**U.S. Oil & Gas Production and Disposition Analysis (2015–2025)**

**Author**: Joel Akoun – Data Scientist

**Date**: October 2025

1. **Executive Summary**

**Business Perspective:**

This project analyzes U.S. Oil & Gas production (2015–2025) from the Office of Natural Resources Revenue (ONRR) to understand and predict national production trends across states, commodities, and land categories. By turning historical records into predictive insights, the work supports better operational planning, resource allocation, and energy policy decisions.

**Technical Perspective:**

Using supervised learning, we built two complementary solutions:

* Time Series Forecasting to predict monthly production volumes (BOE)
* Classification to detect high vs low output operations

**Key Outcomes:**

* **Forecasting**: ARIMA/SARIMA captured long-term trends and seasonality, producing low errors and balanced bias suitable for planning.
* **Classification**: Extra Trees achieved high accuracy and excellent balance between precision and recall, reliably identifying high-producing operations.
* **Drivers of production**: The most influential factors include Commodity Type (Oil/Gas), Disposition Group (e.g., Sales, Inventory, Transferred), and operational indicators such as Is Positive Volume, Volume Class, and Land Category (Onshore).

1. **Business Problem**

The U.S. oil and gas industry operates across diverse regions, commodities, and land types, where production levels fluctuate due to market conditions, regulations, and operational efficiency.

Despite the abundance of historical production data, companies often lack predictive tools to anticipate output trends and identify underperforming operations early.

This gap limits effective budgeting, planning, and resource deployment, especially when decisions depend on future production expectations rather than past performance.

The objective of this project is to use machine learning to:

* Forecast production volumes (timeseries) to support operational and financial planning.
* Classify operations by performance (high vs. low output) to guide field optimization.
* Uncover key drivers such as commodity type, disposition, and land characteristics that explain production variability across regions and time.

By addressing these challenges, the project enables data-driven forecasting, smarter investment allocation, and more transparent production management within the U.S. energy sector.

1. **Data Overview**

**Data Source:**

The dataset comes from the Office of Natural Resources Revenue (ONRR) and contains monthly U.S. Oil and Gas Production and Disposition records from 2015 to 2025.

It includes information on commodities, operational categories, and geographic locations, representing both onshore and offshore production across U.S. states and counties.

The data covers over 500,000 observations with two modeling targets:

* Time Series Forecasting : Predicting monthly production volumes (in Barrels of Oil Equivalent, BOE) to understand trends and seasonality over time.
* Classification: Output Class - high vs. low production category

|  |  |  |
| --- | --- | --- |
| **Category** | **Example Features** | **Description** |
| Commodity | Commodity Type, Commodity Unit | Identifies what is produced (e.g., oil, gas). |
| Operational | Disposition Group, Volume Class, Is Positive Volume | Describes how production is reported and categorized. |
| Geographic | State, County, Offshore Region, Land Category, Land Class | Defines where production occurs. |
| Temporal | Production Date, Year, Month | Captures seasonal and long-term trends. |

**Data Quality & Preprocessing Notes:**

* Missing values were primarily concentrated in the Offshore Region column (~97% missing) and handled appropriately (drop or encode).
* Text-based fields were standardized and encoded for model readiness.
* Numerical fields such as Volume were converted and scaled for comparability across models.
* Derived features such as Year, Month, and Days Since Start were created to capture temporal patterns.

1. **Data Preparation**

Before modeling, the dataset was carefully cleaned, standardized, and transformed to ensure consistency and reliability.

Irrelevant identifiers such as *FIPS Code* and *Disposition Code* were removed, and missing values were addressed most notably in the Offshore Region, where about 97% of entries were missing and encoded as “None” or excluded when necessary.

Production volumes were converted to numeric values and standardized for comparability across records.

From the Production Date, new features such as Year, Month, and Days Since Start were derived to capture temporal patterns and production cycles.

Operational indicators including Volume Class and Is Positive Volume were also created to support classification and interpretability.

Categorical variables like *State*, *County*, and *Land Class* were encoded for model compatibility, while numerical features were scaled using StandardScaler for algorithms sensitive to feature magnitude.

Finally, the data was split to reflect each modeling objective: a chronological split (train ≤ 2023, test 2024-2025) for regression to simulate real forecasting, and a random split for classification to maintain balanced representation across classes.

This structured preparation ensured that all models received clean, well-engineered data, optimized for both accuracy and interpretability.

1. **Modeling Approach**

Two supervised learning approaches were developed:

* Time Series Forecasting to model and predict production volume trends over time.
* Classification to identify high- vs. low-output operations.

For the forecasting task, several classical time series models (AR, MA, ARMA, ARIMA, and SARIMA) were tested.

The ARIMA and SARIMA models achieved the best overall results, effectively capturing both long-term production trends and seasonal variations. These models provided robust forecasts with low error rates (MAPE < 8%) and balanced bias, indicating accurate and reliable temporal predictions.

For the classification task, a range of machine learning models were evaluated, including Logistic Regression, Decision Tree, Random Forest, Linear SVM, and Extra Trees Classifier.

The Extra Trees Classifier outperformed all others, achieving the highest accuracy (~95%) and ROC-AUC (~0.99), with excellent balance between precision and recall. This model was retained for its strong performance and interpretability, supporting the feature importance analysis in later sections.

1. **Model Evaluation & Results**

For the time series forecasting, both ARIMA and SARIMA models performed consistently well.

They accurately captured the main production trends and seasonal variations across years.

ARIMA provided stable overall forecasts, while SARIMA slightly improved performance when seasonal patterns were stronger.

Both models showed low prediction errors and balanced bias, confirming that production volumes follow clear long-term and seasonal behaviors.

For the classification task, the Extra Trees Classifier delivered the most reliable results.

It achieved high overall accuracy and a strong balance between precision and recall, effectively distinguishing high-producing operations from lower-output ones.

The confusion matrix showed very few missed high producers, indicating excellent sensitivity and generalization to new data.

Overall, the forecasting and classification models complement each other—

ARIMA and SARIMA support accurate volume prediction over time,

while the Extra Trees model provides clear operational insights for performance classification.

1. **Feature Importance & Insights**

Feature importance analysis from the Extra Trees model revealed that production outcomes are most influenced by operational and commodity-related variables.

The top predictors include:

* Is Positive Volume and Volume Class: strongest indicators of active, high-yield operations.
* Disposition Group (Sales, Inventory, Transferred): how production is reported and managed strongly affects output classification.
* Commodity Type (Oil) and Commodity Unit (BBL/MCF): highlight differences in production behavior across commodities.
* Land Category (Onshore) and temporal features such as Year and Months Since Start add moderate predictive power, capturing regional and time-based trends.

Production performance is primarily driven by volume behavior, disposition type, and commodity characteristics, while spatial and temporal factors play a secondary but consistent role in refining model accuracy.

1. **Model Validation & Reliability**

To ensure reliability, validation was tailored to each modeling objective:

* For time series forecasting, a chronological split (train ≤ 2023, test 2024–2025) was used to simulate real forecasting. This preserved temporal order and avoided data leakage. The models performed consistently across folds, confirming stability and robustness.
* For classification, a randomized split with balanced classes ensured unbiased evaluation.

Metrics such as Accuracy, F1, and ROC-AUC were stable across folds, confirming generalization to unseen data. The use of class weighting further minimized bias toward the majority (non–high-producer) class.

Together, these validation strategies confirmed that both the forecasting and classification models are reliable, reproducible, and well-calibrated for real-world application.

1. **Business Impact**

The developed models demonstrate how combining statistical time series forecasting with machine learning classification can transform raw production data into actionable insights for the U.S. oil and gas industry.

* Forecasting models (ARIMA/SARIMA) provide accurate estimates of future production volumes, enabling better capacity planning, budgeting, and resource allocation.
* Classification models (Extra Trees) enable early detection of high-producing operations, allowing operators to prioritize assets, plan maintenance, and optimize investment.

At the strategic level, these predictive tools support policy development and market analysis, helping stakeholders understand future supply patterns and improve transparency in energy management.

Overall, this work demonstrates the value of integrating predictive analytics into production management for smarter, data-driven decision-making.

1. **Recommendations**

To enhance model performance and operational integration, several improvements are recommended:

1. Model Enhancement:

Explore advanced ensemble and hybrid forecasting models (e.g., Gradient Boosting, XGBoost, Prophet) to capture additional non-linear and seasonal behaviors.

1. Feature Expansion:

Incorporate lagged production features (e.g., previous 3-6 months) and external factors such as market prices, weather, or regulatory changes to capture broader production influences.

1. Automation & Deployment:

Develop an automated model retraining pipeline that updates predictions quarterly as new ONRR data becomes available.

Integrate results into a dashboard or API for operational visibility and real-time decision support.

1. Monitoring & Maintenance:

Implement model drift monitoring to detect performance degradation over time, ensuring continuous reliability and transparency.

These actions would transition the project from analysis to a scalable, production-ready forecasting system, providing sustained value for decision-making in the oil and gas sector.

1. **Conclusion**

This project successfully demonstrated how time series forecasting and classification models can predict and explain U.S. oil and gas production trends using ONRR data from 2015–2025.

The ARIMA and SARIMA models effectively captured production trends and seasonality, while the Extra Trees Classifier achieved high precision and recall in identifying top-performing operations.

Key production drivers’ volume behavior, disposition type, commodity characteristics, and land category were identified as critical factors influencing national output.

Overall, the models improved forecasting accuracy and operational intelligence, proving that data-driven forecasting can enhance efficiency, transparency, and strategic decision-making across the U.S. energy sector.