etc. ML tasks are the following: classification, regression, dimensionality reduction, clustering, ranking and others. Also, ML is subdivided into supervised/unsupervised and reinforcement learning.

Supervised ML problem can be formulated as constructing a target function $\hat{f} : X \to Y$ approximating f given a learning sample $S_m = \{(x_m, y_m)\}$, where

$x_m \in X$, $y_m \in Y$ with $y_i = f(x_i)$.

To avoid overfitting (discussed in the next section), it is also very important to select an ML model properly. This choice largely depends on the size, quality and nature of the data, but often without a real experiment it is very difficult to answer which of the algorithms will be really effective.

The lack of data becomes one of the most common problems when dealing with field data. Some ML models can manage it (decision trees), some are very sensitive to sparse data (ANNs). A number of the most popular algorithms such as linear models or ANNs do not cope with the lack of data, SVMs have a large list of parameters that need to be set, and trees are prone to overfitting.

The choice of the model and the choice of the initial sample can highly affect the final results and the correct interpretation.

There are articles with results on application of ML to HF data that describe models with high predictive accuracy. However, the authors use small samples with rather homogeneous data and complex models prone to overfitting. Therefore, more research is needed to evaluate prediction accuracy and stability separately for different fields and well types.

1.1.5.1 Overfitting

Today, there is an increasing trend in the number of papers on the application of ML in the field of HF data processing. However, many of them could lead a reader to question the validity of the results, which could be erroneous due to overfitting.

Overfitting is a negative phenomenon that occurs when the learning algorithm generates a model that provides predictions mimicking a training dataset too accurately, but have very inaccurate predictions on the test data Hastie et al. [2009]. In other words, overfitting is the use of models or procedures that violate the so-called Occam Razor Hawkins [2004]: the models include more terms and variables than necessary, or use more complex approaches than necessary. Figure 1-1 shows how the pattern of training for test and training datasets changes dramatically if overfitting takes place.

There are several reasons for this phenomenon Hawkins [2004], Baumes et al.

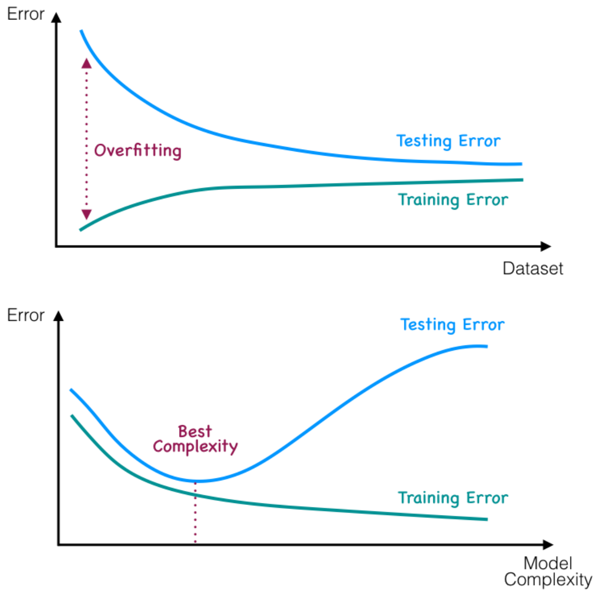


Figure 1-1: Dependence of prediction error on data set size and model complexity

[2006]:

- Traditional overfitting: learning a complex model on a small amount of data without validation. This is a fairly common problem, especially for industries that not always have access to big datasets, such as medicine, due to the difficulties with data collection.

- Parameter tweak overfitting: usage of a learning algorithm with too many hyperparameters. Selection of the parameters based on the performance on the test set.

- Bad statistics: misuse statistics to overstate confidence. Often some known-false assumptions about some system are made and then excessive confidence of results is derived. E.g. we use Gaussian assumption when estimating confidence.

- Incomplete prediction: use an incorrectly chosen target variable or its incorrect representation. E.g. there is a data leak and inputs already contain target variable.

- Human-loop overfitting: a human is still a part of the learning process, he or she selects hyperparameters, creates a database from measurements, so we should take into account overfitting by the entire human/computer interaction.

For example, in the article Alimkhanov et al. [2014] only 289 wells, each described by 178 features, were considered for the analysis. This number of points is too small compared to the number of input features, so a sufficiently complex predictive model simply “remembers” the entire dataset, but it is unlikely that the model is robust enough and can provide reliable predictions. This is also confirmed by a very large scatter of results: the coefficient of determination varies from 0.2 to 0.6.

In this context you can find many articles, which used small data: e.g. in Mohaghegh et al. [2002] — 150 wells, or in Esmaili and Mohaghegh [2016] — 135 wells, etc. Also, each of the mentioned articles uses a very limited choice of input features, which exclude some important stages of the HF. For example, article Lolon et al. [2016] uses the following parameters to predict the quality of the HF performed: stage spacing, cementing, number of stages, average proppant pumped, mass of liquid pumped, maximum treatment rate, water cut, gross layer thickness, oil gravity, Lower Bakken Shale TOC, Upper Bakken Shale TOC, total vertical depth. Such set of parameters does not take into account many nuances, such as the geomechanical parameters of the formation or the completion parameters of the well.

Quite good results were shown by the authors of the article Schuetter et al. [2015]; they also investigated various models. But as noted in the article from 476 wells, only 171 have no NaN values.

In addition to the problems described above, overfitting may be caused by using too complex models: in many articles they use one of the most popular ML methods, the artificial neural network (ANN). But it is known that a neural network is a highly non-linear model that very poorly copes with the lack of data and is extremely prone to overfitting. Lack of data is a fairly frequent case when it comes to a real field data, which makes the use of ANNs unreliable.

The authors of the article use the SVM algorithm Xiaofeng et al. [2016]; the main disadvantage of SVM is that it has several key hyperparameters that need to be set correctly to achieve the best classification results for each given problem. The same hyperparameters can be ideal for one task and not fit at all for another. Therefore, when working with SVM a lot of experiments should be made, and the calculation takes a fairly large amount of time. Moreover, a human-loop overfitting can occur.

The above algorithms work very poorly with missing values, and so additional tricks are needed, which often leads to data leak or to various types of overfitting. Among other things these models are not easily interpretable.

In conclusion, to reduce overfitting and construct a robust predictive model, the necessary condition is to develop a big and reliable training dataset that contains all required input features.

1.1.5.2 Dimensionality reduction

When a dataset has large number of features (high dimensionality), it can lead to a long ML algorithm computation time as well as difficulties in finding a good solution due to excessive noise in data. In addition, for higher dimensionality we need more examples in the data set to construct a reliable and accurate predictive model. In addition, this problem greatly increases the likelihood that two input points are too far away, which, like in case of outliers, leads to overfitting. Therefore, in order to decrease input dimensionality and at the same time to keep the completeness of information with its decreasing, we can use dimensionality reduction and manifold learning methods, see Ma and Fu [2011], Kuleshov et al. [2018]. Lastly, this trick helps us visualizing multidimensional data. In our work, we used T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm van der Maaten and Hinton [2008] for visualization after dimensionality reduction and missing values imputation.

1.1.5.3 Clustering

Clustering methods Hastie et al. [2009] are used to identify groups of similar objects in multivariate datasets. In other words, our task is to select groups of objects as close as possible to each other, which, by virtue of the similarity hypothesis, will form our clusters. The clustering belongs to the class of unsupervised learning tasks and can be used to find structures in data. Since our database includes 23 different oilfields, horizontal and vertical wells, as well as different types of fracture design, it would be naive to assume that data is homogeneous and can be described by a single predictive model.

Thus, by dividing dataset in clusters we can obtain more homogenuous subsamples, so that ML algorithms can easily construct more accurate models on subsamples Grihon et al. [2013]. In addition, clustering can be used for detecting outliers Aggarwal and Sathe [2017], Smolyakov et al. [2019] in a multidimensional space. We utilise this for further analysis. In our case, we used t-SNE to visualize a low-dimensional structure of the data set to extract clusters and identify outlying measurements.

1.1.5.4 Regression

After selecting a specific sample of data, it is necessary to solve the regression problem, i.e., to restore a continuous target value y from the original input vector of features x Forrester et al. [2008], Belyaev et al. [2016]. The dependence of the mean value $\mu = f(x)$ of y on x is called the regression of y on x.

In the reviewed articles other authors considered different approaches how to define a target variable. In particular, they considered cumulative production for 3, 6 and 12 months. However, we noted a strong correlation between values of cumulative production for 3, 6 and 12 months. Thus, as a target variable we consider only values of cumulative production for 3 months.

After building a regression model we assess its accuracy on a separate test sample. As a prediction accuracy measure we use the coefficient of determination and mean absolute percentage error (MAPE).

1.1.5.5 Ensemble of models

The ensemble of models Hastie et al. [2009], Burnaev and Prikhod'ko [2013] uses several models in order to obtain better prediction efficiency than could be obtained from each trained model individually.

Ensembles, due to their high flexibility, are very prone to overfitting, but in practice, some assembly techniques, such as bagging, tend to reduce overfitting. The ensemble method is a more powerful tool compared to stand-alone forecasting models, since it minimizes the influence of randomness, averaging the errors of each basic model and reduces the variance.

1.1.5.6 Feature analysis

The use of tree-based models makes it easy to identify features that are of zero importance, because they are not used when calculating prediction. Thus, it is possible to gradually discard unnecessary features, until the calculation time and the quality of the prediction becomes acceptable, while the database does not lose its information content too much.

There is the Boruta method Kursa and Rudnicki [2010] which is a test of the built-in solutions for finding important parameters. The essence of the algorithm is that features are deleted that have a Z-measure less than the maximum Z-measure among the added features at each iteration. Also, the Sobol method Sobol [2001a] is widely used for feature importance. The method is based on the representation of the function of many parameters as the sum of functions of a smaller number of variables with special properties.

In addition, testing and verifying feature importance can be done with the one-variable-at-a-time (OVAT) method Daniel [1973]. It is a method of creating experiments involving testing of parameters one at a time instead of multiple factors simultaneously. It is primarily used when data is noisy and it is not obvious which features affect a target.

1.1.5.7 Hyperparameter search

Hyperparameter optimization is the problem of choosing a set of optimal hyperparameters for a learning algorithm. Whether the algorithm is suitable for the data directly depends on hyperparameters, which directly influence overfitting or underfitting. Each model requires different assumptions, weights or training speeds for different types of data under the conditions of a given loss function.

The most common method for optimizing hyperparameters is a grid search, which simply does a full search on a manually specified subset of the hyperparameter space of the training algorithm. But before using the grid search, a random search can be used to estimate the boundaries of a region where parameters are selected. Moreover, according to the Vapnik-Chervonenkis theory, the more flexible a model is, the worse its generalizing ability. Therefore, it is very important to select the model complexity conforming to the given data set, otherwise prediction will be unreliable. To check the generalization ability we can use a cross-validation procedure.

1.1.5.8 Uncertainty Quantification

Uncertainty comes from errors made by an ML algorithm and from noise in a data set. Hence, predicting an output only is not sufficient to be certain with results. Therefore we should also quantify uncertainty of the prediction. This can be done by using prediction intervals providing probabilistic upper and lower bounds on an estimate of the output variable.

The prediction interval depends on some combination of the estimated variance of the model and the variance of the output variable caused by noise. The variance of the model is due to variance of model parameters estimates, resulted from noise in the original data set. By building confidence intervals for the parameters and propagating them through the model we can estimate the variance of the model. In practice to build prediction interval for a general nonlinear model we can use the bootstrap resampling method, although it is rather computationally demanding Hastie et al. [2009].

It is important to note the difference between prediction and confidence interval: the former quantifies the uncertainty on a single observation, estimated from the population, and the latter quantifies the uncertainty on an estimated population variable, such as a mean or a standard deviation. Also, it is necessary to quantify uncertainty on ML model performance which we can do by estimating the corresponding confidence intervals.

Besides prediction or confidence intervals another important type of uncertainty quantification is related to forward uncertainty propagation when we estimate how the variability of input variables of the model influence the output variability to select the most important input parameters Sobol [2001a], Burnaev et al. [2017].

1.1.6 Optimization of HF parameters

There are published studies on optimization of HF parameters based on some relatively moderate data sets with hundreds of wells, e.g., see a study on 570 wells in Clinton sand Mohaghegh S [1996]. The first stage is a neural network which takes as input fracturing completions data and production history and predicts post-frac deliverability. The model serves as a screening tool for excluding wells, which cannot be considerably enhanced after a frac job. The second stage is based on a neural network, where the branch of the network, responsible for the fracture design parameters, is connected to the optimization algorithm to obtain the optimized frac design for each well and the expected post frac deliverability (see Fig. 1-2).

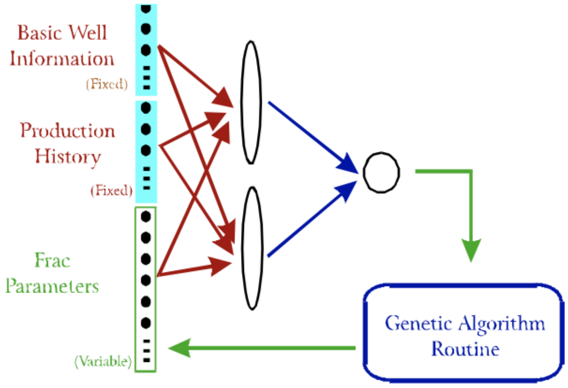


Figure 1-2: A schematic diagram of the neuro-genetic approach

The work E. T. Woldemariam, H. G. Lemu [2019] also utilizes a neuro-genetic algorithm in the optimization pipeline instead of a computationally complex implicit numerical hydrodynamic model. In particular, the article presents a case study of neuro-genetic optimization using the example of a cross-flow microturbine. In another paper Wang and Chen [2019a] an ML regression analysis on more than 3500 wells was performed, the authors optimized the design by visual analysis with taking a pair of the most important parameters and selecting a region of the most optimal design.

Speaking about other industries, the article Shi et al. [2019] uses neural networks and a hybrid multisubgradient descent method with adaptive learning rate to solve multicriteria optimization of multiple tasks (generation of too large droplets and too low droplet speed) in the field of bioprinting. The proposed method can improve both printing accuracy and stability, and is useful for realizing precise cell arrays and complex biological functions.

Another application of machine learning and optimization is the design optimization of thin multilayer solar cells to maximize external quantum efficiency Kaya and Hajimirza [2019]. The authors utilize the concept of transfer learning which implies the use of experience gained in solving one problem to solve another, similar problem. The reason why the transfer learning is applied to the problem is because the problem involves the change of design specifications. Therefore, the transfer learning model acts as a function of surrogate optimization, which refits the surrogate more efficiently. In particular, the procedure improved the results by 2-3 times using only half of the training samples compared to the usual model.

1.2 Objectives, novelty and significance of the work.

HF technology is a complex process, which involves accurate planning and design using multi-discipline mathematical models based on coupled solid Detournay [2016] and fluid Osiptsov [2017] mechanics. At the same time, the comparison of flow rate prediction from reservoir simulators using fracture geometry predicted by HF simulators vs. real field data suggests there is still significant uncertainty in the models.

In contrast to the traditional approach to making the design of fracturing technology based on parametric studies with an HF simulator, it is proposed to investigate the problem of design optimization using ML algorithms on field data from HF jobs, including reservoir, well, frac design, and production data. Training database would be real field data collected on fracturing jobs in Western Siberia, Russia.

The entire database from real fracturing jobs can be conventionally split into the input data and the output data. The input data consists of the parameters of the reservoir and the well (permeability, porosity, hydrocarbon properties, etc.) and the frac job design parameters (pumping rate, proppant mass and concentraion, etc.). The output is a production parameter.

The usefulness of hybrid modeling is widely reported in the literature Helmy et al. [2010]. Numerous efforts have been made by researches to implement data science to lab cost reduction issues. PVT correlations correction for crude oil systems were comparatively studied between ANN and SVM algorithms El-Sebakhy et al. [2007].

Finally, the problem at hand is formulated as follows: one may suppose that a typical hydraulically-fractured well does not reach its full potential, because the fracturing design is not optimal. Hence, a scientific question can be posed within the big data analysis discipline: what is the set of fracturing design parameters, which for a given set of the reservoir characterization-well parameters yields maximum post-fracturing production? In order to answer this question it is proposed to develop a ML algorithm, which would allow one to determine the optimal set of HF design parameters based on the analysis of the available field data.

The recommendations should be made on

- oil production forecast based on well information;
- the optimum frac design;
- data acquisition systems, which are required to improve the quality of data analytics methods.

The following hypotheses and research questions will be checked in the frame of the current research:

1. Are there systematic problems with HF design?
2. What is the objective function for optimization of HF design? What are various metrics of success?
3. How to validate the input database?
4. What database is full (sufficient)? (Optimum ratio of number of data points vs. number of features for the database?)
5. What are the essential parameters one can get from the field data to construct a predictive model and optimize the HF design?

6. Is there a reliable ML-based methodology for finding the optimum set of parameters to design a successful HF job?

   As a result, a complete HF design optimization pipeline should be developed. This pipeline should include the collection of all necessary data, its preprocessing, and the prediction of the target value for further optimization.

   It is expected that the time to implement such a pipeline should be reasonably short. It should outperform existing methods that involve time-consuming hydrodynamic modeling.

Chapter 2

Data

Following the report by McKinsey&Company from 2016 the majority of companies get real profit from annually collected data and analytics Henke et al. [2016]. However, the main problem companies usually face while getting profit from data lies inside the organizational part of the work.

Most of the researches skip the phase of data mining, considering the ready-made dataset as a starting point for ML. Nevertheless, we can get misleading conclusions from false ML predictions due to learning on the low-quality dataset. As follows from results of Gaurav et al. [2017] the most important thing when doing the ML study is not only a representative sample of the wells, but also a complete set of parameters that can fully describe the fracture process with the required accuracy.

As can be seen from Section 1.1.5.1, where we describe various types of overfitting, the most important one is related to a poor quality of the training dataset. In addition, if in case of a non-representative training dataset we use a subsample of it to train the model, corresponding results will be very unstable and will hide the actual state of affairs.

It is known that data preprocessing actually takes up to 74% of the entire time in every data-based project Press [2016]. Having a good, high-quality and validated database is the main key to obtain the interpretable solution using ML. The database must include all the parameters that are important from the point of view of the physics of the process, be accurate in its representation and verified by subject domain experts in order to avoid the influence of errors in database maintenance.

Unfortunately, in field conditions each block of information about different stages of the HF is recorded in a separate database; as a result of this work there is no integrated database containing information about sufficient number of wells that would include all factors for decision making. So, a right way for data preprocessing should be used in order to make a given data set more useful, work – more efficient and results – more reliable.

2.1 Data description

All necessary information is collected from the following sources (Fig. 2-1):

- Frac-list — a document with a general description of the process and the main stages of proppant pumping;
- Monthly production report (MPR) — a table with a production history data collected monthly after the conducted fracturing;
- Operating practices — geological and operational data collected monthly;
- Geomechanics data — stress contrasts, Poisson's ratio, strain modules for each of the formations;
- PVT — physical properties of the fluids for each of the formation;
- Layers intersection data;
- Well logs interpretation data.

Frac-list was selected as the primary source due to the volume of crucial stage-by-stage data and the existence of all ID keys such as field, well, layer and date of HF. It is worth mentioning that frac-list is full of manually filled entries. Moreover, the fact that operations where pumping was interrupted (STOP, or a screen-out) are not necessarily tagged, makes the problem more complex.

Each source from the above list was processed individually, depending on the specifics before merging them together. Particularly, monthly data have been consolidated in 3-, 6- and 12-months slices. Fig. 2-2 shows the distribution of cumulative fluid production for 12 months.

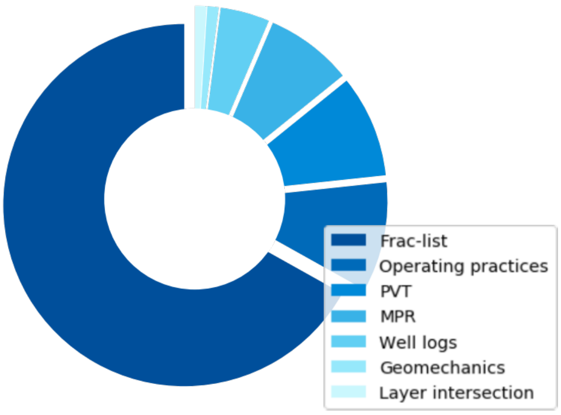


Figure 2-1: Distribution of the initial data

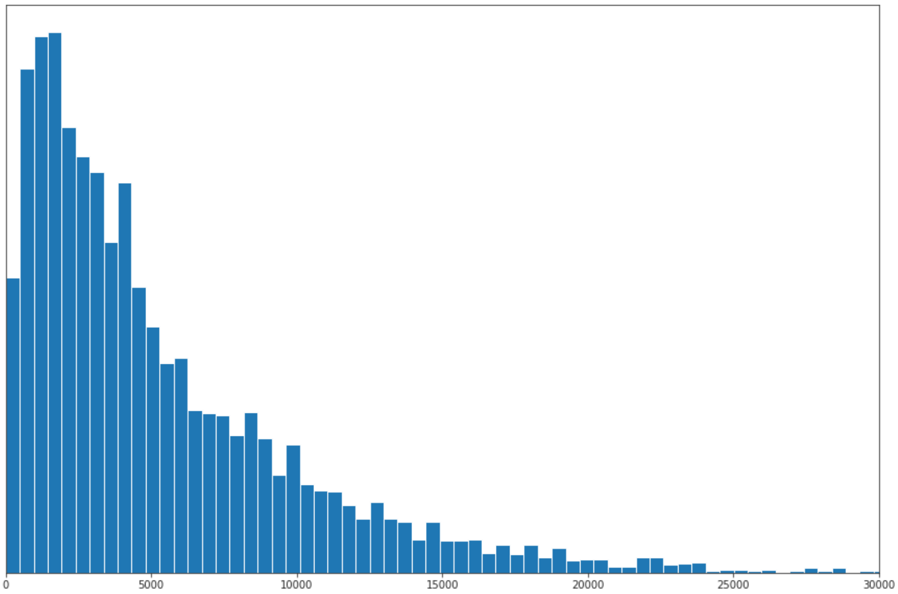


Figure 2-2: Distribution of the target function values

Some illustrative numbers of the initial database are presented in Table 2.1 and Figure 2-3. Each oilfield is coded with a number (we avoid specific oilfield names for confidentiality reasons, in agreement with the operator). It is worth mentioning that the word 'operation' (in legend and tables) refers to the entire stimulation treatment, which may be a single stage fracturing on a vertical well or a multistage fracturing on a near-horizontal well. Then, a multi-stage treatment (operation) is divided into different stages. Each stage is characterized by the set of fracturing design parameters.

Parameter	Numerical value
Observation period	2013 – 2019
Number of oil fields	22
Number of wells	5425
– vertical & directional	4111
– horizontal	1314
Number of fracturing operations	6687
– single-stage treatment	3177
– multi-stage treatment	3510
– refracturing operations (out of total)	1460
Number of STOPs (e.g. screenout)	797
Initial number of input parameters	296
Final x vector of input parameters	92
– formation	36
– well	12
– frac design	44
Number of production parameters	16

Table 2.1: Statistics of the database

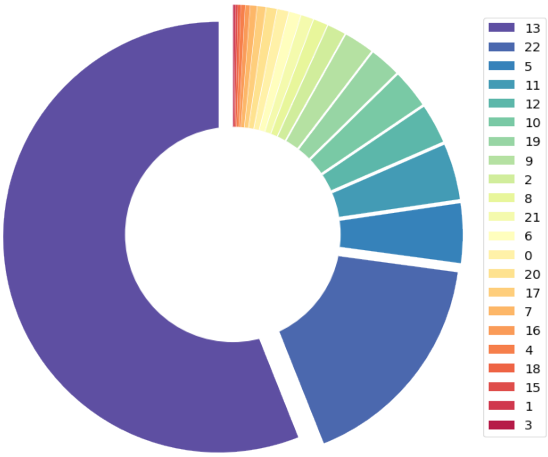


Figure 2-3: Distribution of wells by the field code number

2.2 Preprocessing

2.2.1 Handling outliers

Outliers, i.e. objects that are very different from the most of observations in the data set, often introduce additional noise to an ML algorithm Aggarwal and Sathe [2017]. Outliers can be classified into three types: data errors (measurement inaccuracies, rounding, incorrect records), which occur especially often in case of field data; the presence of noise in objects descriptions; suspiciously “good” or “bad” wells in terms of production; the presence of objects from other populations, e.g., corresponding to significantly different field geologies.

To effectively detect such observations several techniques were used. First of all, statistical methods to analyse data distribution along different dimensions and detected outliers by estimating the kurtosis measure and other statistics.

The second approach is clustering. Clustering has been carried out using the Density-based spatial clustering of applications with noise (DBSCAN) algorithm Ester et al. [1996], because it does not require an a priori number of clusters to be specified in advance, and is able to find clusters of arbitrary shape, even completely surrounded, but not connected. Even more importantly, DBSCAN is quite robust to outliers. As a result, outliers are concentrated in small blobs of observations.

Another method that eliminated more than a hundred questionable values was an anomaly detection method called Isolation Forest Liu et al. [2008], which is a variation of a random forest. The idea of the algorithm is that outliers fall into the leaves in the early stages of a tree branching and can be isolated from the main data sample.

2.2.2 Filling missing values

It is very often that certain features of some objects of the field data sets are absent or corrupted. Moreover, many of ML algorithms like SVM regression or ANNs require all feature values to be known. Considering the structure of the data sources, we could expect that the frac-list has contributed to the majority of such cases of data incompleteness, since this document contains most of the useful data yet it is typically filled in manually, hence it is highly dependent on the quality of the filling process (Fig. 2-1 & Fig. 2-4).

As a result there is a number of methods that allow one to fill in the missing or Not a Number values (NaNs). However, it should also be noted that most approaches can be overly artificial and may not improve the final quality.

We test several approaches to fill in missing values within the framework of the regression problem under consideration:

- dropping objects containing more NaNs within an object (j-th row of a matrix representing the data set) than a certain threshold (65%). For example, if a well has 33 missing feature values out of 50, then we drop it. Among other imputation methods described below, this would keep the database as original as possible;
- filling NaNs of i-th parameter by the average for the wells in a well cluster, that are grouped by geography. The reason for selecting this method of filling in is that the wells of the same cluster have similar frac designs and geology properties of the reservoir layer;
- filling missing values by applying imputation via collaborative filtering (CF) Adomavicius and Tuzhilin [2005]. CF is often used by recommender systems, which makes prediction of absent values with the use of mathematical methods. According to our research, the best results were shown by non-negative matrix factorization (NNMF) and truncated singular value decomposition (TSVD). Worth noting, NNMF cannot handle negative values, such as skin-factor.
- applying unsupervised learning to define similarity clusters and filling NaNs of i-th parameter by the mean of the cluster. In other words, the average of the feature is taken not from the entire database, but from the cluster, which allows us to estimate the missing value more accurately.

2.3 Summary

The chapter focuses on describing the data used in the study and the preprocessing steps undertaken. The data description provides an overview of the data set, while the preprocessing section discusses handling outliers and filling missing values. Throughout this process, several problems were encountered, including data integration, heterogeneity, and consistency. Data integration posed challenges due to the need to combine information from multiple sources, each with its own structure and format. The heterogeneity of the data, characterized by variations in data

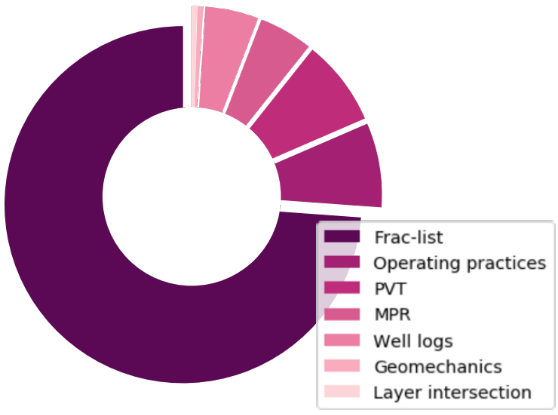


Figure 2-4: Distribution of missing values

types, units, and scales, required careful standardization, verification and normalization. Ensuring data consistency across different sources proved to be a significant issue, as inconsistencies in naming conventions, coding schemes, and data quality were addressed during the preprocessing stage. Despite these challenges, the chapter highlights the steps taken to address the encountered problems and prepare the data for further analysis and use.

Chapter 3

Methodology

The methodology chapter provides an extensive overview of the methods and algorithms employed in the study. In this section, we will delve into the intricacies of the research approach, highlighting the tools and techniques utilized to address the problem of hydraulic fracturing optimization. The objective is to provide a clear understanding of the analytical framework that underpins the study.

This chapter discusses the various methods and algorithms employed to tackle the challenges associated with HF optimization. The focus lies in harnessing the power of big data analytics to extract valuable insights from vast and diverse datasets. The application of advanced data processing, machine learning, and statistical modeling techniques is explored to uncover hidden patterns, establish predictive models, and guide the decision-making process in HF design.

The specific algorithms and techniques employed in the study are examined. This includes the application of clustering techniques and Euclidean distance calculation to identify similar wells and establish interval limits for design parameters.

By presenting a comprehensive account of the methods and algorithms utilized in this research, a solid foundation is established for the subsequent analyses and findings presented in this study.

Figure 3-1 visually illustrates general workflow of this work and also emphasizes several challenges and describes methods that emerged in the course of the study.

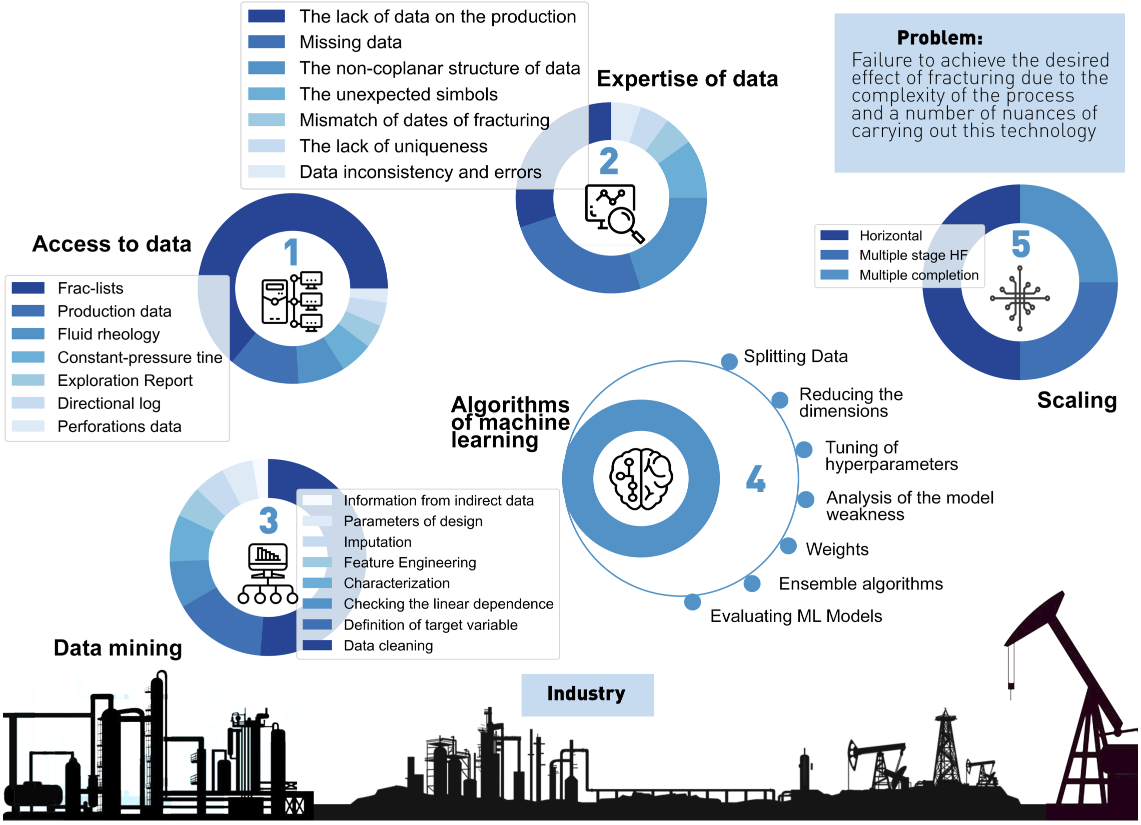

The lack of data on the production
Missing data
The non-coplanar structure of data
The unexpected simbols
Mismatch of dates of fracturing
The lack of uniqueness
Data inconsistency and errors
Expertise of data
2
Problem:
Failure to achieve the desired effect of fracturing due to the complexity of the process and a number of nuances of carrying out this technology
Access to data
1
Frac-lists
Production data
Fluid rheology
Constant-pressure tine
Exploration Report
Directional log
Perforations data
Horizontal
Multiple stage HF
Multiple completion
5
Scaling
Splitting Data
Reducing the dimensions
Tuning of hyperparameters
Analysis of the model weakness
Weights
Ensemble algorithms
Evaluating ML Models
Algorithms of machine learning
4
3
Information from indirect data
Parameters of design
Imputation
Feature Engineering
Characterization
Checking the linear dependence
Definition of target variable
Data cleaning
Data mining
Industry

Figure 3-1: General workflow

3.1 Forward problem. Predicting production after hydraulic fracturing

The first goal of the work is to predict production after hydraulic fracturing. After that we can use the resulting model in further optimization process. The model should include all necessary parameters we can get from a well, characterization of its geological environment and HF design parameters. The vector of these parameters should be sufficient to predict the target production variable.

3.1.1 Target variable

Selection of the right target variable is crucial for the success of the entire optimization workflow. Based on the available data and discussion of current fracture design development strategies by reservoir stimulation engineers, it was concluded that fracturing operations in the field are optimized based on the metric of maximizing reservoir contact: the larger the fracture, the better. Hence, bigger fractures yield larger cumulative volume of recovered total fluid (both oil and reservoir water), and there is a strong correlation. It is also important to consider that the oversized fractures may break through into upper or lower strata, causing additional water production.

A model aimed at predicting the cumulative volume of total fluid produced was trained on existing data with slightly higher accuracy (as there is a direct correlation between the input parameters characterizing the fracture design and the total cumulative fluid) compared to a model aimed at cumulative oil only, where the prediction accuracy was lower as the oil production is less correlated with the design parameters governing the fracture dimensions. The present realization of the model does not take into account the presence of water-bearing layers, so the model better predicts the total volume of produced fluid. During testing, we deal with the particular oilfield where there are no bottom waters nearby target formation, so there is no risk of a breakthrough into the aquifer.

At the same time, generalization of the present workflow to oilfields with the presence of aqueous layers will require modifications: some features characterizing the presence of upper/lower water bearing formations should be included to be able to predict the production of oil and water separately, and also the target variable should be composed of two components: maximum total fluid and maximum pure oil (or minimum water cut, which is equivalent).

The 3-month cumulative fluid production data was used to utilize as much wells as possible, including the most recently fractured wells, where the production history is short.

The later fracturing was carried out, the less production history we know. So, the shorter the considered period - the more HF experience is expressed in the dataset. Therefore, 3-month (or precisely 90 days) production data slices were chosen as representative sets.

An interpolation of the monthly data was used in order to come up with 90-day production period. This approach is suitable, for instance, if a well worked for 20 days in the first month after fracturing treatment and 30 days in the following months, the desired value is in 80-110 days or 3-4 months range and can be found through the interpolation.

Last but not least, the target variable should be chosen with respect to the further business metric for the entire project. This metric should be money related. The main goal of the project might be to increase the revenue for a particular well while keeping the cost of HF treatment as low as possible. In this case, the cumulative fluid production could be easily converted to the cost of oil produced, minus the operating costs. Then, using the workflow presented, the design of the HF can be optimized to meet all the necessary business requirements.

3.1.2 Feature selection

To construct the models and increase interpretability of the problem feature importance analysis was used. It can be done with or without the involvement of an approximation model.

One of the methods is a statistical sensitivity analysis via Sobol indices Sobol [2001b]. It decomposes the variance of the target variable into parts attributed to input features. This method does not involve a constructed approximation model.

On the contrary, SHAP method Lundberg, Scott M., and Su-In Lee [2017] is based on an approximation model, utilizes the concept of Shapley values Shapley [1953] and measures features importances in terms of predictive power of each feature. It shows the true importance of the features more accurately Song et al. [2016]. The SHAP values can be calculated for tree-based models (which were used).

Another way of analyzing parameters is the feature elimination procedure which implies a reduction of feature space which in its turn may improve the performance of the approximation model. Feature elimination can be achieved by applying a pairwise correlation of parameters via the Spearman correlation and by the Recursive Feature Elimination (RFE) method which involves the use of the approximation model. Speaking of the former, the reason of choosing the Spearman correlation instead of the Pearson correlation is because most of the features correlations are non-linear. So, by estimating correlations between features we can remove perfectly correlated features and features with zero variance. Regarding the second method, RFE is a procedure for backward selection of features which works as follows. First, a model is built on an the entire set of input features, and features' importance is calculated. Then, the least important features are removed. After that, the model is rebuilt on the reduced set of features and we repeat the process.

Eventually, the following procedure for feature elimination was used:

1. Select 3-month slices; (6- and 12-months production & geological and technical data are removed due to the lack of data from the latest treatments);
2. Remove parameters which are not relevant for the considered target variable, or for which more than 80% of the observations are missing;
3. Estimate Spearman correlations to identify perfectly correlated features; remove features with almost zero variance;
4. Apply RFE to select features, which are the most important for the target prediction.

As the result, initial set of features has been reduced from 387 to 35 features.

3.1.3 Regression

Several ML models have been chosen to predict cumulative fluid production over 3 months.

Several ML regression algorithms were used, including: SVM, KNN, ANN, Decision Trees, and various types of ensembles based on decision trees such as Random Forest, ExtraTrees, CatBoost, LGBM and XGBoost.

Each model was trained on a subsample with cross validation on 5 folds. Then, models were tested on a separate (hold-out) sample. All these sets were shuffled and had similar target value distributions. Most of the ML models are decision tree-based and, hence, have important advantages: they are fast, able to work with any number of objects and features (including categorical ones), can process data sets with Not a Number values (NaNs) and have a small number of hyperparameters.

Each experiment is conducted two times on four data sets constructed using different imputation techniques:

- on the entire dataset containing information about 5425 wells. Here, we used hyperparameters of the regression algorithms set to their default values first. Then, after figuring out the best imputation technique, we proceed to the next experimental setup described below;

- on wells from one field only; again, we used default hyperparameters of the regression algorithms.

The reason to use two experimental setups is to check if more homogeneous dataset enhances predictive performance of the model.

Then, we take the best performing methods based on the R^2 on test set of each experiment, tune their hyperparameters via the grid search, and combine them into an ensemble to further improve the results. If the result of the ensemble of models is worse than the single best regressor, then we are taking the results of the best regressor.

In the stacked approach, multiple models are combined to leverage their strengths and compensate for their weaknesses. By using an ensemble of models, each with its own strengths and characteristics, a more comprehensive and accurate representation of the real-world behavior of the process can be achieved.

The stacked approach allows for the incorporation of different algorithms or techniques that complement each other. For example, the inclusion of a linear regression model can capture linear dependencies that may be overlooked by the gradient boosting trees algorithm. By combining these models, the ensemble can better capture the complexities and nuances of the process, resulting in improved predictive performance.

The stacked approach provides a more robust and flexible framework for modeling the process, allowing for a better understanding of its behavior and more accurate predictions. By leveraging the strengths of multiple models, the limitations of any single algorithm can be overcome, leading to a more reliable and comprehensive solution.

3.1.4 Feature analysis

Feature importance analysis is performed for an ensemble of the best algorithms. OVAT analysis is carried out to see how the target varies with the variation of the design parameters. In addition, if the feature rankings of both methods are more or less similar, then we may proceed to parameter reduction. With the available feature importance values, we iterate over a range to remove less important parameters and then calculate the R^2 score. This procedure is important for the design optimization, because it reduces the dimensionality of the problem while keeping the best score.

3.2 Inverse problem. Choosing the optimal design

An inverse problem can be formulated as optimizing a high dimensional black box (BB) with respect to inputs constrained by boundaries. In this case BB is a function with unknown expression or internal structure that, given a list of inputs, returns corresponding outputs. The high dimensionality of the input presents an exponential difficulty for problem modeling and optimization (so-called "curse-of-dimensionality") Bellman and Dreyfus [2015]. To optimize a high-dimensional computationally expensive black box (HEB) function, it is required to iteratively evaluate an objective function, which can be costly and so becomes unacceptable. In our case, the HEB function is represented by the constructed ML regression. To optimize HEB, the following optimization methods were used: *surrogate-based optimization (Sec. 3.2.2), sequential least squares programming Fu et al. [2019], particle swarm optimization Bonyadi and Michalewicz [2017] and differential evolution Storn and Price [1997] algorithms.* The advantages of these methods are that they make no assumptions about the problem being optimized and can perform searches in very large spaces of candidate solutions.

For presented methodology, the goal is to maximize the cumulative fluid production by finding a set of optimal design parameters constrained by boundaries (see Sec. 3.2.1.1) for the specified parameters of the pilot well environment.

3.2.1 Selection of optimization intervals

After constructing the model, the inverse problem is formulated as finding a set of optimal fracturing design parameters to maximize the target. Since the model is multi-dimensional, a valid selection of the design parameter values is only possible if they change within the relevant intervals during the optimization procedure. These intervals are caused by various constraints, arising in the field.

The additional restrictions on design parameter values can be divided into geological and technological constraints. Geological constraints include those related to the geological structure of the formation. For example, proximity of gas or water bearing formations leads to the necessity to limit fracture height growth, which automatically leads to limitation of the maximum volume of injected proppant and requires change of the perforation strategy. Geological constraints can also be related to waterflooding cases. For example, when an injection well, operating at bottomhole pressures higher than the formation breakdown pressure, is located relatively close to the production well and is in the direction of fracture propagation, it is necessary to limit the maximum fracture half-length on the production well.

Technological limitations include those related to the technical capabilities of the equipment and chemicals used. For instance, the maximum fracturing pressure (first of all, it is connected with the capabilities of the wellhead equipment and pumping units) can lead to the restriction on the maximum fracture width. Those limitations at a first approximation could be obtained by offset (similar) well search methods.

In the field, a fracturing job is pumped with variable proppant concentration, so another limitation we add is a parameter characterising the rate of increase in the proppant concentration in the fracturing fluid from initial (c_{start}) to final (c_{fin}) concentration $[kg/m^3]$ with respect to increase to average c_{avg}. The parameter is denoted as ϵ and is defined as:

$$\epsilon = \frac{c_{fin} - c_{start}}{c_{avg} - c_{start}} - 1. \tag{3.1}$$

Boundaries for this parameter are the same for all pilot wells: from 0.5 to 1.5. This is not a parameter that needs to be optimized, so these bounds are introduced as a constraint for optimization algorithms.

3.2.1.1 Clustering for offset wells selection

To find optimization intervals for fracturing design parameters we define a pilot cluster as a set of wells which are similar to the pilot one (the same field, layer, face and direction). We estimate design parameters limits as 5^{th} and 95^{th} percentiles of the values of the parameters for wells, belonging to the constructed cluster. Thus, the results of the optimization would be more robust, as we do not use extrapolated model's predictions, which could be a problem for tree-based ML algorithms.

Another approach to obtain optimization boundaries for the specified pilot well is to utilize unsupervised ML methods like clustering and dimensionality reduction. The procedure is as follows (See also figs. 3-2, 3-3):

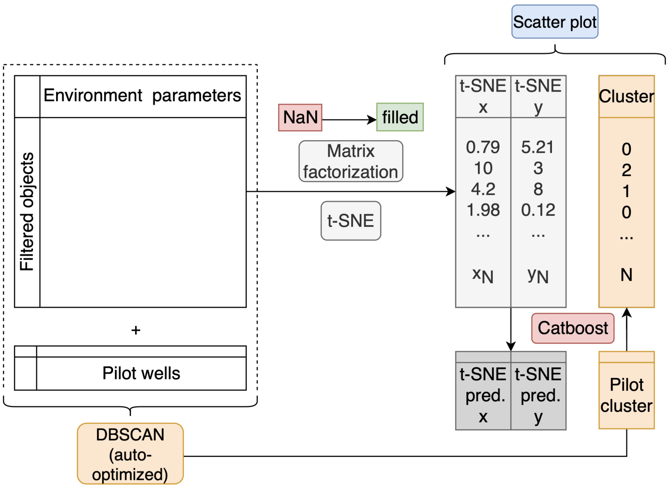


Figure 3-2: Cluster procedure: steps 1 – 5, 9

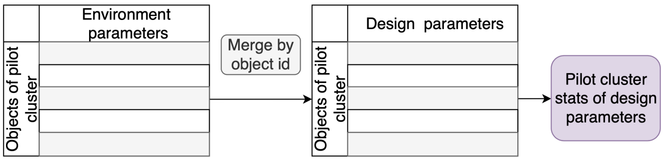


Figure 3-3: Cluster procedure: steps 6 – 8

1. Remove all features from the database except the environment parameters: PVT, geomechanics, well log interpretation, well id;

2. Filter the obtained subset by the number of stages, layer id and face of the pilot well. This step is needed in order to collect wells technically similar to the pilot well;

3. Apply a clustering algorithm to the environment parameters of the filtered data subset. Hyperparameters of the clustering method are optimized via a gradient-free optimization algorithm by maximizing the mean silhouette coefficient, calculated for a particular cluster as follows:

   $$S_i = \frac{b_i - a_i}{max(a_i, b_i)}, \tag{3.2}$$

   where a_i is a mean intra-cluster distance, b_i is a mean nearest-cluster distance;

4. Find a cluster to which the pilot well belongs to. Assert this cluster as a pilot well cluster;

5. Using id-s of the objects from the pilot cluster, add the design parameters values to the objects' descriptions;

6. Analyze the pilot cluster statistics: minimum, maximum and mean values of the design parameters. The minimum values and the maximum values serve as the optimization boundaries while the mean values of the design parameters are used as an initialization for the optimization problem.

7. Perform a visual analysis via the t-distributed stochastic neighbor embedding (t-SNE) algorithm based on the environment parameters. t-SNE can only be applied to data without missing values. Hence, the missing values are imputed by the matrix factorization algorithm Morozov et al. [2020a]. After that we obtain the t-SNE embedding for the selected subset of data. Here t-SNE embedding is a mapping of the multidimensional input features to the two-dimensional x and y coordinates for each observation of the data subset. These two-dimensional coordinates can be used for visualization of the considered data subset.

8. After that, using the considered subset of the data and the obtained t-SNE

coordinates build an ML regression model to predict t-SNE x and y coordinates for any new object. The corresponding ML model takes as input environment parameters and returns as output t-SNE coordinates x and y. The reason of predicting the t-SNE coordinates is because any change in the data like appending a new pilot well to the data set will require running the t-SNE algorithm from scratch;

9. Predict the t-SNE coordinates x and y for the pilot well using the constructed regression model. This allows to visualize the position of a new well with respect to the selected data subset;

10. Create a t-SNE scatter plot from the x and y t-SNE coordinates. Label the corresponding clusters of each object by some colors and label the pilot well by a star symbol. This procedure visually verifies how the data is clustered and how the pilot well is located with respect to the remaining data set.

The example of this method implementation is shown in Fig. 3-4.

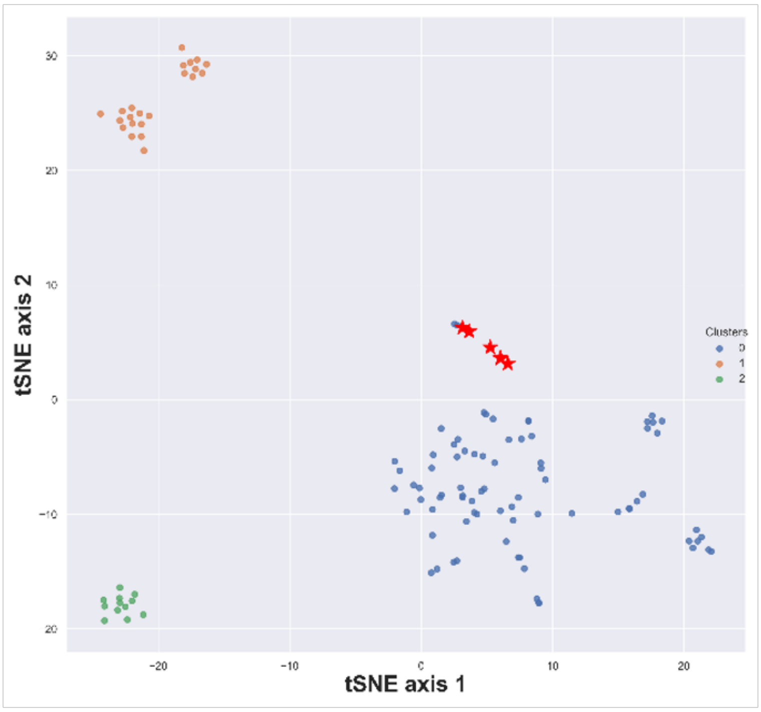


Figure 3-4: t-SNE scatter plot for cluster visualization

3.2.1.2 Missing pilot parameters imputation

In practice some of the parameters values of a well can be missing. Imputation is needed for wells parameters values because gradient-free optimization algorithms are based on constructing regression models and so they cannot work with missing values. For better results, it is important to impute these values not just by filling them with corresponding mean values, but to do it in a smarter way. Several strategies are proposed:

- a matrix factorization method, described in Sec. 2.2.2;
- impute missing values by averaging parameters values of the top-N similar wells. The method to find similar wells is described in Sec. 3.2.1.3;
- use the mean pilot cluster parameters values as described in step 6 of the algorithm in the previous subsection.

3.2.1.3 Offset wells selection by Euclidean distance

Offset wells are the wells, similar to the pilot one in terms of their geological surroundings. One may look for these wells both in terms of their geological and geographical (closest wells within certain radius from the pilot one) similarities. These analogue wells search is very useful for a petroleum engineer as it allows to analyse fracturing operations, conducted previously, their design parameters values, check whether an operation was successful or not, etc. We can also extract additional features from the neighbouring wells, which increase predictive power of the models Erofeev et al. [2021]. For example, in this work average fluid production divided by distance from the pilot well was tested as a feature. Wells within 1 km from the pilot one were considered.

As a similarity metric we can use the Euclidean distance between the pilot well and the other wells. To calculate it, we firstly need to normalize values in the database. Here, linear min-max normalization seems to be a valid choice, where min and max values are the 1^{st} percentile and the 99^{th} percentile respectively. The usage of percentiles is due to possible outliers in the initial data set.

Certain input features exhibit a low variance, while others do not. In order to account for this, the similarity metric calculation incorporates weights that correspond to each feature's standard deviation values over the subset.

Then, the Euclidean distance between two vectors of parameters, characterizing wells, is calculated as

$$d(p,q) = \sqrt{w_1(p_1 - q_1)^2 + \cdots + w_n(p_n - q_n)^2}, \tag{3.3}$$

where p_i and q_i are the i-th parameter's values of the corresponding wells and w_i is its standard deviation. The distances are calculated between the pilot well and all other wells, belonging to the same cluster (within the same field, layer and face). The results of such similar wells search are considered robust and sustainable by field geologists. One can see an example of a result of such search in Fig. A-6.

It is worth noting, that some features require special consideration. For example, if we use the well's azimuth, we need to consider, that 0 and 180 degrees would be the same.

In this work a combination of the clustering method for selecting offset wells along with the Euclidean similarity search was used. The clustering is used for obtaining the set of similar wells, then we reduce the size of this set, leaving only top-N wells by the Euclidean distance. This method allows us to look for optimal values of the design parameters in a certain vicinity where the prediction model works well. Also, in some cases the top-N similar wells can be used for imputing missing parameters for the pilot well.

3.2.2 Surrogate-based optimization (SBO)

Surrogate models (approximation models) Belyaev et al. [2016] can be used with multi-dimensional input design spaces. As we know, the more parameters the surrogate model takes as input the more computational resources (training time) it is required to construct the model.

A particular example of Surrogate-based Optimization (SBO) or sequential model-based global optimization is Bayesian optimization. The algorithm utilizes a probabilistic surrogate function which approximates the expensive objective function and an acquisition function which, in its turn, allows to select a new candidate input design point within the domain. At the beginning, a surrogate function is built on a small number of samples from the original objective function, which provides a target value. Then, the algorithm maximizes the acquisition function and the surrogate is updated with a new input point and its actual output value. After repeating the process, an optimum can be achieved by taking the information about the samples from the past iterations. A typical choice of the surrogate is Gaussian processes and random forest, while typical acquisition functions are *the Expected Improvement* (EI) and *the Probability of Improvement* (PI).

The SBO algorithm implemented in the industrial software was used. The SBO methodology is based on Gaussian processes modeling technique Burnaev et al. [2016], Zaytsev and Burnaev [2017], Burnaev and Zaytsev [2015]. The particular numerical realization roots in the scientific works published in Burnaev and Panov [2015].

3.3 Summary

The methodologies discussed in this chapter serve as a foundation for the subsequent analyses and findings presented in this study. By integrating predictive modeling techniques, feature analysis, and optimization methods, a comprehensive understanding of hydraulic fracturing processes could be developed. These approaches could enable informed decision-making regarding HF design, ultimately leading to improved production outcomes and increased efficiency in reservoir development.

Overall, the methodologies presented in this chapter establish a robust framework for addressing the challenges of HF optimization and provide a pathway for the subsequent chapters, where a deeper analysis will be conducted, and the results and implications of the study will be discussed.

Chapter 4

Validation and results

Throughout this chapter, a detailed analysis of the experimental results is presented, highlighting the key findings and their significance in the context of HF optimization. By examining the validation and experimental outcomes, we aim to validate the effectiveness of presented methodologies and contribute to the existing body of knowledge in the field.

The chapter provides empirical evidence and a comprehensive understanding of the implications and effectiveness of the approaches utilized in this research. The outcomes of this analysis lay the groundwork for the subsequent discussions and conclusions presented in the next chapter.

4.1 Introductory remarks

During the research, the model was trained on wells from the field of interest only, thereby reducing the database from 6687 to 3308 fracturing operations.

Additionally, the database is divided into two parts: primary and repeat stimulations. During testing, only primary stimulations (on new wells) were used.

Field tests have been carried out on 21 wells, which were not included in the training dataset. 9 wells were horizontal, 7 — vertical multilateral, operating on several layers simultaneously, the rest 5 wells were regular vertical ones. We were testing the accuracy of our prediction models, as well as HF design optimization overall pipeline. This pipeline includes:

1. obtaining design parameters optimization boundaries by similar wells search with euclidian distance;

2. imputing missing parameters with mean corresponding parameters values, calculated using top-10 similar wells, obtained within the pilot cluster via the Euclidean distance search;

3. performing the design optimization, using the predictive model and optimization algorithms.

Models accuracy checks were performed firstly on a hold-out set (separated from the training data set) and secondly, on the wells from field tests.

4.2 Forward problem

4.2.1 Filling missing values and clustering

By applying the first method of missing data imputation, rows with more than 42 NaNs have been dropped. Then, the rest of the NaNs for other objects in the data set are filled with their mean values.

Fig. 4-1 shows how the entire database is clustered with DBSCAN algorithm. Since the algorithm itself cannot handle missing values, we recover them using the collaborative filtering. Then we assign cluster labels for each well to the original data set. To visualize clusters, t-SNE is applied to transform data space into 2D and build a scatter plot. As seen from the figure, there are 3 groups in total with the biggest cluster marked as “2”

Once the database is created, four imputation methods of filling NaNs are evaluated in terms of their performance. After applying these imputation methods to the database, the results of the best tuned ML algorithms for regression problem are as follows (R^2):

- Matrix factorization: $R^2 = 0.64$;

- Dropping the entire row, if NaN’s count more than 65% in that row: $R^2 = 0.56$;

- Filling with mean values of the well pad: $R^2 = 0.47$;
- Filling with mean values of the cluster: $R^2 = 0.49$.

Comparison between different ML algorithms with these filling methods can be seen in Figs. A-2, A-3, A-4 and A-5. Matrix factorization appeared to be the most effective method, so it was applied to the entire dataset. Handling negative values (skin-factor) has been done via introducing a binary parameter, which shows whether the skin is negative or not.

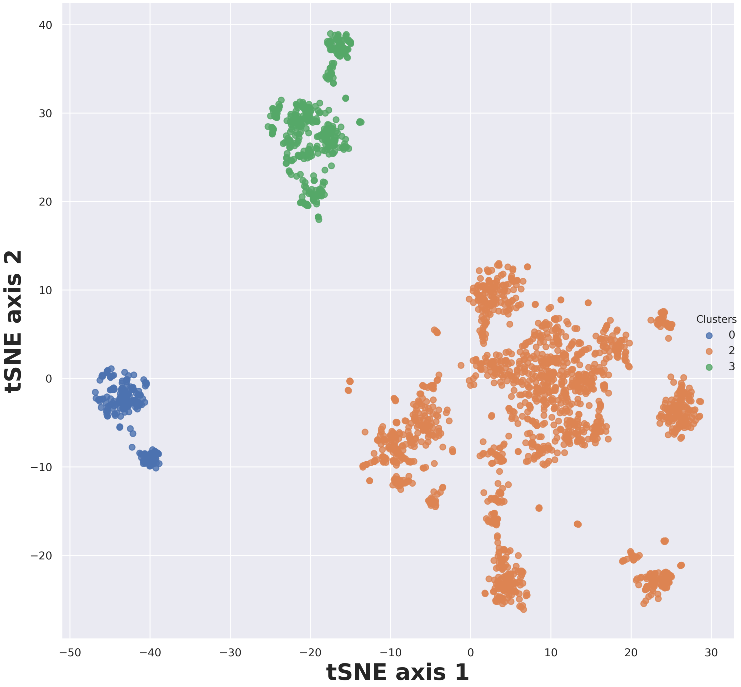


Figure 4-1: Wells clustering by DBSCAN algorithm, represented as t-SNE visualization plot

To conclude, filling missing values is useful for further work with predicting models. The best imputation technique is matrix factorization.

4.2.2 Regression

The results are shown in the table 4.1 and on the regression plot 4-3.

Comparison of various regression algorithm performance can also be seen on Figs. A-2, A-3, A-4 and A-5. It is worth noting that:

- The family of decision-tree based algorithms show better accuracy than other approaches. CatBoost algorithm (based on gradient boosted decision trees) outperforms all other methods;

- Some of the ML algorithms like SVM or ANN resulted in negative R^2, which is interpreted as poor prediction accuracy. The possible explanation is that some methods are preferred when there are homogeneous/hierarchical features like images, text, or audio, which is not our case;

- The best imputation technique is collaborative filtering;

- Based on the log scale regression plot, a relatively large amount of errors comes from the points with too low or too high oil production rates. The possible solution of the problem is to perform regression for different clusters;

However, during testing, we found that using only the CatBoost algorithm did not yield feasible results. When examining the relationship between the target variable (fluid production) and the proppant mass feature, for example, we observed non-monotonic dependencies with peaks in the CatBoost model. This is not desirable, as increasing proppant should lead to increased production due to longer fractures. This phenomenon can be attributed to the limitations of gradient boosting models, which may not always capture obvious dependencies due to the complexity of an algorithm.

To address this issue, the following pipeline was implemented: first, the Ridge regression model (a form of linear regression with L2 regularization) was trained on a smaller subset of features to capture any linear dependencies that are critical to building a robust model for hydraulic fracturing processes. Next, we subtracted the linear regression predictions from the actual target variable values and attempted to predict the residuals using CatBoost on all available features.

As a result, despite the slight decrease in predictive power (Tab. 4.1), it shows more physically meaningful predictions from the resulting ensemble of the models, as depicted in the accompanying figure 4-2.

Table 4.1: Metrics of the prediction models on the hold-out (30%) test set

Metric	**CatBoost**	**Stacked**
RMSE	1670	1713
MAE	1131	1165
R2	0.64	0.62
MAPE	36.22%	37.78%
wMAPE	29.07%	29.95%

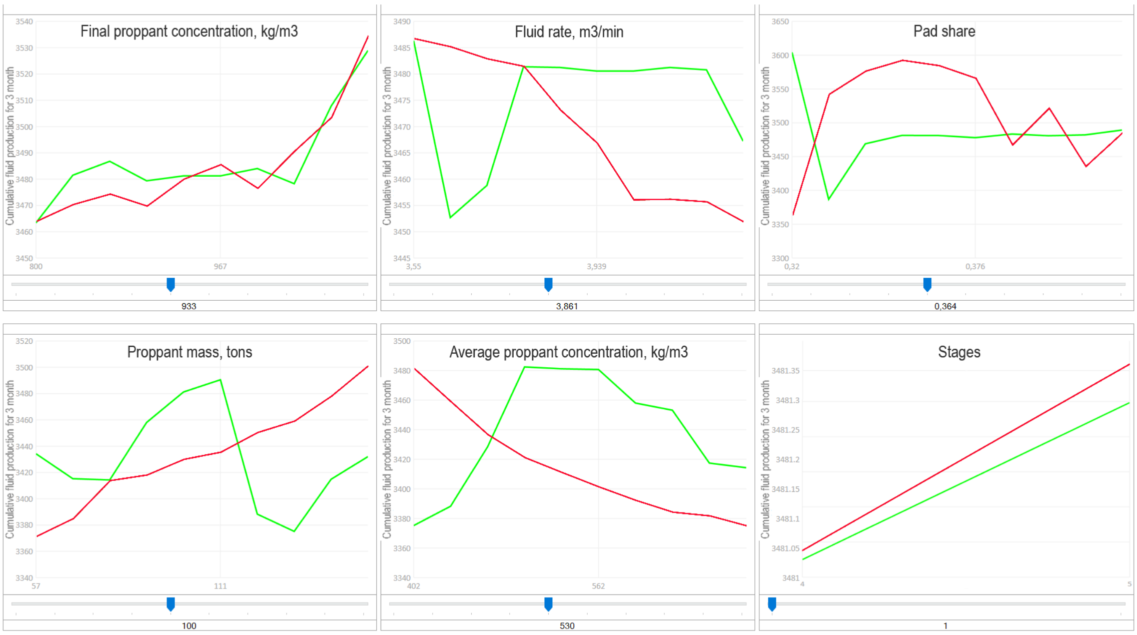


Figure 4-2: Comparing two models for predicting production: CatBoost (green) vs more smooth Stacked (Ridge+CatBoost) dependences

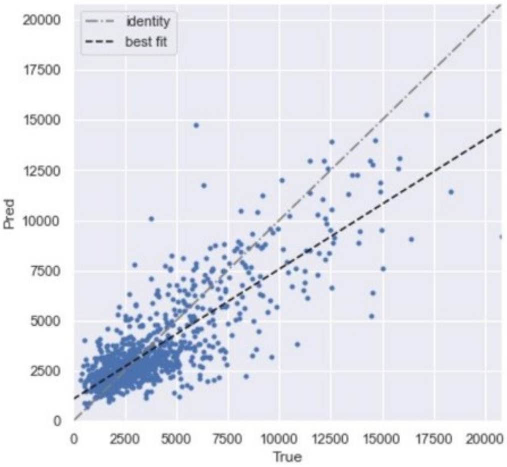


Figure 4-3: Regression plot for the best model on test set

4.2.3 Feature analysis

The most important features, calculated via Sobol indeces can be seen in Fig. 4-4.

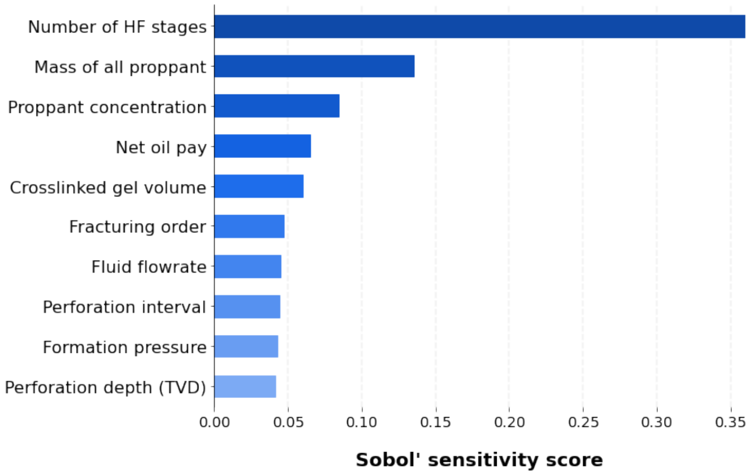


Figure 4-4: Statistical feature importance: Sobol sensitivity scores for the entire database.

In this context, the crucial features are associated with the hydraulic fracturing (HF) design. Alongside the design factors, two significant features are the net oil pay and formation pressure, which characterize the reservoir characteristics.

Then, feature importance analysis using Shapley method was introduced. At first, the concept of Shapley values has been used to visually point out the difference between the two models for primary stimulation vs refracturing. (Figs. 4-5, 4-6). (In case of categorical features one-hot encoding was used. The corresponding features are prefixed with “cat.”.) Particularly, for refracturing operations the key feature having major impact on the target is the level of production before the refracturing treatment (which was not captured in previous analysis), which is absent in case of primary operations. Having data on production prior to refracturing makes the production forecast problem easier to solve, compared to the case of production forecast after primary fracturing operations. This has also been noted in other studies Erofeev et al. [2021].

From the analysis conducted, it can be deduced that certain proppant properties, specifically density and size, do not play a significant role in predicting fluid production. Conversely, the crucial features for prediction are linked to, for example, a pad (fluid used to initiate hydraulic fracturing that does not contain proppant).

Figure 4-7 shows a Tornado chart of the OVAT analysis for design features. These features are essential to further optmization pipeline. The most relevant features

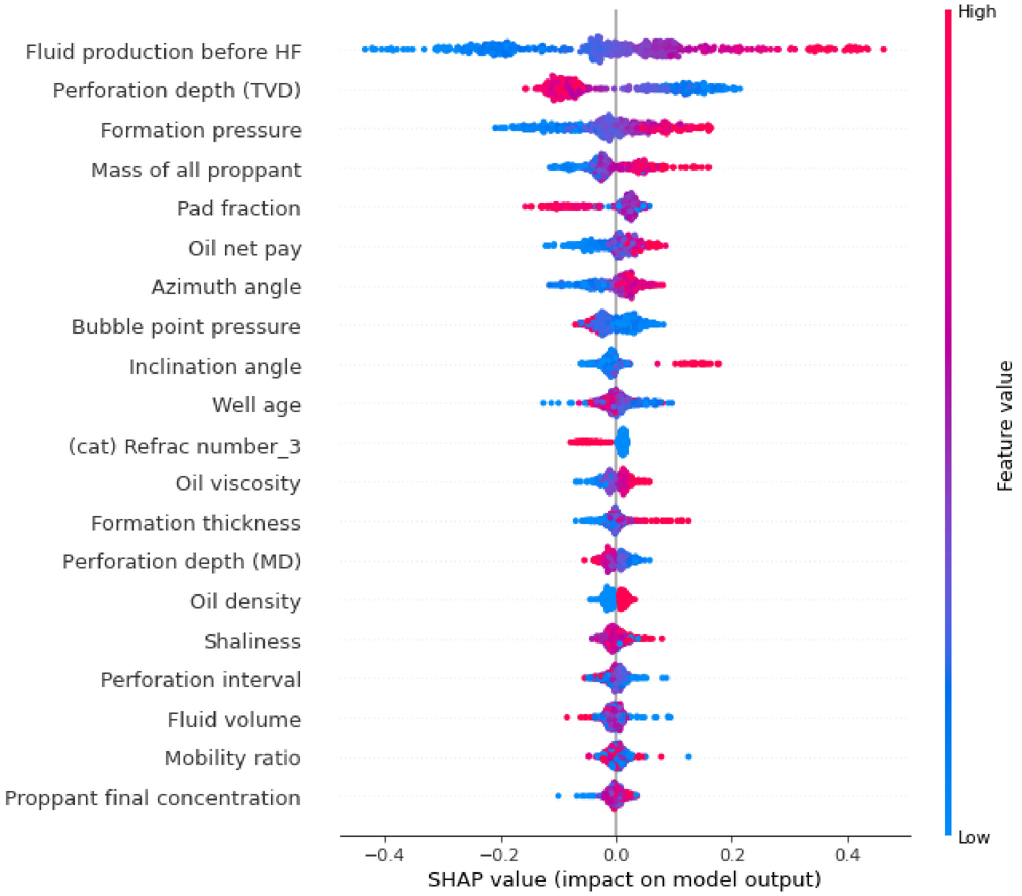


Figure 4-5: SHAP feature importance for refracturing operations

are indicated as bars, and the red dashed line is a target where all parameters are taken at their mean values over the data set. To interpret the graph, consider the top feature, pad share (which is the ratio of the pad volume to the overall volume of injected fluid). The dark blue means the difference between the target with the "average parameter value" and the target with the pad share parameter decreased by 50%, while the rest of the parameters are kept at their average values.

The feature importance analysis within the Catboost model is carried out for the entire feature list, while the OVAT analysis is conducted for the design features only. The reason for performing the OVAT on the design features only is due to the problem objective, where our goal is to vary design parameters to maximize the target. Moreover, the design features deviate from their means, which is oftentimes not true for different types of wells. In other words, the limitation of OVAT is that we have to deviate the i-th parameter from its average, while some wells have the

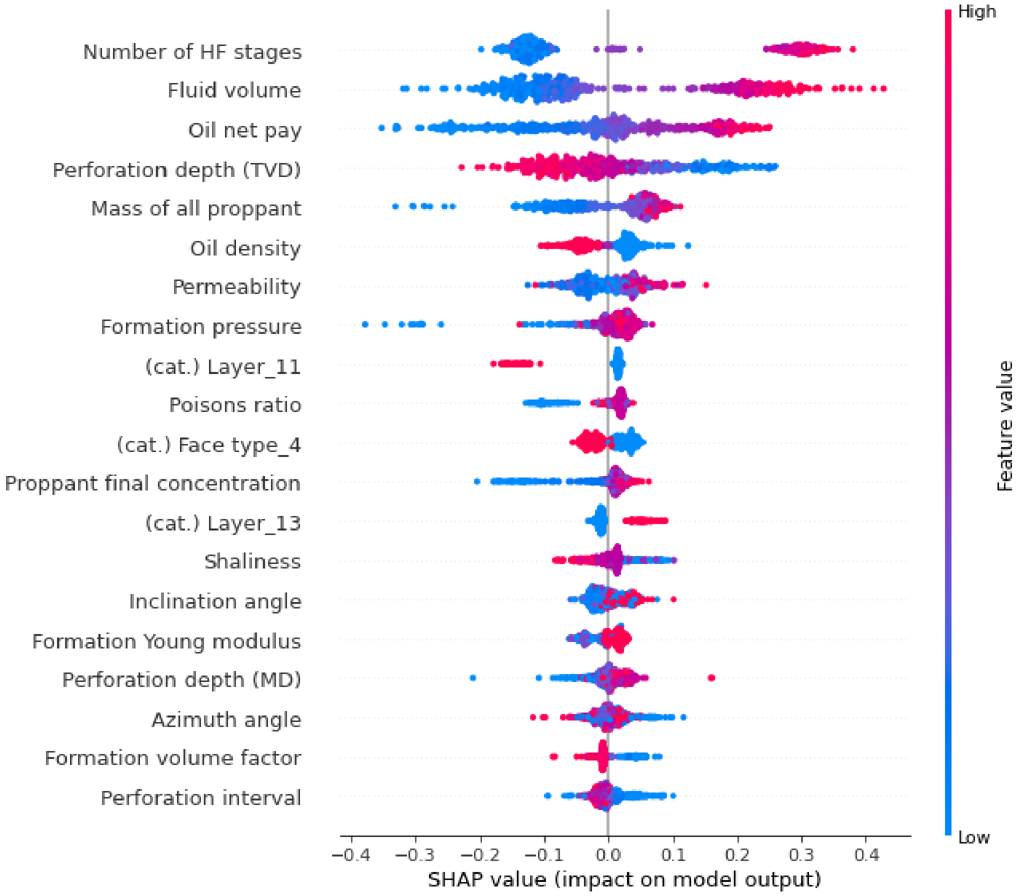


Figure 4-6: Shap feature importance for primary fracturing operations on new wells

value of the i-th feature far from the mean. Hence, the OVAT is more applicable for the design optimization problem and is not consistent with the feature importance analysis. An example of inconsistency is the number of HF stages, where its ranking within the feature importance (Fig 4-6) is the 1st while its ranking on the tornado chart (Fig 4-7) is the 4th among other design parameters. To summarize, OVAT is not representative for verifying feature importance, while it is suitable for getting target value sensitivity to the variation of a single parameter on a particular HF operation (with all other features fixed at their mean values).

4.2.3.1 Parameter selection

The relationship between the model's predictive capability and the number of parameters considered is analyzed using the feature importance analysis discussed earlier (refer to Fig 4-8 for reference). The findings indicate that reducing the

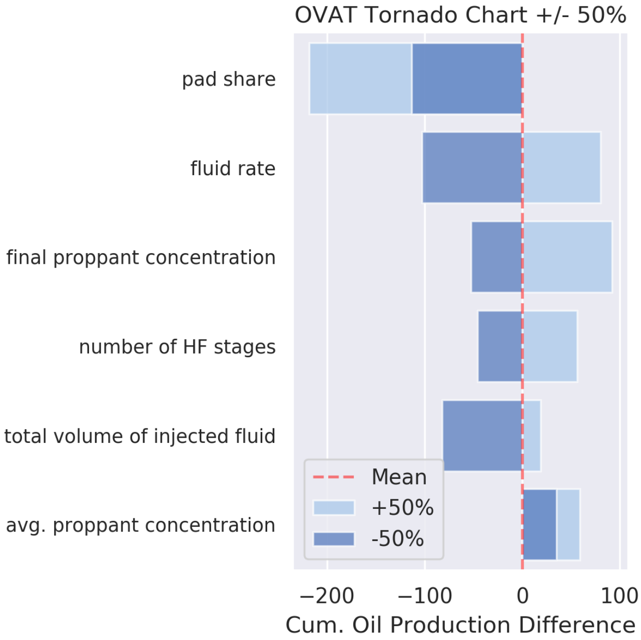


Figure 4-7: Model sensitivity analysis: tornado chart of the OVAT analysis

dimensionality of the problem from 50 to 35 parameters improves the tractability of the design optimization task without compromising the R^2-score. Interestingly, not all design features depicted on the OVAT tornado chart rank among the top 10 important features. Nevertheless, the design parameters will still be included in the input data as they are subject to optimization, which remains the primary objective of the second part of this study.

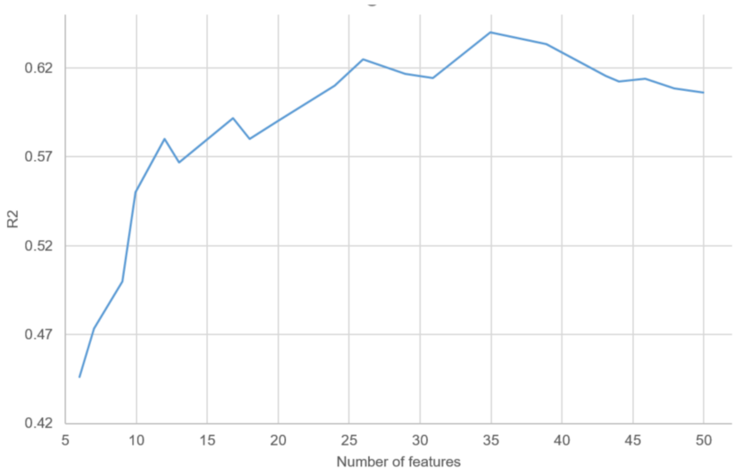


Figure 4-8: Recursive feature elimination: model's score vs. number of input parameters

4.2.4 Field test

The production prediction errors for all 21 tested wells are shown in Fig. 4-9. Here we see a relatively low percentage error for horizontal wells, which can be explained by the higher production rates of such wells. What is important here is the high error for multilateral wells, which is most likely caused by data distortion for this type of wells. Currently, the data point for a multilateral well is represented as a multistage fracture treatment with the number of stages equal to the number of laterals with fractures. The disadvantages of this method are obvious when hydraulic fracturing is performed at different points in time. In addition, different reservoir parameters must be considered for each operational production formation.

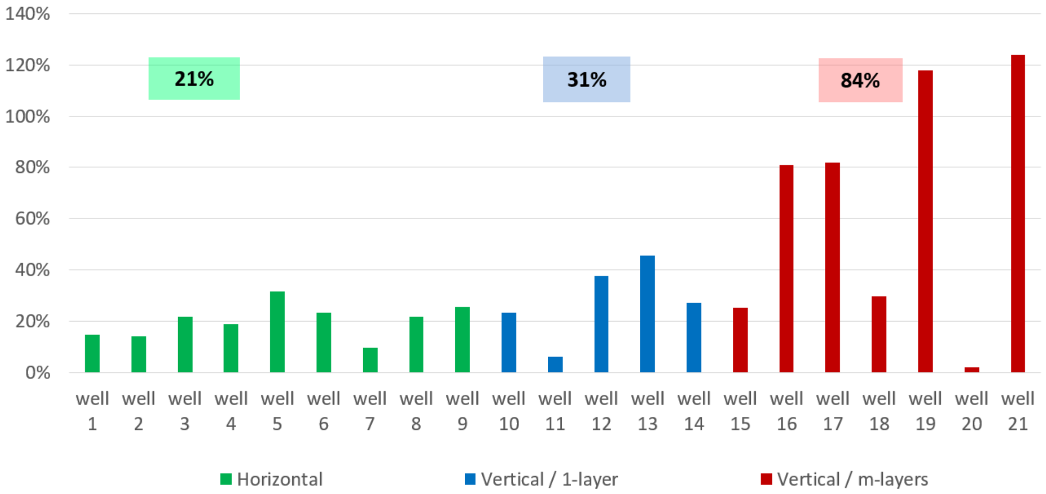


Figure 4-9: Real vs Predicted production on the real design

Generally, the accuracy of the prediction model ($MAPE = 37.28\%$, $wMAPE = 27.46\%$) in field tests is close to the hold-out set accuracy check ($MAPE = 37.78\%$, $wMAPE = 29.95\%$). The distribution of well types (vertical, horizontal) for field tests is close to that presented in the hold-out set.

4.3 Inverse problem

Overall, four approaches to the problem of design optimization were formulated:

1. *SBO*: Surrogate-Based Optimization (Sec. 3.2.2),

2. *SLSQ*: Sequential Least Squares Programming,

3. *PSO*: Particle Swamp Optimization,

4. *DE*: Differential Evolution.

These methods were compared with each other in terms of maximum produced fluid and physically-grounded recommendations for the design parameters for 3 well types: horizontal, vertical and vertical multilateral.

During the testing, we encountered a problem regarding missing parameter values for some wells. Some optimization algorithms, which rely on constructing regression models, are unable to handle missing values. Therefore, imputation techniques are necessary to address this issue and improve the accuracy of the optimization process. Simply filling missing values with the corresponding mean values may not yield optimal results. Instead, several strategies to impute missing values in a more intelligent manner were proposed.

One such strategy is matrix factorization, as discussed in Section 2.2.2. This approach has demonstrated greater success in terms of enhancing the predictive power of the model. Consequently, we employed this method to train the production prediction model, allowing for more accurate and reliable predictions.

Another method involves imputing missing values by averaging the parameter values of the top-N similar wells, as described in Section 3.2.1.3. This method was used during the design optimization stage, as it is particularly advantageous for petroleum engineers, because it provides a physically explainable approach to the task.

4.3.1 Design optimization results

Comparison of the efficiency of optimization algorithms in Figure A-7 shows the average cumulative fluid production across all pilot wells. A larger value means higher efficiency of an algorithm. Figure A-15 shows the extended results of the optimization for each well individually.

The results of the optimization for design parameters are shown in Figures A-8-A-13. Here the optimum values are indicated in percentages within the optimization limits for each parameter. 0% is a lower and 100% is an upper bound for each well, obtained by the offset wells search. Color bars represent the average value of the analyzed parameter for all wells in each of the three groups.

All the methods were operating under similar conditions: boundaries, constraints, maximum allowable number of function evaluations (200). Under these conditions, SBO proved to be the most efficient algorithm in terms of maximizing the objective function (Figures A-7 and A-15). It is also possible to draw conclusions about general optimization trends. First of all, the most of the approaches for all well types maximized the target by increasing the proppant mass and reducing the average proppant concentration. It makes sense to increase the amount of pumping fluid and proppant to get maximum production while we have neither WOC nor GOC.

We can clearly see the recommendation trends difference between horizontal and both types of vertical wells even though vertical multilateral wells are represented in our database as multistage fracturing operations (like horizontal wells). Pad share here is the most interesting parameter, which does not follow any particular trend. The final proppant concentration is lower for horizontal wells, which can be explained by the high probability of ineffective treatments with high concentration values in these types of wells, as such fluids are difficult to pump through well sections with high wellbore curvature. The value of the fluid flow rate varies widely and is probably not important for the purpose of maximizing production. This can be proved by the low feature importance of this parameter, and may also be due to the fact that the values of this parameter are chosen properly in most of HF treatments presented in the database.

Once we have the results of the optimization, we can carry out a kind of retrospective analysis to find out how we can change usual HF designs to improve its results: if the value is below 50% – the optimum value is less than usual one used during HF treatments and vice versa.

We tested the SBO method to find the optimal set of design parameters for each well from the pilot tests. In addition, we limited the proppant mass to a value from the actual fracture design (the actual values of the design parameters have been chosen by the contractor). Then the fluid production was calculated for the optimal parameters and for the actual ones. Optimal parameters showed a higher production rate, *theoretically* increased by 38%. (Fig. A-14).

Chapter 5

Conclusions and Discussions

5.1 Data gathering

The resulting database compiled during this study encompasses over 5,000 oil wells located in Western Siberia. These wells, which were drilled between 2013 and 2019, include a diverse range of vertical, directional, and horizontal configurations and have undergone fracturing and refracturing treatments. This is a remarkably representative dataset, compared to the majority of open literature on the subject. The overall x-vector characterizing a well (data point) contains 92 parameters, including 36 parameters describing formation properties, 12 for the wellbore, and 44 for the fracturing design. The input vector is reduced to 35 parameters for model training after recursive feature importance analysis and elimination.

During the research, various challenges were identified that arise when constructing a digital field database that integrates three major components originating from distinct sources: reservoir geology, HF design, well construction and production data.

The challenges include:

- Data integration: Bringing together data from reservoir geology, HF design, and production requires overcoming compatibility issues, differing formats, and varying levels of data quality. Integrating these diverse datasets into a cohesive and unified database poses a significant technical challenge.

- Data heterogeneity: The data obtained come from fundamentally different sources, each with its own unique characteristics and structures. Dealing with the heterogeneity of these datasets, including differences in data types, formats, and scales, requires careful consideration and robust data transformation techniques.

- Data consistency: Ensuring the consistency and accuracy of the integrated data is crucial. Inconsistencies, errors, and missing information can adversely affect the reliability of analyses, interpretations and predictions based on the database. Implementing data validation and quality control procedures becomes essential to maintain data integrity.

- Data volume and scalability: The amount of data generated from reservoir geology, HF design, and production can be substantial and continue to grow over time. Managing and storing large volumes of data, as well as ensuring scalability of the database infrastructure, is a critical aspect to consider by oil and gas production companies.

Addressing these challenges requires a comprehensive approach that includes robust data integration techniques, effective data management strategies, and the utilization of appropriate technologies for data storage, processing, and analysis. Additionally, collaboration and coordination between different stakeholders, including geologists, engineers, and data scientists, are crucial for successful implementation of a digital field database within the IT perimeter of a company.

5.2 Forward problem

The forward problem of predicting the production rate based on fracturing design parameters is solved using the widely used ML algorithms. Cross-comparison revealed that decision tree based models outperformed the others due to high heterogeneity of the input parameters. As the result of solving the forward problem, the accuracy of predicting cumulative oil production is R^2=0.64 achieved by the stacked Ridge Regression + CatBoost algorithm. The stacked approach was chosen as a solution because the single gradient boosting trees algorithm was found to be insufficient in capturing the real-world behavior of the process.

Regarding relative feature importance within the model, the top ten important parameters are:

- number of stages in a multistage treatment;
- volume of injected fluid;
- proppant mass per meter of perforated interval;
- perforation true vertical depth;
- perforation zenith angle;
- reservoir net pay;
- geological facies;
- reservoir layer;
- perforated interval;
- formation permeability.

From the OVAT analysis, it follows that the following possible patterns in the average multi-stage HF treatment can be identified (with all parameters at its mean values):

- Mean value of the pad share is optimal for an average treatment. Deviations from this value have negative effect on production;
- Increasing the fluid rate increases production and vice versa;
- Mean final proppant concentration possibly was selected below optimum, comparing to the optimum value, which is less than average. In addition, average proppant concentration is probably systematically selected below optimum value, too (by frac design engineers planning the treatment);

- Building conclusions on the results of both OVAT and feature importance analysis, we come to the conjecture that the volume of injected fluid is one of the most important features.

Cleaning the data and handling the missing data on real field data set appeared to be one of the most important tasks due to huge amount of errors and missing records in the original raw data. Missing values imputation via collaborative filtering technique (NNMF) allowed to improve predictability of a model. The higher predictive capability of the model proved to be based on a data base of wells with re-fracturing (where production before treatment is known).

The following important points need to be emphasized:

- ML model completely depends on input data (data completeness, data quality, and preprocessing);
- Collection of field data is the most important step for the ML project aimed at an optimization of a stimulation treatment. A database, which has been properly validated, filled and verified with subject matter experts, allows one to build high-quality predictive models and make well-informed decisions, based on all the advantages of modern ML techniques;
- Data pre-processing and use of complex tuned ML models allow to achieve higher accuracy. However, the results of the study on small train sets versus full data set show that this accuracy does not always indicate the model capability to generalize the obtained results. High accuracy reported in the literature on relatively small data is typically the consequence of overfitting. Presented results were validated on a proper hold-out set, which was not used in training of the model;
- The test accuracy of the model highly depends on the number of samples and on the complexity of the ML algorithm. If either of these two does not fit together, it would lead to overfitting or underfitting. In this study the best available ML practices are selected and tuned.

Thus, an accurately formed, verified and validated field database on stimulation treatments may lead to the results that are not "ideal" (in terms of the predictive power), because of its inherent heterogeneities. The database in the study comprises various well types, including horizontal, vertical, and multilateral configurations, each of which differs significantly in their predictability of production outcomes. The limitations observed in the database primarily stemmed from an underrepresentation of multilateral wells. Prediction error is dominated by poor prediction of this type of the wells. In the data set they are represented as a multistaged treatment, similarly to the horizontal wells.

Faced with this segregation, we experimented with a training pipeline that utilized only a subset of the database (e.g., exclusively vertical or horizontal wells). This approach, however, resulted in reduced prediction capability. A more successful method involved representing each fracturing operation within the same data space and training the model on the entire dataset accordingly.

Speaking about the forward problem in determining the cumulative production, several approaches available in ML nowadays were used, including clustering, model ensembles and tuning, feature importance, and uncertainty quantification.

5.3 Inverse problem

5.3.1 Optimization interval limits

To optimize the hydraulic fracturing (HF) design, a method for searching similar wells has been developed. This method employs a clustering technique and utilizes the Euclidean distance as a metric of similarity. The results obtained from this method assist engineers in reviewing previously conducted treatments, enabling them to estimate interval limits for the design parameters to be optimized.

The lower and upper limits are determined based on the 5^{th} and 95^{th} percentiles, respectively, of the cluster parameter values. To refine the set of wells, it is possible to reduce its size by selecting only the top-N wells based on the Euclidean distance. By doing so, engineers can focus on exploring optimal values of the design parameters within a specific vicinity where the prediction model performs well.

Furthermore, the top-N similar wells can also be utilized for imputing missing parameters in the pilot well. This approach leverages the information from the most similar wells to fill in any gaps or missing values, enhancing the accuracy and completeness of the data for the pilot well.

A number of optimization techniques were tested during the pilot testing: *differential evolution, sequential least squares programming, particle swarm optimization and surrogate-based optimization.*

5.3.2 Testing optimization algorithms

The *SBO* approach on average maximized the 3-month fluid production better than other methods. General optimization trend of 21 pilot wells is to increase the amount of fracturing fluid and proppant mass. Trends for final, average proppant concentration and fluid rate are different for each type of well (horizontal, vertical and vertical multilateral). The calculated production from the fracturing with the optimal set of design parameters, compared to the treatments with an actual HF design, gives a theoretical target improvement of 38%.

5.3.3 Optimal target selection

The primary objective of hydraulic fracturing design optimization is to maximize hydrocarbon recovery by enhancing reservoir connectivity and productivity. However, this optimization neglects an important component — economic considerations. Without integrating economics, the optimization process tends to favor larger and more numerous fracturing operations, assuming that more significant efforts lead to higher oil recovery. This approach disregards the economic constraints that inherently affect hydraulic fracturing operations.

In the approach presented in this study, the focus initially centered on testing the model's capability to maximize cumulative fluid production. This choice served as a rigorous test case to assess the model's physicality, ensuring it could capture all relevant dependencies in the hydraulic fracturing process. The decision to omit economics at this stage stemmed from the recognition that, while maximizing fluid production is an essential technical goal, the primary objective in real-world hydraulic fracturing operations is the economics.

Economics therefore introduces a dynamic interplay between technical and financial targets. When optimizing hydraulic fracturing designs, economic factors often lead to trade-offs. For instance, while increasing the number of fracturing stages may boost production, it also escalates operational costs. In the context of this study, the introduction of economic criteria, for instance, NPV and ROI, into the optimization process is expected to result in a reduction in the number of fracturing stages, proppant mass, and fluid volume being targeted for optimization. This balance between technical and economic factors requires careful consideration.

5.3.4 Retrospective analysis

After testing optimization algorithms, a retrospective analysis was conducted to determine how one can modify typical hydraulic fracturing (HF) design to enhance its outcome. By comparing the optimized value with the usual value used in HF treatments, we can assess whether the optimum value is below or above 50

If the optimized value is below 50%, it suggests that the optimal HF design requires a lower value compared to the conventional practice. This finding implies that by implementing the optimized design, we can potentially improve the results of HF treatments.

Conversely, if the optimized value is above 50%, it indicates that the optimal HF design requires a higher value than the usual approach. In this case, modifying the HF design according to the optimized parameters may lead to improved outcomes.

Indeed, conducting a retrospective analysis in hydraulic fracturing design optimization is crucial as it allows us to evaluate the performance of regular or standard designs when compared to the optimized ones. By comparing the outcomes of the optimization process with the results achieved using conventional designs, we can assess whether the regular designs were well-optimized or not.

By assessing the performance of regular designs in light of the optimization results, we gain valuable insights into the strengths and weaknesses of different design approaches. This information can guide future decision-making processes, allowing for continuous improvement in HF design practices and optimization techniques.

5.4 Future research

In future work, several extensions to the presented workflow could be considered to enhance its capabilities:

- Expand the list of features: Including the distance to upper and lower water-bearing layers as additional features would enable separate predictions for the production of pure oil and water. This extension would involve extending the target variables to maximum total fluid and maximum pure oil production (or minimum water cut);

- Implement an economics model: Introducing an economics model would enable the workflow to operate under metrics such as production quantity (Q) divided by capital expenditure (CAPEX). By considering the economic aspect, decision-making regarding the HF design can be guided by a more comprehensive understanding of the costs and benefits associated with different HF design choices;

- Account for the influence of injection wells: Considering the impact of injection wells on production rates is crucial. Different injection rates from the neighbouring wells can significantly affect the overall performance of a well after the hydraulic fracturing;

- Extend the target to multi-criteria optimization: To address a broader range of objectives, the target could be extended to include multiple criteria. In addition to maximizing production, minimizing the total proppant load can be included as an additional objective.

Furthermore, combining synthetic data from fracture design simulations, such as those conducted using commercial simulators, can be highly beneficial. By integrating primary fracture design and reservoir properties data with simulated fracture propagation data, essential parameters such as fracture length, width, and height, which directly impact production, can be obtained. This integration of synthetic data could enhance the predictive power of the models and improve their ability to capture real-world behavior.

Bibliography

A Abdulraheem, M Ahmed, A Vantala, T Parvez, et al. Prediction of rock mechanical parameters for hydrocarbon reservoirs using different artificial intelligence techniques. 2009.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

C.C. Aggarwal and S. Sathe. *Outlier ensembles: an introduction.* Springer, 2017.

Basil Al-Shamma, Helene Nicole, Peyman R Nurafza, Wei Cher Feng, et al. Evaluation of multi-fractured horizontal well performance: Babbage field case study. In *SPE Hydraulic Fracturing Technology Conference.* Society of Petroleum Engineers, 2014.

Rustam Alimkhanov, Irina Samoylova, et al. Application of data mining tools for analysis and prediction of hydraulic fracturing efficiency for the bv8 reservoir of the povkh oil field. In *SPE Russian Oil and Gas Exploration & Production Technical Conference and Exhibition.* Society of Petroleum Engineers, 2014.

Roger N Anderson*, Boyi Xie, Leon Wu, Arthur A Kressner, Joseph H Frantz Jr, Matthew A Ockree, Kenneth G Brown, Peter Carragher, and Mark A McLane. Using machine learning to identify the highest wet gas producing mix of hydraulic fracture classes and technology improvements in the marcellus shale. In *Unconventional Resources Technology Conference, San Antonio, Texas, 1-3 August 2016*, pages 254–266. Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers, 2016.

Obadare Awoleke, Robert Lane, et al. Analysis of data from the barnett shale using conventional statistical and virtual intelligence techniques. *SPE Reservoir Evaluation & Engineering*, 14(05):544–556, 2011.

R Mark Balen, HZ Mens, Michael J Economides, et al. Applications of the net present value (npv) in the optimization of hydraulic fractures. In *SPE Eastern Regional Meeting.* Society of Petroleum Engineers, 1988a.

R Mark Balen, HZ Mens, Michael J Economides, et al. Applications of the net present value (npv) in the optimization of hydraulic fractures. In *SPE Eastern Regional Meeting.* Society of Petroleum Engineers, 1988b.

LA Baumes, JM Serra, P Serna, and A Corma. Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications. *Journal of combinatorial chemistry*, 8(4):583–596, 2006.

Richard E Bellman and Stuart E Dreyfus. *Applied dynamic programming*, volume 2050. Princeton university press, 2015.

M. Belyaev, E. Burnaev, E. Kapushev, M. Panov, P. Prikhodko, D. Vetrov, and D. Yarotsky. Gtapprox: Surrogate modeling for industrial design. *Advances in Engineering Software*, 102:29–39, 2016.

M. G. Belyaev and E. V. Burnaev. Approximation of a multidimensional dependency based on a linear expansion in a dictionary of parametric functions. *Informatics and its Applications*, 7(3):114–125, 2013.

Jack Betz et al. Low oil prices increase value of big data in fracturing. *Journal of Petroleum Technology*, 67(04):60–61, 2015.

Mohammad Reza Bonyadi and Zbigniew Michalewicz. Particle swarm optimization for single objective continuous space problems: A review. *Evolutionary Computation*, 25(1):1–54, 2017. doi:10.1162/EVCO_r_00180.

E. Burnaev and P. Erofeev. The influence of parameter initialization on the training time and accuracy of a nonlinear regression model. *Journal of Communications Technology and Electronics*, 61(6):646–660, Jun 2016.

E. Burnaev and M. Panov. Adaptive design of experiments based on gaussian processes. In Alexander Gammerman, Vladimir Vovk, and Harris Papadopoulos, editors, *Statistical Learning and Data Sciences*, pages 116–125, Cham, 2015. Springer International Publishing. ISBN 978-3-319-17091-6.

E. Burnaev and A. Zaytsev. Surrogate modeling of multifidelity data for large samples. *Journal of Communications Technology and Electronics*, 60(12):1348–1355, Dec 2015. ISSN 1555-6557.

E. Burnaev, M. Panov, and A. Zaytsev. Regression on the basis of nonstationary gaussian processes with bayesian regularization. *Journal of Communications Technology and Electronics*, 61(6):661–671, Jun 2016. ISSN 1555-6557.

E. Burnaev, I. Panin, and B. Sudret. Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions. *Annals of Mathematics and Artificial Intelligence*, 81(1):187–207, Oct 2017. ISSN 1573-7470. doi:10.1007/s10472-017-9542-1.

E. V. Burnaev. Algorithmic foundations of predictive analytics in industrial engineering design. *Journal of communications technology and electronics*, 64(12): 1485–1492, 2019.

E. V. Burnaev and P. V. Prikhod'ko. On a method for constructing ensembles of regression models. *Automation and Remote Control*, 74(10):1630–1644, Oct 2013.

Artem Chashchin et al. Metamodelling of transient hydraulic fracture growth and proppant transport with machine learning on synthetic planar3d data (to be submitted). *Journal of Petroleum Science and Engineering*, 2020.

C. Daniel. One-at-a-time plans. *Journal of the American Statistical Association*, 68: 353–360, 1973.

Emmanuel Detournay. Mechanics of hydraulic fractures. *Annual Review of Fluid Mechanics*, 48:311–339, 2016.

V. Duplyakov, A. Morozov, D. Popkov, A. Vainshtein, A. Osiptsov, E. Burnaev, E. Shel, G. Paderin, P. Kabanova, I. Fayzullin, R. Uchuev, A. Mukhametov, A. Prutsakov, I. Vikhman, and M. Staritsyn. Practical aspects of hydraulic fracturing design optimization using machine learning on field data: Digital database, algorithms and planning the field tests. 09 2020. doi:10.2118/203890-MS.

V. Duplyakov, A. Morozov, D. Popkov, E. Shel, A. Vainshtein, E. Burnaev, A. Osiptsov, and G. Paderin. Data-driven model for hydraulic fracturing design optimization. part ii: Inverse problem. *Journal of Petroleum Science and Engineering*, 208:109303, 2022. ISSN 0920-4105. doi:https://doi.org/10.1016/j.petrol.2021.109303.

V. Duplyakov, V. Vanovskii, D. Popkov, A. Morozov, A. Vainstein, A. Osiptsov, S. Kaygorodov, V. Kotezhekov, B. Belozerov, and E. Burnaev. Building a permeability map of an oil reservoir by combining data from logging, well testing and seismic surveys. Intelligent data analysis in oil and gas industry. Novosibirsk, Russia, 2022.

V. Duplyakov, A. Morozov, D. Popkov, A. Vainshtein, A. Osiptsov, E. Burnaev, E. Shel, G. Paderin, P. Kabanova, I. Fayzullin, R. Uchuev, A. Mukhametov, A. Prutsakov, I. Vikhman, and M. Staritsyn. Practical aspects of hydraulic fracturing design optimization using machine learning on field data: Digital database, algorithms and planning the field tests. page 24, SPE Symposium: Hydraulic Fracturing in Russia. Experience and Prospects. Moscow, Russia, 2020.

E. T. Woldemariam, H. G. Lemu. A machine learning based framework for model approximation followed by design optimization for expensive numerical simulation-based optimization problems. In *Proceedings of the Twenty-ninth International Ocean and Polar Engineering Conference www.isope.org Honolulu, Hawaii, USA, June 16-21.* International Society of Offshore and Polar Engineers (ISOPE), 2019.

Michael J Economides, Kenneth G Nolte, et al. *Reservoir stimulation*, volume 2. Prentice Hall Englewood Cliffs, NJ, 1989.

Emad Ahmed El-Sebakhy, Tarek Sheltami, Said Y Al-Bokhitan, Yasser Shaaban, Putu D Raharja, Yaman Khaeruzzaman, et al. Support vector machines framework for predicting the pvt properties of crude oil systems. 2007.

Andrei Erofeev, D. Orlov, D. Perets, and D. Koroteev. AI-Based Estimation of Hydraulic Fracturing Effect. *SPE Journal* , 04 2021. doi:10.2118/205479-PA.

Soodabeh Esmaili and Shahab D Mohaghegh. Full field reservoir modeling of shale assets using advanced data-driven analytics. *Geoscience Frontiers*, 7(1):11–20, 2016.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *KDD*, pages 226–231. AAAI Press, 1996.

Alexander Forrester, Andreas Sobester, and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley and Sons, 2008.

Zhengqing Fu, Goulin Liu, and Lanlan Guo. Sequential Quadratic Programming Method for Nonlinear Least Squares Estimation and Its Application. *Mathematical Problems in Engineering*, 2019:1–8, 06 2019. doi:10.1155/2019/3087949.

Jiyao Gao and Fengqi You. Design and optimization of shale gas energy systems: Overview, research challenges, and future directions. *Computers & Chemical Engineering*, 106:699–718, 2017.

Abhishek Gaurav et al. Horizontal shale well eur determination integrating geology, machine learning, pattern recognition and multivariate statistics focused on the permian basin. 2017.

S. Grihon, E. Burnaev, M. Belyaev, and P. Prikhodko. *Surrogate Modeling of Stability Constraints for Optimization of Composite Structures*, pages 359–391. Springer New York, New York, NY, 2013.

Jianchun Guo, Yong Xiao, and Haiyan Zhu. A new method for fracturing wells reservoir evaluation in fractured gas reservoir. *Mathematical Problems in Engineering*, 2014, 2014.

Geir Hareland, Paul Rampersad, Jirapong Dharaphop, Sunthan Sasnanand, et al. Hydraulic fracturing design optimization. 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

Youwei He, Shiqing Cheng, Jiazheng Qin, Jianwen Chen, Yang Wang, Naichao Feng, Haiyang Yu, et al. Successful application of well testing and electrical resistance tomography to determine production contribution of individual fracture and water-breakthrough locations of multifractured horizontal well in changqing oil field, china. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 2017.

Tarek Helmy, Anifowose Fatai, and Kanaan Faisal. Hybrid computational models for the characterization of oil and gas reservoirs. *Expert Systems with Applications*, 37(7):5353–5363, 2010.

Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, 4, 2016.

Mine Kaya and Shima Hajimirza. Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies. *Nature*, 2019.

Reza Keshavarzi, Reza Jahanbakhshi, et al. Real-time prediction of complex hydraulic fracture behaviour in unconventional naturally fractured reservoirs. 2013.

A. Kuleshov, A. Bernstein, and E. Burnaev. Kernel regression on manifold valued data. In *Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics*, pages 120–129, 2018.

Miron Kursa and Witold Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software, Articles*, 36(11):1–13, 2010. ISSN 1548-7660. URL `https://www.jstatsoft.org/v036/i11`.

F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

E Lolon, K Hamidieh, L Weijers, M Mayerhofer, H Melcher, O Oduba, et al. Evaluating the relationship between well parameters and production using multivariate statistical models: a middle bakken and three forks case history. In *SPE Hydraulic Fracturing Technology Conference*. Society of Petroleum Engineers, 2016.

Lundberg, Scott M., and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.

Yunqian Ma and Yun Fu. *Manifold Learning Theory and Applications*. CRC Press, Inc., 2011.

Ivan Makhotin, Dmitry Koroteev, and Evgeny Burnaev. Gradient boosting to boost the efficiency of hydraulic fracturing. *Journal of Petroleum Exploration and Production Technology*, pages 1–7, 2019.

Camron K Miller, George A Waters, Erik I Rylander, et al. Evaluation of production log data from horizontal wells drilled in organic shales. In *North American Unconventional Gas Conference and Exhibition*. Society of Petroleum Engineers, 2011.

SD Mohaghegh, R Gaskari, M Maysami, et al. 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, 2017.

Shahab D Mohaghegh, Andrei Popa, Razi Gaskari, Sam Ameri, SL Wolhart, et al. Identification of successful practices in hydraulic fracturing using intelligent data mining tools; application to the codell formation in the dj-basin. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 2002.

Shahab D Mohaghegh, Ognjen Srecko Grujic, Saeed Zargari, Amirmasoud Kalantari Dahaghi, et al. Modeling, history matching, forecasting and analysis of shale reservoirs performance using artificial intelligence. 2011.

Ameri S Mohaghegh S, Balan B. A hybrid, neuro-genetic approach to hydraulic fracture treatment design and optimization. 1996.

A. Morozov, D. Popkov, V. Duplyakov, E. Mutalova, A. Osiptsov, A. Vainshtein, E. Burnaev, E. Shel, and G. Paderin. Data-driven model for hydraulic fracturing design optimization: focus on building digital database and production forecast. *Journal of Petroleum Science and Engineering*, 194:107504, 2020a. ISSN 0920-4105. doi:https://doi.org/10.1016/j.petrol.2020.107504.

A. Morozov, D. Popkov, V. Duplyakov, A. Osiptsov, A. Vainshtein, E. Burnaev, E. Shel, and G. Paderin. Machine learning on field data for hydraulic fracturing design optimization: Digital database and production forecast model. pages 1–5, 01 2020b. doi:10.3997/2214-4609.202032068.

Andrei A Osiptsov. Fluid mechanics of hydraulic fracturing: a review. *Journal of Petroleum Science and Engineering*, 156:513–535, 2017.

G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. A way to select the optimal fracturing design based on intelligent analysis of field data to increase hydrocarbon production, №2775034, Russian Federation, 2022.

Piyush Pankaj, Steve Geetan, Richard MacDonald, Priyavrat Shukla, Abhishek Sharma, Samir Menasria, Han Xue, Tobias Judd, et al. Application of data science and machine learning for well completion optimization. 2018.

DK Poulsen, MY Soliman, et al. A procedure for optimal hydraulic fracturing treatment design. 1986.

Gil Press. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes, March*, 23:15, 2016.

(In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "well analog selection in terms of hydraulic fracturing", Russian Federation, 2023a.

(In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "production prediction after hydraulic fracturing based on geology and hydraulic fracture parameters", Russian Federation, 2023b.

(In progress) G. Paderin, E. Shel, A. Osiptsov, E. Burnaev, A. Vainstein, V. Duplyakov, A. Morozov, and D. Popkov. Computer module "fracturing design optimization based on well production prediction", Russian Federation, 2023c.

(In progress) V. Duplyakov, A. Morozov, D. Popkov, K. Pavlenko, A. Vainshtein, V. Kotezhekov, S. Kaygorodov, B. Belozerov, V. Vanovskiy, A. Osiptsov, and E. Burnaev. Data fusion of well logs, build-up test interpretations and seismic data for reservoir permeability field estimation. *Computers and Geotechnics*, 2023.

Nestor V Queipo, Alexander J Verde, José Canelón, and Salvador Pintos. Efficient global optimization for hydraulic fracturing treatment design. *Journal of Petroleum Science and Engineering*, 35(3-4):151–166, 2002.

MM Rahman, MK Rahman, and SS Rahman. An integrated model for multiobjective design optimization of hydraulic fracturing. *Journal of Petroleum Science and Engineering*, 31(1):41–62, 2001.

Pedro M Saldungaray, Terry T Palisch, et al. Hydraulic fracture optimization in unconventional reservoirs. In *SPE Middle East unconventional gas conference and exhibition*. Society of Petroleum Engineers, 2012.

Jared Schuetter, Srikanta Mishra*, Ming Zhong, and Randy LaFollette. Data analytics for production optimization in unconventional reservoirs. In *Unconventional Resources Technology Conference, San Antonio, Texas, 20-22 July 2015*, pages 249–269. Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers, 2015.

Jared Schuetter, Srikanta Mishra, Ming Zhong, Randy LaFollette, et al. A data-analytics tutorial: Building predictive models for oil production in an unconventional shale reservoir. *SPE Journal*, 2018.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Jia Shi, Jinchun Song, Bin Song, and Wen F. Lu. Multi-objective optimization design through machine learning for drop-on-demand bioprinting. *Engineering*, 5(3):586 – 593, 2019. ISSN 2095-8099. doi:https://doi.org/10.1016/j.eng.2018.12.009.

D. Smolyakov, N. Sviridenko, V. Ishimtsev, E. Burikov, and E. Burnaev. Learning ensembles of anomaly detectors on synthetic data. *LNCS: ISNN 2019 — Advances in Neural Networks*, pages 292–306, 2019.

I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *MATH COMPUT SIMULAT*, 5(1-3):271–280, 2001a.

I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* , 2001b. doi:10.1016/S0378-4754(00)00270-6.

Eunhye Song, Barry L Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

Rainer Storn and Kenneth Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359, 01 1997. doi:10.1023/A:1008202821328.

C Temizel, S Purwar, A Abdullayev, K Urrutia, Aditya Tiwari, et al. Efficient use of data analytics in optimization of hydraulic fracturing in unconventional reservoirs. In *Abu Dhabi International Petroleum Exhibition and Conference*. Society of Petroleum Engineers, 2015.

L.J.P. van der Maaten and G.E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9, 2008.

S. Wang and S. Chen. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering*, 2019a.

Shuhua Wang and Shengnan Chen. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering*, 174:682–695, 2019b.

Shuhua Wang, Shengnan Chen, et al. A comprehensive evaluation of well completion and production performance in bakken shale using data-driven approaches. 2016.

Zhou Xiaofeng, AB Zolotukhin, An Guangliang, et al. A post-fracturing evaluation method of fracture properties in multi-stage fractured wells (russian). 2016.

Wang Yanfang and Saeed Salehi. Refracture candidate selection using hybrid simulation with neural network and data analysis techniques. *Journal of Petroleum Science and Engineering*, 123:138–146, 2014.

A. Zaytsev and E. Burnaev. Large scale variable fidelity surrogate modeling. *Annals of Mathematics and Artificial Intelligence*, 81(1):167–186, Oct 2017. ISSN 1573-7470.

Mansoor Zoveidavianpoor, Ariffin Samsuri, Seyed Reza Shadizadeh, et al. Development of a fuzzy system model for candidate-well selection for hydraulic fracturing in a carbonate reservoir. In *SPE Oil and Gas India Conference and Exhibition*. Society of Petroleum Engineers, 2012.

Appendix

Formation parameters		
Layer	Porosity average per perforation	Median clay content per perforation
Net pay	Porosity average per layer	Median clay content per layer
Facies type	Porosity median per perforation	Oil saturation average per perforation
Formation thickness	Porosity median per layer	Oil saturation average per layer
Formation pressure	Permeability average per perforation	Oil saturation median per perforation
Bubble point pressure	Permeability average per layer	Oil saturation median per layer
Oil formation volume factor	Permeability median per perforation	NTG per perforation
Permeability from well flow test	Permeability median per layer	NTG per layer
Oil viscosity	kh median per perforation	Stratification factor per perforation
Water viscosity	kh median per layer	Stratification factor per layer
Oil density	Average clay content per perforation	Formation temperature

Formation pressure	Average clay content per layer	Well intersection data
Well structure		
Perforation depth (MD)	Inclination angle	Tubing diameter
Perforation depth (TVD)	Well's drift direction	Perforation density
Perforation interval	Skin before/after HF	Perforation type
Drainage radius	Dimensionless productivity index (Jd)	Inclination angle from well-logs
HF design parameters		
Number of HF stages	Fracture permeability	Proppant per gross height
Multifrac stage	Closure gradient	Shut-in pressure
Polymer type	ISIP for displacement	Breaker #1 amount
Polymer concentration	ISIP on DFIT	Breaker #2 amount
Crosslinker type	ISIP on main work	Breaker #3 amount
Crosslinker concentration	Delta ISIP	Pad volume
Polymer concentration for pad	Effective pressure on DFIT	Fracture length
Fluid type for main work	Effective pressure on main work	Fracture height
Breaker type #1	Fluid efficiency	Fracture width
Breaker type #2	Proppant concentration	Mass of proppant type #1
Breaker type #3	Pressure loss on friction	Mass of proppant type #2
Average pressure on main work	Pressure loss in BH area	Mass of proppant type #3
Dimensionless fracture conductivity	Proppant per oil-saturated height	Mass of proppant type #4
Fracture conductivity	Proppant per effective height	Mass of proppant type #5
Mass of all proppant	Fluid volume	
Production data		

Bottom-hole pressure	Watercut before HF	**Cumulative oil production** (target)
Productivity index	Suspended solids concentration	Cumulative fluid production
Dimensionless productivity index (Jd)	Fluid rate after HF	Cumulative gas production
Fluid rate	Oil rate after HF	Watercut average during production
Gas rate before	Watercut after HF	Operational hours during production
Fluid rate before HF		

Table A.1: Features used to describe a well

**features averages and medians per layer and perforation sourced from well log interpretation data*

***these are all parameters in the data base before feature selection*

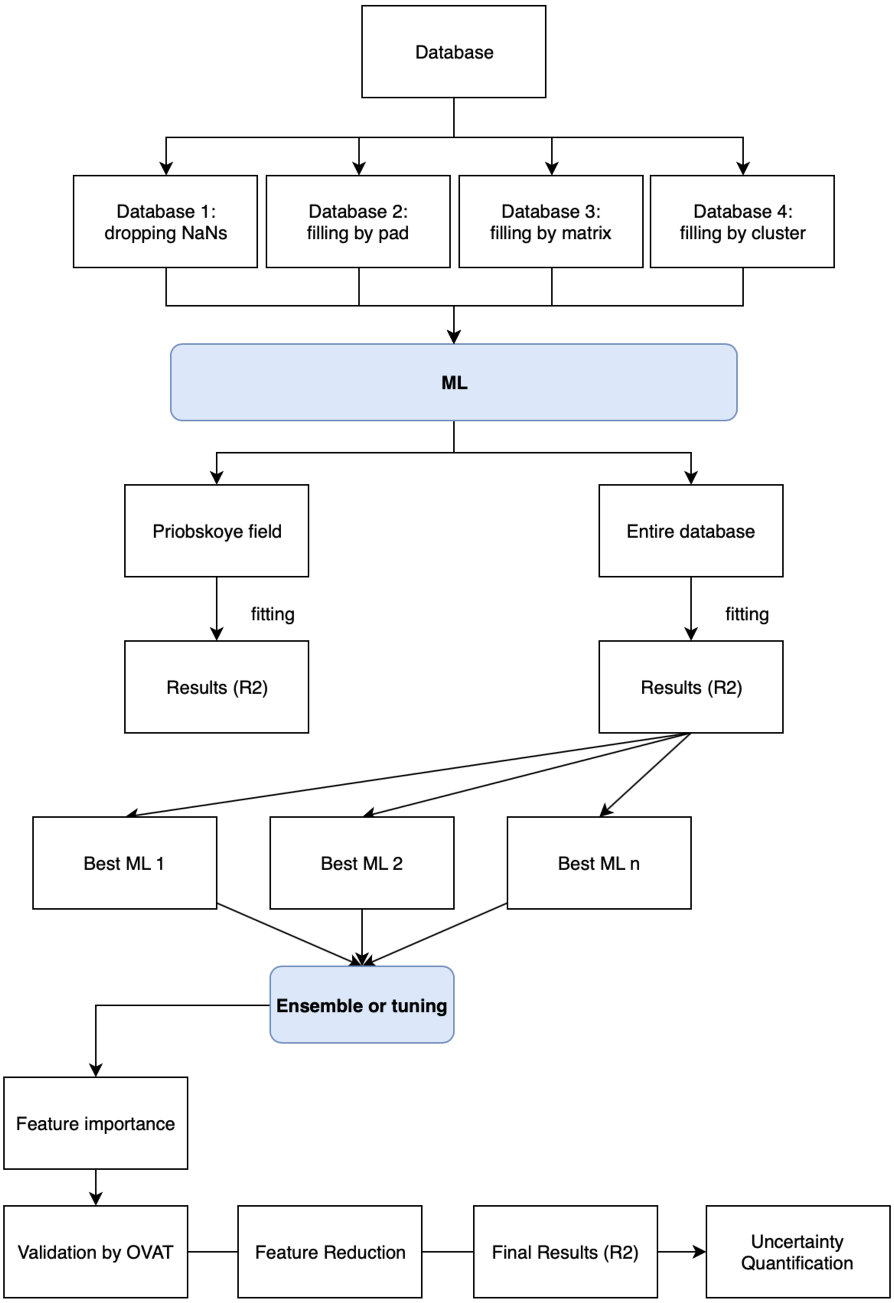


Figure A-1: Forward model algorithm

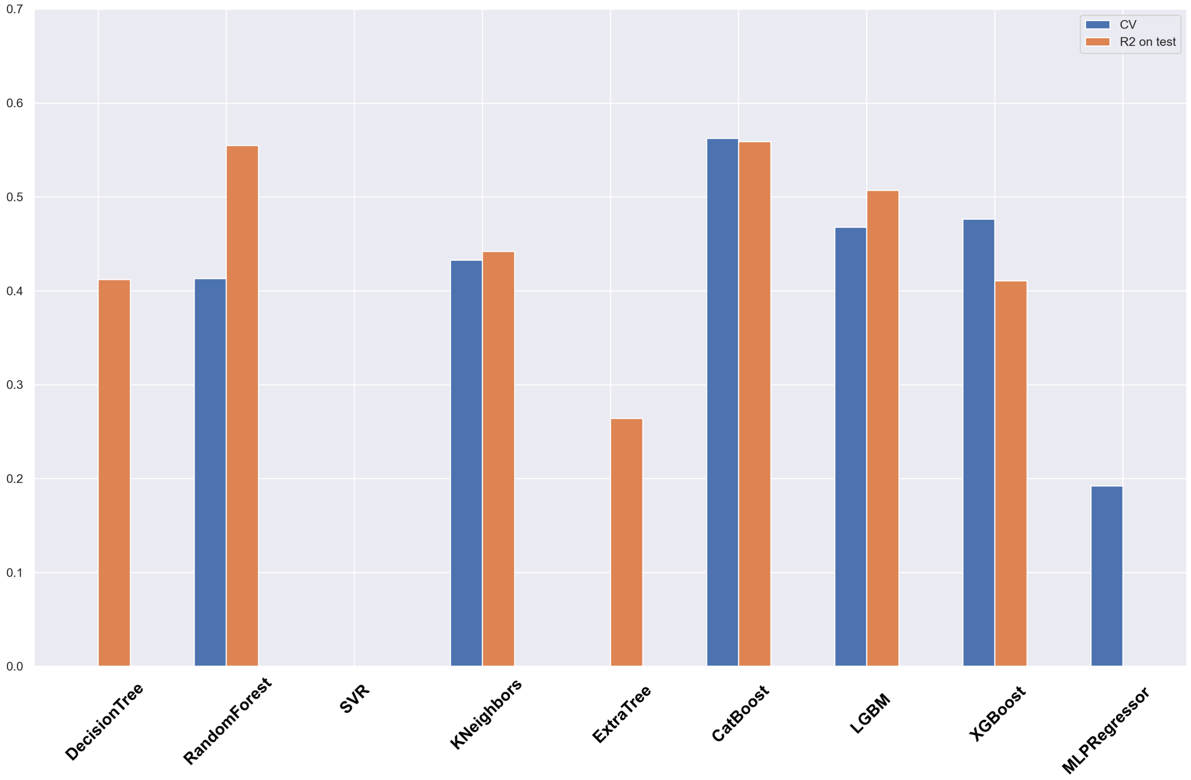


Figure A-2: Algorithms' performance on a test set (untuned): dropping NaNs method

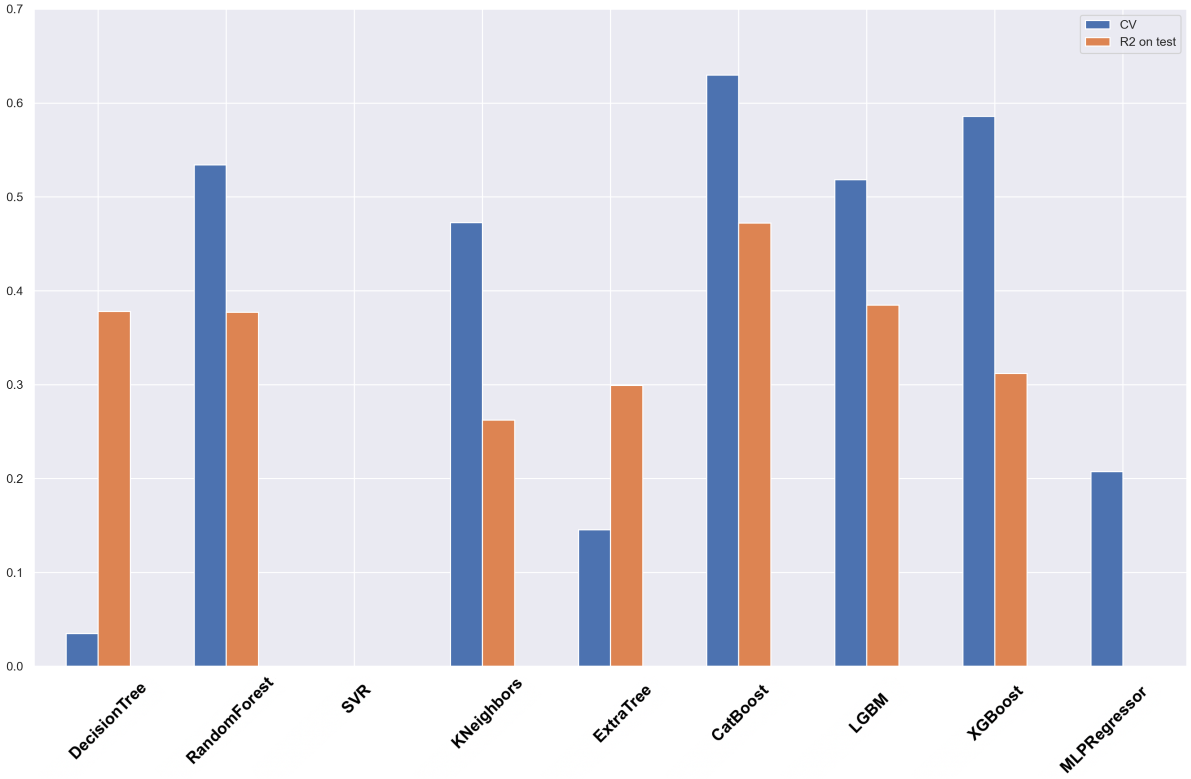


Figure A-3: Algorithms' performance on a test set (untuned): filling by well pad

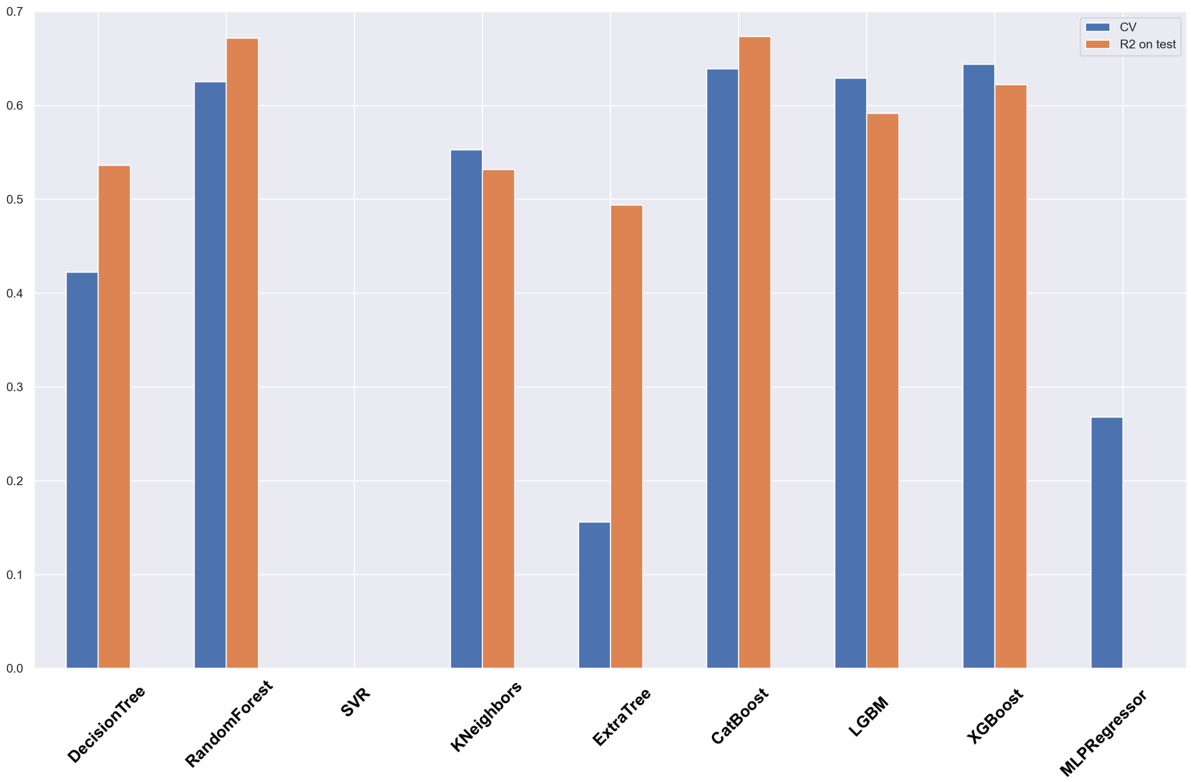


Figure A-4: Algorithms' performance on a test set (untuned): matrix imputation (collaborative filtering)

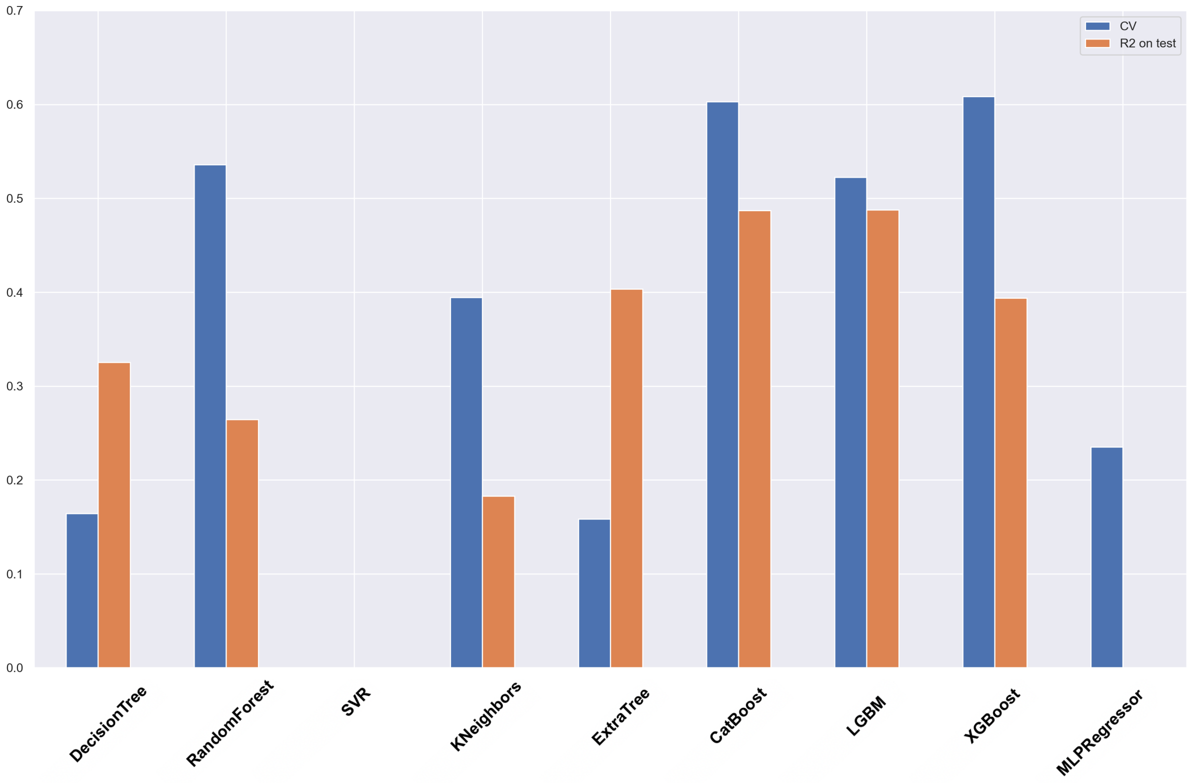
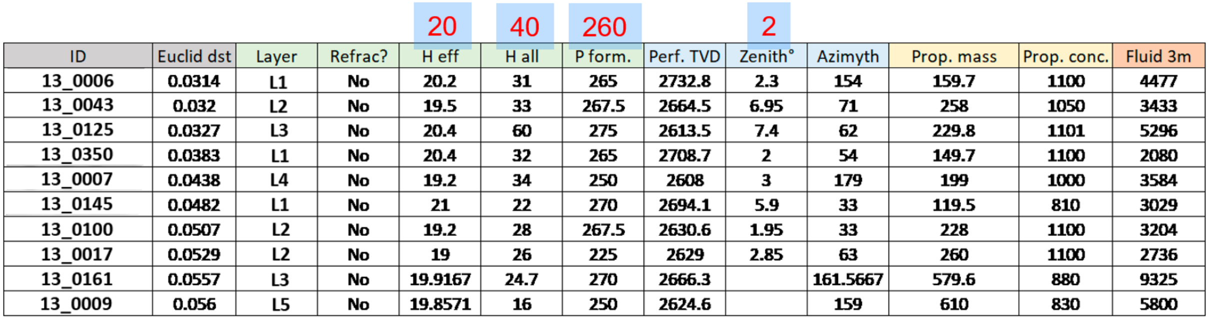


Figure A-5: Algorithms' performance on a test set (untuned): filling by mean values, calculated in each corresponding cluster



				20	40	260		2				
ID	Euclid dst	Layer	Refrac?	H eff	H all	P form.	Perf. TVD	Zenith°	Azimyth	Prop. mass	Prop. conc.	Fluid 3m
13_0006	0.0314	L1	No	20.2	31	265	2732.8	2.3	154	159.7	1100	4477
13_0043	0.032	L2	No	19.5	33	267.5	2664.5	6.95	71	258	1050	3433
13_0125	0.0327	L3	No	20.4	60	275	2613.5	7.4	62	229.8	1101	5296
13_0350	0.0383	L1	No	20.4	32	265	2708.7	2	54	149.7	1100	2080
13_0007	0.0438	L4	No	19.2	34	250	2608	3	179	199	1000	3584
13_0145	0.0482	L1	No	21	22	270	2694.1	5.9	33	119.5	810	3029
13_0100	0.0507	L2	No	19.2	28	267.5	2630.6	1.95	33	228	1100	3204
13_0017	0.0529	L2	No	19	26	225	2629	2.85	63	260	1100	2736
13_0161	0.0557	L3	No	19.9167	24.7	270	2666.3		161.5667	579.6	880	9325
13_0009	0.056	L5	No	19.8571	16	250	2624.6		159	610	830	5800

Figure A-6: Example of offset wells selection

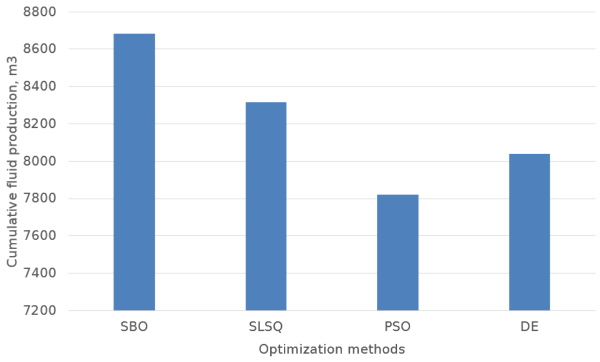


Figure A-7: Average optimized production for all wells by different optimization algorithms

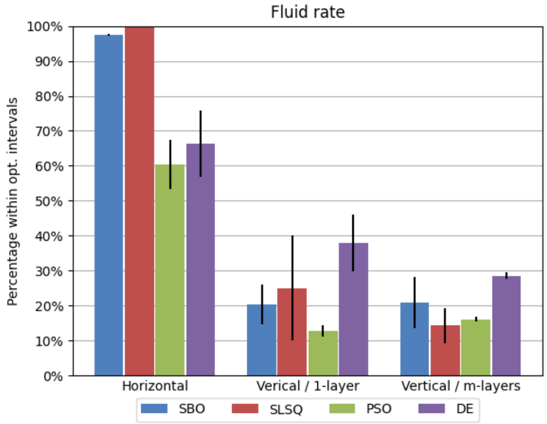


Figure A-8: Average recommended fluid rate for the wells

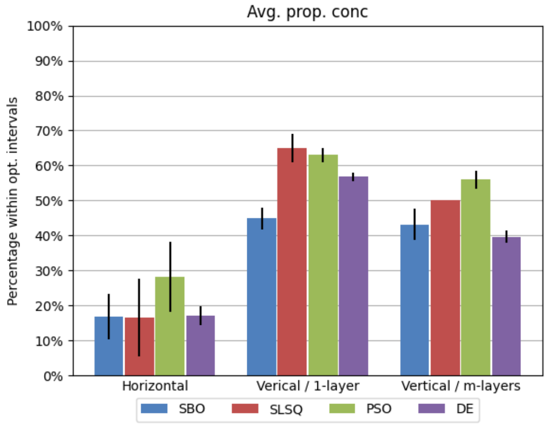


Figure A-9: Average recommended mean proppant concentration for the wells

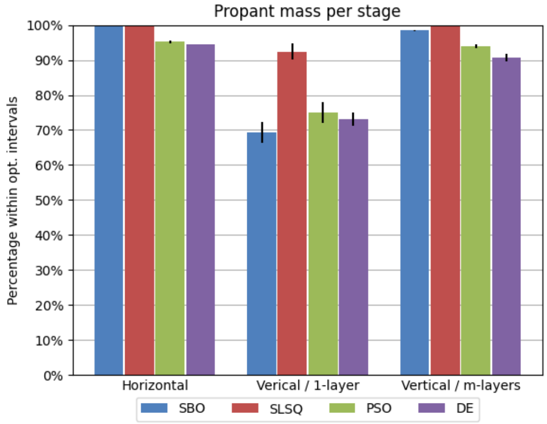


Figure A-10: Average recommended proppant masses (per stage) for the wells

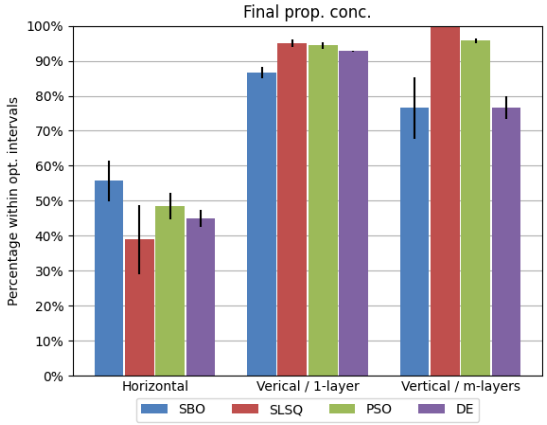


Figure A-11: Average recommended final proppant concentration for the wells

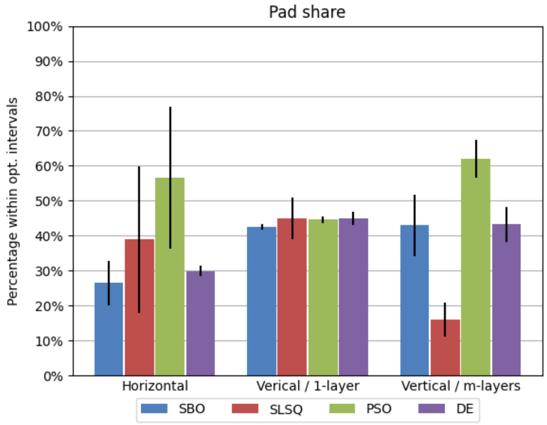


Figure A-12: Average recommended pad share for the wells

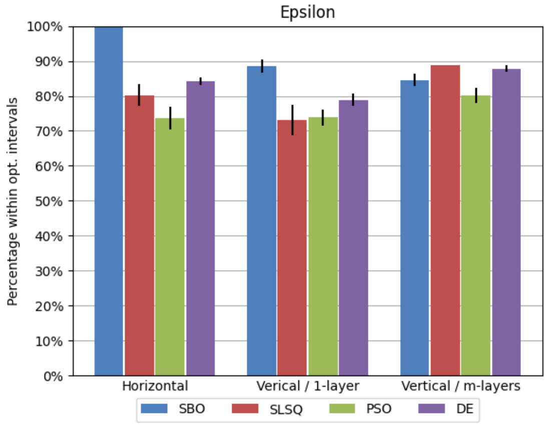


Figure A-13: Average recommended calculated epsilon

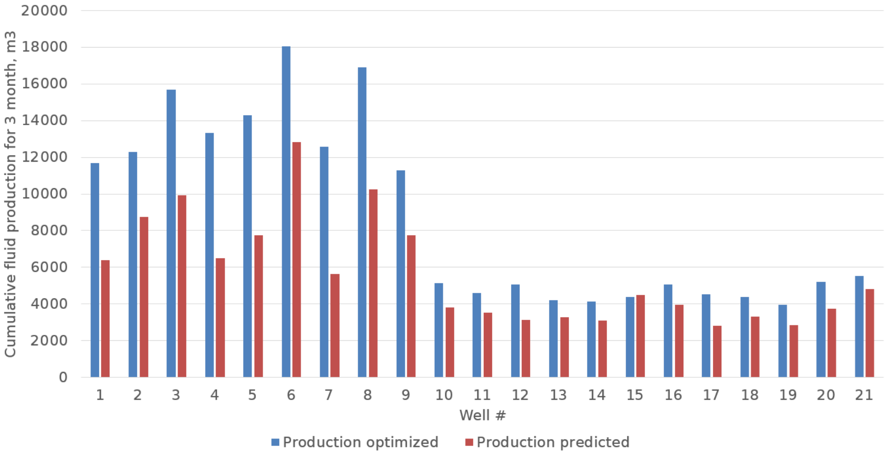


Figure A-14: Production increase with optimal parameters set

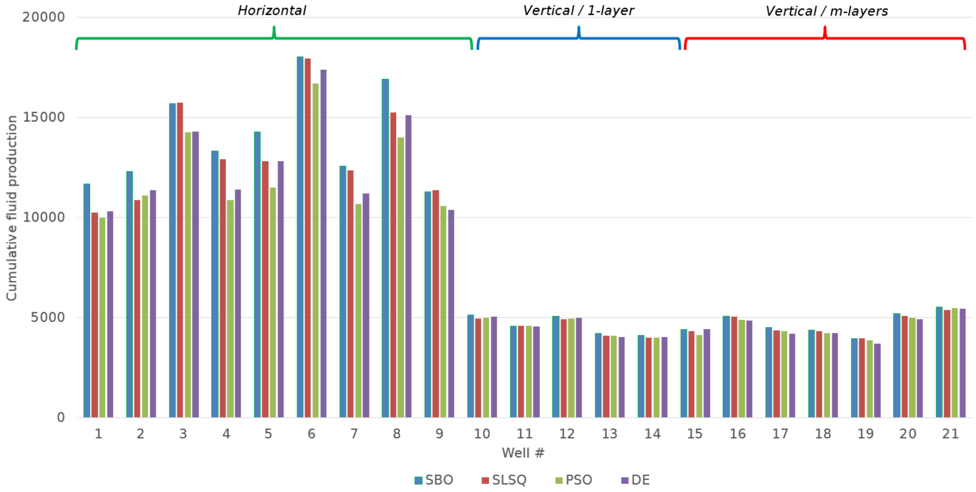


Figure A-15: Optimized production comparison between methods