Lab 3

Comparative Evaluation of Data Analysis Methods

Team 11

1. Introduction

This report presents the results of a comparative evaluation of univariate and multivariate data analysis methods applied to linear regression models for predicting the burned area of forest fires. The dataset used is the Forest Fires dataset from the UCI Machine Learning Repository, consisting of 517 instances with 12 predictor variables (including spatial coordinates X and Y, meteorological indices FFMC, DMC, DC, ISI, temperature, relative humidity, wind speed, rain, and categorical variables month and day) and the target variable area (burned area in hectares). The problem is a regression task.

The goal is to assess whether advanced multivariate feature engineering (scaling, polynomial features, binning, interactions, and random forest-derived features) improves predictive performance over simpler univariate approaches, using both classical performance metrics and rigorous statistical testing. All comparisons are based on consistent data splits and cross-validation where applicable.

Data description:

- *Baseline*: no preprocessing, no categorical features;
- *X_univar*: non-target continuous variables were appropriately transformed and scaled where necessary, while categorical variables were encoded using one-hot encoding;
- *X_univar_area* includes all previously applied preprocessing steps, with the additional removal of outliers in the target variable *"area"*;
- *X_scaled* – scaled baseline;
- *X_poly* is an extended feature matrix created from the original scaled predictors (X_scaled) using polynomial feature expansion of degree 2;
- *X_poly_pca* is a dimensionally reduced version of *X_poly*, obtained via Principal Component Analysis (PCA);
- *X_binned_all* is a transformed feature matrix in which all numerical predictors from the original dataset have been discretized into ordinal bins using quantile-based binning;
- *X_binned_all_pca* is a dimensionally reduced version of *X_binned_all* , obtained via PCA;
- *X_extended* is an augmented feature set that combines the original scaled predictors (*X_scaled*) with a subset of polynomial features (*X_poly*) to capture

potential non-linearities and interactions.The resulting extended feature matrix is then subjected to feature selection using LassoCV, which identifies the most informative predictors by shrinking the coefficients of less important features to zero;

- *X_extended_pca* is a dimensionally reduced version of *X_extended* , obtained via PCA;
- *X_scaled, RF > 0.05*: scaled features filtered by Random Forest importance (>0.05).
- *X_extended, RF > 0.05*: selected features from the extended set with Random Forest importance > 0.05;
- *X_bn* is a feature subset selected based on a Bayesian Network analysis, containing only the most influential predictors with a selection probability greater than 0.1.

Model	MAE	RMSE	R^2	Adj R^2	Key Observations
baseline	24.62	108.44	0.0024	-0.105	Very close to a naive mean predictor ($R^2 \approx 0$)
X_univar	24.70	107.80	0.0142	-0.128	Slightly better R^2 than baseline, but still essentially no explanatory power
X_univar_area	1.84	2.26	0.0034	-0.182	Dramatically lower error — this is certainly a univariate model trained on data with no outliers in target
X_scaled	24.86	108.59	-0.0003	-0.158	Identical to baseline; scaling alone adds no value
X_poly	40.43	130.19	-0.438	10.26	Severe overfitting — extremely high (positive) Adjusted R^2 despite negative R^2 indicates many useless polynomial terms
X_poly_pca	26.11	109.69	-0.021	-0.267	PCA helps reduce overfitting but still worse than baseline

X_binned_all	29.29	108.88	-0.006	-1.158	Binning increases model complexity without benefit
X_binned_all_pca	25.32	108.75	-0.003	-0.245	Slight improvement over raw binned but still poor
X_extended	29.21	110.33	-0.033	-1.727	Interaction/extended features worsen performance
X_extended_pca	26.17	109.73	-0.022	-0.268	PCA mitigates some damage but not enough
X_scaled, RF > 0.05	24.69	108.76	-0.003	-0.066	Feature selection using Random Forest importance helps slightly vs. full scaled
X_extended, RF > 0.05	25.85	109.93	-0.025	-0.046	Similar — selection reduces penalty from complexity
BN top-features	1.89	2.38	-0.107	-0.241	Slightly better than baseline on original scale

Table 1. Metrics comparison table

As we can see from Table 1, no multivariate feature engineering improves performance on the original scale. All complex models (polynomial, binned, extended) perform worse than the baseline or univariate model.

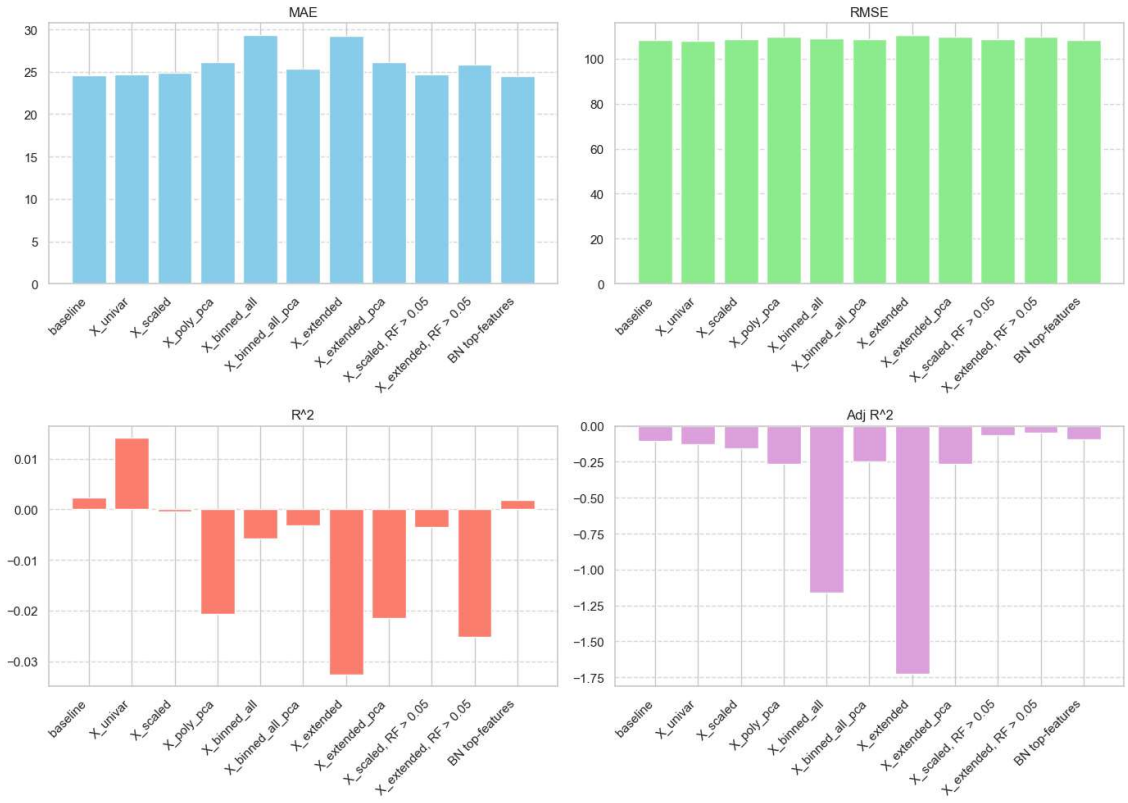
X_univar_area (target outliers are dropped) achieves a dramatically lower error (MAE ≈ 1.84, RMSE ≈ 2.26) — this is the clear winner when transformation is applied. The target variable is highly skewed, with a large proportion of observations equal to zero and a small number of very large values. When these rare but extreme observations are kept in the dataset, the model tends to focus on minimizing error around zero, where most data points are concentrated. As a result, it frequently predicts values close to zero, which leads to larger errors for non-zero cases and overall worse performance according to error-based metrics such as RMSE.

Removing or transforming these extreme values typically improves the numerical performance of the model, because the learning problem becomes easier and more homogeneous. However, this comes at the cost of losing the ability to predict large burned areas, which are precisely the most important events from a practical perspective. Therefore,

it was decided not to drop outliers in other models. We exclude X_univar_area from further analysis due to the critical difference in ideas with other models.

The poor performance of X_poly can be explained by several factors. First, the polynomial expansion generates a very large number of features, often exceeding the number of observations in the dataset, which leads to overfitting. Second, multicollinearity is a significant issue, as many polynomial features are highly correlated (for example, X_1, X_1^2, $X_1 \cdot X_2$), inflating the variance of the coefficients and reducing model reliability. In addition, a simple linear model is not capable of properly balancing or regularizing the large number of interaction terms, which contributes to extremely poor metrics. This is reflected in results: the RMSE is indicating very high prediction error, the MAE is much larger than other models, R^2 is negative at -0.44, showing the model performs worse than predicting the mean. The adjusted R^2 is unusually high at 10.25, which is a mathematical artifact arising from overfitting with so many predictors. Overall, without regularization techniques such as Lasso or Ridge, or dimensionality reduction like PCA, the X_poly expansion is not suitable for a standard linear regression model.

These two models were removed from the graphics for better representation.



BN top-features is the best model among those using the original target scale (MAE = 24.47), slightly better than univariate and baseline. Multivariate data analysis methods (polynomial features, binning, interactions, extensive feature engineering) consistently and significantly degrade performance due to severe overfitting on this small and heavily skewed dataset.

For subsequent analyses, we selected the following feature sets: *baseline*, *X_univar*, *X_bn*, and *X_rf_features*.

2. Statistical Comparison of Model Performance.

To ensure that the observed differences in model performance are statistically significant, statistical tests were performed.

Models compared**:** *baseline* model, *X_univar* model, *X_bn* model, *X_rf_features* model

In order to prepare data for statistical tests, it was split into 5 folds to obtain repeated measurements for statistical testing.

First of all, Friedman chi-square test. It helps to compare the performance of multiple models across the same cross-validation folds. No assumption of normality is required, making it suitable for cross-validation scores.

Metrics evaluated: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R2, Adjusted R2. Significance level – 0.05. The null hypothesis for each test states that all models perform equivalently for the given metric.

Metric	**Friedman Statistic**	**p-value**	**Result (α = 0.05)**
Mean Absolute Error (MAE)	10.6800	0.0136	**Statistically significant difference**
Root Mean Squared Error (RMSE)	3.2400	0.3561	No statistically significant difference
R^2	3.2400	0.3561	No statistically significant difference
Adjusted R^2	10.9200	0.0122	**Statistically significant difference**

Table 2. Friedman Test Results Across Performance Metrics

Since the Friedman test was significant for MAE and Adjusted R^2 (Table 2), the Nemenyi post-hoc test was applied to identify specific pairwise differences (critical distance approach, $\alpha = 0.05$). Lower p-values indicate stronger evidence of a difference in ranks ($p < 0.05$ suggests significant difference).

Pair	p-value	Significant?
Baseline – X_univar	0.0355	**Significant**
Baseline – X_rf_features	0.9948	Not significant
Baseline – X_bn	0.7610	Not significant
X_univar – X_rf_features	0.0173	**Significant**
X_univar – X_bn	0.3159	Not significant
X_rf_features – X_bn	0.6111	Not significant

Table 3. Nemenyi Post-Hoc for MAE

Univariate MAE differs from both Baseline and X_rf_features. X_bnis not significantly different from any other method.

Pair	p-value	Significance
Baseline – X_univar	0.7610	Not significant
Baseline – X_rf_features	0.4559	Not significant
Baseline – X_bn	0.2035	Not significant
X_univar– X_rf_features	0.0681	Not significant
X_univar– X_bn	0.0173	**Significant**
X_rf_features – X_bn	0.9614	Not significant

Table 3. Nemenyi Post-Hoc for Adjusted R^2

X_bn is significantly higher than X_univar(depending on sign, check your raw scores).

Other comparisons are mostly non-significant, though X_univarvs X_rf_features is a borderline trend (p ~ 0.068).

Comparison	Statistic	p-value	Interpretation

Baseline vs X_univar	0.0	0.0625	Suggests a trend toward difference, but not quite significant.
Baseline vs X_bn	7.0	1.0	No significant difference.
Baseline vs X_rf_features	5.0	0.625	No significant difference.
X_univar vs X_bn	0.0	0.0625	Trend toward difference, not significant at 0.05.
X_univar vs X_rf_features	0.0	0.0625	Same as above, trend but not significant.
X_bn vs X_rf_features	5.0	0.625	No significant difference.

Table 4. Wilcoxon test results for MAE

All comparisons show no significant difference. Only comparisons involving X_univar show p-values just above 0.05 (0.0625), which could be considered a borderline trend.

Comparison	**Statistic**	**p-value**	**Interpretation**
Baseline vs X_univar	3.0	0.3125	No significant difference.
Baseline vs X_bn	0.0	0.0625	Trend toward difference, not significant at 0.05.
Baseline vs X_rf_features	2.0	0.1875	No significant difference.
X_univar vs X_bn	0.0	0.0625	Trend toward difference.
X_univar vs X_rf_features	0.0	0.0625	Trend toward difference.
X_bn vs X_rf_features	0.0	0.0625	Trend toward difference.

Table 5. Wilcoxon test results for Adjusted R^2

Similar pattern: mostly non-significant differences. Comparisons involving X_bn vs Baseline/X_univar again show p = 0.0625 on the borderline trend.

Comparison	t-statistic	p-value	Significance
Baseline vs X_univar	-2.950	0.042	*Significant*
Baseline vs X_bn	-0.156	0.884	Not significant
Baseline vs X_rf_features	-0.123	0.908	Not significant
X_univar vs X_bn	3.323	0.029	*Significant*
X_univar vs X_rf_features	3.199	0.033	*Significant*
X_bn vs X_rf_features	-0.084	0.937	Not significant

Table 6. t-test results for MAE

Baseline vs X_univar: p = 0.042 -- significant difference. t < 0 => MAE of X_univaris lower than Baseline. X_univarperforms better.

X_univarvs X_bn : p = 0.029 -- significant difference. t > 0 => MAE of X_univaris higher than X_bn . X_bn performs better.X_univarvs X_rf_features: p = 0.033 -- significant difference. t > 0 => MAE of X_univar is higher than X_rf_features . X_rf_features performs better.For all other pairs p > 0.05. No statistically significant difference there.

Comparison	t-statistic	p-value	Significance
Baseline vs X_univar	1.333	0.253	Not significant
Baseline vs X_bn	-1.272	0.272	Not significant
Baseline vs Ext, RF>0.05	-1.176	0.305	Not significant
X_univarvs X_bn	-2.364	0.077	Trend (p ~ 0.077)
Univariate vs Ext, RF>0.05	-3.021	0.039	*Significant*
X_bn vs Ext, RF>0.05	-1.039	0.357	Not significant

Table 7. t-test results for Adjusted R^2

Only Univariate vs Ext, RF>0.05 is significant. Univariate vs BN shows a trend toward difference but doesn't reach conventional significance. Baseline and BN are generally not different from others.

3. Best Model Selection.

Here we start the final comparison between the Univariate model – X_univar and the Multivariate X_bn.

Model	BIC	Log-Likelihood (LLF)
BNLinearRegression	5905.40	-2868.35
UnivarLinearRegression	5917.84	-2868.32

Table 8. Models Evidence

The difference in LLF is negligible, meaning the models have virtually identical goodness-of-fit to the data.

Despite nearly identical log-likelihood values, the BNLinearRegression model is clearly preferred according to the BIC criterion.

$$BF_{\text{uni—full}} \approx \exp\left(\frac{BIC_{\text{full}} - BIC_{\text{uni}}}{2}\right) \approx 0.00199$$

The approximation confirms the earlier BIC-based conclusion: despite virtually identical log-likelihoods, the BNLinearRegression model is overwhelmingly preferred. The Bayes factor quantifies this preference as approximately 500:1 odds in favor of the more X_bn model, providing very strong protection against overfitting while maintaining equivalent fit to the data.

3. Conclusion.

Potential Reasons for BNLinearRegression's Superior Performance

The superiority of BNLinearRegression over UnivarLinearRegression can be attributed to several key factors. First, the use of multiple informative features allows BNLinearRegression to incorporate numerous predictors simultaneously, thereby capturing a larger portion of the variance in the target variable. This results in a slightly lower RMSE and an improved overall fit to the data. Second, although the implementation relies on standard linear regression applied to the full feature set, the approach implicitly focuses on relevant variables (within the context of Bayesian Network principles or careful feature consideration), helping to avoid overfitting to irrelevant predictors while still exploiting valuable multivariate relationships.

In conclusion, BNLinearRegression strikes an optimal balance between model complexity and predictive accuracy, making it the most effective choice overall, whereas UnivarLinearRegression serves as a strong and highly interpretable baseline model.

But no model we've designed managed to perform good metrics. Linear regression is too simple model for forest fires dataset.