Model Optimization Report

Overview

This report summarizes the results of multiple regression models applied to a feature-engineered dataset for predicting the `actual_time` target variable. The workflow included data preprocessing, model training, hyperparameter tuning, and performance evaluation using several metrics.

Data Preparation

- **Feature Engineering:** Loaded a processed dataset (`finaldata.csv`) with both numeric and categorical features.
- **Encoding:** All non-numeric features were label-encoded.
- **Scaling:** StandardScaler was used to normalize features before model fitting.
- **Train-Test Split:** 80% training, 20% testing, with a fixed random seed for reproducibility.

Evaluation Metrics

- **MAE** (Mean Absolute Error)
- **MSE** (Mean Squared Error)
- **RMSE** (Root Mean Squared Error)
- **R^2 Score** (Coefficient of Determination)
- **Relative Accuracy (±20%)**: Percentage of predictions within ±20% of the true value

Model Performance Comparison

Model	MAE	MSE	RMSE	R^2 Score	Relative Accuracy (±20%)
Ridge Regression	0.8111	2.882	1.6977	0.971	63.58 %
Linear Regression	0.8112	2.882	1.6977	0.971	63.57%
Lasso Regression	0.8054	2.9061	1.7047	0.9708	64.69%
ElasticNet Regression	0.8122	2.9226	1.7096	0.9706	63.89%
Random Forest	0.1407	0.0911	0.3019	0.9991	98.06 %
XGBoost	0.1745	0.1127	0.3357	0.9989	95.67 %

Key Findings

- **Best Performance:**
    - **Random Forest** achieved the lowest errors and highest R^2 score, with nearly perfect relative accuracy (98.06%).
    - **XGBoost** also performed exceptionally well, slightly behind Random Forest.
- **Linear Models:**
    - **Ridge, Lasso, ElasticNet, and Linear Regression** performed similarly, with considerably higher errors and lower relative accuracy compared to ensemble methods.
- **Hyperparameter Tuning:**
    - **Ridge Regression:** Best alpha (λ) found was 0.01.
- **Warnings:**
    - Lasso and ElasticNet models triggered convergence warnings, suggesting further tuning or more iterations may be needed for optimal results.

Recommendations

- **Deploy Random Forest** for production use given its superior accuracy and robustness.
- **Consider XGBoost** as an alternative, especially if interpretability or speed is a concern.
- **Further Tuning:** For linear models, increase the number of iterations or adjust regularization parameters to address convergence issues.

Conclusion

Ensemble models, particularly Random Forest and XGBoost, significantly outperformed linear regression approaches on this dataset. Model selection should prioritize these algorithms for the most accurate predictions.