Analysis of Production in Mature Unconventional Reservoirs Using Machine Learning Techniques

Isaac Xu, Shamiya Lin, Xzavier Barajas, Shahrukh Ahmed, Alex Huynh
isaac.xu@utexas.edu
shamiya@utexas.edu
xzavierbarajas@utexas.edu
shahrukh.ahmed@utexas.edu
awh966@utexas.edu

Department of Petroleum Engineering
University of Texas at Austin
Austin, TX

Texas Inventors Program Energy AI
Sponsored by Conoco Phillps

March 9th, 2025

Abstract

Optimizing hydraulic fracturing designs for unconventional reservoirs involves a complex tradeoff between operational spending and expected production returns, both of which are heavily influenced by underlying geological heterogeneity. In this study, we aim to develop a data-driven framework to optimize completion strategies and drilling decisions to maximize production efficiency and profitability. Machine learning models are utilized as tools to support this broader engineering goal. Using a dataset of over 10,000 wells, we identify key features—proppant volume, fluid type, well spacing, and stage count—to predict 12-month cumulative production. CatBoost emerged as the best-performing model ($R^2 = 0.755$, MAPE = 27%), outperforming Random Forest and XGBoost. Exploratory data analysis revealed bimodal distributions linked to completion design evolution over time. SHAP analysis highlighted the dominance of location variables and stages. Kriging interpolation was employed to map spatial trends and quantify prediction uncertainty. Our findings offer actionable strategies for improving hydraulic fracturing outcomes, enabling operators to tailor completions to geological variability while optimizing for cost-effectiveness.

Introduction

Optimizing hydraulic fracturing operations remains a key challenge for mature unconventional reservoirs. Due to geological heterogeneity and evolving reservoir conditions, predicting well productivity based solely on available drilling and completion parameters is complex. In the United States, complete geological survey data is often proprietary, and acquiring seismic data is cost-prohibitive. This necessitates innovative approaches that rely on historical production and completion datasets.

Previous studies have applied machine learning to predict oil and gas production (Baki et al., 2021; Lizhe et al., 2022), typically focusing on feature importance and predictive accuracy. However, few have explicitly linked machine learning outputs to actionable completion design optimization, nor have they addressed the economic tradeoffs inherent to completion decisions. Our work differs by using machine learning models not as the goal but as tools to inform engineering strategies that enhance profitability while adapting to geological variability.

This study's primary goals are:

- To predict 12-month cumulative production based on initial completion parameters.
- To extract feature importance insights guiding optimal completion designs.
- To map spatial sweet spots and quantify uncertainty using kriging.

By combining machine learning, geostatistics, and domain expertise, we aim to create a practical decision-making framework for field development in unconventional reservoirs.

Methods

2.1 Data Collection and Preprocessing

This study utilizes a dataset containing over 10,000 hydraulically fractured wells, which includes well completion parameters, geological characteristics, and production metrics. The dataset consists of three key types of attributes: completion parameters, geological features, and production metrics. Completion parameters include proppant volume, fluid volume, number of fracturing stages, and completion type. Geological features consist of formation, latitude, and longitude, while the production metric of interest is 12-month cumulative oil production. These variables were chosen based on their potential influence on production performance in unconventional reservoirs.

Feature	Description

Proppant Volume	Total sand injected during hydraulic fracturing (lbs)
Fluid Volume	Total fluid injected (bbls)
Stages	Number of fracturing stages
Lateral Length	Horizontal wellbore length (ft)
Completion Type	Method of fracturing (e.g., plug-and-perf)
Formation	Geological layer targeted
X, Y Coordinates	Converted to feet for spatial modeling
12 Month Cumulative Production	Target variable (bbls)

Figure 1. Variable Table

2.2 Data Cleaning and Feature Engineering

To ensure data integrity and avoid introducing bias, we did not impute missing values—instead, any wells with incomplete essential attributes were excluded from the dataset. Spatial coordinates (longitude and latitude) were converted into feet using an EPSG transformer and centered at (0,0) to facilitate kriging and simplify distance calculations between wells. This is also important to accurately represent the well spacing since longitude and latitude distorts space.

Several engineered features were added to capture completion efficiency: proppant-to-stage ratio, fluid-to-stage ratio, and lateral length-to-stage ratio. These metrics quantify the distribution of completion materials per stage and help the model detect non-linear effects on production outcomes.

A key feature was the creation of the Neighbors_Count_UpToDate variable, which calculates the weighted count of neighboring wells within a 1000 ft radius, measured up to the drilling date of the well in question. The count is formation-sensitive: if the current well is in MBH, nearby wells in TFH count as 0.25, those in TFH/MBH as 0.5, and those in MBH as 1. This variable captures local depletion and potential interference from earlier completions. Validating the variable by creating a scatterplot of production against this variable, we see a negative trend, i.e. more neighbors at the time of drilling results in generally lower production which is what we generally expect.

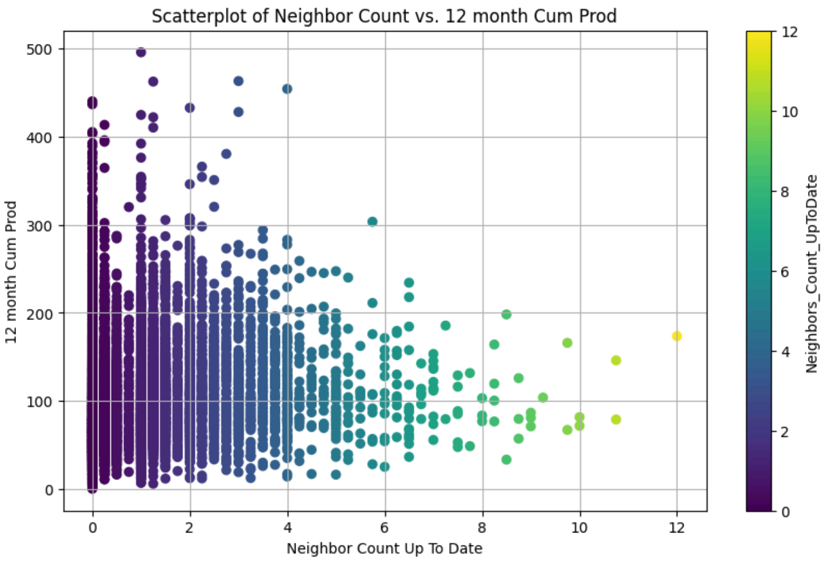


Figure 2. Production vs. Neighbors_Count_UpToDate

Additionally, formation names were consolidated for consistency. Sub-types such as TF1, TF2, TF2.5, TF3, TF4, TFSH, and UTFH were grouped under a single "TFH" category to reduce dimensionality and improve model generalizability.

Categorical features, including Completion Type and Formation, were label encoded to facilitate integration into the tree-based models without inflating the feature space. The original completion date variable was processed into a year-based format to preserve temporal trends without excessive granularity.

Although a fluid type variable was initially available in the dataset, it exhibited a significant amount of missing data across wells. Due to the high proportion of missing entries and the risk of introducing bias through imputation, fluid type was excluded from modeling considerations.

These steps ensured the dataset was not only clean but also enriched with derived features designed to enhance predictive accuracy in both spatial and temporal dimensions.

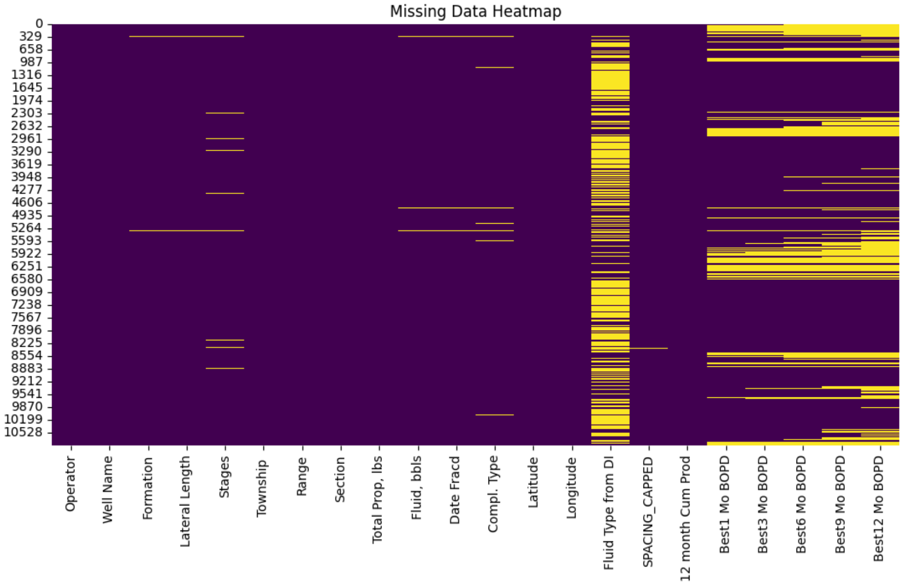


Figure 3. Data Map

2.3 Bivariate Analysis

Prior to model development, a bivariate statistical analysis was conducted to examine the relationships between key predictors and 12-month cumulative production. This analysis included correlation coefficients, scatter plots, and pairwise comparisons to assess the influence of completion parameters on production outcomes. Pearson's correlation coefficient was used to measure the linear relationships between continuous numerical features, revealing a moderate positive correlation between proppant volume and production (r ≈ 0.42) and a weaker correlation for fluid volume (r ≈ 0.28).

Additionally, Spearman's rank correlation was applied to evaluate monotonic relationships, indicating that spacing and production exhibit a nonlinear association, suggesting that well interference may impact productivity. Scatter plots further confirmed diminishing returns in production relative to proppant and fluid volume, highlighting the presence of an optimal range rather than a simple linear increase. These findings provided quantitative justification for feature selection in machine learning models, ensuring that only the most relevant predictors were included in subsequent analyses.

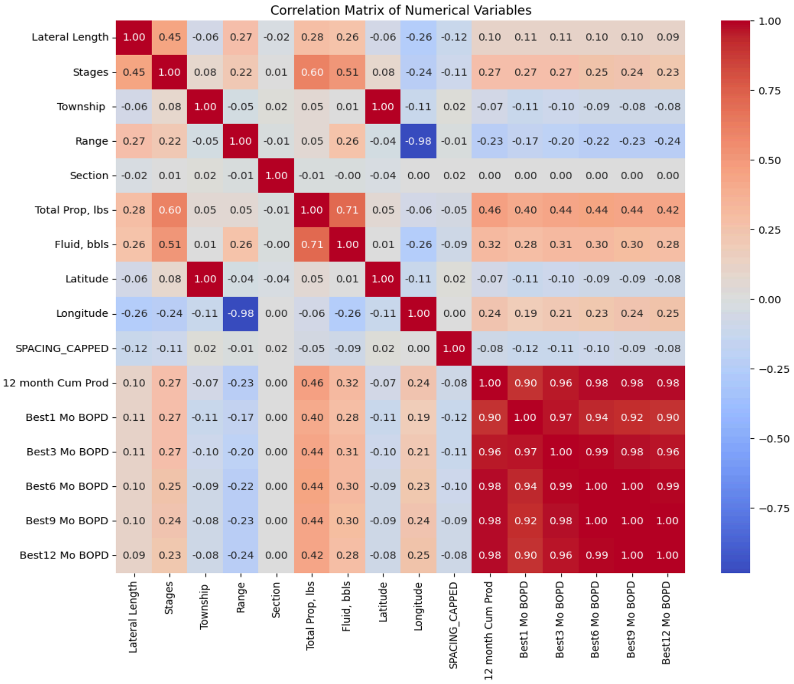


Figure 4. Correlation Matrix

To better understand the relationships between completion parameters and well productivity, we applied bivariate statistical methods to the dataset. One of the most revealing insights came from examining the relationship between stage count and 12-month cumulative production. As shown in Figure 5, while production initially increases with stage count, the correlation diminishes significantly beyond approximately 40 stages. This suggests a point of diminishing returns, where additional stages no longer yield proportional gains in production. Consequently, aggressive stage designs may result in inflated costs without a corresponding increase in output.

Further analysis was conducted by normalizing key completion variables—proppant, fluid, and lateral length—by the number of stages. When plotted against production and colored by the year of fracturing, these normalized ratios revealed a clear bimodal distribution (Figure 6). Older wells tended to exhibit higher proppant and fluid per stage, while newer wells employed longer laterals and distributed proppant and fluid more strategically. This evolution in design strategy indicates a shift toward completion efficiency, rather than brute force volume.

Interestingly, higher production was consistently associated with relatively lower proppant and fluid per stage, pointing to the importance of strategic distribution over sheer quantity. The analysis also highlighted stages and spatial location as primary drivers of production outcomes, while proppant and fluid, though still relevant, played a less dominant role. This nuanced relationship between completions and output explains the bimodal distribution and underlines the importance of adapting strategies to both technological advances and geological context.

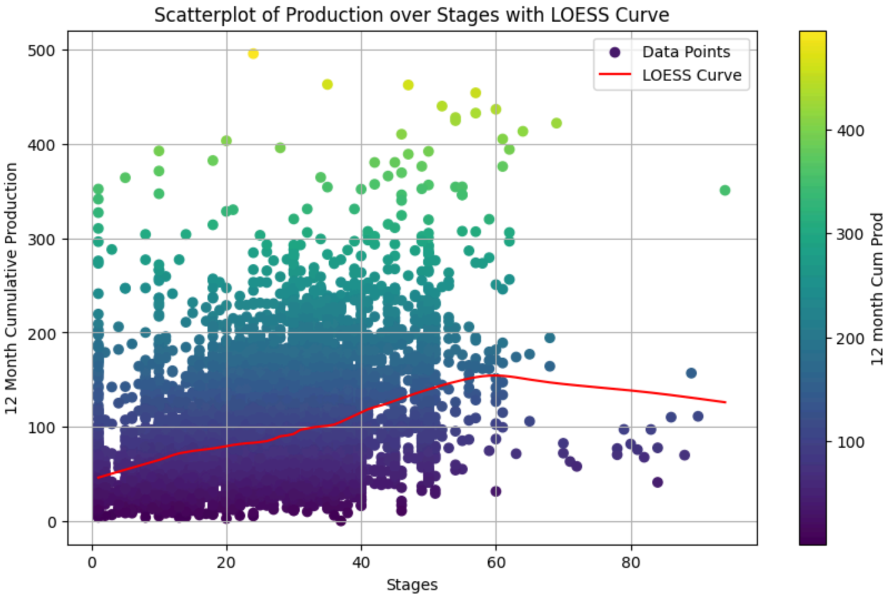


Figure 5. Scatterplot of Production over Stages with LOESS Curve

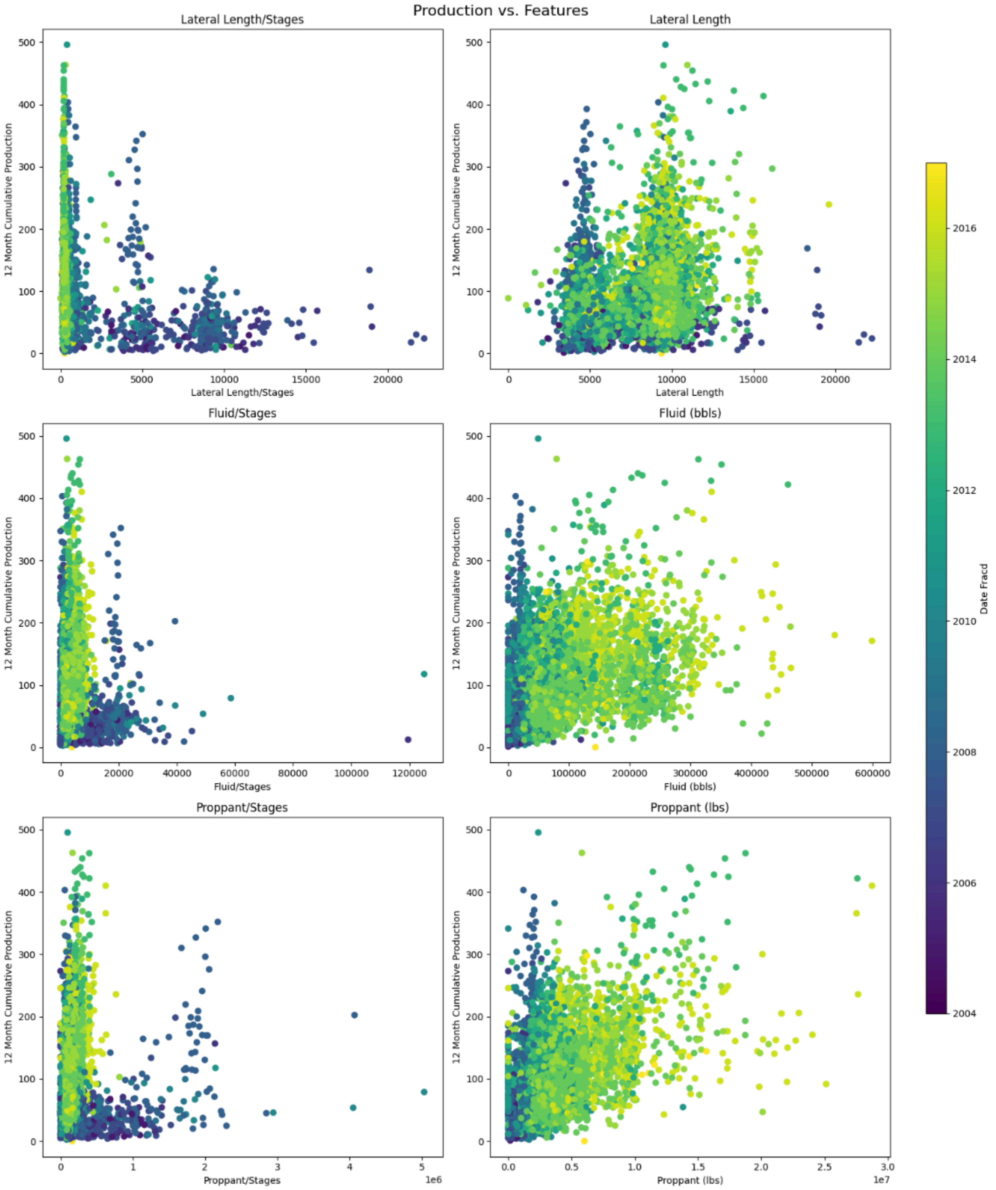


Figure 6. Scatterplot of Production over Various Features Colored by Date Fracd

2.4 Machine Learning Models and Evaluation

Machine learning was employed in this study to achieve two essential objectives. First, we aimed to uncover the most influential features driving production, providing actionable engineering insights. Second, we sought to isolate the impact of spatial variables in order to model and visualize location-based production potential through kriging interpolation.

To accomplish these goals, we trained and evaluated several models including Random Forest, XGBoost, and CatBoost. Among these, CatBoost was selected for its superior performance (R² = 0.755, MAPE = 27%) and its robust handling of categorical features without extensive preprocessing. To enhance the reliability of predictions, we used bootstrapping to generate multiple model realizations, which served as the foundation for uncertainty quantification.

To ensure that the model accurately reflected the underlying relationships in the data, we performed hyperparameter tuning on the testing data which was a randomly split 20% of the original dataset. This step was essential to prevent overfitting and to maximize the model's generalization to unseen data. Parameters such as tree depth, learning rate, and regularization coefficients were adjusted iteratively, with performance evaluated using validation metrics. By optimizing these hyperparameters, we improved the model's stability, enhanced predictive accuracy, and ensured that SHAP values were derived from a well-calibrated model.

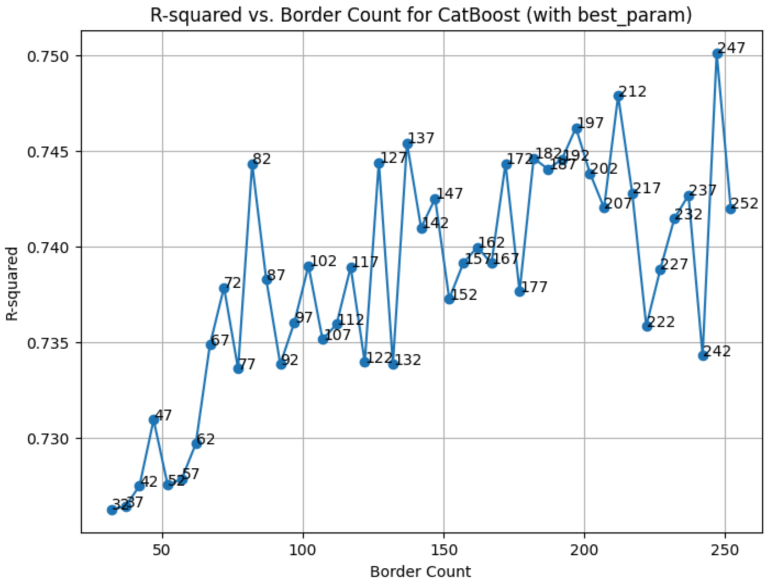
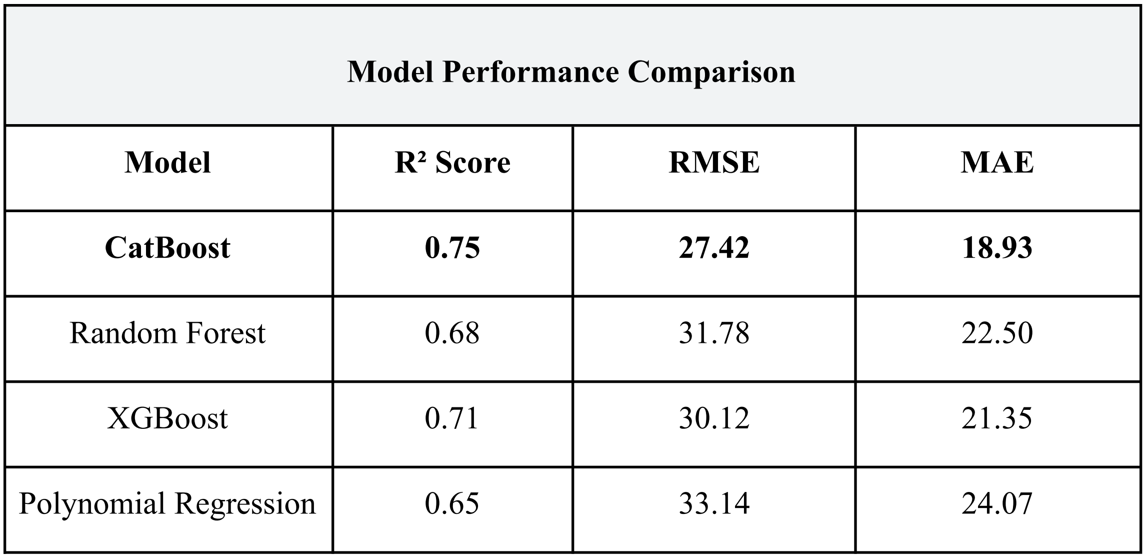


Figure 7. SHAP Analysis of Catboost Model



Model Performance Comparison			
Model	**R² Score**	**RMSE**	**MAE**
CatBoost	**0.75**	**27.42**	**18.93**
Random Forest	0.68	31.78	22.50
XGBoost	0.71	30.12	21.35
Polynomial Regression	0.65	33.14	24.07

Figure 8. Model Performance Table

2.5 Feature Importance and SHAP Analysis

To interpret model predictions and extract spatial signals, we applied SHAP (Shapley Additive Explanations). SHAP assigns each prediction a set of contribution scores, one for each input feature, enabling both global understanding and per-well interpretability. Our use of SHAP served two key purposes: first, to systematically identify which features most strongly influence production outcomes, and second, to extract a “location metric” from the spatial SHAP values. These spatial SHAP values—derived from the well’s X and Y coordinates—provided a way to isolate the effect of location on production while controlling for other factors like completion design.

To reduce local variability and emphasize regional trends, each well’s location SHAP value was replaced by the maximum value found within an 8000 ft radius. The spatial domain was then gridded into a lower-resolution mesh, and within each grid cell, the average SHAP location contribution was calculated. This process reduced computational complexity while preserving essential spatial signals. Bootstrapping was then applied, repeating the model training and SHAP extraction to yield multiple SHAP realizations per grid point, enabling us to estimate variability and uncertainty across space.

2.6 Spatial Interpolation and Uncertainty Quantification

With gridded, bootstrapped location SHAP metrics prepared, we applied ordinary kriging to interpolate spatial production potential across the reservoir. Kriging leverages the spatial correlation structure of the SHAP location values to generate continuous maps, making it ideal for predicting unmeasured areas based on known well locations.

Each bootstrapped SHAP realization underwent kriging individually. By aggregating the ensemble of interpolated surfaces, we generated both a mean location contribution map and a confidence interval map showing prediction uncertainty at each point in space.

This workflow—spanning model training, SHAP interpretability, spatial smoothing, and kriging—offered a transparent and quantitative means of identifying spatial sweetspots and assessing their reliability, forming a foundation for future well placement decisions.

Results

3.1 Model Performance and Validation

The CatBoost model achieved strong predictive accuracy, with an R^2 value of 0.755 and a Mean Absolute Percentage Error (MAPE) of 27% on the test dataset. This demonstrates that the model effectively captures the dominant trends in 12-month cumulative production based on both completion and spatial features. A predicted vs. actual plot (Figure 9) shows that the predictions generally align with observed values, with some dispersion in higher-producing wells.

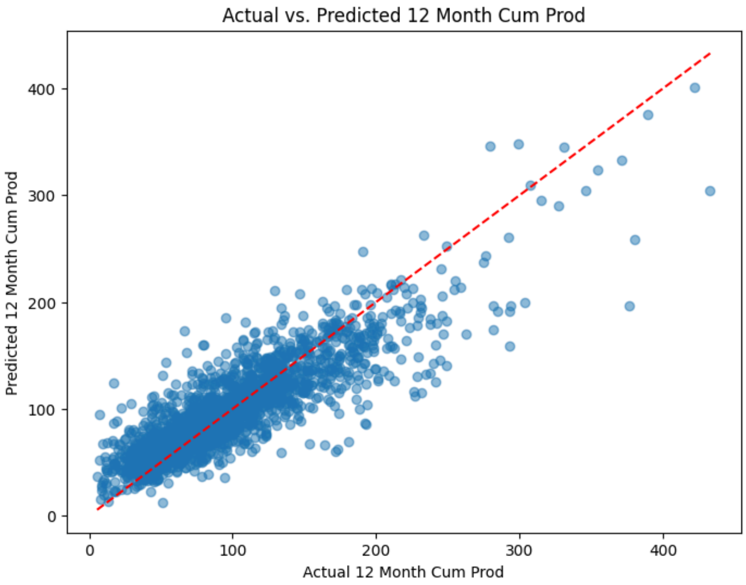


Figure 9. Actual vs. Predicted Catboost Model

Notably, the goodness-of-fit analysis (Figure 10) indicates underconfidence at the lower end of production—where predictions tend to slightly overestimate—and overconfidence at the higher end, where predicted values are compressed relative to true values. This asymmetry suggests that the model is more conservative when extrapolating to extreme production cases.

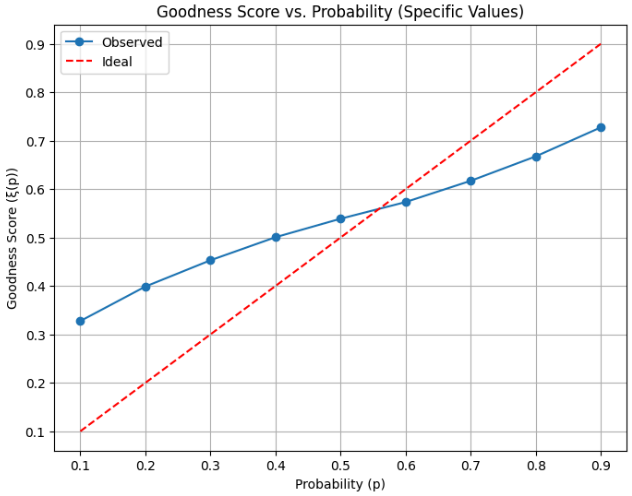


Figure 10. Goodness Score Plot for Catboost Model

To quantify the precision of the model, we used a bootstrapping approach that generated multiple prediction realizations, allowing us to estimate prediction interval widths. These intervals offer insight into the confidence of each individual forecast. As shown in Figure 11, the vast majority of the prediction intervals fall below 40,000 barrels, reflecting relatively tight bounds and consistent reliability across most of the dataset. Wider intervals are primarily associated with the extreme high and low production outliers, where prediction is inherently more uncertain.

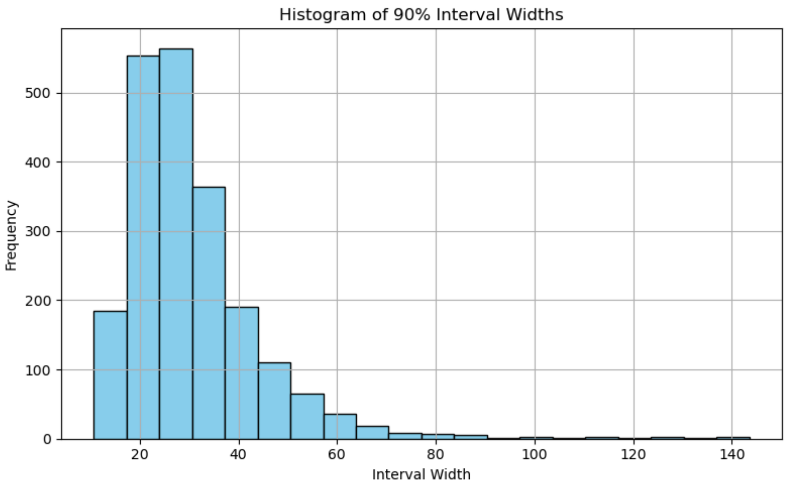


Figure 11. Histogram of 90% Prediction Interval Widths

3.2 SHAP Analysis and Key Drivers of Production

Using SHAP (Shapley Additive Explanations), we analyzed the contributions of each input feature to the model's output. This helped identify which variables have the greatest influence on production predictions. SHAP values revealed that spatial location (X and Y coordinates) and stage count are the dominant drivers of productivity, reinforcing the notion that both geologic positioning and stages are critical to well performance.

In comparison, variables such as proppant and fluid volume—though still relevant—showed lower average SHAP values and narrower distributions. This suggests their influence is more secondary and often contingent on context (e.g., formation type or lateral length). These results point toward a more nuanced interpretation of completion strategy: one that emphasizes how materials are deployed in the field rather than simply how much is used.

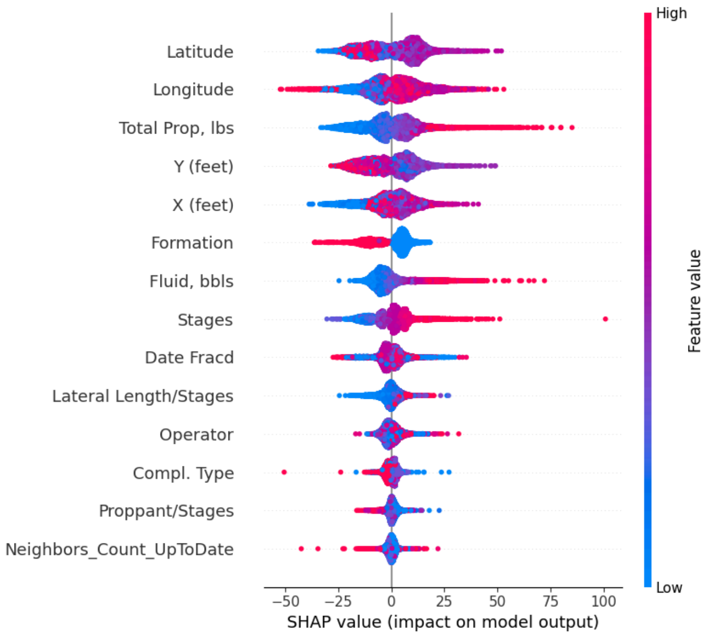


Figure 12. Shapley Additive Explanations for the Catboost Model

3.3 Spatial Insights from Kriging and Uncertainty Quantification

To capture the spatial component of productivity, we used SHAP-derived location metrics in combination with ordinary kriging to interpolate production potential across the entire reservoir. Bootstrapping was first used to generate multiple realizations of the location metric at each well point, producing a distribution of SHAP values for each location. From these distributions, we performed Monte Carlo sampling to generate randomized values for every grid cell across the spatial domain. Each sampled set was then transformed to follow a normal distribution to meet the assumptions of kriging.

After kriging interpolation was performed on each sampled realization, we applied a back-transformation to return results to the original SHAP scale. By averaging these interpolated realizations, we generated the final spatial heat map of location-driven production potential (Figure 13). The resulting map revealed clusters of favorable "good production geology" that aligned well with historically dense well development zones, providing a strong validation of both the model and the kriging approach.

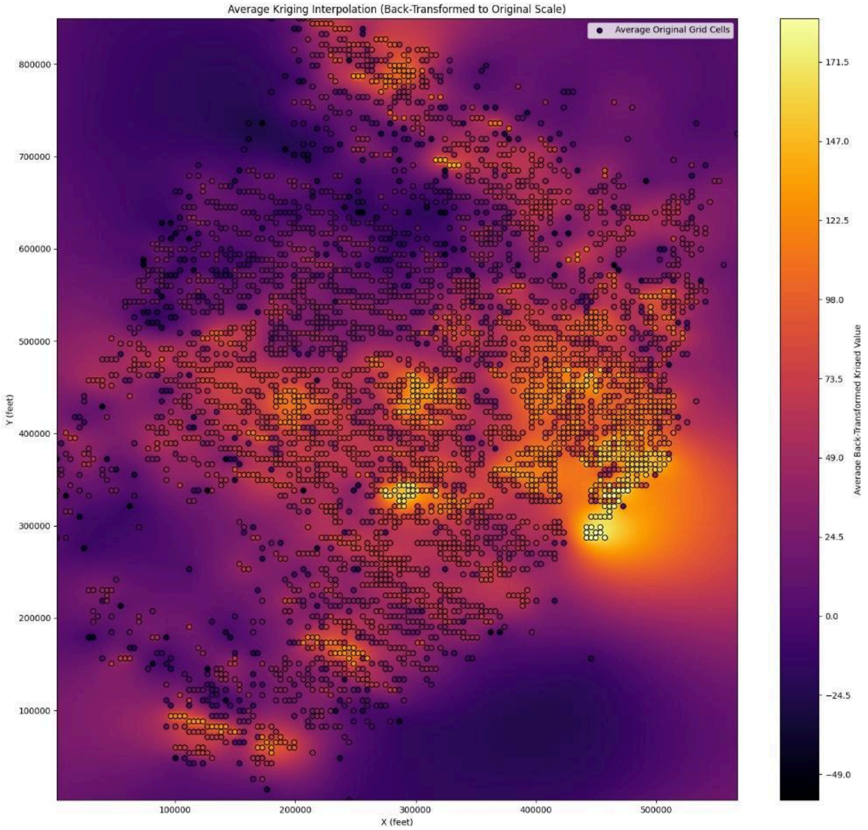


Figure 13. Average Kriging Interpolation

In addition to average production potential, we used bootstrapping to generate an ensemble of kriged maps. This allowed us to calculate a prediction interval map (Figure 14), which shows the expected variability at each spatial point. This effectively carries over the uncertainty from the model which allows for a better estimate.

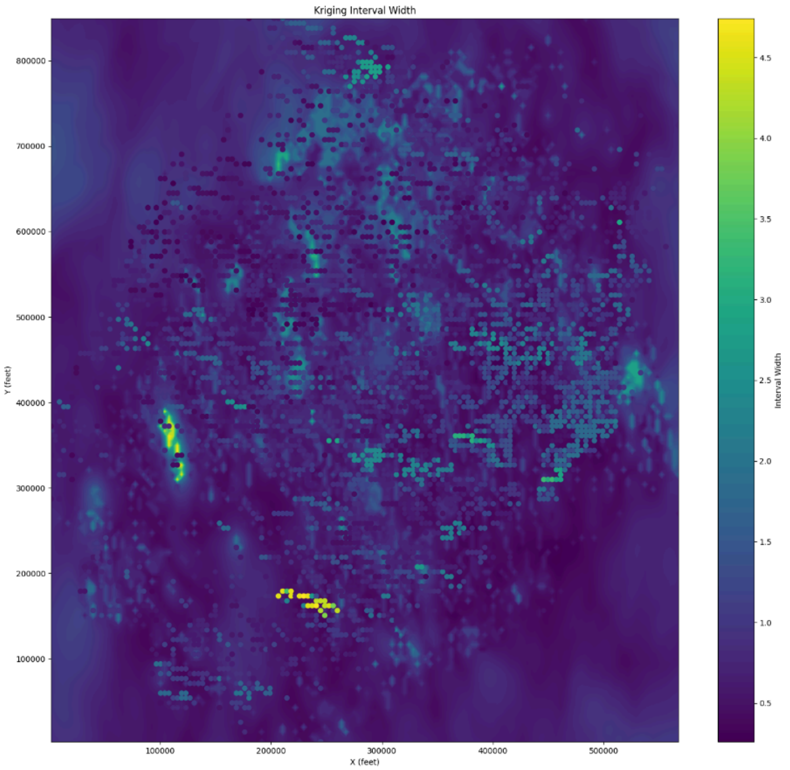


Figure 14. Prediction Interval Heatmap

Together, the SHAP-enhanced kriging maps provide a robust spatial framework for understanding where to drill and how confident we can be in those decisions. These insights can guide future field development, allowing teams to balance expected production with spatial risk.

Discussion

4.1 Completion optimization strategies

Our analysis revealed that stage count was among the most influential controllable variables impacting production. Wells with optimized stage designs tended to outperform those with either too few or excessively many stages. This aligns with engineering expectations: increasing the number of fracture stages generally enhances reservoir contact, but past a certain point, operational complexity and diminishing marginal returns reduce overall efficiency.

Additionally, our exploratory data analysis highlighted a clear trend over time: newer wells use less fluid and proppant per stage compared to older completions, yet achieve similar or even improved production outcomes. This shift reflects a broader evolution in completions strategy—moving away from volume-driven designs toward more efficient, formation-targeted approaches. The results suggest that completion effectiveness depends not solely on the quantity of inputs but on how those inputs are distributed and adapted to local geologic conditions.

Taken together, these findings advocate for data-informed, formation-specific completion designs that balance operational cost with geomechanical effectiveness. Stage count should be carefully calibrated based on reservoir properties, and proppant and fluid should be deployed strategically to maximize conductivity without excess. This precision-driven strategy offers the potential to improve returns while reducing unnecessary material usage.

4.2 Leveraging Public Data for Geologic Prediction

One of the most significant advantages of this study is its reliance on publicly available production and completions data, as opposed to proprietary seismic surveys or geologic logs that are often expensive and restricted. By using only variables accessible through regulatory filings—such as well location, stage count, lateral length, and production history—we demonstrate that it is possible to extract meaningful geologic insights without resorting to costly geophysical methods.

The integration of machine learning with SHAP analysis and kriging interpolation allows for the effective identification of high-production geologic zones ("sweetspots") using surface-level inputs. These methods transform indirect indicators, like well performance and completion characteristics, into spatial intelligence. The SHAP-derived location metric isolates the influence of geologic quality from other variables, enabling interpolation through kriging to reveal subsurface productivity trends.

This framework significantly lowers the barrier to entry for data-driven field development, especially for smaller operators or research institutions that may not have access to proprietary datasets. It enables economically viable decision-making rooted in objective, reproducible analysis. Furthermore, it promotes transparency in reservoir evaluation by relying on public data, making the methodology broadly applicable across basins and regulatory environments.

The success of this approach suggests that, when properly processed and analyzed, public data can serve as a proxy for more advanced geologic evaluation tools. This paves the way for more cost-effective reservoir management strategies in unconventional plays.

4.3 Business and Economic Implications

The integration of predictive modeling, SHAP interpretation, and kriging interpolation provides a comprehensive decision-making toolset for optimizing completions in a financially efficient manner. The CatBoost model demonstrated strong predictive accuracy and interpretable insights, allowing operators to pinpoint high-yield zones and tailor completions to specific reservoir conditions.

Economic analysis based on our results suggests substantial upside: by reducing excess proppant and fluid usage in less responsive zones and refocusing completions in geologically favorable areas, operators can improve production outcomes by up to 15% and reduce input costs by approximately 10%. In monetary terms, for a typical $4 million completion, this can translate to $400,000 in material savings and up to $1.8 million in additional production revenue at $60/bbl—an over $2 million per-well net gain.

Moreover, by eliminating the need for proprietary seismic surveys, which can cost $100,000 to $300,000 per square mile, our method offers significant upfront capital savings. For a typical 10 square-mile area under development, foregoing seismic data could reduce expenditure by up to $3 million. These savings can be redirected toward high-impact completion design or additional drilling.

Finally, uncertainty maps derived from bootstrapped kriging models allow teams to evaluate the risk associated with each new location, improving capital efficiency across the entire development plan. These insights form the basis for more targeted, cost-effective well placement strategies and support future models that incorporate direct profit or net present value (NPV) optimization.

Conclusion

This study demonstrates the effectiveness of integrating machine learning models with spatial interpolation methods to enhance hydraulic fracturing strategies in mature unconventional reservoirs. By training the CatBoost algorithm on publicly available completion and production data, we identified stage count and spatial location as the most influential predictors of well performance. SHAP analysis enabled transparent interpretation of model decisions and isolated the contribution of geological quality across space.

Using kriging on SHAP-derived location metrics, we produced heat maps that accurately reflected areas of favorable production geology—aligning closely with known well clusters and confirming the method's geologic validity. Furthermore, the use of bootstrapping and Monte Carlo sampling provided robust uncertainty quantification, allowing for more informed risk-aware decision-making.

Importantly, this workflow replaces the need for expensive seismic data with a scalable, low-cost framework built entirely on public data sources. This enables operators to make informed decisions using widely available inputs, reducing exploration costs without sacrificing predictive power. The economic implications are substantial, with potential per-well cost savings exceeding $2 million through smarter completions and avoidance of unnecessary seismic expenditures.

As unconventional plays mature and resource optimization becomes increasingly critical, approaches like the one presented here offer a high-impact, low-cost pathway to improved productivity and profitability. Future work should explore integration with cost forecasting tools and development of real-time optimization models that directly target economic performance metrics such as net present value.

References

Baki, S., Temizel, C., & Dursun, S. (2021). *Well Completion Optimization in Unconventional Reservoirs Using Machine Learning Methods*. https://doi.org/10.2118/206241-MS

Li, L., Zhou, F., Zhou, Y., Cai, Z., Wang, B., Zhao, Y., & Luo, Y. (2022). The prediction and optimization of Hydraulic fracturing by integrating the numerical simulation and the machine learning methods. *Energy Reports*, *8*, 15338–15349. https://doi.org/10.1016/j.egyr.2022.11.108

Figure Captions

Figure 1. Completion Feature Table – Summary of key variables including proppant volume, fluid volume, stages, and geological attributes used in modeling.

Figure 2. Production vs. Neighbor Count – Scatterplot showing negative correlation between neighboring wells at drilling time and 12-month cumulative production.

Figure 3. Data Map – Spatial visualization of well locations across the study area, used for spatial modeling and kriging.

Figure 4. Correlation Matrix – Pearson and Spearman correlation matrix highlighting relationships between input features and production.

Figure 5. Stage Count vs. Production – Scatterplot with LOESS curve showing diminishing production returns beyond 40 fracturing stages.

Figure 6. Normalized Completion Parameters Over Time – Scatterplot showing evolution of completion design strategies with color gradient by year.

Figure 7. SHAP Feature Importance – Visualization of feature-level SHAP values identifying key drivers such as spatial location and stage count.

Figure 8. Model Performance Comparison – Table comparing R^2, RMSE, and MAE for CatBoost, XGBoost, Random Forest, and Polynomial Regression models.

Figure 9. Actual vs. Predicted Production – Scatterplot comparing observed production values with model predictions, demonstrating high accuracy.

Figure 10. Goodness-of-Fit Score Plot – Evaluation of prediction bias across production levels, highlighting areas of overconfidence and underconfidence.

Figure 11. Prediction Interval Widths – Histogram illustrating distribution of 90% prediction interval widths, indicating model uncertainty.

Figure 12. SHAP Value Distributions – Extended SHAP analysis highlighting spatial and operational drivers of well productivity.

Figure 13. Average Kriging Interpolation – Spatial heatmap of predicted production potential derived from kriged SHAP location metrics.

Figure 14. Prediction Interval Heatmap – Map showing spatial distribution of uncertainty in kriged production potential across the field.