FORECASTING DAMAGES CAUSED BY PIPE BURSTS

PROBLEM

Oil pipeline leaks and spills can have severe and far-reaching consequences, both environmentally and economically.

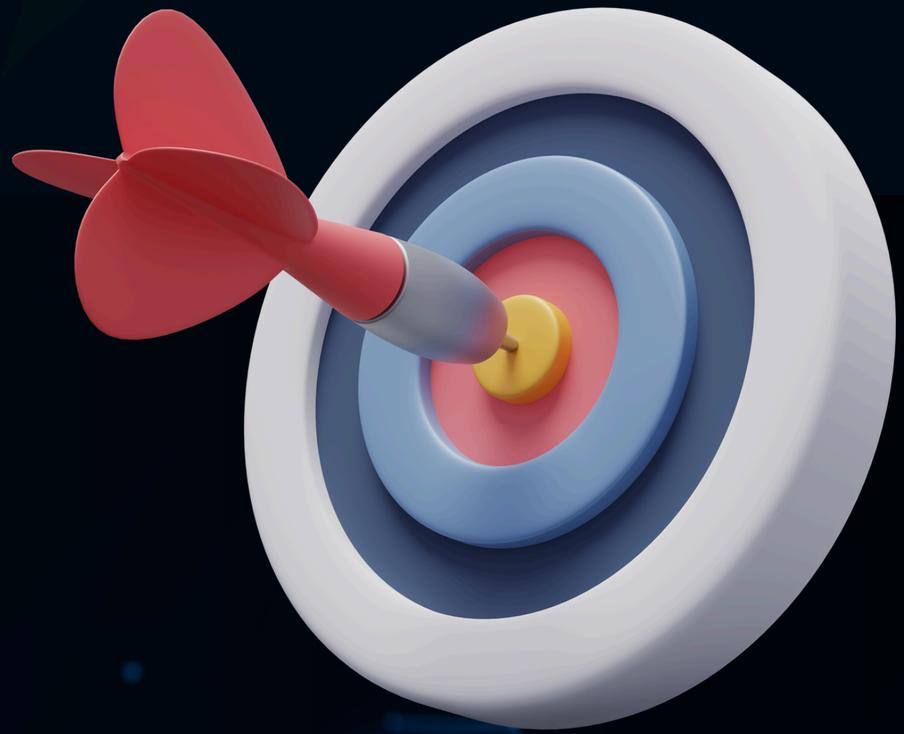
Environmental Damage

- **Soil and Groundwater Contamination**
- **Water Pollution**
- **Air Pollution**
- **Habitat Destruction**

Economic Loss

- Cleanup Costs
- Private/PublicProperty Damage Costs
- Legal Liabilities
- Environmental Remediation

PURPOSE



We will be focused on economic loss and our aim is to develop a predictive model that can estimate the potential cost of damage based on various features and factors present in the dataset.

DATA OVERVIEW

TARGET VARIABLES

Property Damage Costs

Lost Commodity Costs

Private/Public Property Damage Costs

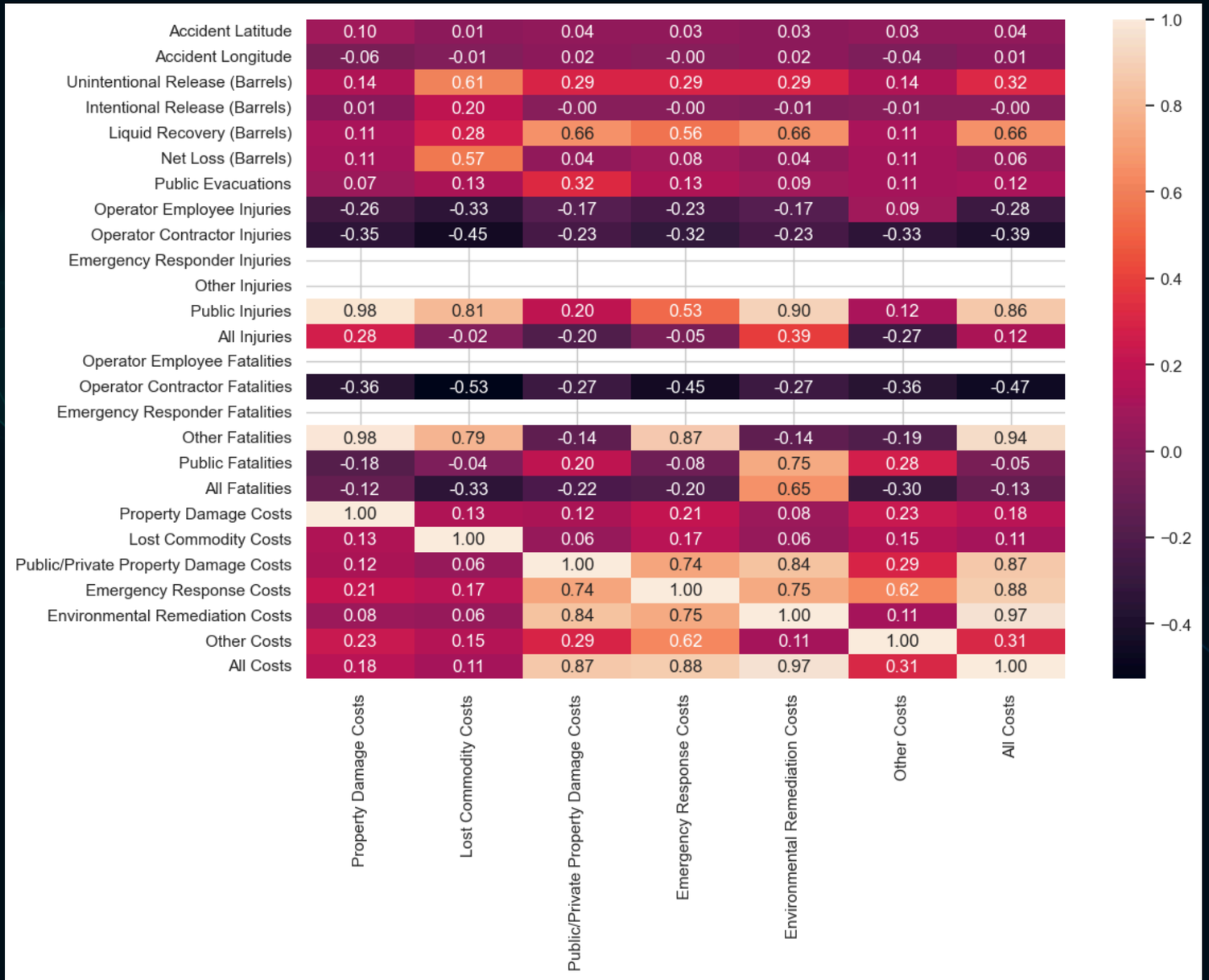
Emergency Responce Costs

Other Costs

Environmental Remediation Costs

All Costs

CORRELATION MATRIX



EDA RESULTS

1. A lot of outliesrs
2. A lot of missing data
3. Class imbalance
4. Target varriables distribution is skewed to right
5. Slight correlation between target variables and features

DATA CLEANING

1. Droping of columns (Identification, a lot of unique values)
2. Deletion of rows with missing values in target variables
3. Imputing of missing values in other variables

DATA PREPROCESSING

CATEGORICAL DATA

feature	preprocessing
Pipeline Type Accident State	SimpleImputer (most freq.)
Liquid Subtype Liquid Name	BinaryEncoder

model

DATETIME

feature

preprocessing

Accident Date/Time

DateToMonth
RepeatingBasisFunction

DateToMonth
RepeatingBasisFunction
'n_periods': 7

DateToNSM

month

day

seconds

model

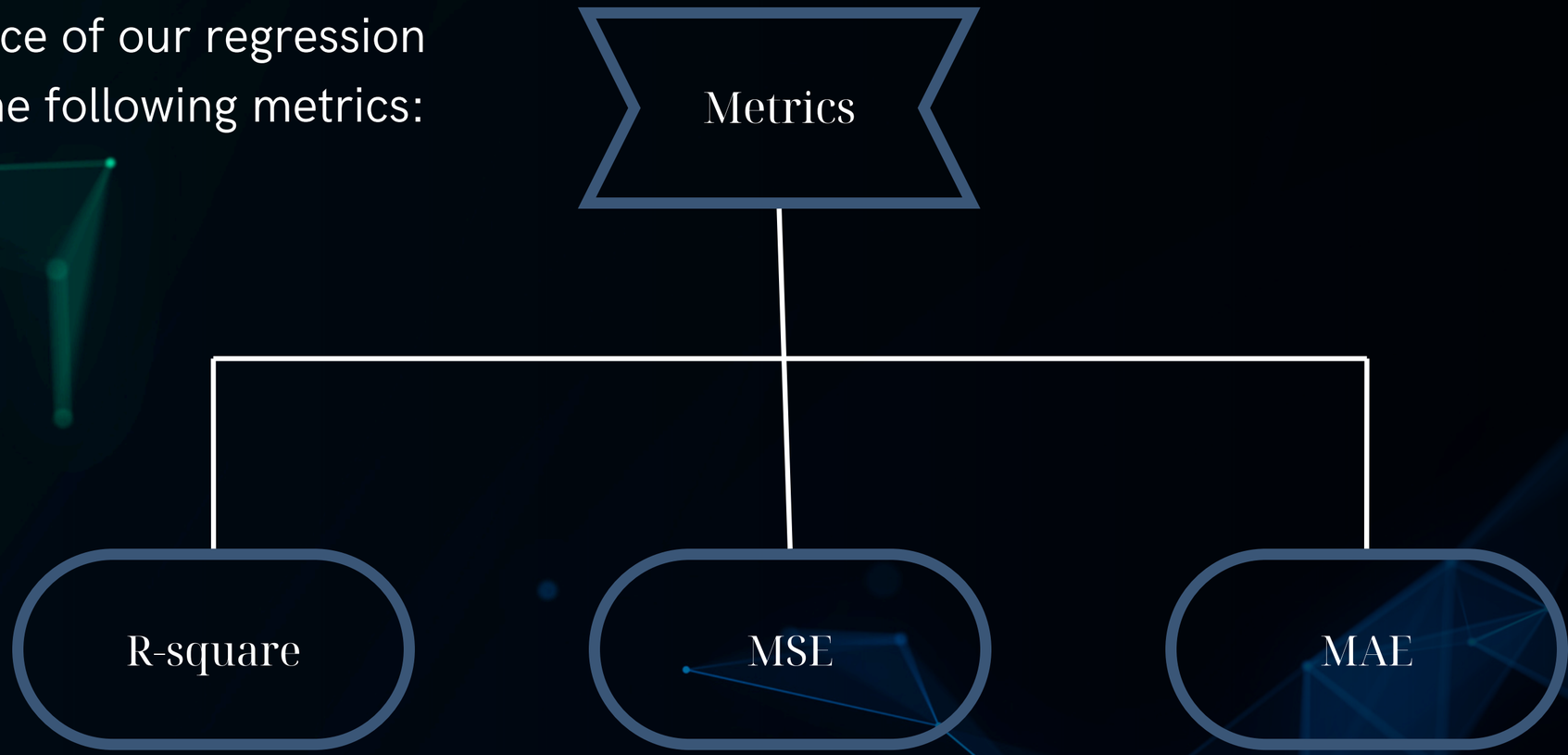
NUMERICAL DATA

feature	preprocessing
Pipeline Shutdown Shutdown DT Restart DT	DateDiff
Public Evacuations Unintentional Release Liquid Recovery Net Loss	KNNImputer FunctionTransformer PolynomialFeatures
Injuries&Fatalities	None to 0
Intentional Release	BinaryEncoder

model


METRICS

For evaluating the performance of our regression model, we have selected the following metrics:



By considering both R-squared and error metrics (MSE and MAE), we can make informed decisions about model selection and potential improvements.


MODELS

Hyperparameters tuning using Cost-Frugal Optimization

XGBoost

Hyperparameters:

- n_estimators,
- max_leaves,
- min_child_weight,
- learning_rate,
- subsample,
- colsample_bylevel,
- colsample_bytree,
- reg_alpha,
- reg_lambda.

RandomForest

Hyperparameters:

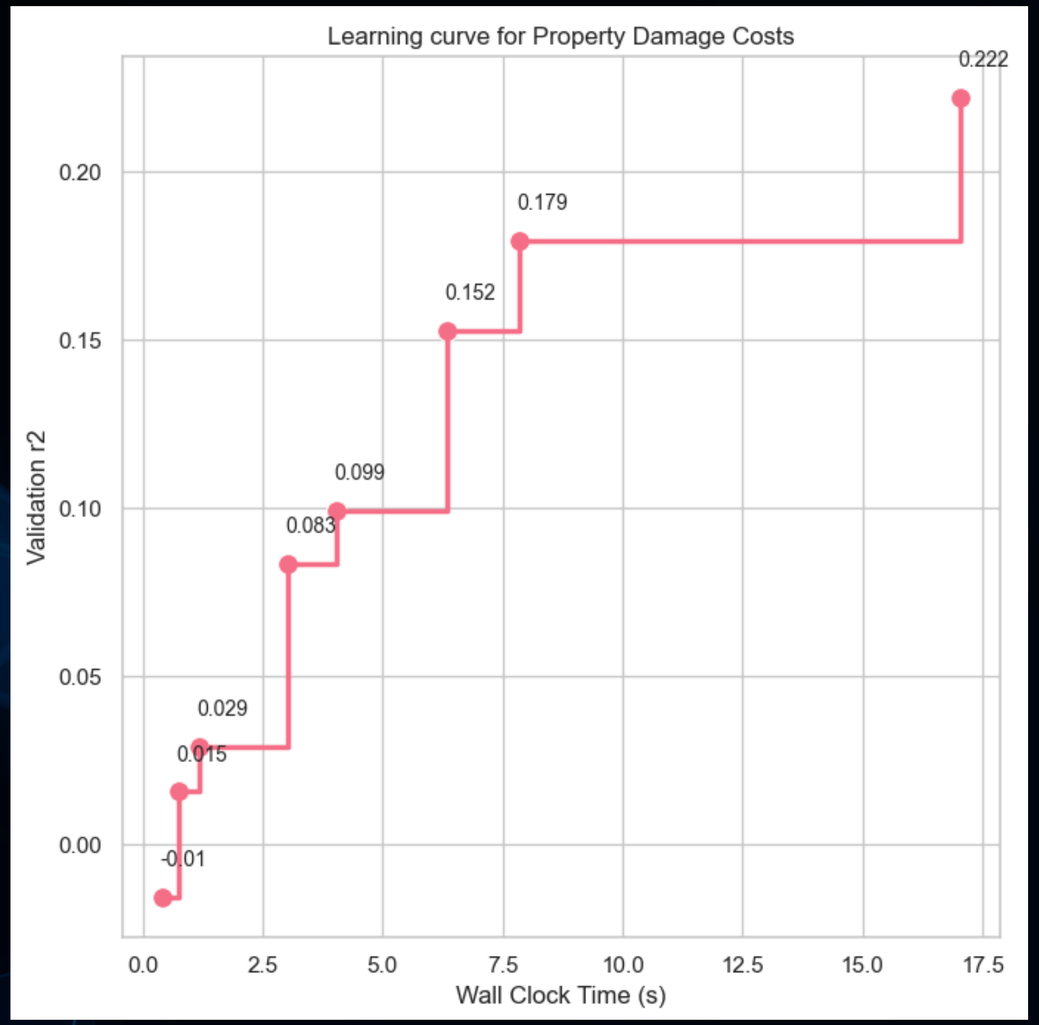
- n_estimators,
- max_features,
- max_leaves

ExtraTree

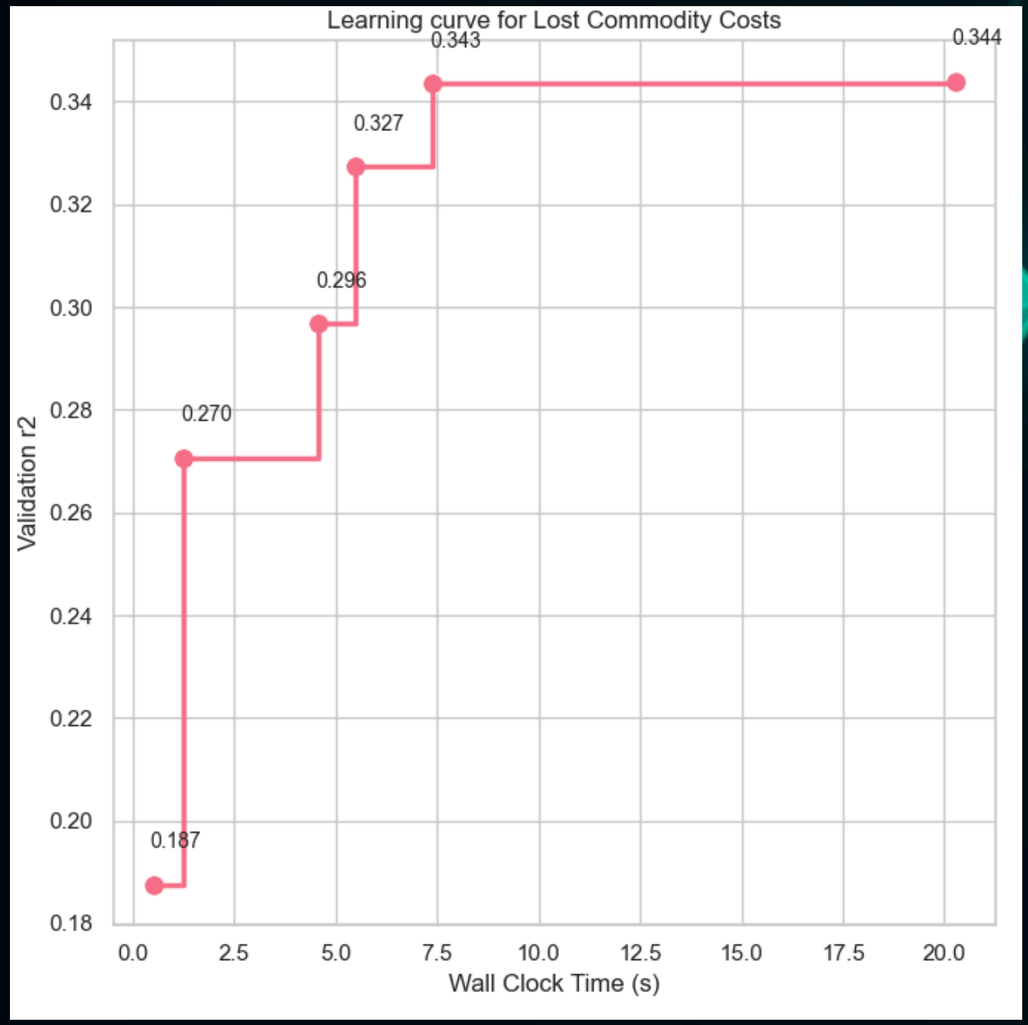
Hyperparameters:

- n_estimators,
- max_features,
- max_leaves

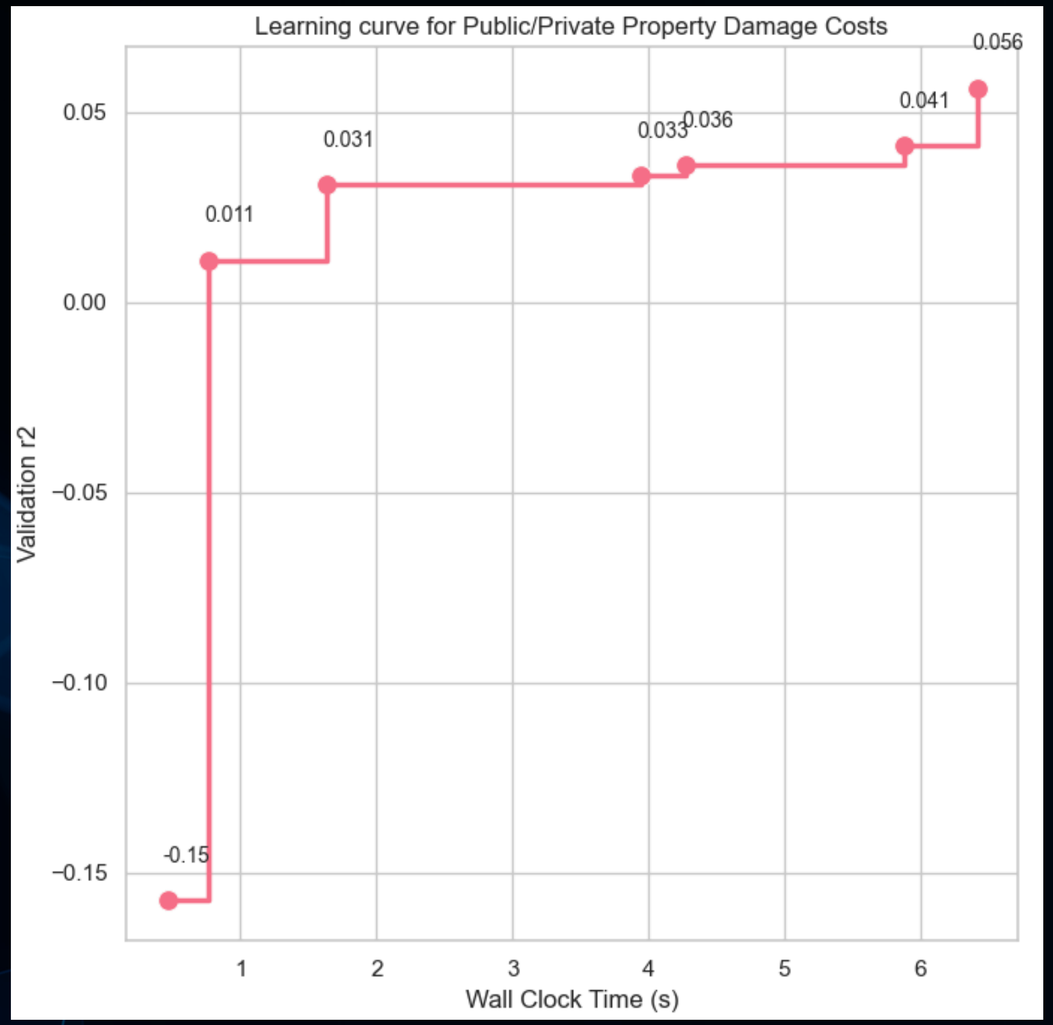
Property Damage Costs
XGBoost



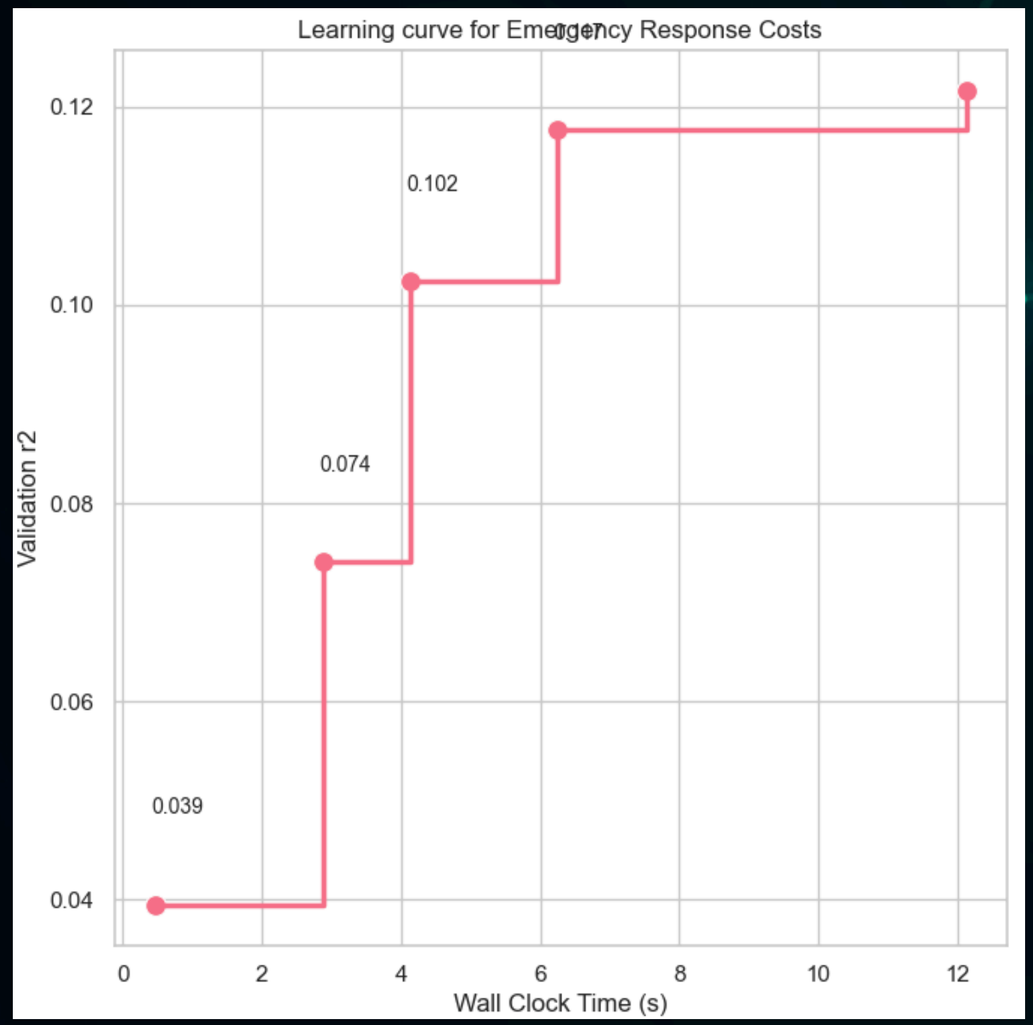
Lost Commodity Costs
XGBoost



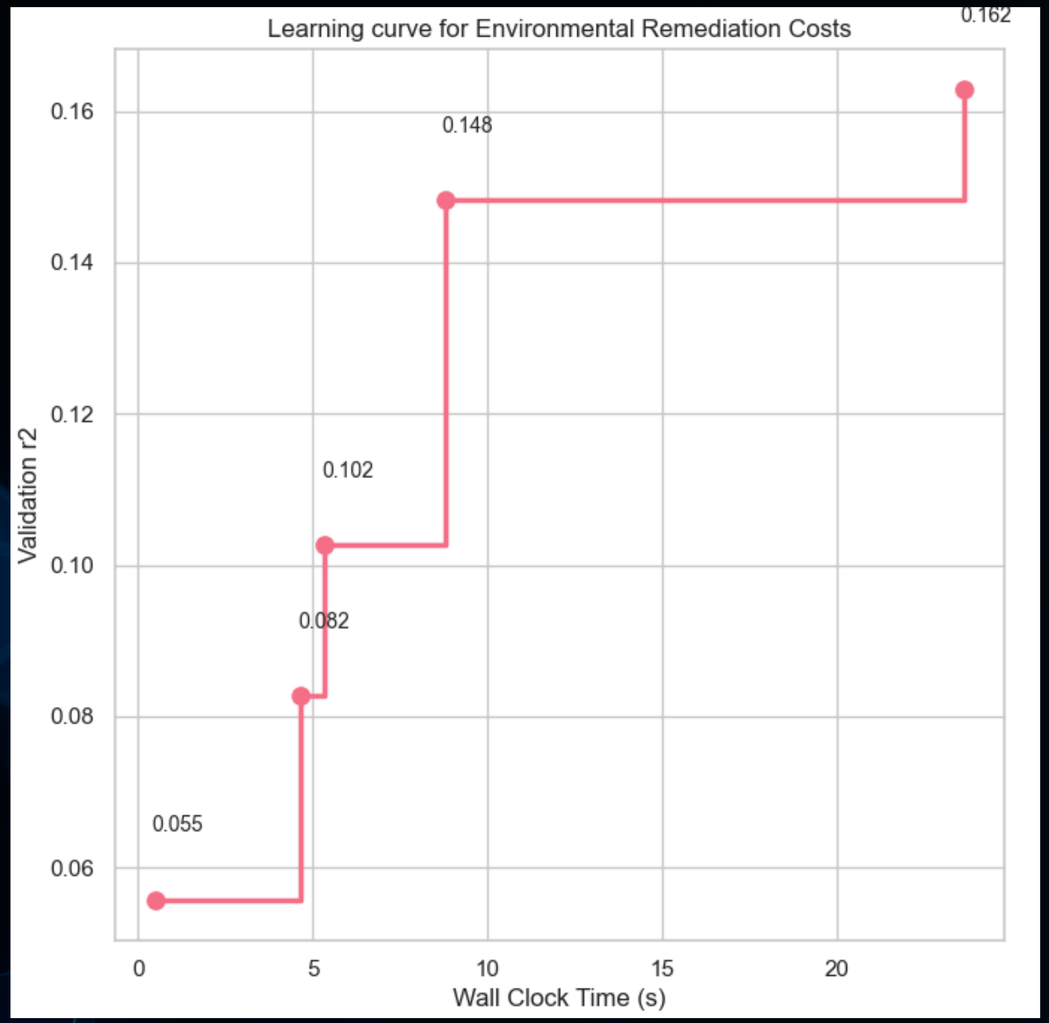
Public/Private Property Damage Costs XGBoost



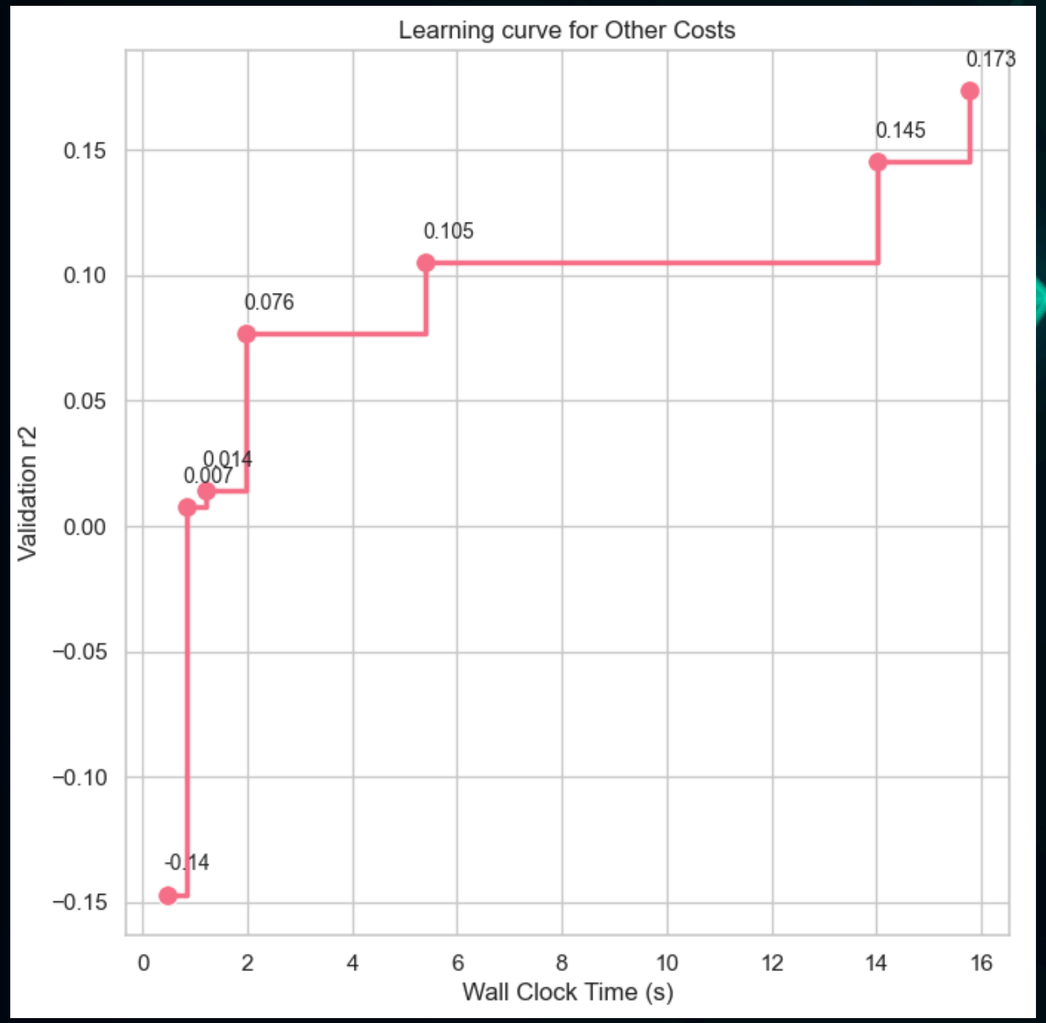
Emergency Response Costs XGBoost



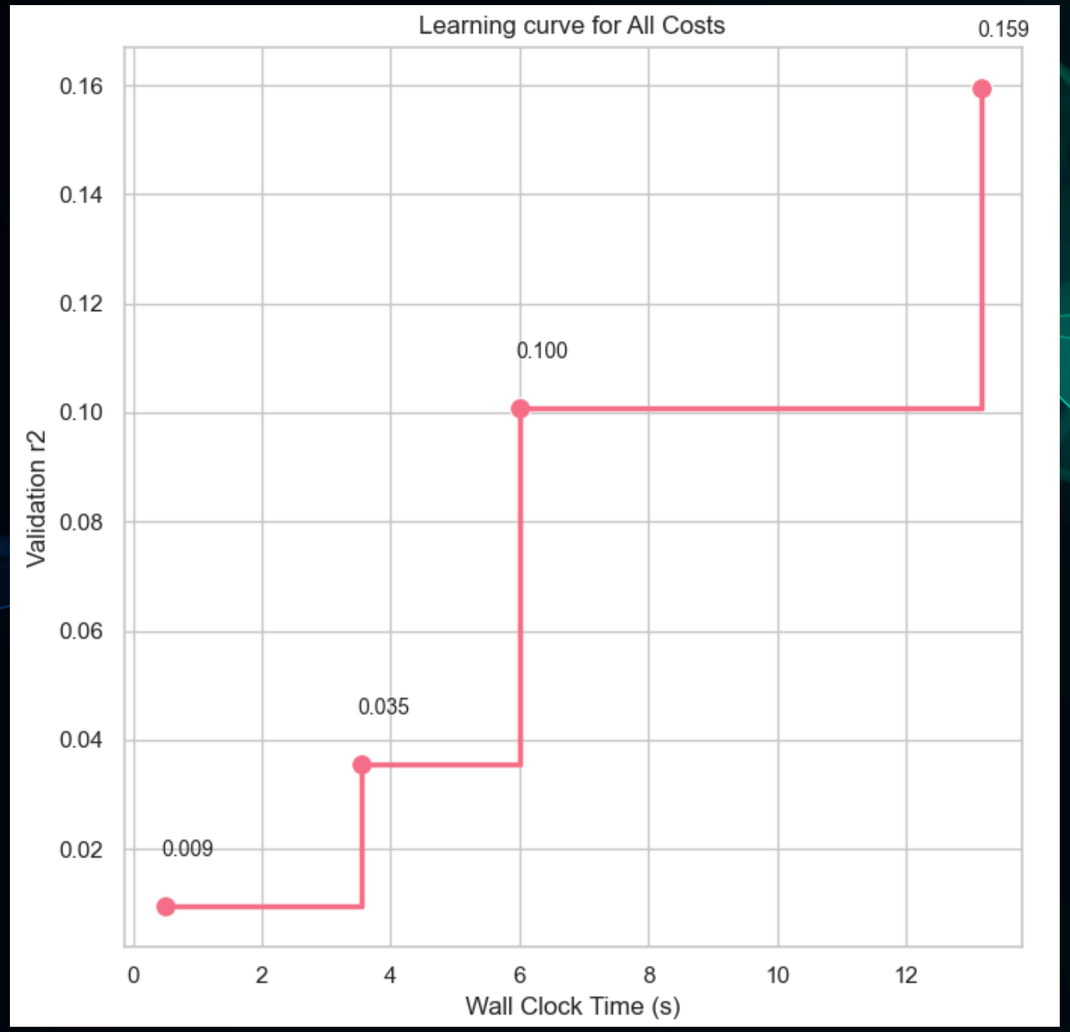
Environmental Remediation Costs
XGBoost



Other Costs
XGBoost

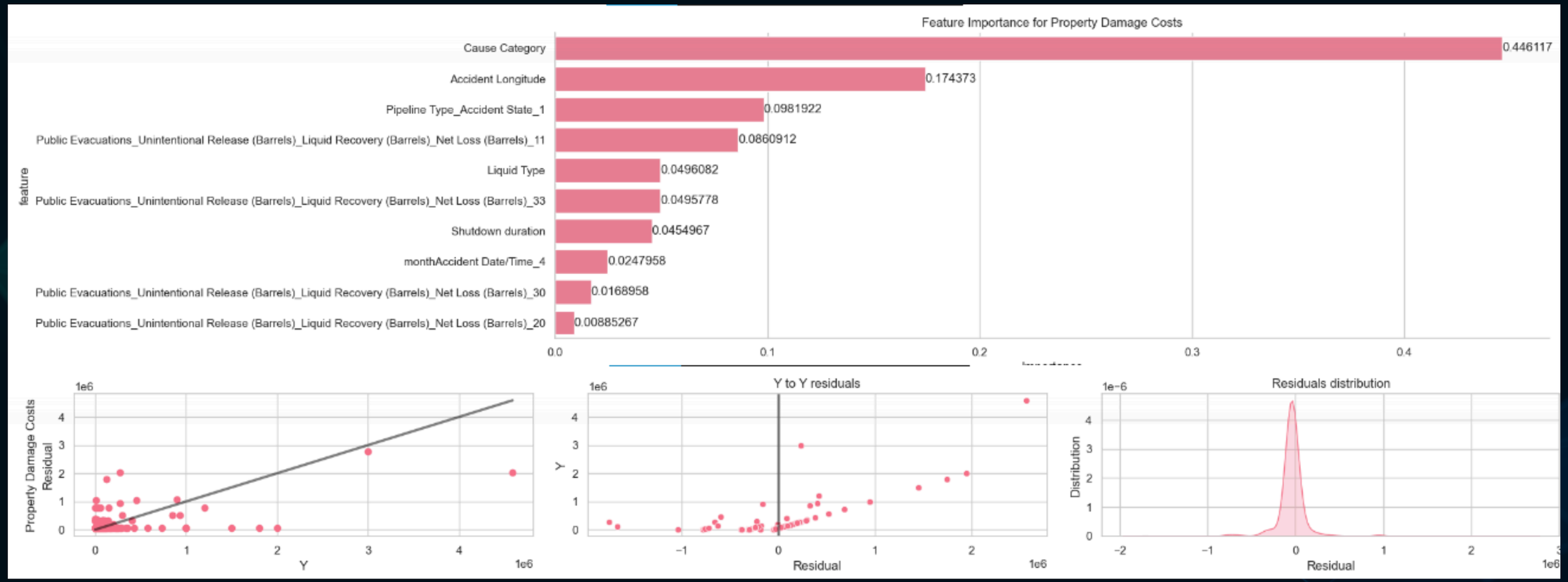


All Costs
ExtraTree



RESULTS

PROPERTY DAMAGE COSTS

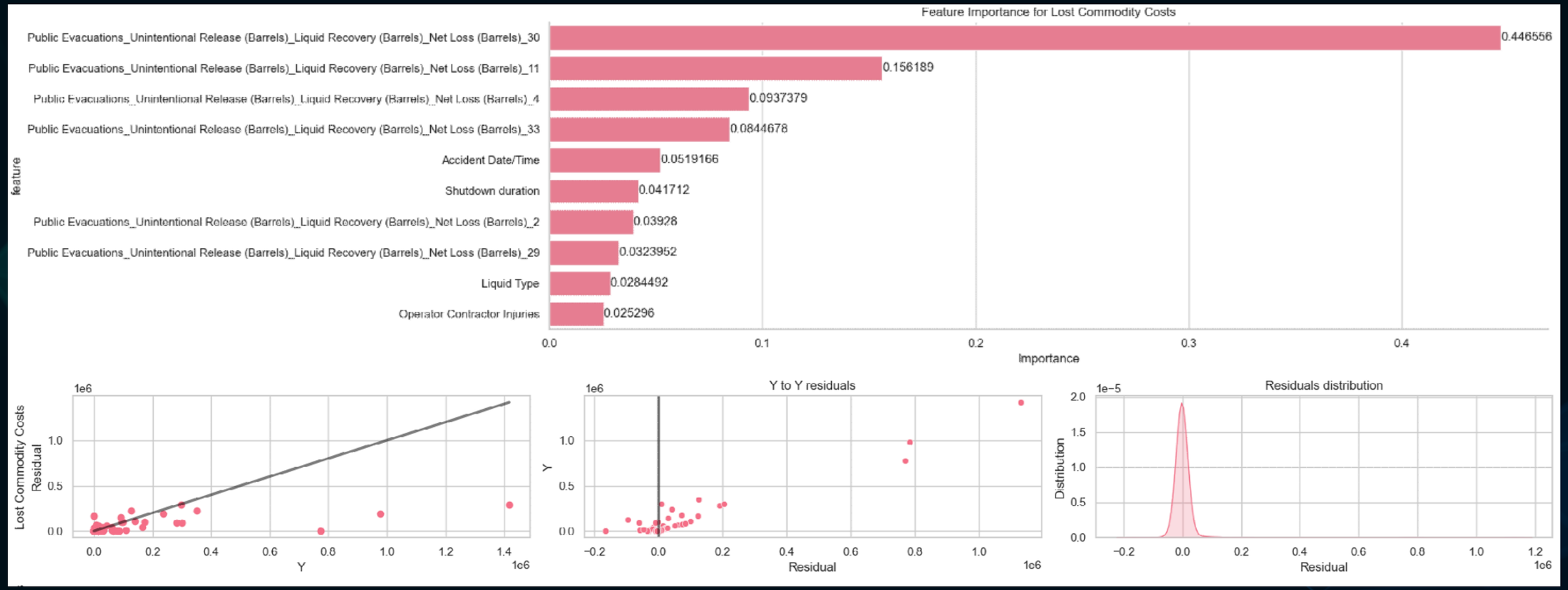


R2: 0.262

MSE: 6.4 x 10^10

MAE: 1.02 x 10^5

LOST COMMODITY COSTS

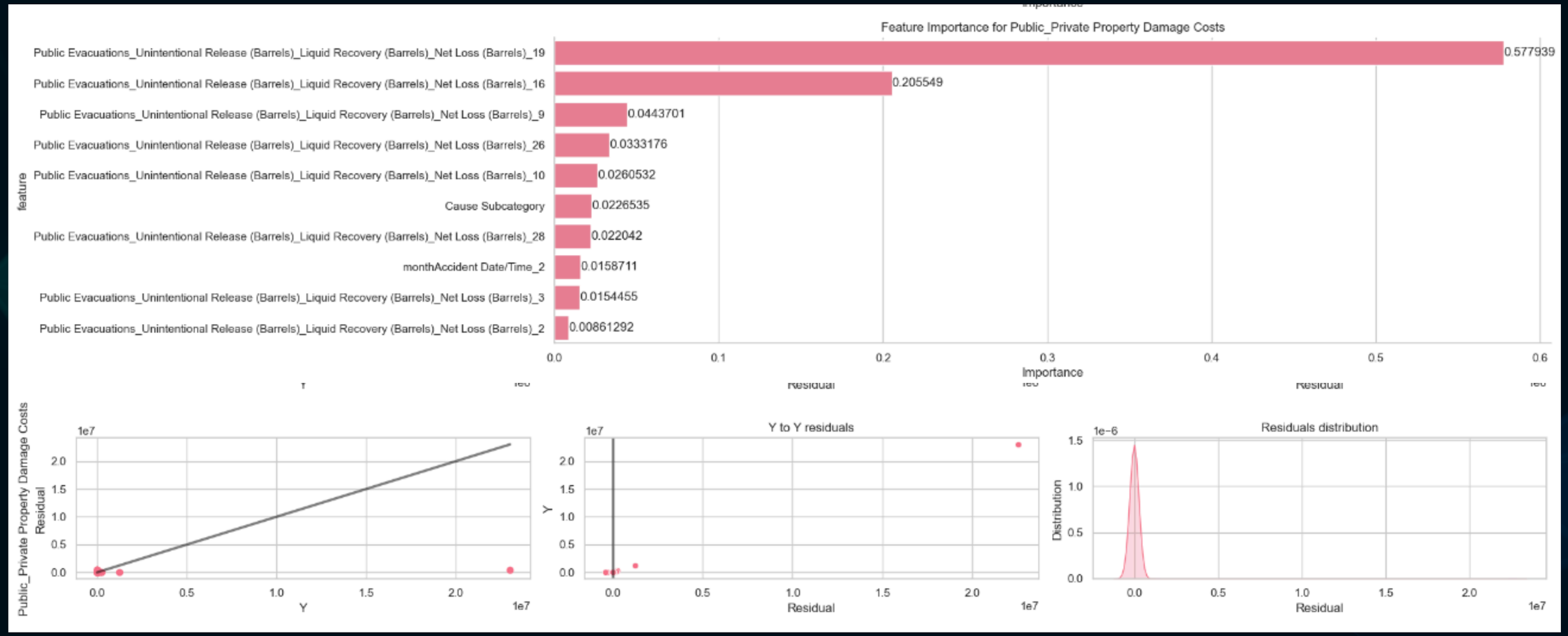


R2: 0.342

MSE: 4.84 x 10^10

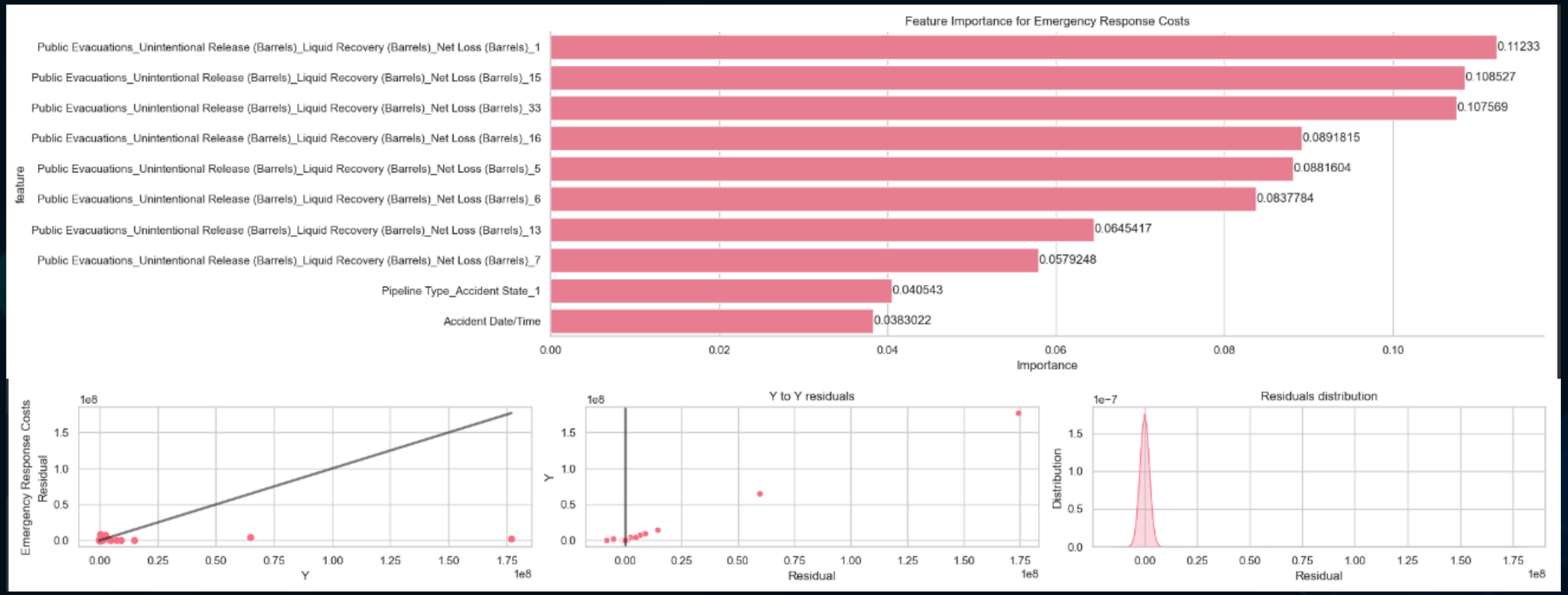
MAE: 1.16 x 10^5

PUBLIC/PRIVATE PROPERTY DAMAGE COSTS



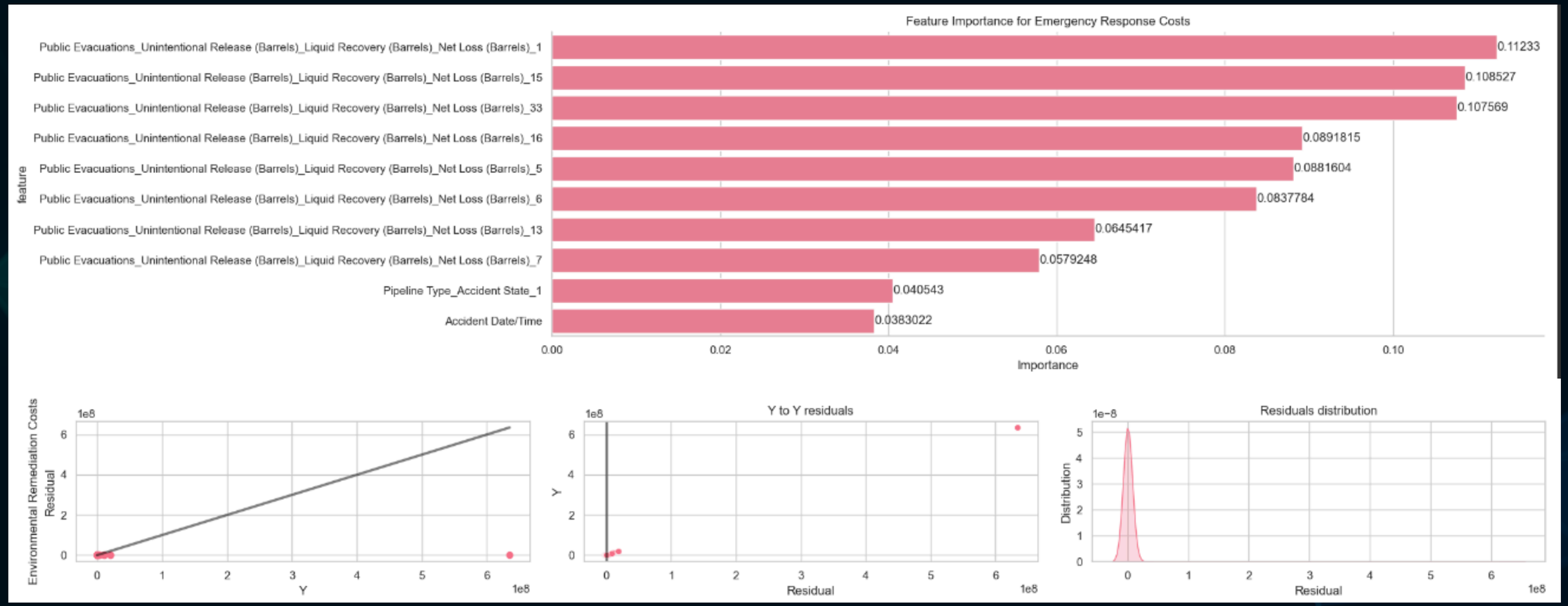
R2: 0.032 *MSE: 9.21 x 10^11* *MAE: 6.20 x 10^4*

EMERGENCY RESPONSE COSTS



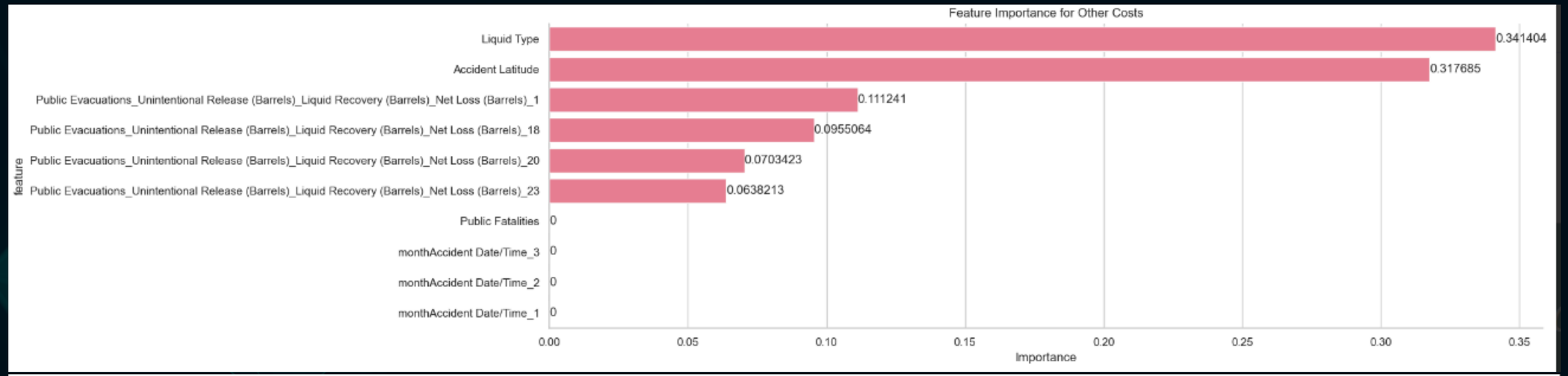
R2: 0.038 *MSE: 6.18 x 10^13* *MAE: 6.67 x 10^5*

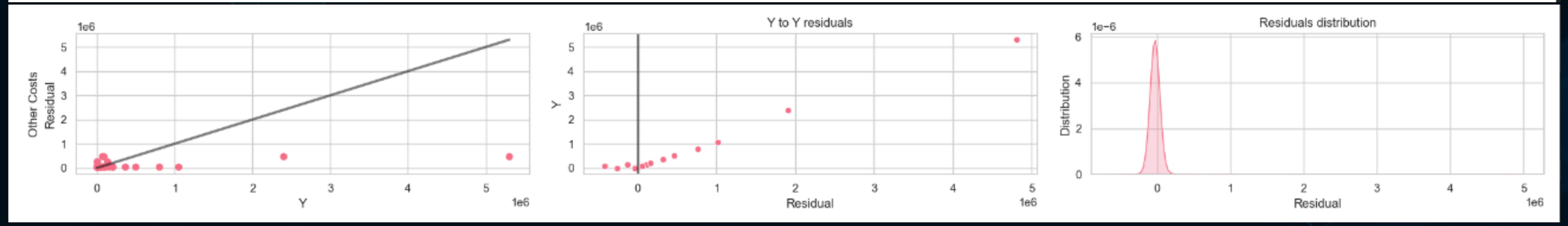
ENVIRONMENTAL REMEDIATION COSTS



R2: 0.002 *MSE: 7.23 x 10^14* *MAE: 1.34 x 10^6*

OTHER COSTS



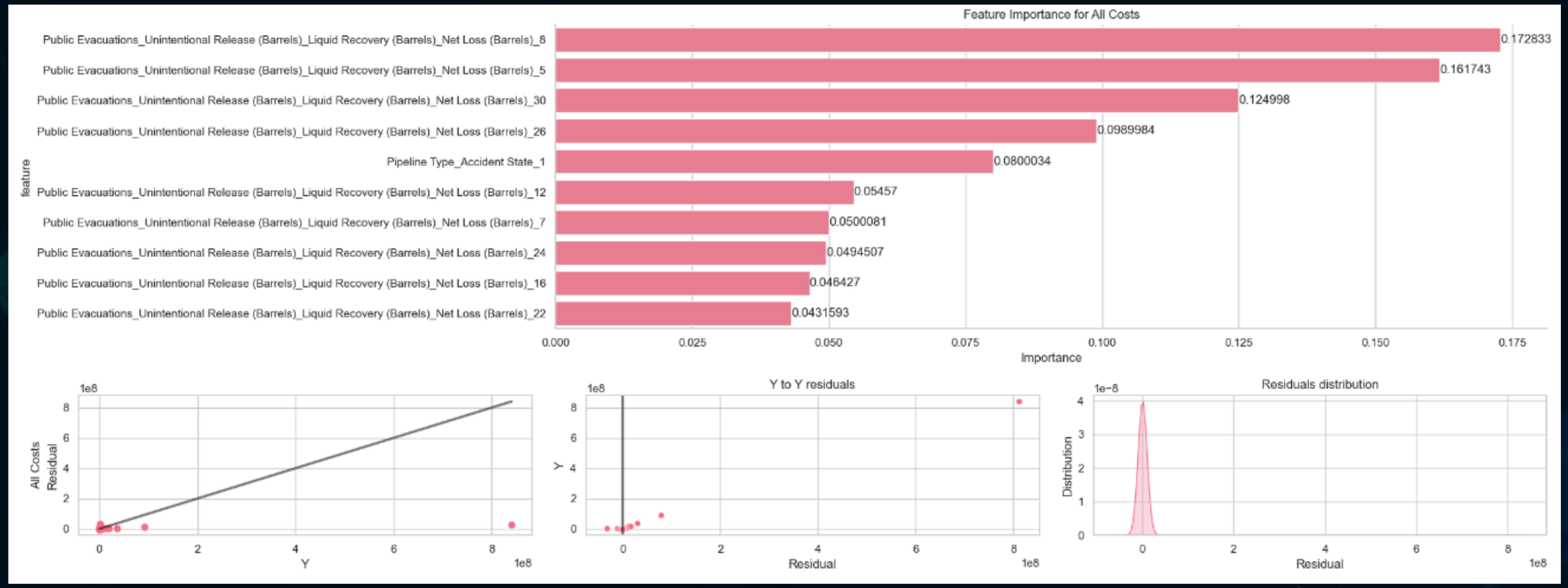


R2: 0.166

MSE: 5.38 x 10^10

MAE: 5.46 x 10^4

ALL COSTS



R2: 0.066 *MSE: 1.19 x 10^15* *MAE: 2.18 x 10^6*

THANK YOU!

FOR YOUR ATTENTION