EP2420 *Project 1: task III*

Yiyi Miao

November 13, 2025

Project Overview

In this project, we train regression models that map IT and network infrastructure measurements X to predictions of service-level metrics Y. The measurements are obtained from a Video-on-Demand (VoD) service and a Key-Value store (KV) service, and the service-level metrics are Video Frame Rate and Response Time experienced by clients.

Using machine-learning techniques, the problem is to find a function (i.e., train a model) $M : X \to \hat{Y}$, such that $\hat{Y}$ closely approximates Y for a given X. Formally, we predict the expectation of $P(Y|X)$. If Y is numeric, the problem is referred to as a regression problem; if Y is a class or category, the problem is called a classification problem. This project considers both of these problems.

The machine-learning methods you will use are linear regression, random forest regression and classification, and neural network regression.

To estimate M, measurement pairs (or observations) of the form (x, y) are needed. A list of such measurement pairs ordered by time stamps is called a *trace*. The traces in this project are based on observations collected once per second during several hours from a running system.

Task III

1 Methodology

1.1 Pre-processing Variants

The following design matrices are evaluated:

- X_0: Raw data (no pre-processing)
- X_1: L2 normalization by column
- X_2: L2 normalization by row
- X_3: Min–Max normalization by column
- X_4: Min–Max normalization by row
- X_5: Standardization by column
- X_6: Standardization by row

For each version, the dataset is split into 70% for training and 30% for testing. The NMAE on the test set is calculated as:

$$\text{NMAE} = \frac{1}{\bar{y}} \cdot \frac{1}{m} \sum_{j=1}^{m} |y_j - \hat{y}_j|$$

where $\bar{y}$ is the mean of the observed target values in the test set.

1.2 Outlier Detection and Removal

Outlier removal is based on the standardized feature matrix X_5. A sample is marked as an outlier if any feature value exceeds a threshold T in absolute magnitude. By varying integer T in the range $[5, 120]$, we count the number of outliers removed and measure the NMAE of both regression models after retraining on the filtered datasets.

1.3 Regression Model

Linear regression and random forest regression are adopted in Task III. Specifically, we use the following hyper parameters for the decision tree to speed up the training process given the large amount of data (full feature matrix).

```
RandomForestRegressor(random_state=42,
                      n_estimators=50,
                      max_features=20,
                      n_jobs=-1)
```

2 Results and Discussion

2.1 Comparison of Pre-processing Methods

Table 1 summarizes the NMAE values for all seven pre-processing schemes on the KV and VoD datasets.

Table 1: NMAE comparison of different pre-processing methods for KV and VoD datasets.

Dataset	Pre-processing	Linear Regression	Random Forest
KV	X_0	5.63e+04	0.0181
	X_1	9.26e+03	0.0181
	X_2	1.33e+05	0.0182
	X_3	242	0.0182
	X_4	3.99e+04	0.0192
	X_5	9.26e+03	0.0182
	X_6	1.31e+05	0.0206
VoD	X_0	0.127	0.0869
	X_1	0.122	0.0871
	X_2	0.128	0.0932
	X_3	0.123	0.0871
	X_4	0.129	0.0984
	X_5	0.122	0.0870
	X_6	0.124	0.0964

Observations:

- For the KV dataset:
  - Linear regression performs poorly on raw and row-normalized data due to the unscaled magnitude of features.
  - Random forest regression is robust to scaling and shows nearly constant performance (NMAE ≈ 0.018) across all pre-processing methods.

  - Min-Max normalization by column (X_3) greatly reduces NMAE of the linear regression model and performs almost the best for the decision tree model, hence working well for both regression methods.

- For the VoD dataset
  - Linear regression results remain relatively stable (NMAE: $0.0122 \sim 0.129$) while random forest results remain relatively consistent (NMAE: $0.0087 \sim 0.0964$).
  - Standardization by column (X_5) achieves the optimal (minimum NMAE) of both regression models at the same time, which can be considered as the proper preprocessing method.

2.2 Outlier Statistics

Figure 1 illustrates how the number of outliers decreases with the threshold T. The number of outliers declines exponentially with increasing T, meaning that only a small subset of samples dominate extreme deviations. The VoD dataset initially contains more extreme samples than KV, but close to each other for $T > 50$ and finally converging to nearly zero outliers for $T > 80$.

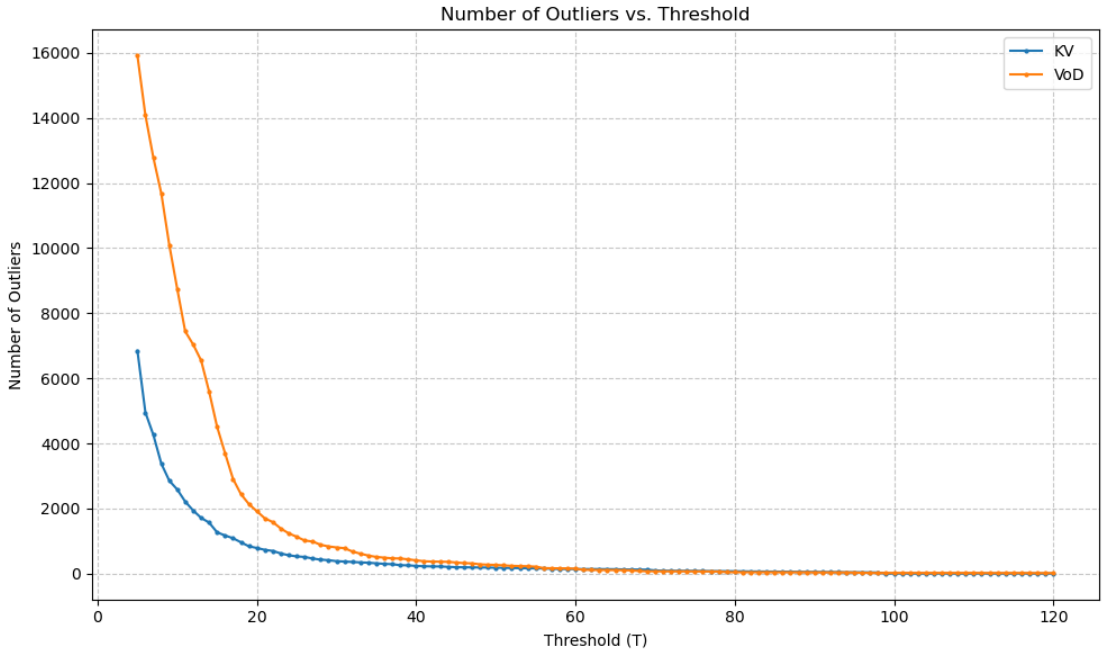


Figure 1: Number of detected outliers vs. threshold T for KV and VoD datasets.

2.3 Effect of Outlier Removal on Model Accuracy

Figures 2 and 3 show the NMAE values as a function of the threshold T on both datasets KV and VoD. Discussion:

- Linear regression occasionally exhibits spikes in NMAE for specific thresholds (e.g., NMAE $\approx 2e4, T = 70$; NMAE $\approx 9e3, T = 100$ for KV, NMAE $\approx 5.60, T = 20$ for VoD), suggesting that removing samples with certain threshold disturbs the overall balance of the dataset.
- In For both datasets, random forest remains nearly unaffected by outlier removal, confirming its robustness to outliers in the dataset.

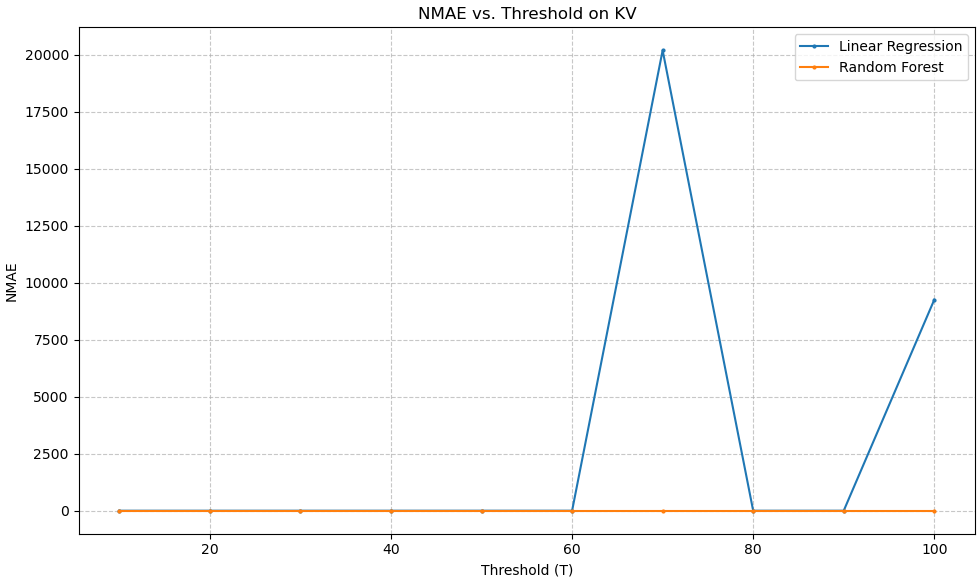


Figure 2: NMAE vs. threshold T on the KV dataset for linear regression and random forest regression.

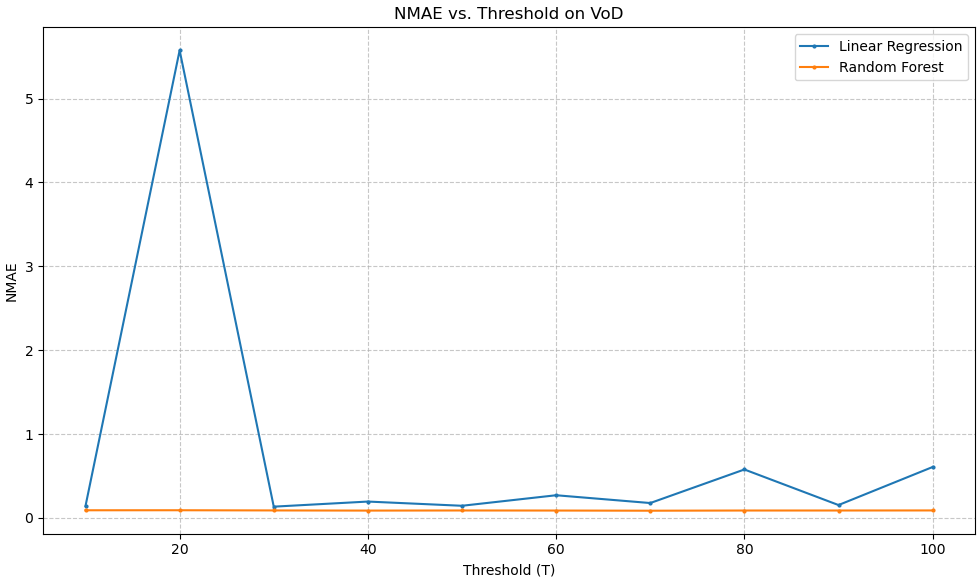


Figure 3: NMAE vs. threshold T on the VoD dataset for linear regression and random forest regression.

2.4 Cross-validation

Table 2 summarizes the mean NMAE obtained from 10-fold cross validation for both the KV and VoD datasets across different outlier thresholds T.

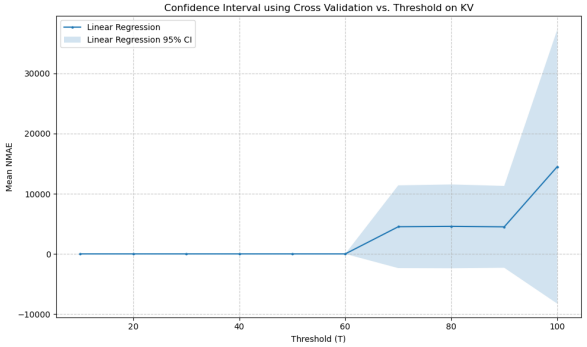
For the **KV dataset**, the Random Forest regressor consistently achieves a very low and stable error around 0.019 across all thresholds, indicating strong robustness to outliers. In contrast, the Linear Regression model performs reasonably well for $T \leq 60$ with NMAE between 0.12 and 0.19, but its error increases drastically for $T \geq 70$ (up to 1.45×10^4). This sharp rise demonstrates the sensitivity of linear regression to

extreme values, which distort the model coefficients and lead to instability.

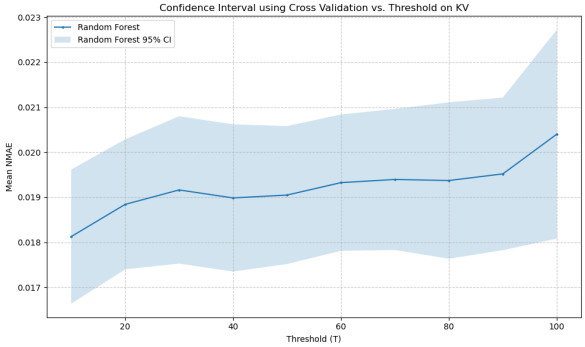
For the **VoD dataset**, the Random Forest again maintains a stable NMAE in the range $0.108 \sim 0.113$ for all thresholds, whereas the Linear Regression model produces significantly higher errors (around 4–6), suggesting that the mapping between device measurements and service metrics in this dataset is highly non-linear. These results confirm that Random Forest captures complex dependencies more effectively than Linear Regression.

Table 2: Mean NMAE over 10-fold Cross Validation for Different Thresholds T

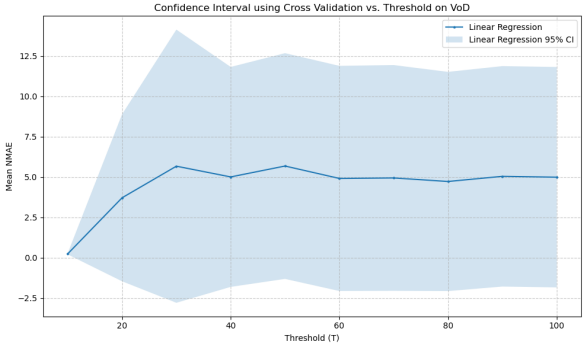
T	**KV Dataset**		**VoD Dataset**	
	Linear Regression	Random Forest	Linear Regression	Random Forest
10	0.189	0.018	0.234	0.113
20	0.137	0.019	3.71	0.110
30	0.120	0.019	5.67	0.108
40	0.194	0.019	5.01	0.108
50	0.127	0.019	5.68	0.108
60	0.121	0.019	4.91	0.109
70	4.52e+03	0.019	4.94	0.109
80	4.58e+03	0.019	4.72	0.110
90	4.50e+03	0.020	5.04	0.109
100	1.45e+04	0.020	4.99	0.109



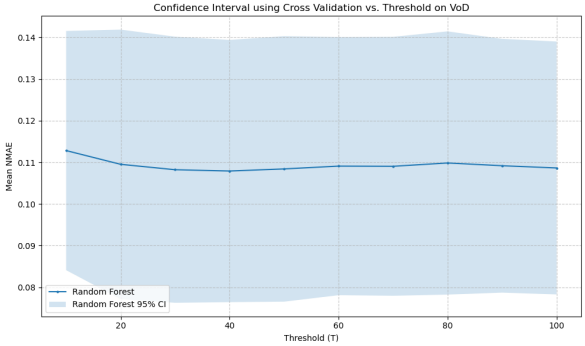
(a) Linear Regression Model on KV dataset



(b) Random Forest Model on KV dataset



(c) Linear Regression Model on VoD dataset



(d) Random Forest Model on VoD dataset

Figure 4: 95% Confidence Interval using Cross Validation on KV and VoD datasets.

Figure 4 illustrates the mean NMAE and the corresponding 95% confidence intervals obtained from 10-fold cross validation for both datasets and regression models. The shaded regions indicate the variability of the model's performance across folds: narrower intervals represent higher stability and stronger generalization, while wider intervals reflect greater sensitivity to training data or outliers.

- **KV dataset:**
  - The Linear Regression model shows a rapidly increasing mean NMAE and a widening confidence interval when the threshold T exceeds 60. This suggests that the model becomes highly unstable under the influence of outliers, and the prediction error varies greatly among folds.
  - The Random Forest model maintains a consistently low mean NMAE (around 0.019) and a narrow confidence interval across all thresholds, demonstrating excellent robustness and stable generalization.
- **VoD dataset:**
  - The Linear Regression model exhibits high error values (around 4–6) with wide confidence intervals across all thresholds, indicating poor consistency and weak predictive capability.
  - The Random Forest model achieves both low mean NMAE (around 0.11) and narrow confidence intervals, showing strong resistance to variations in data splits and excellent generalization across folds.

Discussion

Quiz

$$\begin{aligned}\mathbb{E}[Y \mid X = x_0] &= \sum_{i=1}^{9} y_i \cdot P(Y = y_i \mid X = x_0) \\ &= (1 \cdot 0.05) + (2 \cdot 0.10) + (3 \cdot 0.15) + (4 \cdot 0.10) + (5 \cdot 0.05) \\ &\quad + (6 \cdot 0.15) + (7 \cdot 0.25) + (8 \cdot 0.10) + (9 \cdot 0.05) \\ &= 5.25\end{aligned}$$

- The expected value $\mathbb{E}[Y \mid X = x_0]$ is 5.25.
- The most probable value (the mode) is the value of Y that has the highest probability. Looking at the histogram, the highest bar corresponds to $P(Y = 7) = 0.25$.
- The 90th percentile is the value of Y at or below 90%. At $Y = 7$, the cumulative probability is 0.85 (or 85%), which is below 90%. At $Y = 8$, the cumulative probability is 0.95 (or 95%), which is the first value to cross the 90% threshold. So the 90th percentile is 8.

Based on the results from Task III, we can draw the following conclusions:

- **Impact of Pre-processing:**
  - **KV dataset:** Feature scaling had a major impact on the linear regression model but almost no influence on the random forest model. Column-wise Min–Max normalization (X_3) yielded the best balance across both models, greatly improving linear regression stability.
  - **VoD dataset:** Results were generally stable across all preprocessing methods, with column-wise standardization (X_5) slightly outperforming the others.
  - These findings indicate that tree-based models are inherently scale-invariant, whereas linear regression relies heavily on normalization to ensure meaningful coefficient estimation.

- **Effect of Outlier Removal:**
  - The number of detected outliers decreased exponentially as the threshold T increased.
  - A similar pattern was observed on both datasets: linear regression performance fluctuated with outlier removal, but random forest performance stayed nearly unchanged.
  - This demonstrates that linear regression is highly sensitive to extreme values, since a few outliers can distort coefficient estimation.
- **Cross-validation and Model Robustness:**
  - Ten-fold cross validation confirmed the above trends.
  - **KV dataset:** Random forest regression achieved a very low and stable NMAE (approximately 0.019) with narrow 95% confidence intervals, demonstrating strong generalization. In contrast, linear regression showed increasing error and widening intervals when $T > 60$, indicating poor robustness to outliers.
  - **VoD dataset:** Random forest maintained stable performance with NMAE around 0.11, whereas linear regression exhibited large errors (NMAE $\approx$ 4–6) and wide confidence intervals across all thresholds, confirming its sensitivity to non-linear patterns.
- **Overall Conclusions:**
  - Random forest regression consistently outperformed linear regression across all preprocessing variants and outlier thresholds.
  - Proper feature scaling (e.g., Min–Max normalization or standardization) is essential for linear regression but offers limited benefits for random forests.
  - Moderate outlier removal can enhance stability, but overly aggressive filtering eliminates meaningful variation and degrades model performance.