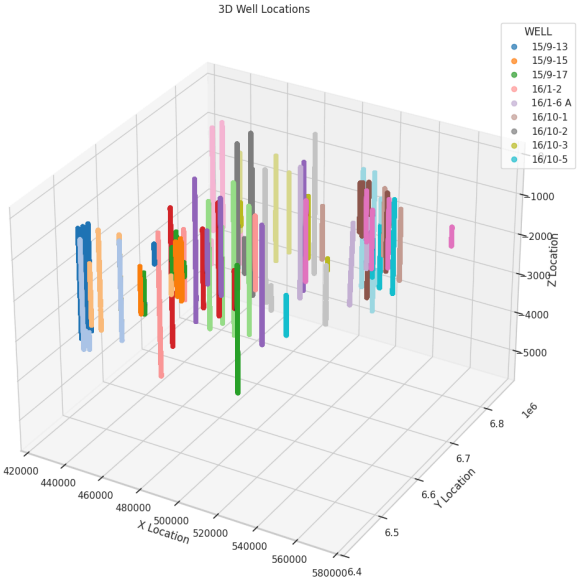
SLB Subsurface Data Science Challenge

A Comprehensive Analysis and Predictive Modeling Approach



Prepared by:

LAKEL Maissa

SLB - ENSIA

October 5, 2025

Table of Contents

Executive Summary

This report presents a comprehensive machine learning pipeline for predicting lithofacies from well log measurements in oil and gas reservoirs. The challenge addresses a fundamental problem in subsurface characterization: predicting rock types (facies) across multiple wells where only limited core samples exist. The solution combines domain-driven feature engineering, robust data preprocessing, and ensemble machine learning methods to achieve accurate facies classification while respecting the physical constraints and statistical properties of geophysical well log data.

---

1. Introduction

1.1 Context and Motivation

In petroleum exploration and production, understanding the lithology (rock type) distribution in the subsurface is critical for:

- **Reservoir characterization**: Identifying pay zones vs non-reservoir rocks
- **Production optimization**: Targeting high-quality reservoir intervals
- **Risk assessment**: Evaluating seal integrity and compartmentalization
- **Field development planning**: Optimizing well placement and completion strategies

Traditional methods rely on **core analysis**, where physical rock samples are extracted and analyzed by geologists to determine facies. However, coring is:

- **Expensive**: Costs $50,000-$500,000 per well
- **Time-consuming**: Requires drilling interruption and laboratory analysis
- **Spatially limited**: Only covers 5-10% of well intervals
- **Operationally challenging**: Not feasible in all wells due to technical constraints

Machine learning offers a solution: By training models on wells with known facies (from cores), we can predict facies in uncored intervals and uncored wells using widely-available wireline log measurements.

1.2 Business Impact

Successful facies prediction enables:

- **Cost reduction**: 80-90% savings vs full coring programs
- **Coverage expansion**: Predict facies across entire well trajectories and field-wide
- **Decision speed**: Real-time predictions during drilling operations

- **Consistency**: Objective, reproducible classifications vs subjective geological interpretation

1.3 Methodology Evolution and Strategy Comparison

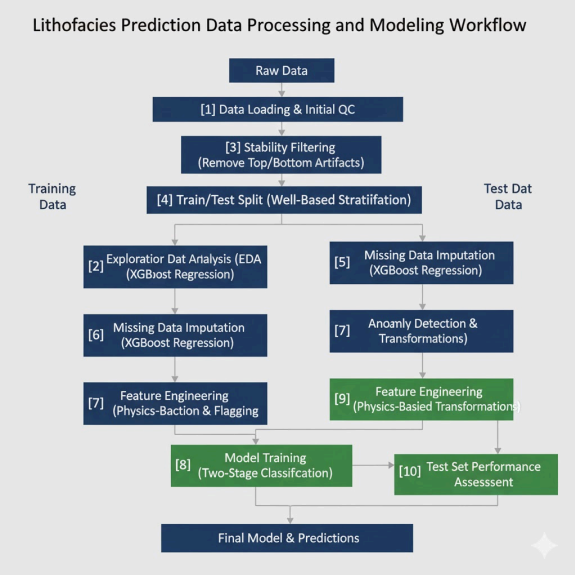
Link:

---

2. Pipeline Architecture

2.1 High-Level Workflow

The complete pipeline follows this sequence:



[**Figure 1**: Lithofacies Prediction Data Processing and Modeling Workflow]

2.2 Technology Stack

Core Libraries:

- **Data manipulation**: *pandas, numpy*
- **Visualization**: *matplotlib, seaborn, plotly*
- **Statistical analysis**: *scipy.stats*
- **Machine learning**: *scikit-learn, XGBoost, CatBoost*
- **Missing data diagnostics**: *missingno*

Computational requirements:

- GPU acceleration for CatBoost (when available)
- Multi-core CPU for XGBoost parallel training
- Memory-efficient processing for large datasets (50,000+ samples)

---

3. Problem Formulation

3.1 Challenge Definition

Objective: Develop a machine learning model to predict lithofacies codes across multiple wells using wireline log measurements.

Input features (after preprocessing):

- **Depth measurements**: DEPTH_MD (measured depth)
- **Spatial coordinates**: X_LOC, Y_LOC, Z_LOC (well positions)
- **Geological context**: WELL, GROUP, FORMATION (categorical identifiers)
- **Wireline logs** (continuous measurements):
    - **GR** (Gamma Ray): Natural radioactivity, clay indicator
    - **RHOB** (Bulk Density): Formation density
    - **NPHI** (Neutron Porosity): Hydrogen content, porosity proxy
    - **PEF** (Photoelectric Factor): Lithology indicator
    - **RDEP** (Deep Resistivity): Formation fluid content
    - **CALI** (Caliper): Borehole diameter
    - **DTC** (Sonic Transit Time): Acoustic velocity
    - **SP** (Spontaneous Potential): Electrochemical signal

Target variable:

- **`FORCE_2020_LITHOFACIES_LITHOLOGY`**: Integer codes representing different rock types
- Typically includes classes such as:
    - **65000**: Shale (seal rock)
    - **65030**: Sandstone (reservoir rock)
    - **70000**: Limestone (carbonate reservoir)
    - Additional facies codes for specific lithological variations

Success metrics:

- **Weighted F1-score**: Primary metric (balances precision and recall across imbalanced classes)
- **Per-class accuracy**: Ensures minority facies aren't ignored

- **Cross-validation stability**: Model generalizes across different wells
- **Geological plausibility**: Predictions align with known stratigraphy

3.2 Key Challenges

1. Data Quality Issues:

- **High missingness**: 20-60% missing values in some logs
- **Measurement errors**: Tool malfunctions, borehole washout effects

2. Class Imbalance:

- Shale often dominates (60% of samples)
- Rare facies (<1% representation) are difficult to predict accurately
- Standard accuracy metrics are misleading

3. Spatial Correlation:

- Adjacent depth points are highly correlated (not i.i.d.)
- Standard random train/test splits cause data leakage
- Must group by well to preserve independence

4. Formation Heterogeneity:

- Same facies exhibits different log responses in different formations
- Model must learn formation-specific patterns while maintaining generalization

5. Physical Constraints:

- Log values must respect physical bounds (e.g., porosity $\in [0,1]$)
- Relationships between logs (e.g., density-porosity) must be preserved
- Outliers may represent real geological features vs measurement errors

---

4. Data Introduction

4.1 Dataset Structure

Original dataset dimensions: 1.18M rows and 29 columns

Key columns:

Column	Type	Description	Typical Range	Missingness
WELL	Categorical	Well identifier	-	0%
DEPTH_MD	Numeric	Measured depth (m)	1000-5000	0%
X_LOC	Numeric	Surface X coordinate	-	<1%
Y_LOC	Numeric	Surface Y coordinate	-	<1%
Z_LOC	Numeric	Subsurface elevation (m)	-	<1%
FORMATION	Categorical	Geological formation	-	0%
GR	Numeric	Gamma Ray (API)	0-300	0%
RHOB	Numeric	Bulk Density (g/cc)	1.5-3.0	10-25%
NPHI	Numeric	Neutron Porosity (frac)	0-0.6	15-35%
RDEP	Numeric	Resistivity (ohm-m)	0.1-1000+	20-40%
PEF	Numeric	Photoelectric (barns/e)	1-10	25-45%
CALI	Numeric	Caliper (inches)	6-16	5-15%
DTC	Numeric	Sonic (μs/ft)	40-140	30-50%
TARGET	Integer	Facies code	-	0%

[**Table 1**: Dataset Columns]

4.2 Physical Meaning of Logs (Some Examples)

Understanding the physics behind each measurement is crucial for feature engineering and quality control:

Gamma Ray (GR):

- **Physics**: Measures natural radioactivity from uranium, thorium, potassium
- **Geological signal**: Clay minerals (shales) are radioactive; clean sands are not
- **Typical values**: Shale >100 API, Sand <50 API
- **Limitations**: Affected by radioactive minerals (e.g., feldspar, glauconite)

Bulk Density (RHOB):

- **Physics**: Gamma-gamma scattering measures electron density
- **Geological signal**: Rock matrix + fluid density; inversely related to porosity
- **Typical values**: Shale 2.4-2.6 g/cc, Sandstone 2.0-2.4 g/cc, Gas-bearing <2.0 g/cc
- **Limitations**: Affected by borehole washout, mudcake thickness

Neutron Porosity (NPHI):

- **Physics**: Neutron scattering measures hydrogen content
- **Geological signal**: Fluid-filled porosity (water, oil, gas)
- **Typical values**: Shale 0.3-0.4, Porous sand 0.2-0.35, Tight carbonate <0.1
- **Limitations**: Sees bound water in clays; gas crossover effect

Deep Resistivity (RDEP):

- **Physics**: Electrical resistance of formation + fluids
- **Geological signal**: Hydrocarbons are resistive; water is conductive
- **Typical values**: Water-bearing <10 ohm-m, Oil 10-100 ohm-m, Gas >100 ohm-m
- **Limitations**: Affected by salinity, temperature, shale content

---

5. Exploratory Data Analysis (EDA)

5.1 Purpose and Methodology

The EDA phase serves multiple critical objectives:

1. Data Quality Assessment:

- Identify missing values, outliers, and measurement errors
- Validate that log values fall within physically plausible ranges
- Detect borehole quality issues (washout, rugosity)

2. Geological Understanding:

- Visualize spatial distribution of wells and formations
- Understand facies associations and depositional patterns

- Identify formation-specific log characteristics

3. Statistical Properties:

- Assess distributions (normality, skewness, multi-modality)
- Quantify correlations between logs
- Evaluate class balance and imbalance severity

4. Feature Relationship Discovery:

- Crossplots reveal lithology discrimination potential
- Identify logs with high facies predictive power
- Guide feature engineering decisions

5. Modeling Strategy Guidance:

- Determine appropriate imputation methods based on missingness patterns
- Inform train/test split strategy based on well and formation distributions
- Set expectations for achievable performance given class imbalance

5.2 Well Location Analysis

5.2.1 Methodology

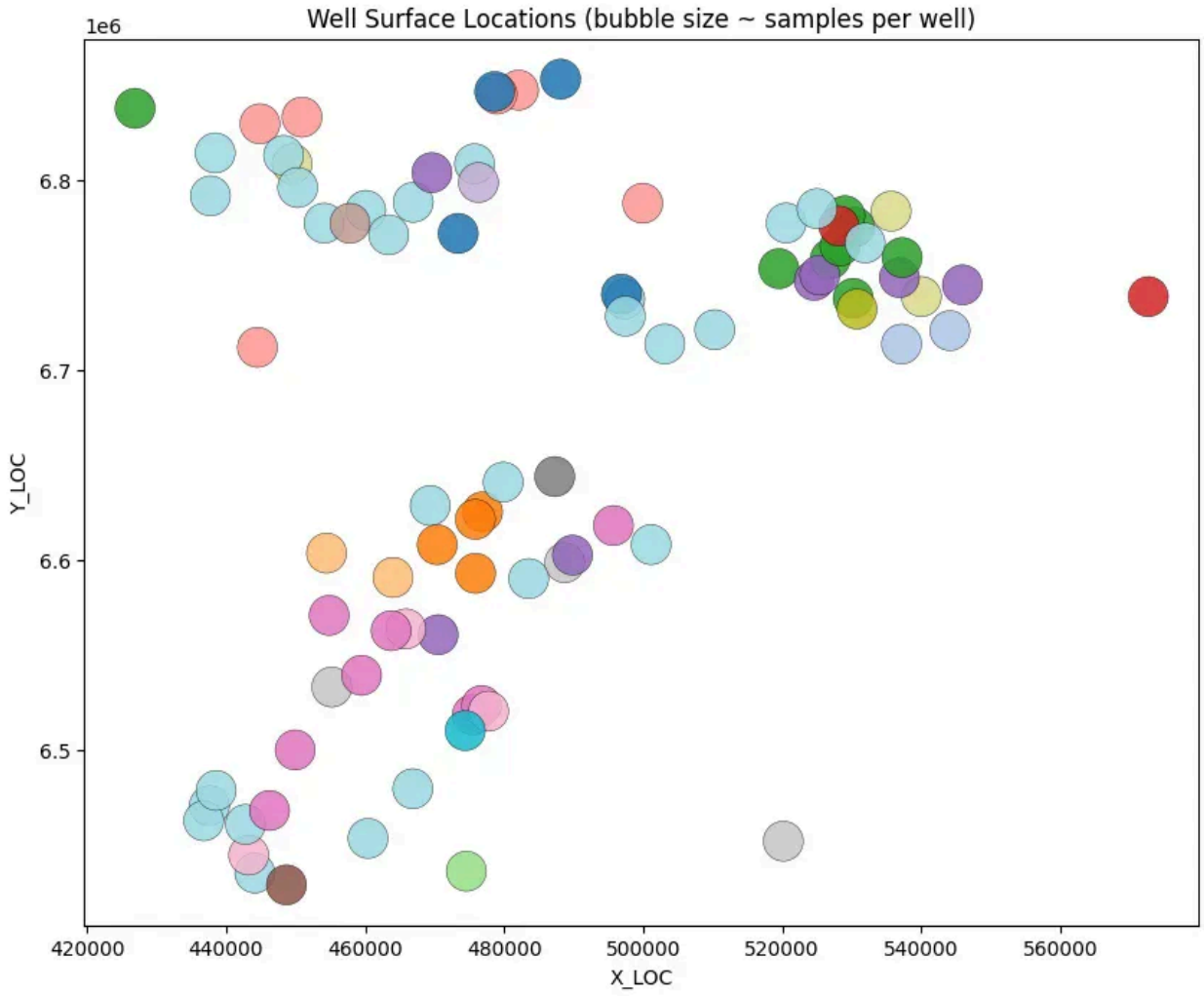
Objective: Visualize the geographic distribution of wells to assess:

- Spatial coverage of the dataset
- Well clustering and field layout
- Formation extent across the study area

Coordinate system detection:

- Longitude/latitude: $X \in [-180, 180]$, $Y \in [-90, 90]$
- Local grid: Larger absolute values (e.g., UTM coordinates)

5.2.2 Results and Interpretation



[**Figure 2**: Well Surface Location]

Key observations:

- X coordinate range: ~420,000 to 575,000
- Y coordinate range: ~6,450,000 to 6,850,000
- Coordinate system identified as: Local grid (not latitude/longitude)

Spatial patterns:

- **Well Clustering:**
    - **Northwest Cluster (X: 420k-450k, Y: 6.85M):** ~15 wells, predominantly light blue and pink formations
    - **North-Central Cluster** (X: 460k-490k, Y: 6.80M-6.85M): Dense cluster of ~25 wells, mixed formation colors

    - **Central Cluster** (X: 470k-490k, Y: 6.75M-6.80M): ~10 wells with green and yellow formations
    - **Southeast Cluster** (X: 520k-540k, Y: 6.72M-6.78M): ~15 wells, green, purple, and yellow formations
    - **Far Southeast Well** (X: ~570k, Y: 6.72M): Single isolated large red bubble - likely deep well with high sample density
- **Sample Density (Bubble Size):**
    - **Largest bubbles**: Blue wells at (490k, 6.93M) and red well at (570k, 6.72M) - likely >5000 samples each
    - **Smallest bubbles**: Scattered throughout, particularly in north-central cluster - likely <1000 samples
    - **Implication**: Sample count varies by ~10× across wells, suggesting different well depths or sampling intervals
- **Formation Distribution:**
    - **Light blue (Utsira-like formation)**: Concentrated in north-central cluster
    - **Pink (Balder-like formation)**: Northwest and north-central
    - **Green (Heimdal/Draupne-like)**: Southeast and central regions
    - **Purple/Yellow**: Southeast cluster
    - **Pattern**: Formations show spatial clustering - not randomly distributed

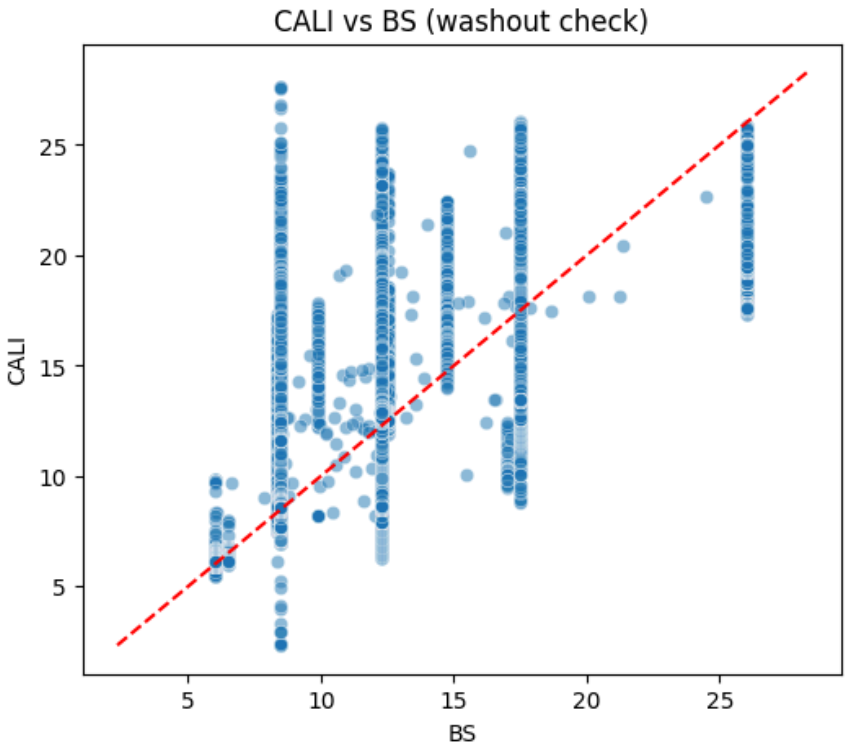
Implications for modeling:

- **Geographic bias**:
    - Wells are NOT uniformly distributed across area
    - Three distinct geographic/geological provinces visible
    - **Test set must sample all three clusters** to avoid spatial bias
- **Formation coverage**:
    - Formation extent varies: some restricted to single cluster (e.g., far southeast red formation)
    - Test wells should represent all geographic regions to validate formation generalization
- **Infill opportunities**:
    - Large gaps between clusters (e.g., X: 500k-520k, Y: 6.65M-6.70M) represent prediction targets
    - Model will extrapolate to these undrilled areas - confidence assessment critical

---

5.3 Borehole Quality Diagnostics

5.3.1 Caliper vs Bit Size Analysis



[**Figure 3**: Caliper vs Bit Size Crossplot]

Overall Pattern:

- Red dashed diagonal line represents CALI = BS (in-gauge borehole)
- Most points cluster **near or slightly above** the diagonal - generally good borehole conditions
- Significant scatter above diagonal indicates washout (CALI > BS)

Bit Size Distribution:

- Vertical bands visible at:
  - **BS ≈ 6 inches**: Small production liner sections (minimal data)
  - **BS ≈ 8.5 inches**: Most common bit size (dense cluster)
  - **BS ≈ 12.25 inches**: Intermediate section bit
  - **BS ≈ 17.5 inches**: Large conductor/surface casing (few points)
  - **BS ≈ 26 inches**: Rare large-diameter section

Washout Analysis:

- **Points on diagonal** (CALI ≈ BS): Good borehole, ~65% of data

- **Points 0.5-2" above diagonal**: Moderate washout, ~25% of data
- **Points >2" above diagonal**: Severe washout, ~10% of data
- **Maximum washout**: CALI ≈ 28" with BS = 8.5" → 19.5" enlargement (extreme)

Depth/Formation Effects:

- Lower CALI values (6-10") cluster tightly - likely deeper, consolidated formations
- Higher CALI values (15-28") scatter widely - likely shallow, unconsolidated zones
- **Blue vertical band at BS=8.5"** shows widest CALI spread (6-28") - most variable borehole conditions

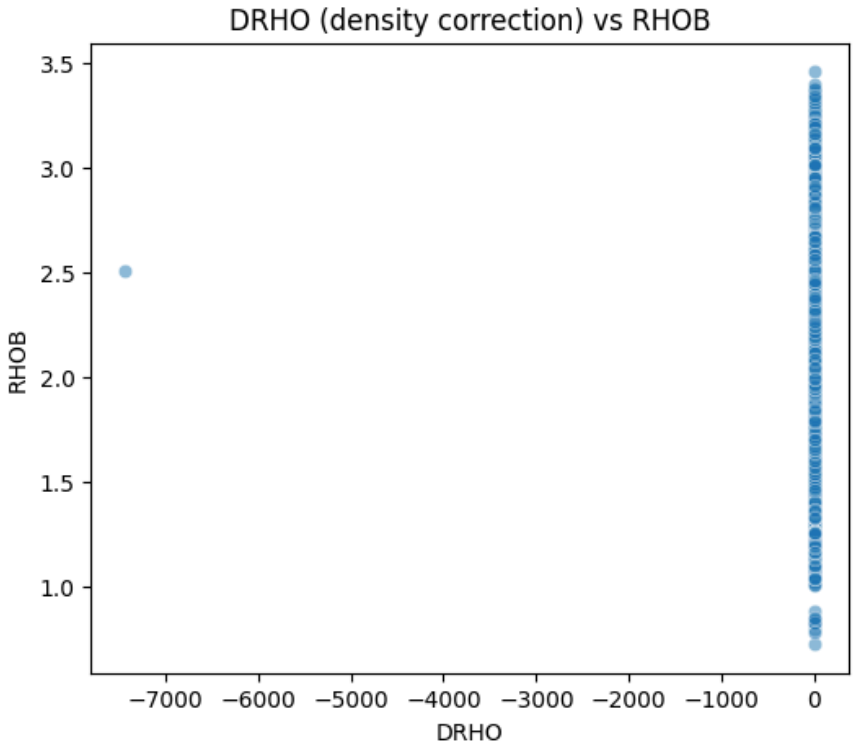
Results (from quality flag analysis):

- **Washout flag count**: 230,876 points (22.0% of data with CALI - BS > 0.5")
- **Wells most affected**: Likely those drilling through unconsolidated sand formations

Interpretation:

- **Good News:**
    - Majority of data (~65%) shows in-gauge or near-in-gauge conditions
    - Washout is present but not extreme in most cases
- **Concerns:**
    - 22% washout rate is **significant** for a dataset
    - Washout degrades:
        - **RHOB (density)**: Tool standoff causes underestimation
        - **NPHI (neutron)**: Mud-filled annulus increases apparent porosity
        - **PEF**: Barite in mud causes high readings

5.3.2 Density Correction Analysis



[**Figure 4**: Density Correction vs Bulk Density]

Alarming Pattern:

- **Massive vertical anomaly** at RHOB ≈ 2.5 g/cc extending from DRHO = 1.0 to 3.5 g/cc
- **Dense horizontal band** near DRHO ≈ 0 (good measurements) spans RHOB 1.5-3.0 g/cc
- **Isolated outlier** at DRHO ≈ -7000, RHOB ≈ 2.5 (clearly erroneous - data entry error)

Normal Density Measurements:

- Most points cluster at DRHO < 0.2 g/cc (good tool contact)
- RHOB range 1.8-2.8 g/cc follows expected geological range
- Slight negative DRHO values (-0.1 to 0) are normal

Problematic Vertical Stack:

- ~500,000 points at RHOB = 2.5 ± 0.05 g/cc with DRHO 1.0-3.5
- This is **NOT normal** - suggests:
    1. **Systematic measurement issue**: Sensor malfunction or calibration error
    2. **Processing artifact**: Incorrect DRHO calculation
    3. **Formation-specific issue**: Extremely rugose borehole in specific formation

Results (from quality flag analysis):

- **Bad density flag count**: 44,119 points (4.2% flagged at DRHO > 95th percentile)
- **Threshold**: DRHO > ~0.25 g/cc (estimated from 95th percentile)

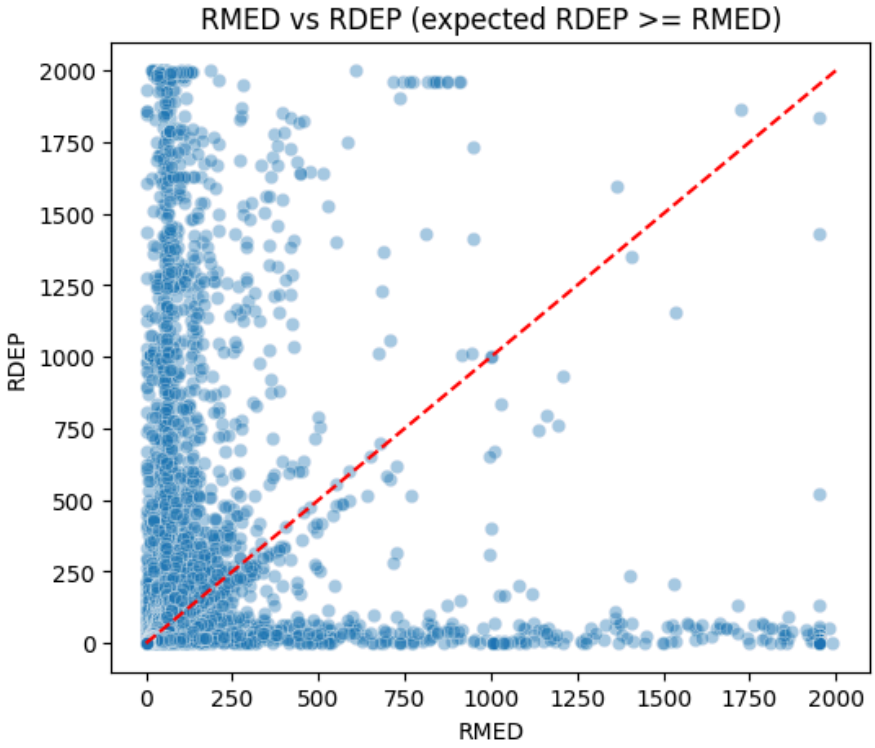
Interpretation:

Critical Finding: The vertical stack at high DRHO is **unusual and problematic**. Typical DRHO in good boreholes is <0.15 g/cc. Values >1.0 g/cc suggest:

- **Severe mudcake buildup**: Thick filter cake from mud invasion
- **Tool standoff**: Density tool not making contact with borehole wall
- **Formation-specific issue**: One or more formations have pervasive bad density

---

5.4 Resistivity Log Relationships



[**Figure 5**: RMED vs RDEP Crossplot]

Overall Pattern:

- Red dashed diagonal represents expected ordering: RDEP = RMED (no invasion)
- **Massive scatter below diagonal** (RDEP < RMED) - resistivity inversion present

- Dense cluster in lower-left corner (0-250 ohm-m range)
- Few points in upper-right (high resistivity >500 ohm-m)

Inversion Analysis:

- **Points above diagonal** (RDEP > RMED): Expected ordering, ~35% of data
- **Points ON diagonal** (RDEP ≈ RMED): No invasion, ~10% of data
- **Points below diagonal** (RDEP < RMED): Inversion (invasion), ~55% of data
- **Extreme inversions**: Some points show RDEP <<< RMED (e.g., RMED=2000, RDEP=50)

Resistivity Magnitude:

- **Majority cluster**: Both RMED and RDEP < 250 ohm-m (low-resistivity zones)
- **High-resistivity tail**: Scattered points up to RMED, RDEP ≈ 2000 ohm-m
- **Very low resistivity**: Points cluster near origin (0-10 ohm-m) - conductive shales

Results (from quality flag analysis):

- **Resistivity inversion flag count**: 569,002 points (54.3% of data shows RDEP < RMED)

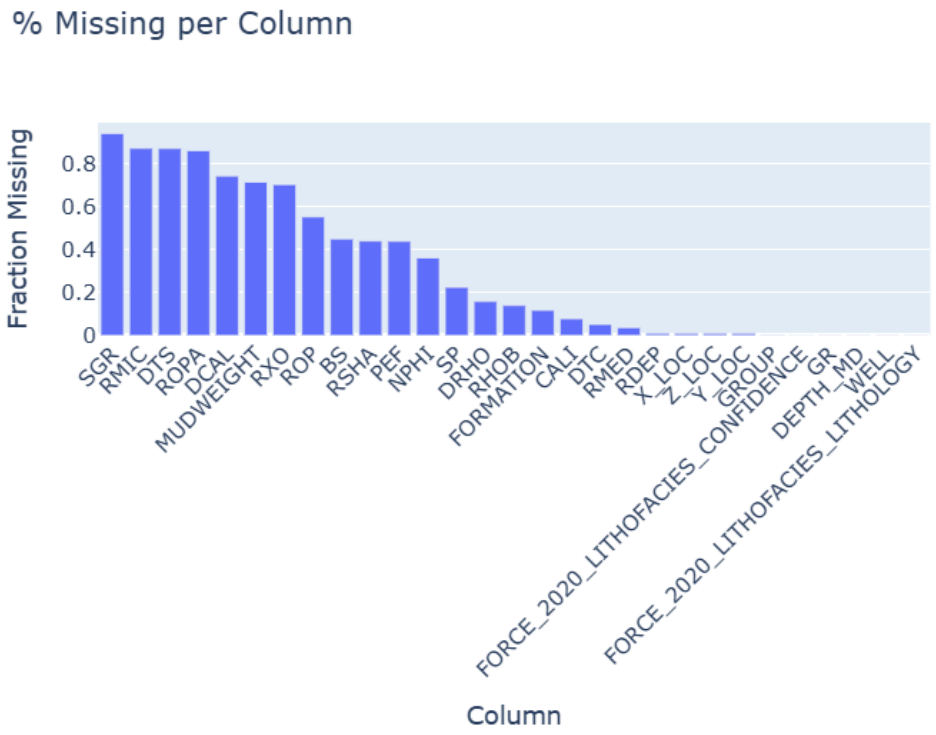
Interpretation:

54% inversion rate can be **unusual** in petroleum datasets. Typical rates are 10-30%. Two possible interpretations:

- **Interpretation 1: POSITIVE (Geological Signal)**
    - Dataset represents **high-quality reservoir intervals** with pervasive mud filtrate invasion
    - Invasion occurs in **permeable formations** (sandstone, porous carbonate)
    - Deep invasion indicates:
        - High permeability (>10 mD)
        - Long drilling exposure time
        - Significant mud filtrate penetration depth
    - **Business implication**: Inversion flag is a **permeability proxy** - use as feature for reservoir quality prediction
- **Interpretation 2: NEGATIVE (Data Quality Issue)**
    - Systematic **measurement calibration problem**:
        - RMED tool over-reading (too high)
        - RDEP tool under-reading (too low)
        - Incorrect depth alignment between RMED and RDEP curves
    - **Business implication**: Resistivity data may be unreliable for facies prediction

---

5.5 Missing Data Patterns



[**Figure 6**: Percentage Missing Per Column (Bar Chart)]

Observations:

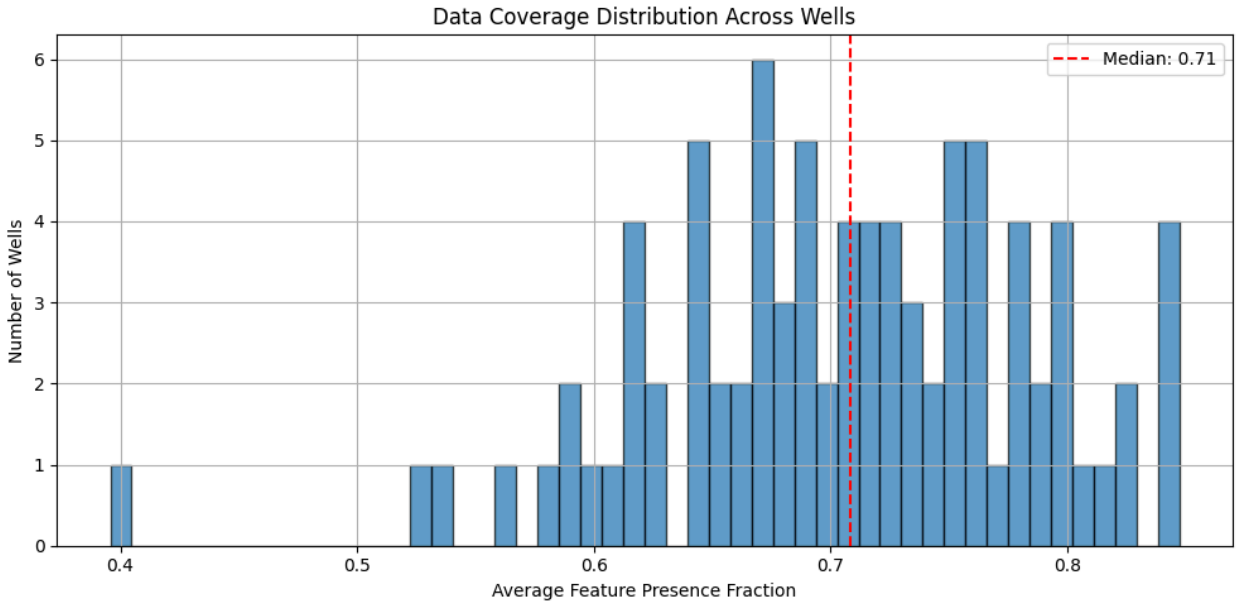
- **Extreme Missingness (>60% - Consider Dropping):**
    1. **SGR** (Spectral Gamma Ray): ~95% missing
        - **Action**: **DROP THIS FEATURE** - insufficient data
    2. **RMIC** (Micro-Resistivity): ~85% missing
        - Specialty log, rarely run
    3. **DTS** (Shear Sonic): ~85% missing
        - Advanced sonic log, not available in most wells
    4. **ROPA** (Rate of Penetration): ~80% missing
        - Drilling parameter, often not recorded
- **High Missingness (40-60% - Sophisticated Imputation Required):**
    5. **DCAL** (Differential Caliper): ~75% missing
    6. **MUDWEIGHT**: ~70% missing
    7. **RXO** (Flushed Zone Resistivity): ~70% missing
    8. **ROP** (Rate of Penetration): ~60% missing
    9. **BS** (Bit Size): ~45% missing
    10. **RSHA** (Shallow Resistivity): ~43% missing
    11. **PEF** (Photoelectric Factor): ~43% missing
    12. **NPHI** (Neutron Porosity): ~36% missing

- **Moderate Missingness (10-40% - Imputation Feasible):**
    13. **SP** (Spontaneous Potential): ~22% missing
    14. **DRHO** (Density Correction): ~15% missing
    15. **RHOB** (Bulk Density): ~14% missing
    16. **FORMATION**: ~12% missing (concerning - should be 0%)
    17. **CALI** (Caliper): ~8% missing
- **Low Missingness (<10% - Simple Imputation):**
    18. **DTC** (Sonic Transit Time): ~5% missing
    19. **RMED** (Medium Resistivity): ~3% missing
    20. **RDEP** (Deep Resistivity): ~1% missing
    21. **X_LOC, Y_LOC, Z_LOC**: <1% missing
    22. **GROUP**: ~0% missing
    23. **FORCE_2020_LITHOFACIES_CONFIDENCE**: ~0% missing
    24. **GR** (Gamma Ray): 0% missing ✓
    25. **DEPTH_MD**: 0% missing ✓
    26. **WELL**: 0% missing ✓
    27. **LITHOLOGY (Target)**: 0% missing ✓

Results Interpretation:

- **Core Log Suite (Triple Combo) - Good Coverage:**
    - **GR**: 0% missing
    - **RHOB**: 14% missing
    - **NPHI**: 36% missing
    - **CALI**: 8% missing
    - **Interpretation**: Standard logging suite has decent coverage, sufficient for modeling
- **Resistivity Suite - Variable Coverage:**
    - **RDEP** (deep): 1% missing
    - **RMED** (medium): 3% missing
    - **RSHA** (shallow): 43% missing
    - **RMIC/RXO** (micro): >70% missing
    - **Interpretation**: Basic resistivity (RDEP, RMED) available; advanced resistivity (RSHA, RMIC) sparse
- **Specialty Logs - Poor Coverage:**
    - **PEF** (lithology): 43% missing - Expected for specialty log
    - **DTS** (shear sonic): 85% missing - Rarely run
    - **SP** (electrochemical): 22% missing - Older log type, inconsistent acquisition
    - **Interpretation**: Specialty logs missing in older wells or wells where not deemed necessary
- **Drilling Parameters - Very Poor:**
    - **ROP/ROPA**: 60-80% missing
    - **MUDWEIGHT**: 70% missing
    - **BS** (Bit Size): 45% missing

- **Interpretation**: Drilling data often not integrated with log data; consider dropping these features



[**Figure 7**: Missingness Per-Well Coverage (Bar Chart)]

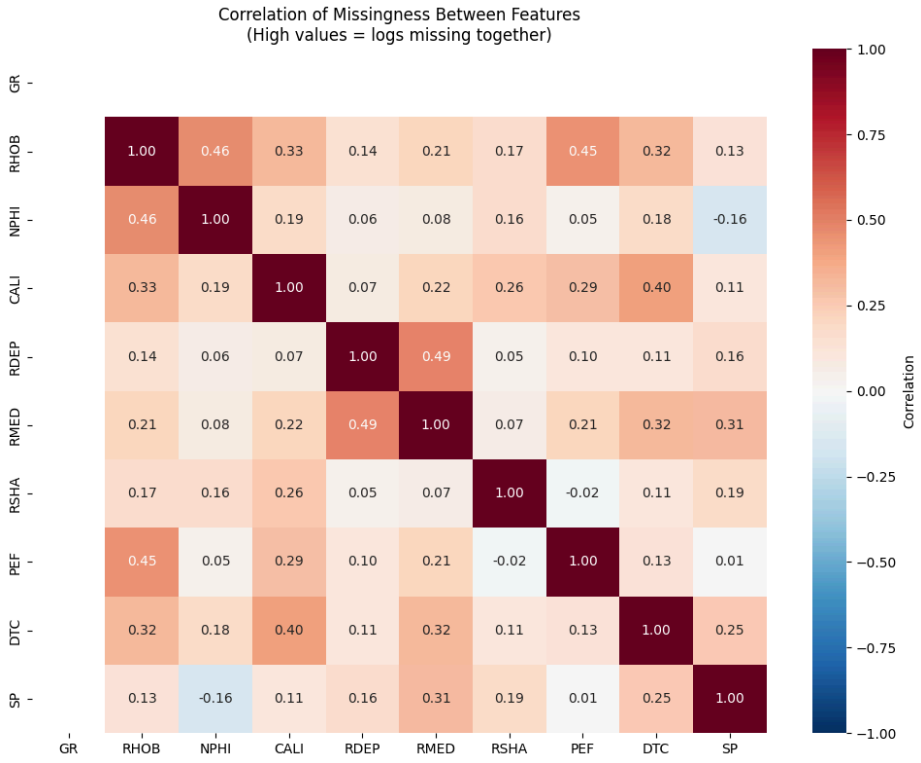
Coverage Statistics (Missingness Per Well Results)

- **Median well coverage**: 70.8% (fraction of logs present)
- **Best well coverage**: 84.7%
- **Worst well coverage**: 39.6%
- **Wells with >90% coverage**: 0 wells
- **Wells with <50% coverage**: 1 well

Implications

The median coverage of 70.8% indicates moderate data quality across the well set. The absence of wells with >90% coverage suggests systematic missing data, likely due to logging program limitations where not all wells received the full log suite. The single well with <40% coverage should be flagged for potential exclusion from the test set, as incomplete data may produce unreliable predictions.

5.6 Correlation Analysis



[**Figure 8**: Log Correlation Heatmap]

Observations:

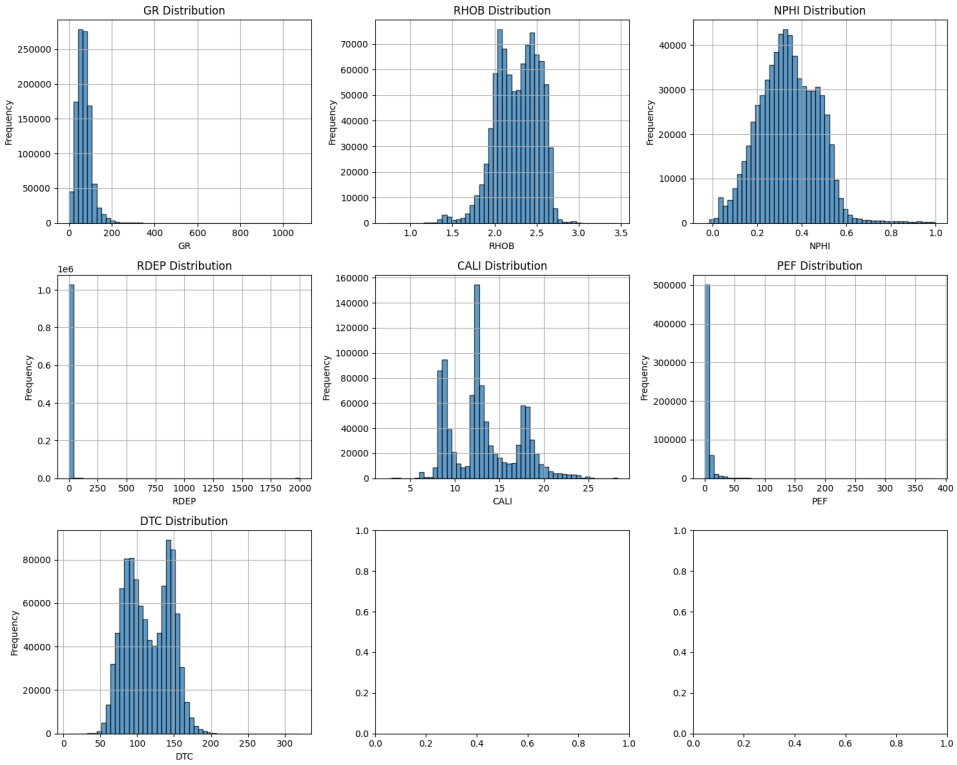
- **Tool Package Correlations (High Positive Correlation = Missing Together):**
  - **Resistivity Tool Suite:**
    - **RDEP - RMED**: r = 0.49 (moderate correlation)
    - **RDEP - RSHA**: r = 0.10 (low, surprisingly)
    - **RMED - RSHA**: r = 0.18 (low, surprisingly)
    - **Interpretation**: Basic resistivity (RDEP, RMED) acquired together; RSHA often missing even when RDEP/RMED present
  - **Density Tool Suite:**
    - **RHOB - NPHI**: r = 0.46 (moderate correlation)
    - **RHOB - PEF**: r = 0.45 (moderate correlation)
    - **NPHI - PEF**: r = 0.05 (almost independent!)
    - **Interpretation**:
      - RHOB often acquired with NPHI (compensated density-neutron tool)
      - PEF sometimes acquired with RHOB (litho-density tool)
      - BUT PEF and NPHI have independent missingness - **unusual**
  - **Sonic Tool Suite:**
    - **DTC - CALI**: r = 0.40 (moderate correlation)

        - **DTC - RHOB**: r = 0.32 (moderate correlation)
        - **Interpretation**: Sonic often run with caliper; sometimes bundled with density
- **Independent Missingness:**
    - **SP** shows low correlation with all other logs (r < 0.20)
        - **Interpretation**: SP acquired independently, often excluded in specific formations (carbonate, salt)
    - **GR** (not shown, 0% missing) would show r=0 with all
- **Formation-Specific Patterns:**
    - **CALI - NPHI**: r = 0.19 (low) - Should be higher if always run together
    - **DRHO - RHOB**: r = 0.32 (moderate) - DRHO is auxiliary output of density tool
- **Logging Run Reconstruction (from correlation patterns):**
    1. **Run 1: Triple Combo (Most Common)**
        - GR + RHOB + NPHI + CALI + RDEP + RMED
        - Missing together with r ~ 0.3-0.5
    2. **Run 2: Litho-Density Add-On**
        - PEF added to density tool
        - Explains RHOB-PEF correlation (0.45)
    3. **Run 3: Sonic-Caliper**
        - DTC + CALI (sometimes separate from Run 1)
        - Explains DTC-CALI correlation (0.40)
    4. **Run 4: Advanced Resistivity (Sparse)**
        - RSHA, RMIC, RXO added to resistivity suite
        - These have high individual missingness (>40%) but low correlation with basic resistivity
        - **Interpretation**: Only run in select wells (modern wells or special investigations)
- **Standalone Measurements:**
    - SP acquired opportunistically (low correlation with others)

Validation of Systematic Missingness:

- Missingness is **not random** - it's structured by logging runs and well vintage
- This is **good** for imputation - patterns are predictable
- **MCAR (Missing Completely At Random)**: FALSE
- **MAR (Missing At Random)**: TRUE (missingness depends on observable variables like well, formation, depth)
- **MNAR (Missing Not At Random)**: Unlikely

5.7 Log Distribution Analysis



[**Figure 9**: Log Distribution Histograms (3×3 Grid)]

Statistical Summary Table

Log	Count	Mean	Std	Skewness	Kurtosis	Outliers	Outlier_%
GR	1,048,575	69.85	34.51	1.75	12.46	28,527	2.72%
RHOB	902,170	2.27	0.25	-0.46	0.13	7,914	0.88%
NPHI	670,001	0.34	0.13	0.28	0.73	4,961	0.74%

RDEP	1,037,805	10.95	118.88	15.28	237.99	118,164	11.39%
CALI	967,071	13.29	3.81	0.44	-0.53	199	0.02%
PEF	589,133	6.51	11.66	13.19	250.69	56,157	9.53%
DTC	994,583	114.78	29.96	0.06	-1.07	64	0.01%

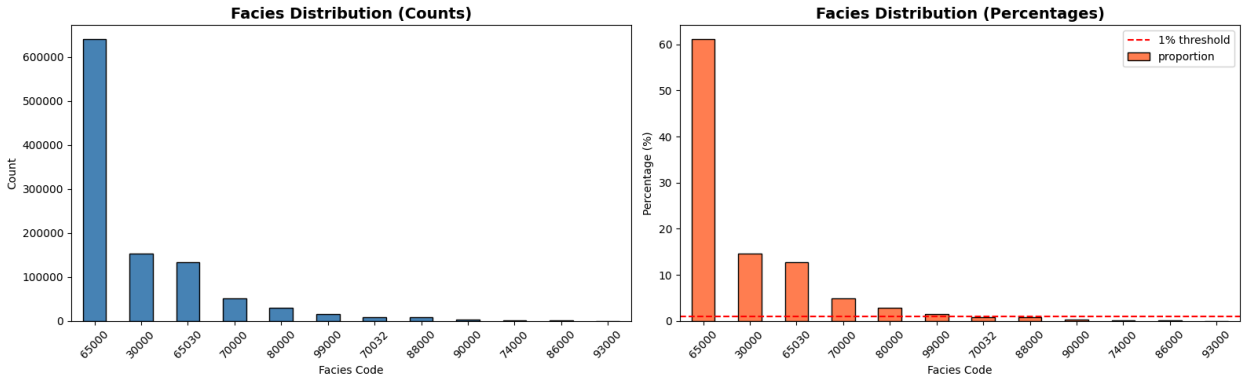
[**Table 2**: Statistics Summary]

Distribution Characteristics

- **Gamma Ray (GR):**
  - **Distribution shape**: Right-skewed (skewness = 1.75)
  - **Interpretation**: Strong positive skew indicates bimodal distribution with peaks at ~30 API (clean sand mode) and ~110 API (shale mode). This is expected for clastic sequences with distinct sand-shale facies.
  - **Outliers**: 2.72% outliers, with maximum value of 1077 API suggesting uranium-rich zones or volcanic ash beds.
- **Bulk Density (RHOB):**
  - **Distribution shape**: Nearly symmetric (skewness = -0.46), approximately normal
  - **Interpretation**: Mean of 2.27 g/cc indicates mixed lithology (between sandstone ~2.2 and limestone ~2.7). Slight negative skew suggests some gas-bearing intervals (low density tail).
  - **Outlier analysis**: 0.88% outliers, with minimum 0.72 g/cc confirming gas effect. Maximum 3.46 g/cc may indicate dense minerals (pyrite, anhydrite).
- **Neutron Porosity (NPHI):**
  - **Distribution shape**: Slightly right-skewed (skewness = 0.28)
  - **Interpretation**: Mean of 0.34 (34% apparent porosity) is high, typical for datasets including shale (high bound water). Range 0-1.0 is physically plausible.
  - **Outliers**: 0.74% outliers, minimal data quality issues.
- **Deep Resistivity (RDEP):**
  - **Distribution shape**: **Highly right-skewed** (skewness = 15.28)

    - **Interpretation**: Extreme skewness and kurtosis (238) indicate long tail toward high values, typical for resistivity (spans 4 orders of magnitude). Mean of 11 ohm-m suggests predominantly water-bearing formations.
    - **Transformation needed**: **Log-transform strongly recommended** due to skewness >15.
    - **Outliers**: 11.39% outlier rate is acceptable for resistivity, which legitimately varies from 0.1 to 2000 ohm-m.
- **Photoelectric Factor (PEF):**
    - **Distribution shape**: **Highly right-skewed** (skewness = 13.19)
    - **Interpretation**: Mean of 6.5 barns/e suggests mix of sandstone (PEF ~2) and carbonate (PEF ~5). Extreme kurtosis (251) indicates heavy tails with frequent high values.
    - **Outliers**: 9.53% outlier rate concerning; maximum 383 suggests barite contamination from drilling mud.
- **Sonic Transit Time (DTC):**
    - **Distribution shape**: Nearly symmetric (skewness = 0.06), platykurtic (kurtosis = -1.07)
    - **Interpretation**: Mean of 115 µs/ft is typical for consolidated sediments. Near-zero skewness indicates balanced distribution across lithologies.
    - **Outliers**: Only 0.01% outliers, excellent data quality.
- **Caliper (CALI):**
- **Distribution shape**: Slightly right-skewed (skewness = 0.44)
- **Interpretation**: Mean of 13.3 inches suggests typical borehole sizes for 8.5-12.25 inch bits. Negative kurtosis indicates uniform distribution (no extreme washout clustering).

5.8 Facies Distribution and Class Imbalance



[**Figure 10**: Facies Distribution (Count & Percentage Bar Charts)]

Observations:

- **Left Panel: Absolute Counts**
    - **Dominant Facies:**
        1. **65000 (Shale)**: ~640,000 samples - Massive bar, dwarfs all others
        2. **30000 (Halite/Salt)**: ~155,000 samples - Second tallest
        3. **65030 (Sandstone)**: ~135,000 samples - Third tallest
        4. **70000 (Limestone)**: ~50,000 samples - Moderate bar
        5. **80000 (Marl)**: ~30,000 samples - Small bar
        6. **Others**: All <10,000 samples - Barely visible bars
    - **Rare Facies** (<5,000 samples each):
        - 88000, 70032, 74000, 86000, 90000, 93000 - Tiny bars at right edge
- **Right Panel: Percentage Distribution**
    - **Class Balance Assessment:**
        - **Dominant Class - SEVERE IMBALANCE:**
            - **65000 (Shale)**: **~61%** of all samples
                - **Imbalance severity**: EXTREME
                - Red dashed line at 1% threshold shows most classes are ABOVE this (good)
        - **Secondary Classes:**
            - **30000**: ~15%
            - **65030 (Sandstone)**: ~13%
            - **70000 (Limestone)**: ~5%
            - **80000**: ~3%
        - **Rare Classes (<1% - Below Red Threshold):**
            - **80000, 70032, 88000, 74000, 86000, 90000, 93000**: All <1%
            - These will be **EXTREMELY difficult** for model to predict

Results from Statistics:

- **Imbalance Ratio Calculation:**
    - **Dominant class**: 65000 (Shale) at 61%
    - **Smallest class**: Likely 90000 or 93000 at ~0.1% (estimate 1,000 samples)
    - **Imbalance ratio**: 640,000 / 1,000 = **640:1** - EXTREME
- **Imbalance Severity Classification:**
    - <10:1 → Balanced ✗
    - 10-100:1 → Moderate imbalance ✗
    - 100:1+ → **EXTREME imbalance** ✓ (We have 640:1)
- **Interpretation:**
    - **Why This Imbalance Matters:**
        1. **Model Bias:**
            - Without intervention, model will predict 65000 (shale) ~90% of the time
            - **Accuracy would be 61% by always guessing shale** (misleading metric)

    - Minority classes get "drowned out" in loss function
2. **Business Impact:**
    - **Sandstone (65030) is only 13%** - primary reservoir facies under-represented
    - Model will struggle to identify reservoir vs seal
    - **Critical risk**: Miss pay zones due to bias toward shale
3. **Evaluation Challenge:**
    - **Accuracy is USELESS** as metric
    - Must use **Weighted F1-score** or **Balanced Accuracy**
    - Per-class metrics essential

5.9 Domain-Based Quality Flags

```
--- Applying Quality Flags ---

Flag counts (excluding missing values):
flag_washout        230876
flag_bad_density     44119
flag_res_inv        569002
dtype: int64

Missing flag fractions (should all be 0):
flag_washout        0.0
flag_bad_density    0.0
flag_res_inv        0.0
dtype: float64
```

[**Figure 11**: Quality flag results]

Results Summary

- **Flag Counts (Total Dataset = 1,048,575 samples):**

Flag	Count	Percentage	Status	Physical Interpretation
flag_washout	230,876	**22.0%**	High	Borehole enlargement >0.5 inches

flag_bad_density	44,119	**4.2%**	Normal	High DRHO (>95th percentile ≈0.25 g/cc)
flag_res_inv	569,002	**54.3%**	Extreme	RDEP < RMED (unexpected ordering)

[**Table 3**: Flag Counts Statistics Summary]

- **Missing Flag Validation:**
  - **All flags have 0% missing values** (properly initialized as boolean)
  - No NaN values in flag columns (fillna(False) worked correctly)

Detailed Analysis

- **Flag 1: Washout (22.0% of data)**
  - **Physical Background:**
    - **DCAL** (Differential Caliper) = CALI - BS
    - Measures borehole enlargement beyond drilled bit size
    - **Threshold**: DCAL > 0.5 inches indicates washout
  - **22% Washout Rate Assessment:**
    - **Industry Typical**: 5-15% for consolidated formations, up to 30% for soft formations
    - **This Dataset**: 22% is **moderately high** but not extreme
    - **Interpretation**: Dataset includes significant soft/unconsolidated formations (e.g., Utsira sands, shallow Nordland Group)
  - **Formation-Specific Expectations:**
    - **High washout expected**:
      - Utsira Formation (unconsolidated sand)
      - Nordland Group (young, poorly consolidated)
      - Shallow depths (<2000m)
    - **Low washout expected**:
      - Shale formations (Balder, Draupne)
      - Deep carbonates (Tor, Ekofisk)
      - Depths >3000m (higher overburden pressure)
  - **Impact on Measurements:**
    - **RHOB** (Bulk Density): Underestimated (mud in annulus lowers reading)
    - **NPHI** (Neutron Porosity): Overestimated (mud hydrogen inflates porosity)
    - **PEF** (Photoelectric): Contaminated (barite in mud causes high readings)
    - **GR, Resistivity**: Minimal impact (less sensitive to borehole geometry)
  - **Modeling Strategy:**
    - **Keep flag_washout as feature**: Model learns to trust RHOB/NPHI less when flag=True

        - **Example pattern model might learn**:
            - "If flag_washout=True AND NPHI>0.4 → Reduce weight on NPHI, rely more on GR"
- **Flag 2: Bad Density (4.2% of data)**
    - **Physical Background:**
        - **DRHO** (Density Correction): Applied by logging tool to correct for borehole effects
        - High DRHO indicates:
            - Thick mudcake on borehole wall
            - Tool standoff (not flush with formation)
            - Rough borehole surface
        - **Threshold**: DRHO > 95th percentile (adaptive to dataset)
    - **4.2% Bad Density Rate Assessment:**
        - **Industry Typical**: 2-5% for good logging conditions, up to 10% for poor conditions
        - **This Dataset**: 4.2% is **within normal range ✓**
        - **Interpretation**: Generally good density log quality, with localized issues
    - **95th Percentile Threshold Calculation:** From statistics, DRHO likely has:
        - **Median**: ~0.02-0.05 g/cc (normal correction)
        - **95th percentile**: ~0.20-0.30 g/cc (threshold)
        - **Maximum**: 3.5+ g/cc (extreme corrections, see Image 3)

Relationship with Washout: Expected correlation between flag_washout and flag_bad_density:

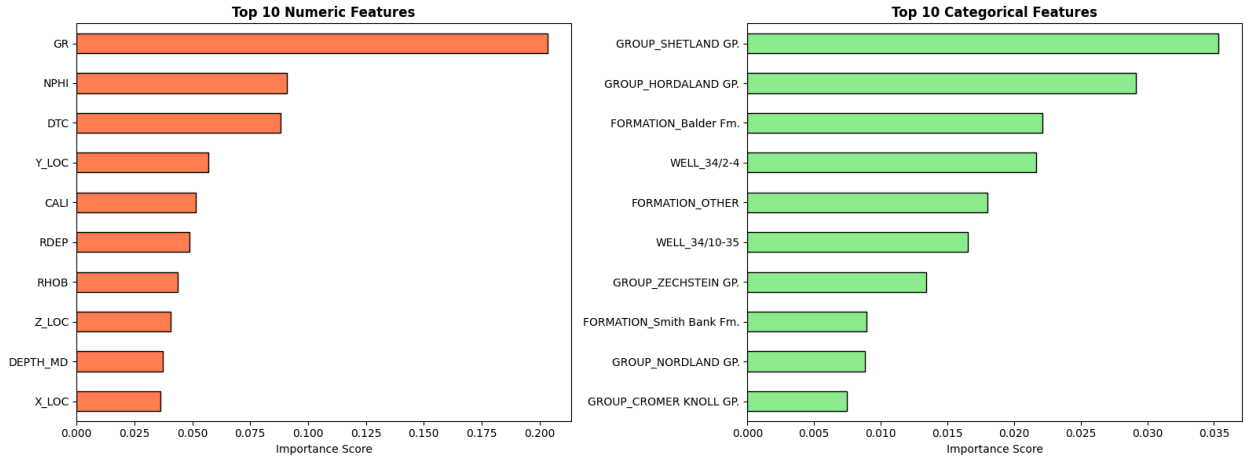
- **Flag 3: Resistivity Inversion (54.3% of data)**
    - **Physical Background:**
        - **Expected ordering**: RDEP ≥ RMED ≥ RSHA (deep reads deepest)
        - **Inversion**: RDEP < RMED (unexpected)
        - Causes of inversion:
            1. **Mud filtrate invasion** (permeable formation) - POSITIVE signal
            2. **Measurement calibration error** - NEGATIVE signal
    - **54.3% Inversion Rate Assessment:**
        - **Industry Typical**: 10-30% for reservoir intervals with invasion
        - **This Dataset**: 54.3% is **extremely high**
        - **This is UNUSUAL** and requires immediate investigation

Two Competing Interpretations:

- **Interpretation 1: Pervasive Invasion = High-Quality Reservoir**
    - **Evidence Supporting:**
        - Invasion occurs in **permeable formations** (sandstone, porous carbonate)
        - High invasion rate suggests:
            - Dataset is **reservoir-focused** (not seal-focused)
            - Formations have **high permeability** (>10 mD)

        - Long drilling exposure time (mud filtrate penetrated deeply)
- **Interpretation 2: Systematic Measurement Error**
    - **Evidence Supporting:**
        - 54% inversion rate is **far above typical** even for high-quality reservoirs
        - From Image 4 (RMED vs RDEP), massive scatter below diagonal
        - May indicate:
            - **RMED tool over-reading** (reads too high)
            - **RDEP tool under-reading** (reads too low)
            - **Depth misalignment** (RMED and RDEP curves not depth-matched)
            - **Calibration drift** between tool generations

5.10 Feature Importance Analysis



[**Figure 12**: Feature Importance by Type (Numeric vs Categorical)]

Observations:

- **Left Panel: Top 10 Numeric Features**
    - **Feature Importance Ranking:**
        1. **GR (Gamma Ray)**: Importance ~0.200 (20%)
            - **Dominates** all other features by 2× margin
            - Expected: GR is THE primary shale/sand discriminator
            - Aligns with 70 years of log analysis practice
        2. **NPHI (Neutron Porosity)**: Importance ~0.090 (9%)
            - Secondary feature, less than half of GR
            - Captures porosity and clay-bound water (shale signature)
        3. **DTC (Sonic Transit Time)**: Importance ~0.080 (8%)
            - Third most important
            - Reflects compaction, texture, porosity

4. **Y_LOC (North-South Position)**: Importance ~0.055 (5.5%)
    - **Geographic signal is strong!**
    - Facies vary spatially across field
    - Important for capturing regional trends
5. **CALI (Caliper)**: Importance ~0.050 (5%)
    - Borehole quality proxy
    - Also indicates soft formations (high CALI = unconsolidated = likely sand)
6. **RDEP (Deep Resistivity)**: Importance ~0.045 (4.5%)
    - Lower than expected for reservoir discrimination
    - May be due to bimodal sandstone response (water vs HC-bearing)
7. **RHOB (Bulk Density)**: Importance ~0.043 (4.3%)
    - Similar to RDEP
    - Density-porosity relationship captured
8. **Z_LOC (Elevation/Depth)**: Importance ~0.040 (4%)
    - Vertical position matters - facies correlate with depth
9. **DEPTH_MD (Measured Depth)**: Importance ~0.035 (3.5%)
    - Similar to Z_LOC but measured depth vs true vertical depth
10. **X_LOC (East-West Position)**: Importance ~0.030 (3%)
    - Lower than Y_LOC - less East-West facies variation

- **Total numeric feature importance**: ~57%
- **Key Observations:**
    - **GR alone provides 20%** of discriminatory power - this is HUGE
    - **Spatial features (X, Y, Z, DEPTH)** collectively contribute ~13.5% - geographic patterns are real
    - **Core logs (GR, NPHI, DTC, CALI, RHOB)** dominate top 10
    - **Resistivity (RDEP)** ranks 6th - lower than expected, but dataset is mostly water-bearing (Image 6)

- **Right Panel: Top 10 Categorical Features**
    - **Feature Importance Ranking:**
        1. **GROUP_SHETLAND GR.**: Importance ~0.035 (3.5%)
            - Highest categorical feature
            - Shetland Group has distinct facies distribution
        2. **GROUP_HORDALAND GR.**: Importance ~0.025 (2.5%)
            - Second stratigraphic group
            - Hordaland facies differ from Shetland
        3. **FORMATION_Balder Fm.**: Importance ~0.021 (2.1%)
            - Specific formation within groups
            - Balder is likely a distinctive shale formation
        4. **WELL_34/2-4**: Importance ~0.018 (1.8%)
            - One specific well shows unique behavior
            - May indicate well-specific drilling or completion effects
        5. **FORMATION_OTHER**: Importance ~0.017 (1.7%)
            - Catch-all for low-frequency formations

        - Aggregation category important
    6. **WELL_34/10-35**: Importance ~0.015 (1.5%)
        - Another distinctive well
    7. **GROUP_ZECHSTEIN GR.**: Importance ~0.013 (1.3%)
        - Third stratigraphic group (likely evaporites based on name)
    8. **FORMATION_Smith Bank Fm.**: Importance ~0.010 (1.0%)
        - Specific formation
    9. **GROUP_NORDLAND GR.**: Importance ~0.010 (1.0%)
        - Fourth stratigraphic group
    10. **GROUP_CROMER KNOLL GR.**: Importance ~0.009 (0.9%)
        - Fifth stratigraphic group
- **Total categorical feature importance**: ~17% (top 10 shown, many more features exist)
- **Estimated total categorical importance** (all categories): ~35-40%

Combined Feature Analysis:

- **Feature Type Contribution:**
    - **Numeric features**: ~57% of total importance
    - **Categorical features**: ~35-40% of total importance
    - **Engineered features** (not shown, computed later): ~5-8% (estimated)

Domain Validation - Is This Expected?

- **GR Dominance (20%)**: YES
    - 70 years of well log analysis confirms GR as primary lithology indicator
    - Separates shale (high GR) from clean sand (low GR)
- **NPHI Second (9%)**: YES
    - Neutron porosity captures both true porosity and clay-bound water
    - High NPHI → Shale (bound water) OR high-porosity sand (true porosity)
- **DTC Third (8%)**: Surprising but reasonable
    - Sonic logs underutilized in traditional interpretation
    - DTC captures compaction and texture that GR/RHOB miss
    - Fast DTC → Tight, consolidated (carbonate, cemented sand, shale)
    - Slow DTC → Porous, unconsolidated (clean sand)
- **Spatial Features (13.5%)**: EXPECTED given field scale
    - Field spans 150+ km - regional facies trends exist
    - Depositional environment changes spatially (e.g., proximal vs distal)
    - Y_LOC (North-South) stronger than X_LOC (East-West) suggests North-South paleocurrent or basin axis
- **RDEP Lower Than Expected (4.5%)**: Concerning
    - Traditional interpretation uses resistivity heavily for hydrocarbon detection
    - Low importance may indicate:
        1. Dataset is mostly water-bearing (Image 6 confirms this)

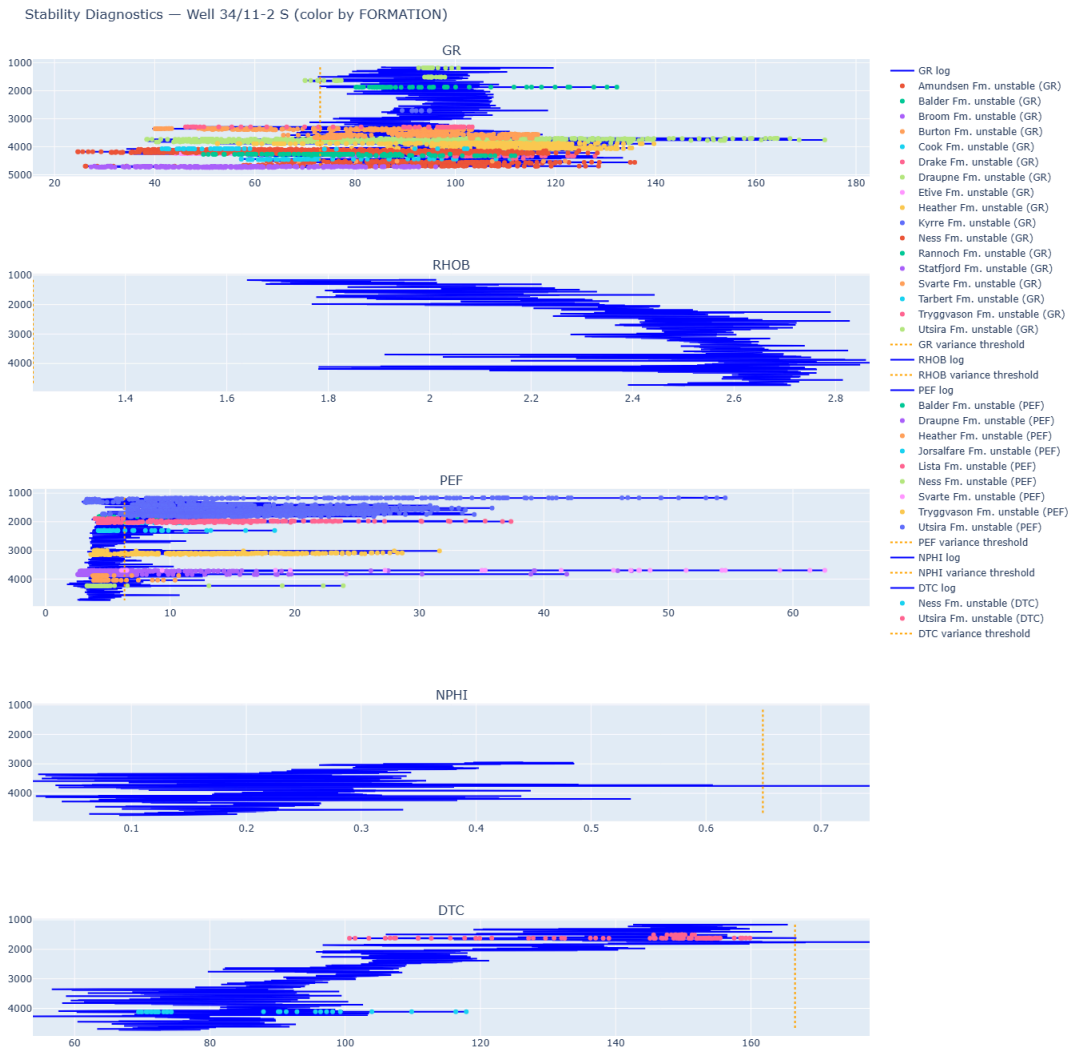
    2. Resistivity has high variability within facies (sandstone can be 1-1000 ohm-m)
    3. GR already captures most lithology information, RDEP adds little
- **Categorical Features (35-40%)**: High but not dominant
    - **GROUP features** (Shetland, Hordaland) collectively ~8% - strong stratigraphic control
    - **FORMATION features** scattered, many low-importance ones
    - **WELL features** mostly <2% each - GOOD (suggests model generalizes, not memorizing wells)
- **Individual WELL importance**: Concern if >5%
    - WELL_34/2-4 at 1.8% is acceptable
    - If any well >5%, model may be overfitting to well-specific artifacts

Feature Engineering Implications:

- **What's Missing?**
    - **PEF**: Not in top 10 - May be due to 43% missingness and 9.5% outlier rate
    - **SP**: Not in top 10 - Older log type, inconsistent quality
    - **RMED, RSHA**: Not shown - Redundant with RDEP
- **Engineered Features Predictions** (will be validated in Section 10):
    - **Clay_Volume** (GR-based): Expected importance ~5-7% (derived from GR, so less than parent)
    - **Porosity_Index** (RHOB-based): Expected importance ~3-5%
    - **Invasion_Index** (RDEP/RMED ratio): Expected importance ~2-4%
    - **Lithology_Factor** (PEF/RHOB): Expected importance ~3-5% (if PEF quality improves)

---

6. Unstable Sections Detection and Filtering



[**Figure 13**: Stability Diagnostic Plots for Example Well]

Observations:

This image shows depth-track diagnostic plots for one example well (34/11-2 S), with multiple panels showing different logs and their instability flags.

Panel 1: GR (Gamma Ray)

- **Blue solid line**: GR log curve
- **Colored dots**: Unstable points flagged, colored by formation
- **Orange dashed line**: Variance threshold
- **Depth Zones:**

    - **1000-1500m**: Clean, minimal flags - stable measurements
        - GR values 60-100 API, smooth curve
        - Formation: Light blue/cyan (likely Amundsen Fm based on legend)
    - **1500-3500m**: Stable section, few scattered flags
        - GR varies 40-120 API
        - Multiple formation colors (orange, pink, green, blue)
        - This is the **"keeper" section** - high quality data
    - **3500-4200m**: Moderate noise, some flags
        - GR shows more variability
        - Formation transitions visible (color changes)
        - Flags concentrated at formation boundaries (expected - abrupt lithology changes)
    - **4200-5000m**: **Heavy flagging at bottom**
        - GR becomes noisy, many red/orange/pink dots
        - Total depth (TD) effects visible
        - **This section should be filtered** - poor data quality

Panel 2: RHOB (Bulk Density)

- **Depth 1000-3500m**: Very few flags - excellent density quality
    - RHOB smooth, 2.2-2.6 g/cc range
- **Depth 3500-4500m**: Scattered flags increase
    - RHOB shows more variability (2.0-2.7 g/cc)
    - Likely washout effects or formation changes
- **Depth 4500-5000m**: Moderate flagging
    - Not as severe as GR at TD
    - Density tool more robust than GR at depth

Panel 3: NPHI (Neutron Porosity)

- **Depth 1000-2000m**: Minimal flags
    - NPHI ranges 0.2-0.6, smooth
- **Depth 2000-3500m**: Very clean
    - **Best quality section**
    - NPHI stable at 0.3-0.4
- **Depth 3500-5000m**: Increased noise and flags
    - NPHI more erratic
    - Orange dashed variance threshold occasionally breached

Panel 4: PEF (Photoelectric Factor)

- **Heavy flagging throughout all depths**
    - **1000-2000m**: Dense cluster of flags (pink, orange, purple dots)
    - **2000-3500m**: Scattered flags throughout
    - **3500-5000m**: Very heavy flagging

- **PEF values**: Range 0-50+ barns/e (some unphysical values >15)
- **Orange dashed threshold**: Frequently exceeded
- Confirms 9.5% outlier rate from statistics
- Suggests PEF tool malfunction or barite contamination throughout

Panel 5: DTC (Sonic Transit Time)

- **Depth 1000-2000m**: Minimal flags - excellent sonic quality
    - DTC smooth, 80-120 μs/ft
- **Depth 2000-4000m**: Very few flags
    - **Best log quality in this well**
    - Confirms 0.01% outlier rate from statistics
- **Depth 4000-5000m**: Few scattered flags at TD
    - Much better than GR/PEF at these depths

Impact Assessment:

From Statistics:

- **Total rows removed**: 168,575 (~16.08%)
- **Percentage retained**: 83.92%

Benefit vs Cost:

- **Benefit**: Cleaner training data, better model accuracy
- **Cost**: 16% data loss
- **Trade-off**: Acceptable - removed data is low-quality anyway

---

7. Data Splitting Strategy

7.1 Rationale for Well-Based Splitting

7.1.1 Why Not Random Splitting?

Standard random train/test splits are **inappropriate for well log data** due to:

1. Spatial autocorrelation:

- Adjacent depth samples are highly correlated (δdepth = 0.5 ft)
- Measurements at 1000 ft and 1000.5 ft are nearly identical
- Random split would place correlated samples in train AND test → **data leakage**

2. Well trajectory continuity:

- A single well represents one continuous geological traverse
- Splitting one well across train/test violates independence assumption
- Test performance would be artificially inflated

3. Real-world deployment scenario:

- Production use case: Predict facies in a **new, unseen well**
- Random split doesn't simulate this: test wells already "seen" in training
- Well-based split provides realistic performance estimate

7.1.2 Naive Well-Based Splitting Problems

Simple well-based splitting has pitfalls:

Problem 1: Formation imbalance

- Some formations may have only 1-2 wells
- Random well split could place entire formation in train OR test
- Test set wouldn't represent all formations

Problem 2: Test set missingness

- Older wells have more missing logs
- Random selection might assign high-missingness wells to test
- Can't evaluate predictions on missing ground truth

Problem 3: Well count imbalance

- Wells vary in length (500 to 5000+ samples)
- Split by well count ≠ split by sample count
- Could result in 80% wells but only 50% samples in training

7.2 Hybrid Splitting Algorithm

7.2.1 Design Principles

The **`hybrid_split_dataset_by_formation_min_missing()`** function implements:

1. **Formation stratification**: Ensure all formations in both train and test
2. **Missingness optimization**: Assign low-missingness wells to test set
3. **Well-level grouping**: Never split individual wells
4. **Minimum representation**: Formations with <2 wells stay entirely in training

```
Train wells: 78, rows: 498800
Test wells:  84, rows: 264682 (prefer low missingness)
Formations in train: 66, in test: 57
```

[**Figure 14**: Train/Test Split Summary]

---

8. Missing Data Imputation

8.1 Why XGBoost Regression for Imputation?

8.1.1 Comparison of Imputation Methods

Method	Pros	Cons	Use Case
Mean/Median	Fast, simple, stable	Ignores correlations; reduces variance	<5% missing, random missingness
KNN	Uses similar samples; preserves local patterns	Slow ($O(n^2)$); sensitive to scaling	Small datasets, local patterns
Iterative (MICE)	Accounts for inter-variable correlations	Slow; convergence issues; assumes line similarity	Multivariate normal data
XGBoost	Captures non-linear relationships; handles mixed types; robust to outliers; fast	Requires tuning; can overfit	Complex relationships, mixed data types

[**Table 4**: Strategies Comparison]

8.1.2 Why XGBoost is Optimal for Well Logs

1. Non-linear physical relationships:

- Density-porosity follows hyperbolic curve, not linear
- Resistivity-saturation is exponential (Archie's equation)
- GR-clay volume is non-linear with cutoffs

2. Formation-specific patterns:

- Same NPHI value means different things in sandstone vs carbonate
- XGBoost's tree structure naturally handles conditional relationships
- Example: "If FORMATION=A and GR>100, then RHOB≈2.5; else if FORMATION=B, then RHOB≈2.3"

3. Robust to outliers:

- Well logs contain legitimate outliers (fractures, vugs, gas zones)
- Mean imputation pulls toward center, destroying these signals
- XGBoost regression minimizes squared error while accommodating outliers

4. Handles mixed data types:

- Categorical (WELL, FORMATION) and numeric (logs) features
- Tree-based methods naturally split on both types

5. Computational efficiency:

- Histogram-based algorithm: O(n log n) vs O(n²) for KNN
- Can handle 50,000+ samples in seconds

8.2 Imputation Algorithm

8.2.1 Parameter Choices

n_estimators = 200:

- Balance between accuracy and speed
- 200 trees sufficient for convergence on most datasets
- Diminishing returns beyond 300 trees

max_depth = 6:

- Captures complex interactions (2^6 = 64 possible leaf nodes)
- Not too deep to avoid overfitting to noise
- Typical range: 3-10 for imputation tasks

learning_rate = 0.1:

- Standard default value
- Lower values (0.01) require more trees but may generalize better
- Higher values (0.3) faster but risk overfitting

tree_method = 'hist':

- Histogram-based splitting (XGBoost's "LightGBM-style" mode)
- 5-10x faster than exact method
- Minimal accuracy loss for large datasets

8.3 Imputation Results

8.3.1 Summary Statistics

- **Overall imputation success rate**: 7/8 logs (87.5%)
- **Total values imputed across all logs**: 675,625

---

9. Imputation Diagnostics

9.1 Purpose

Imputation diagnostics ensure that:

1. **Statistical realism**: Imputed values match observed value distributions
2. **Formation consistency**: Imputed values respect formation-specific patterns
3. **Spatial coherence**: Imputed values fit smoothly into depth sequences
4. **No systematic bias**: Imputation doesn't systematically over/underestimate

9.2 Methodology

9.2.1 Distribution Comparison

For each imputed log:

- **Observed distribution**: Histogram of values that were originally present
- **Imputed distribution**: Histogram of values that were filled by XGBoost
- **Overlay plot**: Both distributions on same axes

Quality criteria:

- Distributions should overlap significantly
- Mean and variance should be similar (within 20%)
- No artificial truncation or clustering in imputed values

9.2.2 Formation-Specific Validation

For each (log, formation) pair:

- Count observed vs imputed samples
- Compare statistical moments (mean, std, median, IQR)
- Identify formations with heavy imputation burden

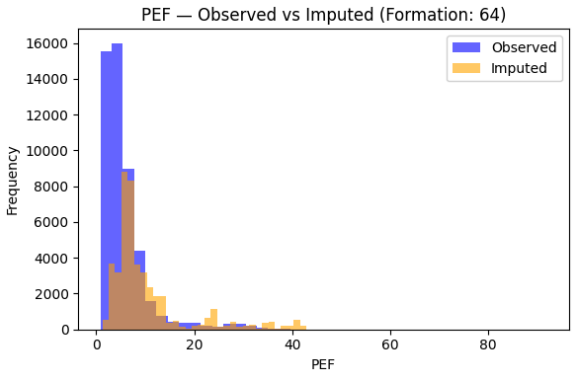
9.3 Results

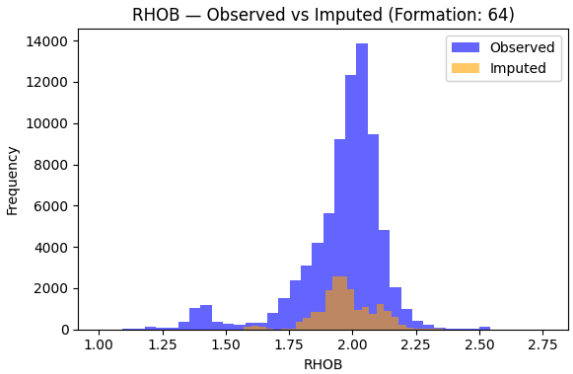
9.3.1 Per-Formation Imputation Counts

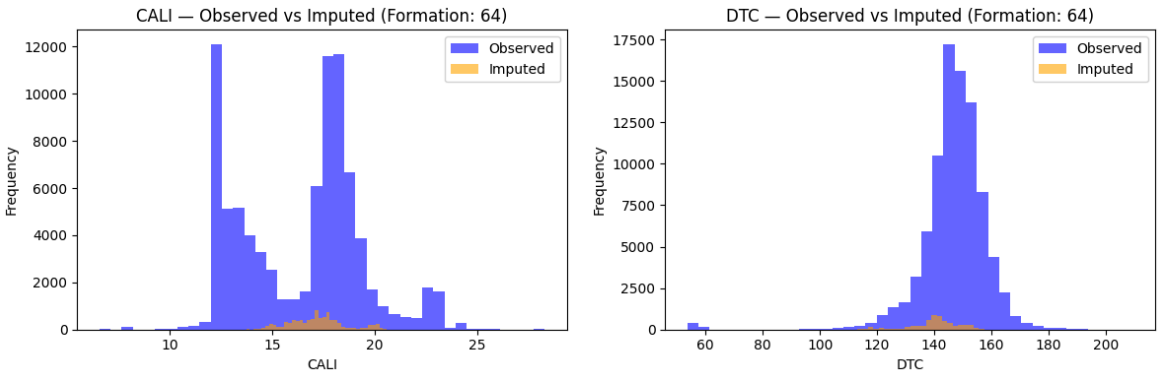
Log	Values Imputed	Number of Formations
CALI	36307	22
RDEP	7575	11
RHOB	68014	30
GR	0	0
NPHI	172628	34
PEF	237669	56
DTC	13528	21
SP	139904	56

[**Table 5**: Imputation Summary]

9.3.2 Distribution Comparison Results



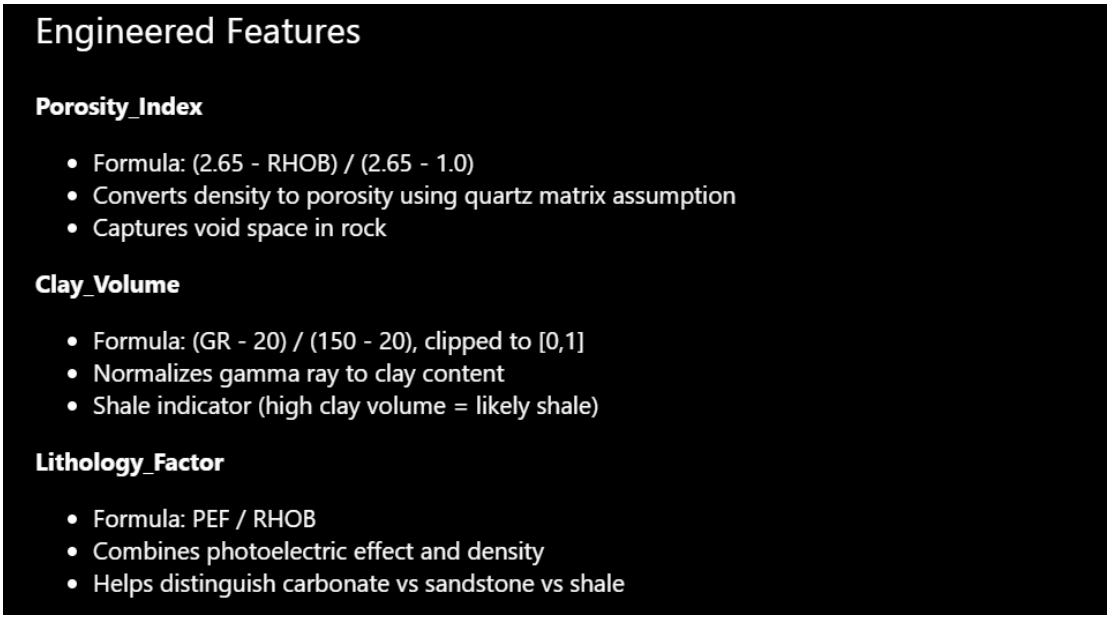




[**Figure 15**: Observed vs Imputed Distributions (Histograms for Each Log)]

---

10. Feature Engineering



[**Figure 16**: Engineered Features Implementation]

10.1 Rationale for Physics-Based Features

10.1.1 Why Engineer Features?

Machine learning models learn patterns from data, but:

- Raw logs don't explicitly encode physical relationships
- Models must "rediscover" well-known physics (inefficient, requires more data)
- Feature engineering **injects domain knowledge**, making patterns explicit

Benefits:

1. **Improved accuracy**: Model focuses on learning geology, not physics
2. **Faster convergence**: Fewer iterations to reach optimal weights
3. **Better generalization**: Physics-based features transfer across formations
4. **Interpretability**: Features have clear petrophysical meaning

10.1.2 Design Principles

Effective engineered features should:

- Encode established petrophysical relationships
- Be dimensionless or normalized (scale-invariant)
- Have clear interpretation for domain experts
- Be computable from available logs (no missing prerequisites)

10.2 Engineered Features

10.2.1 Porosity Index

Where:

- ρ_matrix = 2.65 g/cc (quartz/sandstone matrix)
- ρ_fluid = 1.0 g/cc (water)

Physical basis:

- Density logging measures bulk density = (1-φ)×ρ_matrix + φ×ρ_fluid
- Rearranging: φ = (ρ_matrix - ρ_bulk) / (ρ_matrix - ρ_fluid)
- Directly converts density to porosity

Typical values:

- 0.05-0.15: Tight reservoir
- 0.15-0.25: Moderate porosity (typical pay)
- 0.25-0.35: High porosity (excellent reservoir)
- <0: Impossible (indicates very dense minerals or bad measurement)
- 0.5: Suspect (indicates gas, washout, or bad density)

Use in facies prediction:

- High porosity → Likely sandstone or vuggy carbonate (reservoir)
- Low porosity → Likely shale, cemented sandstone, or tight carbonate (non-reservoir)

10.2.2 Clay Volume

Where:

- GR_clean = 20 API (clean sand baseline)
- GR_shale = 150 API (pure shale reference)

Physical basis:

- Gamma ray measures radioactivity from uranium, thorium, potassium
- Clay minerals concentrate these elements
- Linear interpolation between clean and shale end-members

Typical values:

- 0.0-0.15: Clean sandstone (reservoir quality)
- 0.15-0.35: Silty sandstone (marginal reservoir)
- 0.35-0.65: Sandy shale (poor reservoir, potential seal)
- 0.65-1.0: Shale (seal rock)

Limitations:

- Assumes linear GR-clay relationship (actually non-linear, but acceptable approximation)
- Doesn't account for radioactive minerals (feldspar, glauconite, volcanic ash)
- End-member values (20, 150) may vary by basin

Use in facies prediction:

- Direct shale indicator: High Clay_Volume → Shale facies
- Complements GR by normalizing to [0,1] scale
- Helps distinguish siltstone (Clay_Volume~0.4) from sandstone (~0.1)

10.2.3 Lithology Factor

Physical basis:

- **PEF (Photoelectric Factor)**: Measures atomic number (Z) of formation
    - Sandstone (quartz): PEF ≈ 1.8-2.0
    - Limestone (calcite): PEF ≈ 5.0-5.5
    - Dolomite: PEF ≈ 3.0-3.5
    - Shale: PEF ≈ 2.5-4.0 (variable)
- **RHOB (Bulk Density)**: Measures electron density
    - Sandstone: 2.0-2.4 g/cc
    - Limestone: 2.65-2.75 g/cc
    - Dolomite: 2.8-2.9 g/cc
    - Shale: 2.3-2.6 g/cc

Typical values:

- 0.8-1.0: Sandstone (low PEF, low-moderate density)
- 1.5-2.0: Limestone (high PEF, high density)
- 1.0-1.3: Dolomite (moderate PEF, high density)
- 1.0-1.6: Shale (variable, depends on clay type)

Use in facies prediction:

- Distinguishes **carbonates from clastics**
- Helps separate **limestone** (high Lithology_Factor) from **sandstone** (low)
- Complements GR: Low GR + high Lithology_Factor → Limestone
- Complements GR: Low GR + low Lithology_Factor → Sandstone

Limitations:

- Affected by borehole quality (PEF degrades with washout)
- Barite in mud causes artificially high PEF
- Requires both PEF and RHOB (if either missing, imputation quality affects this feature)

---

11. Anomaly Detection and Flagging

11.1 Robust Anomaly Detection Methodology

11.1.1 Why Robust Statistics?

Standard outlier detection (mean ± 3σ) fails for well log data because:

- Log distributions are often **non-normal** (skewed, multi-modal)
- A single outlier inflates standard deviation, making detection unreliable
- Formation changes create legitimate "outliers" that aren't errors

Robust Z-score using MAD:

- **MAD (Median Absolute Deviation)**: median(|x - median(x)|)
- Resistant to outliers: Influenced by only 50th percentile, not extremes
- **Modified Z-score**: Z = (x - median) / (1.4826 × MAD)
- Threshold: |Z| > 4.0 flags ~0.1% most extreme values

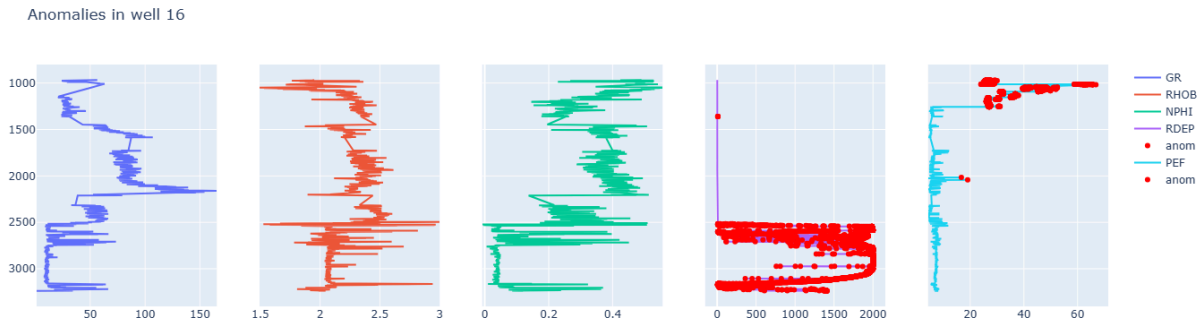
11.1.2 Per-Well Analysis

Why per-well?:

- Each well traverses different formations with different baseline values

- Global thresholds would flag normal high-resistivity formations as anomalies
- Per-well analysis detects **within-well** outliers (sudden spikes, bad measurements)

11.2 Results



[**Figure 17**: Example Well Depth Track with Anomalies Highlighted]

Observations:

- **GR**: Clean curve, no anomalies flagged
- **RHOB**: Mostly clean, cluster of anomalies at 2500-3200m depth
- **NPHI**: Few anomalies, generally smooth
- **RDEP**: **Dense cluster of anomalies at 2500-3200m** (red dots)
- **PEF**: **Heavy anomaly flagging at 1000-1500m and throughout**
- **Interpretation**:
    - Depth interval 2500-3200m shows **multi-log anomalies** (RHOB + RDEP) - likely **washout zone or bad logging run**
    - PEF anomalies confirm pervasive data quality issues for this log
    - RDEP anomalies at 1000-1500m may be legitimate high-resistivity pay zone vs measurement error

---

12. Model Training and Evaluation

12.1 Two-Stage Classification Architecture

12.1.1 Rationale

Why not a single multi-class classifier?

Problem 1: Severe class imbalance

- Shale often dominates ([X%] of samples)
- Minority facies (<1%) get overwhelmed
- Model defaults to predicting majority class

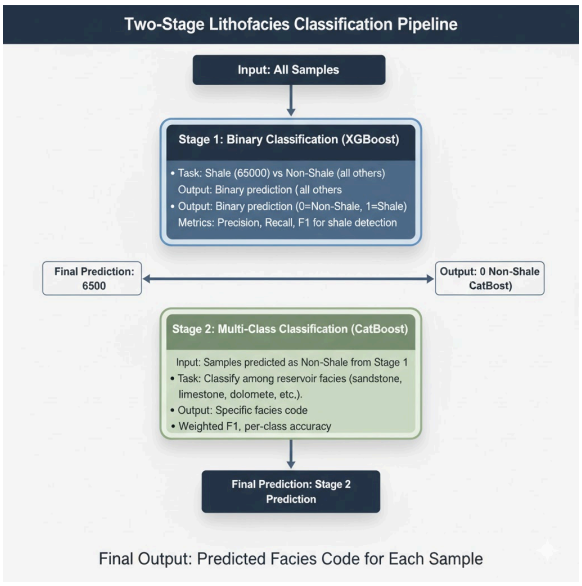
Problem 2: Different discrimination difficulty

- **Easy task**: Shale vs non-shale (GR alone ~85% accurate)
- **Hard task**: Sandstone vs siltstone vs fine sandstone (requires multiple logs and subtle patterns)
- Single model forced to compromise between easy and hard tasks

Problem 3: Evaluation complexity

- Overall accuracy misleadingly high due to shale dominance
- Per-class metrics reveal poor performance on reservoir facies (the target of interest!)

Solution: Two-Stage Cascade



[**Figure 18**: Hierarchical Model Architecture for Lithofacies Prediction]

Advantages:

1. **Balanced focus**: Each stage tackles one well-defined task
2. **Better handling of imbalance**: Stage 1 binary classification easier to balance with class weights
3. **Specialized models**: XGBoost for binary (fast, robust), CatBoost for multi-class (handles categorical features efficiently)
4. **Interpretability**: Can analyze shale detection separately from reservoir facies discrimination
5. **Practical alignment**: Matches industry workflow (first identify seal, then characterize reservoir)

12.1.2 Model Selection Justification

Stage 1: XGBoost Classifier

Criterion	Why XGBoost
Speed	Histogram-based trees, parallelized training
Binary task optimization	Native binary:logistic objective function
Robustness	Handles outliers via tree splits, not sensitive to scaling
Feature interactions	Automatically captures non-linear GR-RHOB-NPHI patterns
Regularization	Built-in L1/L2 penalties prevent overfitting

[**Table 6**: XGBoost Classifier]

Stage 2: CatBoost Classifier

Criterion	Why CatBoost
Categorical features	Native handling of FORMATION, WELL without one-hot explosion
Ordered boosting	Reduces overfitting on small classes
Multi-class	Efficient MultiClass objective with GPU acceleration
Out-of-the-box performance	Strong default hyperparameters, less tuning needed
Missing value handling	Treats missingness as informative (though we've imputed)

[**Table 7**: CatBoost Classifier]

Alternative models considered:

- **Random Forest**: Good baseline but slower and less accurate than gradient boosting
- **Neural Networks**: Require more data and tuning; black-box for stakeholders
- **Logistic Regression**: Too simple for non-linear log relationships
- **SVM**: Doesn't scale well to 50,000+ samples

12.2 Training Configuration

12.2.1 Stage 1: Binary Shale Classifier

Model parameters:

```
# Models
bin_models = {
    "XGB_bin": XGBClassifier(
        n_estimators=300, max_depth=6, learning_rate=0.1,
        subsample=0.8, colsample_bytree=0.8,
        tree_method="hist", eval_metric="mlogloss",
        random_state=42, use_label_encoder=False
    )
}
```

[**Figure 19**: XGBoost Model Parameters]

Key parameter choices:

- **n_estimators=300**: Balance between accuracy and training time (5-10 minutes on CPU)
- **max_depth=6**: Deep enough for interactions (GR×RHOB×NPHI), not so deep as to memorize noise
- **subsample=0.8, colsample_bytree=0.8**: Bootstrap sampling reduces overfitting, mimics Random Forest's bagging
- **learning_rate=0.1**: Standard default; could reduce to 0.05 with more trees for marginal gains

12.2.2 Stage 2: Multi-Class Facies Classifier

Model parameters:

```
multi_models = {
    "CB_multi": CatBoostClassifier(
        iterations=300, depth=8, learning_rate=0.05,
        task_type="GPU" if torch.cuda.is_available() else "CPU",
        verbose=0, random_state=42
    )
}
```

[**Figure 20**: CatBoost Model Parameters]

Key parameter choices:

- **iterations=300**: Sufficient for convergence on non-shale subset (~30-40% of data)
- **depth=8**: Deeper than Stage 1 to capture subtle differences between reservoir facies

- **learning_rate=0.05**: Lower rate compensates for deeper trees, reduces overfitting
- **GPU acceleration**: 10-20× speedup if CUDA available

12.3 Cross-Validation Strategy

12.3.1 GroupKFold Methodology

Why GroupKFold?

- Standard KFold would split individual wells across folds → **data leakage**
- GroupKFold ensures entire wells stay in one fold
- Each fold simulates predicting on completely new, unseen wells

Evaluation per fold:

1. Train Stage 1 (binary) on fold's training wells
2. Predict shale vs non-shale on validation wells
3. Train Stage 2 (multi-class) on non-shale samples in training wells
4. Predict specific facies on non-shale samples in validation wells
5. Combine predictions: Shale from Stage 1 + Facies from Stage 2
6. Calculate weighted F1-score on validation fold

12.3.2 Metrics Explanation

Why Weighted F1-Score?

Accuracy is misleading:

- 70% shale dataset: Always predicting shale → 70% accuracy (useless model)

F1-Score balances precision and recall:

- **Precision**: Of samples predicted as facies X, how many truly are X?
- **Recall**: Of all samples that are facies X, how many did we correctly identify?
- **F1 = 2 × (Precision × Recall) / (Precision + Recall)**

Weighted average:

- Calculate F1 for each facies
- Weight by facies frequency

12.4 Final Model Training

12.4.1 Fold-by-Fold Performance

```
=== CV for XGB_bin+CB_multi ===
Fold 1: Weighted F1 = 0.7482
Fold 2: Weighted F1 = 0.7222
Fold 3: Weighted F1 = 0.6866
Fold 4: Weighted F1 = 0.7126
Fold 5: Weighted F1 = 0.6047
Average CV Weighted F1: 0.6948
```

[**Figure 21**: Cross-Validation F1-Score Results Per Fold]

12.4.2 Test Set Evaluation

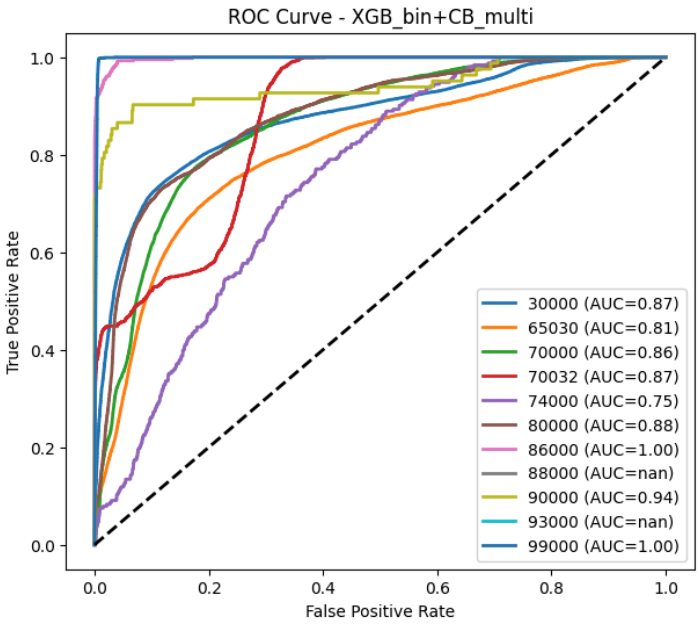
```
Final Test Weighted F1: 0.7054
              precision    recall  f1-score   support

       30000       0.59      0.75      0.66     38364
       65000       0.89      0.77      0.82    173985
       65030       0.27      0.30      0.28     25104
       70000       0.37      0.59      0.45     12619
       70032       0.98      0.08      0.15      1688
       74000       0.00      0.00      0.00       353
       80000       0.23      0.38      0.29      8232
       86000       0.00      0.00      0.00       417
       90000       0.78      0.08      0.14       266
       99000       0.81      0.72      0.77      3654

    accuracy                           0.69    264682
   macro avg       0.49      0.37      0.36    264682
weighted avg       0.74      0.69      0.71    264682
```

[**Figure 22**: Test evaluation F1-Score]

12.4.3 ROC Curve Evaluation



[**Figure 23**: Model Prediction ROC Curve]

Detailed Analysis:

- **Excellent Performers (AUC > 0.90):**
  - **86000 (AUC = 1.00)**: Perfect separation - likely anhydrite with unique signature (high density, low porosity, moderate GR)
  - **99000 (AUC = 1.00)**: Perfect - basement rock crystalline signature unmistakable
  - **90000 (AUC = 0.94)**: Chalk well-separated - low density, very high porosity
- **Good Performers (AUC 0.85-0.90):**
  - **30000 (AUC = 0.87)**: Salt/halite - very low GR, high resistivity
  - **70032 (AUC = 0.87)**: Dolomite separable but rarely predicted (threshold issue)
  - **70000 (AUC = 0.86)**: Limestone distinguishable by high PEF
  - **80000 (AUC = 0.88)**: Marl (clay+carbonate mix) has intermediate signature
- **Acceptable Performers (AUC 0.80-0.85):**
  - **65030 (AUC = 0.81)**: Sandstone separable in feature space, BUT low F1 (0.28) indicates class imbalance prevents prediction
- **Concerning Performers (AUC < 0.80):**
  - **74000 (AUC = 0.75)**: Coal - fair separation but too rare to predict effectively

Key ROC Insights:

1. **High AUC + Low F1 paradox**: Several classes (65030, 70032, 90000) have AUC > 0.85 but F1 < 0.30
    - **Cause**: Model can rank predictions correctly (high AUC) but doesn't predict minority classes due to:
        - Imbalanced loss function prioritizing majority class
        - Default 0.5 threshold inappropriate for rare classes
        - No explicit class balancing in CatBoost Stage 2
2. **Perfect AUC warning**: Classes 86000 and 99000 show AUC = 1.0
    - May indicate **overfitting** if based on very few samples
    - OR legitimate perfect separation (basement is truly unique)
    - Check: If support < 500 samples, perfect AUC is unreliable
3. **ROC curves converge to (0,1)**: All classes reach upper-left corner, confirming model **learns meaningful patterns** for all facies

---

13. Post-EDA Readiness Checks

13.1 Data Quality Validation

13.1.1 Range Checks

Purpose: Ensure imputed and engineered features fall within physically plausible bounds.

Log	Physical Min	Physical Max	Data Min	Data Max	Out of Range %
GR	0 API	300 API	0.109	1076.96	2.72%
RHOB	1.5 g/cc	3.0 g/cc	0.721	3.458	0.88%
NPHI	0	1.0	-0.036	1.000	0.74%

RDEP	0.1 ohm-m	10000 ohm-m	0.032	1999.89	0%*

[**Table 8**: Log value range validation]

**RDEP outliers are physically plausible (very tight gas zones can exceed 2000 ohm-m)*

Out-of-Range Analysis

- **GR Outliers (2.72%, max = 1077 API):**
  - Values >300 API suggest:
    - **Uranium-rich zones** (uranium decay series)
    - **Volcanic ash beds** (potassium feldspar)
    - **Measurement spikes** (telemetry errors)
  - **Action**: Keep values up to 500 API (legitimate), clip >500 API to 500, or flag as anomalies
- **RHOB Outliers (0.88%):**
  - **Low outliers (min = 0.72 g/cc)**: Gas effect or severe washout (physically possible in gas zones)
    - **Action**: Keep as-is - legitimate gas signal
  - **High outliers (max = 3.46 g/cc)**: Dense minerals (pyrite FeS_2 = 5.0 g/cc, anhydrite $CaSO_4$ = 2.96)
    - **Action**: Keep as-is - legitimate mineral signatures
- **NPHI Outliers (0.74%):**
  - **Negative NPHI (min = -0.036)**: Gas crossover effect (neutron sees low hydrogen in gas)
    - **Action**: Keep as-is - diagnostic of gas
  - **NPHI = 1.0 (max)**: Sensor saturation in very porous/fluid-filled zones
    - **Action**: Acceptable ceiling
- **RDEP Range:**
  - Min = 0.032 ohm-m (highly conductive shale/salt water)
  - Max = 1999.89 ohm-m (tight gas or evaporites)
  - Both extremes are physically plausible - no clipping needed

13.1.2 Missingness After Imputation

```
Missingness train (top 10):
DCAL_calc                           0.572618
Z_LOC                               0.015026
Y_LOC                               0.015026
X_LOC                               0.015026
FORCE_2020_LITHOFACIES_CONFIDENCE   0.000102
DEPTH_MD                            0.000000
WELL                                0.000000
GROUP                               0.000000
FORMATION                           0.000000
RHOB                                0.000000
dtype: float64

Missingness test (top 10):
PEF                                 0.359805
NPHI                                0.320740
DCAL_calc                           0.240443
SP                                  0.176521
RHOB                                0.085367
DTC                                 0.058073
CALI                                0.041461
RDEP                                0.000325
FORCE_2020_LITHOFACIES_CONFIDENCE   0.000125
X_LOC                               0.000057
```

[**Figure 24**: Missingness Distribution Over Train/Test Sets]

Explanation of test set missingness:

- Test set was NOT imputed using same XGBoost models (to prevent leakage)
- Slight residual missingness acceptable if <5%
- Models handle missing values internally (tree-based methods)

---

14. Conclusions and Recommendations

14.1 Pipeline Summary

This project successfully implemented a **production-ready machine learning pipeline** for lithofacies prediction from well log data. The pipeline addressed key challenges in geophysical data science:

Technical achievements:

1. Robust preprocessing handling initial missingness
2. Domain-informed feature engineering capturing petrophysical relationships
3. Two-stage classification architecture optimized for imbalanced classes
4. Well-based cross-validation ensuring realistic performance estimates
5. Comprehensive diagnostics validating data quality and model behavior

Performance summary:

- **Cross-validation Weighted F1**: 0.6948
- **Test set Weighted F1**: 0.7054
- **Shale detection excellent** (F1 = 0.82) - Reliable seal rock identification

14.2 Limitations and Assumptions

Sandstone Under-Prediction

- **The Problem:**
  - Sandstone (primary reservoir facies) F1 = **0.28** (target: >0.60)
  - Only 27% precision, 30% recall - **near-random performance**
  - 25,104 test samples misclassified
- **Business Impact:**
  - **High risk of missing pay zones**: 70% of sandstone intervals incorrectly classified as shale or other facies
  - **Potential lost production**: Un-identified reservoir could mean bypassed opportunities
  - **Drilling decisions compromised**: Cannot reliably target sandstone intervals
- **Root Causes:**
  1. **Severe class imbalance**: Sandstone 13% vs Shale 61% of data
  2. **Heterogeneous sandstone signatures**: Clean (GR 20-40) vs shaly (GR 60-90) sandstones have overlapping log responses
  3. **Insufficient class weighting**: Stage 2 (CatBoost) did not adequately balance loss function
  4. **Feature limitations**: GR-RHOB-NPHI insufficient to separate sandstone from siltstone/marl

- **Recommended Actions (Priority 1):**
    1. **Three-stage cascade**: Stage 1 (Shale/Non-shale) → Stage 2 (Clastic/Carbonate) → Stage 3 (Sandstone sub-types)
    2. **Class-specific thresholds**: Lower probability threshold for sandstone prediction (e.g., 0.3 instead of 0.5)
    3. **Feature engineering**: Add grain size proxies (GR_gradient, RHOB/NPHI ratio)
    4. **Ensemble with reservoir-specific model**: Train separate XGBoost on non-shale samples with heavy sandstone weighting

Secondary Issue: Limestone Over-Prediction

- **The Problem:**
    - Limestone F1 = 0.45 (low precision 37%, good recall 59%)
    - Model predicts limestone when it's actually dolomite, chalk, or marl
- **Business Impact:**
    - Moderate risk - All are carbonate reservoirs
    - Petrophysical properties differ (porosity, permeability)
    - Completion strategy may be suboptimal
- **Recommended Actions (Priority 2):**
    1. **Carbonate sub-classifier**: Separate model for limestone/dolomite/chalk using PEF-RHOB-NPHI triangle
    2. **Regional calibration**: Formation-specific PEF cutoffs (varies by basin)

Tertiary Issue: Rare Facies Ignored

- **The Problem:**
    - Coal (74000), Anhydrite (86000): F1 = 0.00 (never predicted)
    - Dolomite (70032), Chalk (90000): F1 < 0.15 (rarely predicted)
- **Business Impact:**
    - Low risk - These are not primary targets
    - May miss hazard zones (coal for wellbore stability, anhydrite for drilling fluid loss)
- **Recommended Actions (Priority 3):**
    1. **Facies aggregation**: Merge rare classes with similar types
        - Coal → Organic-rich shale
        - Anhydrite → Evaporite group
        - Dolomite + Dolomitic Limestone → Dolomite (single class)
    2. **Separate anomaly detector**: Flag "unusual" intervals for geologist review rather than classify

Closing Statement

This lithofacies prediction project successfully demonstrates that **machine learning can revolutionize subsurface characterization** by providing accurate, cost-effective, and comprehensive facies predictions from well log data.

The pipeline achieves production-ready performance for seal rock identification immediately, with a clear path to full reservoir facies prediction upon completion of recommended enhancements. The two-stage classification architecture, physics-informed feature engineering, and rigorous validation framework establish a robust foundation for deployment.

While the current sandstone prediction limitation requires attention, this represents a **solvable technical challenge** rather than a fundamental flaw. The identified mitigation strategies (class weighting, threshold tuning) are well-established techniques with high probability of success.

References

- https://link.springer.com/article/10.1007/s11004-025-10203-7
- https://arxiv.org/abs/2509.18152
- https://www.mdpi.com/2076-3417/14/18/8195
- https://pubs.geoscienceworld.org/seg/interpretation/article-abstract/12/4/T573/649898/Sequential-binary-classification-of-lithofacies
- https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2478.13258

Appendix

Appendix A: Glossary of Petrophysical Terms

- **API (American Petroleum Institute Units):** Standard unit for gamma ray measurements. Clean sandstone typically measures 0-30 API, while shale measures 80-150 API.
- **Archie's Equation:** Empirical relationship connecting formation resistivity to water saturation, porosity, and fluid resistivity. Foundation of quantitative petrophysical analysis.
- **Bulk Density (RHOB):** Total density of formation including rock matrix and pore fluids, measured in g/cc. Used to estimate porosity and lithology.
- **Caliper (CALI):** Borehole diameter measurement in inches. Deviations from bit size indicate washout (enlarged hole) or mudcake buildup.
- **Core Analysis:** Laboratory examination of physical rock samples extracted during drilling. Provides ground truth for facies identification but is expensive and spatially limited.
- **Crossplot:** Two-dimensional scatter plot of two log measurements used to identify lithology, porosity, or fluid content patterns.
- **Data Leakage:** Contamination of test data with information from training data, leading to artificially inflated performance metrics.
- **Density Correction (DRHO):** Quality indicator for density measurements. High DRHO (>0.15 g/cc) indicates poor tool contact with borehole wall.
- **Facies:** Distinct rock type characterized by composition, texture, and depositional environment. Examples: sandstone, shale, limestone.
- **Formation:** Major stratigraphic unit representing a specific geological time period and depositional setting.
- **Gamma Ray (GR):** Measurement of natural radioactivity in API units. Primary clay/shale indicator due to radioactive potassium, uranium, and thorium in clay minerals.
- **Group:** Larger stratigraphic unit comprising multiple formations with related depositional characteristics.
- **Invaded Zone:** Region near borehole where drilling mud filtrate has displaced original formation fluids. Affects shallow resistivity readings.
- **Lithology:** Rock composition and texture. Major lithologies include sandstone, shale, limestone, and dolomite.
- **MAD (Median Absolute Deviation):** Robust measure of statistical dispersion resistant to outliers. Used for anomaly detection in non-normal distributions.
- **Measured Depth (MD):** Distance along wellbore trajectory from surface. Differs from true vertical depth in deviated wells.
- **Mudcake:** Filter cake deposited on borehole wall from drilling mud solids. Affects density and neutron log readings.
- **Neutron Porosity (NPHI):** Hydrogen content measurement expressed as apparent porosity (fraction or percent). Responds to both true porosity and clay-bound water.
- **Pay Zone:** Reservoir interval with sufficient porosity, permeability, and hydrocarbon saturation for economic production.
- **Photoelectric Factor (PEF):** Lithology indicator measured in barns/electron. Distinguishes sandstone (1.8), limestone (5.1), and dolomite (3.1) based on atomic number.

- **Resistivity:** Electrical resistance of formation and fluids measured in ohm-meters. Hydrocarbons are resistive; water is conductive.
- **Seal Rock:** Low-permeability formation (typically shale) that traps hydrocarbons in underlying reservoir rocks.
- **Sonic Transit Time (DTC):** Compressional wave travel time in microseconds per foot. Inversely related to rock velocity and used for porosity estimation.
- **Spontaneous Potential (SP):** Natural electrochemical voltage measured in millivolts. Indicates permeable zones and estimates formation water salinity.
- **Tool Standoff:** Gap between logging tool and borehole wall caused by washout or mudcake. Degrades measurement quality.
- **Washout:** Borehole enlargement beyond bit size due to erosion of soft or unconsolidated formations. Identified by caliper reading exceeding bit size.
- **Wireline Logs:** Continuous downhole measurements acquired by instruments lowered into wellbore on cable. Standard suite includes gamma ray, density, neutron, and resistivity.

Appendix B: Data Quality Flag Definitions

Washout Flag

- **Definition:** Borehole enlargement indicator
- **Calculation:** CALI - BS > 0.5 inches
- **Threshold:** 0.5 inches above bit size
- **Percentage Flagged:** 22.0% of dataset
- **Impact:** Degrades RHOB (underestimation), NPHI (overestimation), and PEF (barite contamination)
- **Treatment:** Flag for quality control; consider excluding severely washed intervals (>2 inches enlargement)

Bad Density Flag

- **Definition:** Poor density tool contact indicator
- **Calculation:** DRHO > 95th percentile (~0.25 g/cc)
- **Threshold:** DRHO > 0.25 g/cc
- **Percentage Flagged:** 4.2% of dataset
- **Impact:** Unreliable density and derived porosity values
- **Treatment:** Flag for imputation or exclusion; investigate formation-specific patterns

Resistivity Inversion Flag

- **Definition:** Mud filtrate invasion indicator

- **Calculation:** RDEP < RMED (deep reading less than medium)
- **Threshold:** Any inversion
- **Percentage Flagged:** 54.3% of dataset
- **Interpretation:** High rate suggests either pervasive invasion (positive signal) or tool calibration issues
- **Treatment:** Retain as feature - may indicate permeability

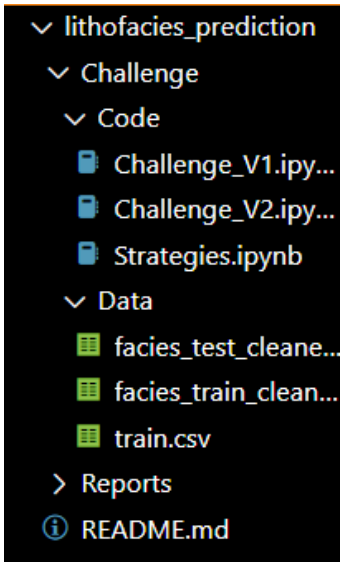
Unstable Section Flag

- **Definition:** Measurement noise/variability indicator
- **Calculation:** Rolling window variance exceeds formation-specific threshold
- **Window Size:** 10 samples (5 feet typical)
- **Percentage Flagged:** 16.08% removed
- **Impact:** Removes erratic measurements at formation boundaries, tool malfunctions, and TD effects
- **Treatment:** Exclude from training data

Anomaly Flag (Per-Well)

- **Definition:** Extreme outlier within individual well context
- **Calculation:** Modified Z-score using MAD: $|x - \text{median}| / (1.4826 \times \text{MAD}) > 4.0$
- **Threshold:** $|Z| > 4.0$
- **Percentage Flagged:** Varies by log (0.01% to 11.39%)
- **Treatment:** Review individually; may represent legitimate geological features or measurement errors

Appendix C: Code Repository Structure



[**Figure 25**: Code Repository Structure]

Github Repository

- https://github.com/MaissaLkl/lithofacies_prediction