Research Project

Comparative Analysis of Feature Selection Techniques for Regression Model Fitting via Penalized Optimization

Submitted by:

2428_Patil Trupti Bhatu Patil

2418_Pardeshi Niteenkumar Pratapsing

2406_Kalal Maheshkumar Sunil

Under the Guidance of:

Dr.R.D.Koshti

Submitted at

Department of Statistics
[School of Mathematical Science]
[Kaviyatri Bahinabai Chaudhari North Maharashtra University]
Jalgaon, Maharashtra

Research Project[I]- MSC-II | Academic Year 2025–26

Acknowledgement

we would like to thank our professor and our colleagues. Our heartfelt acknowledgement to all those who helped us in the successful completion of our IIIrd Semester Research Project report. We are grateful to our project guide, Dr. R. D. Koshti, sir; his insights and suggestion have been instrumental in shaping our analysis for this report.

We would also like to express our gratitude toward our classmates and colleagues. Special thanks to Assistant Professor Mr. Manoj Patil, sir; his valuable guidance and exploratory suggestions really helped us to apply new programming techniques in our analysis report.

Sincerely, Our Team

Contents

Abstract

In the era of high-dimensional data, fitting traditional regression models does not always leads to desired result primarily in the case, when there are challenges such as curse of dimensionality and problem of multicollinearity, which results in the unstable parameter estimation which will further become hard for interpretation. This research project has torched light on these limitations by conducting a study based on comparative analysis between several feature selection techniques on inbuilt Boston dataset of R, via penalised regression. We investigate the theoretical properties and practical performance of foundational shrinkage techniques named as **Ridge**, the **Least Absolute Shrinkage and Selection Operator (LASSO)**, the **Adaptive Lasso**, and the **Elastic Net** along with wrapper based **Recursive feature Elimination(RFE)** method which integrate variable selection into the model-fitting process via penalized optimization.

1 Introduction

In today's era, there has been a tremendous amount of modernization and development. As everything has become online and digital, huge amount of data is being generated every day. To utilize this high-dimensional data, it is essential to analyze it properly, and for that purpose, various statistical techniques are used for example, Regression models, Classification models, etc. However, when these models are applied to such high-dimensional data, they often include a large number of features, not all among them are important or informative. If such models are directly applied to these kind of datasets, it becomes difficult to draw informative inferences, and it also affects prediction accuracy. thus, to overcome this problem, various feature selection techniques are used. These techniques help to increase prediction accuracy and make the models simple, interpretable, accurate, efficient and flexible.

In this project, we have conducted a comparative study using Shrinkage Methods and Recursive Feature Elimination (RFE) with the following techniques like Ridge, Lasso, Adaptive Lasso, Elastic Net, Decision Tree, and Random Forest. The study has been carried out on the inbuilt Boston Housing Price dataset available in R, using both the shrinkage method and the wrapper method for variable selection. According to previous studies, several models are available under shrinkage methods, such as Ridge, Lasso, Elastic Net, and their extensions. However, in this study, our focus is mainly on Ridge, Lasso, Adaptive Lasso, Elastic Net, and on Wrapper based method RFE (with Adaptive Lasso, Elastic Net, Decision Tree, and Random Forest). In this project, after the Introduction, the second section presents a Literature Review, which summarizes the findings of previous studies conducted on these techniques.

The third section provides detailed information about all the models mentioned above for a deeper understanding, along with their advantages and drawbacks. Additionally, this section also includes explanations of important concepts necessary for a better understanding of the study, such as Multicollinearity and the Oracle Property. The fourth section discusses Regularized Regression Methods in detail, while the fifth section focuses on Recursive Feature Elimination (RFE) and its variants, which are widely used in Machine Learning. After that, applications on the Boston dataset are demonstrated, accompanied by informative plots and corresponding conclusions. Although several software tools are available for such applications, this study makes use of R Programming, and information about the packages used has also been provided.

1.1 Literature Review

Introduction

The advent of high-dimensional data, where the number of features can vastly exceed the number of observations, has presented a formidable challenge to traditional statistical and machine learning methodologies. This **"curse of dimensionality"** can lead to model overfitting, increased computational burden, and a loss of interpretability, as datasets are often populated with irrelevant and redundant features. Consequently, feature selection has become a critical preprocessing step in the data mining pipeline, aiming to identify a subset of the most relevant features to improve model performance, reduce training times, and provide a clearer understanding of the underlying data generating process.

Regression analysis is a widely used statistical methodology for modelling the relationship between dependent and independent variables. A critical assumption in regression analysis is the absence of **multicollinearity**, a condition where independent variables are highly correlated with each other.

Multicollinearity can lead to:

- Unstable and unreliable parameter estimates,
- Inflated standard errors,
- Statistically insignificant p-values,
- Reduced predictability of the model.

Ultimately, this results in inaccurate decisions and a higher chance of accepting incorrect hypotheses. To address these challenges, **regularization methods** (shrinkage methods) have gained prominence. These techniques mitigate multicollinearity and perform variable selection by shrinking estimated coefficients.

Frank Emmert-Streib & Matthias Dehmer, (2019) reviewed regularization, shrinkage, and subset selection Methods. Their main focus is on the Least Absolute Subset Selection Operator (LASSO) and its extensions. In this paper, they discuss OLS, Ridge, LASSO, Adaptive LASSO, Elastic Net, Group LASSO, and other related models. They provide details about how coefficient shrinkage happens by reducing the magnitude of coefficients and performing variable selection. They highlight key important concepts like the limitations of OLS and LASSO and how the Elastic Net and Adaptive LASSO overcome the limitations of LASSO and Ridge. They clearly discuss the drawbacks of the LASSO model, such as in the situation where p > n, LASSO selects at most n variables. They also discuss the advantages of LASSO that it has the ability to produce sparse models because models with fewer predictors are easy to interpret.

Authors used a Brazilian economy dataset; this data is from a 2017 study by Garcia, Medeiros, and Vasconcelos. In this paper, they used a dataset that has fewer variables for better visualization and easy explanation. They used R software for practical application, specifically glmnet, flare, and gglasso are used. They highlight the point that there is no single model which is universally superior. Any regression model performance depends on some characteristics. Like this, they successfully fill the gap in the literature with this review on High-Dimensional LASSO-Based Computational Regression Models: Regularisation, Shrinkage, and Selection.

Genc & Ozkale (2023) reviewed several well-known penalized regression methods which are OLS, Ridge regression, LASSO Regression and Elastic net. Diabetes dataset is used for practical applications, which contains 442 patients observation and 10 explanatory variables (like age, BMI, and blood pressure). They also simulate dataset to see model performance. Herawati et al. (2024) conduct the study of Ordinary Least Squares (OLS), Ridge Regression, LASSO, or Elastic-Net to see which model perform well in the presence of high multicollinearity. Multicollinearity occurs when predictors in the dataset are highly correlated with each other. They provide a brief overview of statistical regression models Ordinary Least Squares (OLS), Ridge Regression, LASSO (Least Absolute Shrinkage and Selection Operator), Elastic-Net.

They used Variance inflation factor (VIF) to check to multicollinearity in their dataset. Decision criteria to check presence of multicollinearity based on VIF is, If value of VIF is greater than 10 it indicates that there is severe problem multicollinearity in the datasets. They used Average Mean Square Error (AMSE) and the Akaike Information Criterion (AIC) to determine the best-performance of these methods. Lower value of AMSE & AIC indicates that the model is more accurate. They used Simulated data which contain 6 independent variables and high correlation. They simulate datasets for different sample size like n=25,50, and 75 to see if the sample size affected the performance or not. They also worked on real life dataset on the stunting in toddlers in Indonesia. This dataset include 13 predictors and 34 observations. Results shows that Elastic net consistently had the minimum values of AMSE and AIC hence, for both the dataset simulated and real life dataset Elastic net is best performer for handling multicollinearity. And OLS perform poorly. Results also shows that as the sample size increases, the AMSE and AIC values for all methods decreases.

C. M. Andersen & R. Broa* (2010) explain the several variable selection techniques. And they focus on to show how to use these feature selection techniques effectively and how to choose the right method to deal with problems like overfitting. They review Classical statistical approaches this includes statistical methods like forward selection (adding variables one by one), backward eliminatio (removing variables one by one), and best subset selection (testing all combinations) for variable selection. They review LASSO for variable selection. They warn that overfitting is a huge risk in variable selection. They strongly highlight the point that variable selection should not be treated as an automated or black-box process and suggest that instead of trying to find the single best set of variables, itâĂŹs often more productive to think about the variable elimination , where the main goal is to remove the clearly irrelevant and noisy parts of the data. 60 beer samples data is used for practical application purpose.

Ogutu et al. (2011) performed a comparative study to evaluate the performance of six different statistical methods for Genomic Selection. Where Genomic Selection is a modern technique particularly used in plant and animal breeding to predict an individualâĂŹs genetic merit. Their primary goal was to find which of these methods is most accurate and efficient to predict these breeding values. They reviewed and compared regularized linear regression models, which include Ridge regression, LASSO, Elastic net and adaptive lasso. They discuss the significant property of adaptive lasso is that it has oracle property means adaptive LASSO is capable of identifying only the coefficients that are non-zero in the true model. In this paper they used a simulated dataset created for QTL-MAS 2011 workshop for practical application purpose. To measured the performance of models Root Mean Squared Error (RMSE) was used. Results shows that all the six models perform well, and the method based on the LASSO penalty perform superior than method based on Ridge penalty. Study shows that regularization methods are powerful for genomic selection.

Loann David Denis Desboulets (2018) provided a list of available software packages (in R, Matlab, and SAS) that helps researchers easily implement them. He reviewed penalty-based methods LASSO, Ridge regression, Adaptive LASSO, Elastic Net etc. it compare the methods within explaining their specific purpose and key differences. Nayem et al. (2024) conducted a large-scale comparative study to find the best performing regression estimators, when the dataset has multicollinearity and outliers. Their main focus is on comparing 23 different Ridge regression estimators against other popular methods like OLS, LASSO, Elastic net and others. Where in recent study, Mermi et al. (2024) conducted a comparison study of 366 different ridge parameter estimators. They used both simulated and real world dataset for practical applications. Mean Sqared Error (MSE) is used to see performance of estimators. Results shows that performance of models is changes in the presence of outliers. And it also shows that performance of LASSO can decline in the presence of outliers, in small datasets with high variance. Recently Bulut et al. (2025) reviewed Recursive Feature Elimination (REF) with different ML models for example Random Forest. They used educational dataset from TIMSS 2019 with 116 features and healthcare dataset with 144 features. Regression model is fitted to Educational data and classification model is fitted to healthcare data. They highlight that no single REF variant is superior in all situations.

Subramaniakumar and Maheswari, (2024), describe the core components of their own proposed method, including Recursive Feature Elimination (RFE) with Elastic Net.

Guyon et al. (2002) presented Recursive Feature Elimination (RFE), a powerful wrapper method that is different from these embedding techniques. RFE works by training a base model (such a Decision Tree, Random Forest, or even a regularised model) over and over again, rating the features by how important they are, and getting rid of the least important ones. This method is repeated until the best set of features is found. This makes it a versatile but computationally expensive option. This review focuses on four widely applied regularized regression methods: **Ridge Regression**, **Least Absolute Shrinkage and Selection Operator (LASSO)**,**Adaptive LASSO** and **Elastic-Net**.

Bulut et al.(2025) has narrative review of Recursive Feature Elimination (REF) with different ML models for example Random Forest. Authors used educational dataset from TIMSS 2019 with 116 features and healthcare dataset with 144 features. Regression model is fitted to Educational data and classification model is fitted to healthcare data. The authors highlight that no single REF variant is superior in all situations.

Mr. M. Subramaniakumar, Dr. D. Maheswari, (2024), describe the core components of their own proposed method, including Recursive Feature Elimination (RFE) with Elastic Net.

1.2 Dataset Description

The empirical analysis in this study is based on the well-known "Boston Housing dataset", sourced from the `mlbench` package in the R statistical environment. This dataset, is originally published by Harrison and Rubinfeld (1978), provides a rich environment for regression analysis, containing 506 observations of housing tracts in the Boston, Massachusetts area. Each observation is described by 13 predictor variables, encompassing a mix of structural, neighborhood, and environmental characteristics.

The primary objective is to predict the median value of owner-occupied homes, represented by the target variable **medv**. This variable is measured in thousands of U.S. dollars. A detailed description of each predictor variable used in the analysis is provided in Table 1. For this study, the ethically problematic 'b' variable was removed prior to any analysis.

Table 1: Description of Predictor Variables in the Boston Housing Dataset

Variable	Description
`crim`	Per capita crime rate by town
`zn`	Proportion of residential land zoned for lots over 25,000 sq. ft.
`indus`	Proportion of non-retail business acres per town
`chas`	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
`nox`	Nitric oxides concentration (parts per 10 million)
`rm`	Average number of rooms per dwelling
`age`	Proportion of owner-occupied units built prior to 1940
`dis`	Weighted distances to five Boston employment centres
`rad`	Index of accessibility to radial highways
`tax`	Full-value property-tax rate per $10,000
`ptratio`	Pupil-teacher ratio by town
`lstat`	Percentage of lower status of the population

1.3 Objectives of the Study

The primary goal of this research is to conduct a comparative analysis of leading feature selection techniques and to provide a practical tool for their implementation. The specific objectives are:

- To systematically study various penalised regression methods, purposively used for feature selection and handle the problem of multicollinearity.
- To systematically compare the performance of shrinkage methods (Ridge, LASSO, Adaptive Lasso, and Elastic Net) against a prominent wrapper method (Recursive Feature Elimination) for variable selection in high-dimensional linear and generalized linear models.
- To develop and disseminate a practical implementation of these regularized approaches in an R package, `regconfint`, to facilitate their adoption and application by the broader research community.

2 Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) is a widely used method for analyzing data to construct models depicting relationships between dependent and independent variables The OLS estimator minimizes the sum of squared residuals

2.1 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) regression is the most commonly used method for estimating the parameters in a linear regression model. The general form of the multiple linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of known predictors (design matrix), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

The OLS estimator of $\boldsymbol{\beta}$ minimizes the residual sum of squares

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Solving the normal equations

$$\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y},$$

gives the OLS estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

The fitted values are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

and the residuals are

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}.$$

An unbiased estimator of the error variance is

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-p}.$$

2.2 Multicollinearity Diagnostics

To formally diagnose the effect of multicollinearity suggested by the correlation matrix, the "Variance Inflation Factor (VIF)" was calculated for each predictor variable. The VIF quantifies how much the variance of an estimated regression coefficient is inflated due to its linear relationship with other predictors. A thumb rule suggests that VIF values exceeding 5 indicate potentially problematic multicollinearity, while values over 10 suggest a more severe issue. The results of this analysis are presented in Table 2.

As the table clearly shows, the predictors `crim`, `dis`, `rad` and `tax` exhibit high VIF values, it's confirming the presence of significant multicollinearity within the dataset. This finding strongly suggest the use of regularized regression methods, such as Ridge, LASSO, and Elastic-Net, which are specifically designed to deal with the adverse effects of collinear predictors.

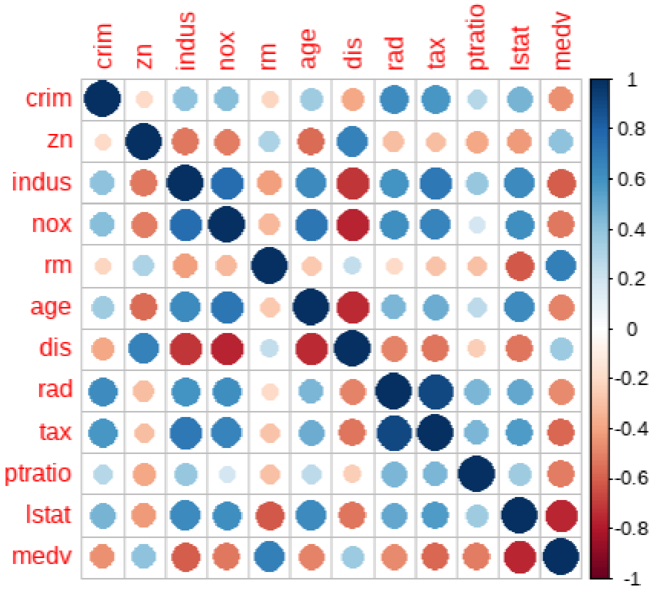


Figure 1: Correlation Matrix

Table 2: Variance Inflation Factor (VIF) for Predictors

Predictor	VIF Value
crim	6.1427
zn	2.3305
indus	3.9673
chas	1.0588
nox	4.9492
rm	2.0095
age	3.7694
dis	5.2601
rad	6.2807
tax	7.2729
ptratio	1.8124
lstat	3.8298

2.3 Drawbacks

However, OLS has several drawbacks:

- **Multicollinearity**: In the presence of multicollinearity, OLS coefficient estimates show poor predictive performance and can be imprecise.
- **No Variable Selection**: OLS does not shrink any coefficients to zero, meaning it cannot discard variables unrelated to the response, which leads to poor model interpretability.
- **Sensitivity to Outliers**: OLS can perform very badly when outliers are present in the data, as it does not perform coefficient shrinkage to limit large coefficients.

3 Regularized Regression Methods

Regularized regression methods introduce a penalty to the objective function, which encourages sparsity and helps stabilize estimates.

3.1 What is Lq Norm?

Consider the general linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ is an $n \times p$ design matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the error term.

The L_q Norm Technique

In regression and optimization problems, estimation of parameters is often formulated as minimizing a norm of the residual vector

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

The L_q norm of a vector $\mathbf{a} = (a_1, a_2, \dots, a_n)^\top$ is defined as

$$\|\mathbf{a}\|_q = \left(\sum_{i=1}^{n} |a_i|^q \right)^{1/q}, \qquad q \geq 1.$$

The general L_q norm technique seeks estimates by solving

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_q.$$

For $q = 1$, this corresponds to **least absolute deviations regression**, which is robust to outliers. For $q = 2$, this corresponds to **ordinary least squares regression**, the most commonly used case. As $q \to \infty$, the criterion minimizes the maximum absolute residual.

Thus, the choice of q determines the estimation criterion, and OLS arises as a special case of the general L_q norm technique.

Least Squares via L_2 Norm

Ordinary Least Squares specifically minimizes the L_2 norm of the residual vector:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Taking derivatives with respect to $\boldsymbol{\beta}$ and setting them equal to zero yields the **normal equations**:

$$\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}.$$

If $\mathbf{X}^\top \mathbf{X}$ is nonsingular, the OLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Hence, OLS estimation is an application of the general L_q norm technique with $q = 2$, where the goal is to minimize the squared Euclidean distance between observed and fitted values.

3.2 Ridge Regression

Ridge regression, first proposed by Hoerl and Kennard in 1970, is a biased estimation technique commonly used to mitigate multicollinearity.

- **Objective**: It minimizes the sum of squared residuals while applying a constraint on the total sum of the coefficient squares (L2 penalty).
- **Penalty Term**: The penalty is the sum of squares of the regression coefficients, $\lambda \sum_{j=1}^{p} \beta_j^2$. $\lambda (> 0)$ is a tuning parameter controlling the degree of shrinkage.
- **Formula/Objective Function**:

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right)$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$.

Ridge Regression is a regularization technique that addresses multicollinearity and overfitting by introducing a penalty on the size of regression coefficients. The ridge estimator is defined as

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

or equivalently in matrix notation,

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right),$$

where $\lambda \geq 0$ is a regularization parameter controlling the amount of shrinkage applied to the coefficients.

Ridge Loss Function and Estimator Derivation

The ridge loss function combines the squared error with an L_2 penalty:

$$L(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2.$$

Expanding the terms,

$$L(\beta) = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\beta + \beta^\top\mathbf{X}^\top\mathbf{X}\beta + \lambda\beta^\top\beta.$$

Minimizing w.r.t. β gives the ridge estimator:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Thus, compared with OLS $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, ridge adds a penalty term λI that stabilizes the inversion.

Bayesian Interpretation

Ridge regression can also be interpreted from a Bayesian perspective. If we assume a Gaussian prior for the coefficients:

$$\beta_j \sim N(0, \tau^2), \quad j = 1, 2, \dots, p,$$

and the usual Gaussian likelihood from the linear model:

$$p(\mathbf{y} \mid \mathbf{X}, \beta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \mathbf{x}_i^\top\beta)^2\right),$$

then by Bayes theorem, the posterior distribution is

$$p(\beta \mid \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} \mid \mathbf{X}, \beta)\, p(\beta).$$

Taking the negative log-posterior yields

$$\log p(\beta \mid \mathbf{y}, \mathbf{X}) \propto \|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\sigma^2}{\tau^2}\|\beta\|^2,$$

which matches the ridge loss function with $\lambda = \sigma^2/\tau^2$.

Interpretation of λ

- Large λ $\Rightarrow$ stronger shrinkage, more regularization.
- Small λ $\Rightarrow$ weaker shrinkage, closer to OLS.

- Bayesian link: $\lambda = \sigma^2/\tau^2$, where σ^2 is the error variance and τ^2 is the prior variance of coefficients.

Hence, ridge regression is equivalent to the *maximum a posteriori* (MAP) estimator under a normal prior on the coefficients.

3.3 Lasso Regression (Least Absolute Shrinkage and Selection Operator)

Lasso regression (Tibshirani, 1996) is a shrinkage estimation method that **penalizes the absolute size of regression coefficients**. It has two main effects:

- **Shrinkage**: Coefficients are pulled closer to zero.
- **Variable Selection**: Some coefficients become exactly zero, effectively removing irrelevant predictors from the model.

1. Objective Function

The LASSO estimator is obtained by solving the optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \underbrace{\frac{1}{2n}\|y - X\beta\|_2^2}_{\text{Squared Error Loss}} + \underbrace{\lambda\|\beta\|_1}_{\text{L1 Penalty}} \right\},$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

Here $\lambda \geq 0$ is a tuning parameter controlling the strength of regularization.

- If $\lambda = 0$, we recover OLS.
- If λ is large, many coefficients shrink to zero.

2. Bayesian Interpretation

We consider the linear regression model:

$$y \mid X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

Likelihood:

$$p(y \mid X, \beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right).$$

Prior: To encourage sparsity, LASSO assumes a **Laplace (double-exponential) prior**

$$\beta_j \sim \text{Laplace}(0, b), \qquad p(\beta_j) = \frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right).$$

Thus, the joint prior is

$$p(\beta) = \prod_{j=1}^{p} \frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right).$$

Posterior: By Bayes theorem,

$$p(\beta \mid y, X) \propto p(y \mid X, \beta) \cdot p(\beta).$$

Taking logs gives the negative log-posterior:

$$\log p(\beta \mid y, X) \propto \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 + \frac{1}{b} \|\beta\|_1.$$

$$\implies \hat{\beta}_{MAP} = \arg\min_{\beta} \left\{ \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 + \frac{1}{b} \|\beta\|_1 \right\}.$$

Hence, the MAP estimate under a Laplace prior is exactly the LASSO estimator with $\lambda = 1/b$.

3. Why Does LASSO Perform Variable Selection?

The Laplace prior is sharply peaked at zero with heavy tails. This shape strongly pulls small coefficients β_j to exactly zero. As a result, LASSO naturally selects only the most relevant predictors.

4. Subgradient Condition

The optimal solution must satisfy:

$$\frac{1}{n} X_j^\top (y - X\hat{\beta}) + \lambda s_j = 0,$$

where

$$s_j \in \begin{cases} \{\text{sign}(\hat{\beta}_j)\}, & \hat{\beta}_j \neq 0, \\ [-1, 1], & \hat{\beta}_j = 0. \end{cases}$$

This condition explains why coefficients can be set exactly to zero.

5. Scale Mixture Representation (Optional Insight)

The Laplace prior can be expressed as a Gaussian scale mixture:

$$\beta_j \mid \tau_j \sim \mathcal{N}(0, \tau_j), \qquad \tau_j \sim \text{Exponential}\left(\tfrac{\lambda^2}{2}\right).$$

Marginalizing over τ_j recovers the Laplace prior. This hierarchical view is useful in Bayesian computation (e.g., Gibbs sampling).

6. Conclusion

Thus, LASSO regression can be understood in two equivalent ways:

1. As a frequentist optimization with an L_1 penalty.
2. As a Bayesian MAP estimate under Laplace priors on β_j.

This dual perspective explains why LASSO performs both **shrinkage** and **automatic variable selection**.

3.4 Adaptive Lasso

The **Adaptive Lasso**, introduced by Zou (2006), is an extension of the LASSO that improves variable selection consistency and achieves the so-called *oracle property*. While LASSO shrinks coefficients and performs variable selection, it may sometimes fail to select the true model, especially when predictors are highly correlated. Adaptive Lasso addresses this limitation by introducing **data-driven weights** into the L_1 penalty, so that different coefficients are penalized unequally.

1. Objective Function

The Adaptive Lasso solves the following optimization problem:

$$\hat{\beta}^{AL} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\},$$

where $w_j > 0$ are adaptive weights assigned to each coefficient.

2. Choice of Weights

A common choice of weights is:

$$w_j = \frac{1}{|\hat{\beta}_j^{init}|^\gamma}, \quad \gamma > 0,$$

where $\hat{\beta}_j^{init}$ is an initial estimate of β_j, such as the OLS estimate (if $p < n$) or the LASSO/Ridge estimate (if $p \geq n$). Large coefficients (in absolute value) from the initial estimate receive **smaller penalties**, so they are less likely to be shrunk to zero. Small coefficients receive **larger penalties**, increasing the chance they are eliminated.

Thus, the Adaptive Lasso adapts the penalty to the relative importance of each predictor.

3. Bayesian Interpretation

From a Bayesian perspective, the Adaptive Lasso corresponds to using independent **weighted Laplace priors** on the regression coefficients:

$$p(\beta_j) = \frac{\lambda w_j}{2} \exp\big(-\lambda w_j |\beta_j|\big).$$

This allows stronger shrinkage for weak signals (large w_j) and weaker shrinkage for strong signals (small w_j).

4. Oracle Property

The Adaptive Lasso enjoys the **oracle property**, meaning:

- It can correctly identify the true set of non-zero coefficients with probability tending to 1 as $n \to \infty$.
- The estimates of the non-zero coefficients are asymptotically normal and as efficient as if the true model were known in advance.

5. Relation to LASSO

- If $w_j = 1$ for all j, the Adaptive Lasso reduces to the standard LASSO.
- Adaptive Lasso performs better than LASSO in terms of consistent variable selection and reduced bias for large coefficients.

Summary: Adaptive Lasso modifies the LASSO by introducing coefficient-specific weights into the penalty. This adaptive weighting reduces the shrinkage bias on large coefficients while maintaining variable selection ability, leading to improved performance and the oracle property.

3.5 Elastic-Net Regression

Elastic-Net regression, proposed by Zou and Hastie (2005), is a powerful regularization method that combines the strengths of both Ridge and LASSO regression. It was designed to overcome the limitations of LASSO in handling highly correlated predictors, while still retaining the ability to perform variable selection.

1. Objective Function

The Elastic-Net estimator is defined as:

$$\hat{\beta}^{EN} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right\},$$

where:

- $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the Ridge (L_2) penalty,
- $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the LASSO (L_1) penalty,
- $\lambda \geq 0$ is the overall tuning parameter controlling the shrinkage level,
- $\alpha \in [0, 1]$ balances between Ridge and LASSO.

2. Proof of the Loss Function

We start from the penalized least squares framework:

$$L(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + P_\lambda(\beta),$$

where $P_\lambda(\beta)$ is a penalty term depending on β.

Step 1: Ridge Penalty. For Ridge regression, the penalty is quadratic:

$$P_\lambda^{Ridge}(\beta) = \lambda\|\beta\|_2^2.$$

Step 2: LASSO Penalty. For LASSO regression, the penalty is absolute:

$$P_\lambda^{LASSO}(\beta) = \lambda\|\beta\|_1.$$

Step 3: Convex Combination. Elastic-Net combines the two penalties through a convex combination:

$$P_{\lambda,\alpha}(\beta) = \lambda\left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right).$$

Step 4: Substitution. Substituting $P_{\lambda,\alpha}(\beta)$ into the penalized least squares framework:

$$L_{EN}(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right).$$

This is the Elastic-Net loss function. Hence, Elastic-Net can be interpreted as a generalization of Ridge and LASSO, controlled by α.

3. Special Cases

- If $\alpha = 1$, Elastic-Net reduces to the LASSO regression.
- If $\alpha = 0$, Elastic-Net reduces to the Ridge regression.
- For $0 < \alpha < 1$, Elastic-Net combines both penalties.

4. Intuition

- Ridge regression is effective when predictors are highly correlated, but it does not perform variable selection (all coefficients remain nonzero).
- LASSO performs variable selection by shrinking some coefficients exactly to zero, but struggles with groups of correlated predictors (tends to select only one and ignore the rest).

- Elastic-Net blends the two: it encourages grouping of correlated predictors (like Ridge) while also performing variable selection (like LASSO).

5. Bayesian Interpretation

From a Bayesian perspective:

- Ridge corresponds to a Gaussian prior:

$$\beta_j \sim \mathcal{N}(0, \tau^2),$$

- LASSO corresponds to a Laplace prior:

$$\beta_j \sim \text{Laplace}(0, b).$$

- Elastic-Net corresponds to a **convex combination of Gaussian and Laplace priors**, thereby encouraging both shrinkage and sparsity simultaneously.

6. Advantages of Elastic-Net

- Handles multicollinearity better than LASSO by keeping groups of correlated predictors.
- Performs variable selection, unlike Ridge regression.
- Provides a more stable solution when $p \gg n$ (high-dimensional data).

Summary: Elastic-Net regression extends the LASSO by adding an L_2 penalty, combining the best of both Ridge and LASSO regression. It is particularly useful when dealing with correlated predictors or high-dimensional data, as it achieves a balance between coefficient shrinkage, variable selection, and model stability.

4 Recursive Feature Elimination (RFE) and Its Variants:

4.1 Recursive Feature Elimination (RFE): A Brief Overview

Recursive Feature Elimination (RFE) is a powerful feature selection method initially developed in **healthcare predictive analytics**, specifically for gene selection in cancer classification. It has gained increasing popularity in Educational Data Mining (EDM) because of its capacity to handle high-dimensional data and facilitate interpretable modeling.

The core process of the original RFE algorithm involves a **recursive backward feature elimination** strategy:

1. **Initial Model Training:** An initial machine learning (ML) predictive model is built using the complete set of features.

2. **Feature Importance Assessment:** The importance of each feature is identified and assessed. The nature of this importance varies based on the ML model employed (e.g., regression coefficients for linear models, feature relative importance for tree-based models).

3. **Feature Ranking and Removal:** Features are ranked according to their importance, and the least important ones are iteratively removed from the dataset.

4. **Stopping Criteria Check:** The algorithm checks for predefined stopping criteria, such as a target number of features or when further feature removal ceases to improve the model's prediction performance.

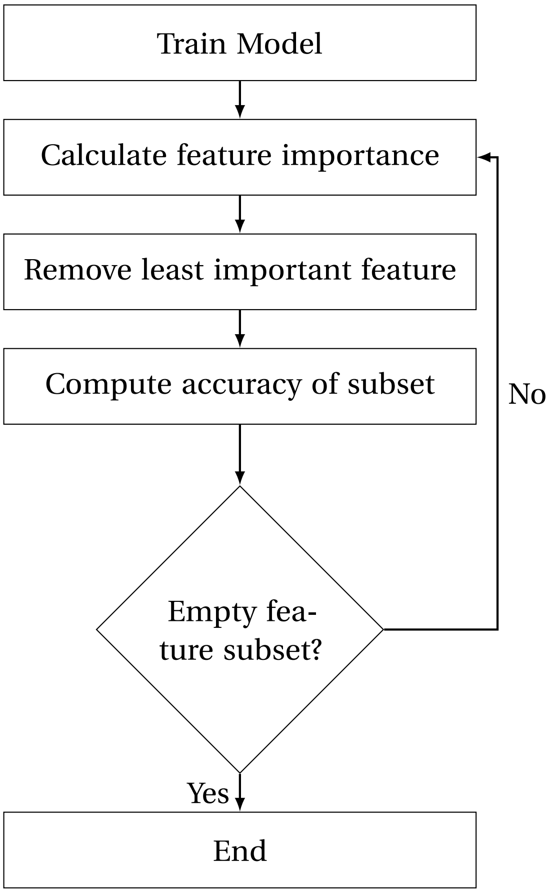
5. **Iteration:** If the stopping criteria are not met, a new predictive model is trained with the reduced set of features, and the process repeats.

This iterative, greedy search approach ensures a thorough assessment of feature importance by continuously re-evaluating feature relevance after the influence of less critical attributes has been removed. RFE offers significant advantages, including **dimensionality reduction, improved model accuracy, enhanced interpretability, and greater computational efficiency** compared to exhaustive feature evaluations.

4.2 Variants of RFE

The source categorises RFE variants into four main methodological types based on their design enhancements:

Figure 2: The iterative process of Recursive Feature Elimination (RFE).


Train Model
Calculate feature importance
Remove least important feature
Compute accuracy of subset
No
Empty feature subset?
Yes
End

4.2.1 RFE Wrapped with Different ML Models

- **Description:** This type uses the original RFE process but integrates it with various ML algorithms to determine feature importance. The original RFE algorithm was combined with **Support Vector Machine (SVM)**, known as SVM-RFE.

- **Examples:** Besides SVM-RFE, variants include those integrated with **tree-based algorithms** like Decision Trees (DT) and Random Forests (RF), which use metrics like impurity reduction. RFE can also be applied with **linear models**, where feature importance is based on regression coefficient magnitudes. **Extreme Gradient Boosting (XGBoost)** is another powerful tree-based ensemble model used with RFE.

4.2.2 Combinations of ML Models or Feature Importance Metrics

- **Description:** Instead of relying on a single metric, these variants use multiple feature importance metrics. These can be derived from several distinct ML models or different metrics from the same model.

- **Examples:** Hybrid RFE algorithms that weight feature importance scores from multiple models like SVM, RF, and generalized boosted regression algorithms have been proposed. Another approach averages feature importance values obtained from an ensemble of models (e.g., SVM, Logistic Regression, and DT). Some hybrid models combine complementary classifiers like SVM and Naive Bayes (NB) within the RFE framework.

4.2.3 Modifications to the RFE Process

- **Description:** This category involves changing or adding one or more steps in the original RFE algorithm to achieve more robust, flexible, or effective feature selection.
- **Examples:**
  - **RFE with Cross-Validation (CV):** Enhances robustness and generalisability by obtaining multiple sets of feature importance scores, which are then averaged to guide elimination.
  - **Dynamic RFE:** Allows for the elimination of more than one feature per iteration, offering greater flexibility.
  - **Enhanced RFE:** Modifies the elimination criteria by reinstating features if their removal significantly degrades model performance, ensuring that weak but potentially complementary features are not prematurely discarded.
  - **Local Search RFE:** Introduces an optimization layer that explores neighbouring feature subsets to refine the suggested feature set, addressing the greedy nature of the original RFE.
  - **Marginal Improvement-based RFE:** Determines feature importance by assessing how the elimination of each feature influences model performance when removed one at a time.

4.2.4 RFE Hybridized with Other Feature Selection or Dimension Reduction Methods

- **Description:** These variants combine RFE with other feature selection techniques or dimensionality reduction methods to accelerate the process or identify important features more effectively.
- **Examples:** Combining RFE with **Principal Component Analysis (PCA)** for initial dimensionality reduction, or incorporating **K-means clustering** to identify feature clusters from which RFE selects representative features. Other hybrids include combining SVM-RFE with the **Gini index (GI)** or a **chi-square test**.

4.3 Key Considerations

When writing research paper on RFE and its variants, it is crucial to understand the **trade-offs** associated with each approach. The choice of RFE variant should be guided by your specific research objectives, data characteristics, and practical constraints.

- **Predictive Performance vs. Interpretability and Computational Cost:**
  - **RF-RFE and XGBoost-RFE** often deliver strong predictive performance, particularly in capturing complex, non-linear relationships. However, they tend to retain larger feature subsets, which can reduce model interpretability, and they typically incur significantly higher computational costs.
  - **Enhanced RFE** offers a favourable balance, achieving substantial feature reduction with only a marginal loss in predictive accuracy and often demonstrating higher stability. This makes it particularly suitable when interpretability and efficiency are prioritised.
  - **RFE with local search operators** can offer slight improvements in predictive metrics but sometimes at the expense of computational time or feature selection stability.
- **Data Characteristics:** Consider the nature of your data (e.g., linearity, correlations, presence of outliers, dimensionality) when selecting a variant. For instance, if you anticipate complex, non-linear interactions, tree-based RFE variants might be more effective. If multicollinearity is a concern, some methods might be better at handling groups of correlated variables.
- **Application Context:** The emphasis on prediction accuracy, feature reduction, or interpretability varies across domains. For example, in healthcare, high recall (minimizing false negatives) might be paramount, while in educational policy, clear interpretability of influential factors might be more critical.

5 Methodological Advances in Recursive Feature Elimination:

Recursive Feature Elimination (RFE) has emerged as a powerful wrapper-based feature selection method, valued for its ability to handle high-dimensional data and produce interpretable models. Originally developed in healthcare analytics, its iterative backward elimination process effectively identifies influential features by leveraging the predictive power of an underlying machine learning (ML) algorithm. While the original implementation with Support Vector Machines (SVM-RFE) is common, numerous variants have been proposed to enhance its performance, robustness, and flexibility. This section provides a detailed examination of two significant categories of RFE variants: those integrated with tree-based algorithms and those that modify the core RFE process.

5.1 RFE Integrated with Tree-Based Algorithms

This class of variants integrates the RFE process with tree-based ML models, which are particularly adept at capturing complex, non-linear relationships among predictors. In these models, feature importance is typically assessed based on metrics such as impurity reduction or permutation scores.

5.1.1 Random Forest RFE (RF-RFE)

- **Mechanism:** RF-RFE employs Random Forest (RF) models to rank features. Its importance metric is often based on the decrease in predictive performance when the values of an individual feature are randomly permuted. This permutation-based approach makes RF-RFE robust to noise and highly effective at identifying features involved in complex, non-linear interactions within the data.
- **Performance and Trade-offs:**
  - Achieved highest predictive performance ($R^2 = 0.379$, stability = 0.966).
  - Retained a large number of features (108 of 116), lowering interpretability.
  - Computationally very expensive.
  - In healthcare tasks, tended to prioritise precision over recall, which is risky in clinical settings.
- **Suitability:** Recommended when predictive accuracy is the priority and computational resources are not a constraint. Not ideal for interpretable or time-sensitive analyses.

5.1.2 Decision Tree RFE (DT-RFE)

- **Mechanism:** This variant integrates RFE with Decision Trees (DT), where feature importance is assessed based on metrics like impurity reduction. It operates on the same principle as other tree-based variants but uses a single DT instead of an ensemble.

- **Performance and Trade-offs:**
  - Faster and simpler than RF-RFE.
  - Less accurate and stable due to reliance on a single tree.
  - Useful within ensemble frameworks (e.g., intrusion detection).
- **Suitability:** Good for exploratory analysis and interpretability. For robust prediction, ensemble methods like RF-RFE are preferred.

5.2 Modifications to the RFE Process

This category integrates advanced regularization techniques into the RFE framework to improve feature selection, particularly in challenging data scenarios like high dimensionality and multicollinearity. These methods leverage penalized models not just for their embedded feature ranking but also for their inherent statistical properties that lead to more robust and reliable feature subsets.

5.2.1 RFE with Elastic Net (RFE-ElasticNet)

- **Mechanism and Formulation:** This method uses the coefficients from an Elastic Net model to determine feature importance at each RFE step. Elastic Net is a penalized regression model designed as a compromise between Ridge and Lasso regression. Its objective is to minimize:

  $$\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2)$$

  The hyperparameter α acts as a mixing parameter ($0 \le \alpha \le 1$). When $\alpha = 1$, the model is equivalent to LASSO; when $\alpha = 0$, it is equivalent to Ridge. For values in between, it blends their properties. This dual penalty allows it to perform feature selection (like LASSO) while also handling correlated predictors gracefully (like Ridge).

- **The Grouping Effect:** A significant advantage of Elastic Net is its **grouping effect**. When a group of features are highly correlated, LASSO tends to arbitrarily select only one feature from the group while zeroing out the others. In contrast, the Ridge component of Elastic Net encourages the model to shrink the coefficients of correlated features together, effectively selecting or discarding them as a group. This leads to more stable and interpretable results in the presence of multicollinearity.

- **Performance and Trade-offs:**

  - By design, it excels at handling datasets with high multicollinearity. The grouping effect ensures that valuable, correlated predictors are not prematurely or arbitrarily eliminated during the RFE process.
  - It provides a more stable feature selection path than RFE-Lasso, especially when $p \gg n$ (more features than samples), as the Ridge component adds stability to the coefficient estimates.
  - The primary trade-off is the increased computational complexity. Effective implementation requires tuning two hyperparameters: the overall penalty strength (λ) and the L1/L2 mixing parameter (α). This typically necessitates a more intensive search, often using a nested cross-validation grid search, to find the optimal combination for the model at each elimination step.

- **Suitability and Application:** Ideal for high-dimensional data where feature correlation is common and expected. Its robustness makes it a superior choice in fields like genomics (where genes operate in correlated pathways), financial modeling (where economic indicators are often linked), and chemometrics.

5.2.2 RFE with Adaptive LASSO (RFE-AdaptiveLASSO)

Methodological Contribution: RFE with Adaptive Lasso

A key methodological contribution of this research project is the evaluation of a hybrid feature selection technique combining "Recursive Feature Elimination (RFE) with an Adaptive Lasso core model. While both RFE and Adaptive Lasso are powerful methods individually, their synergistic application is not widely documented in existing literature. This study, therefore, provides a novel performance benchmark and explores the viability of this combined approach, which leverages the iterative elimination strength of RFE with the intelligent, weighted penalty of Adaptive Lasso.

- **Mechanism and Formulation:** This is a sophisticated enhancement of LASSO that addresses one of its key limitations. The Adaptive LASSO applies unique, data-driven weights to the L1 penalty for each feature, modifying the penalty term to $\lambda \sum_{j=1}^{p} w_j |\beta_j|$. The weights are defined as:

$$w_j = \frac{1}{|\hat{\beta}_{j,initial}|^{\gamma}}$$

  where $\hat{\beta}_{j,initial}$ is a coefficient estimate from an initial, consistent model (often Ridge regression for stability), and $\gamma > 0$ is a tuning parameter. This mechanism heavily penalizes features with small initial coefficients (pushing them toward zero) while lightly penalizing features with large initial coefficients (protecting them from elimination).

- **The Oracle Property Explained:** The primary theoretical advantage of Adaptive LASSO is that it possesses the **oracle property**. In simple terms, an "oracle" model would know in advance exactly which features are truly important. The Adaptive LASSO, under certain conditions, can emulate this by simultaneously:

    1. Identifying the correct subset of true predictors with a probability approaching one as the sample size increases (selection consistency).
    2. Estimating the coefficients of the true predictors as accurately and unbiasedly as if the correct model were known from the start.

  This makes it statistically more powerful than the standard LASSO.

- **Performance and Trade-offs:**

    - It significantly reduces the estimation bias present in the standard LASSO, which can over-penalize important predictors that happen to have smaller effect sizes. This leads to more accurate coefficient estimates and a more reliable final feature set.
    - The two-stage process (requiring an initial model fit to generate weights, followed by the weighted LASSO fit within each RFE step) adds computational overhead and implementation complexity.
    - The quality of the final feature set is sensitive to the quality of the initial estimator used to generate the weights. Using a stable estimator like Ridge is crucial, especially with collinear data.

- **Suitability and Application:** Best suited for scenarios where achieving the most accurate and parsimonious model is paramount, and there is reason to believe that predictors have a wide range of effect sizes. It is highly valued in fields requiring high-confidence variable selection, such as biostatistics, clinical trial analysis, and econometrics.

6 Conclusion:

6.1 Conclusion based over Ridge Regression:

Fitting regularised models over model Boston dataset leads to the folowing reults obtained:

Figure 3 shows that for ridge regression models none among the set of 12 feature is eliminated from the model. addressing the minimum value of lambda corresponding to which minimum value of Mean-Squared Error has obtained.

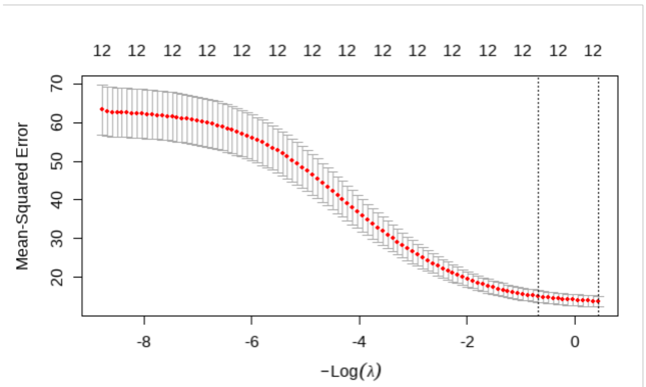


Figure 3: Cross-validation results for Ridge regression.

The cross-validation plot confirms that we've found the optimal penalty (lambda) to minimize prediction error, ensuring the model is accurate and from Fig. 4 The coefficient path plot shows this penalty shrinks the influence of all features, which creates a more stable and robust model that avoids overfitting. This confirms that predictors like the number of rooms (rm) and socio-economic status (lstat) are the most important factors in determining house prices.

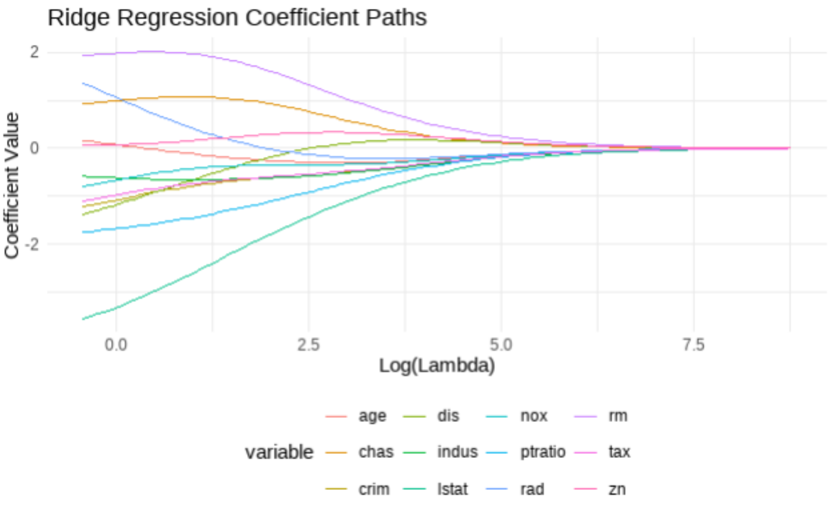


Figure 4: Coefficient path for Ridge regression.

6.2 Conclusion based over LASSO Regression:

Figure 5 shows that we've got a very low error once we have about 10-12 features in the LASSO model. The dotted line around "10" suggests that this is a great choiceâĂŤit gives us a simple, but powerful model that is almost as accurate as the most complex one. we've successfully found a sweet spot that balances accuracy and simplicity.

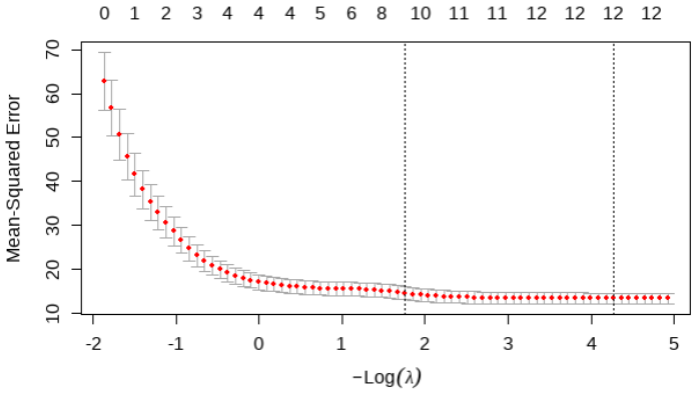


Figure 5: Cross-validation results for LASSO regression.

In Fig. 6, on the right side, the penalty is very high, and the model is forced to be simpleâĂŤall features are eliminated (their importance is zero). As we move to the left, the penalty gets weaker, and the most impactful features are allowed to "enter in model." This plot beautifully visualizes LASSO's key strength: it doesn't just reduce feature influence, it performs automatic feature selection by kicking the unimportant ones out entirely. can see that some features, like rm (number of rooms) and lstat (percentage lower status), are star players; their lines break away from zero early and become very influential

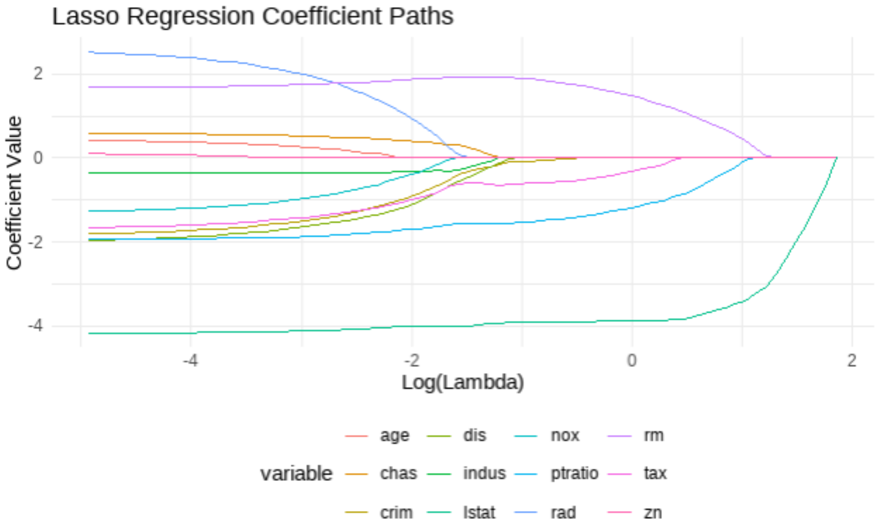


Figure 6: Coefficient path for LASSO regression.

6.3 Conclusion based on CV plot for Adaptive LASSO:

From Fig. 7 we can easily see that the error drops quickly as the model adds the first few, most powerful features. After that, the improvement becomes minimal as the curve flattens out. The first dotted line highlights a model with just 7 features. This is an excellent choice. It represents the simplest model that is statistically just as powerful as the more complex models with 10 or 11 features. It has successfully filtered out the noise and identified a small, potent set of variables. from the previous plots we can also conclude that the nature of flexibility is higher in adaptive lasso as compared to previous ones.

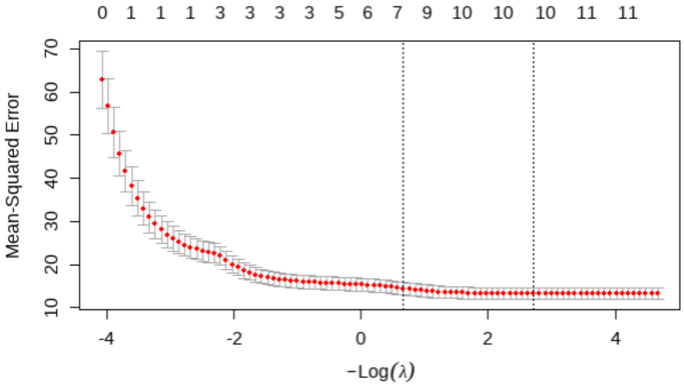


Figure 7: Cross-validation for Adaptive LASSO.

6.4 Results obtained for wrapper based RFE methods:

This analysis was conducted to identify the optimal number of predictive features for the Boston housing dataset and to compare the effectiveness of four different machine learning models. Figure 8 displays the results from a Recursive Feature Elimination (RFE) process. RFE works by systematically building models with progressively fewer features, removing the least important one at each step. This allows us to find the "sweet spot" where the model achieves the lowest prediction error (measured by Root Mean Squared Error, or RMSE) with the simplest possible feature set.

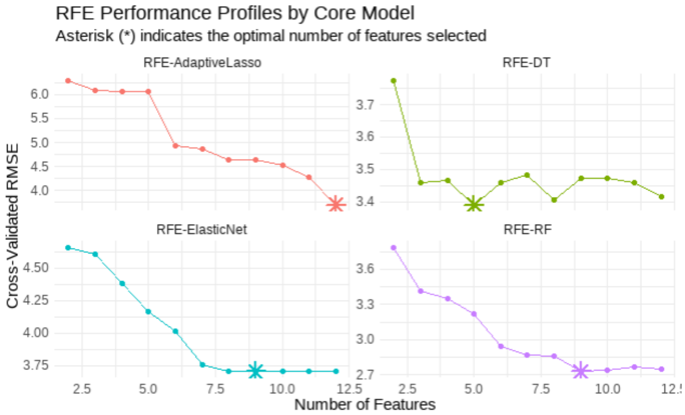
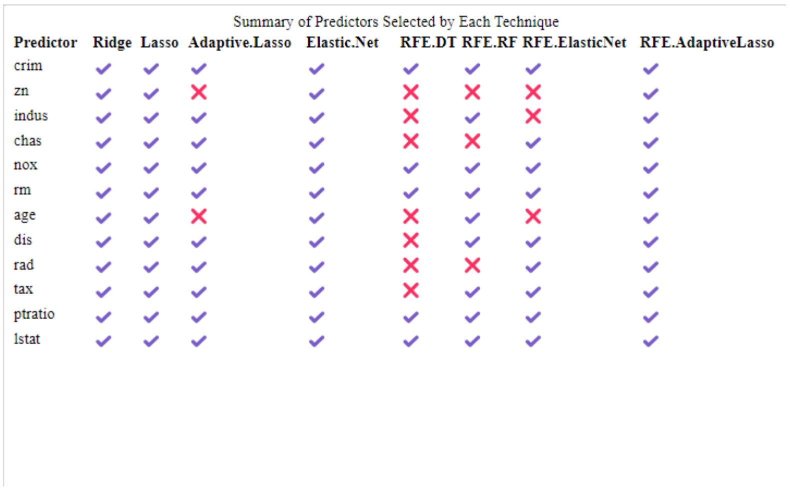


Figure 8: Recursive Feature Elimination (RFE) performance.

The optimal number of features for each model is marked with an asterisk (*). As we can see **Random Forest (RFE-RF)**: This model is the top performer among all the wrapper based models. It achieved the lowest overall RMSE of approximately 2.7 when using an optimal subset of 9 features. the issue arise with Random Forest (RFE-DT) is this model is highly flexible to interpretation because of selecting set of 5 features only, but looking at the RMSE, it ranked after the RFE-RF.

6.5 Table for selected Feature:

In Fig. 9, table provides a comprehensive view of the obtained feature selection results, summarizing which predictors were considered important (Yes) versus those that were eliminated (Cross) by each of the different feature selection techniques employed. This cross-model comparison is crucial for understanding the relative importance of each feature and for building confidence in our final model's structure.



Summary of Predictors Selected by Each Technique

Predictor	Ridge	Lasso	Adaptive.Lasso	Elastic.Net	RFE.DT	RFE.RF	RFE.ElasticNet	RFE.AdaptiveLasso
crim	✓	✓	✓	✓	✓	✓	✓	✓
zn	✓	✓	✗	✓	✗	✗	✗	✓
indus	✓	✓	✓	✓	✗	✓	✗	✓
chas	✓	✓	✓	✓	✗	✗	✓	✓
nox	✓	✓	✓	✓	✓	✓	✓	✓
rm	✓	✓	✓	✓	✓	✓	✓	✓
age	✓	✓	✗	✓	✗	✓	✗	✓
dis	✓	✓	✓	✓	✗	✓	✓	✓
rad	✓	✓	✓	✓	✗	✗	✓	✓
tax	✓	✓	✓	✓	✗	✓	✓	✓
ptratio	✓	✓	✓	✓	✓	✓	✓	✓
lstat	✓	✓	✓	✓	✓	✓	✓	✓

Figure 9: Table of features selected by different models.

Consistently Important Predictors: A core set of features were identified as important by nearly every analytical technique. These include *lstat, rm, ptratio, tax, nox, indus*, and *crim.* The strong consensus across diverse methods highlight that these features have fundamental importance in predicting Boston housing prices.

Contrasting Model Behaviors The table highlights the different philosophies of each model.

- **Ridge Regression** serves as a baseline, retaining all features as it is not designed for feature elimination.
- **Adaptive Lasso** proved to be more selective than the standard Lasso, dropping three predictors.
- **Recursive Feature Elimination (RFE)** methods were the most discerning. RFE with a **Decision Tree** was the most aggressive, creating the smallest model with only 5 features. The top-performing **RFE with Random Forest** identified a potent subset of 9 features.

6.6 Comparative Analysis of Regularized Regression Models

The following figure presents a multi-stage comparative study of models developed to predict Boston housing prices. The primary focus lies in evaluating the effectiveness of various shrinkage (regularization) methods, followed by a comprehensive performance ranking across all tested models.

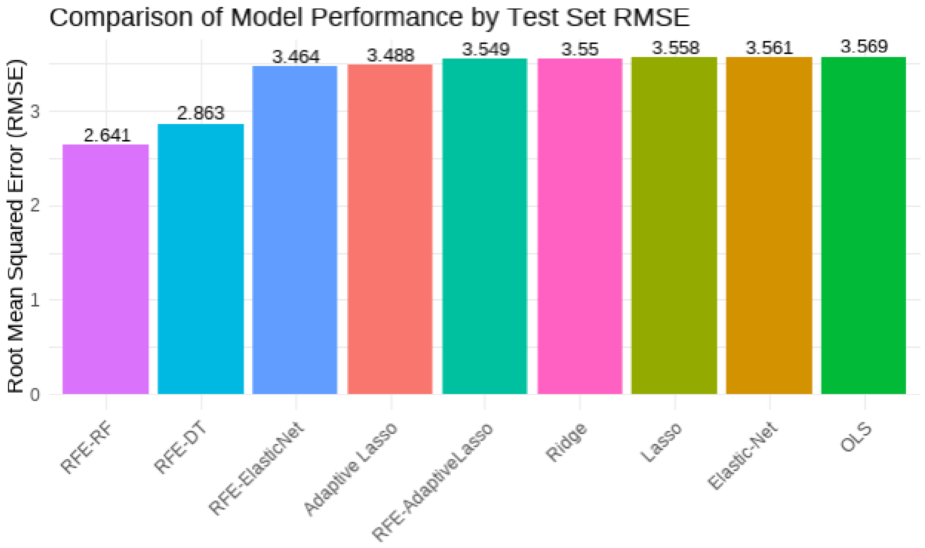


Figure 10: Bar plot addressing RMSE for each method.

6.6.1 Comparative Study of Core Shrinkage Methods:

This initial analysis focuses on four key shrinkage techniques designed to improve upon standard linear regression by preventing overfitting: **Ridge**, **Lasso**, **Adaptive Lasso**, and **Elastic Net**.

Ridge vs Lasso Both Ridge (RMSE: 3.55) and Lasso (RMSE: 3.558) offered a marginal but meaningful improvement over the baseline OLS model. While Ridge shrinks the influence of all features, Lasso can eliminate features entirely. However, for this dataset, the optimal Lasso model did not eliminate any feature which results in similar performance to Ridge.

Elastic Net As a mixture of Ridge and Lasso, Elastic Net (RMSE: 3.561) also performed within the same narrow range. From this we can conclude that for this dataset, combining the L1 and L2 penalties did not yield a significant advantage.

The Standout Performer: Adaptive Lasso The Adaptive Lasso model (RMSE: 3.488) was the clear winner among this group. Unlike standard Lasso, which applies a uniform penalty, Adaptive Lasso uses a weighted approach that intelligently penalizes coefficients based on their estimated importance. This enabled it to correctly identify and eliminate three less-important features (*zn*, *age*, and *dis*), as confirmed by the feature selection summary table. This refined selection resulted in a lower prediction error, making Adaptive Lasso the most effective among the standard shrinkage models.

Among the base regularization techniques, Adaptive Lasso demonstrates superior ability to produce a more accurate model through its advanced feature selection mechanism.

Interestingly, from Table 3 we can infer that some of the models that perform automatic feature selection, such as Adaptive Lasso and RFE-ElasticNet, shrink the coefficients of certain features like `zn`, `indus`, and `age` to exactly zero. This implies that these features contribute the least to the predictive power of the model and could potentially be excluded to create a simpler, more interpretable model without a significant loss in accuracy. Overall, the comparison demonstrates a general consensus among the models on the key drivers, while also highlighting the benefits of methods that can simplify the model by identifying and removing less relevant features. The most significant predictor appears to be `lstat` (the percentage of the population with lower status), which consistently shows a strong negative coefficient (ranging from -4.00 to -4.16). This indicates that as the lstat value increases, the predicted outcome value strongly decreases. Conversely, `rm` (the average number of rooms per dwelling) is a strong positive predictor, with all models assigning it a high positive coefficient. This suggests that properties with more rooms tend to have a significantly higher outcome value.

Table 3: Regression Coefficients for Various Linear Regularized Models

	Model						
Feature	**OLS**	**Ridge**	**Lasso**	**A-Lasso**[a]	**E-Net**[b]	**RFE-ENet**[c]	**RFE-ALasso**[d]
(Intercept)	87.5039	21.6179	21.6369	21.6308	21.6345	21.6382	21.6318
crim	–1.8080	–1.2305	–1.7748	–1.7368	–1.7845	–1.7746	–1.6085
zn	0.0790	0.0639	0.0838	0.0000	0.1070	0.0000	0.0641
indus	–0.0523	–0.5850	–0.3637	–0.3186	–0.3854	0.0000	–0.4350
chas	0.5954	0.0351	0.5784	0.5242	0.6240	0.5543	0.6742
nox	–11.2047	–0.7952	–1.2267	–1.0555	–1.2427	–1.2722	–1.0894
rm	2.4915	1.9404	1.6916	1.7853	1.6932	1.8163	1.7594
age	0.0156	0.1542	0.3892	0.0000	0.4124	0.0000	0.3305
dis	–4.9588	–1.3742	–1.9180	–1.9603	–1.9400	–1.9870	–1.7662
rad	3.4579	1.3571	2.4525	2.3702	2.4607	2.5783	2.1215
tax	–0.0099	–1.0931	–1.6152	–1.6405	–1.6164	–1.8797	–1.4460
ptratio	–16.7799	–1.7490	–1.9236	–1.9168	–1.9188	–1.9742	–1.8803
lstat	–8.0775	3.5531	–4.1615	–4.0270	–4.1454	–4.0050	–4.0356
RMSE	**3.5691**	**3.5500**	**3.5580**	**3.4880**	**3.5610**	**3.4640**	**3.5490**

[a] A-Lasso: Adaptive Lasso
[b] E-Net: Elastic-Net
[c] RFE-ENet: Recursive Feature Elimination with Elastic-Net
[d] RFE-ALasso: Recursive Feature Elimination with Adaptive Lasso

6.6.2 Analysis of RFE with Advanced Shrinkage Methods:

In this phase we've evaluated the performance of combining Recursive Feature Elimination (RFE) with modern shrinkage techniques, treating them as updated versions of traditional methods.

RFE with Elastic Net (RFE-ElasticNet) This combination proved highly effective. By iteratively removing weaker predictors at each step of model building, the RFE process allowed Elastic Net to focus on a potent subset of only nine features. This team up achieved an RMSE of 3.464, making it the best-performing linear model in the entire study. It outperformed all stand-alone shrinkage methods, highlighting the power of integrating iterative feature elimination with regularization.

RFE with Adaptive Lasso (RFE-AdaptiveLasso) However, this approach was less effective for this dataset because the resulting model (RMSE: 3.549) performed slightly worse than both RFE-ElasticNet and standard Adaptive Lasso. But in our study it is found that we haven't seen any of the research paper where this kind of integration had been used. Maybe for other datasets this model will play a significant role. According to the feature summary table, this method was overly cautiousâĂŤeliminating only one feature (*rad*)âĂŤand failed to deliver the accuracy gains achieved by its Elastic Net counterpart.

6.6.3 Comprehensive Study and Final Ranking:

A Methodological Comparison of Penalized Regression and Recursive Feature Elimination for Predictive Modeling

1. **Baseline Performance (OLS):** The Ordinary Least Squares model (RMSE: 3.569) served as the baseline, confirming the advantage of regularized methods.

2. **The Shrinkage Tier:** Standard penalized methods, including Lasso, Ridge, Adaptive LASSO and Elastic Net, offered consistent improvements over OLS. Within this tier, Adaptive Lasso and especially RFE-ElasticNet has resulted in notable accuracy gains due to effective feature selection.

3. **The Top Tier (Tree-Based Methods):** Even if the focus of this analysis was on linear shrinkage models, the tree-based approaches paired with RFEâĂŤparticularly RFE-DT (RMSE: 2.863) and RFE-RF (RMSE: 2.641)âĂŤproduced substantially lower errors than any other linear model.

References

C. M. Andersen and R. Bro. Variable selection in regression—a tutorial. *Journal of Chemometrics*, 2010.

Okan Bulut, Bin Tan, Elisabetta Mazzullo, and Ali Syed. Benchmarking variants of recursive feature elimination: Insights from predictive tasks in education and healthcare. *Information*, 16:476, 2025.

Loann David Denis Desboulets. A review on variable selection in regression analysis. *Econometrics*, 6:45, 2018.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, USA, 2nd edition, 2009.

Netti Herawati, Ameliana Wijayanti, Agus Sutrisno, Nusyirwan, and Misgiyati. The performance of ridge regression, lasso, and elastic-net in controlling multicollinearity: A simulation and application. *Journal of Modern Applied Statistical Methods*, 23, 2024.

Takahiro Kitano and Hisashi Noma. Ridge, lasso, and elastic-net estimations of the modified poisson and least-squares regressions for binary outcome data. *The Graduate Institute for Advanced Studies, The Graduate University for Advanced Studies*, 2024. Working Paper/Preprint. ORCID: https://orcid.org/0009-0001-6614-4336; ORCID: http://orcid.org/0000-0002-2520-9949.

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 5th edition, 2012.

H. M Nayem, Sinha Aziz, and B. M. Golam Kibria. Comparison among ordinary least squares, ridge, lasso, and elastic net estimators in the presence of outliers: Simulation and application. *International Journal of Statistical Sciences*, 24:25–48, 2024.

J. O. Ogutu, T. Schulz-Streeck, and H. P. Piepho. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6:S10, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58:267–288, 1996.

R codes can be accessible through : `https://github.com/niteenishere/Research-Project/blob/main/updated_research_codes.Rmd`