Enhanced Vehicle Efficiency Analysis – Technical Report

This report documents the data, methods, modeling, fine-tuning, and results for predicting vehicle efficiency across EV and ICE cohorts.

# Abstract

We analyze a vehicle dataset, engineer domain-informed features, select predictive subsets, train and evaluate multiple regressors, and fine-tune the top candidates. The pipeline outputs ranked results, serialized models, and comprehensive visualizations. This document explains the process and summarizes results, including any tuned-vs-base improvements.

# Data & Preprocessing

* Input CSV: data/vehicle\_comparison\_dataset\_030417.csv
* Target: efficiency = mileage\_km / energy\_consumption (computed)
* Outlier removal: IQR filter on efficiency
* Cohorts: split by vehicle\_type → EV and ICE; drop vehicle\_type thereafter
* Clipping: co2\_emissions\_g\_per\_km clipped at lower=0

# Feature Engineering & Selection

* Engineered ratios: power\_efficiency, storage\_per\_torque, cost\_efficiency
* Maintenance & lifespan: maintenance\_per\_year, maintenance\_per\_torque, lifespan\_torque\_ratio
* Environmental: eco\_efficiency (EV), green\_performance (EV), emission\_intensity (ICE), emission\_per\_storage (ICE)
* Categories → codes: torque\_category\_num, acceleration\_category\_num
* Polynomial/interactions: torque\_squared, cost\_squared, torque\_x\_lifespan, storage\_x\_lifespan
* Transforms/normalization: log\_\* features and normalized\_\* features
* Leakage prevention: exclude mileage\_km and energy\_consumption from features
* Selection: correlation thresholds (EV>0.02, ICE>0.03), variance filter, multicollinearity pruning (|corr|<0.85), cap 20 features

# Modeling & Evaluation

* Models: Linear, Ridge, Lasso, Random Forest, Gradient Boosting, Decision Tree; optional XGBoost, LightGBM, CatBoost
* Preprocessing: PowerTransformer for linear models; raw passthrough for tree/boosting
* Validation: 5-fold KFold (shuffle, random\_state=42), metric: MAE
* Hold-out: 80/20 train/test split; metrics: MAE, RMSE, R²
* Ranking: configurable by R² or MAE (default R²)

# Fine-Tuning Strategy

* Select top 2 models per cohort (by chosen rank metric)
* Boosting models: Optuna search if available (trials configurable), fallback to sklearn
* Non-boosting: RandomizedSearchCV by default, or GridSearchCV
* Tuning metric: MAE or R² (default MAE)

# Results Summary

Best EV Model: LightGBM (Test R²: 0.003139743802415995)

Best ICE Model: XGBoost (Test R²: -0.027397995790911622)

## EV Model Rankings (Top 5)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unnamed: 0 | cv\_mae\_mean | cv\_mae\_std | train\_mae | train\_r2 | test\_mae | test\_rmse | test\_r2 | features\_used |
| LightGBM | 2185.2304799038675 | 120.28560706743316 | 1900.278969215036 | 0.0161239207913248 | 2038.2033869018917 | 2524.6968918162793 | 0.0031397438024159 | 7.0 |
| CatBoost | 2076.6190266239328 | 101.13315969158197 | 1885.0428436252416 | 0.0311369485496748 | 2039.545692491102 | 2526.838046962327 | 0.0014481842623957 | 7.0 |
| Random Forest | 2026.0014103740607 | 101.8764204858578 | 1795.89059603618 | 0.1229589866862321 | 2023.0345347361977 | 2527.313074161853 | 0.001072707995248 | 7.0 |
| XGBoost | 2166.0068613547473 | 137.40428497657953 | 1899.3562627825147 | 0.0167020854948504 | 2040.023243469181 | 2527.5749584595587 | 0.0008656763228492 | 7.0 |
| Gradient Boosting | 2024.032936117876 | 85.3134371617958 | 1899.972500953693 | 0.0188556189445214 | 2043.3519919343712 | 2528.4506351259806 | 0.0001732575532905 | 7.0 |

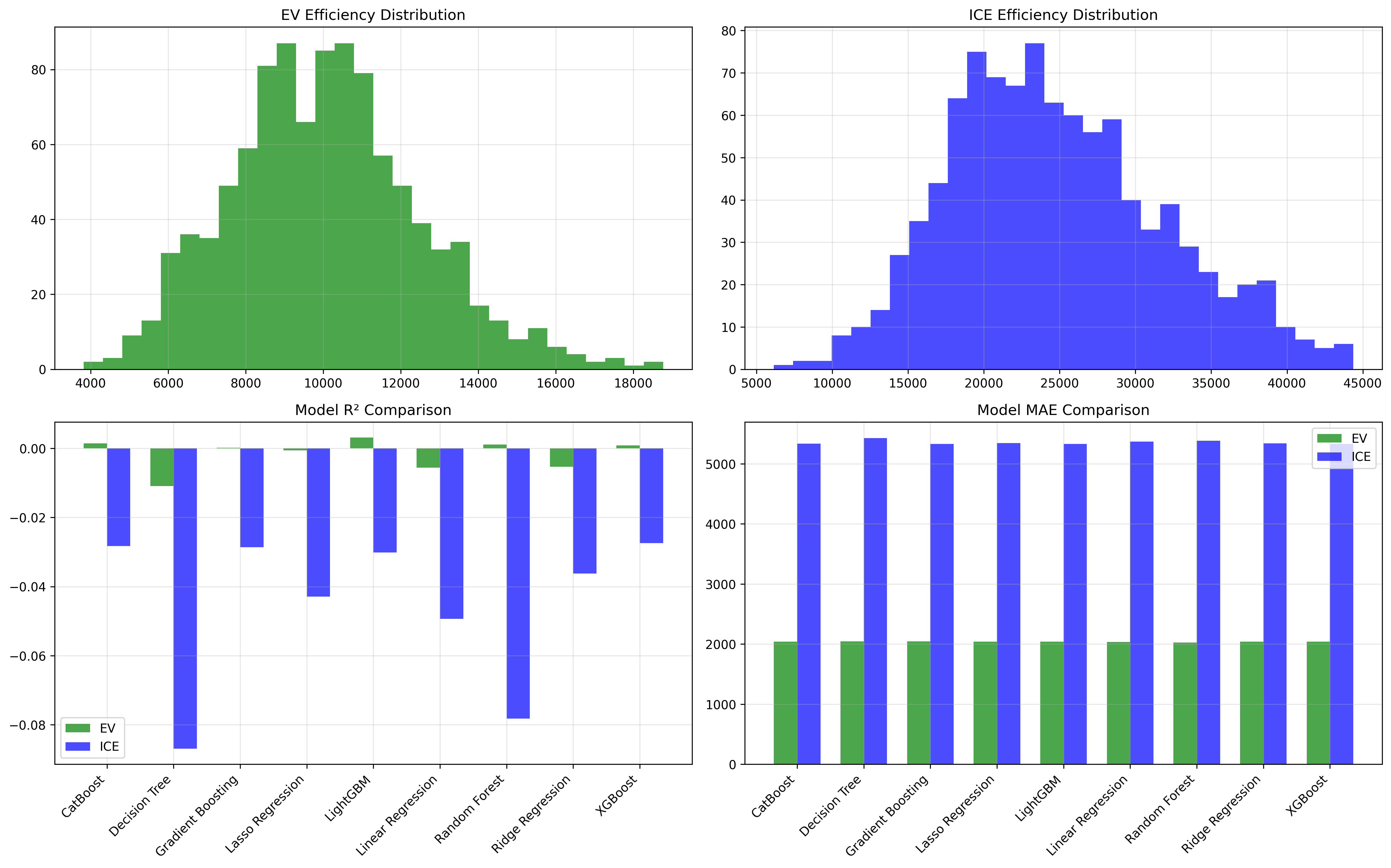
## ICE Model Rankings (Top 5)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unnamed: 0 | cv\_mae\_mean | cv\_mae\_std | train\_mae | train\_r2 | test\_mae | test\_rmse | test\_r2 | features\_used |
| XGBoost | 6885.776490802518 | 334.26978159350824 | 5788.0529214468625 | 0.0142997695716727 | 5330.633211288183 | 6627.677052375655 | -0.0273979957909116 | 5.0 |
| CatBoost | 6376.821707306675 | 311.0111459825889 | 5804.002942542261 | 0.0081239711791722 | 5333.690944432039 | 6630.480106719207 | -0.0282672178070877 | 5.0 |
| Gradient Boosting | 6157.7732460564175 | 353.67881561750096 | 5780.965539710824 | 0.0173204170704952 | 5331.997244813508 | 6631.533571872517 | -0.0285939904446268 | 5.0 |
| LightGBM | 6675.805816648157 | 241.79190980292103 | 5780.961331948532 | 0.0163211496526102 | 5327.974156167079 | 6636.481093083114 | -0.0301293484053521 | 5.0 |
| Ridge Regression | 5834.843299820534 | 184.52721080689085 | 5805.542462715296 | 0.0061105896488415 | 5340.504710209606 | 6656.007166708573 | -0.0362000270727385 | 5.0 |

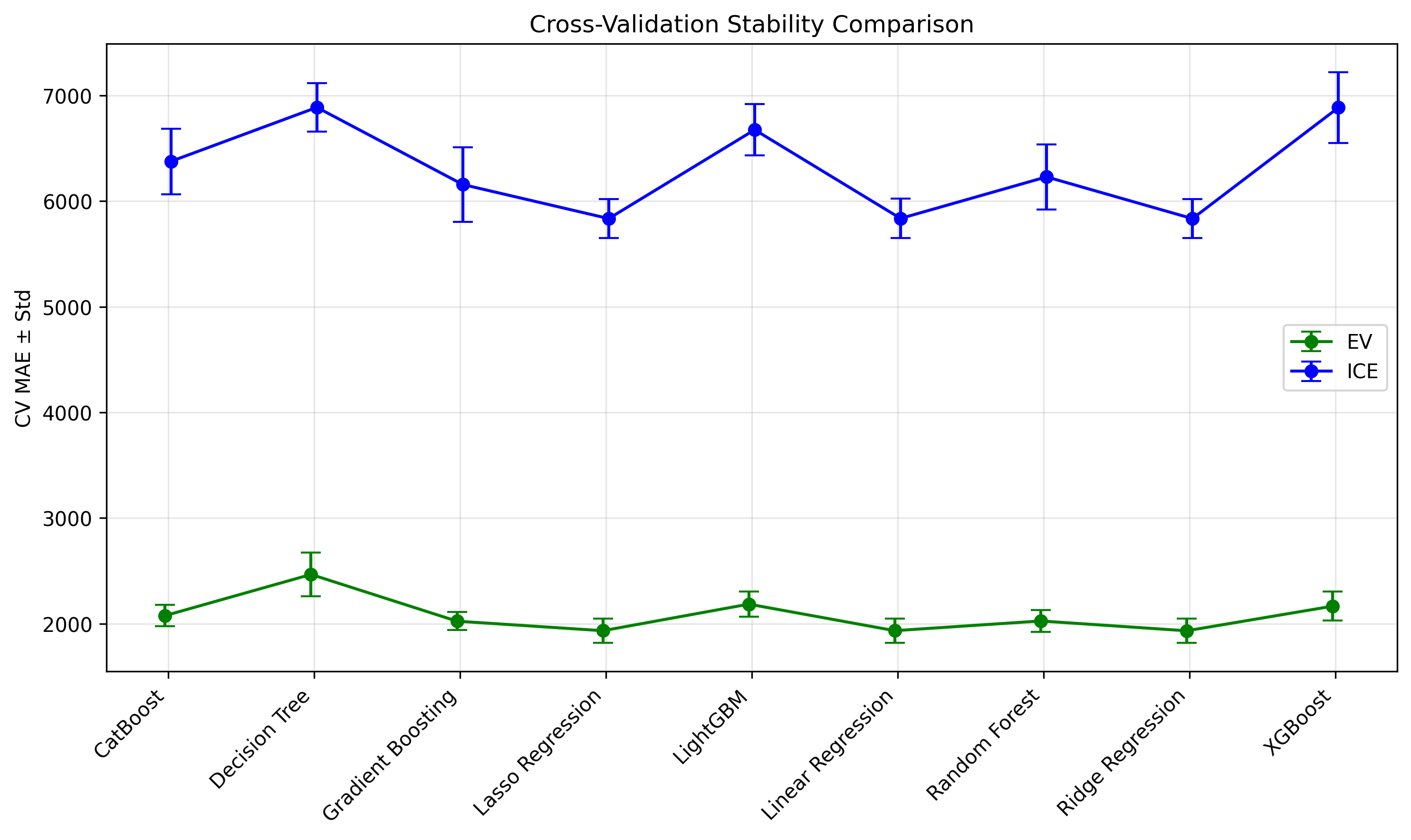
# Visualizations

The following figures summarize distributions, model performance, and correlation structures.

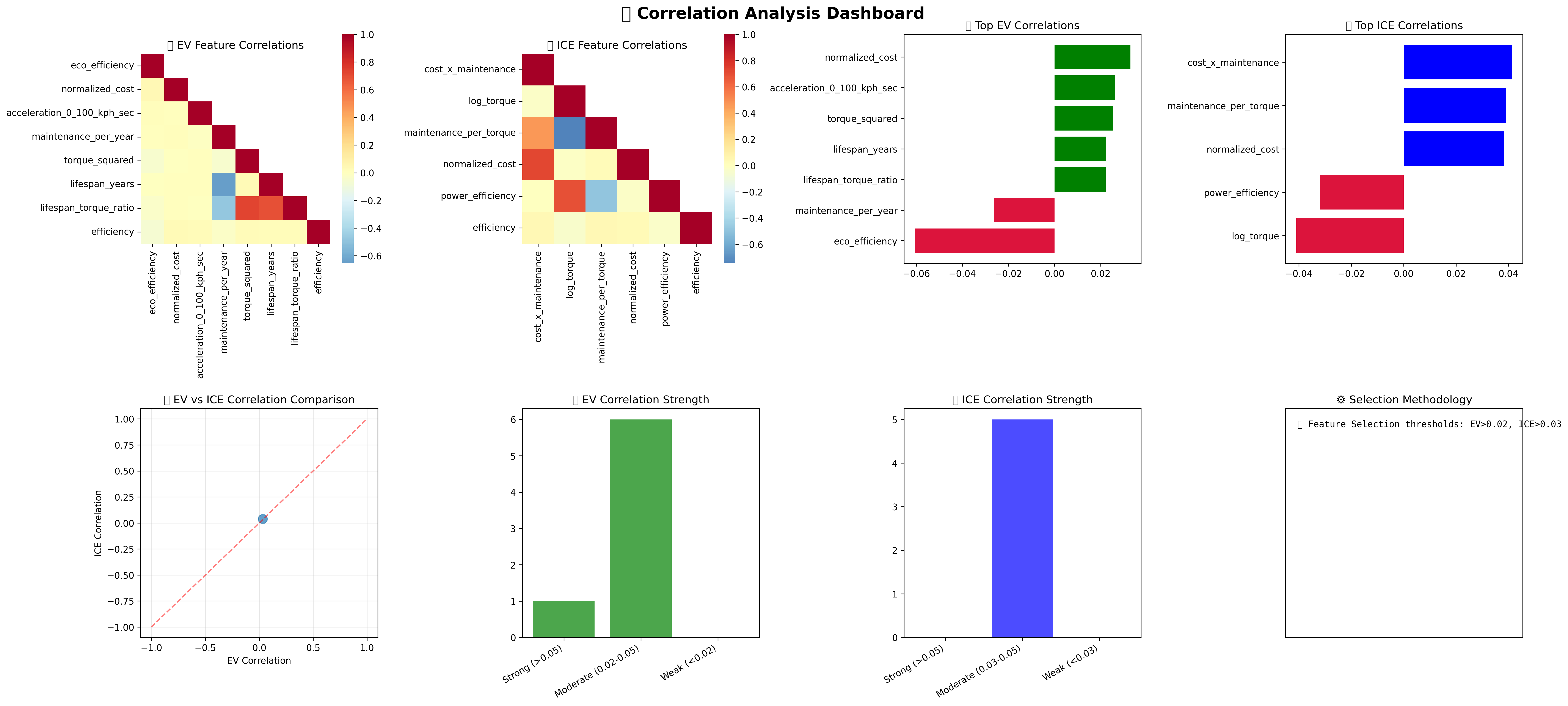
## Main Dashboard

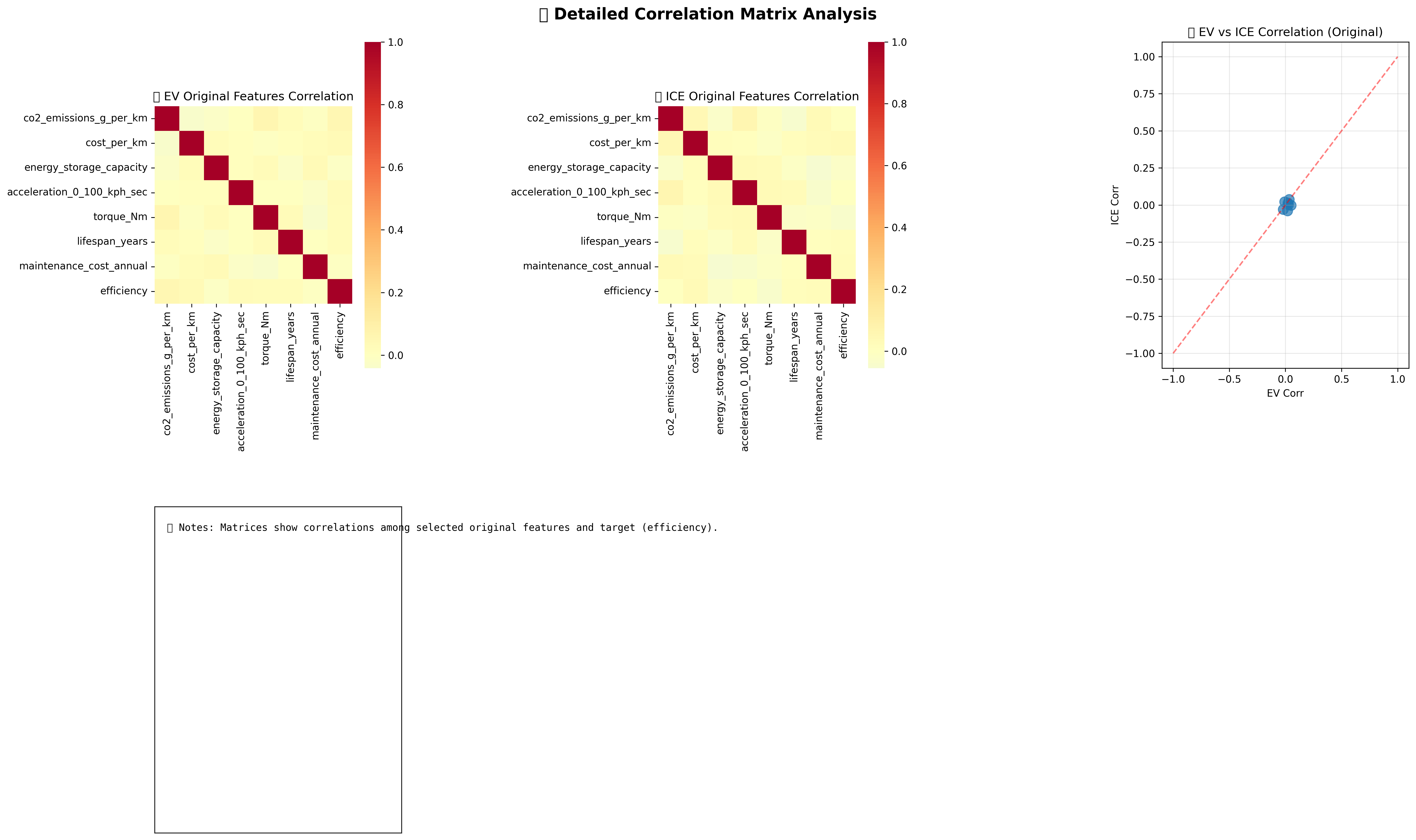


## CV Stability



## Correlation Dashboards





# System & Reproducibility

- python\_version: 3.12.9 | packaged by Anaconda, Inc. | (main, Feb 6 2025, 12:55:12) [Clang 14.0.6 ]

- numpy\_version: 1.26.4

- pandas\_version: 2.1.4

- sklearn\_version: 1.4.2

Random state used in KFold and model constructors where applicable.