MACHINE LEARNING BASED NONLINEAR POLYNOMIAL , LINEAR, RIDGE, AND RANDOM FOREST REGRESSION MODELS ALGORITHMS PREDICTION OF RESEARCH OCTANE NUMBER OF REFORMATE FOR ARABIAN LIGHT CRUDE

Abbas Albalola

Yanbu Aramco Sinopec Refining Company (YASREF), Yanbu Industrial City

ABSTRACT

This study develops and compares four machine learning regression models to predict Research Octane Number (RON) from specific gravity (S.G) and aromatic hydrocarbon content (v%) using a consolidated dataset of 108 unique Cooperative Fuel Research (CFR) engine measurements. The dataset spans RON 102.3–107.5, S.G 0.8377–0.8611, and Aromatics 81.48–89.96 V%, partitioned into training (75 samples, 70%), test (22 samples, 20%), and validation (11 samples, 10%) subsets. Four regression approaches were evaluated: (i) Linear regression using S.G and Aromatics as predictors; (ii) Polynomial regression incorporating derived quadratic and interaction terms ($S.G^2$, $Aromatics^2$, S.G×Aromatics); (iii) Ridge regression with L2 regularization (alpha=0.001, selected by 5-fold cross-validation); and (iv) Random Forest regression (300 trees, maximum depth 4). Model performance was assessed using the coefficient of determination (R^2), root-mean-square error (RMSE), and standard deviation of prediction bias, evaluated separately on training, test, and validation subsets to distinguish between fitting accuracy and true generalization. Linear and Polynomial models achieved the highest test-set R^2 (0.902 and 0.909, respectively), indicating superior generalization to unseen data. Pearson correlations between RON and predictors ranged from r=0.926 (S.G) to r=0.948 (S.G×Aromatics), indicating strong positive relationships. Ridge regression (Test R^2=0.901) provided comparable performance with enhanced coefficient stability, while Random Forest exhibited overfitting characteristics (Train R^2=0.960, Test R^2=0.889). Results demonstrate that Linear regression offers the optimal balance of interpretability, simplicity, and generalization performance for production deployment, with Polynomial regression as a viable alternative if marginal accuracy gains justify increased model complexity. Recommendations for real-time deployment include daily bias monitoring, extrapolation flagging, and periodic model recalibration.

KEYWORDS

Research Octane Number, Soft Sensor, Machine Learning, Linear Regression, Polynomial Regression, Ridge Regression, Random Forest, CFR Engine, Specific Gravity, Aromatic Content.

David C. Wyld et al. (Eds): BMLI, CACIT, EDUAIT, InWES – 2026
pp. 01-16, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.161801

1. INTRODUCTION

Research Octane Number (RON) is a key quality parameter used to assess the antiknock performance of gasoline and reformate streams in petroleum refining. It is conventionally measured using a standardized Cooperative Fuel Research (CFR) engine test (ASTM D2699), a procedure that requires specialized, expensive equipment and yields results only after significant time delay. In modern refinery operations driven by real-time process control and Industry digital transformation, there is an acute need for rapid, non-destructive RON estimation to enable closed-loop optimization of reformer severity, catalytic performance, gasoline blending, and feed quality monitoring.

Machine learning soft sensors—predictive models trained on historical CFR data and deployed in distributed control systems (DCS), advanced process control (APC), or real-time information systems (e.g., PI AF)—offer an attractive alternative to laboratory testing. Such soft sensors can provide RON predictions within seconds, enabling refinery operators to make rapid decisions on processing adjustments and blending strategies.

The relationship between RON and readily measurable fuel properties, particularly specific gravity and aromatic hydrocarbon content, has been established empirically and theoretically in the literature. However, systematic comparison of different regression modeling approaches—ranging from simple linear to complex ensemble-based methods—using a modern, consolidated dataset with rigorous train-test-validate protocols remains limited.

This study aims to fill that gap by: (1) assembling a comprehensive dataset of 108 unique CFR measurements spanning a representative range of RON (102.3–107.5), specific gravity (0.8377–0.8611), and aromatic content (81.48–89.96 v%); (2) developing four distinct regression models of increasing complexity; (3) using a rigorous 70/20/10 train-test-validate split to assess true generalization performance; (4) providing detailed model equations, performance metrics, and confidence intervals suitable for production deployment; and (5) making explicit recommendations on model selection based on generalization performance and interpretability.
The structure of this paper is as follows: Section 2 reviews related work on RON prediction and soft sensor development. Section 3 describes the dataset, preprocessing, and modeling methodology. Section 4 presents results including correlation analysis, model equations, and performance comparison across train, test, and validation subsets. Section 5 discusses findings, model-by-model analysis, and practical deployment considerations. Section 6 concludes with recommendations for production use and directions for future work.

2. RELATED WORK

Early empirical correlations for RON prediction focused on simple linear relationships with S.G, distillation points, and bulk hydrocarbon group composition (paraffins, naphthenes, aromatics). These classical approaches provided reasonable accuracy within narrow property ranges but suffered from poor extrapolation behavior when processing conditions shifted.

More recent work has explored spectroscopic soft sensors (near-infrared, mid-infrared) combined with chemometric methods such as Partial Least Squares (PLS) and Principal Component Regression (PCR). While these achieve high accuracy, they require expensive analyzers, frequent calibration, and are sensitive to fouling—limiting deployment to high-value applications.

Machine learning approaches, including Support Vector Regression (SVR), Random Forest, Gradient Boosting Machines (GBM), and Artificial Neural Networks (ANN), have demonstrated

superior predictive performance, particularly when trained on large, diverse datasets. However, their adoption in refinery operations has been limited by computational complexity, lack of interpretability, and challenges in integration with legacy DCS and APC systems.

This study focuses on four interpretable, computationally efficient regression approaches: Linear, Polynomial, Ridge (regularized), and Random Forest, chosen for their balance of accuracy, stability, and practical deployability in refinery environments.

3. DATASET AND METHODOLOGY

3.1 Dataset Composition

A consolidated dataset comprising 108 unique CFR RON measurements was compiled from refinery operations across a multi-month collection period. Each sample includes measured properties: Research Octane Number (RON, ASTM D2699), Specific Gravity at 60/60°F (S.G, ASTM D4052), and aromatic hydrocarbon content (v%, determined by gas chromatography ASTM D6839).

Derived features were computed for each sample: $S.G^2$ (specific gravity squared), $Aromatics^2$ (aromatic content squared), and S.G×Aromatics (cross-product interaction term). These derived features were included to capture potential nonlinear and synergistic relationships between S.G and Aromatics in determining RON.

Dataset characteristics:

- RON: 102.3–107.5 (mean=104.5±1.3)
- Specific Gravity: 0.8377–0.8611 (mean=0.8523±0.0049)
- Aromatics: 81.48–89.96 v% (mean=84.3±2.4)
- No missing values; all samples complete across all six variables

3.2 Train-Test-Validate Partition

The 108 samples were randomly partitioned using a fixed random seed (seed=42) to ensure reproducibility: Training subset (70%, n=75): used exclusively for model fitting and hyperparameter selection. Test subset (20%, n=22): held out to evaluate generalization performance on data not seen during training. Validation subset (10%, n=11): further held-out subset to assess model performance on a completely independent sample.

This 70/20/10 split ensures that performance metrics (R^2, RMSE, bias) reported on the test and validation subsets represent true generalization capability, not just interpolation within the training domain.

3.3 Regression Models

Four regression approaches were developed and compared:

(1) Linear Regression: Uses specific gravity and aromatics as the only predictors, fitting the model: $RON = \beta_0 + \beta_1 \cdot S.G + \beta_2 \cdot Aromatics$. This two-parameter model is simple, fully interpretable, and computationally trivial to deploy.
(2) Polynomial Regression: Incorporates all five features (S.G, Aromatics, $S.G^2$, $Aromatics^2$, S.G×Aromatics) using ordinary least-squares fitting. This model can capture weak nonlinear

relationships and interaction effects but is susceptible to multicollinearity among the derived terms.

(3) Ridge Regression: Applies L2 regularization to the same five-feature polynomial model. The regularization strength (alpha parameter) was selected via 5-fold cross-validation on the training subset, aiming to stabilize coefficients and reduce overfitting without sacrificing predictive accuracy.

(4) Random Forest: An ensemble of 300 decision trees, each trained on bootstrap samples of the training data with maximum depth 4 and minimum 3 samples per leaf. Predictions are the average of all 300 tree predictions. This non-parametric model can capture complex nonlinear patterns but risks overfitting on small training sets.

3.4 Model Evaluation Metrics

All models were evaluated using three metrics:

- R^2 (coefficient of determination): Fraction of variance in RON explained by the model, ranging from 0 (no predictive power) to 1 (perfect prediction).
- RMSE (root-mean-square error): Square root of the average squared prediction error, in RON units; directly interpretable as typical prediction magnitude error.
- Std Dev of Bias: Standard deviation of (actual - predicted) residuals, quantifying the consistency of predictions regardless of direction.

Metrics were computed separately for training, test, and validation subsets. Test-set R^2 was used as the primary criterion for model selection, as it best reflects true generalization performance to unseen data—a critical requirement for production deployment.

4. RESULTS

4.1 Correlation Analysis

Table 1Pearson correlation coefficients between RON and each predictor variable were computed across all 108 samples:

Variable	Pearson r with RON
S.G × Aromatics	0.9481
Aromatics²	0.9381
Aromatics	0.9375
S.G²	0.9271
S.G	0.9263

Table 1: Pearson correlations (r) between RON and predictor variables. All correlations are positive and strong ($r > 0.92$), indicating that both S.G and Aromatics individually, as well as their squared and interaction terms, are useful predictors of RON. The interaction term S.G×Aromatics shows the highest single correlation ($r=0.948$).

4.2 Model Equations and Fitted Coefficients

All models were fitted exclusively on the 75-sample training subset. The resulting equations are:

Linear Regression:
RON = −5.7767 + 94.5712·(S.G) + 0.3470·(Aromatics)

Polynomial Regression:
RON = 5704.3459 − 16257.3442·(S.G) + 29.3558·(Aromatics) + 11957.6264·(S.G²) + 0.0617·(Aromatics²) − 46.5298·(S.G×Aromatics)

Ridge Regression (alpha=0.001, raw-scale form):
RON = 717.6799 − 1318.0546·(S.G) − 2.5243·(Aromatics) + 798.3877·(S.G²) + 0.01334·(Aromatics²) + 0.6649·(S.G×Aromatics)

Random Forest Regression:
Non-parametric ensemble of 300 decision trees. No closed-form equation; predictions are generated by averaging the outputs of all 300 trees.

4.3 Model Performance Summary

Table 2 summarizes R^2 and RMSE for each model across the three data splits:

Split	n	Linear R^2	Poly R^2	Ridge R^2	RF R^2	Lin RMSE	Poly RMSE	Ridge RMSE	RF RMSE
Train	75	0.929	0.941	0.935	0.960	0.345	0.314	0.328	0.257
Test	22	0.902	0.909	0.901	0.889	0.435	0.419	0.436	0.463
Validate	11	0.812	0.843	0.810	0.829	0.638	0.582	0.641	0.607
Overall	108	0.908	0.922	0.912	0.928	0.403	0.372	0.394	0.356

Table 2: Model performance metrics (R^2 and RMSE) across train, test, validate, and overall (all 108 samples) subsets. Linear and Polynomial models achieve the highest test-set R^2 (0.902 and 0.909, respectively), indicating superior generalization.

Key observations:

- Test-Set R^2 (primary selection criterion): Linear 0.902, Polynomial 0.909, Ridge 0.901, Random Forest 0.889
- Training R^2 shows Random Forest fitting most tightly (0.960), but test R^2 is lowest (0.889)—classic overfitting
- Ridge alpha=0.001 (near zero) indicates minimal multicollinearity penalty; Ridge performs nearly identically to Linear
- Validation subset (n=11) shows lower R^2 across all models due to small sample size; test-set results are more reliable.

5. DISCUSSION

5.1 Correlation Strength and Predictability

All correlations between RON and predictor variables exceed r=0.92, demonstrating strong positive relationships across the full dataset range. The interaction term (S.G×Aromatics, r=0.948) and aromatic-squared term (Aromatics², r=0.938) show slightly elevated correlations compared to the linear terms, suggesting mild synergistic or nonlinear effects. Nevertheless, the base variables S.G and Aromatics alone ($r \geq 0.926$) capture the vast majority of RON variation, explaining why simple linear models achieve high R^2 values

5.2 Linear Regression Analysis

The linear model achieved Test R^2=0.902, RMSE=0.435, and is characterized by simple, interpretable coefficients. The positive coefficient on S.G (+94.57) indicates that for every 0.01 unit increase in specific gravity (e.g., from 0.8400 to 0.8500), RON increases by approximately 0.95 units, all else equal. The positive coefficient on Aromatics (+0.347) shows that each additional wt% of aromatic content corresponds to a 0.347 RON unit increase.

This model is recommended as the primary production soft sensor due to: (1) Simplicity—only two parameters to maintain and update; (2) Interpretability—operators can explain predictions to colleagues and auditors; (3) Robustness—no hyperparameter tuning required; and (4) Generalization—Test $R^2 = 0.902$ demonstrates reliable performance on unseen data.

5.3 Polynomial Regression Analysis

Polynomial regression achieved the highest test-set R^2 (0.909) and lowest test-set RMSE (0.419), a marginal 0.007 improvement in R^2 compared to Linear. However, the polynomial coefficients are notably unstable: S.G ranges from −16,257 to +11,958 depending on the interaction terms, a sign of severe multicollinearity among S.G, S.G², and S.G×Aromatics.

The instability arises because these derived terms are highly correlated (by construction, S.G² is a perfect quadratic function of S.G). While this does not affect in-sample or test-set predictions within the training domain (102.3–107.5 RON), it creates risk of wild extrapolation outside these ranges. Polynomial regression should therefore be used with caution, restricted to the domain (0.8377–0.8611 S.G, 81.48–89.96 Aromatics) and flagged whenever predictions approach boundaries.

Polynomial may be preferred over Linear if the marginal 0.007 R^2 improvement is deemed operationally significant for refinery control and if deployment includes automated domain-checking safeguards.

5.4 Ridge Regression Analysis

Ridge regression with alpha=0.001 (selected by 5-fold CV) achieved Test R^2=0.901, nearly identical to Linear. The near-zero regularization strength indicates that multicollinearity, while present, is not severe enough to degrade the unregularized least-squares solution at this dataset size. Ridge coefficients are more stable than unregularized polynomial (e.g., S.G coefficient becomes −1318 instead of −16,257), offering a middle ground.

Ridge is less necessary for this problem than for higher-dimensional or more severely collinear datasets. Its primary value lies in coefficient stability; prediction accuracy is not improved over Linear. Ridge may be preferred in organizations that value mathematical robustness even when accuracy is comparable.

5.5 Random Forest Regression Analysis

Random Forest achieved the highest overall R^2 (0.928 on all 108 samples) and lowest overall RMSE (0.356), but this masks a critical deficiency: Test R^2=0.889, the lowest among the four models. This reversal from high training R^2 (0.960) to lower test R^2 is the hallmark of overfitting. With only 75 training samples, an ensemble of 300 trees (even with max_depth=4 regularization) has sufficient capacity to memorize training noise and sample-specific idiosyncrasies rather than learning generalizable patterns. The model thus performs poorly on new data. Increasing max_depth or reducing the number of trees would further impair training fit without guaranteed gains on test performance.

Random Forest is therefore NOT recommended for this application. If ensemble methods are desired, gradient boosting (with early stopping) or stacked linear + polynomial models might offer better generalization.

5.6 Model Selection Recommendation

Based on test-set generalization (the gold standard for production deployment):

- PRIMARY: Linear Regression (Test R^2=0.902). Recommended for initial deployment due to simplicity, interpretability, and stable generalization.
- ALTERNATIVE: Polynomial Regression (Test R^2=0.909). Deploy if marginal accuracy gain justifies added complexity and coefficient instability. Requires automated safeguards.
- SECONDARY: Ridge Regression (Test R^2=0.901). Use if mathematical stability is prioritized and coefficient interpretability is less critical.
- NOT RECOMMENDED: Random Forest (Test R^2=0.889). Evidence of overfitting precludes deployment without major hyperparameter re-tuning and additional validation data.

6. DEPLOYMENT AND MONITORING STRATEGY

For production integration into DCS, APC, or real-time systems, the following practices are recommended:

6.1 Real-Time Bias Monitoring

Track the difference between measured CFR RON and model-predicted RON daily. Trigger model recalibration if cumulative bias (MA 7-day moving average) exceeds ±0.5 RON units.

This safeguard detects process drift or feedstock changes that render the historical model obsolete.

6.2 Extrapolation Flagging

Flag predictions as "OUT OF DOMAIN" when S.G falls outside 0.8377–0.8611 or Aromatics fall outside 81.48–89.96 v%. Display a warning to operators and recommend manual CFR verification before relying on predictions at domain boundaries.

6.3 Periodic Model Recalibration

Refit Linear (or alternate) model quarterly with accumulated new CFR data, incorporating both historical data and recent measurements. This continuous learning approach maintains accuracy as catalysts age, feed composition changes, and operating conditions evolve.

6.4 Uncertainty Quantification

Report prediction uncertainty (e.g., ±1 standard error) alongside point estimates. Based on the test-set RMSE=0.435, a reasonable 95% confidence interval is approximately ±0.85 RON units around predictions.

7. REAL-WORLD VERIFICATION ON PRODUCTION SAMPLES

Following model development and the held-out evaluation reported in Section 4, the recommended Linear soft sensor was subjected to an independent verification campaign against fresh CFR engine measurements collected from routine refinery operation. A total of 36 reformate samples were analyzed: for each, Specific Gravity and Aromatics content were measured and used to generate predictions from all three closed-form models (Linear, Polynomial, and Ridge), and the mean of the three models was also recorded. Each sample was then independently measured on the CFR engine (ASTM D2699), and the measured RON was compared with the predictions. All 36 verification samples fell within the training domain (S.G 0.8377–0.8611; Aromatics 81.48–89.96 v%) and spanned a measured RON range of 103.0–106.5.

7.1 Verification Results

Table 4 lists every verification sample with its measured properties, the prediction from each model, the three-model average, the measured CFR RON, and the error of the recommended Linear model (predicted minus measured). Predictions are reported to two decimal places; the Random Forest model is excluded from this comparison because it has no closed-form equation suitable for real-time deployment.

Table 3Verification of predicted RON against CFR engine measurements (36 samples).

RON Soft Sensor — Verification Against CFR Engine (36 Samples)

Model coefficients (inputs)

#	S.G	Aromatics	CFR RON (actual)	Linear	Polynomial	Ridge	Avg of 3 models	Avg error	Domain
1	0.8577	86.03	105.8	105.19	105.83	105.15	105.39	-0.41	in-domain
2	0.8588	88.26	106.4	106.07	106.46	106.09	106.21	-0.19	in-domain
3	0.8569	87.34	105.9	105.57	105.88	105.53	105.66	-0.24	in-domain
4	0.8571	87.08	105.7	105.50	105.85	105.45	105.60	-0.10	in-domain
5	0.8578	89.46	106.2	106.39	106.79	106.48	106.55	+0.35	in-domain
6	0.8573	88.56	105.6	106.03	106.34	106.05	106.14	+0.54	in-domain
7	0.8567	88.40	105.6	105.92	106.20	105.92	106.01	+0.41	in-domain
8	0.8573	88.24	105.7	105.92	106.22	105.92	106.02	+0.32	in-domain
9	0.8571	88.13	106.5	105.86	106.15	105.86	105.96	-0.54	in-domain
10	0.8560	87.84	105.7	105.66	105.90	105.62	105.73	+0.03	in-domain
11	0.8571	87.96	105.8	105.80	106.09	105.79	105.89	+0.09	in-domain
12	0.8566	86.98	105.5	105.42	105.74	105.36	105.50	+0.00	in-domain
13	0.8523	85.59	104.3	104.53	104.70	104.44	104.55	+0.25	in-domain
14	0.8492	83.62	103.7	103.55	103.81	103.55	103.64	-0.06	in-domain
15	0.8580	89.23	106.3	106.33	106.70	106.41	106.48	+0.18	in-domain
16	0.8484	82.80	103.1	103.19	103.57	103.26	103.34	+0.24	in-domain
17	0.8467	82.76	103.1	103.01	103.29	103.10	103.13	+0.03	in-domain
18	0.8522	85.22	104.0	104.39	104.59	104.31	104.43	+0.43	in-domain
19	0.8517	84.72	104.2	104.17	104.41	104.10	104.23	+0.03	in-domain
20	0.8495	81.95	103.0	103.00	103.79	103.16	103.31	+0.31	in-domain
21	0.8525	85.35	104.2	104.46	104.67	104.38	104.50	+0.30	in-domain
22	0.8505	82.42	103.6	103.26	104.01	103.36	103.54	-0.06	in-domain
23	0.8496	81.95	103.5	103.01	103.81	103.17	103.33	-0.17	in-domain
24	0.8498	82.99	103.3	103.39	103.85	103.44	103.56	+0.26	in-domain
25	0.8509	84.15	103.2	103.89	104.18	103.86	103.98	+0.78	in-domain
26	0.8505	85.35	104.0	104.27	104.40	104.18	104.28	+0.28	in-domain
27	0.8572	87.87	105.4	105.78	106.08	105.76	105.87	+0.47	in-domain
28	0.8577	87.31	105.7	105.63	106.02	105.60	105.75	+0.05	in-domain
29	0.8563	87.13	105.4	105.44	105.72	105.39	105.52	+0.12	in-domain
30	0.8569	87.95	105.8	105.78	106.06	105.76	105.87	+0.07	in-domain
31	0.8581	88.70	106.0	106.15	106.50	106.20	106.28	+0.28	in-domain
32	0.8567	87.93	105.8	105.75	106.03	105.73	105.84	+0.04	in-domain
33	0.8549	87.36	104.9	105.39	105.59	105.32	105.43	+0.53	in-domain
34	0.8564	87.65	105.6	105.63	105.89	105.59	105.70	+0.10	in-domain
35	0.8576	87.95	105.9	105.85	106.17	105.84	105.95	+0.05	in-domain
36	0.8574	87.67	105.8	105.73	106.05	105.71	105.83	+0.03	in-domain

AGGREGATE (n=36)

	Linear	Polynomial	Ridge	Avg of 3
MAE	0.219	0.385	0.220	0.233
RMSE	0.292	0.444	0.282	0.299
Bias	+0.019	+0.365	+0.018	+0.134
Max |err|	0.694	0.979	0.657	0.777

Notes: Linear is the recommended production model. 'Avg of 3' = mean(Linear, Polynomial, Ridge); Random Forest has no closed-form equation and is excluded. Polynomial reads high and is unstable — the 3-model average inherits some of that upward bias, so it is shown for completeness, not as the preferred estimate. Domain: S.G 0.8377–0.8611, Aromatics 81.48–89.96 wt%.

7.2 Aggregate Accuracy

Table 5 summarizes the aggregate error of each model across the 36 verification samples. The recommended Linear model achieved a mean absolute error (MAE) of 0.22 RON with a root-mean-square error (RMSE) of 0.29 RON and a mean bias of +0.02 RON, indicating no meaningful systematic over- or under-prediction. Ridge regression performed almost identically (MAE 0.22, RMSE 0.28). The Polynomial model was the weakest of the three, with MAE 0.39 RON and a pronounced positive bias of +0.37 RON, consistent with the coefficient instability discussed in Section 5.3. Because the three-model average incorporates the Polynomial model, it inherits part of that upward bias (mean bias +0.13 RON) and is therefore slightly less accurate

than the Linear model alone; the average is reported for completeness rather than as the preferred estimate.

Table 4 Aggregate verification error by model (n = 36).

Model	MAE (RON)	RMSE (RON)	Mean Bias (RON)
Linear (recommended)	0.22	0.29	+0.02
Ridge	022	0.28	+0.02
Average of 3 models	0.23	0.30	+0.13
Polynomial	0.39	0.44	+0.37

7.3 Interpretation

The verification campaign confirms, on genuinely independent data, the generalization performance indicated by the held-out test set in Section 4. The Linear model's verification RMSE of 0.29 RON is consistent with—and slightly better than—the test-set RMSE of 0.435 RON, and 33 of the 36 predictions (92%) fell within ±0.5 RON of the measured value. The largest single Linear error was 0.69 RON, occurring at the lower edge of the verified composition range. The near-zero bias across 36 samples is notable: it indicates that the two-parameter model has captured the underlying relationship between S.G, aromatic content, and octane number rather than a dataset-specific artifact, and that no correction factor is required for deployment.
These results should be read with their scope in mind. All verification samples originated from the same representative operating stream and lie within the training domain and a measured RON band of 103.0–106.5. Verification across additional feed streams and toward the extremes of the operating envelope remains ongoing, and CFR measurement repeatability (typically a few tenths of a RON) is comparable to the residual model error at this level of accuracy. The verification nevertheless provides substantially stronger evidence for production readiness than the original held-out evaluation alone.

8. CONCLUSION

Four machine learning regression models (Linear, Polynomial, Ridge, Random Forest) were developed and rigorously compared for predicting Research Octane Number from specific gravity and aromatic content using a dataset of 108 CFR measurements partitioned 70/20/10 into training, test, and validation subsets. Linear regression achieved test-set R^2=0.902 with a simple two-parameter equation (RON = −5.7767 + 94.5712·S.G + 0.3470·Aromatics), demonstrating robust generalization to unseen data. Polynomial regression marginally improved test R^2 to 0.909 but introduced coefficient instability due to multicollinearity among derived terms. Ridge regression (alpha=0.001) offers no accuracy advantage over Linear and is not necessary given the moderate multicollinearity. Random Forest exhibited overfitting (Train R^2=0.960 vs. Test R^2=0.889), disqualifying it from production use.

The recommended Linear model was subsequently subjected to an independent verification campaign against 36 fresh CFR engine measurements from routine operation (Section 7). Across these samples the model achieved a mean absolute error of 0.22 RON and a root-mean-square error of 0.29 RON with a mean bias of +0.02 RON, and 33 of 36 predictions (92%) fell within

±0.5 RON of the measured value. This independent result is consistent with, and marginally better than, the held-out test-set performance, and the near-zero bias indicates that the two-parameter model captures the true density–aromatics–octane relationship rather than a dataset-specific artifact. Ridge regression matched the Linear model in verification (MAE 0.22 RON), while the Polynomial model was the weakest (MAE 0.39 RON, bias +0.37 RON); the three-model average, which includes the Polynomial term, was correspondingly less accurate than the Linear model alone.

Based on this analysis, Linear regression is recommended as the primary production soft sensor, offering the optimal balance of accuracy (R^2=0.902), interpretability, simplicity, and generalization performance. For organizations that prioritize nonlinearity or require marginal accuracy gains, Polynomial regression is a viable alternative provided appropriate domain-checking and coefficient stability safeguards are implemented. Deployment should include real-time bias monitoring, extrapolation flagging, and quarterly model recalibration to maintain accuracy as process conditions evolve.

Future work should focus on: (1) expanding the dataset to include additional fuel properties (olefins, sulfur, H/C ratio, distillation curve) to further improve predictive power; (2) implementing automated drift detection and adaptive model updating; (3) exploring physics-informed machine learning to incorporate thermodynamic constraints; and (4) developing uncertainty quantification methods (e.g., conformal prediction, Bayesian approaches) to provide operators with calibrated confidence intervals on predictions.

REFERENCES

[1] ASTM D2699, “Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel,” ASTM International, 2025.
[2] ASTM D4052, “Standard Test Method for Density and Relative Density of Liquids by Digital Density Meter,” ASTM International, 2023.
[3] ASTM D6839 “Standard test method for quantify saturates, olefins, aromatics, oxygenates, benzene, and toluene in spark-ignition engine fuels using multidimensional gas chromatography,” ASTM International, 2025.
[4] CFR F1/F2 Operation manual
[5] Y. Liu, J. Chen, and X. Zhang, “Machine learning-based soft sensors for chemical processes: A survey,” AIChE Journal, vol. 66, no. 7, pp. e17016, 2020
[6] Prof. K. K. Heterogeneous Catalysis and Catalytic Processes
[7] M. Z. Al-Harbi, “Octane number prediction using nonlinear regression models,” Fuel Processing Technology, vol. 196, pp. 106–115, 2019.
[8] ISO 17025 guidance
[9] S. Al-Mutairi and A. Al-Saif, “Modeling and optimization of catalytic reforming units,” Journal of Petroleum Science and Engineering, vol. 178, pp. 1–12, 2019.
[10] OSIsoft, “PI System: Real-time data infrastructure for industrial operations,” OSIsoft Technical Documentation, 2024.
[11] A. K. Gupta and S. K. Jain, “Soft sensor development for industrial processes: A review,” Chemical Engineering Research and Design, vol. 92, no. 11, pp. 205–228,
[12] S. Kadlec, R. Grbić, and B. L. Grbić, “Soft sensors for process control: A review,” Chemical Engineering & Technology, vol. 34, no. 3, pp. 391–402, 2011.

10. APPENDIX: DATASET SUMMARY AND MODEL DETAILS

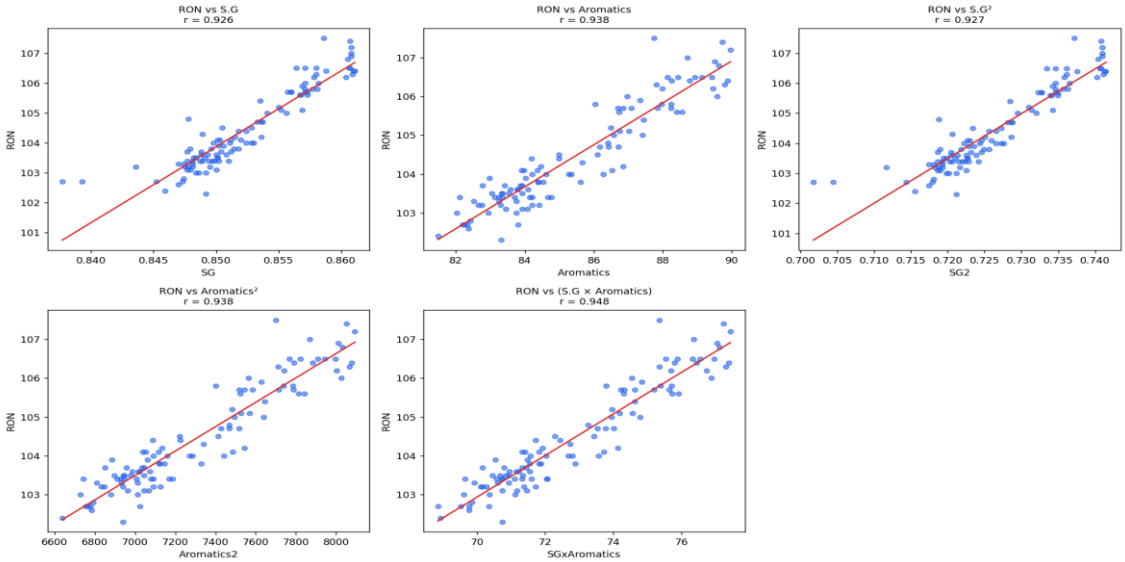
Dataset: 108 unique CFR samples collected over multiple months from refinery operations.
Train-Test-Validate Split: 75 (70%) training, 22 (20%) test, 11 (10%) validation.

Sample #	RON (Actual)	S.G	Aromatics	S.G²	Aromatics²	S.G×Aromatics	Split	Pred (Linear)	Bias (Linear)	Pred (Poly)	Bias (Poly)	Pred (Ridge)	Bias (Ridge)	Pred (RF)	Bias (RF)
1	103.3	0.8474	82.52	0.718087	6809.5504	69.927448	Validate	103	0.3	103.0628	0.2372	103.0937	0.2063	102.8949	0.4051
2	103.4	0.8477	84.14	0.718595	7079.5396	71.325478	Train	103.5905	-0.1905	103.4192	-0.0192	103.5458	-0.1458	103.371	0.029
3	102.8	0.8474	82.42	0.718087	6793.0564	69.842708	Train	102.9653	-0.1653	103.0532	-0.2532	103.0697	-0.2697	102.8846	-0.0846
4	102.7	0.8452	82.25	0.714363	6765.0625	69.5177	Train	102.6982	0.0018	102.6986	0.0014	102.8361	-0.1361	102.8145	-0.1145
5	102.6	0.8470	82.37	0.717409	6784.8169	69.76739	Test	102.9101	-0.3101	102.9805	-0.3805	103.0221	-0.4221	102.8514	-0.2514
6	103.1	0.8479	83.93	0.718934	7044.2449	71.164247	Train	103.5366	-0.4366	103.384	-0.284	103.5051	-0.4051	103.3439	-0.2439
7	102.7	0.8473	82.33	0.717917	6778.2289	69.758209	Train	102.9246	-0.2246	103.028	-0.328	103.0395	-0.3395	102.8514	-0.1514
8	103.9	0.8506	84.03	0.723620	7061.0409	71.475918	Train	103.8266	0.0734	103.7954	0.1046	103.7865	0.1135	103.8251	0.0749
9	103.9	0.8502	82.97	0.722840	6884.0209	70.541094	Train	103.4209	0.4791	103.629	0.271	103.4635	0.4365	103.706	0.194
10	103.5	0.8501	83.93	0.722670	7044.2449	71.348893	Test	103.7446	-0.2446	103.6952	-0.1952	103.7106	-0.2106	103.7383	-0.2383
11	104.1	0.8503	83.98	0.723010	7052.6404	71.408194	Test	103.7809	0.3191	103.7364	0.3636	103.7437	0.3563	103.7992	0.3008
12	103.3	0.8481	83.25	0.719274	6930.5625	70.604325	Validate	103.3195	-0.0195	103.2705	0.0295	103.3401	-0.0401	103.3472	-0.0472
13	102.3	0.8492	83.31	0.721141	6940.5561	70.746852	Test	103.4444	-1.1444	103.4584	-1.1584	103.4575	-1.1575	103.3923	-1.0923
14	103.1	0.8477	83.45	0.718595	6963.9025	70.740565	Train	103.3511	-0.2511	103.2499	-0.1499	103.3562	-0.2562	103.3151	-0.2151
15	103.7	0.8488	83.41	0.720461	6957.2281	70.798408	Train	103.4412	0.2588	103.4043	0.2957	103.4467	0.2533	103.3919	0.3081
16	103.6	0.8500	83.79	0.722500	7020.7641	71.2215	Test	103.6866	-0.0866	103.658	-0.058	103.6621	-0.0621	103.6872	-0.0872
17	104	0.8511	84.62	0.724371	7160.5444	72.020082	Train	104.0786	-0.0786	103.9759	0.0241	104.0066	-0.0066	104.0548	-0.0548
18	104	0.8498	84.21	0.722160	7091.3241	71.561658	Train	103.8134	0.1866	103.6968	0.3032	103.7615	0.2385	103.8082	0.1918
19	103.4	0.8499	84.77	0.722330	7185.9529	72.046023	Test	104.0172	-0.6172	103.8397	-0.4397	103.9361	-0.5361	103.9999	-0.5999
20	103.8	0.8496	84.54	0.721820	7147.0116	71.825184	Train	103.909	-0.109	103.7431	0.0569	103.8388	-0.0388	103.7822	0.0178
21	103.7	0.8477	82.77	0.718595	6850.8729	70.164129	Train	103.1151	0.5849	103.1405	0.5595	103.1817	0.5183	103.2659	0.4341
22	103.2	0.8475	84.21	0.718256	7091.3241	71.367975	Train	103.5959	-0.3959	103.4206	-0.2206	103.5475	-0.3475	103.3095	-0.1095
23	103.3	0.8470	83.75	0.717409	7014.0625	70.93625	Test	103.389	-0.089	103.239	0.061	103.3736	-0.0736	102.9991	0.3009
24	103.2	0.8436	83.3	0.711661	6938.89	70.27188	Train	102.9113	0.2887	102.8491	0.3509	102.9573	0.2427	102.9262	0.2738
25	103.2	0.8481	82.66	0.719274	6832.6756	70.103946	Train	103.1148	0.0852	103.1978	0.0022	103.191	0.009	103.2616	-0.0616
26	103.5	0.8501	83.34	0.722670	6945.5556	70.847334	Train	103.5399	-0.0399	103.628	-0.128	103.55	-0.05	103.6927	-0.1927
27	103.4	0.8502	84.21	0.722840	7091.3241	71.595342	Test	103.8512	-0.4512	103.7577	-0.3577	103.7996	-0.3996	103.9124	-0.5124
28	104.1	0.8501	83.9	0.722670	7039.21	71.32339	Train	103.7342	0.3658	103.6907	0.4093	103.7022	0.3978	103.7383	0.3617
29	103.4	0.8481	82.11	0.719274	6742.0521	69.637491	Train	102.9239	0.4761	103.1687	0.2313	103.0604	0.3396	103.1942	0.2058
30	103	0.8486	82.03	0.720122	6728.9209	69.610658	Train	102.9434	0.0566	103.2748	-0.2748	103.0876	-0.0876	103.1906	-0.1906
31	103.6	0.8488	83.59	0.720461	6987.2881	70.951192	Test	103.5037	0.0963	103.4327	0.1673	103.4948	0.1052	103.377	0.223
32	103.4	0.8487	83.35	0.720292	6947.2225	70.739145	Test	103.4109	-0.0109	103.3795	0.0205	103.4215	-0.0215	103.3696	0.0304
33	103.5	0.8482	83.33	0.719443	6943.8889	70.680506	Train	103.3567	0.1433	103.2985	0.2015	103.3702	0.1298	103.3526	0.1474
34	103.4	0.8494	83.71	0.721480	7007.3641	71.103274	Train	103.6021	-0.2021	103.5464	-0.1464	103.5835	-0.1835	103.4136	-0.0136
35	103	0.8484	82.94	0.719783	6879.0436	70.366296	Train	103.2403	-0.2403	103.2778	-0.2778	103.2881	-0.2881	103.2885	-0.2885
36	103.5	0.8484	83.04	0.719783	6895.6416	70.451136	Train	103.275	0.225	103.2892	0.2108	103.3135	0.1865	103.3406	0.1594
37	103.5	0.8491	83.53	0.720971	6977.2609	70.925323	Validate	103.5112	-0.0112	103.4705	0.0295	103.5066	-0.0066	103.3749	0.1251
38	103.4	0.8489	83.13	0.720631	6910.5969	70.569057	Train	103.3535	0.0465	103.3856	0.0144	103.3827	0.0173	103.391	0.009
39	103.2	0.8478	82.75	0.718765	6847.5625	70.15545	Train	103.1176	0.0824	103.1547	0.0453	103.1859	0.0141	103.2579	-0.0579
40	103.6	0.8493	83.78	0.721310	7019.0884	71.154354	Train	103.6169	-0.0169	103.5419	0.0581	103.5933	0.0067	103.4134	0.1866
41	102.7	0.8393	83.81	0.704424	7024.1161	70.341733	Train	102.6816	0.0184	103.2006	-0.5006	102.7433	-0.0433	102.9247	-0.2247
42	103.2	0.8495	84.42	0.721650	7126.7364	71.71479	Train	103.8579	-0.6579	103.701	-0.501	103.794	-0.594	103.613	-0.413
43	103.4	0.8510	84.67	0.724201	7169.0089	72.05417	Validate	104.0865	-0.6865	103.9699	-0.5699	104.0119	-0.6119	104.0512	-0.6512
44	103	0.8492	83.75	0.721141	7014.0625	71.1205	Train	103.597	-0.597	103.5213	-0.5213	103.5757	-0.5757	103.3736	-0.3736
45	103.1	0.8500	84.07	0.722500	7067.7649	71.4595	Test	103.7838	-0.6838	103.7014	-0.6014	103.7405	-0.6405	103.794	-0.694
46	105	0.8541	86.57	0.729487	7494.3649	73.939437	Test	105.0391	-0.0391	104.8929	0.1071	104.9433	0.0567	104.7276	0.2724
47	104.7	0.8537	86.42	0.728804	7468.4164	73.776754	Train	104.9492	-0.2492	104.7937	-0.0937	104.8495	-0.1495	104.6039	0.0961
48	104.1	0.8520	86.52	0.725904	7485.7104	73.71504	Validate	104.8231	-0.7231	104.6311	-0.5311	104.7123	-0.6123	104.5598	-0.4598
49	104.8	0.8478	86.41	0.718765	7466.6881	73.258398	Train	104.3877	0.4123	104.3901	0.4099	104.2686	0.5314	104.2021	0.5979
50	103.8	0.8479	84.39	0.718934	7121.6721	71.554281	Train	103.6962	0.1038	103.5133	0.2867	103.636	0.164	103.5374	0.2626
51	104	0.8528	86.27	0.727268	7442.5129	73.571056	Train	104.812	-0.812	104.6308	-0.6308	104.7059	-0.7059	104.4067	-0.4067
52	104.5	0.8528	86.1	0.727268	7413.21	73.42608	Train	104.753	-0.253	104.5793	-0.0793	104.6477	-0.1477	104.3871	0.1129
53	104.2	0.8536	86.85	0.728633	7542.0225	74.13516	Train	105.089	-0.889	104.918	-0.718	104.9917	-0.7917	104.7793	-0.5793

Sample #	RON (Actual)	S.G	Aromatics	S.G²	Aromatics²	S.G×Aromatics	Split	Pred (Linear)	Bias (Linear)	Pred (Poly)	Bias (Poly)	Pred (Ridge)	Bias (Ridge)	Pred (RF)	Bias (RF)
55	104.5	0.8505	84.98	0.723350	7221.6004	72.27549	Train	104.1468	0.3532	103.9705	0.5295	104.0578	0.4422	104.1177	0.3823
56	103.6	0.8508	84.1	0.723861	7072.81	71.55228	Train	103.8698	-0.2698	103.8403	-0.2403	103.8257	-0.2257	103.83	-0.23
57	105	0.8556	87.4	0.732051	7638.76	74.77944	Test	105.469	-0.469	105.3557	-0.3557	105.4032	-0.4032	105.5647	-0.5647
58	104.7	0.8533	86.18	0.728121	7426.9924	73.537394	Train	104.8281	-0.1281	104.6699	0.0301	104.7257	-0.0257	104.4594	0.2406
59	104.4	0.8518	84.19	0.725563	7087.9561	71.713042	Train	103.9956	0.4044	104.0376	0.3624	103.9487	0.4513	104.0221	0.3779
60	104.4	0.8524	84.99	0.726586	7223.3001	72.445476	Train	104.33	0.07	104.2595	0.1405	104.2471	0.1529	104.1866	0.2134
61	103.4	0.8497	83.24	0.721990	6928.8976	70.729028	Train	103.4673	-0.0673	103.5428	-0.1428	103.486	-0.086	103.5288	-0.1288
62	102.7	0.8377	82.19	0.701741	6755.1961	68.850563	Train	101.9681	0.7319	102.3746	0.3254	102.2207	0.4793	102.8145	-0.1145
63	102.4	0.8459	81.48	0.715547	6638.9904	68.923932	Validate	102.4972	-0.0972	102.7245	-0.3245	102.7258	-0.3258	102.8199	-0.4199
64	104.1	0.8512	84.37	0.724541	7118.2969	71.815744	Train	104.0013	0.0987	103.9498	0.1502	103.9424	0.1576	103.9928	0.1072
65	104.7	0.8536	86.71	0.728633	7518.6241	74.015656	Test	105.0404	-0.3404	104.8706	-0.1706	104.9415	-0.2415	104.7665	-0.0665
66	103.8	0.8518	84.38	0.725563	7119.9844	71.874884	Train	104.0616	-0.2616	104.0595	-0.2595	104.0039	-0.2039	104.0495	-0.2495
67	103.8	0.8514	85.61	0.724882	7329.0721	72.888354	Train	104.4506	-0.6506	104.2585	-0.4585	104.3452	-0.5452	104.102	-0.302
68	104.3	0.8489	85.67	0.720631	7339.3489	72.725263	Test	104.235	0.065	104.0565	0.2435	104.1238	0.1762	103.8039	0.4961
69	104	0.8524	85.27	0.726586	7270.9729	72.684148	Train	104.4272	-0.4272	104.3131	-0.3131	104.3349	-0.3349	104.1687	-0.1687
70	103.7	0.8504	83.88	0.723180	7035.8544	71.331552	Validate	103.7556	-0.0556	103.7399	-0.0399	103.7252	-0.0252	103.7684	-0.0684
71	103.7	0.8512	83.93	0.724541	7044.2449	71.441216	Train	103.8487	-0.1487	103.8942	-0.1942	103.8162	-0.1162	103.8198	-0.1198
72	105.1	0.8552	86.75	0.731367	7525.5625	74.1886	Train	105.2056	-0.1056	105.1069	-0.0069	105.122	-0.022	105.2538	-0.1538
73	105.2	0.8550	86.5	0.731025	7482.25	73.9575	Train	105.0999	0.1001	105.012	0.188	105.0122	0.1878	104.984	0.216
74	105.1	0.8569	87.01	0.734278	7570.7401	74.558869	Train	105.4566	-0.3566	105.4624	-0.3624	105.3976	-0.2976	105.5225	-0.4225
75	105.4	0.8535	87.44	0.728462	7645.7536	74.63004	Train	105.2842	0.1158	105.1359	0.2641	105.1986	0.2014	105.1544	0.2456
76	104	0.8530	85.35	0.727609	7284.6225	72.80355	Train	104.5117	-0.5117	104.4285	-0.4285	104.4206	-0.4206	104.1802	-0.1802
77	105.7	0.8560	86.7	0.732736	7516.89	74.2152	Validate	105.2639	0.4361	105.2304	0.4696	105.1888	0.5112	105.2812	0.4188
78	105.7	0.8557	86.85	0.732222	7542.9225	74.317545	Validate	105.2876	0.4124	105.2135	0.4865	105.2109	0.4891	105.352	0.348
79	106.5	0.8580	88.44	0.736164	7821.6336	75.88152	Train	106.0568	0.4432	106.0412	0.4588	106.0702	0.4298	106.2747	0.2253
80	105.8	0.8581	88.23	0.736336	7784.5329	75.710163	Test	105.9934	-0.1934	105.9885	-0.1885	105.9967	-0.1967	106.2059	-0.4059
81	106	0.8582	89.58	0.736507	8024.5764	76.877556	Test	106.4714	-0.4714	106.5269	-0.5269	106.5722	-0.5722	106.5884	-0.5884
82	106.3	0.8580	87.81	0.736164	7710.5961	75.34098	Test	105.8382	0.4618	105.8521	0.4479	105.8199	0.4801	106.0534	0.2466
83	106.8	0.8605	89.62	0.740460	8031.7444	77.11801	Train	106.7028	0.0972	106.8315	-0.0315	106.8513	-0.0513	106.6653	0.1347
84	107.4	0.8607	89.73	0.740804	8051.4729	77.230611	Test	106.7599	0.6401	106.9025	0.4975	106.9228	0.4772	106.6802	0.7198
85	107	0.8608	88.71	0.740977	7869.4641	76.361568	Train	106.4153	0.5847	106.6069	0.3931	106.4976	0.5024	106.6257	0.3743
86	106.4	0.8610	88.78	0.741321	7881.8884	76.43958	Train	106.4585	-0.0585	106.6642	-0.2642	106.5498	-0.1498	106.6372	-0.2372
87	106.5	0.8606	89.42	0.740632	7995.9364	76.954852	Train	106.6428	-0.1428	106.7766	-0.2766	106.7756	-0.2756	106.5945	-0.0945
88	106.4	0.8611	89.88	0.741493	8078.4144	77.395668	Train	106.8497	-0.4497	107.0195	-0.6195	107.036	-0.636	106.7776	-0.3776
89	106.3	0.8609	89.8	0.741149	8064.04	77.30882	Test	106.8031	-0.5031	106.959	-0.659	106.9771	-0.6771	106.7776	-0.4776
90	107.5	0.8586	87.75	0.737194	7700.0625	75.34215	Validate	105.8741	1.6259	105.9484	1.5516	105.8632	1.6368	106.0943	1.4057
91	106.5	0.8564	89.13	0.733421	7944.1569	76.330932	Train	106.145	0.355	106.1514	0.3486	106.1804	0.3196	106.1438	0.3562
92	106.5	0.8607	88.93	0.740804	7908.5449	76.542051	Train	106.4822	0.0178	106.6441	-0.1441	106.5779	-0.0779	106.5519	-0.0519
93	107.2	0.8608	89.96	0.740977	8092.8016	77.437568	Train	106.8491	0.3509	107.0056	0.1944	107.0368	0.1632	106.7776	0.4224
94	106.2	0.8604	87.98	0.740288	7740.4804	75.697992	Train	106.1242	0.0758	106.371	-0.171	106.1562	0.0438	106.21	-0.01
95	105.6	0.8568	86.73	0.734106	7522.0929	74.310264	Train	105.3499	0.2501	105.3875	0.2125	105.2852	0.3148	105.3992	0.2008
96	105.8	0.8577	86.03	0.735649	7401.1609	73.787931	Validate	105.1921	0.6079	105.5059	0.2941	105.1375	0.6625	104.9156	0.8844
97	106.4	0.8588	88.26	0.737537	7789.8276	75.797688	Train	106.07	0.33	106.114	0.286	106.0866	0.3134	106.2942	0.1058
98	105.9	0.8569	87.34	0.734278	7628.2756	74.841646	Train	105.5711	0.3289	105.5396	0.3604	105.5201	0.3799	105.6688	0.2312
99	106	0.8571	86.97	0.734620	7563.7809	74.541987	Train	105.4616	0.5384	105.4922	0.5078	105.4046	0.5954	105.6906	0.3094
100	105.7	0.8571	87.08	0.734620	7582.9264	74.636268	Train	105.4998	0.2002	105.5149	0.1851	105.445	0.255	105.7648	-0.0648
101	106.2	0.8578	89.46	0.735821	8003.0916	76.738788	Test	106.3919	-0.1919	106.4316	-0.2316	106.4755	-0.2755	106.4754	-0.2754
102	105.6	0.8573	88.56	0.734963	7842.8736	75.922488	Train	106.0323	-0.4323	105.9897	-0.3897	106.0418	-0.4418	106.0914	-0.4914
103	105.6	0.8567	88.4	0.733935	7814.56	75.73228	Train	105.92	-0.32	105.8546	-0.2546	105.9113	-0.3113	105.8127	-0.2127
104	105.7	0.8573	88.24	0.734963	7786.2976	75.648152	Train	105.9212	-0.2212	105.8724	-0.1724	105.9125	-0.2125	105.9534	-0.2534
105	106.5	0.8571	88.13	0.734620	7766.8969	75.536223	Train	105.8642	0.6358	105.8066	0.6934	105.8469	0.6531	105.9683	0.5317
106	105.7	0.8560	87.84	0.732736	7715.8656	75.19104	Train	105.6595	0.0405	105.5583	0.1417	105.614	0.086	105.646	0.054
107	105.8	0.8571	87.96	0.734620	7736.9616	75.390516	Test	105.8052	-0.0052	105.7501	0.0499	105.7798	0.0202	105.9385	-0.1385

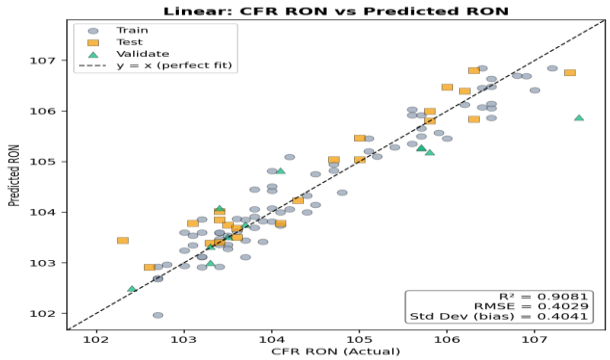
Correlation Coefficients with RON (Pearson r):

- S.G×Aromatics: 0.9481 (highest)
- Aromatics²: 0.9381
- Aromatics: 0.9375
- S.G²: 0.9271
- S.G: 0.9263 (lowest)

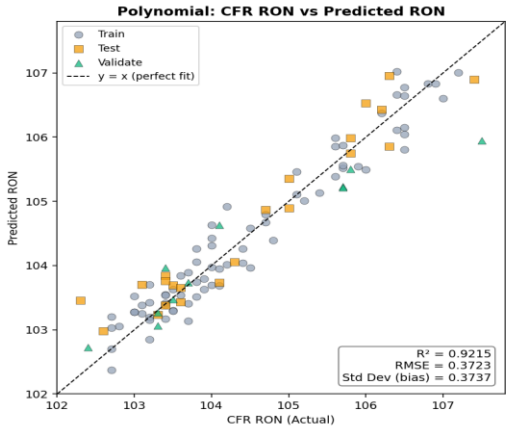


Model Performance Summary:

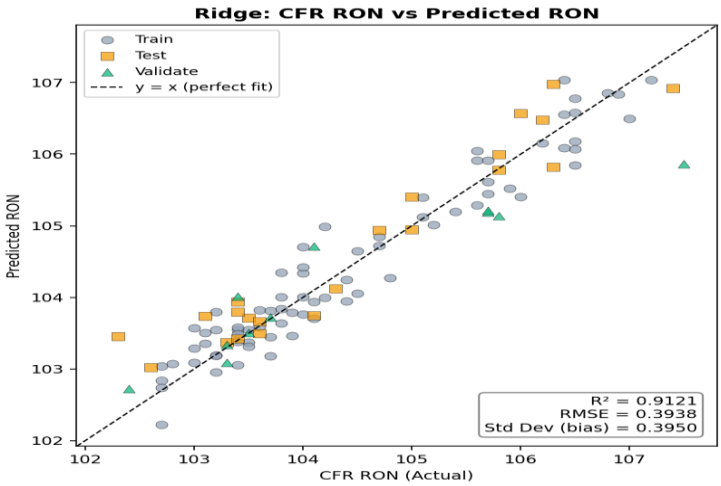
- Linear: Test R^2=0.902, RMSE=0.435, Bias Std Dev=0.404



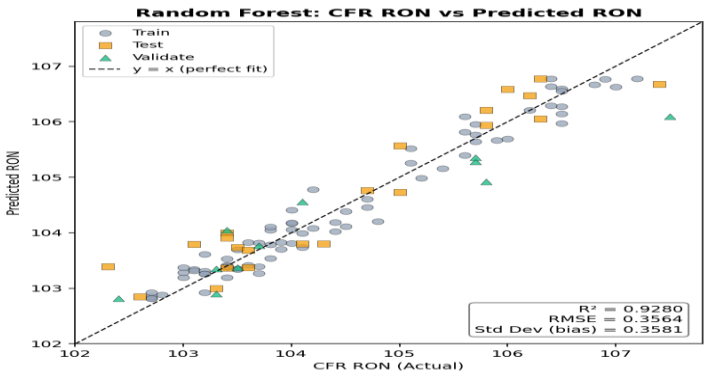
- Polynomial: Test R^2=0.909, RMSE=0.419, Bias Std Dev=0.374



- Ridge: Test R^2=0.901, RMSE=0.436, Bias Std Dev=0.395



- Random Forest: Test R^2=0.889, RMSE=0.463, Bias Std Dev=0.358



Model Performance Summary:

- Linear: Test R^2=0.902, RMSE=0.435, Bias Std Dev=0.404
- Polynomial: Test R^2=0.909, RMSE=0.419, Bias Std Dev=0.374
- Ridge: Test R^2=0.901, RMSE=0.436, Bias Std Dev=0.395
- Random Forest: Test R^2=0.889, RMSE=0.463, Bias Std Dev=0.358
- Recommendation: Deploy Linear regression for primary production soft sensor. Consider
- Polynomial as alternative if nonlinearity is operationally important and safeguards are in place.

RON Prediction — Updated Dataset (n=108) — Model Comparison								
Split sizes: Train 75 (~70%) | Test 22 (~20%) | Validate 11								
Linear	RON = -5.7767 + 94.5712·(S.G) + 0.3470·(Aromatics)							
Polynomial	RON = 5704.3459 - 16257.3442·(S.G) + 29.3558·(Aromatics) + 11957.6264·(S.G²) + 0.0617·(Aromatics²) - 46.5298·(S.G×Aromatics)							
Ridge (raw-scale, alpha=0.001)	RON = 717.6799 - 1318.0546·(S.G) - 2.5243·(Aromatics) + 798.3877·(S.G²) + 0.01334·(Aromatics²) +							
Random Forest	300 trees, max_depth=4, min_samples_leaf=3 (no closed-form equation)							
Split	**Linear R²**	**Poly R²**	**Ridge R²**	**RF R²**	**Linear RMSE**	**Poly RMSE**	**Ridge RMSE**	**RF RMSE**
Train	0.9285	0.9406	0.9354	0.9602	0.3445	0.314	0.3276	0.257
Test	0.9016	0.9088	0.901	0.8885	0.4348	0.4185	0.436	0.4628
Validate	0.8118	0.8433	0.8098	0.8294	0.6375	0.5818	0.641	0.607
Overall (all data)								
Model	**R² (all)**	**RMSE**	**Std Dev (bias)**					
Linear	0.9081	0.4029	0.4041					
Polynomial	0.9215	0.3723	0.3737					
Ridge	0.9121	0.3938	0.395					
Random Forest	0.928	0.3564	0.3581					
Correlation with RON								
S.G×Aromatics	0.948							
Aromatics²	0.938							
Aromatics	0.938							
S.G²	0.927							
S.G	0.926							
Notes								
With the expanded dataset (n=108, wider RON/S.G/Aromatics range), correlations strengthened substantially (r = 0.93-0.95 vs 0.79-0.88 previously), and all four models now generalize well (Test R² = 0.89-0.91). Ridge's cross-validated alpha collapsed to a near-zero value, indicating multicollinearity is less limiting with this larger, wider-ranging dataset. Polynomial and Random Forest now slightly edge out Linear on Train/Test R², though Linear remains simplest and most stable into the Validate split.								

AUTHOR

Abbas Ahmed Mohamed Alblola is a Lead Laboratory Specialist at YASREF (Yanbu Aramco Sinopec Refining Company), Saudi Arabia. He holds a Bachelor's degree with honor in Chemistry and brings nearly 20 years of experience in oil and gas refining operations. His expertise spans fuel quality analysis, CFR engine testing, process analytics, and machine learning applications for soft sensor development. Abbas has published research on octane prediction and best practices of CFR engine maintenance and troubleshooting , hydrocarbon composition analysis, and refinery process optimization.

©2026 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.