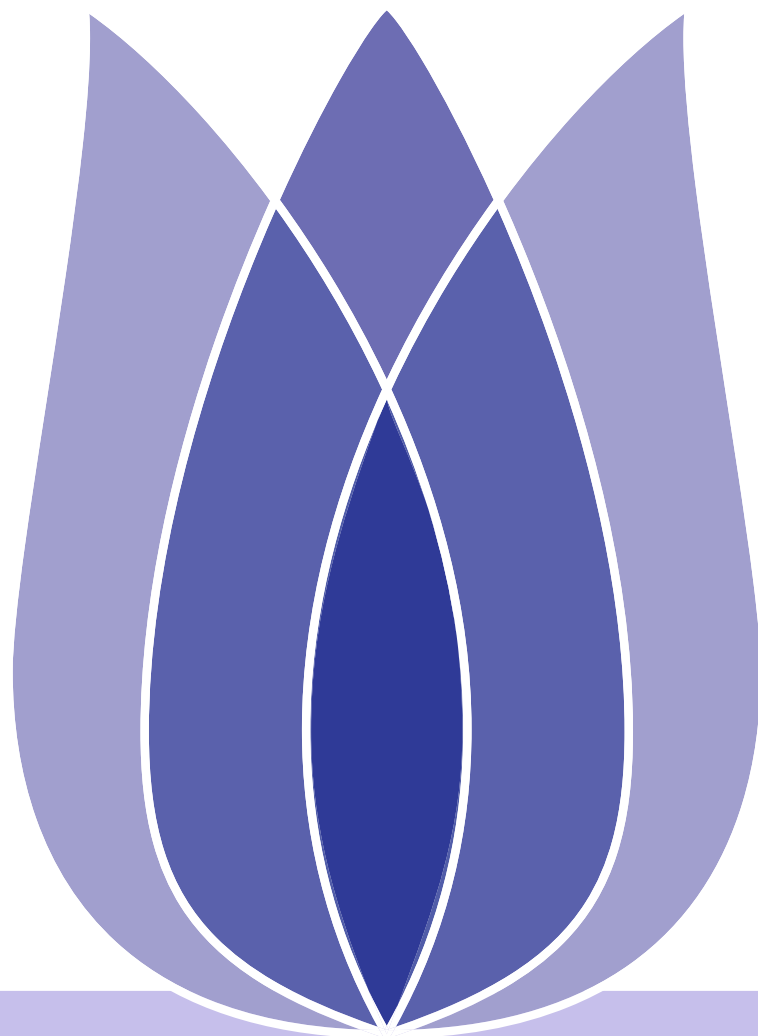


Air Pollution Prediction based on multicollinearity

Shanika Iroshi Nanayakkara
Deakin University

03/06/2022





Overview

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)

Problem Definition

Air pollution predictors

Challenges

Challenges

Proposed Model

Model Prediction and Evaluation

Model Prediction and Evaluation

Evaluation Results

Conclusion



Problem Definition

Air pollution predictors

Challenges

Proposed Model

Evaluation Results

Conclusion

Problem Definition



Air pollution predictors

Problem Definition
Air pollution predictors
Challenges
Proposed Model
Evaluation Results
Conclusion

Defn

- Predict air pollution composition in atmosphere in the future.
- Use predictors such as Temperature, Humidity, Sensor data.
 - Response variables are carbon monoxide, benzene, notrous oxide



Linear Regression vs Rigid Regrsson

Problem Definition
Air pollution predictors
Challenges
Proposed Model
Evaluation Results
Conclusion

Linear Regression

- Linear regression presents relationship as a straight line.
- Show correlation between two variables (one predictor for response variable/variables).
- Response should be continuous and independent variable(s) (predictor variables) can be continuous or discrete.

Rigid Linear Regression

- Use to implement multicollinearity of predictor variables which are highly correlated each other **objects** in the whole dataset.
- Add penalty values to reduce the loss or error of linear regression cause by bias and/or variance of the variables.



- [Problem Definition](#)
- [Challenges](#)**
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)

Challenges



Challenges

- Problem Definition
- Challenges
- Challenges
- Proposed Model
- Evaluation Results
- Conclusion

multicollinearity

- Focus on correlation between predictors.

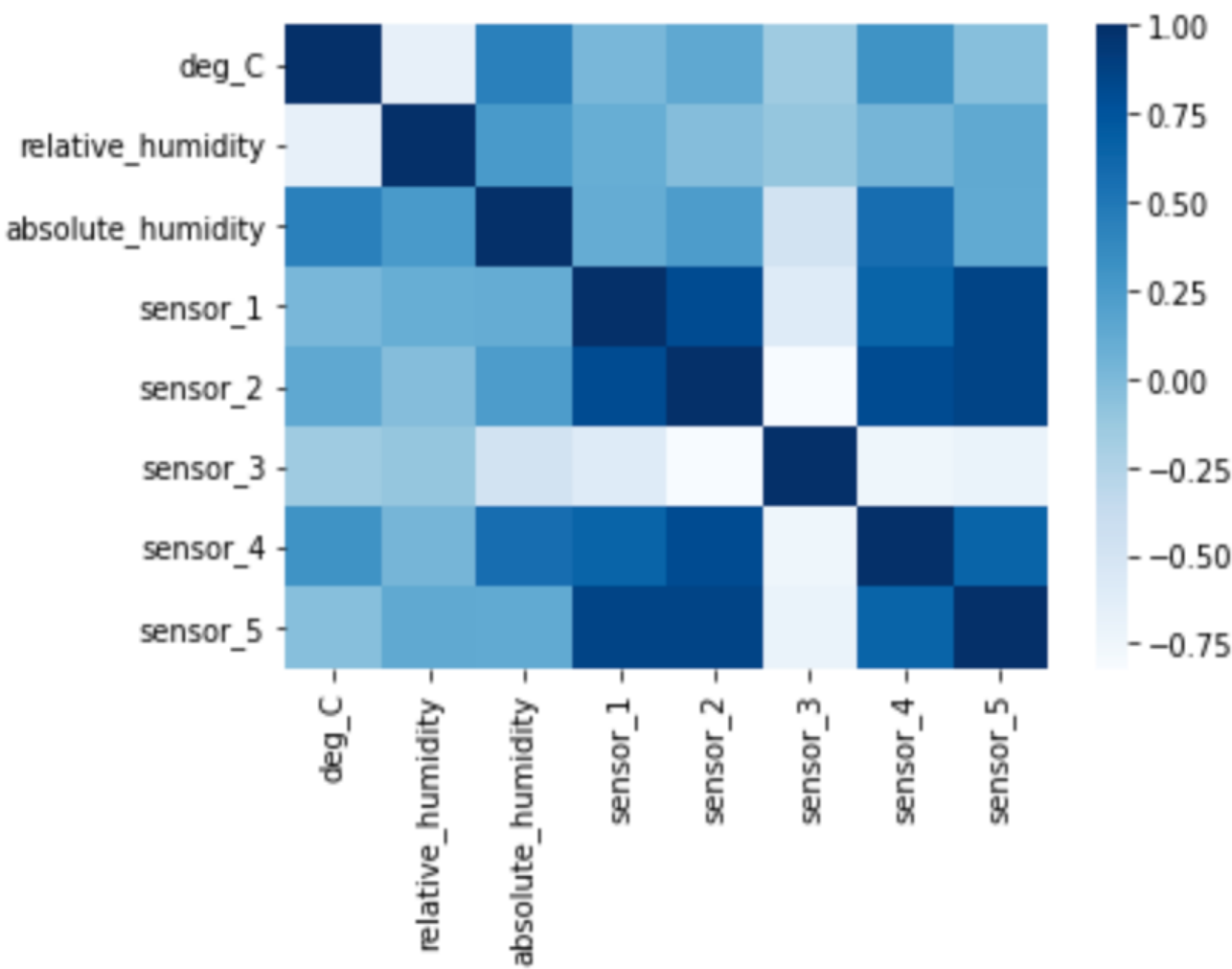


Figure 1: Framework of Model

Reasons for multicollinearity in predictors.

- Inaccurate use of different types of variables.
- Poor selection of questions or null hypothesis.
- Variable repetition.
- A dependent variable selection.
- High correlation.
- Use of dummy variables.



Challenges

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)

Fixing multicollinearity

- Obtain more data.
- Utilize a ridge regression.
- Utilize a partial squares regression
- Removing a variable.
- Do nothing.



- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)**
- [Model Prediction and Evaluation](#)
- [Model Prediction and Evaluation](#)
- [Evaluation Results](#)
- [Conclusion](#)

Proposed Model



- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Model Prediction and Evaluation](#)
- [Model Prediction and Evaluation](#)
- [Evaluation Results](#)
- [Conclusion](#)

Framework of Proposed Model:

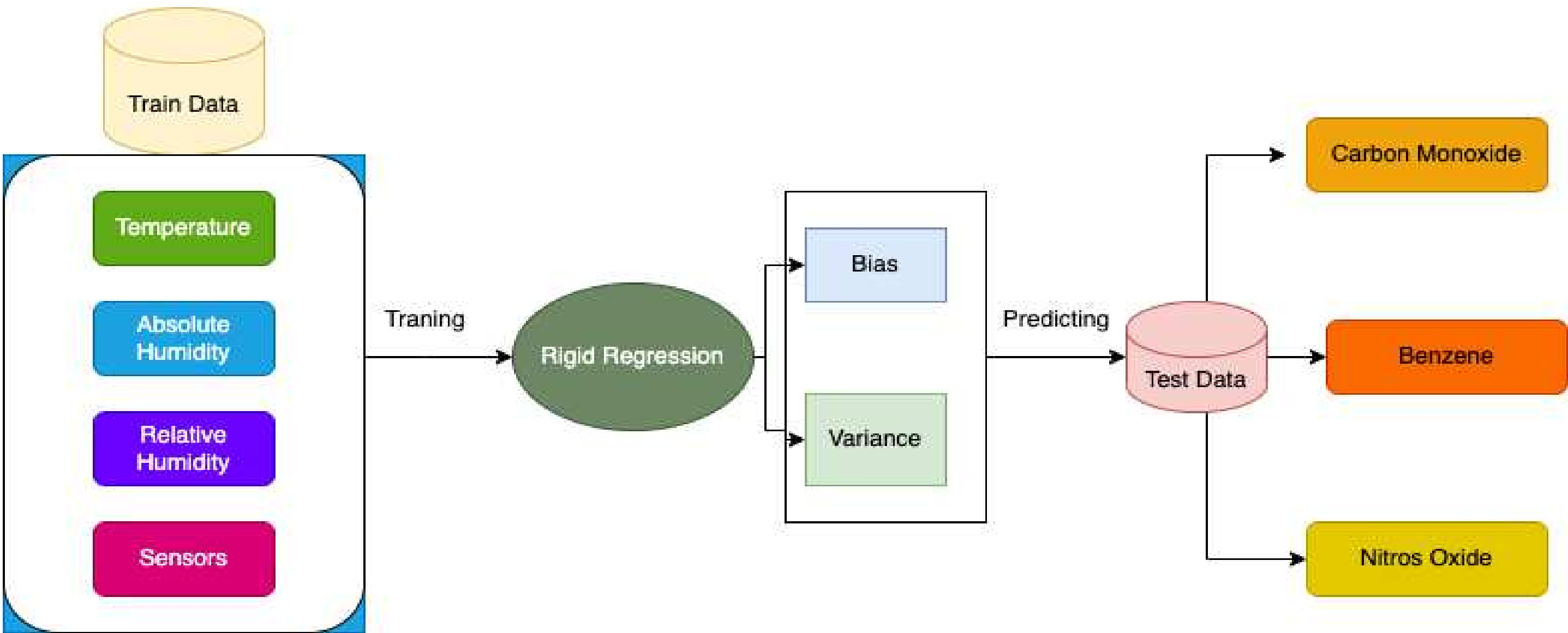


Figure 2: Framework of Model

Preprocessing and Model training

Problem Definition

Challenges

Proposed Model

Model Prediction and Evaluation

Model Prediction and Evaluation

Evaluation Results

Conclusion

Bias and Variance

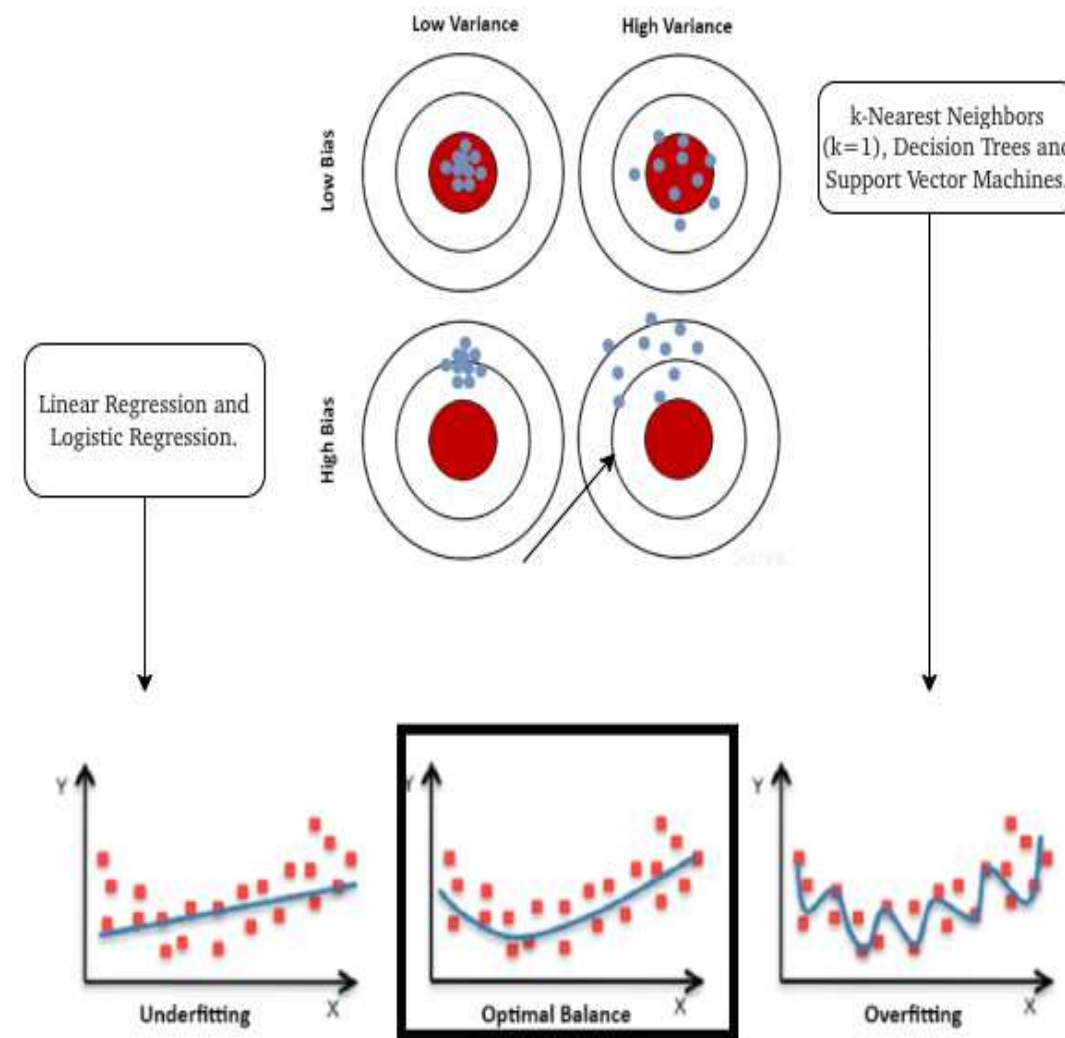


Figure 3: Bias and VAriance

Reasons for **multicollinearity** in predictors.

- Remove Duplicates.
- Remove unwanted columns.
- Scale for data consistency.
- Remove Outliers.



TULIP

Team for Universal Learning and Intelligent Processing



Preprocessing and Model training

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Model Prediction and Evaluation](#)
- [Model Prediction and Evaluation](#)
- [Evaluation Results](#)
- [Conclusion](#)

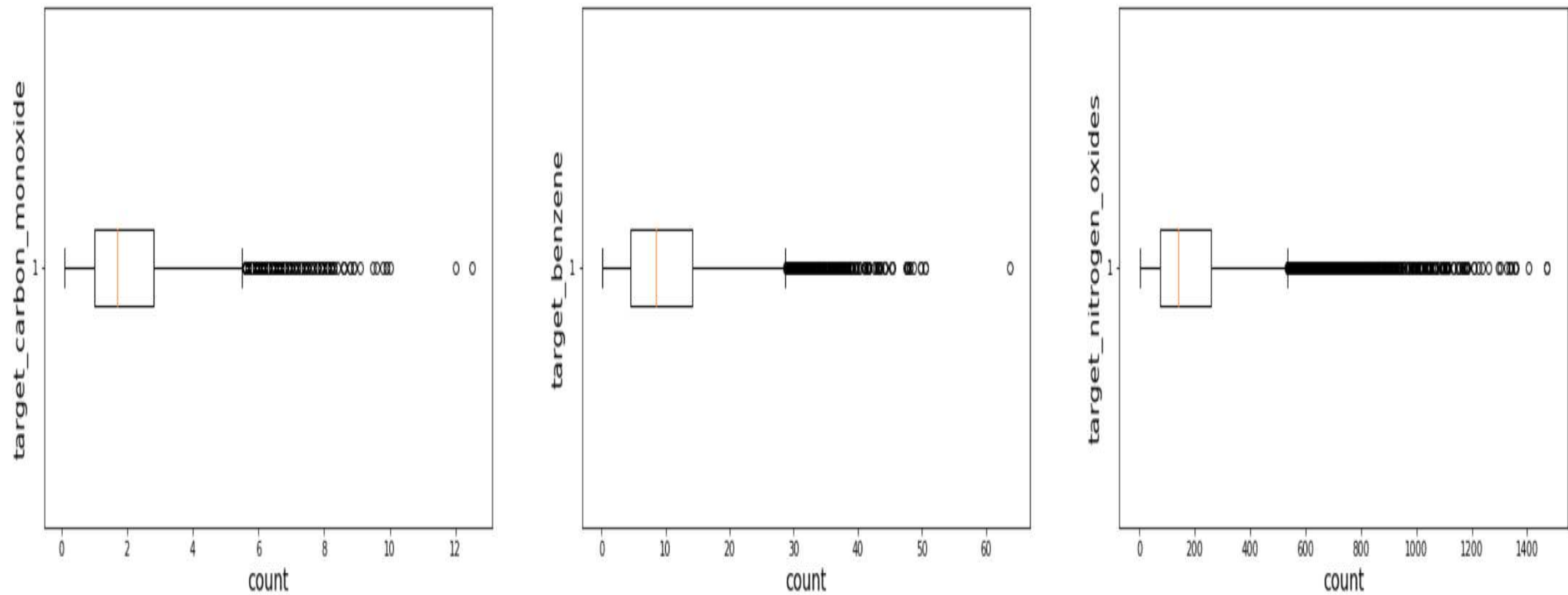


Figure 4: Response data distribution and Outlier Detection.



Model Prediction and Evaluation

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Model Prediction and Evaluation](#)**
- [Model Prediction and Evaluation](#)
- [Evaluation Results](#)
- [Conclusion](#)

■ Relationships between predictors and response variables over time.

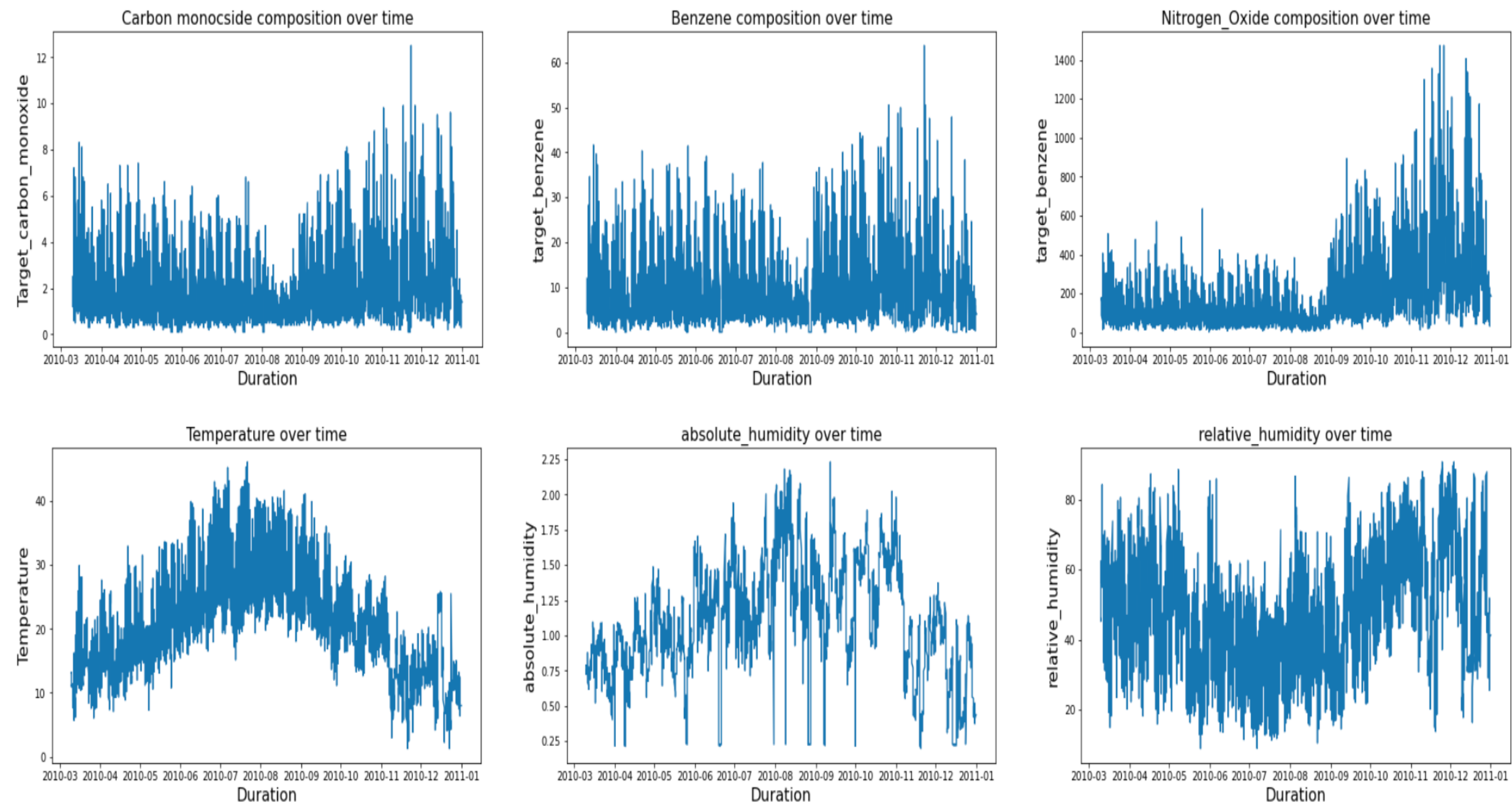


Figure 5: Relationships over the time



- Problem Definition
- Challenges
- Proposed Model
- Model Prediction and Evaluation
- Model Prediction and Evaluation**
- Evaluation Results
- Conclusion

■ Determination of coefficient.

	CM	Benzene	NO
Bias	0.38	3.379	12916.6
Variance	0.001	0.009	30.591
CD Training	0.82	0.95	0.68
CD Testing	0.83	0.94	0.66

CD : Coefficient Determination
CM : Carbon Monoxide
NO : Nitrous Oxide

Figure 6: performance Evaluation



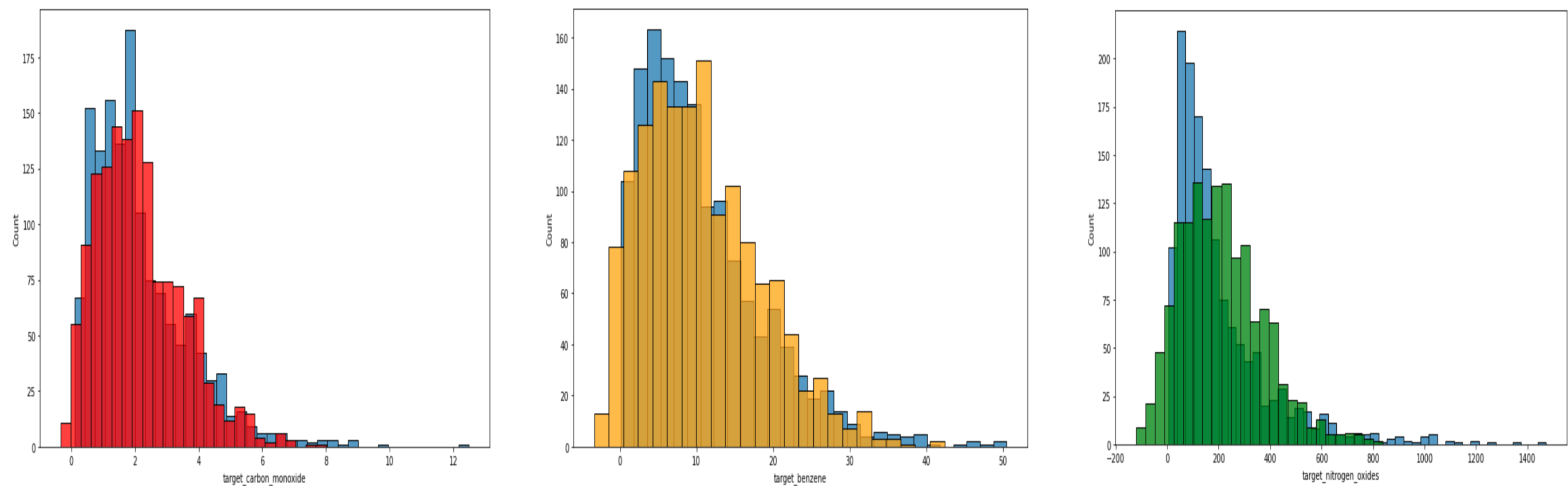
- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)**
- [Conclusion](#)

Evaluation Results



Evaluation

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)



(a) Carbon_monoxide_prediction (b) Benzene_Prediction (c) Nitrous Oxide Prediction

Figure 7: Histogram of three test feature predictions



- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)

Conclusion



Conclusion

Problem Definition

Challenges

Proposed Model

Evaluation Results

Conclusion

- Problem Definition: we propose prediction model to predict air pollution components which are specified as carbon monoxide, benzene and nitrous oxide. Algorithm Consequently, we use rigid linear regression instead general linear regression method.
- Strategies :We uses multiple predictors such as temperature, absolute and relative humidity and five sensor data. We identify that these predictors are correlated among each other. Therefore, we reveals the need of handling multicollinearity. We clearly show the performance efficiency of proposed models using determination of coefficient, bias and variance scores for each models.
- Recommendations : Evaluate how well these prediction models behave on adding more noisy data. How to affect prediction accuracy by noisy data. Increase the amount of predictors to predict air pollution



TULIP

Team for Universal Learning and Intelligent Processing



Questions?

- [Problem Definition](#)
- [Challenges](#)
- [Proposed Model](#)
- [Evaluation Results](#)
- [Conclusion](#)



Contact Information

Associate Professor Gang Li
School of Information Technology
Deakin University, Australia



GANGLI@TULIP.ORG.AU



TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

