Report on Screening task for Autumn Internship under AI/ML Surrogate Modelling for Binary Distillation

Task: Build and compare ML surrogates for distillate purity & energy

Flowsheet and Simulation setup details

1. **Simulation Environment and Thermodynamic Model**
   a. **Process Simulator:** The Flowsheet was designed and simulated using DWSIM, an open-source chemical process simulator.
   b. **Property package:** The **NRTL (Non-Random Two Liquid)** activity coefficient model was selected as the thermodynamic property package.
      i. **Justification**: The selected Ethanol & Water system is a highly non-ideal aqeous mixture that forms a minimum boiling azeotrope at approximately 95.6% mass fraction of ethanol.

2. **Flowsheet Components**
   The flowsheet consists of one primary unit operation and its associated material and energy streams:
   a. **Unit Operation:** A single rigorous **Distillation Column** (DCOL-1).
   b. **Material Streams**:
      i. feed : An inlet stream representing the Ethanol Water mixture fed to the column
      ii. 3: The top outlet stream, representing the high-purity **Distillate** product.
      iii. 4: The bottom outlet stream, representing **Bottoms** product.
   c. **Energy Streams**:
      a. E1: The energy removed from the system by the condenser, representing the **Condenser Duty.**
      b. E2: The energy supplied to the system by the reboiler, representing the **Reboiler Duty (QR).** This was a key target variable for the surrogate model.

3. **Distillation Column (DCOL-1) Configuration**
   The distillation column was configured with the following parameters:
   a. **Number of Stages:** 20. Including Total Condenser (Stage 1) and Kettle Reboiler (Stage 20).
   b. **Feed Stage:** The feed stream was introduced at **stage 10**, the following represents a mid-column feed point.
   c. **Column Pressure:** The column was modelled to operate under an **isobaric assumption** at an atmospheric pressure.
      i. Condenser/Top Pressure: **1 atm** (as instructed in the pdf)
      ii. Column Pressure Drop: **0 atm**
   d. Column Specifications (Operating Variables):
      i. **Condenser Specifications**: The primary control variable for the top of the column was the **Reflux Ratio (R).** This was varied systematically from **0.8** to **5.0** during the process of data generation.
      ii. **Reboiler Specifications:** The primary control variable for the bottom of the column was the **Boilup Ratio (B).** This was varied systematically from **1.0** to **3.0** during data generation.

4. **Feed Stream (feed) Conditions**

The feed stream was defined with the following properties:

a. **Composition (Mole Fraction, xF):** This was a key independent variable. The mole fraction of Ethanol was systematically varied through the set **[0.2,0.3,0.4,0.5,0.65,0.7,0.8,0.95]** to generate a comprehensive dataset.

b. **Flow Rate (F):** A constant base case molar flow rate was used for all simulations.

c. **Thermal Condition:** The feed was specified at **1 atm** and **80 degree Celsius**. This resulted in a two phase (Vapor-Liquid) feed, which is a realistic condition for a distillation column.

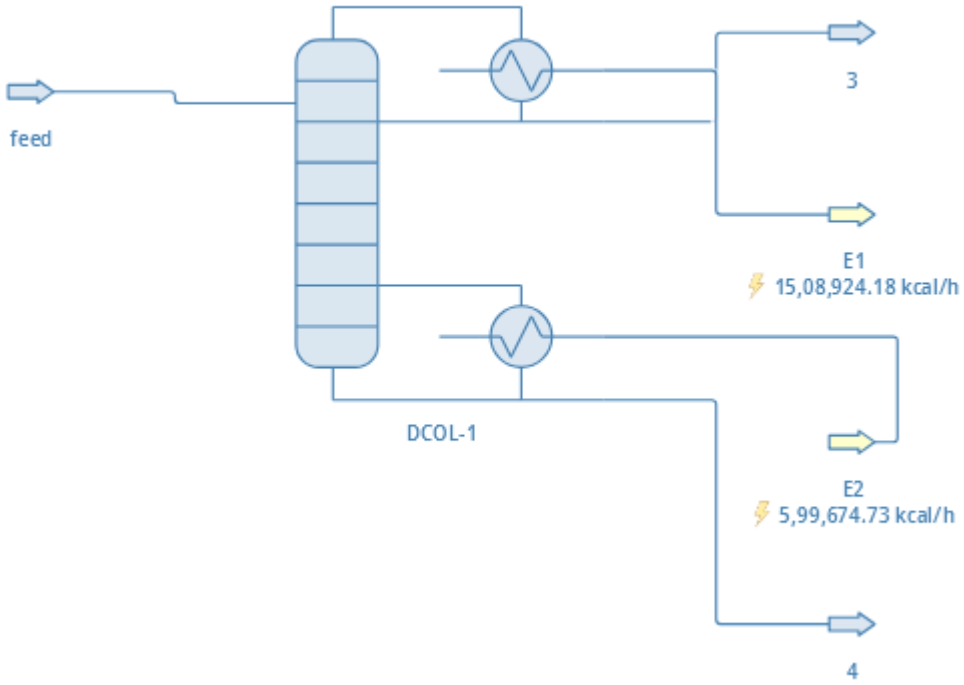
5. **Solver Configuration**

To ensure robust and reliable convergence of the simulation across the wide range of operating conditions:

a. **Column Solver:** The **Wang-Henke (Bubble Point)** algorithm was used.

b. **Initial Estimates Provider:** The default provider was changed to **Internal 2 (Experimental).** This was a critical step taken after initial convergence failures, demonstrating a necessary adjustment to handle the non-ideal thermodynamics of the following system.

Using the following described Flowsheet and Simulation Setup in DWSIM Software, after simulating the Binary Distillation process, 392 data points were generated in total. Following data points were used for the further Machine Learning task.



Screenshot of the Flowsheet constructed in DWSIM simulator

Data Generation and Analysis

Variable Ranges & Data Generation Protocol

Inorder to create a comprehensive dataset for surrogate modelling, the key input variables of the DWSIM simulation were systematically varied across their typical operating ranges. The data generation protocol was as follows:

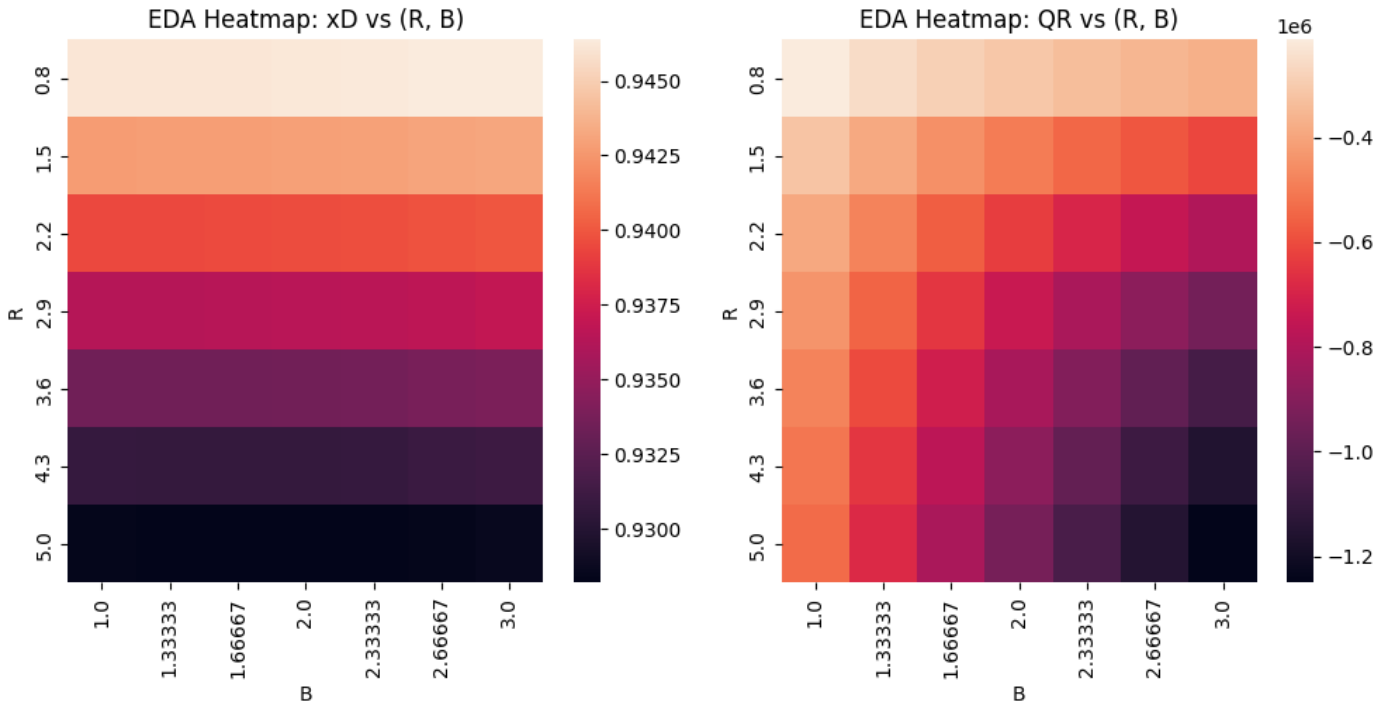
a. **Independent Variables:**
   i. **Feed Mole Fraction of Ethanol (xF):** Varied through the discrete set: [0.2,0.3,0.4,0.5,0.65,0.7,0.8,0.95].
   ii. **Reflux Ratio (R):** Varied from 0.8 to 5.0 in 7 steps.
   iii. **Boilup Ratio (B):** Varied from 1.0 to 3.0 in 7 steps.

b. **Generation Protocol:**
   Due to a two variable limitation in DWSIM sensitivity study tool, I had to employ a semi-automated protocol. The feed composition (xF) was set manually. For each fixed xF, a full 2D sensitivity study (7x7 = 49 simulations) was executed by varying R and B. The resulting data was exported to a separate CSV file. This process was repeated for all 7 values of xF.

c. **Final Dataset:**
   The 7 individual CSV files were then programmatically combined into a single master dataset, with a new column for xF added to preserve the feed condition for each run. The final dataset consists of **392 unique data points.**

Exploratory Data Analysis (EDA)

An exploratory analysis was performed to understand the underlying physics and complexity of the dataset. Heatmaps of the outputs (xD and QR) versus the primary inputs (R and B) reveal a highly non-linear system. The xD(purity) heatmap, in particular, demonstrates that achieving high purity is most sensitive to changes in R and B at lower concentrations, with diminishing returns as the system approaches the azeotrope.



The Extrapolation Test: Block Split Strategy

In order to evaluate the model's ability to generalize beyond their training domain, a critical requirement for a reliable surrogate – a challenging **block based splitting strategy** was implemented. Instead of random split, the data was partitioned based on the feed mole fraction (xF):

a. **Training Set:** All data points where xF < 0.8 (294 samples)
b. **Test Set:** All data points where xF >= 0.8 (98 samples)

This forces the following models to train on the lower to mid concentration range and then **extrapolate** into the high concentration region, which exhibits different physical behaviour as it nears the ethanol water azeotrope.

Modelling & Evaluation Framework

Models Evaluated

Three distinct ML models were developed to predict distillate purity (xD) and reboiler duty (QR):

a. **Polynomial Regression:** A polynomial model was used as a statistical baseline. The polynomial degree was treated as a hyperparameter. Bayesian Optimization determined that **degree 2** provided the best fit for both xD and QR targets.
b. **XGBoost:** A gradient boosted tree-based ensemble model
c. **Artificial Neural Network (ANN):** A feedforward neural network with two hidden layers (64 neurons and 32 neurons , ReLU activation).

Hyperparameter Tuning & Training

Model hyperparameters were tuned using Bayesian Optimization to minimize Mean Squared Error. For each algorithm, two separate models were trained: one to predict xD and another for QR. After the data split, input features were scaled using a StandardScaler that was fit exclusively on the training data to prevent data leakage.

The Feature Engineering Experiment

To investigate the impace of physics informed features, two distinct experiments were conducted:

a. **"Bare" Model Experiment:** Models were trained using only the three raw inputs (xF, R, B).
b. **"Featured" Model Experiment:** Models were trained on an expanded set of eight features, including the raw inputs plus physics informed features like an approximated relative volatility (alpha_approx) and a dimensionless reflux factor (R_factor).

Results & Diagnostic Analysis

Statistical Performance:

When evaluated on the extrapolation test set (xF>=0.8), all models demonstrated a critical failure in predicting distillate purity (xD), as evidenced by highly negative R2 scores. The R2 scores for QR remained strong, as it is a less complex, more linear function of the inputs.

Model & Target	Experiment	R2 score	MSE	MAE	Bounds Violations
ANN - xD	**Bare**	**-0.32**	0.0055	0.056	0
ANN - xD	Featured	-2.09	0.0128	0.093	49
XGB - xD	**Bare**	**-0.80**	0.0075	0.062	0
XGB - xD	Featured	-0.71	0.0071	0.061	0
POLY - xD	**Bare**	**-53.84**	0.2274	0.413	77
POLY - xD	Featured	-4802145.5	19907.5	100.7	98

Model & Target	Experiment	R2 score	MSE	MAE
ANN - QR	**Bare**	**0.93**	**6179022336.0**	**65491.91**
ANN - QR	Featured	0.63	35648712704.0	158133.31
XGB - QR	**Bare**	**0.88**	**11445347328.0**	**84370.96**
XGB - QR	Featured	0.85	14313678848.0	96876.2
POLY - QR	**Bare**	**0.74**	**24979406362.6**	**145085.4**
POLY - QR	Featured	-4985519.5	4.82	492063910.4

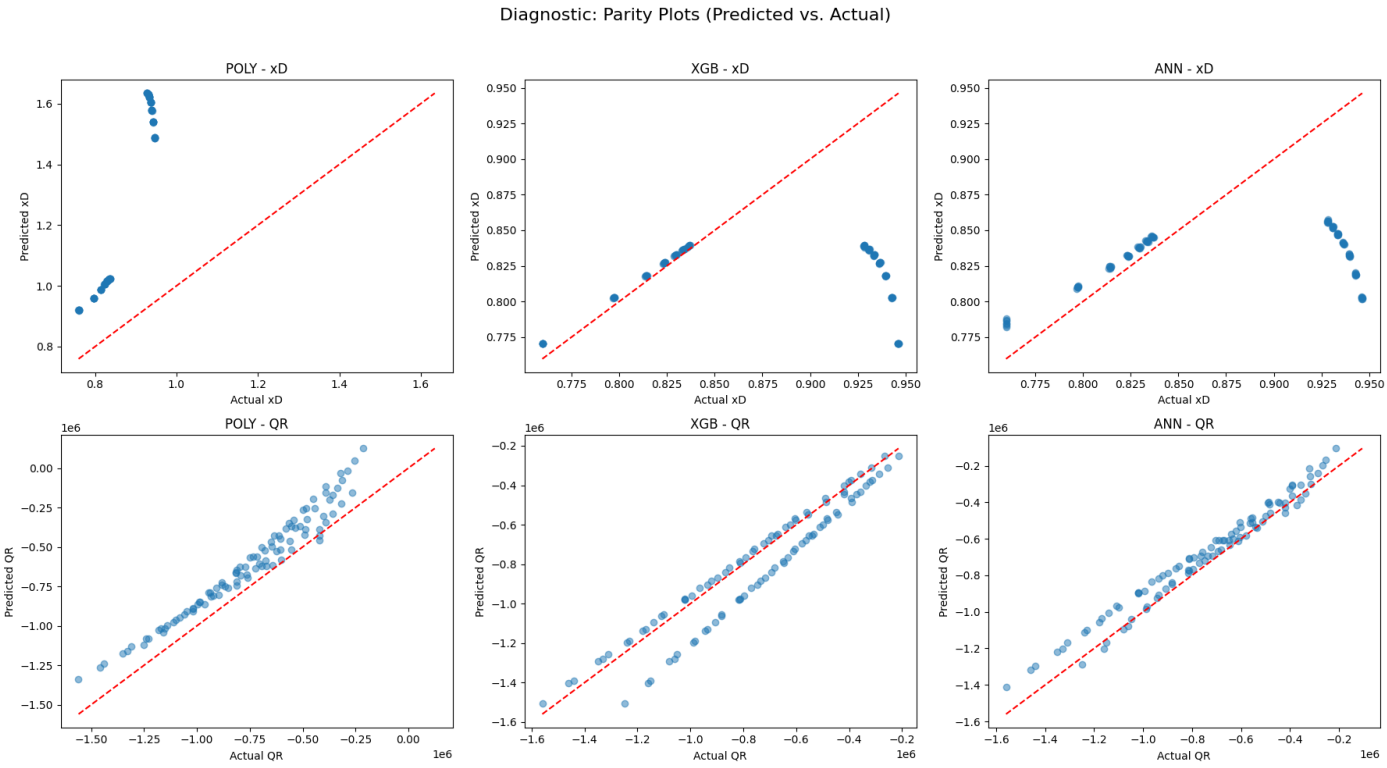
Key finding: The introduction of feature engineering, contrary to initial expectations, **significantly worsened the extrapolation performance for the ANN and Polynomial models for xD**. While the XGBoost scores are comparable, the "Bare" ANN model was substantially less wrong than its "Featured" counterpart. This counter-intuitive result is the central finding of this investigation.

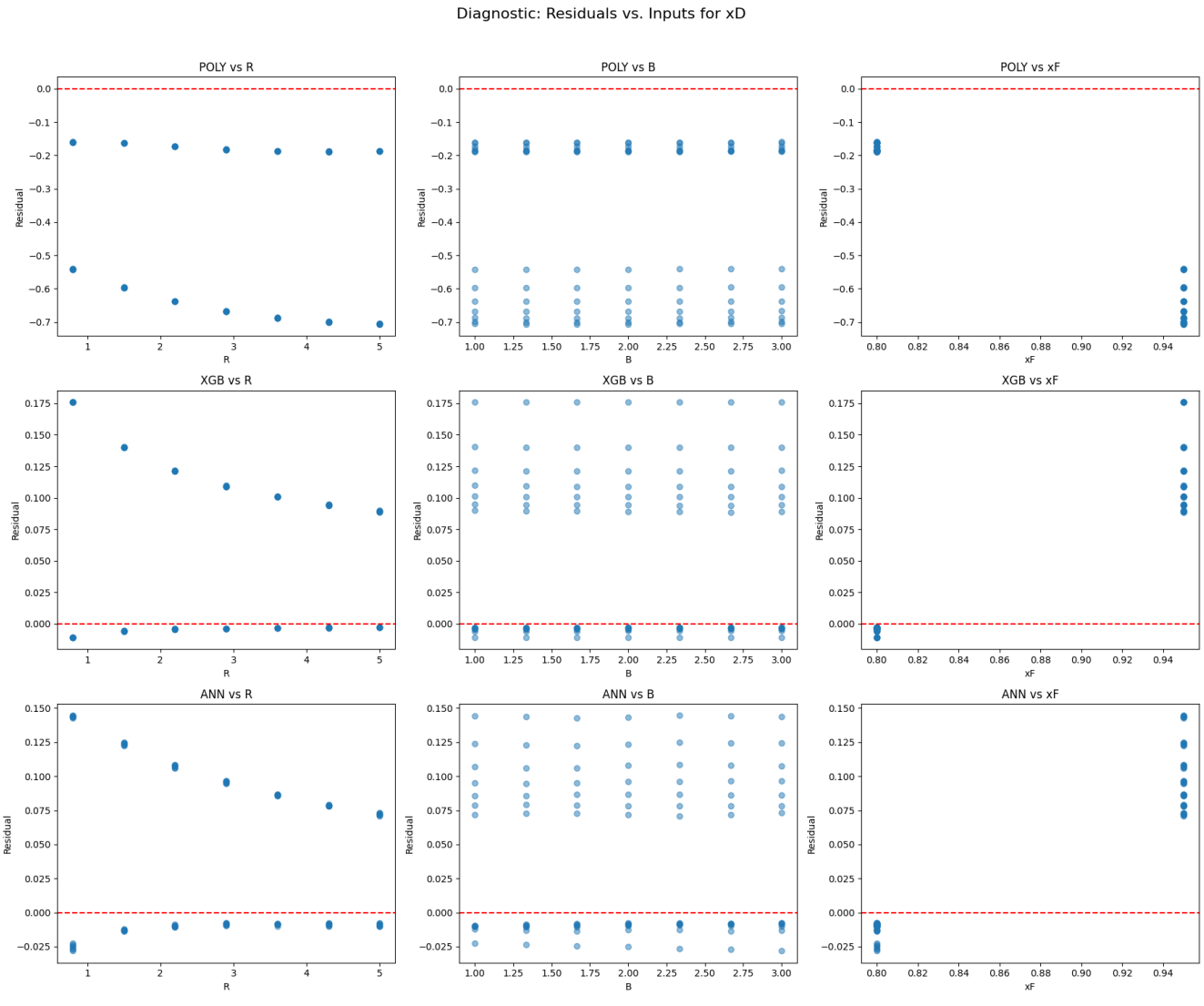
a. **Purity (xD) Prediction: Complete Failure.**
   All models failed to predict xD (negative R2 Scores), providing they cannot extrapolate.
b. **Energy (QR) Prediction: Success (with a catch).**
   The "Bare" models (without extra features) successfully predicted QR. The ANN was the best (R2 = 0.93).
c. **Feature Engineering was Detrimental**
   Adding features made xD predictions significantly worse and ruined the accurate QR predictions. The simpler "Bare" models were universally better.

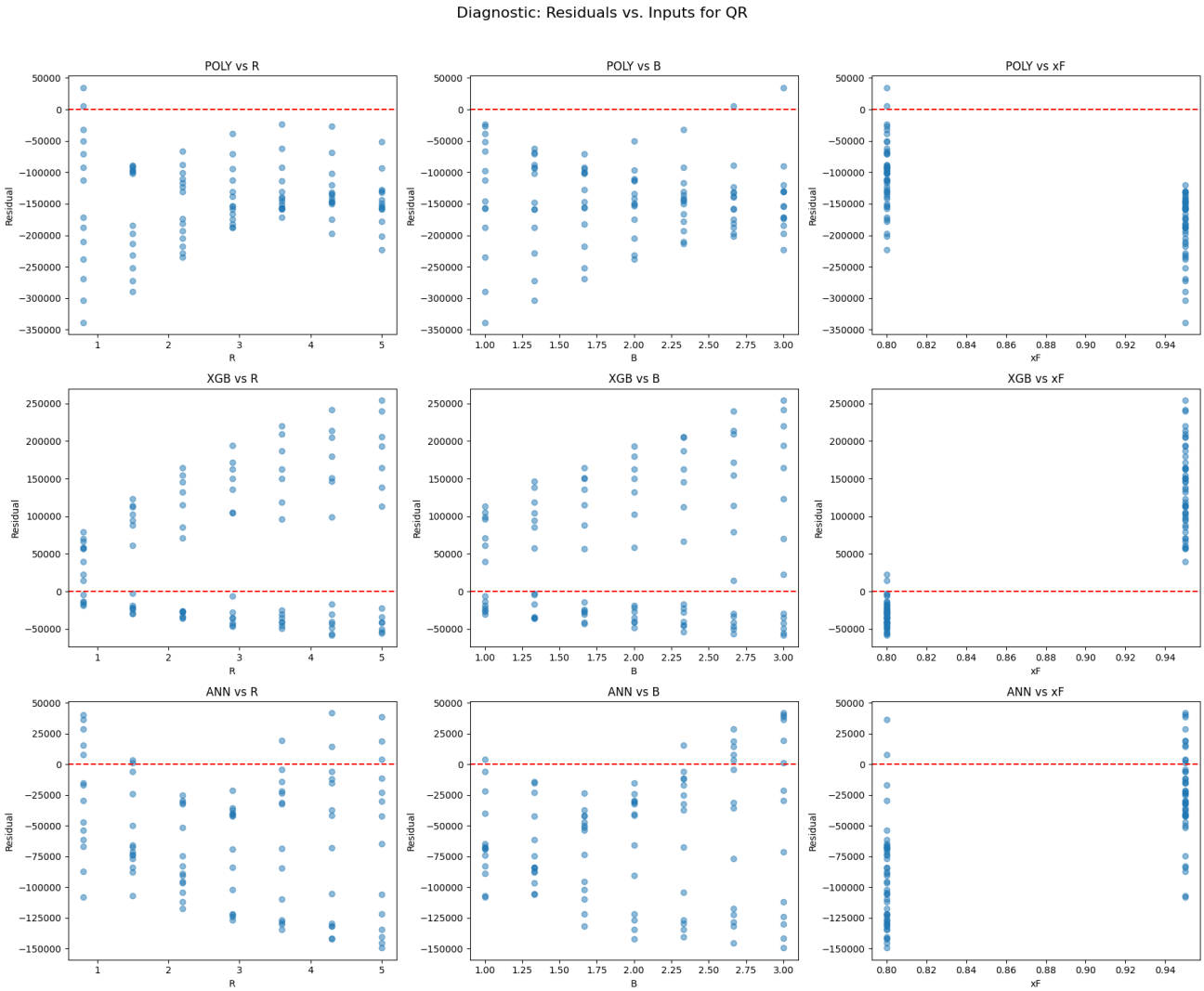
The noteworthy paradox arises from these metrics, highlighting the importance of interpreting R2 and error (MAE/MSE) in the context of the target variable's physical scale. The xD models, despite their negative R2 (e.g., MAE of ~0.06), meaning their predictions are physically close to the true values. The negative R2 reveals that even these small errors are larger than the tiny variance of the high-purity test data, indicating a complete failure to capture the underlying trend. Conversely, the successful QR models have high R2 scores but very large absolute errors (e.g., MAE>65,000). This is because, I believe the error is small relative to the massive scale and variance of reboiler duty. The high R2 correctly shows the QR model has learned the process relationship, while the negative R2 for xD correctly flags that its model has not.

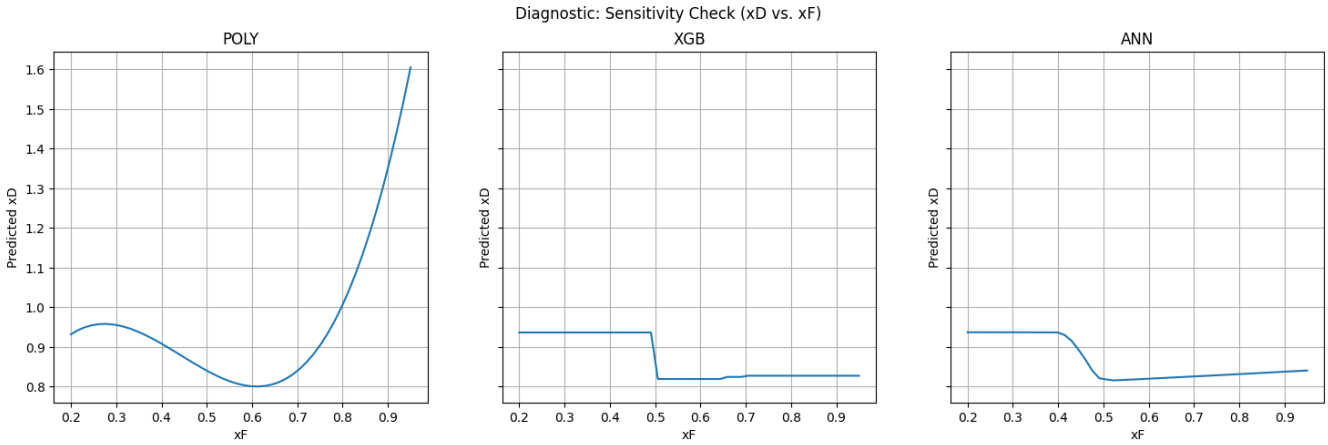
Root Cause Analysis

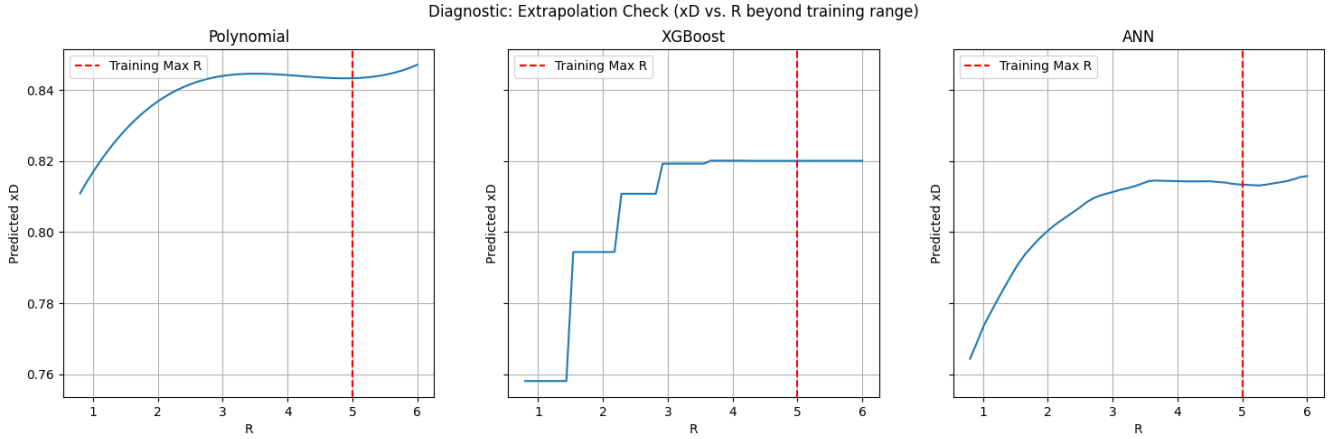
The diagnostic plots of Base case-

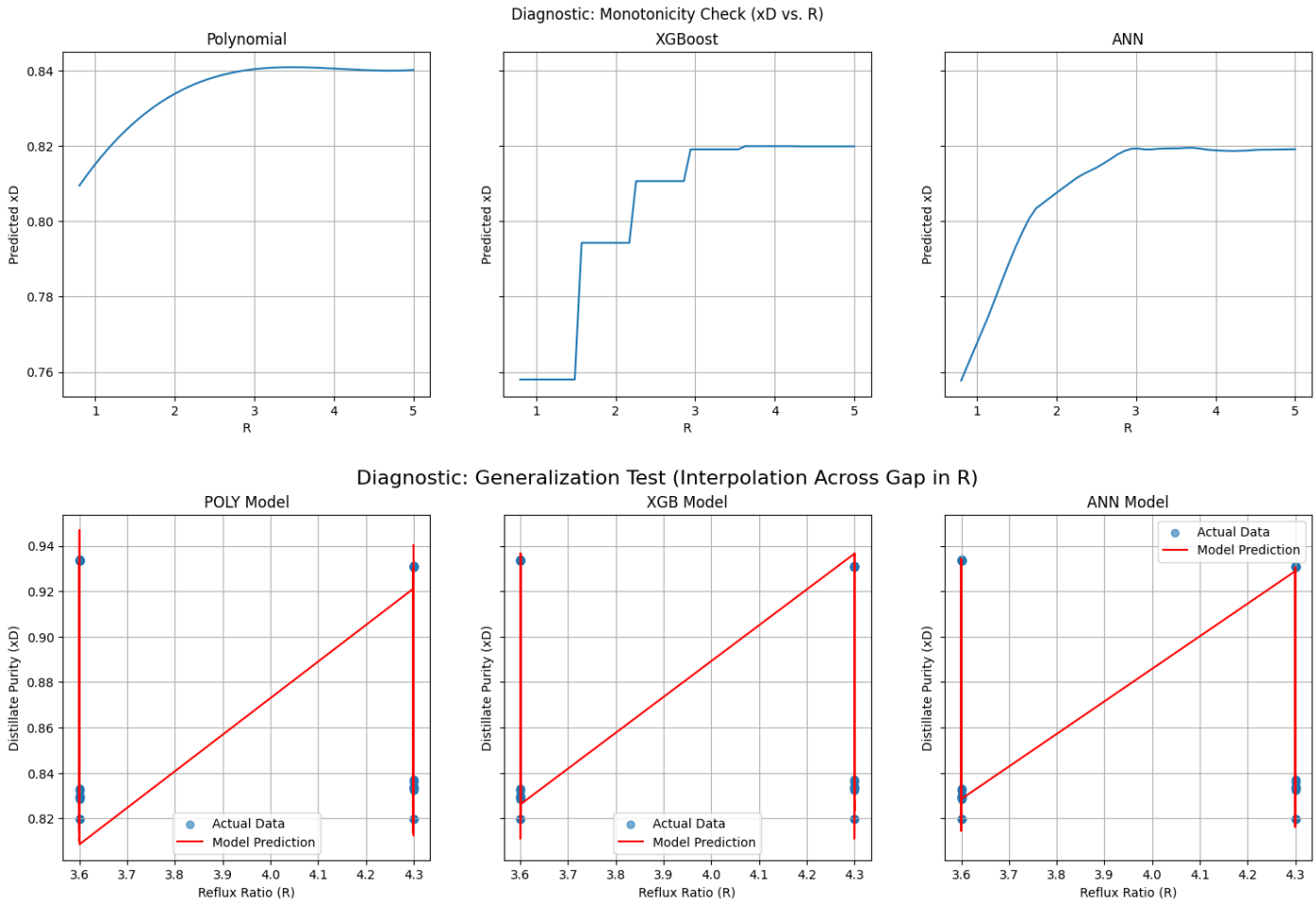





Diagnostic: Residuals vs. Inputs for QR
POLY vs R
POLY vs B
POLY vs xF
XGB vs R
XGB vs B
XGB vs xF
ANN vs R
ANN vs B
ANN vs xF
Residual
R
B
xF


Diagnostic: Sensitivity Check (xD vs. xF)
POLY
XGB
ANN
Predicted xD
xF


Diagnostic: Extrapolation Check (xD vs. R beyond training range)
Polynomial
XGBoost
ANN
Training Max R
Predicted xD
R



a. **Parity & Residual Plots:** The following plots visually confirm the poor performance on xD. The data points for the test set deviate significantly from the ideal y=x line, and the residuals show clear, structured error patterns, proving the models have not learned the correct physical relationship in the extrapolation region.
b. **Extrapolation & Sensitivity Checks:** The Sensitivity check (xD vs xF) and Extrapolation Check (xD vs R) plots are the most crucial evidence.
   i. **Polynomial:** This model exhibits classic instability with predictions plunging to physically impossible negative values. It is unreliable.
   ii. **XGBoost:** As a tree based model, it is algorithmically incapable of predicting values outside the target range seen during training. The plots show its predictions "flat-lining" at the maximum purity value it learned from the xF<0.8 data. It can be classified as failure.
   iii. **ANN:** The ANN learns a spurious, non-physical trend and continues it into the extrapolation region.

Conclusion

Analysis of Model Trade-offs

While the statistical evaluation revealed that all models failed to accurately extrapolate distillate purity (xD), a deeper analysis of the diagnostic plots reveals crucial difference in their model of failure. This is the most important factor in selecting a surrogate model for engineering appliactions.

a. The **Polynomial Model** is rejected completely. It is statistically and physically poor, exhibiting instability outside its training domain.

b. The **ANN**, despite its high accuracy on QR, learned non-physical relationship for xD. The monotonicity plot shows it incorrectly predicting that purity decreases at high reflux. This makes it fundamentally untrustworthy, using this model could make incorrect decisions.
c. The **XGBoost model** , while also inaccurate in extrapolation region, was the only model to demonstrate the **robustness & physical consistency.** It passed the monotonicity check, and its extrapolation behaviour was a safe, predictable saturation. It did not invent non-physical trends.

Final Recommendation: XGBoost

Despite its poor R2 score on extrapolation test, **XGBoost is a trustworthy engineering surrogate**. The justification is based on **engineering reliability.** The XGBoost model was the only one to demonstrate both high accuracy on the simpler QR target and consistent adherence to the physical laws of the process for the more complex xD target. Its inherent structure makes it robust against learning untrustworthy correlations.

Therefore, while its quantitative predictions for xD in the xF >= 0.8 region shouldn't be trusted, its qualitative understanding of the process is sound. It is reliable interpolator & a safe, predictable extrapolator.

Name: Shivaprasad B. Gowda

Contact: 9113299141

College: Indian Institute of Information Technology Nagpur.